Learning Entity and Relation Representation for Low-Resource Medical Language Processing

A dissertation submitted in fulfillment of the requirements of the degree

Doctor of Philosophy

from the

Department of Language Science and Technology,

Saarland University

by

Saadullah Amin, B.Sc.

from the

Department of Multilinguality and Language Technology,

German Research Center for Artificial Intelligence (DFKI)

Saarbrücken, 2024

Saadullah Amin: *Learning Entity and Relation Representation for Low-Resource Medical Language Processing*, a dissertation submitted in fulfillment of the requirements of the degree Ph.D. in Computational Linguistics from the Department of Language Science and Technology at Saarland University.

Dean: Prof. Dr. Stephanie Haberzettl
Chairperson: Prof. Dr. Ingo Reich

First Reviewer: Prof. Dr. Günter Neumann
Second Reviewer: Prof. Dr. Josef van Genabith

University Professor: Prof. Dr. Vera Demberg
Academic Assistant: Dr. Cristina España i Bonet

Oral Defence: 05.06.2024
Location: Saarbrücken, Germany

To my parents
Fouzia & Amin

DECLARATION

I, Saadullah Amin, declare that:

- The work reported in the dissertation "Learning Entity and Relation Representation for Low-Resource Medical Language Processing" is my own work.
- Where I consulted previous work, proper reference to such previous work is provided in the dissertation.
- For joint publications resulting from the research presented in the dissertation, my contribution is stated clearly.
- No part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution.

Saarbrücken, Germany, June 2024

Saadullah Amin

ABSTRACT

Recent advancements in natural language processing have led to growing interest in critical domains such as legal, finance, and healthcare. In particular, medical language processing has emerged as a focus of its own. As the number of unstructured text and structured ontologies increases, applying information extraction techniques becomes an essential first step for downstream applications in healthcare. To partially meet these needs, this dissertation studies knowledge acquisition for medical language processing under real-world low-resource conditions. This includes limited-to-no labeled data, multilingualism, domain specificity, and missing knowledge. The focus is on the fundamental building blocks of information extraction, entities and relations, where the proposed methods derive representations from pre-trained language models for unstructured text and knowledge graph embedding models for structured data.

First, we consider entity-centric learning in the clinical domain, starting with multilingual and unsupervised concept extraction from text for semantic indexing to named entity transfer for privacy-preserving cross-lingual de-identification. We demonstrate the effectiveness of transfer learning with multilingual and domain-specific language models in supervised, unsupervised, and few-shot settings. In particular, we follow a pre-train and then fine-tune paradigm to achieve better performance compared to state-of-the-art neural architectures for concepts extraction from multilingual clinical texts. Whereas for unsupervised extraction, we propose a hybrid framework, Dense Phrase Matching, which combines embedding-based matching with concepts string matching, showing strong improvements on lexically rich texts, with further application to multilingual clinical texts. We then propose a Transformer based transfer learning framework, T2NER, that offers to bridge the gap between growing research in deep transformer models, NER transfer, and domain adaptation. We use T2NER for the task of identifying protected health information by empirically investigating the few-shot cross-lingual transfer property of multilingual BERT, which primarily has been the focus of zero-shot transfer, and propose an adaptation strategy that significantly improves clinical de-identification for code-mixed texts with few samples.

Second, we consider relation-centric learning in the biomedical domain, starting with distantly supervised relation extraction from text for knowledge base enrichment to multi-relational link prediction for discovering missing facts in the knowledge graph. We showcase the utility of scientific language models for relation extraction and efficient tensor factorization for knowledge graph completion. We first propose entity-enriched relation classification BERT for multi-instance learning, whereby knowledge sensitive data encoding scheme is introduced that significantly reduces noise in distant supervision. We then investigate existing broad-coverage biomedical relation extraction benchmarks to identify a notable shortcoming of overlapping training and test relationships which we address by introducing a more accurate benchmark MedDistant19. Lastly, we propose an efficient knowledge graph completion model, LowFER, that achieves on par or state-of-the-art on several datasets in general and biomedical domains. We show that LowFER's representation capacity is fully expressive to handle arbitrary relation types and its low-rank generalization of Tucker decomposition encapsulates existing models as special cases.

ZUSAMMENFASSUNG

Durch die Einführung elektronischer Patientenakten (EHR) können Krankenhäuser und klinische Einrichtungen auf große Mengen heterogener Patientendaten zugreifen. Diese Daten bestehen aus strukturierten Daten (Versicherungen, Abrechnungen und Laborergebnisse) und unstrukturierten Daten (Arztbriefe, Aufnahme-/Entlassungsdaten und Medikationsschritte). Unstrukturierte Texte sind von großer Bedeutung für die Anwendung von Methoden der Informationsextraktion, um relevante Konzepte, Entitäten, Ereignisse und Interaktionen zu lernen. Darüber hinaus verfügen wir über zahlreiche strukturierte taxonomische Ressourcen im medizinischen Bereich. Diese medizinischen Wissensquellen werden von Fachleuten kuratiert und können ein reichhaltiges Lernsignal liefern, bei dem die medizinischen Konzepte in eine hierarchische Struktur eingeordnet werden.

Da Text- und Wissensdatenbanken eine gemeinsame Schnittstelle für Anwendungen zur Informationsextraktion im Gesundheitswesen darstellen, spielen sie in dieser Dissertation als Eingabequellen eine zentrale Rolle. Da die medizinische Sprachverarbeitung darauf abzielt, aussagekräftige Informationen aus unstrukturiertem Text zu extrahieren und zu analysieren, um sie in strukturierte Daten umzuwandeln, ist ein häufig verwendetes Format das der Entität und der Beziehungen zwischen ihnen, welches hier im Mittelpunkt des Lernens von Repräsentationen steht. Insbesondere werden in dieser Dissertation die folgenden Bedingungen betrachtet, die sich aus der Verfügbarkeit von *geringen Ressourcen* ergeben:

- **Limitierte bis keine annotierte Daten:** Eine immer wiederkehrende Herausforderung von überwachten Lernmethoden ist die Verfügbarkeit von annotierten Daten, die sich in Bezug auf Zeit und Kosten für die medizinische Sprachverarbeitung noch verschärft, da für die Erstellung der Daten Experten benötigt werden.

- **Mehrsprachigkeit:** Trotz der Fortschritte im Bereich der Analyse von elektronischen Patientenakten konzentrieren sich die meisten Forschungsarbeiten bisher auf die englische Sprache. Aufgrund zu geringer oder kaum verfügbarer Daten in anderen Sprachen, stellt die Erforschung und Entwicklung mehrsprachiger Gesundheitssysteme immer noch eine sehr große Herausforderung dar.

- **Domänenspezifität:** Eine unmittelbare Folge der Arbeit mit medizinischer Sprachverarbeitung ist die Spezifität der Domäne, die bei der Arbeit in Teilgebieten wie Krebs oder Schlaganfall noch ausgeprägter ist.

- **Fehlendes Wissen:** Mit dem kontinuierlichen Wachstum der biomedizinischen Literatur besteht ein ständiger Bedarf an der Erstellung, Anreicherung und Aktualisierung der Wissensdatenbanken, um fehlende Fakten zu entdecken und neue hinzuzufügen.

Was die Modellierung betrifft, so verwenden wir in erster Linie *vortrainierte Sprachmodelle* zur Darstellung unstrukturierter Texte und *Modelle für Einbettungen von Wissensgraphen* zur Darstellung strukturierter multirelationaler Daten. In den nächsten Abschnitten geben wir einen Überblick über jeden Teil der Arbeit, der in entsprechende Kapitel unterteilt ist, einschließlich der relevanten Forschungsfragen und Beiträge.

Teil I: Entitätszentriertes Lernen

In diesem Teil konzentrieren wir uns auf das Lernen von Repräsentationen für Entitäten, die die Grundlage für das relationale Lernen bilden. Das Auffinden von Entitäten umfasst die Erkennung, Verknüpfung, Typisierung und deren Abgleich, wobei wir uns hauptsächlich auf die Entdeckung von Konzepten und Entitäten aus dem Text für nachgelagerte Anwendungen oder den Aufbau neuer Taxonomien und Wissensdatenbanken konzentrieren.

Kapitel 2: Konzeptextraktion

Wie bereits erwähnt, haben sich die meisten Forschungsarbeiten trotz des digitalen Fortschritts im Gesundheitswesen auf die englische Sprache konzentriert, mit nur wenigen neueren Möglichkeiten für andere Sprachen (Névéol et al., 2018). Wir gehen dieses Problem an, indem wir uns auf die Aufgabe der Extraktion von klinischen Konzepten aus mehrsprachigen elektronischen Patientendaten (EHRs - Electronic Health Records) konzentrieren. Die hier betrachteten Konzepte sind eine Teilmenge einer gegebenen Wissensbasis und einem relevanten Teilgebiet, z. B. krebsbezogene Konzepte aus ICD-10 oder schlaganfallbezogene Konzepte aus SNOMED CT (Donnelly et al., 2006). Ein *Konzept* ist definiert als eine semantische Texteinheit, die im Text explizit angegeben werden kann oder auch nicht, sich aber auf einen zugrunde liegenden Begriff aus einer Wissensbasis bezieht, wobei eine explizit erwähnte benannte Entität als Sonderfall eines Konzepts betrachtet wird. Sobald wir klinisch relevante Konzepte aus dem Text extrahiert haben, kann ein impliziter Abgleich zwischen Text und Wissensbasis für die semantische Indizierung von Texten für die Suche und Wissensentdeckung verwendet werden.

Um den Abgleich zu erlernen, gehen wir von zwei Szenarien aus, die von der Verfügbarkeit von annotierten Daten abhängen. Wenn wir eine Sammlung von *(Text, Konzepte)*-Paaren haben, reduziert sich das Problem auf eine überwachte Multi-Label-Konzeptklassifikation. Traditionelle Methoden zur Konzeptextraktion haben von Feature-Engineering und dem Nachschlagen in domänenspezifischen Wörterbüchern profitiert (Bounaama and Amine, 2018; Gobeill and Ruch, 2018), wobei neuronale Netzwerke den Stand der Technik für englische klinische Texte verbessert haben (Baumel et al., 2018). Im Gegensatz dazu hat das Vor-Trainieren tiefer neuronaler Modelle (Qiu et al., 2020) viele Sprachverstehensaufgaben deutlich verbessert, einschließlich der Feinabstimmung bei Problemen mit geringen Ressourcen. In Kapitel 2 untersuchen wir zunächst die Auswirkungen des Transferlernens in einer überwachten Umgebung für deutsche klinische Texte. Wir betrachten eine umfassende Reihe von neuronalen Baselines, die modernste Textklassifizierer auf der Basis von Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM) und Gated Recurrent Unit (GRU) Kodierern umfassen, während wir neuartige Beschreibungs- und Self-Attention-Modelle vorschlagen. Um die durch die Behandlung von Deutschen Texten entstehenden Lücken teilweise aufzuzeigen, führen wir außerdem die automatische maschinelle Übersetzung ein, um Deutsche Texte ins Englische zu übersetzen, und untersuchen die durch vortrainierte statische und kontextualisierte Einbettungen verursachten Domänenlücken. Wir betrachten folgende vortrainierte Sprachmodelle: multilinguale bidirektionale Encoder-Repräsentationen von Transformatoren (mBERT) (Devlin et al., 2019a) für Deutsch und seine domänenspezifische englische Variante BioBERT (Lee et al., 2020) unter dem Paradigma *vortrainieren, dann feinabstimmen*.

Unsere experimentelle Auswertung ergibt einen F1-Score von 73%, was die Effektivität der impliziten Konzeptrepräsentation der BERT-Text-Encoder unterstreicht und gleichzeitig den Aufwand verringert, domänenspezifische Anpassungen durchzuführen, etwa durch spezielle Wörterbücher, Feature-Engineering oder durch die Spezifikation komplexer neuronaler Architekturen. Wir haben auch festgestellt, dass die automatische maschinelle Übersetzung die absoluten Modellwerte im Durchschnitt um 6% verbessert hat, was die Sprachlücke verdeutlicht. Schließlich wirkten sich domänenspezifische Worteinbettungen stärker auf die statischen Modelle aus als auf die kontextbezogenen. Diese Modelle und Ergebnisse sind in Amin et al. (2019) veröffentlicht.

Das Erlernen des Abgleichs (engl. Alignment) als unüberwachtes Problem ähnelt jedoch einer "Zero-Shot" Klassifizierungsaufgabe, bei der wir nur eine Sammlung von *Texten* und einen vordefinierten Satz von *Begriffen* haben. Trotz aktuellster Forschungsergebnisse erfordert die Feinabstimmung manuell annotierte Daten in Form von *(Text, Konzepte)*-Paaren, die für viele Sprachen und klinische Teilbereiche nicht ohne weiteres verfügbar sind. Inspiriert von aktuellen Arbeiten im Bereich des Zero-Shot Entity Linking (ZSEL) (Wu et al., 2020a), schlagen wir einen hybriden Ansatz für die unüberwachte Konzeptextraktion vor, der Wörterbuch- und Kontextabgleiche verwendet, um die genannten Kandidaten zu generieren. Für die Generierung von Kandidaten verwenden wir Modelle zur unüberwachten Keyphrase Extraction (KPE) und Contextual Span Detection (CSD). Die extrahierten Bereiche werden mithilfe von CPMerge (Okazaki and Tsujii, 2010) und Dense Nearest Neighbour Search mit FAISS (Johnson et al., 2019) abgeglichen, wobei die Einbettungen mit einem kontextuellen satzbasierten Paraphrasen-Modell berechnet werden (Reimers and Gurevych, 2019). Schließlich wenden wir eine einfache Filterschwelle an, um die Teilmenge der Konzepte zu erhalten. Wir evaluieren unseren Ansatz an einem englischsprachigen Datensatz für Arzneimittelrezensionen und erreichen dabei einen beträchtlichen Leistungszuwachs von 12% F1-Score im Vergleich zu QuickUMLS (Soldaini and Goharian, 2016), und verwenden ihn in einem unbeaufsichtigten mehrsprachigen Setup für den deutschen klinischen Text wie im überwachten Fall, wobei wir einen F1-Score von 11% erzielen.

Kapitel 3: Erkennung von Entitäten

Ein Sonderfall eines Konzepts ist eine benannte Entität, die in biomedizinischer Literatur oder klinischen Berichten vorkommen, wie z. B. der Name eines medizinischen Gerätes (*MRI*), einer Krankheit (*Schlaganfall*) oder eines Proteins (*PEX-13*). Die Extraktion von benannten Entitäten aus Text und die anschließende Typisierung der Entitäten ermöglicht den Aufbau einer Wissensbasis mit einer sauberen Taxonomie. Die Erstellung umfangreicher annotierter Korpora ist jedoch eine Herausforderung, weshalb wir häufig auf Methoden des Transferlernens zurückgreifen. Wie in Kapitel 2 gesehen, ermöglichen vortrainierte Sprachmodelle zwar einen effektiven Konzepttransfer, ohne jedoch eine spezielle Experimentierumgebung für die Erkennung von benannten Entitäten (NER) bereitzustellen. Gleichzeitig hat die Forschung auf dem Gebiet des NER-Transfers in den letzten Jahren erhebliche Fortschritte gemacht, und es wurden viele neue Verfahren vorgeschlagen.

In Kapitel 3 stellen wir zunächst ein allgemeines Transformer-basiertes Transfer-Learning-Framework für die Named Entity Recognition (T2NER) vor, das in PyTorch (Paszke et al., 2019) für die Aufgabe der Entitätenerkennung mit tiefen Transformatormodellen entwickelt wurde, die traditionell von LSTM-Netzwerken profitiert hat. Das Framework basiert auf der Transformer-Bibliothek (Wolf et al., 2020) und unterstützt verschiedene Transfer-Learning-Szenarien, von sequentiellem Transfer bis hin zu Domänenanpassung, Multi-Task-Learning und semi-supervised Learning. T2NER zielt darauf ab, die Lücke zwischen den algorithmischen Fortschritten in diesen Bereichen zu schließen, indem es diese mit dem Stand der Forschung bei Transformer-Modellen kombiniert, um eine einheitliche, leicht erweiterbare Plattform zu bieten. Es kann für die Transfer-Learning-Forschung in der ressourcenarmen NER und für reale Anwendungen, wie die biomedizinische NER, verwendet werden. Das Framework ist zusammen mit seinen Konstruktionsprinzipien und der Systembeschreibung in Amin and Neumann (2021) veröffentlicht.

Neben biomedizinischen Konzepten und Entitäten enthalten klinische Texte in der Regel auch geschützte Gesundheitsinformationen (PHI - Patient Health Information), was ein Risiko für die Identifizierung der Patienten darstellt , wenn diese gegenüber Informationsextraktionswerkzeugen für nachgelagerte Anwendungen offengelegt werden. Daher ist es von entscheidender Bedeutung, solche sensiblen Daten zu entfernen, um eine ethische und sicherheitsrelevante natürlich-sprachliche Verarbeitung gemäß der GDPR-Verordnung zu gewährleisten. Bestehende Arbeiten zur De-Identifizierung (wo es als ein spezielles NER-Problem betrachtet wird) beruhen auf der Verwendung großer annotierter Korpora in englischer Sprache (Stubbs et al., 2017; Stubbs and Uzuner, 2015), was für reale mehrsprachige Umgebungen in der Regel ungeeignet ist. Vorgefertigte Sprachmodelle besitzen ein großes Potenzial für den sprachenübergreifenden Transfer in ressourcenarmen Umgebungen, einschließlich NER. Allerdings ist das Modellverhalten bezüglich einer "Few-shot" sprachübergreifenden NER-Übertragung mit dem Potenzial für eine domänenspezifische klinische De-Identifizierung unklar.

Um diese Frage zu untersuchen, zeigen wir empirisch die *few-shot* sprachenübergreifende Transfereigenschaft von mBERT für NER und setzen diese in einer realen ressourcenarmen Anwendung ein, wo zweisprachige (spanisch-katalanisch) klinische Notizen im Schlaganfallbereich de-identifizert werden müssen. Wir *(a)* schlagen eine optimale sprachübergreifende "Few Shot" Transferstrategie vor, *(b)* annotieren ein Zielentwicklungsset und *(c)* konstruieren einen annotierten "Few shot" Zielkorpus für effektives sprachübergreifendes Transferlernen. Das Entwicklungsset wird für die Auswahl von Modellen mit wenigen Stichproben eingesetzt, bei denen wir nur einige hundert annotierte Beispiele für das Training verwenden. Unser Modell verbessert den Zero-Shot-F1-Score von 73,7% auf 91,2% im Gold-Evaluierungsset, wenn mBERT aus dem synthetischen MEDDOCAN-Korpus (Marimon et al., 2019) in Spanisch mit unserem sprachübergreifenden "Few-shot" Zielkorpus angepasst wird. Bei der Verallgemeinerung auf einen Testsatz außerhalb der Stichprobe erreicht das beste Modell einen F1-Score von 97,2% bei einer manuell durchgeführten Auswertung. Diese Ergebnisse sind in Amin et al. (2022b) veröffentlicht.

Teil II: Relationen-Zentriertes Lernen

In diesem Teil gehen wir davon aus, dass wir eine Wissensbasis (Knowledge Base - KB) mit einer sauberen und aussagekräftigen Taxonomie semantischer Typen (oder Klassen) haben und dass diese Typen mit einem umfassenden Satz normalisierter (d.h. eindeutig identifizierter) Entitäten befüllt sind. Eine solche KB kann mit den in Teil I erörterten entitätszentrierten Lernansätzen gegeben oder teilweise befüllt werden.

Kapitel 4: Relationsextraktion

Computer können nur begrenzt Wissen generieren, da das gesamte Faktenwissen über unsere Welt von Menschen geschaffen und in Enzyklopädien, wissenschaftlichen Publikationen, Büchern oder Tagesnachrichten dokumentiert wird (Weikum et al., 2021). Relationsextraktion (RE) ist eine solche Aufgabe der Wissensentdeckung, die darauf abzielt, Interaktionen zwischen Entitäten aus Texten zu lernen, um strukturiertes Wissen in Form von (Subjekt, Prädikat, Objekt) (SPO)-Tripeln auszugeben, die entweder eine bestehende, von Menschen kuratierte KB bereichern oder halbautomatisch eine neue aufbauen können.

Im biomedizinischen Bereich ist dies jedoch eine Herausforderung, da es an annotierten Daten mangelt und die Kosten für eine Annotation hoch sind, da hierfür Experten benötigt werden. Distant Supervision (DS) wird häufig verwendet, um den Mangel an annotierten Daten zu beheben, indem Beziehungen zwischen der Wissensbasis und Rohtexten automatisch berechnet werden (Mintz et al., 2009). Eine solche Pipeline ist anfällig für Rauschen aufgrund einer großen Anzahl falsch positiver Ergebnisse, wobei frühere Arbeiten von Dai et al. (2019) ein Framework für gegenseitiges Lernen von Han et al. (2018a) mit biomedizinischem Text und Knowledge Graph (KG) unter Verwendung von Hilfsaufgaben, einschließlich Knowledge Graph Completion (KGC) und Klassifizierung von Entitätstypen, erweiterten. Sie zeigten, dass die Verwendung von "Attention-Mechanismen" mit einem Wissensgraphen bei der Entfernung von Rauschen mit einem Piece-wise Convolutional Neural Network (PCNN) Satz-Encoder helfen kann. In Teil I haben wir gezeigt, dass vortrainierte Sprachmodelle und ihre domänenspezifischen Varianten effektive Entitätsrepräsentationen mit Transferlernen liefern. Darauf aufbauend und inspiriert von Dai et al. (2019), die eine Wissensbasis verwenden, und von Alt et al. (2019), die das OpenAI GPT für Bag-level Multi-Instance Learning (Bag-Level-MIL) (Surdeanu et al., 2012) für DSRE (Distant Supervision Relation Extraction) für die allgemeinen Domäne fein-justierten, konzentrieren wir uns auf wissensbasierte DSRE in der biomedizinischern Domäne (Bio-DSRE).

In Kapitel 4 schlagen wir zunächst vor, das Rauschen von DS zu reduzieren, indem wir ein mit Entitäten angereichertes Klassifizierungsmodell für Relationen (RBERT) durch ein einfaches KB-kontrolliertes Kodierungsschema auf das Problem der Bag-Level-MIL für DSRE erweitern. Die Kodierung identifiziert die Head- und Tail-Rolle von Entitäten aus dem Wissensgraphen (Knowledge Graph - KG) durch Markierung und durchläuft tiefe Transformator-Schichten mit Attention. Die daraus resultierende textuelle Repräsentation können in eine Relation-Repräsentation zusammengeführt werden, die implizit das gegenseitige Lernen mit Text und KG datengesteuert kodiert. Das vorgeschlagene Model MIL-RBERT reduziert das Rauschen signifikant und erreicht eine State-of-the-Art-Leistung mit 68,4% AUC und 64,9% F1-Score, mit einem absoluten 7% P@2k-Gewinn im Vergleich zu Dai et al. (2019). Die Datenpipeline und das Modell sind in Amin et al. (2020a) veröffentlicht.

Neben dem Rauschen besteht die zweite große Herausforderung in der Skalierung auf eine große Anzahl von Konzepten, um eine breite Abdeckung zu erreichen. Die bestehenden Arbeiten im Bereich der breit-angelegten Biomedical Distant Supervision Relation Extraction (Bio-DSRE) liefern jedoch sehr genaue Ergebnisse (Amin et al., 2020a; Hogan et al., 2021; Xing et al., 2020), einschließlich unseres Beitrags aus dem letzten Abschnitt, was uns veranlasst hat, die verwendeten Benchmarks genauer zu untersuchen. Außerdem veranlasst uns das Fehlen einer gründlichen Evaluierung von domänenspezifischen Sprachmodellen für die biomedizinische Relationsextraktion dazu, die domänenspezifischen Sprachmodelle gründlich zu untersuchen.

Wir beginnen mit der Untersuchung bestehender Benchmarks auf mögliche Trainings- und Testlecks von KG-Tripeln und finden signifikante Überschneidungen von 26% bis 86%. Solche Lecks wirken sich auf die Leistung des Modells aus, da es durch einfaches Auswendiglernen der Trainingsrelationen eine höhere Punktzahl erreichen kann, anstatt auf neue, bisher unbekannte Relationen zu generalisieren. Als Ursachen für diese Probleme erkennen wir die Normalisierung der Textform von Begriffserwähnungen auf ihre eindeutigen Bezeichner und die unsachgemäße Behandlung inverser Beziehungen. Im Gegensatz dazu gibt es genauere Benchmarks, die sich jedoch auf engere Arten von Interaktionen konzentrieren. Um die Probleme mit der breiten Abdeckung von Benchmarks zu lindern und diese Lücke zu schließen, stellen wir einen neuen Benchmark MedDistant19 vor, der seinen Wissensgraphen aus der weit verbreiteten Gesundheitsontologie SNOMED CT (Chang et al., 2020) bezieht. Mit dem Erfolg von domänenspezifischen vortrainierten Sprachmodellen (Gu et al., 2021) und inspiriert von bestehenden gründlichen Studien zur Relationsextraktion in der allgemeinen Domäne (Alt et al., 2020; Gao et al., 2021a; Peng et al., 2020), führen wir eine umfassende Evaluierung mit MedDistant19 für die biomedizinische Domäne durch. Der Benchmark und die Ergebnisse sind in Amin et al. (2022a) veröffentlicht.

Kapitel 5: Vervollständigung von Wissensgraphen

Trotz manuell oder halbautomatisch erstellter KBs sind die meisten von Natur aus unvollständig und enthalten nur eine begrenzte Anzahl von beobachteten Fakten aus dem allgemeinen oder biomedizinischen Bereich, die als strukturierte Beziehungen zwischen Entitäten dargestellt werden. Um dieses Problem teilweise zu lösen, konzentrieren wir uns auf eine weit verbreitete Aufgabe beim statistischen relationalen Lernen, nämlich die der Vorhersage von Verbindungen bzw. der Vervollständigung von Wissensgraphen, wobei wir von einer unvollständigen, von Menschen erstellten KB ausgehen. Als ein Ergebnis erhalten wir eine gering-dimensionale multirelationale graphische Darstellung von Entitäten und ihren Beziehungen, die zur Aufdeckung fehlender Fakten (Vollständigkeit) oder zur Überprüfung der Gültigkeit von Fakten (Verifizierung) verwendet werden kann. Zur Lösung dieses Problems wurden bisher sowohl lineare als auch nichtlineare Einbettungsmodelle vorgeschlagen. Bilineare Modelle sind aussagekräftig, neigen aber zur Überanpassung und können zu einem quadratischen Wachstum der Parameter bei der Anzahl der Beziehungen führen, was uns veranlasst hat, diese Richtung zu untersuchen.

Da einfachere Modelle zum Standard geworden sind, mit bestimmten Einschränkungen für die bilineare Abbildung als Parameter für Relationen, schlagen wir in Kapitel 5 ein faktorisiertes bilineares Pooling-Modell vor, das üblicherweise beim multimodalen Lernen (Yu et al., 2017) verwendet wird, um eine bessere Fusion von Entitäten und Relationen zu erreichen, was zu einem effizienten und constraint-freien Tensor Faktorisierungsmodell, LowFER, führt. Empirisch evaluieren wir Teilmengen von Freebase (Bollacker et al., 2008), WordNet (Miller, 1992) und YAGO (Rebele et al., 2016) für die allgemeine Domäne und UMLS (Bodenreider, 2004) und SNOMED CT (Donnelly et al., 2006) für die biomedizinischen Domäne und erreichen dabei eine Leistung, die dem Stand der Forschung entspricht.

Aus formaler Sicht ist eine wichtige theoretische Eigenschaft von KGC-Modellen ihre volle Ausdrucksfähigkeit. Ein vollständig ausdrucksstarkes Modell kann Beziehungen jeden Typs darstellen, einschließlich symmetrischer, asymmetrischer, reflexiver und transitiver Beziehungen, um nur einige zu nennen. Modelle wie RESCAL, HolE, ComplEx, SimplE und TuckER haben gezeigt, dass sie voll ausdrucksstark sind (Balažević et al., 2019b; Kazemi and Poole, 2018; Trouillon and Nickel, 2017; Wang et al., 2018b). Darüber hinaus wurde von Kazemi and Poole (2018) gezeigt, dass RESCAL, DistMult, ComplEx und SimplE zu einer Familie von bilinearen Modellen mit unterschiedlichen Mengen von Constraints gehören. Später stellte Balažević et al. (2019b) fest, dass TuckER alle diese Modelle als Spezialfälle verallgemeinert. Diese theoretischen Untersuchungen helfen uns auch, LowFER formal besser zu verstehen.

Wir beweisen, dass LowFER voll aussagekräftig ist, indem wir Schranken für die Dimensionalität der Einbettung und den Rang der Faktorisierung angeben. Es verallgemeinert das auf der Tucker-Zerlegung basierende TuckER-Modell (Balažević et al., 2019b) als eine effiziente Approximation mit niedrigem Rang (Low-Rank Approximation), ohne die Leistung wesentlich zu beeinträchtigen. Aufgrund der Low-Rank-Approximation kann die Modellkomplexität durch den Faktorisierungsrang kontrolliert werden, wodurch das mögliche kubische Wachstum von TuckER vermieden wird. Bei extrem niedrigen Rängen bewahrt LowFER die Leistung und bleibt dabei parametereffizient. Das Modell und die Ergebnisse sind in Amin et al. (2020b) veröffentlicht.

PUBLICATIONS

The dissertation is composed of the following peer-reviewed articles:

Saadullah Amin, Günter Neumann, Katherine Ann Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted (2019). MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In *Proceedings of the 20th Conference and Labs of the Evaluation Forum (Working Notes)* (pp. 1–15). Lugano, Switzerland. CEUR Workshop Proceedings.

`CLEF’19`
Concept Extraction
Chapter 2

Saadullah Amin, Stalin Varanasi, Katherine Ann Dunfield & Günter Neumann (2020). LowFER: Low-rank Bilinear Pooling for Link Prediction. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 257-268). Online. Proceedings of Machine Learning Research.

`ICML’20`
Knowledge Graph Completion
Chapter 5

Saadullah Amin*, Katherine Ann Dunfield*, Anna Vechkaeva & Günter Neumann (2020). A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing* (pp. 187-194). Online. Association for Computational Linguistics.

`BioNLP’20`
Relation Extraction
Chapter 4

Saadullah Amin & Günter Neumann (2021). T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 212-220). Online. Association for Computational Linguistics.

`EACL’21`
Named Entity Recognition
Chapter 3

Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Kelly Wixted, Alejandro García-Rudolph, Catalina Martínez-Costa & Günter Neumann (2022). Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts. In *Proceedings of the 21st Workshop on Biomedical Language Processing* (pp. 200-211). Dublin, Ireland. Association for Computational Linguistics.

`BioNLP’22`
Named Entity Recognition
Chapter 3

Saadullah Amin*, Pasqaule Minervini*, David Chang, Pontus Stenetorp & Günter Neumann (2022). MedDistant19: Towards an Accurate Benchmark for Broad-Coverage Biomedical Relation Extraction. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2259-2277). Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

`COLING’22`
Relation Extraction
Chapter 4

* *The authors contributed equally.*

ACKNOWLEDGEMENTS

I am grateful to Günter for providing the freedom to conduct this research in an independent way and for his time, support, feedback, and discussions as supervisor. It instilled a certain degree of confidence in me that resulted in gradual scientific growth.

I am grateful to Josef for giving me the opportunity to conduct this work at his Multilinguality and Language Technology (MLT) department at DFKI. I sincerely appreciate his continued support to carry out the research and for his time to provide feedback.

As a colleague and friend, I thank Stalin for the support, discussions, brainstorming sessions, and fun talks. It helped me with a gentle start to this journey while transitioning to the new work-life environment in Saarbrücken. I am thankful to Koel for her unwavering support, encouragement, and countless talks that helped me gradually see the end of this journey. I found a sound Ph.D. batchmate and a valuable friend in her. I am thankful to Noam for his friendship, unconditional help, and ample support that guided me towards this research programme. I am grateful to be reached out following his recommendation as it stepped me closer to my academic goals.

I am thankful to Sheheryaar for the inspiring intellectual exchanges shared over the years. I am thankful for the support, talks, and friendship Kathryn offered. I am grateful to Katie Ann and Anna for always being a fun bunch to talk to and work with, and to Iza for her kind help and support. I am also thankful to Tom for his constant eagerness to help and to Dominik for the pragmatic research talks.

During this time, I had the opportunity to work with and learn from several bright students including Kathryn, Jannis, Noon, Katie Ann, Anna, Morgan, and Mei. Also at the MLT Lab, I am thankful to several colleagues including Stephan, Corinna, Simon, and Lucia for keeping the administrative issues at bay. For research collaborations, I am grateful to Pasquale, Pontus, David, Cati, and Alejandro. In the industry, I am thankful to Jo for providing me with a timely opportunity and openly supporting me towards my ambitions. Furthermore, I thank my former colleagues Bala, Adnan, and Todd for the interesting roles offered and to my academic referees Amir, Awais, and Murtaza.

I am humbly grateful for the unparalleled support, love, prayers, and kindness Ammi and Abbu have offered. I am *only* able to fulfill my ambitions because of their hardships and innumerable sacrifices. I am forever thankful to them for putting me on an educational path and for the values taught that helped me weather the hardest times. I am grateful to Baji, Sid, Bhai Jaan, Saif, and Assad for completing the support block in the family. I greatly appreciate my family's patient presence while I focused on my goals. I have been lucky to discover mindfulness practice that helped me maintain a healthy state of mind and provided clarity when I needed the most.

I am thankful for the generous funding from the EU project Precise4Q (777107) and partial funding from the BMBF through the projects DEEPLEE (01IW17001) and CoRA4NLP (01IW20010). I am grateful to all the reviewers and the organizers of the scientific conferences in which the publications of this dissertation have appeared. I am thankful to the open-source communities for making this a more accessible pathway. Lastly, I am thankful for the computing resources provided by the DFKI. Without these, this research would not have been possible for practical reasons.

CONTENTS

1 INTRODUCTION

We will start the dissertation by providing a conceptual overview of the work, along with examples. We introduce each Chapter in terms of motivation, problem, research questions, and contributions. We use two sources of knowledge as input, text as unstructured data and knowledge base as structured data. These input sources provide a commonly used setting for medical language processing, but other modalities exist, such as images, videos, audio, and sensor data. However, they are considered out of the thesis' scope. We focus on two medical domains, biomedical and clinical, and where applicable, we discuss the methods both in a more general setting and for a specific sub-domain. We study four low-resource dimensions: labeled data, domain specificity, multilingualism, and missing knowledge, and split the work into entity-centric and relation-centric learning as shown in Figure 1. Our goal in this Chapter is to highlight the dissertation's overall scope and organization.

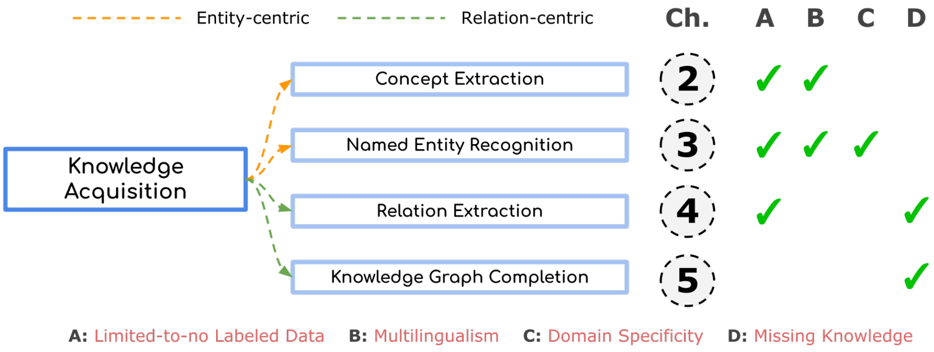
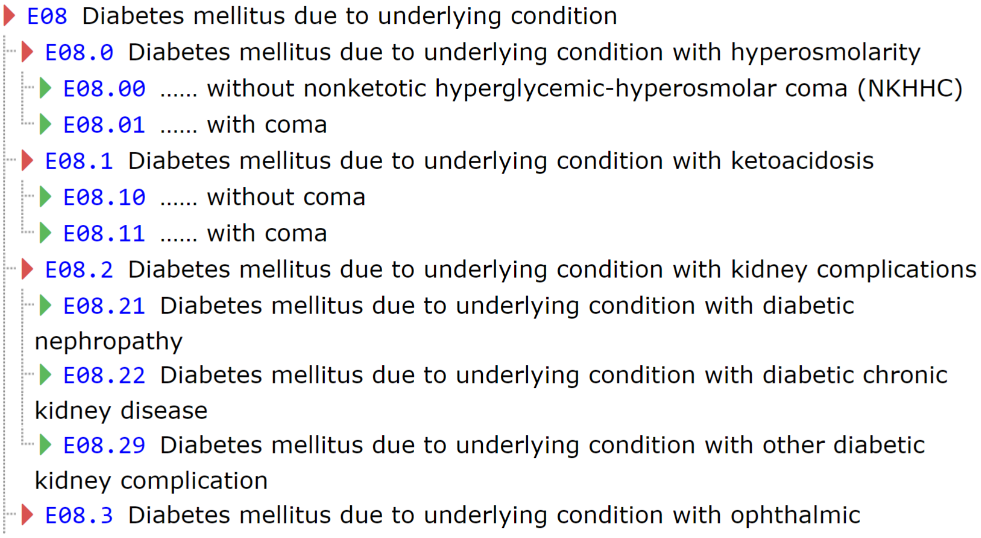


Figure 1: Knowledge Acquisition for Low-Resource Medical Language Processing: An overview of the dissertation across core Chapters. Knowledge acquisition tasks are divided into entity discovery, relation extraction, and knowledge graph completion (Ji et al., 2021). Entity discovery and relation extraction aim to construct or enrich a knowledge base from the text or use an existing one for knowledge-aware applications, whereas knowledge graph completion extends an incomplete one. Entities and relations between them form the basis of these tasks, which we study here to address the emerging needs of low-resource medical language processing.

1.1 OVERVIEW

By adopting Electronic Health Record (EHR) systems, hospitals, and clinical institutes can access large amounts of heterogeneous patient data. Such data consists of structured inputs (insurance, billing, and lab results) and unstructured inputs (doctor notes, admission/discharge details, and medication steps). Unstructured texts are often more challenging for applying information extraction methods to learn relevant concepts, entities, events, and interactions. Furthermore, we have a large number of taxonomic resources in the medical domain. These structured resources are curated by domain experts and provide a rich learning signal, where the medical concepts are classified into a hierarchical structure. Figure 2 shows an example snippet from one such KB.



- E08 Diabetes mellitus due to underlying condition
  - E08.0 Diabetes mellitus due to underlying condition with hyperosmolarity
    - E08.00 ...... without nonketotic hyperglycemic-hyperosmolar coma (NKHHC)
    - E08.01 ...... with coma
  - E08.1 Diabetes mellitus due to underlying condition with ketoacidosis
    - E08.10 ...... without coma
    - E08.11 ...... with coma
  - E08.2 Diabetes mellitus due to underlying condition with kidney complications
    - E08.21 Diabetes mellitus due to underlying condition with diabetic nephropathy
    - E08.22 Diabetes mellitus due to underlying condition with diabetic chronic kidney disease
    - E08.29 Diabetes mellitus due to underlying condition with other diabetic kidney complication
  - E08.3 Diabetes mellitus due to underlying condition with ophthalmic

Figure 2: A partial snippet of hierarchical taxonomy for the concept `E08` under sub-chapter `E08-E13` (Diabetes mellitus) under chapter `E00-E89` (Endocrine, nutritional and metabolic diseases) from International Classification of Diseases 10th revision (ICD-10) (Organization, 2004).

Since text and knowledge bases provide a common setting for information extraction applications in healthcare, they play a central role in this dissertation as input sources. Further, as *medical language processing*[1] aims to extract and parse meaningful information from unstructured text to transform it into structured data (Ananiadou and McNaught, 2006), one commonly used format is through *entity* and *relation* information, which is the focus of our *represeantation learning-based* approaches in this thesis. Lastly, the following *low-resource* conditions are considered in this dissertation:

- **Limited-to-no Labeled Data:** A recurring challenge of supervised learning methods is availability of labeled data, which is exacerbated in terms of time and costs for medical language processing needing domain experts.

- **Multilingualism:** Despite the EHR advances, most research has focused on English, with only a few recent studies for other languages (Névéol et al., 2018), making multilingual health systems a major challenge.

- **Domain Specificity:** A direct consequence of working with medical language processing is the specificity of the domain, which is more pronounced when working with sub-domains such as cancer or stroke.

- **Missing Knowledge:** With the continuous growth of biomedical literature, there is a constant need to construct, enrich, and update the knowledge bases for discovering missing facts and adding new ones.

As for modeling, we primarily use *pre-trained language models* to represent unstructured text and *knowledge graph embedding models* to represent structured multi-relational data. The thesis is structured into two main parts entity-centric learning and relation-centric learning. In the next sections, we will provide an overview of each part divided into Chapters, including research questions and contributions.

[1] We coin this term with a purpose to unify literature from biomedical text mining, biomedical language processing (BioNLP), clinical language processing (Clinical NLP), and medical NLP.

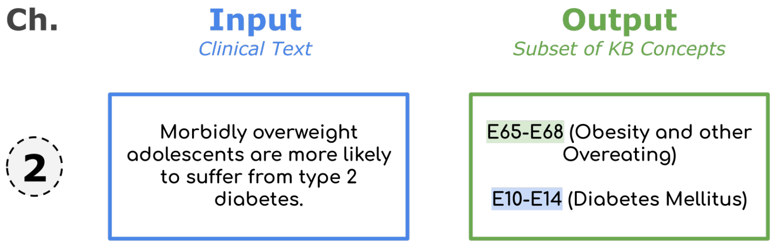

Ch.
Input
Clinical Text
Output
Subset of KB Concepts
2
Morbidly overweight adolescents are more likely to suffer from type 2 diabetes.
E65-E68 (Obesity and other Overeating)
E10-E14 (Diabetes Mellitus)

Figure 3: Chapter 2 targets supervised and unsupervised clinical concept extraction for semantic indexing. In the example input, mentions *overweight* and *diabetes* trigger the assignment of ICD-10 concepts E65-E68 and E10-E14 to text. The text is translated from German to English for readability.

1.1.1 *Part-I: Entity-centric Learning*

In this part, we focus on entity representation learning, which forms the basis of relational learning. Entity discovery includes recognition, linking, typing, and alignment (Ji et al., 2021), where we mainly focus on discovering concepts and entities from the text for downstream applications or constructing new taxonomies and knowledge bases.

1.1.1.1 *Chapter 2: Concept Extraction*

As discussed earlier, despite the digital advances in healthcare, most research has focused on English, with only a few recent studies for other languages (Névéol et al., 2018). We address this by focusing on the task of multilingual clinical concept extraction from text in EHRs. The concepts considered here are a subset of a given knowledge base and domain of interest, e.g., cancer-related concepts from ICD-10 or stroke-related concepts from SNOMED CT (Donnelly et al., 2006). A *concept* is defined as a semantic unit of text that may or may not be explicitly stated in the text but refers to an underlying term from a KB, where a named entity mentioned explicitly is considered a special case of a concept. Once we have extracted clinically relevant concepts from the text, an implicit alignment between text and knowledge base can be used for semantically indexing texts for search and knowledge discovery.

To learn the alignment, we assume two scenarios depending on the availability of labeled data. When we have a collection of *(text, concepts)* pairs, the problem reduces to a supervised multi-label concept classification. Traditional methods in concept extraction have benefited from feature engineering and dictionary look-ups (Bounaama and Amine, 2018; Gobeill and Ruch, 2018), where neural networks have improved the state-of-the-art for English clinical texts (Baumel et al., 2018). In contrast, pre-training deep neural models (Qiu et al., 2020) has significantly improved many language understanding tasks, including fine-tuning on low-resource problems, which motivates us to investigate our first research question.

RQ1: How effective is supervised transfer with pre-trained language models for multilingual clinical concept extraction?

Chapter 2 first studies the impact of transfer learning in a supervised setting for German clinical text. We consider a comprehensive suite of neural baselines that include state-of-the-art text classifiers based on Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) encoders while proposing novel label descriptions and self-attention models. To partly assess the language gap, we further introduce the use of automatic machine translation for transforming text to English and we study the domain gap by using pre-trained static and contextualized embeddings. We consider deep language models: multilingual Bidirectional Encoder Representation from Transformers (mBERT) (Devlin et al., 2019a) for German and its domain-specific variant BioBERT (Lee et al., 2020) for English under the *pre-train then fine-tune* paradigm. Our experimental evaluation resulted in an F1 score of 73% compared to the baseline of 35% (Neves et al., 2019b), highlighting the effectiveness of implicit concept representation from the BERT text encoders, alleviating the need to construct dictionaries, features, and complex neural architectures. We also find that automatic machine translation into richly resourced English improved absolute model scores by 6% on average over the results achieved in the original German source, highlighting the language gap. Lastly, domain-specific word embeddings impacted the static models more than the contextual ones. These models and findings are published in the following peer-reviewed article (Amin et al., 2019):

Saadullah Amin, Günter Neumann, Katherine Ann Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted (2019). **MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT**. In *Proceedings of the 20th Conference and Labs of the Evaluation Forum (Working Notes)* (pp. 1–15). Lugano, Switzerland. CEUR-WS.

Authors Contribution: I proposed the ideas, designed the experiments, implemented the models, and performed the analysis. I wrote the paper. My co-authors provided feedback on the manuscript. Günter Neumann suggested participating in the CLEF eHealth 2019 shared task (Kelly et al., 2019b; Neves et al., 2019a), contributed to initial brainstorming discussions, and supervised the project. Additionally, Stalin Varanasi provided helpful input for the implementation, including masking operations, residual connections, and a weighted cross-entropy loss in PyTorch (Paszke et al., 2019).

However, solving the text-concept alignment as an unsupervised problem is similar to a zero-shot classification task when we only have a collection of *text* and a predefined set of *concepts*. Despite state-of-the-art results, our fine-tuning based approach as presented above requires manually labeled data as *(text, concepts)* pairs, which may not be readily available for many languages and clinical sub-domains, prompting us to the next research question.

> *RQ2: Can we partly address the labeled data requirement of supervised clinical concept extraction with an unsupervised approach?*

Inspired by Zero-Shot Entity Linking (ZSEL) (Wu et al., 2020a), we propose a hybrid approach for unsupervised concept extraction that uses dictionary and contextual matching to generate mentioned candidates. For candidate generation, we use unsupervised Keyphrase Extraction (KPE) and Contextual Span Detection (CSD) models.

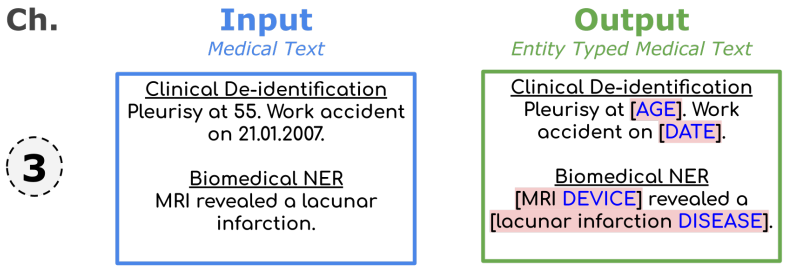

Ch.
Input
Medical Text
Output
Entity Typed Medical Text
3
Clinical De-identification
Pleurisy at 55. Work accident on 21.01.2007.
Biomedical NER
MRI revealed a lacunar infarction.
Clinical De-identification
Pleurisy at [AGE]. Work accident on [DATE].
Biomedical NER
[MRI DEVICE] revealed a [lacunar infarction DISEASE].

Figure 4: Chapter 3 considers recognizing the explicit mention of entity – a special case of concept – in text with a focus on transfer learning and application to de-identification. The relevant Protected Health Information (PHI) is redacted in clinical de-identification to preserve patient privacy, whereas the entity spans are recognized with their types for knowledge base construction and relation extraction.

The extracted spans are matched using CPMerge (Okazaki and Tsujii, 2010) and dense nearest neighbor search with FAISS (Johnson et al., 2019), where the embeddings are computed with a contextual paraphrased sentence embeddings model (Reimers and Gurevych, 2019). Finally, we apply a simple filtering threshold to get the concept subset. We evaluate our approach on a drugs review dataset in English (Yates and Goharian, 2013), reaching a considerable performance gain of 12% F1 score compared to QuickUMLS (Soldaini and Goharian, 2016), and utilize it in an unsupervised multilingual setup for the German clinical text as in the supervised case, obtaining 11% F1 score.

1.1.1.2 *Chapter 3: Named Entity Recognition*

A special case of a concept is a named entity such as a medical device (*MRI*), a disease (*stroke*), a protein (*PEX-13*), etc., in biomedical literature or clinical narratives. Extracting named entities from text followed by entity typing allows us to automatically populate a knowledge base with a clean taxonomy. However, obtaining large-scale annotated corpora is challenging, and we often resort to transfer learning methods. As highlighted in Chapter 2, pre-trained language models allow for an effective transfer of concept representations but lack a dedicated experimental test bed for Named Entity Recognition (NER). Concurrently, the research in NER transfer has made significant progress in recent years leading us to explore the following research question.

> *RQ3: How can we bridge the gap between research in pre-trained language models and algorithmic advances in NER transfer?*

To meet these needs, in **Chapter 3**, we first present a general-purpose Transformer based Transfer Learning Framework for Named Entity Recognition (T2NER) created in PyTorch (Paszke et al., 2019) for the task of NER with deep transformer models, which traditionally has benefited from LSTM networks. As the core modeling engine, the framework is built upon the Transformers (Wolf et al., 2020) library. It supports several transfer learning scenarios from sequential transfer to domain adaptation, multi-task learning, and semi-supervised learning. It aims to bridge the gap between the algorithmic advances in these areas by combining them with the state-of-the-art in transformer models to provide a unified platform that is readily extensible. It can be used for transfer learning research in low-resource NER and real-world applications, such as medical

NER. The framework, along with its design principles, NER transfer algorithms, and system description, is published in the following peer-reviewed article (Amin and Neumann, 2021):

Saadullah Amin & Günter Neumann (2021). **T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 212-220). Online. ACL.

Authors Contribution: I proposed the ideas, identified the design principles, initiated the development project, designed the architecture, and implemented the framework. I wrote the paper. Günter Neumann provided feedback on the manuscript and supervised the project.

Besides biomedical concepts and entities, clinical texts contain Protected Health Information (PHI) that risks patient identification when exposed to information extraction tools for downstream applications. Therefore, it is critical to remove such sensitive information to ensure ethical and private NLP according to GDPR (Regulation, 2016). Existing works in de-identification (solved as NER) rely on using large-scale annotated corpora in English (Stubbs et al., 2017; Stubbs and Uzuner, 2015), which often is unsuitable for real-world multilingual settings. Pre-trained language models have shown great potential for *zero-shot* cross-lingual transfer in low-resource settings (Pires et al., 2019; Wu and Dredze, 2019), including NER. However, to date model behavior is unclear in *few-shot* cross-lingual NER transfer with a potential for domain-specific clinical de-identification. We, therefore, address this gap in the next research question.

RQ4: What characteristic NER transfer property of multilingual pre-trained language models can effectively be applied to low-resource clinical de-identification?

To investigate the question, we empirically show the *few-shot* cross-lingual transfer property of mBERT for NER and apply it to solve a low-resource and real-world challenge of code-mixed (Spanish-Catalan) clinical notes de-identification in the stroke domain. We *(a)* propose an optimal few-shot cross-lingual transfer strategy *(b)* annotate a target development set, and *(c)* construct an annotated few-shot target corpus for effective cross-lingual transfer with T2NER. The development set is used for few-shot model selection, where we only use a few hundred labeled examples for training. Our model improves the zero-shot F1-score from 73.7% to 91.2% on the gold evaluation set when adapting mBERT from the synthetic MEDDOCAN (Marimon et al., 2019) corpus in Spanish with our few-shot cross-lingual target corpus. When generalized to an out-of-sample test set, the best model achieves a human-evaluation F1 score of 97.2%. These findings are published in the following peer-reviewed article (Amin et al., 2022b):

Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Kelly Wixted, Alejandro García-Rudolph, Catalina Martínez-Costa & Günter Neumann (2022). **Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts**. In *Proceedings of the 21st Workshop on Biomedical Language Processing* (pp. 200-211). Dublin, Ireland. ACL.

Authors Contribution: I led the annotation project, proposed the ideas, implemented the models, and developed the annotation toolkit. Noon Pokaratsiri Goldstein did the manual annotation, conducted data inspection, provided clinical insights, reviewed the out-of-sample test set, and contributed to writing. Morgan Kelly Wixted also did the manual annotation. Both annotators adjusted the 2014 i2b2 VA/Challenge (Stubbs and Uzuner, 2015) guidelines to make it suitable for coarse-grained de-identification. Alejandro García-Rudolph reviewed the out-of-sample test set. Catalina Martínez-Costa suggested the MEDDOCAN corpus. I wrote the main paper. My co-authors provided feedback on the manuscript. Günter Neumann presented the need for clinical de-identification and supervised the project. Additionally, Josef van Genabith provided comments that helped improve the final version of the paper.

1.1.2 *Part-II: Relation-centric Learning*

In this part, we assume that we have a knowledge base with a clean and expressive taxonomy of semantic types (or classes) and that these types are populated with a comprehensive set of canonicalized (i.e., uniquely identified) entities. Such a KB may be given or partially populated with entity-centric learning approaches discussed in Part-I.

1.1.2.1 *Chapter 4: Relation Extraction*

Machines are limited in generating knowledge since all factual knowledge about our world is created by humans and documented in an encyclopedia, scientific publications, books, or daily news (Weikum et al., 2021). Relation Extraction (RE) is one such task of knowledge discovery that aims at learning interactions between entities from text to output structured knowledge in the form of (subject, predicate, object) (SPO) triples that can either enrich an existing human-curated KB or semi-automatically construct a new one.

However, in the biomedical domain, this is challenging due to the lack of labeled data and high annotation costs, needing domain experts. Distant Supervision (DS) is commonly used to tackle the scarcity of annotated data by automatically pairing knowledge base relationships with raw texts (Mintz et al., 2009). Such a pipeline is prone to noise due to a large number of false positives, where prior work of Dai et al. (2019) extended a mutual learning framework of Han et al. (2018a) with biomedical text and a Knowledge Graph (KG) using auxiliary tasks, including Knowledge Graph Completion (KGC) and entity type classification. They showed that using attention with a KG can help in denoising using a Piece-wise Convolutional Neural Network (PCNN) sentence encoder. In Part-I, we showed that pre-trained language models and their domain-specific variants provides effective entity representations with transfer learning. Building on this and inspired by Dai et al. (2019), utilizing a knowledge base, and by Alt et al. (2019), who fine-tuned OpenAI GPT for bag-level Multi-Instance Learning (MIL) (Surdeanu et al., 2012) for Distantly Supervised Relation Extraction (DSRE) in the general domain, we arrive at the next research question.

RQ5: Can a KB be utilized for denoising relation representations from domain-specific language models for distantly supervised biomedical RE?

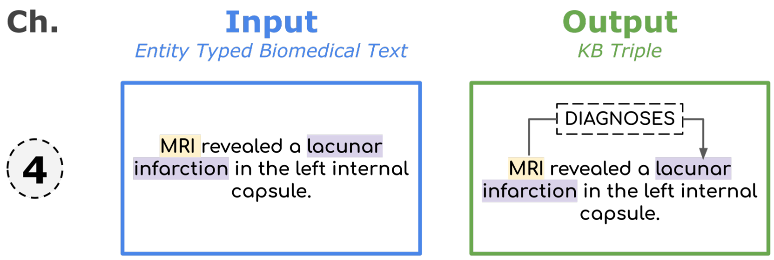

Ch.
Input
Entity Typed Biomedical Text
Output
KB Triple
4
MRI revealed a lacunar infarction in the left internal capsule.
DIAGNOSES
MRI revealed a lacunar infarction in the left internal capsule.

Figure 5: Chapter 4 studies KB enrichment where the biomedical texts are tagged with *(subject, object)* pairs from an existing KB using distant supervision to mine new triples from the text. The sentence tagged with *(MRI, lacunar infarction)* implicitly express the relation *diagnoses*.

In **Chapter 4**, we first propose to reduce the distant supervision noise by extending an entity-enriched Relation classification BERT model (Wu and He, 2019) (RBERT) to the problem of bag-level MIL for DSRE through a KB-guided encoding scheme. Our encoding identifies the *head* and *tail* role of entities from the knowledge graph by entity markers and passes through deep transformer layers with self-attention. The resulting textual representation can be pooled into a relation representation, which implicitly encodes the mutual learning with text and KG in a data-driven manner. The proposed MIL-RBERT significantly reduces noise, reaching state-of-the-art performance with 68.4% AUC and 64.9% F1 score, with an absolute 7% P@2k gain compared to Dai et al. (2019). The data pipeline and model are published in the following peer-reviewed article (Amin et al., 2020a):

Saadullah Amin*, Katherine Ann Dunfield*, Anna Vechkaeva & Günter Neumann (2020). **A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction**. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing* (pp. 187-194). Online. ACL.

Authors Contribution: I proposed the ideas, created the data, designed the experiments, implemented the models, and performed the analysis. Katherine Ann Dunfield worked on the initial implementation of RBERT in PyTorch from TensorFlow, created parts of the earlier dataset, provided helpful references in related work, and refined the model graphics. Anna Vechkaeva created parts of the earlier dataset and drafted the initial version of model graphics. I wrote the paper. My co-authors provided the feedback, and Günter Neumann supervised the project. Additionally, Qin Dai generously guided us through the steps to obtain relevant triples data from the UMLS, and Dominik Stammbach kindly provided his TensorFlow implementation of RBERT developed for the general domain RE.

In addition to noise, the second major challenge comes from scaling to a large number of concepts for broad-coverage. However, the existing works in broad-coverage Distantly Supervised Biomedical Relation Extraction (Bio-DSRE) report very accurate results (Amin et al., 2020a; Hogan et al., 2021; Xing et al., 2020), including our contribution from the last section, prompting us to inspect the benchmarks used in the evaluations. A potential lack of careful and thorough evaluation of domain-specific language models for biomedical relation extraction leads us to the next research question.

> *RQ6: Are there limitations to accurately evaluate domain-specific language models for broad-coverage distantly supervised biomedical RE?*

We start by investigating existing benchmarks for possible train-test leakage of KG triples and find significant portions overlapping from 26% up to 86%. Such leakage impacts the model performance as it allows it to score higher by simply memorizing the training relations rather than generalizing to new, previously unknown ones. We identify the sources of these issues as normalizing the textual form of concept mentions to their unique identifiers and improper handling of inverse relations. In contrast, more accurate benchmarks exist (Hong et al., 2020; Marchesin and Silvello, 2022) but focus on narrower types of interactions. To alleviate the broad-coverage benchmark issues and bridge this gap, we present a new benchmark MedDistant19 which draws its knowledge graph from the widely used healthcare ontology SNOMED CT (Chang et al., 2020). Further, with the success of domain-specific pre-trained language models (Gu et al., 2021), and inspired by existing thorough relation extraction studies in the general domain (Alt et al., 2020; Gao et al., 2021a; Peng et al., 2020), we conduct an extensive evaluation using MedDistant19 for the biomedical domain. The benchmark and findings are published in the following peer-reviewed article (Amin et al., 2022a):

Saadullah Amin*, Pasqaule Minervini*, David Chang, Pontus Stenetorp & Günter Neumann (2022). **MedDistant19: Towards an Accurate Benchmark for Broad-Coverage Biomedical Relation Extraction**. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2259-2277). Gyeongju, Republic of Korea. ICCL.

Authors Contribution: The ideas and issues emerged from an initial meeting on GitHub. I created the benchmark, designed the experiments, proposed the baselines, conducted the language model experiments, and performed the analysis. Pasquale Minervini connected through GitHub, which eventually resulted in collaboration; he participated in all the discussions, implemented the baselines, re-run the AMIL (Hogan et al., 2021) code, reached out to the National Library of Medicine (NLM), invited Pontus Stenetorp, and contributed to writing. David Chang helped with obtaining SNOMED CT triples. Pontus Stenetorp participated in the discussions. I wrote the main paper. My co-authors provided feedback on the manuscript, and Günter Neumann supervised the project. Additionally, William Hogan kindly provided data and code for the AMIL re-run.

1.1.2.2 *Chapter 5: Knowledge Graph Completion*

Despite the fact that manually or semi-automatically constructed KBs are already useful, most are incomplete by nature, with only a limited number of observed facts from the general or biomedical domain represented as structured relations between entities. To partly address this issue, we focus on the important task in statistical relational learning of link prediction or knowledge graph completion, where we assume access to a human-curated incomplete KB. As a result, we obtain a low-dimensional multi-relational graph representation of entities and relations that can be used to discover

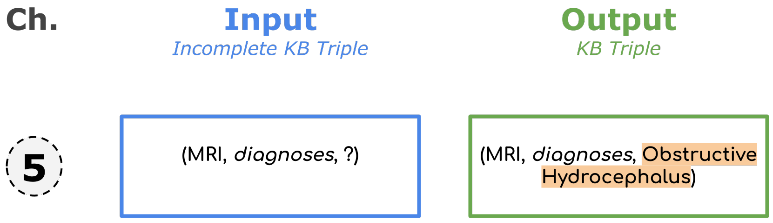


Figure 6: Chapter 5 aims to discover missing knowledge by link prediction. Here the goal is to complete the query *(MRI, diagnoses, ?)* with a target entity from the KB such as *Obstructive Hydrocephalus* assuming we learn from existing facts such as *(MRI, diagnoses, Lacunar Infarction)*. As a result, we obtain a low-dimensional entity and relation representation.

missing facts (completeness) or check the validity of others (verification). Both linear and non-linear embedding models have been proposed to solve the problem. Bilinear models, while expressive, are prone to overfitting and can lead to quadratic growth of parameters in the number of relations, guiding us to our next research question.

RQ7: How to represent entity and relation in a parameter efficient way for knowledge graph completion in the general and biomedical domain?

Since simpler models have become standard, with certain constraints on the bilinear map as relation parameters, in **Chaper 5**, we propose a factorized bilinear pooling model, commonly used in multi-modal learning (Yu et al., 2017), for a better fusion of entities and relations, leading to an efficient and constraint-free tensor factorization model, LowFER. Empirically, we evaluate on subsets of Freebase (Bollacker et al., 2008), WordNet (Miller, 1992), and YAGO (Rebele et al., 2016) in the general domain and UMLS (Bodenreider, 2004) and SNOMED CT (Donnelly et al., 2006) for the biomedical domain, reaching on par or state-of-the-art performance.

More formally, we note that a key theoretical property of KGC models is their ability to be fully expressive. A fully expressive model can represent relations of any type, including symmetric, asymmetric, reflexive, and transitive, among others. Models such as RESCAL, HolE, ComplEx, SimplE, and TuckER have been shown to be fully expressive (Balažević et al., 2019b; Kazemi and Poole, 2018; Trouillon and Nickel, 2017; Wang et al., 2018b). Furthermore, it was shown by Kazemi and Poole (2018) that RESCAL, DistMult, ComplEx, and SimplE belong to a family of bilinear models with a different set of constraints. Later, Balažević et al. (2019b) established that TuckER generalizes all of these models as special cases. These theoretical investigations lead us to our last research question.

RQ8: What theoretical insights can be drawn about the expressivity and generalizability of the efficient parameterization?

We prove LowFER is fully expressive, providing bounds on the embedding dimensionality and factorization rank. LowFER generalizes the Tucker decomposition based

TuckER model (Balažević et al., 2019b) as an efficient low-rank approximation without substantially compromising the performance. Due to low-rank approximation, the model complexity can be controlled by the factorization rank, avoiding the possible cubic growth of TuckER. At extremely low ranks, LowFER preserves the performance while staying parameter efficient. The model and the findings are published in the following peer-reviewed article (Amin et al., 2020b):

Saadullah Amin, Stalin Varanasi, Katherine Ann Dunfield & Günter Neumann (2020). **LowFER: Low-rank Bilinear Pooling for Link Prediction**. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 257-268). Online. PMLR.

Authors Contribution: I proposed the ideas, organized formal understanding, designed the experiments, implemented the models, and performed the analysis. Stalin Varanasi participated in the theoretical discussions, helped the model's understanding in terms of tensor operations, drafted an initial version of the model's interpretation with existing works, provided input for related work, and proofread the paper. Katherine Ann Dunfield participated in the initial discussions of the idea, provided motivation input for entity and relation feature fusion, typeset most tables, and organized the paper's layout and graphics. I wrote the paper. My co-authors provided feedback on the manuscript. Günter Neumann supervised the project. Additionally, the implementation largely benefited from the open-source code released by the TuckER authors.

1.2 CONTRIBUTIONS

Below we summarize contributions in terms of the peer-reviewed articles and the code released. We also declare the funding sources.

Publications

- **Chapter 2: Concept Extraction**

  Paper : Saadullah Amin, Günter Neumann, Katherine Ann Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted (2019). MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In *Proceedings of the 20th Conference and Labs of the Evaluation Forum (Working Notes)* (pp. 1–15). Lugano, Switzerland. CEUR Workshop Proceedings.
  Code : `https://github.com/suamin/ICD-BERT`

  CLEF'19, Precise4Q

- **Chapter 3: Named Entity Recognition**

  Paper : Saadullah Amin & Günter Neumann (2021). T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 212-220). Online. Association for Computational Linguistics.
  Code : `https://github.com/suamin/T2NER`

  EACL'21, Precise4Q, CoRA4NLP

BioNLP'22, Precise4Q, CoRA4NLP

Paper : Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Kelly Wixted, Alejandro García-Rudolph, Catalina Martínez-Costa & Günter Neumann (2022). Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts. In *Proceedings of the 21st Workshop on Biomedical Language Processing* (pp. 200-211). Dublin, Ireland. Association for Computational Linguistics.
Code : `https://github.com/suamin/FewDeid` [2]

- **Chapter 4: Relation Extraction**

BioNLP'20, Precise4Q, DEEPLEE

Paper : Saadullah Amin*, Katherine Ann Dunfield*, Anna Vechkaeva & Günter Neumann (2020). A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing* (pp. 187-194). Online. Association for Computational Linguistics.
Code : `https://github.com/suamin/MIL-RBERT`

COLING'22, Precise4Q, CoRA4NLP

Paper : Saadullah Amin*, Pasqaule Minervini*, David Chang, Pontus Stenetorp & Günter Neumann (2022). MedDistant19: Towards an Accurate Benchmark for Broad-Coverage Biomedical Relation Extraction. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2259-2277). Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
Code : `https://github.com/suamin/MedDistant19`

- **Chapter 5: Knowledge Graph Completion**

ICML'20, Precise4Q, DEEPLEE

Paper : Saadullah Amin, Stalin Varanasi, Katherine Ann Dunfield & Günter Neumann (2020). LowFER: Low-rank Bilinear Pooling for Link Prediction. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 257-268). Online. Proceedings of Machine Learning Research.
Code : `https://github.com/suamin/LowFER`

Funding

First author acknowledges funding from the European Union (EU) Horizon 2020 research and innovation program through the project Precise4Q (777107)[3] and partial funding from the German Federal Ministry of Education and Research (BMBF) through the projects DEEPLEE (01IW17001)[4] and CoRA4NLP (01IW20010)[5].

[2] This is a private repository containing real patient data subject to GDPR and HIPAA compliance.

[3] `https://precise4q.eu/`

[4] `https://www.deeplee.de/`

[5] `https://cora4nlp.github.io/`

* *The authors contributed equally.*

Part I

ENTITY-CENTRIC LEARNING

2 MULTILINGUAL AND UNSUPERVISED CLINICAL CONCEPT EXTRACTION FOR SEMANTIC INDEXING

2.1 INTRODUCTION

EHR systems offer rich data sources that can improve healthcare systems by applying information extraction, representation learning, and predictive modeling techniques (Shickel et al., 2018). Among many other applications, one commonly studied task is the automatic assignment of medical concepts from a knowledge base, e.g. ICD-10 (Organization, 2004), to clinical notes (Crammer et al., 2007). The problem is to learn a mapping from natural language clinical text to a selected subset of concepts from a given domain of interest, such as cancer or stroke. For a new document, the system can assign one or more concepts that allow semantic indexing for search and discovery.

Learning such mapping in a supervised manner is seen as a multi-label classification problem, and is one way to solve the problem, besides hierarchical classification, learning-to-rank, and unsupervised methods. Despite the advances in clinical concept extraction, the challenges of multilingual text have been under-studied mainly due to a lack of labeled corpora with the recent exception of electronic health information extraction shared tasks. Concurrently, NLP has made significant progress in transfer learning that motivates us to study its effectiveness in multilingual clinical concept extraction for **RQ1** in §2.3 as stated:

> *RQ1: How effective is supervised transfer with pre-trained language models for multilingual clinical concept extraction?*

A closely related task for such text-to-knowledge mapping is entity linking, which generally consists of two sub-tasks of surface form extraction, i.e., mention recognition, and named entity disambiguation. A surface form is a contiguous span of text that implicitly refers to a concept. The disambiguation task aims to link the identified named entity to ground truth entities from a given knowledge base. Traditionally, this is a supervised learning task but requires laborious labeling, whereas, in the case of unseen entities, it is referred to as Zero-Shot Entity Linking (ZSEL). Comparatively, the approaches for unsupervised concept extraction use approximate string matching to extract fuzzy candidate mentions in text and align them with the concepts of interest. However, recent advances in medical entity linking have shown the possibility of a hybrid framework combining dictionary and embedding-based matching (Loureiro and Jorge, 2020). Inspired by this and scalable ZSEL (Wu et al., 2020a), we investigate a dense phrase matching framework for unsupervised concept extraction to study the impact on the labeled data requirement of incurred in the supervised counterpart with **RQ2** in §2.4 as stated:

> *RQ2: Can we partly address the labeled data requirement of supervised clinical concept extraction with an unsupervised approach?*

*The contents of §2.2.1 and §2.3 have appeared in the peer-reviewed article of **Amin et al. (2019)** and are included here with minor corrections where appropriate. The contents of §2.2.2 and §2.4 are only appearing in this dissertation.*

2.2 RELATED WORK

2.2.1 *Supervised Methods*

Clinical concept extraction for semantic indexing of health-related documents has been well studied, both in the previous Conference and Labs of the Evaluation Forum (CLEF) eHealth shared tasks and in general (Crammer et al., 2007). Traditional approaches range from rule-based and dictionary look-ups (Bounaama and Amine, 2018) to machine learning models (Gobeill and Ruch, 2018).

More recently, the focus has been on applying deep learning, where several architectures have been proposed using convolutional, recurrent, and hybrid models. Flicoteaux (2018) uses a shallow CNN and improves its predictions for rare labels by dictionary-based lexical matching. Baumel et al. (2018) addresses the challenges of long documents and the high cardinality of the label space (Johnson et al., 2016) by modifying the Hierarchical Attention Network (HAN) (Yang et al., 2016) with label attention.

Ševa et al. (2018) builds a multilingual death cause extraction model using an LSTM encoder-decoder, with concatenated French, Hungarian, and Italian fastText embeddings (Grave et al., 2018) as inputs and causes extracted from concept dictionaries as outputs. Ive et al. (2018) uses a character-level CNN (Zhang et al., 2015) encoder for French and Italian with a bidirectional RNN decoder. Jeblee et al. (2018) enriches word embeddings with language-specific Wikipedia text and creates an ensemble model from a CNN classifier and GRU encoder-decoder. While successful, these approaches make an auto-regressive assumption on output codes, which only holds when there is a single path from parent to child code for a given document. However, in concept extraction, a document can have multiple disjoint paths in a Directed Acyclic Graph (DAG) formed by a concept hierarchy (Silla and Freitas, 2011). Additionally, the decoder suffers from vocabulary sparsity and variance in low-resource datasets.

Contextualized word embeddings, ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019b), derived from pre-trained bidirectional language models trained on large texts, have been shown to improve performance on many NLP tasks: Question Answering (QA), Textual Entailment, Sentiment Classification, Constituency Parsing, and Named Entity Recognition (NER). Sequential transfer with these models involves a *pre-train then fine-tune* strategy for a downstream supervised task to achieve substantial qualitative gains with limited data and task-specific engineering. Hence, they are simple, efficient, and performant. Motivated by this and the recent work of domain-specific variants (Lee et al., 2019), we study BERT for multilingual clinical concept extraction.

2.2.2 *Unsupervised Methods*

Concept extraction pipelines of MetaMap (Aronson and Lang, 2010) and cTAKES (Savova et al., 2010) are widely used in research and industrial applications. The former uses a shallow parser to generate candidate phrases; then, for each candidate phrase, a range of lexical variations are generated; finally, each phrase is scored based on its distance to concepts in UMLS (Bodenreider, 2004). A word disambiguation tool is further used to favor concepts that are semantically consistent with the surrounding text. Similarly, cTAKES matches candidate phrases as well as their permutations and lexical variations with concepts in UMLS and a list of concepts maintained by the Mayo Clinic (Soldaini and Goharian, 2016).

QuickUMLS (Soldaini and Goharian, 2016) is a fast and unsupervised approach based on approximate dictionary matching. It employs a series of filters on the sequence of tokens and part-of-speech tags to extract candidate phrases, which are then matched to UMLS concepts with an approximate dictionary-matching algorithm called CPMerge (Okazaki and Tsujii, 2010). Other unsupervised methods include MaxMatcher (Zhou et al., 2006) and MeTAE (Ben Abacha and Zweigenbaum, 2011).

MedLinker (Loureiro and Jorge, 2020) is a state-of-the-art system trained on a large-scale supervised entity linking dataset MedMentions (Mohan and Li, 2018). It is a hybrid approach of neural models and dictionary-based string matching. It follows a two-step pipeline, a mention recognition followed by entity linking. The entity linking step uses contextual embeddings for entity classification and combines it with semantic similarity matching. To handle zero-shot (unseen) entities, it relies on a dictionary-based CPMerge algorithm, thus offering a hybrid approach. Our unsupervised concept extraction pipeline largely follows MedLinker, where we propose two variants of mention detection and combine them with a dense matching approach, following zero-shot entity linking (Wu et al., 2020a), that provides a framework to study unsupervised extractions.

2.3 multilingual concept extraction

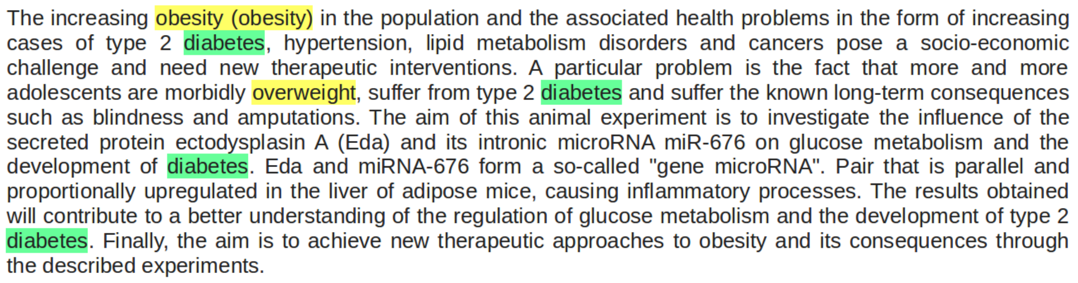
CLEF eHealth 2019 (Kelly et al., 2019a) Task 1 (Neves et al., 2019b) advances research in multilingual concept extraction by focusing on German Non-Technical Summaries (NTS) of animal experiments collected from the AnimalTestInfo database to classify according to ICD-10-GM 2016 [1]. The AnimalTestInfo database was developed in Germany to make the non-technical summaries of animal research studies available in a searchable and easily accessible web-based format (Bert et al., 2017). This task requires an automated approach to classify the NTSs, whereby the data entails challenging attributes of multilingualism, domain specificity, and concept skewness with a hierarchical structure. It serves as the basis for us to study **RQ1** by conducting a thorough evaluation of neural architectures with static word embeddings and fine-tuning pre-trained multilingual and domain-specific language models. In the next section, we will provide details of the models considered.

2.3.1 *Neural Architectures*

2.3.1.1 *Convolutional Networks*

A CNN learns local features of input representations through varying numbers and sizes of filters performing convolution operations. They have succeeded in many text classification tasks (Johnson and Zhang, 2017; Zhang et al., 2015) and while many advanced CNN architectures exist, we only employ a shallow model of Kim (2014).

[1] https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2016/



The increasing obesity (obesity) in the population and the associated health problems in the form of increasing cases of type 2 diabetes, hypertension, lipid metabolism disorders and cancers pose a socio-economic challenge and need new therapeutic interventions. A particular problem is the fact that more and more adolescents are morbidly overweight, suffer from type 2 diabetes and suffer the known long-term consequences such as blindness and amputations. The aim of this animal experiment is to investigate the influence of the secreted protein ectodysplasin A (Eda) and its intronic microRNA miR-676 on glucose metabolism and the development of diabetes. Eda and miRNA-676 form a so-called "gene microRNA". Pair that is parallel and proportionally upregulated in the liver of adipose mice, causing inflammatory processes. The results obtained will contribute to a better understanding of the regulation of glucose metabolism and the development of type 2 diabetes. Finally, the aim is to achieve new therapeutic approaches to obesity and its consequences through the described experiments.

Figure 7: An example document tagged with concepts E10-E14 (*diabetes mellitus*) and E65-E68 (*obesity and other overeating*) containing related words to concepts descriptions.

2.3.1.2 *Recurrent Networks*

In recurrent networks, we focus on attention-based models. The *attention* mechanism, initially proposed in sequence-to-sequence-based Neural Machine Translation (NMT), allows the decoder to attend to encoder states while making predictions (Bahdanau et al., 2015). Attention generates a probability distribution over features, where the model learns to put more weight on relevant features. In our study, we consider three attention-based models.

HAN deals with the problem of document classification by modeling attention at each hierarchical level of the document, i.e., words and sentences (Yang et al., 2016). This lets the model first attend word encoder outputs in a sentence, followed by attending to the sentence encoder outputs to classify a document. Similar to Yang et al. (2016), we use bidirectional GRUs as word and sentence encoders.

We introduce a Self-attention LSTM (SLSTM) network, which is a single-layer model based on a bidirectional LSTM encoder with a dense self-attention and residual connection. An input sequence is first passed through the encoder, and encoded representations are self-attended with a residual path to produce outputs.

ICD-10 concepts have textual descriptions, e.g. concept A80-A89 is about *viral infections of the central nervous system*, which serves as additional meta-data for the model to make predictions. Figure 7 shows a document containing words related to those found in the descriptions of their labeled concepts. Such words may or may not be present in the document, with the potential to be utilized to enrich the encoder representation through attention. To our knowledge, this is the first time the concept descriptions are directly used to align with input text via attention. The closest work t ours is from Baumel et al. (2018), where the author uses label attention, but they directly consider the concept as a unit of representation, creating an embedding lookup. By using description texts, we also create an embedding layer for concepts where a concept representation is obtained via each token's average word embeddings. We call this network Concept-attentive LSTM (CLSTM) and describe it formally below.

Let $X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{n \times d}$ be an n-length input document sequence, where x_i is a d-dimensional embedding vector for input word w_i belonging to document vocabulary $\mathbf{V}_D$. Let $T = \{t_1, t_2, ..., t_m\} \in \mathbb{R}^{m \times l}$ be an m-concepts by l-length description representation matrix, where each $t_i = \{t_{i_1}, t_{i_2}, ..., t_{i_l}\} \in \mathbb{R}^{l \times d}$ and t_{i_j} is a d-dimensional embedding vector for code i's description word j, belonging to the descriptions vocabulary $\mathbf{V}_T$. The embedding matrices are different for documents and concept descriptions since the description words can be missing in the document vocabulary. Similarly, we used different LSTM encoders for document and code words. The network then transforms input as $X_{out} = \text{CLSTM}(X, T)$, with following operations:

$$
\begin{aligned}
X_{enc} &= [x_{1_{enc}}, x_{2_{enc}}, ..., x_{n_{enc}}] \\
x_{i_{enc}} &= \text{LSTM}_{W}(x_i) \\
T_{enc} &= [t_{1_{enc}}, t_{2_{enc}}, ..., t_{m_{enc}}] \\
t_{i_{enc}} &= \frac{1}{l}\sum_{j=1}^{l} \text{LSTM}_{C}(t_{i_j}) \\
X_{out} &= [X_{enc}; T_{enc}] \in \mathbb{R}^{(n+m)\times h} \\
A &= \texttt{softmax}(X_{out}X_{out}^{T}) \in \mathbb{R}^{(n+m)\times(n+m)} \\
X_{out} &= X_{out} + A^{T}X_{out} \\
X_{out} &= \frac{1}{n}\sum_{j=1}^{n} X_{out_j}
\end{aligned}
$$

where X_{enc} is a sequence of word encoder (LSTM_W) outputs and T_{enc} is a sequence of averaged description words encodings by the code encoder (LSTM_C). We concatenate the document word sequence with the description sequence and perform self-attention A, followed by residual connection and average over the resulting sequence to obtain the final representation.

2.3.2 *Transfer Learning*

Pre-training large models on an unsupervised corpus with a language modeling objective and then fine-tuning the same model for a downstream supervised task eliminates the need for heavily engineered task-specific architectures. BERT (Devlin et al., 2019b) is a recently proposed model, following ELMo (Peters et al., 2018) and OpenAI GPT (Radford et al., 2018). It is a multi-layer bidirectional transformer encoder with a feed-forward multi-headed self-attention encoder (Vaswani et al., 2017). BERT is trained with two objectives, *masked language modeling*: predicting a missing word in a sentence from the context and *next sentence prediction*: predicting whether two sentences are consecutive sequences. BERT has improved the state-of-the-art in many language understanding tasks. Recent works show that it sequentially models something akin to an NLP pipeline consisting of POS tagging, parsing, NER, semantic role labeling, and coreference resolution (Tenney et al., 2019). Similar works have been performed to understand and interpret BERT's learning capacity (Goldberg, 2019; Yogatama et al., 2019) but currently to the best of our knowledge there is no study on multilingual concept extraction. Therefore, we investigate BERT in our task and show that it achieves better results than other models. Besides neural baselines and pre-trained language models, we also consider a Term Frequency-Inverse Document Frequency (TF-IDF) weighted bag-of-words based linear Support Vector Machine (SVM) model as a simple baseline.

2.3.3 *Machine Translation*

Since the documents are domain-specific and multilingual, it is difficult for open-domain and multilingual pre-trained models to effectively transfer representations.

Concept	No. of documents (training and validation)
II	1515
C00-C97	1479
IX	930
VI	799
C00-C75	732

Table 1: Top-5 most frequent concepts in German clinical texts.

Furthermore, Amplayo et al. (2018) suggests that each language has linguistic and cultural characteristics that contain different signals to classify a specific class. Based on this and the fact that automatic machine translation is generally available, we consider it for data transformation and show improvements across all models. Since English has readily accessible biomedical literature available as free texts, we use English translations for our documents and fine-tune domain-specific model BioBERT (Lee et al., 2019), showing significant gain while highlighting the language performance gap between the English translation of the data and the German original data.

2.3.4 *Experiments*

2.3.4.1 *Data*

The dataset contains 8,385 training documents, including validation, and 407 test documents, all in German. Each document has six text fields: document title, use goals of the experiment, possible harms caused to animals, and comments about *replacement*, *reduction*, and *refinement* in the scope of 3R principles.

The documents are assigned one or more concepts from ICD-10-GM (German Modification version 2016), which exhibits a hierarchy forming a DAG (Silla and Freitas, 2011), where the highest-level nodes are called *chapters*, and their direct child nodes are called *groups*. The depth of most chapters is one, but in some cases, it goes to the second level (e.g., M00-M25, T20-T32) and, in one case, up to the third level (C00-C97). Documents are assigned heterogeneous concepts such that a parent and child node can co-exist, and a child node can have multiple parents. Moreover, 91 documents are missing one or more of the six text fields, and only 6,472 have labels (5,820 in the training set and 654 in the validation set), while 52 have only chapter-level concepts. Table 1 shows the top-5 most frequent concepts. These concepts account for more than 90% of the dataset leading to a long-tailed distribution. Due to a shallow hierarchy, we formulate the task as a multi-label classification problem instead of a hierarchical one.

Pre-processing: We consider each document as one text field, i.e., all six fields are joined together to form one input text. As mentioned, only 6,472 documents are labeled, out of which 654 are in the validation set from a total of 842. Since there is no ground truth available for 188 documents, we cannot evaluate them, so we ignored them during training. This way we avoided adding an extra *NA* class label as a placeholder for predicting no class for such documents. We assume that all documents must be indexed similarly to MEDLINE auto-indexing of new PubMed articles and, therefore, inherently each one has one or more true ICD-10 concepts assigned to them. However, the official evaluation of CLEF eHealth 2019 penalizes model predictions for 188 documents by considering them false positives. We will cover this in detail in the results section.

Data Transformation: To translate German documents to English, we used automatic translation from the Google Translate API v2[2]. For both German and English, we use language-specific sentence and word tokenizers offered by NLTK (Loper and Bird, 2002) and spaCy[3], respectively. Tokens with document frequencies outside 5 and 60% of the training corpus were removed, and only top-10,000 tokens were kept to limit the vocabulary. This applies to all models other than BERT, which uses the WordPiece tokenizer (Wu et al., 2016) and builds its own vocabulary. Lastly, we remove all the classes with a frequency of less than 15 in training. All the experiments were performed with the validation set to find the best parameters.

2.3.4.2 *Implementation*

TF-IDF + Linear SVM: For the baseline, we use the scikit-learn implementation of `LinearSVC` with one-vs-all training (Pedregosa et al., 2011).

CNN: We configured the CNN with 64 channels and filter sizes of 3, 4, and 5.

HAN: Following Yang et al. (2016), we also used bidirectional GRU encoders with a hidden size of 300. We set the maximum number of sentences in documents and the maximum number of words in a sentence as 40 and 10, respectively.

SLSTM: A bidirectional LSTM encoder with a hidden size of 300.

CLSTM: Similar to SLSTM, but with an additional matrix T of size: total number of descriptions, which is 230 as collected from ICD-10-GM, times maximum description sequence length of 10.

BERT: We used PyTorch's implementation of BERT[4] with default parameters. To avoid memory issues, we used a maximum sequence length of 256 with batch size 6.

Ensemble: Based on the validation set results, we also created an ensemble of the top-2 models (across the model classes) as a weighted combination of their raw scores, where then the prediction for each example is given by:

$$\hat{y} = \mathbb{1}\{\sigma(\kappa \times S_1 + (1 - \kappa) \times S_2) > 0.5\} \in \{0, 1\}^{|C|}$$

S_1 and S_2 are raw probability scores from the first and second-best models, respectively, while σ is the sigmoid function and $|C|$ is the number of classes. We select the best value of κ on the validation set such that the F1-score of the ensemble is higher than individual models. Figure 8 shows κ variation with performance metrics.

For all the models, except BERT, we used a batch size of 64, a sequence length of 256, a learning rate of 0.001 with Adam (Kingma and Ba, 2015a), and 50 epochs with early stopping. We used class-balanced binary cross-entropy loss for training and F1-micro score as the performance metric. Experiments were performed on a single 12GB NVIDIA TitanXp GPU. We implemented these models in PyTorch (Paszke et al., 2019), and our code is publicly available.[5] We used the following pre-trained German models:

- FT_{de}: fastText DE Common Crawl (300d)[6]
- BERT_{de}: BERT-Base, Multilingual Cased (768d)[7]

[2] https://cloud.google.com/translate/docs/translating-text

[3] https://spacy.io/usage/models

[4] https://github.com/huggingface/pytorch-pretrained-BERT

[5] https://github.com/suamin/ICD-BERT

[6] https://fasttext.cc/docs/en/crawl-vectors.html

[7] https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

Models		P	R	F1
Baseline	TF-IDF$_{\text{de}}$	**90.72**	58.73	71.30
	TF-IDF$_{\text{en}}$	90.69	65.45	76.03
CNN	FT$_{\text{de}}$	86.08	57.37	68.85
	FT$_{\text{en}}$	85.76	61.59	71.69
	PubMed$_{\text{en}}$	87.95	65.10	74.82
HAN	FT$_{\text{de}}$	78.86	58.79	67.37
	FT$_{\text{en}}$	83.52	64.50	72.79
	PubMed$_{\text{en}}$	85.10	69.61	76.58
SLSTM	FT$_{\text{de}}$	85.55	64.86	73.76
	FT$_{\text{en}}$	87.53	67.65	76.32
	PubMed$_{\text{en}}$	87.33	70.09	77.77
CLSTM	FT$_{\text{de}}$	83.60	63.97	72.48
	FT$_{\text{en}}$	84.39	69.14	76.01
	PubMed$_{\text{en}}^{\dagger}$	87.87	70.21	78.05
BERT	Multi$_{\text{de}}$	70.96	83.41	76.68
	BERT$_{\text{en}}$	79.63	84.60	82.04
	BioBERT$_{\text{en}}^{\ddagger}$	80.35	**85.61**	82.90
Ensemble (†, ‡)		86.29	83.11	**84.67**

Table 2: Results on the **validation set** where overall best is boldfaced and second best underlined.

and the following for English:

- FT$_{\text{en}}$: fastText EN Common Crawl (300d)
- PubMed$_{\text{en}}$: PubMed word2vec (400d)[8]
- BERT$_{\text{en}}$: BERT-Base, Cased (768d)[9]
- BioBERT$_{\text{en}}$: BioBERT (768d)[10]

2.3.5 *Results*

Table 2 summarizes the results on the validation set for all models with different pretrained embeddings. In all of our experiments, working with translated English texts improved the score by an average of 4.07%, even though automatic translation is likely to introduce some noise. This can be attributed to the fact that there is an abundance of English texts compared to the other languages. However, it could also be supported by English having a stronger linguistic signal (Amplayo et al., 2018) to extract the concepts where the German models underperform.

The bag-of-words baseline showed the highest precision and outperformed neural models, HAN and CNN, for German and English with Common Crawl embeddings. Generally, HAN performs better when documents are relatively long, e.g., Baumel et al.

[8] https://archive.org/details/pubmed2018_w2v_400D.tar

[9] https://storage.googleapis.com/bert_models/2018_10_18/cased_L-12_H-768_A-12.zip

[10] https://github.com/naver/biobert-pretrained/releases/tag/v1.0-pubmed-pmc

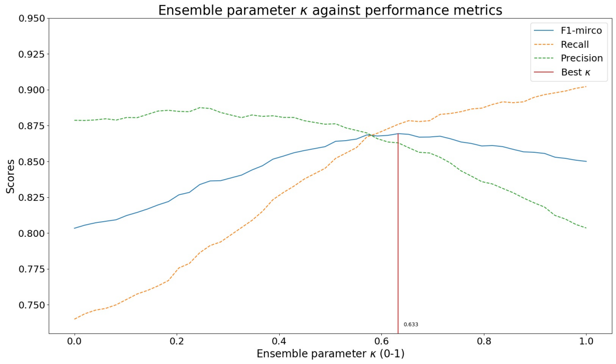


Figure 8: The graph shows the effect of varying κ to create an ensemble of top-2 models. The optimal value is at κ=0.63, represented by the redline's intersection with the x-axis.

(2018) reports strong results with HAN-based models on MIMIC datasets (Johnson et al., 2016), where the average document size exceeds 1,900 tokens. After pre-processing, the average document length in our case was approximately 340. For CNN, advanced variants (with multiple filters) can potentially result in better performance.

SLSTM and CLSTM, being single-layer networks, performed comparably and better than the baseline. SLSTM relies only on self-attention and residual connections, and even better scores are achieved by BERT models composed of stacked multi-headed self-attention and residual blocks. For CLSTM, since many documents in the corpus are missing the description words, the model had weak attention alignments between documents and concepts descriptions meta-data. However, CLSTM performed reasonably well, obtaining the second-best score with PubMed embeddings.

BERT performed better than other models in German and English, with an absolute average score of 6% higher, showing the effectiveness of transfer learning for supervised multilingual clinical concept extraction to answer **RQ1**. BioBERT$_{\text{en}}$ performed just slightly (+0.86%) better than BERT$_{\text{en}}$; this was also noticeable in relation extraction task in Lee et al. (2019), where domain-specific and general BERT performed comparably. This partly shows BERT's ability to generalize and be robust to domain shifts with limited learning from only 5,000 training documents. However, it contradicts the recent findings of Yogatama et al. (2019), where authors study domain shifts and catastrophic forgetting in pre-trained language models. Furthermore, the impact of using in-domain pre-trained models was more significant for static embeddings with PubMed$_{\text{en}}$ word embeddings outperforming open-domain FT$_{\text{en}}$ by an average of 2.77%. Unfortunately, we were not able to conduct a similar analysis for German due to a lack of pre-trained medical domain German word embeddings. BERT models had the highest recall but relatively poor precision. While problematic in general, this is preferable in real-world medical applications, where recall is of much more importance, especially for billing purposes.

We also combined the top-2 models, BioBERT$_{\text{en}}$ and CLSTSM-PubMed$_{\text{en}}$, to obtain an ensemble that performed better than both and achieved the best score of 84.67% on the validation set. The goal was to improve BERT's precision without a substantial loss in the recall. As shown in Figure 8 at $\kappa = 0.63$, the ensemble obtained the highest F1-score. This amounts to an increase in BioBERT$_{\text{en}}$ precision by 7.24% at an expense of 2.5% recall.

Models		Original			Modified		
		P	R	F1	P	R	F1
Baseline	TF-IDF$_{de}$	**89.58**	52.74	66.39	**93.01**	52.74	67.31
	TF-IDF$_{en}$	88.31	60.79	72.01	91.53	60.79	73.06
CNN	FT$_{de}$	80.30	54.66	65.04	86.99	54.66	67.13
	FT$_{en}$	78.09	58.74	67.05	83.33	58.74	68.91
	PubMed$_{en}$	80.89	64.36	71.69	86.74	64.36	73.90
HAN	FT$_{de}$	71.45	54.66	61.93	80.60	54.66	65.14
	FT$_{en}$	75.88	62.70	68.67	82.10	62.70	71.10
	PubMed$_{en}$	79.51	66.41	72.37	84.82	66.41	74.49
SLSTM	FT$_{de}$	79.17	64.11	70.85	85.37	64.11	73.23
	FT$_{en}$	82.53	65.77	73.20	86.26	65.77	74.63
	PubMed$_{en}$	77.13	68.07	72.32	83.15	68.07	74.85
CLSTM	FT$_{de}$	83.60	63.97	72.48	87.52	63.97	73.91
	FT$_{en}$	75.74	65.00	69.96	82.62	65.00	72.76
	PubMed$_{en}$ †	82.15	68.19	74.52	86.82	68.19	76.39
BERT	Multi$_{de}$	54.10	83.39	65.62	68.23	83.39	75.05
	BERT$_{en}$	62.09	83.26	71.11	75.20	83.26	79.03
	BioBERT$_{en}$ ‡	63.68	**85.56**	73.02	76.57	**85.56**	80.82
Ensemble (†, ‡)		74.44	81.86	**77.98**	83.13	81.86	**82.49**

Table 3: Results on the **test set** where the overall best is boldfaced and the second best underlined. The Original column refers to the official evaluation setup, and the Modified refers to the case where we ignore test documents without gold labels for evaluation.

Test Scores: The test set contains 407 documents, which we first translate to English and then run predictions with BioBERT$_{en}$ as our inference model. We obtained a test F1-micro of 73% with 86% recall and 64% precision, as posted by official results. Our model ranked second in CLEF 2019 eHealth Task 1, but there was a significant difference between test and validation set performances, particularly low precision. After the gold set was released, we probed it and found that the official script provided by the shared task considers all predictions on test examples for which there is *no gold label*, 93 of them, as *false positives*. Comparing examples with predictions where the gold standard is unavailable is intrinsically impossible. To emphasize, we give an example, if we take a test document with id 20486 where the gold labels are {C00-C97, C76-C80, II} and our best model predicted {C00-C97, C76-C80, II} i.e., a perfect match with maximum score. Given the official evaluation, if this example did not have the gold standard available, our model predictions would all have been considered false positives. Therefore, it degrades the precision of a model that may have generalized well to predict future examples.

Table 3 shows this comparison on the test set, where in the *Original* column, we use the same evaluation as provided by the task. In the *Modified* column, we remove all documents from the evaluation for which gold labels are unavailable. Table 3 shows that the recall column remains the same as the original, with only the precision column changing, hence improving the F1 score. With the modification, all the models have similar performance as on the validation set where we removed unlabeled examples.

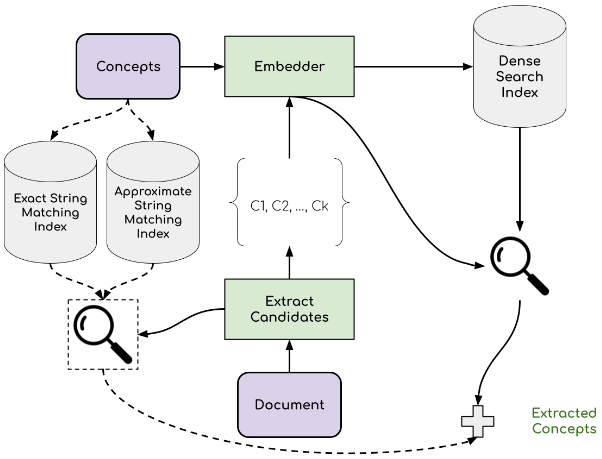


Figure 9: Overview of the unsupervised concept extraction framework with dense phrase matching combined with string matching (Algorithm 1). Dashed lines show optional paths, where the concepts of interest are pre-specified with their textual description, and input is a query document.

Furthermore, the submitted system achieves a test score of 80.82% now compared to that of 82.90% on the validation set. Finally, the ensemble model obtains the highest scores of 77.98% and 82.49% with original and modified evaluations, respectively.

2.4 UNSUPERVISED CONCEPT EXTRACTION

In supervised concept extraction from clinical text (Amin et al., 2019; Crammer et al., 2007), we assume that we have access to manually annotated *(text, concepts)* pairs. However, this might be unfeasible for low-resource languages or challenging focused domains such as cancer or stroke in real-world applications. The key idea behind unsupervised approaches for concept extraction is to use approximate string matching for extracting fuzzy candidate mentions in text and aligning them with the concepts of interest. Prior works have mainly focused on string matching only. However, recent advances in medical entity linking have shown the utility of a hybrid framework combining dictionary and embedding-based matching (Loureiro and Jorge, 2020). Inspired by this and the advances in zero-shot entity linking (Wu et al., 2020a), we investigate a dense phrase matching framework for unsupervised concept extraction to study the labeled data requirement of supervised methods as part of **RQ2**.

2.4.1 *Dense Phrase Matching*

In entity linking, the first step is detecting candidate mentions that might refer to the underlying knowledge base concept. For example, the text *".. more and more adolescents are morbidly overweight and suffer from type 2 diabetes ..."* indicates the presence of concepts *diabetes mellitus* and *obesity and other overeating* from ICD-10 (Organization, 2004). Therefore, a robust mention recognition model is a critical component of concept extraction pipelines.

Secondly, we need to search for each mentioned candidate in a collection of concepts. To scale to modern medical ontologies, we propose a fast and dense nearest neighbor search with FAISS (Johnson et al., 2019). The dense embedding is derived from an off-the-shelf paraphrase embeddings model and can be replaced by any model similar to those discussed in §2.3.4.2. Lastly, we search for candidates in approximate and exact string-matching indices, effectively resulting in a hybrid framework. Figure 9 shows our framework, and Algorithm 1 outlines the Dense Phrase Matching (DPM). We describe each component in detail in the following sections.

2.4.2 *Linguistic Processing*

In this step, we perform basic text processing and run the document through a linguistic pipeline such as spaCy (Honnibal and Montani, 2017). This step involves tokenization, sentence segmentation, and POS tagging. In the case of long documents, we divide them into segments to obtain smaller, more manageable document segments. We use a fixed number of sentences to limit document length.

2.4.3 *Candidate Extraction*

One of the main components in our pipeline is the candidate extractor. While using raw n-grams as candidates can yield high recall, it comes at the expense of noise, affecting precision. To improve this, we use short text spans that are more informative using unsupervised Keyphrase Extraction (KPE) or Concept Span Detection (CSD).

Keyphrase Extraction: The unsupervised keyphrase extraction methods mainly use different features of the document, such as word frequency, position, linguistic features, topic, length, the relationship between words, and external knowledge-based information. Graph-based keyphrase extraction is one of the most effective and widely used methods. Inspired by PageRank (Page et al., 1999), TextRank abstracts the document into a graph, where words or phrases are nodes in the graph and relationships between words are edged (Mihalcea and Tarau, 2004). We use TextRank in this work with plug-in support to apply other approaches.

Concept Span Detection: Besides KPE, we propose a method based on mentions recognition. Unlike MedLinker (Loureiro and Jorge, 2020), we use the MedMentions (Mohan and Li, 2018) dataset to identify text spans of interest without entity type. We train a multilingual model on the MedMentions (in English) and use it to infer candidates from new texts. The approach is language and document length agnostic and can provide noisy candidates. Note that the MedMentions dataset is in English, but the model is multilingual, so we expect it to have a weak zero-shot performance to generalize to unseen languages. Lastly, KPE and CSD focus on noun phrases, i.e., concepts hence limiting the performance when semantic units are non-conceptual.

2.4.4 *Concept Matching*

Embedder: The next important component of our pipeline is the text embedder. Here we propose an improved language-agnostic version of the Sentence BERT (Reimers and Gurevych, 2019) (S-BERT) at the level of short texts. The focus is to enhance Semantic Textual Similarity (STS). More specifically, the authors train an expert monolingual semantic textual similarity model and use knowledge distillation in a teacher-student framework to adapt it to several other languages using a parallel corpus. Besides STS, natural language inference, paraphrase identification, and duplicate detection models have also been proposed in S-BERT. While these embeddings are derived based only

Algorithm 1 Dense Phrase Matching

```
Input: Collection of documents ($\mathcal{D}$), Collection of concepts ($\mathcal{C}$), Threshold ($t$), Embedder ($E$)
$\mathbf{C}_e$ = Embed concepts in $\mathcal{C}$ via $E$
Store $\mathbf{C}_e$ in a dense index I:
if $|\mathcal{C}| \leqslant 500$ then
    I = FullSearchIndex                                ▷ Exact
else
    I = ApproximateSearchIndex                         ▷ k-NN
end if
Update I with $\mathbf{C}_e$                           ▷ Populate index
$\mathcal{R}$ = [ ]                                    ▷ Initialize matches collection
for $d \in \mathcal{D}$ do:                            ▷ Collect candidates
    Pre-process the document
    Parse the document
    (optional) Segment long documents
    X = Extract candidates from the document
    (optional) Filter noisy candidates from X
    $\mathbf{X}_e$ = Embed candidates in X with $E$
    $R_1$ = Run 1-NN search using I
    Filter out matches below the threshold $t$
    (optional) $R_2$ = Run appx. string match
    (optional) $R_3$ = Run Trie exact match
    $R_d = R_1 \cup R_2 \cup R_3$                      ▷ List of document matches
    Append $R_d$ to $\mathcal{R}$
end for
Return $\mathcal{R}$
```

on textual information, they lack the structural information of background ontology. Knowledge graph embeddings, discussed in Chapter 5, provide a more informative solution to capture such information and can easily be integrated with our work; however, we leave this to future work.

Index: We use approximate search in dense space with FAISS, which provides efficient search algorithms and scales up to billions of samples suitable for modern medical ontologies. There are two main components at its core: Vector Quantization (VQ), which significantly reduces initial search space to clustered sub-spaces. Second, the inverted file index points to the centroids with Product Quantization (PQ) to further speed up the search. Once the concepts are embedded, the resulting vectors are indexed, and candidates can be used as queries to find top-1 matches to obtain R1 in Algorithm 1. The overall framework is agnostic to embeddings, scalable, and efficient.

String Matching: Finally, we use a simple and efficient approximate dictionary matching algorithm called CPMerge (Okazaki and Tsujii, 2010) for τ-overlap join of inverted lists. Alternatively, we can also use the Faerie[11] (Li et al., 2011) algorithm based on the single and multi-heap based method with unified similarity and dissimilarity support. We also optionally combine approximate dictionary matching with exact string matching using Trie structure.

[11] https://github.com/suamin/PyNemex

2.4.5 *Experiments*

2.4.5.1 *Data*

We use two benchmark datasets to compare state-of-the-art unsupervised and supervised concept extractors, QuickUMLS and mBERT. A corpus of consumer-generated reviews was used that cover commonly used breast cancer drugs: Anastrozole, Exemestane, Letrozole, Raloxifene, and Tamoxifen; Drugs Review dataset (ADR; Yates and Goharian, 2013). We categorize this dataset as *lexical* since most concepts can be recognized with certain lexical variations. ADR is in English, but our approach can be applied to other languages off-the-shelf, whereas QuickUMLS is limited to English. For the multilingual experiment, we consider the German clinical dataset collected from the AnimalTestInfo database used in supervised concept extraction §2.3.4.1. The state-of-the-art on this dataset uses supervised fine-tuning and this data set is considerably more challenging than the lexical dataset as it requires a deeper understanding of semantics.

2.4.5.2 *Implementation*

Our implementation has the following core modules:

- `flashtext`[12] based exact matching.
- `sentence-transformers`[13] based multilingual XLM-R paraphrase embeddings for semantic matching. We select this based on the assumption that paraphrases capture some character and word order variations and that both semantic and lexical variations can jointly be captured.
- `FAISS`[14] based exact or approximate dense nearest neighbor search similar to Wu et al. (2020a).
- `textaCy`[15] based automatic KPE.
- `simstring`[16] for CPMerge based approximate string matching.
- `T2NER`[17] for training the CSD model, which we will introduce in the next Chapter (Amin and Neumann, 2021) using the BIO-scheme (Farber et al., 2008) while ignoring the entity types for MedMention with multilingual BERT for 3 epochs and 2e-5 learning rate.

2.4.6 *Results*

Table 4 shows the result on the ADR dataset. First, we see that CSD outperforms the KPE method in both cases. On ADR, our system achieved more than 12% F1 score improvement over QuickUMLS, whereas KPE based approach also outperformed QuickUMLS by 6% F1 score. It shall be noted that the KPE is unsupervised, but we use

[12] https://github.com/vi3k6i5/flashtext
[13] https://github.com/UKPLab/sentence-transformers
[14] https://github.com/facebookresearch/faiss
[15] https://github.com/chartbeat-labs/textacy
[16] https://github.com/nullnull/simstring
[17] https://github.com/suamin/T2NER

Method	Threshold (t)	F1	P	R	Speed (ms/doc)
QuickUMLS	0.9	0.48	0.47	0.50	**22**
DPM + KPE	0.7	0.54	0.61	0.53	52
DPM + CSD	0.8	**0.60**	**0.67**	**0.58**	105

Table 4: Unsupervised concept extraction result on *lexical* dataset.

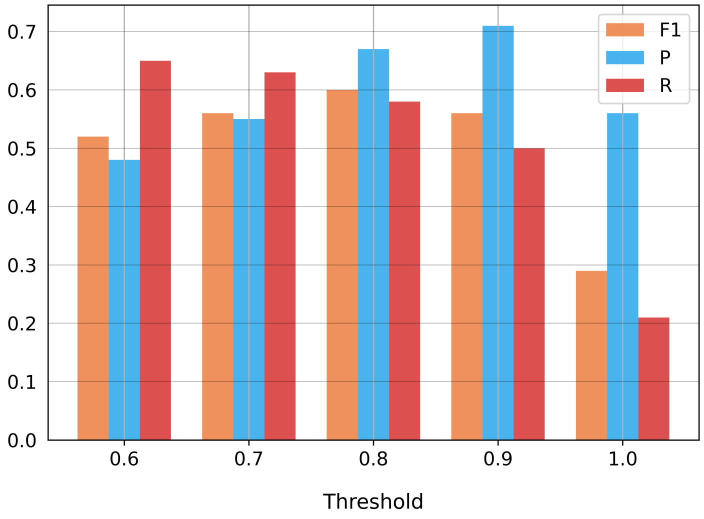


Figure 10: Precision-Recall trade-off with varying matching thresholds.

the external corpus for span detection without entity types in CSD. This limitation can be addressed by utilizing distantly supervised corpora (WikiMed, PubMedDS) as proposed in Vashishth et al. (2021). Therefore, our proposed method (DPM) has the potential to be utilized in an unsupervised fashion for concept extraction to partly answer **RQ2**.

We also note a trade-off between precision and recall as we vary the matching threshold between $[0.6 - 1.0]$ (Figure 10). The higher recall for CSD shows the model's ability to generalize to unseen concepts by understanding the context for enhanced candidate extraction. QuickUMLS is significantly faster as it only applies spaCy for linguistic annotation while we also perform text embedding, indexing, and hybrid search. In the case of CSD, we perform BERT-inference to tag sequences, which is expensive as it requires passing the input sequence through 12 transformer layers.

To directly compare our results from supervised experiments, Table 5 shows multilingual results. Since QuickUMLS lacks such support, we only compare DPM with our state-of-the-art supervised approach that fine-tunes mBERT (Amin et al., 2019). Despite only using noisy candidates and a pre-trained multilingual paraphrase model, our approach showed a zero-shot performance of 11% F1-score, which is dramatically lower than the supervised model with 65% F1-score trained on 5,000 labeled documents. This shows that the labeled data is essential for performant models and in particular for low-resource clinical text to address the remainder of **RQ2**. As we will empirically show in the next Chapter the strong *few-shot cross-lingual* transfer of mBERT for entity representation, a special case of a concept, it can also be investigated for few-shot concept extraction, which we leave as future work.

Method	F1	P	R
Supervised	0.65	0.54	0.83
Unsupervised	0.11	0.11	0.13

Table 5: Concept extraction results on a *semantic* multilingual German dataset, where the supervised results are taken from our work (Amin et al., 2019) on the original test set as shown in Table 3 above.

2.5 CONCLUSION

In this Chapter, we addressed an important task of clinical text mining, concept extraction. We conducted a thorough study of neural methods for supervised concept extraction showing the effectiveness of transfer learning with pre-trained multilingual and domain-specific language models, thus addressing **RQ1**. Such a transferable representation can significantly reduce the engineering required to develop domain-sensitive models. We further demonstrated that the automatic machine translation brought significant performance gain.

Considering our unsupervised setup in poorly resourced scenarios, we presented a hybrid framework utilizing contextual, and dictionary matching approaches with dense phrase matching. In cases where we are more interested in noun-phrased concepts, our approach is highly effective and has the potential to serve in a multilingual setup, thus addressing **RQ2**.

3 TRANSFORMER BASED NER TRANSFER WITH APPLICATION TO CROSS-LINGUAL CLINICAL DE-IDENTIFICATION

3.1 INTRODUCTION

Building on the previous Chapter, we now focus on a special case of concept, named entities in text. Named Entity Recognition (NER) is an important task in information extraction, benefiting the downstream applications such as entity linking (Cucerzan, 2007), relation extraction (Culotta and Sorensen, 2004), and question answering (Krishnamurthy and Mitchell, 2015). NER has been challenging in NLP due to large variations in entity names and flexibility in how entities are mentioned. These challenges are further enhanced in low-resource NER settings, such as for medical and multilingual text, where the added difficulty comes from the difference in text genre and entity names across languages and domains (Jia et al., 2019).

As discussed in the last Chapter, recent successes in transfer learning have mainly come from pre-trained language models with contextualized word embeddings based on deep transformer models (Devlin et al., 2019a; Radford et al., 2019). These models achieve state-of-the-art in several NLP tasks such as named entity recognition, document classification, and question answering. Due to their wide success and community adoption, successful frameworks like *Transformers* (Wolf et al., 2020) have emerged. In NER, the existing framework of *NCRF++* (Yang and Zhang, 2018) lacks the core infrastructure to directly support these models with state-of-the-art transfer learning algorithms.

Therefore, in the first part of this Chapter, we present an adaptable and general-purpose development framework, T2NER, for growing research in transfer learning with deep transformer models for NER. This is in contrast to the standard LSTM-based approaches, which have largely and successfully dominated the NER research. Our framework is aimed to bridge the gap in algorithmic advances in NER transfer with pre-trained language models to address **RQ3** in §3.3 as stated:

> *RQ3: How can we bridge the gap between research in pre-trained language models and algorithmic advances in NER transfer?*

In the second part of the Chapter, we address the real-world needs and challenges of clinical privacy with T2NER. Clinical texts contain rich information about patients, including their gender, age, profession, residence, family, and history, that is useful for record-keeping and billing purposes (Johnson et al., 2016; Shickel et al., 2017). We focus on removing Protected Health Information (PHI) from clinical texts, also called de-identification. We collect real patient data where the target texts are code-mixed (Spanish-Catalan) and domain-constrained (stroke). To avoid high annotation costs, we consider a more realistic setting where we annotate a gold evaluation corpus and a few hundred examples for training. Our approach is motivated by the empirical investigation of the strong performance of multilingual pre-trained language models in *few-shot cross-lingual transfer* for NER with high sample efficiency in comparison to supervised or unsupervised approaches to study **RQ4** in §3.4 as stated:

RQ4: What characteristic NER transfer property of multilingual pre-trained language models can effectively be applied to low-resource clinical de-identification?

The contents of §3.2.1 and §3.3 have appeared in the peer-reviewed article of ***Amin and Neumann (2021)****. The contents of §3.2.2 and §3.4 have appeared in the peer-reviewed article of* ***Amin et al. (2022b)****. These sections are included here with minor corrections where appropriate.*

3.2 RELATED WORK

3.2.1 *NER Transfer*

Transfer learning research in NER is an important and well-studied area due to two challenges. First, NER models show relatively high variance even when trained on the same domain data (Reimers and Gurevych, 2017). Second, these models poorly generalize when tested on data from different domains and languages, and even more so when they contain unseen entity mentions (Agarwal et al., 2020; Augenstein et al., 2017; Wang et al., 2020a). To cater to these issues, research has proposed several advances, including multi-task and joint learning (Jia et al., 2019; Lin et al., 2018; Pan et al., 2017; Peng and Dredze, 2017; Wang et al., 2020a), adversarial learning (Keung et al., 2019; Zhou et al., 2019), feature transfer (Daumé III, 2007; Kim et al., 2015; Wang et al., 2018c), newer architectures (Jia and Zhang, 2020; Lin et al., 2018), parameter sharing (Lee et al., 2018; Lin and Lu, 2018; Yang et al., 2018), parameter generation (Jia et al., 2019), mixture-of-experts (Chen et al., 2018), and usage of external resources (Wang et al., 2020c; Xie et al., 2018). We collectively label them as NER transfer algorithms.

3.2.2 *Clinical De-identification*

2014 i2b2/UTHealth (Stubbs and Uzuner, 2015), and the 2016 CEGS N-GRID (Stubbs et al., 2017) shared tasks explore the challenges of clinical de-identification on diabetic patient records and psychiatric intake records, respectively. Earlier works include machine learning and rule-based approaches (Meystre et al., 2010; Yogarajan et al., 2018), with Liu et al. (2017b) and Dernoncourt et al. (2017) being the first to propose neural models. Friedrich et al. (2019) proposed an adversarial approach to learn privacy-preserving text representations and Yang et al. (2019) used domain-specific embeddings trained on unlabeled corpora. While most works have mainly focused on English, some efforts have been made for Swedish using real patient data (Alfalahi et al., 2012; Velupillai et al., 2009) and for Spanish using a synthetic dataset introduced in the MEDDOCAN shared task (Marimon et al., 2019).

3.3 TRANSFORMER BASED TRANSFER LEARNING FRAMEWORK FOR NER

As discussed in the last Chapter, recent advances in deep transformer models have achieved state-of-the-art in several NLP tasks, where NER has traditionally benefited from LSTM networks. Concurrently, the research in NER transfer is making algorith-

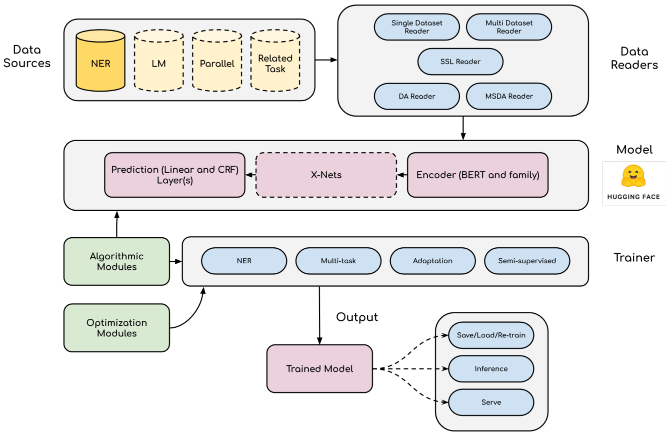


Figure 11: Overview of the T2NER framework.

mic advances, which can potentially benefit from performant encoders. Therefore, we present a Transformer based Transfer Learning Framework for Named Entity Recognition (T2NER) created in PyTorch for the task of NER with deep transformer models that concerns **RQ3**.[1]

3.3.1 *Design Principles*

T2NER is divided into several components, as shown in Figure 11. The core design principle is to seamlessly integrate the *Transformers* (Wolf et al., 2020) library as the backend for modeling while extending it to support different transfer learning scenarios with a range of existing algorithms. *Transformers* offer optimized implementations of several deep transformer models, including BERT (Devlin et al., 2019a), GPT (Radford et al., 2019), RoBERTa (Liu et al., 2019), and XLM (Conneau and Lample, 2019) among others, with multi-GPU, distributed, and mixed precision training.

The second design principle is inspired by transfer learning frameworks in computer vision, `Dassl.pytorch` (Zhou et al., 2020)[2] and `Trans-Learn` (Jiang et al., 2020)[3], that unify domain adaptation, domain generalization, and semi-supervised learning, thus allowing easy benchmarking, fair comparisons, and reproducibility. T2NER builds upon these transfer learning scenarios and offers a range of integrations (Figure 12).

The final design principle aims to unify the NER transfer research and offer a framework to test them with deep transformer models, wherever such an algorithmic abstraction is possible while exploring new paradigms.

3.3.2 *Data Module*

3.3.2.1 *Sources*

The primary data source is NER task data which is expected to be labeled or unlabeled in the CoNLL format. We adopt a widely used Begin-Inside-Outside (BIO) tagging

[1] `https://github.com/suamin/T2NER`
[2] `https://github.com/KaiyangZhou/Dassl.pytorch`
[3] `https://github.com/thuml/Transfer-Learning-Library`

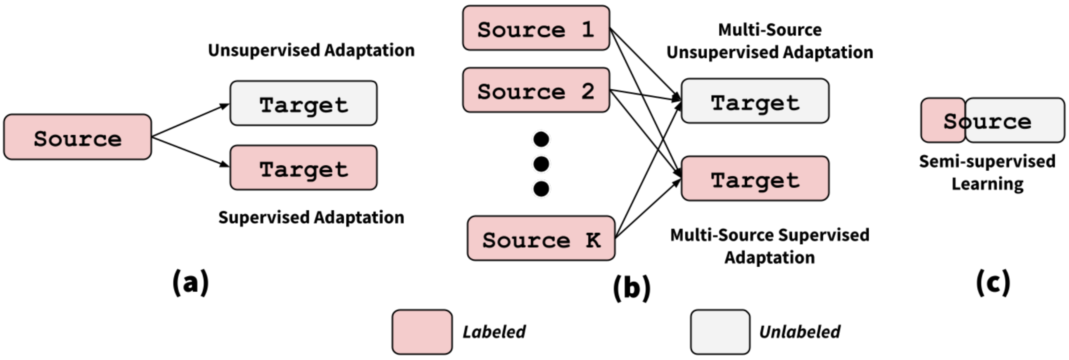


Figure 12: Transfer learning scenarios supported in T2NER. The adaptation scenarios apply to the *cross-domain*, *cross-lingual*, or a *mix* of both. These scenarios can further be complemented with multi-task learning. (a) Single source *supervised* or *unsupervised* domain or language adaptation (b) Multi-source *supervised* or *unsupervised* domain or language adaptation (c) Single source *semi-supervised* learning with partially labeled data. Further new directions in NER, such as multi-source adaptation with semi-supervised or few-shot learning of the target, are possible.

scheme (Farber et al., 2008). In practice, the differences in results arising from different schemas are negligible (Ratinov and Roth, 2009). A simple preprocessing routine is provided to standardize the data files and the required metadata used throughout the framework.

In particular, for a given named collection of the form `domain.datasetname`, possibly split into train, validation, and test files, T2NER creates output data files named as `lang.domain.datasetname-split` and `lang.domain.dataset name.labels`, where language information is provided by the user. A placeholder (`xxx`) can be used in place of missing domain or language metadata. We tokenize using *Transformers* and split sentences longer than the user-defined maximum length for preprocessing. An example output file can be of the form `en.news.conll-train`, referring to the CoNLL 2003 train set from the news domain in English (Tjong Kim Sang, 2002a). Besides NER data, additional task data can be provided, for example, language modeling, POS tagging, and language alignment resources (bilingual dictionaries or parallel sentences).

3.3.2.2 *Readers*

These classes are designed to serve the data needs of a given transfer learning scenario in a modular and extensible way. The framework provides `SimpleData`, `SimpleAdaptationData`, `MultiData`, and `SemiSupervisedData`, which are suitable for single dataset NER, cross-lingual and cross-domain NER, multi-task NER, and single dataset semi-supervised NER, respectively. Each class is derived from a base class `BaseData` and can be extended for other scenarios. As a concrete example, consider a dataset reader class `SimpleAdaptationData` in T2NER, which can provide training data for *source* and *target* language or domain up to a requested number of copies.

3.3.3 *Models*

A model is composed of three main components, a base encoder from *Transformers* (Wolf et al., 2020), an additional network (X-nets) on top of the encoder, useful in feature extraction based methods, and the prediction layers.

An encoder is the model backbone that takes as input tokenized text and returns hidden states such as those from BERT (Devlin et al., 2019a) or RoBERTa (Liu et al., 2019). There are five encoder modes that we support:

- `finetune`: Fine-tunes the encoder and uses the last layer's hidden states.
- `freeze`: Freezes the encoder and uses the last layer's hidden states.
- `firstn`: Freezes only the first n layers of the encoder and uses the last layer's hidden states (Wu and Dredze, 2019).
- `lastn`: Freezes the encoder and uses the aggregated hidden states by summing the outputs from the last n layers (Wang et al., 2020c).
- `embedonly`: Uses and fine-tunes the embedding layer of the encoder only.

X-nets are neural architectures that can optionally be modeled on top of the encoder to act as a feature extractor or pooler to process the hidden states. In T2NER, we provide a multi-layered transformer and bidirectional LSTM by default.

Prediction layers offer the final classification layer for the sequence labeling tasks. Following Devlin et al. (2019a), the default prediction layer in T2NER is linear, with additional support for a linear-chain Conditional Random Field (CRF) (Lafferty et al., 2001). In the multi-task setting, several output layers from different datasets in different domains or languages might be available with partial or exact entity types as outputs. To help the transfer across the tasks, *private* and *shared* prediction layers are also supported (Lin et al., 2018; Wang et al., 2020a).

With these underlying components, models are mainly implemented as single or multi-task architectures. To support a wide range of encoders in a unified API, T2NER adopts the `Auto` classes design from the *Transformers*. Figure 13 shows the class hierarchies, outlining the customized extensions with further possibilities to extend with external model implementations.

3.3.3.1 *Criterions*

For a given sequence of length L with tokens $x = [x_1, x_2, ..., x_L]$, labels $y = [y_1, y_2, ..., y_L]$, where $y_i \in \Delta^{\mathcal{C}}$ is a one-hot entity type vector with C types, and the linear prediction layer, the NER loss is defined as:

$$\mathcal{L}(y;x) = -\sum_{i=1}^{C}\sum_{j=1}^{L} y_{ij} \log p(h_j = i|x_j)$$

where $p(h_j = i|x_j)$ is the probability of token x_j being labeled as entity type i and h_j is the model output. When p is softmax, this becomes a cross-entropy loss. To tackle class imbalance in real-world applications, T2NER also offers two-class sensitive loss functions extensions for token classification:

- Focal loss (Lin et al., 2017) adds a modulating factor to the standard softmax, which reduces the loss contribution from easy examples and extends the range in which an example receives low loss.
- LDAM loss (Cao et al., 2019) is a label-distribution-aware function that encourages the model to have the optimal trade-off between per-class margins by promoting the minority classes to have larger margins.

3.3.3.2 *Auxiliary Tasks*

Multi-task learning has systematically proved beneficial for NER transfer (Jia et al., 2019; Jia and Zhang, 2020; Lin et al., 2018; Wang et al., 2020a). Several auxiliary tasks are supported in a multi-task model by default:

- *Language Classification*: In the cross-lingual setting, this task provides an additional classification signal over the languages (e.g., English and Spanish) used in the training data (Keung et al., 2019).
- *Domain Classification*: In the cross-domain setting, this task provides an additional classification signal over the domains (e.g., News and Biomedical) used in the training data (Wang et al., 2020a).
- *Adversarial Classification*: In the cross-lingual or cross-domain setup, this task provides an additional adversarial classification signal over the languages or domains to learn invariant features used in the training data (Chen et al., 2018; Keung et al., 2019).
- *Language Modeling*: While pre-trained transformer models are already trained on specific corpora, this task adds causal language modeling signal during fine-tuning over the raw task texts (Jia et al., 2019; Jia and Zhang, 2020; Rei, 2017).
- *Entity Type Classification*: To better extract entity type knowledge, an additional linear classifier is added. This performs classification over entity types such as [PER, LOC, ORG, ...] without the segmentation tags such as B/I/E (Jia and Zhang, 2020).
- *Shared Tagging*: In NER settings where the entity types might differ, a shared prediction layer across all the entity types provides an additional signal to the base NER tasks.
- *All-Outside Classification*: This binary classification task predicts if the sentence has entity types other than the outside (O) type.

3.3.3.3 *Optimization Modules*

T2NER provides thin wrappers around the optimizers and learning rate schedulers from PyTorch and *Transformers*.

3.3.4 *Algorithms*

A trainer is the central class concept that binds together all the components and provides a unified setup to develop, test, and benchmark the algorithms. Figure 13 shows the hierarchy of trainer classes. Each transfer learning scenario inherits from the `BaseTrainer` class, where each scenario can further be extended to create an algorithm-specific training regime. This allows the researchers to focus mainly on the algorithms' logic while the framework fulfills the requirements of a chosen transfer scenario. Following Jiang et al. (2020) and Zhou et al. (2020), a few training algorithms are implemented by default which we briefly describe. In the following, a feature extractor is referred to as the base encoder with any X-nets. An optional pooling strategy {`mean`, `sum`, `max`, `attention`, ...} can be applied to aggregate the hidden states.

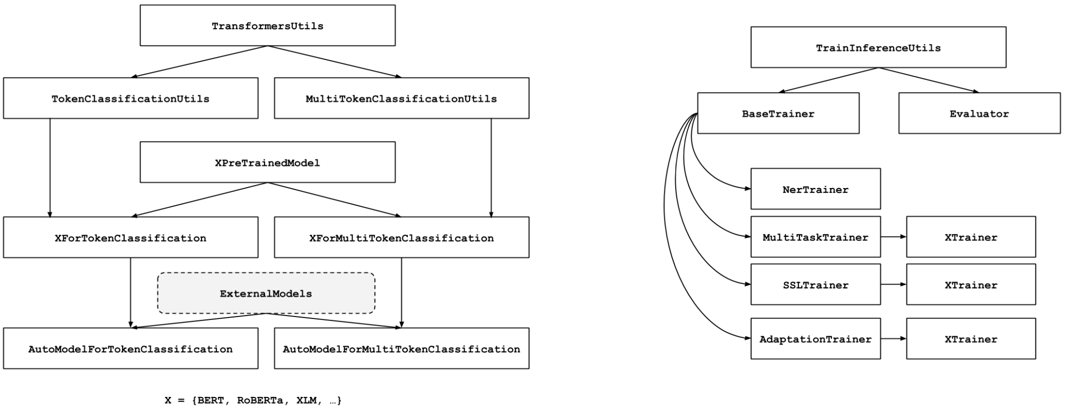


Figure 13: Class hierarchies in T2NER for two main class concepts. (*Left*) Main model architectures in single and multi-task settings with the adoption of `Auto` classes concepts from *Transformers* (Wolf et al., 2020), where customized functionality or new modeling concepts can easily be added. (*Right*) Main trainer classes offer a particular transfer learning scenario and extend it to a specific transferring algorithm.

In the following, domain and language can be used interchangeably. For the sake of the discussion, we use the word domain.

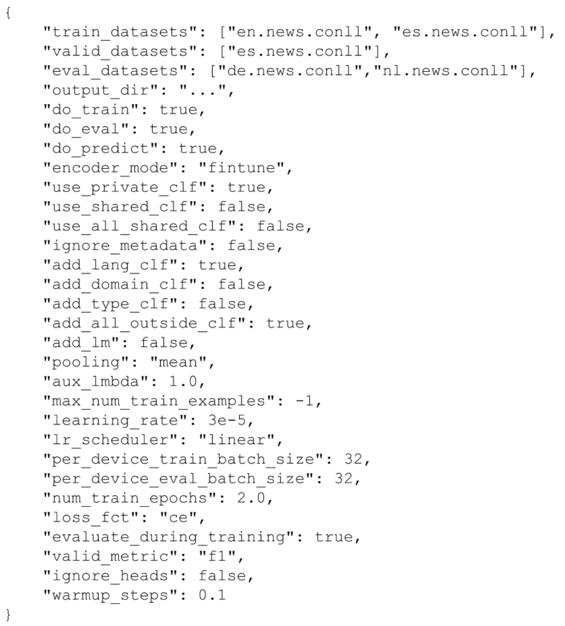
Gradient Reversal Layer (GRL) adds a domain classifier trained to discriminate whether input features come from the source or target domain. In contrast, the feature extractor is trained to confuse the domain classifier into matching feature distributions (Ganin and Lempitsky, 2015).

Earth Mover Distance (EMD) adds a critic that maximizes the difference between unbounded scores of source and target features. This effectively returns the approximation of Wasserstein distance between source and target feature distributions (Arjovsky et al., 2017). The overall objective jointly minimizes NER cross-entropy loss and Wasserstein distance. Theoretically, GRL effectively minimizes Jensen-Shannon (JS) divergence, which suffers from discontinuities and thus provides poor gradients for the feature extractor. In contrast, Wasserstein distance is stable and less prone to hyperparameter selection (Chen et al., 2018). For stable training, a common strategy is to add a gradient penalty (Gulrajani et al., 2017), which is also provided in T2NER.

Keung Adversarial is closely related to GRL but additionally uses the generator loss such that the features are difficult for the discriminator to classify correctly between source and target. The optimization is carried out in a step-wise fashion for the feature extractor, discriminator, and generator (Keung et al., 2019).

Maximum Classifier Discrepancy (MCD) adds a second classifier to measure the discrepancy between the predictions of two classifiers on target samples. This is based on the observation that two different classifiers can measure the target samples outside the support of the source. Overall, MCD solves a *minimax* problem in which the goal is to find two classifiers that maximize the discrepancy on the target sample and a feature generator that minimizes this discrepancy (Saito et al., 2018).

Minimax Entropy (MME) decreases the entropy on unlabeled target features in an adversarial manner by using GRL to obtain high-quality discriminative features (Saito et al., 2019). Besides unsupervised domain adaptation, the method can be used in semi-supervised and few-shot learning scenarios when some labeled target samples are available.

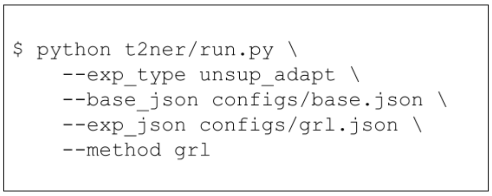


```
{
    "train_datasets": ["en.news.conll", "es.news.conll"],
    "valid_datasets": ["es.news.conll"],
    "eval_datasets": ["de.news.conll","nl.news.conll"],
    "output_dir": "...",
    "do_train": true,
    "do_eval": true,
    "do_predict": true,
    "encoder_mode": "fintune",
    "use_private_clf": true,
    "use_shared_clf": false,
    "use_all_shared_clf": false,
    "ignore_metadata": false,
    "add_lang_clf": true,
    "add_domain_clf": false,
    "add_type_clf": false,
    "add_all_outside_clf": true,
    "add_lm": false,
    "pooling": "mean",
    "aux_lmbda": 1.0,
    "max_num_train_examples": -1,
    "learning_rate": 3e-5,
    "lr_scheduler": "linear",
    "per_device_train_batch_size": 32,
    "per_device_eval_batch_size": 32,
    "num_train_epochs": 2.0,
    "loss_fct": "ce",
    "evaluate_during_training": true,
    "valid_metric": "f1",
    "ignore_heads": false,
    "warmup_steps": 0.1
}
```

Figure 14: An example configuration file showing an instantiation of the multi-task cross-lingual adaptation of CoNLL datasets from English and Spanish to German and Dutch for zero-shot transfer.

Other algorithms, such as classical Conditional Entropy Minimization (CEM) for semi-supervised learning (Grandvalet and Bengio, 2004) or recent works based on Maximum Mean Discrepancy (MMD) for multi-source domain adaptation (Peng et al., 2019a) can be added. In general, extending T2NER to newer algorithms is simple and flexible, addressing the gap between research in pre-trained language models and algorithmic advances as stated in **RQ3**.

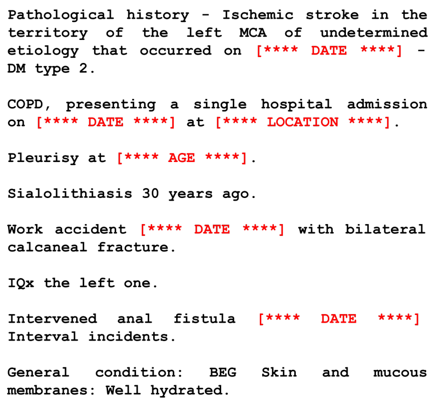
T2NER offers a single entry point to the framework, which relies on a base JSON configuration file, and an experiment-specific JSON configuration file with an optional algorithm name to run. An example experiment-specific configuration file is shown in Figure 14. The command below shows an example run:



```
$ python t2ner/run.py \
    --exp_type unsup_adapt \
    --base_json configs/base.json \
    --exp_json configs/grl.json \
    --method grl
```

Figure 15: Example command showing T2NER usage for unsupervised adaptation with *gradient reversal layer* method.

Similar to other frameworks, it can be further developed and used as a standard Python library. In the next section, we shift our focus to a real-world use case of T2NER for few-shot cross-lingual clinical de-identification in a domain-specific setup.



Pathological history - Ischemic stroke in the territory of the left MCA of undetermined etiology that occurred on [**** DATE ****] - DM type 2.

COPD, presenting a single hospital admission on [**** DATE ****] at [**** LOCATION ****].

Pleurisy at [**** AGE ****].

Sialolithiasis 30 years ago.

Work accident [**** DATE ****] with bilateral calcaneal fracture.

IQx the left one.

Intervened anal fistula [**** DATE ****] Interval incidents.

General condition: BEG Skin and mucous membranes: Well hydrated.

Figure 16: The process of text de-identification, solved as NER, involves the removal of a predefined set of direct identifiers in text (Elliot et al., 2016). For clinical notes, this set is often the PHI categories (or types) defined by the Health Insurance Portability and Accountability Act (HIPAA) (Gunn et al., 2004). The example here shows a de-identified excerpt of a patient note from the Spanish-Catalan stroke dataset used in our study. The text is translated into English for readability.

3.4 CROSS-LINGUAL CLINICAL DE-IDENTIFICATION

With growing interest and innovations in data-driven digital technologies, privacy has become an important legal topic for technology to be regulations compliant. In Europe, the General Data Protection Regulation (GDPR) (Regulation, 2016) requires data owners to have a legal basis for processing personally identifiable information (PII), which also includes the explicit consent of the subjects. In cases where explicit consent is impossible, anonymization is often seen as a resorted-to solution.

GDPR-compliant anonymization requires the complete and irreversible removal of any information that may lead to a subject's data being identified (directly or indirectly) from a dataset (Elliot et al., 2016). However, de-identification is limited to removing specific predefined direct identifiers; further replacement of such direct identifiers with pseudonyms is referred to as pseudonymization (Alfalahi et al., 2012). Generally, de-identification can be seen as a subset of anonymization despite interchangeable usage of the terms in the literature (Chevrier et al., 2019). We focus on solving the problem of de-identification in the clinical domain as a sequence labeling task, specifically NER (Lample et al., 2016).

As outlined in Lison et al. (2021), a significant challenge in clinical text de-identification is the lack of labeled data. These challenges are further pronounced in a multilingual or cross-lingual setup with a clinical sub-domain. Hartman et al. (2020) showed that a small number of manually labeled PHI examples could significantly improve performance. In parallel, prior works in few-shot NER consider the problem where a model is trained on one or more source domains and tested on unseen domains with a few labeled examples per class, some with entity tags different from those in the source domains (Yang and Katiyar, 2020). Models are trained with prototypical methods, noisy supervised pre-training, or self-labeling (Huang et al., 2020). In contrast, we consider a setting where the target and source domains share the *same* entity (PHI) tags but with a few labeled examples in the target language. We take this approach to study the needs of **RQ4**. A similar setup has been employed in few-shot question answering (Ram et al., 2021).

3.4.1 *Problem Definition*

We approach the de-identification problem as an NER task. Given an input sentence $\mathbf{x}$ with N words: $\mathbf{x} = [x_i]_{i=1:N}$, we feed it to a T2NER encoder $f_\phi : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times d}$ to obtain a sequence of hidden representations $\mathbf{h} = [h_i]_{i=1:N}$

$$\mathbf{h} = f_\phi(\mathbf{x}).$$

We then feed $\mathbf{h}$ into an NER classifier, which is a linear classification layer with the `softmax` activation function to predict the PHI label of $\mathbf{x}$:

$$p_\theta(\mathbf{Y}|\mathbf{x}) = \mathtt{softmax}(\mathbf{W}^T\mathbf{h} + \mathbf{b}).$$

$p_\theta(\mathbf{Y}|\mathbf{x}) \in \mathbb{R}^{N \times |\mathcal{P}|}$ is the probability distribution of PHI labels for sentence $\mathbf{x}$ and $\mathcal{P}$ is the PHI label set. $\theta = [\phi, \mathbf{W} \in \mathbb{R}^{d \times |\mathcal{P}|}, \mathbf{b} \in \mathbb{R}^{|\mathcal{P}|}]$ denote the set of learnable parameters and d being the hidden dimension. The model is trained to minimize the per-sample negative log-likelihood:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(Y_i = y_i|x_i). \tag{1}$$

For pre-trained LMs, this setting corresponds to NER fine-tuning (Wu and Dredze, 2019). When we jointly fine-tune on more than one NER dataset, we refer to it as multi-task learning following T2NER.

Definition 1 (**Few-Shot NER**). *Given an entity label set $\mathcal{P}$, we define the task of few-shot NER as having access to $K \leqslant M$ labeled sentences containing each element $p \in \mathcal{P}$ at least once, where K is a small number (e.g., in $[50, 500]$) and M is orders of magnitude larger (e.g., $\geqslant 1000$).*

Definition 2 (**Few-Shot NER Transfer**). *Given an NER dataset in a source domain (or language), we define the task of few-shot cross-domain (or cross-lingual) NER transfer as adapting a model trained on the source domain (or language) to a target domain (or language) with access to a few-shot corpus (Def. 1).*

The Few-Shot NER Transfer setting is different from prior studies in NER transfer, including few-shot (Huang et al., 2020), unsupervised (Keung et al., 2020), and semi-supervised NER (Amin and Neumann, 2021).

3.4.2 *Few-Shot Cross-Lingual NER Transfer*

mBERT (Devlin et al., 2019a) has been shown to achieve robust performance for zero-shot cross-lingual transfer tasks, including NER (Pires et al., 2019; Wu and Dredze, 2019). Adversarial learning has been applied with limited gains (Keung et al., 2019) in unsupervised approaches to improve zero-shot NER transfer, whereas feature alignments have shown better results (Wang et al., 2020c). Meta-learning with minimal resources (Wu et al., 2020c) and word-to-word translation (Wu et al., 2020b) have shown further performance gains. The current state-of-the-art approach of Chen et al. (2021b) combines token-level adversarial learning with self-labeled data selection and knowledge distillation.

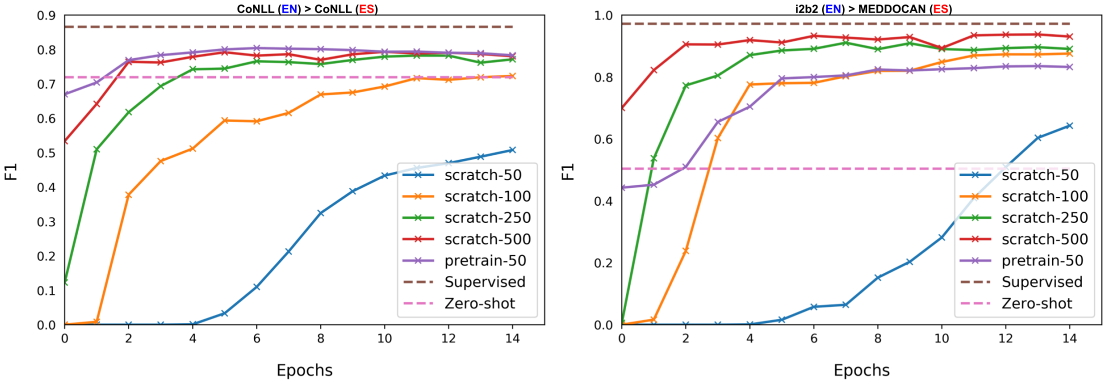


Figure 17: An empirical investigation of few-shot cross-lingual NER transfer in mBERT. We compare different transfer learning scenarios from English (EN) to Spanish (ES) for two pairs of datasets as a preliminary study to investigate the effectiveness of the few-shot cross-lingual transfer in mBERT: CoNLL-2003 to CoNLL-2002 (*left*) and i2b2-2014 to MEDDOCAN (*right*). We use supervised fine-tuning on the entire training set of the target language (ES) as the *upper bound* and the zero-shot score of the model. (pre-trained only on the source language (EN) training set) as the *lower bound* for target language (ES) performance. We then consider 50, 100, 250, and 500 examples from the target language as few-shot training corpora and train the models for 15 epochs. The models without any pre-training on the source corpus (scratch) eventually outperform the *lower bound* as the number of examples grow; with sufficient epochs, the model with only 50 target-language samples reaches more than 10% gain in the de-identification task (*right: scratch-50*). However, we find the cross-lingual transfer-learning strategy to be most sample efficient when pre-trained on *source* language—50-shot performance (*left & right: pretrain-50*) comparable to 500-shot (*left & right: scratch-500*). We apply this strategy to address the real-world challenges of Spanish-Catalan de-identification.

CoNLL (EN) → CoNLL (ES)	F1
Pires et al. (2019)	73.59
Wu and Dredze (2019)	74.96
Keung et al. (2019)	74.30
Wang et al. (2020c)	75.77
Wu et al. (2020c)	77.30
Wu et al. (2020b)	76.75
Chen et al. (2021b)	**79.00**
few-50 (or pretrain-50)	78.30

Table 6: Cross-lingual transfer results on CoNLL. *few-50* represents our fine-tuning of EN-trained mBERT with 50 random labeled samples from ES.

To investigate the few-shot transferability of mBERT, we consider two pairs of datasets with English as the source language and Spanish as the target language: the CoNLL-2003/CoNLL-2002 (Tjong Kim Sang, 2002b,c) in the *general domain* and i2b2/MEDDOCAN (Marimon et al., 2019; Stubbs and Uzuner, 2015) in the *clinical domain*. We report the results of our preliminary study in Figure 17. We observed that with as few as 50 random labeled training samples from the target language, we obtained substantial gains for both datasets, with near state-of-the-art on CoNLL (Table 6). We refer to this as *few-shot cross-lingual transfer* property of mBERT for NER, thus partially addressing **RQ4**. Our study highlights that the property holds for different domains (general and clinical), where the latter focuses on the de-identification task. We leave a large-scale study on more datasets with different languages and domains as future work.

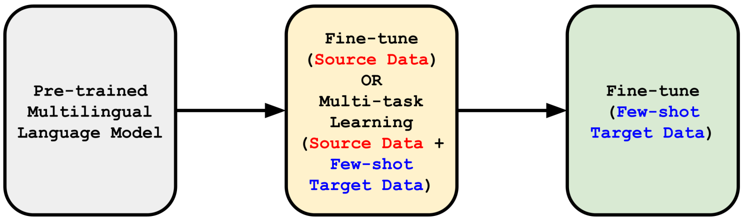

Pre-trained Multilingual Language Model
Fine-tune (Source Data) OR Multi-task Learning (Source Data + Few-shot Target Data)
Fine-tune (Few-shot Target Data)

Figure 18: Our *few-shot cross-lingual transfer* strategy for clinical text de-identification.

Compared to supervised (unsupervised) methods, which use complete labeled (unlabeled) target data, our few-shot approach is *sample-efficient* and alleviates the need for complex pipelines (Chen et al., 2021b; Wu et al., 2020b,c) and large-scale annotations. Furthermore, Keung et al. (2020) highlights the spurious effects of using source data as a development set and recommends using target data as a development set for model selection in NER transfer. Our findings and those in Hartman et al. (2020) motivate us to *(a)* propose an optimal *few-shot cross-lingual transfer* strategy (outlined in Figure 18), *(b)* annotate a target development set, and *(c)* construct an annotated *few-shot target corpus* for effective cross-lingual transfer learning.

3.4.3 *Data and Annotation*

Our dataset consists of stroke patient records collected at Institut Guttmann.[4] Table 7 summarizes the raw data statistics and Table 31 in Appendix §A.1 describes the topics present in the texts. We set aside 100 randomly sampled notes for out-of-sample generalizability evaluation. The remainder of the notes are considered for our development and few-shot corpora sampling; 396k sentences are tokenized in the process.

We follow a protocol similar to Gao et al. (2021b), which uses a fine-tuned BERT model's predictions for constructing a manually annotated dataset to evaluate distantly supervised relation extraction models. In particular, we train mBERT on the MEDDOCAN corpus, using coarse-grained PHI categories {DATE, AGE, LOCATION, NAME, CONTACT, PROFESSION, ID} with the BIO scheme (Farber et al., 2008), for evaluation and few-shot training data selection. We use the trained model to make predictions on the dataset and observe that the model predicts PHI on only 50k out of the 396k sentences. A dataset of 5000 sentences ($<$ 2% of raw sentences) is constructed from a mix of randomly sampled 2500 sentences from this 50k and 2500 from the remaining sentences.

We split the dataset into two partitions of 2500 sentences for independent annotation by two annotators. The annotation is performed one sentence at a time by applying one of the 7 coarse-grained PHI labels to each token using the T2NER-ANNOTATE toolkit (Amin and Neumann, 2021). Each annotator's confidence level between 1-5 is recorded for the token-level labels for each sample. To record the inter-annotator agreement, we use token-level Cohen's kappa (Cohen, 1960) statistic reaching a value of 0.898. The two annotators agreed on 3924 sentences, resulting in our final evaluation set. To save annotation costs for developing a *few-shot target corpus*, we resolved the disagreements to obtain a 384-sentence few-shot corpus for training (see Appendix §A.2 for annotation details).

[4] https://www.guttmann.com/ca/institut-universitari-guttmann-uab

Patients	Notes	ES	CA	Other
1,500	327,775	42.8%	53.0%	4.2%

Table 7: Raw statistics of the Spanish (ES)-Catalan (CA) data from the stroke domain.

Our source dataset (the MEDDOCAN corpus) consists of 1000 synthetically generated clinical case studies in Spanish (Marimon et al., 2019). A practicing physician selected the corpus manually and augmented it with PHI from discharge summaries and clinical records. In contrast, our target corpus focuses on the stroke domain and contains PHI from real-world records. Since the target data is code-mixed between Spanish and Catalan, with the majority (53%) being Catalan, the transfer from Spanish source data (MEDDOCAN) is cross-lingual.[5]

3.4.4 *Experiments*

We conduct our experiments with the T2NER framework (Amin and Neumann, 2021).[6] For the baseline, we consider zero-shot performance on the evaluation set of the mBERT encoder fine-tuned on the MEDDOCAN training set consisting of 16,299 samples. We then fine-tune it on the few-shot target corpus as outlined in Figure 18. Following the multi-task learning (Lin et al., 2018) approach in T2NER, we jointly fine-tune mBERT on the MEDDOCAN and few-shot target corpora. Since the few-shot corpus is much smaller, multi-task learning helps the model transfer. It further acts as a regularization approach by sharing parameters between the datasets. To improve performance on the target data, we further fine-tune with the few-shot target corpus after the first step of fine-tuning to have improved target performance; for the model to be an expert in target, (Cao et al., 2020). All the models are trained for 3 epochs with a learning rate of 3e-5 and linear warm-up of 10%. For few-shot fine-tuning only, the model is trained for 25 epochs.

3.4.5 *Results*

Table 8 shows our results. Fine-tuning the baseline mBERT model with the few-shot target corpus improves the F1-score from 73.7% to 88.6%, a substantial gain of 14.9%, highlighting the effectiveness of *few-shot cross-lingual transfer* with mBERT and addressing the remainder of **RQ4**. The significant increase in recall (26% points) compared to precision (3.4% points) suggests an increase in the model's capacity to recognize domain-specific entities. Multi-task fine-tuning improves the F1-score to 89.5%; further fine-tuning on the few-shot target corpus boosts the best model's performance to 91.2%. Figure 19 shows per-PHI-label scores on the development set and their frequency. The model performs almost perfectly on DATE and AGE since most DATE and AGE labeled segments are similar between Spanish and Catalan as they are simple numbers (for DATE) and numbers followed by the word (for AGE; '*edad*' in both Spanish and Catalan). There are differences in time expressions, e.g., day of the week, as the words are distinctly dissimilar. However, structurally there is only a slight difference. Further, the model struggles with the ID class due to the low sample size (5 instances in the

[5] Although similar, Spanish and Catalan are distinct languages. The domain of MEDDOCAN is missing an explicit mention in Marimon et al. (2019), therefore it is omitted.

[6] `https://github.com/suamin/T2NER`

Transfer Strategy	Precision	Recall	F1
Fine-tune (M)	80.1	68.2	73.7
Fine-tune (M) → Fine-tune (F)	83.5	94.2	88.6
Multi-task (M + F)	86.0	93.3	89.5
Multi-task (M + F) → Fine-tune (F)	**87.7**	**95.0**	**91.2**

Table 8: Results on the development set from the code-mixed stroke data. M denotes the MEDDOCAN (Marimon et al., 2019) training set (source) normalized to 7 PHIs (see Appendix §A.3) at the sentence level, and F denotes our few-shot target corpus. Here multi-task learning refers to the joint fine-tuning of two datasets following T2NER.

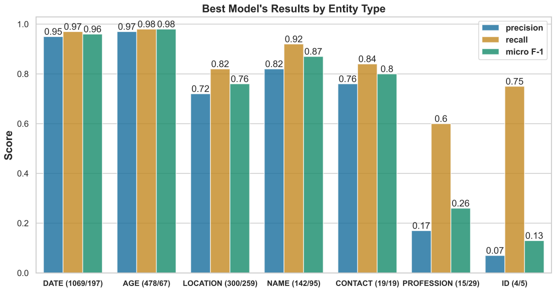


Figure 19: NER metrics on the evaluation set for each entity type with their frequency in development/few-shot training sets.

few-shot corpus). It is generally challenging to disambiguate between an alphanumeric string and a PHI ID, as also noted by the ID class' high recall. Our error analysis reveals high false positives for the PROFESSION label in Catalan, e.g. '*Coloma de Gramenet*' (a LOCATION) and '*Dialogant*' (being able to communicate) are both labeled as PROFESSION. Further, we tokenize the 100 out-of-sample notes into sentences to test model generalization and make predictions with our best model. The resulting annotated sentences are reconstructed into patient notes and manually evaluated by two reviewers (one external and one annotator) for occurrences of true and false positives and negatives. The model achieves precision, recall, and F1 scores of 95.1%, 99.3%, and 97.1%, respectively, on the out-of-sample notes, highlighting the effectiveness of our approach.

3.5 CONCLUSION

In this Chapter, we presented a Transformer based framework for transfer learning research in NER. We laid out the design principles, detailed the architecture, and presented the transfer scenarios and representative algorithms. T2NER offers to bridge the gap between growing research in deep transformer models, NER transfer, and domain adaptation, thus addressing **RQ3**.

We then applied T2NER to the task of clinical notes de-identification in a low-resource scenario where the target texts are code-mixed (Spanish-Catalan), domain-constrained (stroke), and lack a cost-prohibitive large-scale annotation. By empirically investigating the *few-shot cross-lingual transfer* property of mBERT, we proposed an adaptation strategy that significantly boosts zero-shot performance and offers generalizability while keeping the required size of annotated samples low, thus addressing **RQ4**.

Part II

RELATION-CENTRIC LEARNING

4 SCIENTIFIC LANGUAGE MODELS FOR DISTANTLY SUPERVISED BIOMEDICAL RELATION EXTRACTION

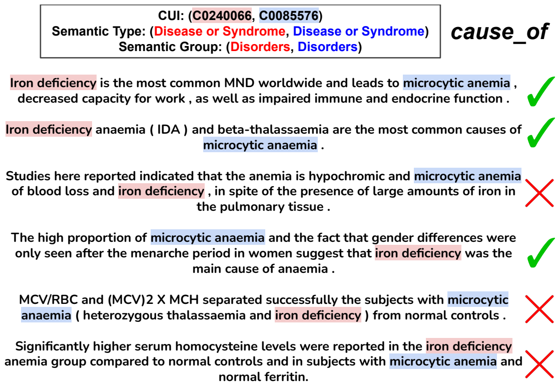
4.1 INTRODUCTION

In Part-I, we developed low-resource entity-centric learning approaches with clinical applications. Starting with this Chapter, we assume that a knowledge base has been constructed partially or entirely where we switch our focus on relation-centric learning now in the biomedical domain. One important information extraction task is the mining of structured data from unstructured text for knowledge discovery and management. In this regard, scientific literature offers rich interactions between entities mentioned in the text (Craven, Kumlien, et al., 1999; Xu and Wang, 2014), which can be helpful for applications such as bio-molecular information extraction, pharmacogenomics, and identifying Drug-Drug Interactions (DDIs), among others (Luo et al., 2017b).

Manually annotating these relations for training supervised learning systems is an expensive and time-consuming process (Kilicoglu et al., 2011; Li et al., 2016; Segura-Bedmar et al., 2011, 2013). Distant Supervision (DS) provides a useful way to obtain large-scale data for RE (Mintz et al., 2009). However, DS for data collection also tends to result in an increased amount of noise, as the target relation may only sometimes be expressed (Ritter et al., 2013; Takamatsu et al., 2012). As individual instance labels are unknown (Wang et al., 2018a), a common strategy is to use Multi-Instance Learning (MIL) by aggregating relational sentences into a bag representation for classification. Since a knowledge base is used for distant supervision and mutual learning between text and KG has been shown to reduce noise (Dai et al., 2019; Han et al., 2018a), we aim to directly encode this implicit knowledge using positional markings and latent relation direction (see Figure 21). Specifically, we propose sentence-level Relation enriched BERT (Wu and He, 2019) to bag-level MIL (MIL-RBERT) for biomedical RE with KB-sensitive markings, *k-tag*, to address **RQ5** in §4.3 as stated:

> *RQ5: Can a KB be utilized for denoising relation representations from domain-specific language models for distantly supervised biomedical RE?*

Secondly, to scale to a large number of biomedical entities, broad-coverage benchmarks have been proposed, including ours in MIL-RBERT, for whom we investigate train-test leakage of knowledge graph triples and find significant portions overlapping. Such leakage impacts the model performance as it allows it to score higher by simply memorizing the training relations rather than generalizing to new, previously unknown ones. We identify the sources of these issues as normalizing the textual form of concept mentions to their unique identifiers and improper handling of inverse relations. More accurate benchmarks exist (Hong et al., 2020; Marchesin and Silvello, 2022) but focus on narrower types of interactions. To alleviate the training-test data leakage in established broad coverage benchmarks and clean but narrow coverage benchmarks, we present a new benchmark MedDistant19 which draws its knowledge graph from the widely used healthcare ontology SNOMED CT (Chang et al., 2020). Further, with



CUI: (C0240066, C0085576)
Semantic Type: (Disease or Syndrome, Disease or Syndrome)
Semantic Group: (Disorders, Disorders)

Iron deficiency is the most common MND worldwide and leads to microcytic anemia , decreased capacity for work , as well as impaired immune and endocrine function . ✓

Iron deficiency anaemia (IDA) and beta-thalassaemia are the most common causes of microcytic anaemia . ✓

Studies here reported indicated that the anemia is hypochromic and microcytic anemia of blood loss and iron deficiency , in spite of the presence of large amounts of iron in the pulmonary tissue . ✗

The high proportion of microcytic anaemia and the fact that gender differences were only seen after the menarche period in women suggest that iron deficiency was the main cause of anaemia . ✓

MCV/RBC and (MCV)2 X MCH separated successfully the subjects with microcytic anaemia (heterozygous thalassaemia and iron deficiency) from normal controls . ✗

Significantly higher serum homocysteine levels were reported in the iron deficiency anemia group compared to normal controls and in subjects with microcytic anemia and normal ferritin. ✗

Figure 20: An example of a bag instance representing the UMLS concept pair (`C0240066`, `C0085576`) expressing the relation *cause_of*. In this example, three out of six sentences express the relation, while others are incorrect labels resulting in noise from the distant supervision.

the success of scientific language models for biomedical and clinical tasks (Gu et al., 2021), and inspired by existing RE evaluation studies in the general domain (Alt et al., 2020; Gao et al., 2021a; Peng et al., 2020), we conduct a thorough analysis to address **RQ6** in §4.4 as stated:

> *RQ6: Are there limitations to accurately evaluate domain-specific language models for broad-coverage distantly supervised biomedical RE?*

> *The contents of §4.2.1 and §4.3 have appeared in the peer-reviewed article of **Amin et al. (2020a)**. The contents of §4.2.2 and §4.4 have appeared in the peer-reviewed article of **Amin et al. (2022a)**. These sections are included here with minor corrections where appropriate.*

4.2 RELATED WORK

4.2.1 *Distantly Supervised RE*

Relation extraction is an important task in NLP. Traditionally, supervised methods require large-scale annotated corpora, whereas distant supervision allows for the automated collection of potentially noisy training examples by aligning a given knowledge base with a collection of text sources (Mintz et al., 2009). Such a form of weak supervision is combined with multi-instance learning by creating a *bag* of instances (Riedel et al., 2010) for corpus-level triple extraction.[1] Earlier works rely on the assumption that at least one of the evidence samples in a bag represents the target relation in a triple (Hoffmann et al., 2011; Riedel et al., 2010; Surdeanu et al., 2012). Recently, Piecewise Convolutional Neural Networks (PCNN) (Zeng et al., 2014) have been applied to DS (Zeng et al., 2015), with notable extensions in selective attention (Lin et al., 2016) and the modeling of noise dynamics (Luo et al., 2017a). Han et al. (2018a) proposed a joint learning framework for Knowledge Graph Completion (KGC), studied in the next

[1] RE is used to refer to two different tasks: sentence-level detection of relational instances and corpus-level triples extraction, i.e. a knowledge graph completion task (Amin et al., 2020b) but from the text.

Chapter, and RE with mutual attention, showing that DS improves downstream KGC performance. At the same time, KGC acts as an indirect signal to filter textual noise.

Relevant to distant RE with pre-trained language models, Alt et al. (2019) extended the OpenAI Generative Pre-trained Transformer (GPT) model (Radford et al., 2019) for bag-level MIL with selective attention (Lin et al., 2016). Sun et al. (2019a) enriched the pre-training stage with KB entity information, resulting in improved performance. For sentence-level RE, Wu and He (2019) proposed an entity marking strategy for BERT (referred to as RBERT) to perform relation classification. Specifically, they mark the entity boundaries with special tokens following the order they appear in the sentence. Likewise, Baldini Soares et al. (2019) studied several data encoding schemes and found marking entity boundaries important for sentence-level RE. With such an encoding, they proposed a novel pre-training scheme for distributed relational learning suited for few-shot relation classification (Han et al., 2018b).

4.2.2 *Broad-Coverage Biomedical RE*

In the biomedical domain, rule-based (Abacha and Zweigenbaum, 2011; Kilicoglu et al., 2020) and weakly supervised approaches (Peng et al., 2016) have been proposed. Roller and Stevenson (2014) first proposed the use of DS with the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) as a KB and PubMed (Canese and Weis, 2013) MEDLINE abstracts as text collection and showed promising results.

One major challenge in biomedical RE is to scale for broad-coverage (Kilicoglu et al., 2011, 2020). First Dai et al. (2019) implemented a knowledge-based attention mechanism (Han et al., 2018a), using improved KG models, ComplEx (Trouillon et al., 2017) and SimplE (Kazemi and Poole, 2018), as well as additional auxiliary tasks to mitigate noise. Xing et al. (2020) introduced a large-scale BioRel benchmark focusing on drug-disease and gene-cancer interactions and showed significant performance gain over a comprehensive selection of baselines. Recent works shifted to focus on using scientific language models for Bio-DSRE. We extend relation-enriched sentence-level BERT to handle bag-level MIL in §4.3 and demonstrate that preserving the direction of the KB relationships can denoise the training signal (Amin et al., 2020a). We also outline the steps to create a broad-coverage benchmark from UMLS. Following this, Hogan et al. (2021) introduced the concept of *Abstractified* MIL (AMIL) by including different argument pairs belonging to the same semantic type pair in one bag, boosting performance on rare triples.

For domain-specific Bio-DSRE, Hong et al. (2020) introduced the BERE framework for latent tree learning and self-attention to use the semantic and syntactic information in the sentence for MIL. They also introduced a Drug-Target Interactions (DTI) Bio-DSRE benchmark, suitable for drug repositioning, drawn from DrugBank (Wishart et al., 2018). Concurrent work of Marchesin and Silvello (2022) introduced a large-scale semi-automatically curated benchmark TGBA for Gene-Disease Associations (GDA). TGBA uses DisGeNET (Piñero et al., 2020), which collects data on human genotype-phenotype relationships.

In supervised RE, ChemProt (Krallinger et al., 2017) and DDI-2013 (Herrero-Zazo et al., 2013) focus on multi-class interactions between chemical-protein and drug-drug, respectively. EU-ADR (Van Mulligen et al., 2012) and GAD (Bravo et al., 2015) focus on binary relations between genes and diseases, while CDR (Li et al., 2016) focuses on binary relations between chemicals and diseases, therefore, being limited in their coverage of entity and relation types.

4.3 MULTI-INSTANCE LEARNING BASED RELATIONAL BERT

4.3.1 *Problem Definition*

We formally define the problem of bag-level MIL for RE. Let $\mathcal{E}$ and $\mathcal{R}$ represent the set of entities and relations from a knowledge base $\mathcal{KB}$, respectively. For $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$, let $(h, r, t) \in \mathcal{KB}$ be a fact triple for an ordered tuple (h, t). We denote all such (h, t) tuples by a set $\mathcal{G}^{+}$, i.e., there exists some $r \in \mathcal{R}$ for which the triple (h, r, t) belongs to the $\mathcal{KB}$, called positive groups. Similarly, we denote by $\mathcal{G}^{-}$ the set of negative groups, i.e., for all $r \in \mathcal{R}$, the triple (h, r, t) does not belong to $\mathcal{KB}$. The union of these groups is represented by $\mathcal{G} = \mathcal{G}^{+} \cup \mathcal{G}^{-}$.[2] We denote by $\mathcal{B}_g = [s_g^{(1)}, ..., s_g^{(m)}]$ an unordered sequence of sentences, called bag, for $g \in \mathcal{G}$ such that the sentences contain the group $g = (h, t)$, where the bag size m can vary. Let f be a function that maps each element of the bag to a low-dimensional relation representation $[\mathbf{r}_g^{(1)}, ..., \mathbf{r}_g^{(m)}]$. With h, we represent the bag aggregation function that maps instance-level relation representation to a final bag representation $\mathbf{b}_g = h(f(\mathcal{B}_g))$. The goal of distantly supervised bag-level MIL for corpus-level RE is to predict the missing relation r given the bag with $p(r|\mathbf{b}_g)$.

4.3.2 *Entity Markers*

Wu and He (2019) and Baldini Soares et al. (2019) showed that using special markers for entities with BERT in the order they appear in a sentence encodes the positional information that improves the performance of sentence-level RE. It allows the model to focus on target entities when other entities are also present in the sentence, implicitly doing entity disambiguation and reducing noise. In contrast, for bag-level distant supervision, the noise sources can be attributed to several factors for a given triple (h, r, t) and bag $\mathcal{B}_g$:

1. Evidence sentences may not express the relation.
2. Multiple entities appear in the sentence, requiring the model to disambiguate target entities, among others.
3. The relation prediction direction between head and tail entity.
4. Discrepancy between the order of the target entities in the sentence and knowledge base.

To address (1), common approaches are to learn a negative relation class *NA* and use better bag aggregation strategies (Alt et al., 2019; Lin et al., 2016; Luo et al., 2017a). For (2), encoding positional information is important, such as in PCNN (Zeng et al., 2014), that takes into account the relative positions of *head* and *tail* entities (Zeng et al., 2015), and in Baldini Soares et al. (2019) and Wu and He (2019) for sentence-level RE. To account for (3) and (4), multi-task learning with KGC and mutual attention has proved effective (Dai et al., 2019; Han et al., 2018a). Simply extending sentence-sensitive marking to bag-level can be adverse, as it enhances (4), and even if the composition is uniform, it distributes the evidence sentence across several bags. On the other hand, expanding relations to multiple sub-classes based on direction (Wu and He, 2019) enhances class imbalance and distributes supporting sentences. To jointly address (2), (3),

[2] The sets are disjoint, $\mathcal{G}^{+} \cap \mathcal{G}^{-} = \varnothing$

and (4), we introduce KB-sensitive encoding for scientific language models suitable for bag-level distant RE to study **RQ5**.

Formally, for a group $g = (h, t)$ and a matching sentence $s_g^{(i)}$ with tokens $(x_0, ..., x_L)$[3], we add special tokens \$ and ^ to mark the entity spans as:

Sentence ordered called *s-tag*, entities are marked in the order they appear in the sentence. Following Baldini Soares et al. (2019), let $s_1 = (i, j)$ and $s_2 = (k, l)$ be the index pairs with $0 < i < j - 1, j < k, k \leqslant l - 1$ and $l \leqslant L$ delimiting the entity mentions $e_1 = (x_i, ..., x_j)$ and $e_2 = (x_k, ..., x_l)$ respectively. We mark the boundary of s_1 with \$ and s_2 with^. Note, e_1 and e_2 can be either h or t.

KB ordered called *k-tag*, entities are marked in the order they appear in the KB. Let $s_h = (i, j)$, and $s_t = (k, l)$ be the index pairs delimiting head (h) and tail (t) entities, irrespective of the order they appear in the sentence. We mark the boundary of s_h with \$ and s_t with^.

Under this terminology, *s-tag* is used by Baldini Soares et al. (2019) and Wu and He (2019) for span identification. In Wu and He (2019), each relation type $r \in \mathcal{R}$ is further expanded to two sub-classes as $r(e_1, e_2)$ and $r(e_2, e_1)$ to capture direction while holding the *s-tag* annotation as fixed. Since the ordered tuple (h, t) is given for DSRE, the task is reduced to relation classification without direction. This side information is encoded in the data with *k-tag*, covering (2) but also (3) and (4). To account for (1), we also experiment with selective attention (Lin et al., 2016), which has been widely used in other works (Alt et al., 2019; Han et al., 2018a; Luo et al., 2017a).

4.3.3 *Model Architecture*

BERT (Devlin et al., 2019a) is used as our base sentence encoder; specifically, BioBERT (Lee et al., 2020), and we extend RBERT (Wu and He, 2019) to bag-level MIL. Figure 21 shows the model's architecture with *k-tag*. Consider a bag $\mathcal{B}_g$ of size m for a group $g \in \mathcal{G}$ representing the ordered tuple (h, t), with corresponding spans $S_g = [(s_h^{(1)}, s_t^{(1)}), ..., (s_h^{(m)}, s_t^{(m)})]$ obtained with *k-tag*, then for a pair of sentences in the bag and spans we have, $(s^{(i)}, (s_h^{(i)}, s_t^{(i)}))$. We can represent the model in three steps, such that the first two steps represent the map f and the final step h, as follows:

1. Sentence Encoding: BERT is applied to the sentence and the final hidden state $\mathbf{H}_0^{(i)} \in \mathbb{R}^d$, corresponding to the `[CLS]` token, is passed through a linear layer[4] $\mathbf{W}^{(1)} \in \mathbb{R}^{d \times d}$ with $\tanh(.)$ activation to obtain the global sentence information as $\mathbf{h}_0^{(i)}$.

2. Relation Representation: For the head entity, represented by the span $s_h^{(i)} = (j, k)$ for $k > j$, we apply average pooling $\frac{1}{k-j+1} \sum_{n=j}^{k} \mathbf{H}_n^{(i)}$, and similarly for the tail entity with span $s_t^{(i)} = (l, m)$ for $m > l$, we get $\frac{1}{m-l+1} \sum_{n=l}^{m} \mathbf{H}_n^{(i)}$. The pooled representations are then passed through a shared linear layer $\mathbf{W}^{(2)} \in \mathbb{R}^{d \times d}$ with $\tanh(.)$ activation to obtain $\mathbf{h}_h^{(i)}$ and $\mathbf{h}_t^{(i)}$. To get the final latent relation representations, we concatenate the pooled entities representation with `[CLS]` as $\mathbf{r}_g^{(i)} = [\mathbf{h}_0^{(i)}; \mathbf{h}_h^{(i)}; \mathbf{h}_t^{(i)}] \in \mathbb{R}^{3d}$.

3. Bag Aggregation: After applying the first two steps to each sentence in the bag, we obtain $[\mathbf{r}_g^{(1)}, ..., \mathbf{r}_g^{(m)}]$. With a final linear layer consisting of a relation matrix $\mathbf{M}_r \in \mathbb{R}^{|\mathcal{R}| \times 3d}$ and a bias vector $\mathbf{b}_r \in \mathbb{R}^{|\mathcal{R}|}$, we aggregate the bag information with h in two ways:

[3] $x_0 =$`[CLS]` and $x_L =$`[SEP]`

[4] Each linear layer is implicitly assumed with a bias vector

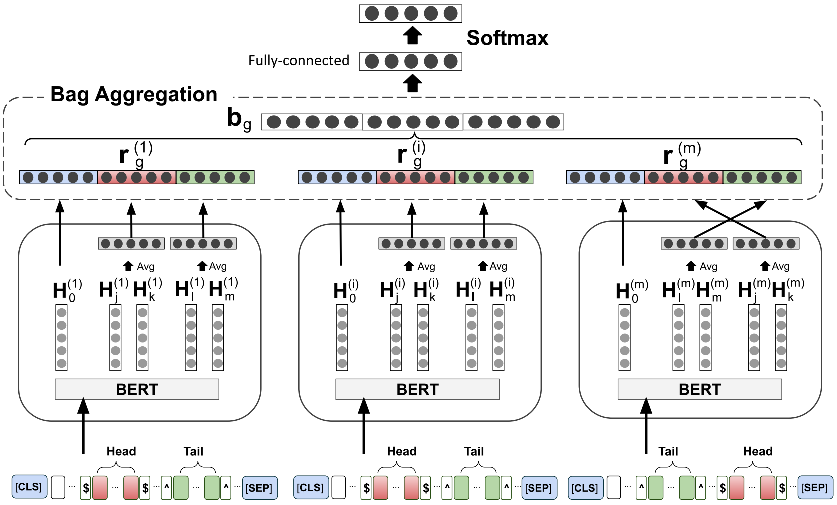

Softmax
Fully-connected
Bag Aggregation
b_g
r_g^(1)
r_g^(i)
r_g^(m)
Avg
Avg
H_0^(1) H_j^(1) H_k^(1) H_l^(1) H_m^(1)
H_0^(i) H_j^(i) H_k^(i) H_l^(i) H_m^(i)
H_0^(m) H_l^(m) H_m^(m) H_j^(m) H_k^(m)
BERT
BERT
BERT
Head
Tail
Head
Tail
Tail
Head
[CLS]
$
^
[SEP]

Figure 21: Multiple Instance Learning (MIL) based bag-level relation classification BERT with KB ordered entity marking (MIL-RBERT). Special markers \$ and ^ always delimit the span of *head* (h_s, h_e) and *tail* (t_s, t_e) entities regardless of their order in the sentence. The markers capture the position of entities and latent relation direction.

Average: The bag elements are averaged as:

$$\mathbf{b}_g = \frac{1}{m}\sum_{i=1}^{m} \mathbf{r}_g^{(i)}$$

Selective Attention: For a row $\mathbf{r} \in \mathbf{M}_r$ representing the relation $r \in \mathcal{R}$, we get the attention weights as:

$$\alpha_i = \frac{\exp(\mathbf{r}^T \mathbf{r}_g^{(i)})}{\sum_{j=1}^{m} \exp(\mathbf{r}^T \mathbf{r}_g^{(j)})}$$

$$\mathbf{b}_g = \sum_{i=1}^{m} \alpha_i \mathbf{r}_g^{(i)}$$

Following $\mathbf{b}_g$, a `softmax` classifier is applied to predict the probability $p(r|\mathbf{b}_g;\theta)$ of relation r being a true relation with θ representing the model parameters. For optimization, we minimize the cross-entropy loss during training.[5]

4.3.4 *Experiments*

For evaluating our proposed entity encoding schemes, we derive our data (UMLS.v2) similar to Dai et al. (2019) and report the data construction details in Appendix §B.2. We compare each tagging scheme, *s-tag* and *k-tag*, with average (*avg*) and selective attention (*attn*) bag aggregation functions. To test the setup of Wu and He (2019), which follows *s-tag*, we expand each relation type (*exprels*) $r \in \mathcal{R}$ to two sub-classes $r(e_1, e_2)$ and $r(e_2, e_1)$ indicating relation direction from the first entity to second and vice versa. For all experiments, we used batch size 2, bag size 16 with sampling (see §B.2.4 for details on bag composition), learning rate 2e-5 with linear decay, and 3 epochs. As

[5] `https://github.com/suamin/MIL-RBERT`

Model	Bag Agg.	AUC	F1	P@100	P@200	P@300	P@2k	P@4k	P@6k
Dai et al. (2019)	-	-	-	-	-	-	.913	.829	.753
s-tag	*avg*	.359	.468	.791	.704	.649	.504	.487	.481
	attn	.122	.225	.587	.563	.547	.476	.441	.418
s-tag+exprels	*avg*	.383	.494	.508	.519	.521	.507	.508	.511
	attn	.114	.216	.459	.476	.482	.504	.496	.484
k-tag	*avg*	**.684**	**.649**	**.974**	**.983**	**.986**	**.983**	**.977**	**.969**
	attn	.314	.376	.967	.941	.925	.857	.814	.772

Table 9: Noise reduction results for different models and data split.

the standard practice (Weston et al., 2013), evaluation is performed by constructing candidate triples by combining the entity pairs in the test set with all relations (except *NA*) and ranking the resulting triples. The extracted triples are matched against the test triples and the Precision-Recall (PR) curve, Area Under the PR Curve (AUC), F1 measure, and Precision@k, for k in {100, 200, 300, 2k, 4k, 6k} are reported.

4.3.5 *Results*

Performance metrics are shown in Table 9 and plots of the resulting PR curves in Figure 22. Since our data differs from Dai et al. (2019), the AUC cannot be directly compared. However, P@k indicates the general ranking performance of extracting the true triples and can therefore be compared. Generally, models annotated with *k-tag* perform significantly better than other models, with *k-tag+avg* achieving state-of-the-art P@{2k,4k,6k} compared to the previous state-of-the-art (Dai et al., 2019). The best model of Dai et al. (2019) uses a PCNN sentence encoder, with additional tasks of SimplE (Kazemi and Poole, 2018) based KGC and KG-attention, entity-type classification, and named entity recognition. In contrast, our data-driven method, *k-tag*, greatly simplifies this by directly encoding the KB information, i.e., order of the *head* and *tail* entities and, therefore, the latent relation direction, thus addressing **RQ5** for denoising. Consider again the example in Figure 20 where our source triple (h, r, t) is (*Iron Deficiency*, *cause_of*, *Microcytic Anaemia*), and only half of the sentences have the same order of entities as KB. This discrepancy is conveniently resolved with *k-tag* (note in Figure 21, for other sentences, the extracted entities sentence order is flipped to KG order when concatenating, unlike *s-tag*). Such knowledge can be seen as learned when jointly modeling with KGC. However, considering the task of bag-level distant RE only, the KG triples are *known* information, and we utilize this information explicitly with *k-tag* encoding.

As PCNN (Zeng et al., 2015) can account for the relative positions of head and tail entities, it also performs better than the models tagged with *s-tag* using sentence order. Similar to Alt et al. (2019)[6], we also note that the pre-trained contextualized models result in improved long-tail performance. *s-tag+exprels* reflects the direct application of Wu and He (2019) to bag-level MIL for distant RE. In this case, the relations are explicitly extended to model entity direction appearing first to second in the sentence and vice versa. This implicitly introduces independence between the two sub-classes of

[6] Their model does not use any entity marking strategy.

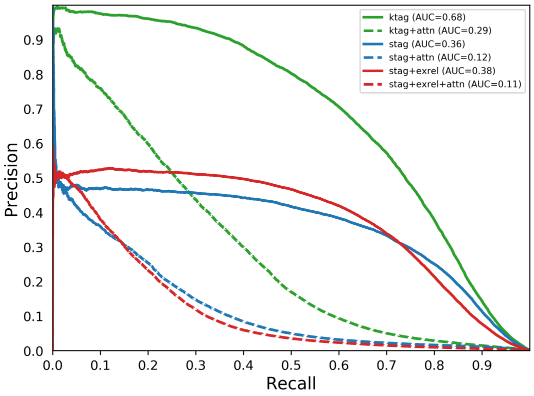


Figure 22: Precision-Recall (PR) curve for entity encoding schemes for noise reduction on UMLS.v2. We see that the models with *k-tag* perform better than the *s-tag* with average aggregation showing consistent performance for long-tail relations.

the same relation, limiting the gain from shared knowledge. Likewise, such expanded relations further enhance class imbalance in more fine-grained classes.

Although selective attention (Lin et al., 2016) has been shown to improve the performance of distant RE (Alt et al., 2019; Han et al., 2018a; Luo et al., 2017a), models in our experiments with such an attention mechanism significantly underperformed, in each case bumping the area under the PR curve and making it flatter. We note that more than 50% of bags are under-sized, in many cases, with only 1-2 sentences, requiring repeated over-sampling to match fixed bag size, making it difficult for attention to learn a distribution over the bag with repetitions and further adding noise. For such cases, the distribution should ideally be uniform, as with averaging, resulting in better performance.

Despite these results, we identify limitations in the UMLS.v2 data construction process and others (Hogan et al., 2021; Xing et al., 2020) in the next section and present a more accurate benchmark MedDistant19 to thoroughly evaluate scientific language models.

4.4 MEDDISTANT19 BENCHMARK

Noting the high accuracy of Bio-DSRE, we first investigate recent results from the broad-coverage Bio-DSRE literature by probing the respective datasets for overlaps between training and test sets. Specifically, in UMLS, each concept is mapped to a *Concept Unique Identifier (CUI)*, and a given CUI might have different surface forms (Bodenreider, 2004). We thus probe for CUI-based KG triples leakage. Our results are shown in Table 11 for UMLS.v2 (Amin et al., 2020a), UMLS.v3 (Hogan et al., 2021), and BioRel (Xing et al., 2020). For UMLS.v2 and UMLS.v3, the triples use surface forms of CUIs rather than the CUIs themselves, which results in an overlap between training and test sets. For example, consider a relationship between a pair of UMLS entities (`C0013798`, `C0429028`). These two entities can appear in different forms within a text, such as (*electrocardiography*, *Q-T interval*), (*ECG*, *Q-T interval*), and (*EKG*, *Q-T interval*); each of these distinct pairs still refers to the same original pair (`C0013798`, `C0429028`). In UMLS.v2, we made sure of eliminating all such text-based leakage, but when canonicalized to their CUIs, this results in leakage across the splits as reported in

Benchmark	Relations	No Train-Test Overlap	Broad-Coverage	Ontology
UMLS.v1 (Roller and Stevenson, 2014)	7	-	✗	UMLS
DTI (Hong et al., 2020)	6	✓	✗	DrugBank
UMLS.v2 (Amin et al., 2020a) (Ours)	355	✗	✓	UMLS
BioRel (Xing et al., 2020)	125	✗	✓	NDFRT, NCI
UMLS.v3 (Hogan et al., 2021)	275	✗	✓	UMLS
TBGA (Marchesin and Silvello, 2022)	4	✓	✗	DisGeNET
MedDistant19	22	✓	✓	SNOMED CT

Table 10: The landscape of Distantly Supervised Biomedical Relation Extraction (Bio-DSRE) benchmarks: all the existing broad-coverage datasets have corpus-level triples overlap between the train and test splits (Table 11), where the Knowledge Graph (KG) is also extracted from multiple ontologies. The DTI and TBGA benchmarks focus on harmonized ontology but are limited to drug-target interactions and gene-disease associations. In contrast, MedDistant19 has a broader coverage of entities and their semantic types and is normalized to a single ontology, SNOMED CT, which has significant clinical relevance. We named the datasets from (Amin et al., 2020a; Hogan et al., 2021; Roller and Stevenson, 2014) to UMLS.v2/3/1 since the original works were missing the names. For UMLS.v1, there is no publicly available code to reconstruct the dataset; thus, the overlap information is missing.

Triples	Train	Valid	Test
BioRel	39,969	17,815 (86.17%)	17,927 (86.37%)
UMLS.v2	211,789	41,993 (26.7%)	89,486 (26.5%)
UMLS.v3	23,163	2,643 (44.38%)	5,184 (40.12%)

Table 11: Training-test leakage we identified in the existing broad-coverage benchmarks. Numbers between parentheses show the percentage overlap of CUI triples.

Table 11. In contrast, BioRel directly splits CUI triples without accounting for inverse relations that can also result in leakage (Chang et al., 2020). Since DSRE aims at corpus-level triples extraction, train-test triples leakage is problematic (see Table 12) compared to supervised sentence-level RE, where we aim to generalize to newer contexts.

We found no such overlap for DTI and TBGA, where the datasets used in (Dai et al., 2019; Roller and Stevenson, 2014) are private. Noting these shortcomings, we introduce a new and accurate benchmark MedDistant19 for broad-coverage Bio-DSRE. Our benchmark utilizes clinically relevant SNOMED CT Knowledge Graph (Chang et al., 2020), extracted from the UMLS, that offers a careful selection of the concept types and is suitable for large-scale biomedical relation extraction. Table 10 summarizes the current landscape of Bio-DSRE benchmarks. Our inspection and its results partially address **RQ6**.

4.4.1 *Documents*

We used PubMed MEDLINE abstracts published up to 2019[7] as our text source, containing 32,151,899 abstracts. Following Hogan et al. (2021), we used ScispaCy[8] (Neumann et al., 2019) for sentence tokenization, resulting in 150,173,169 unique sentences. We further introduce the use of ScispaCy for linking entity mentions to their UMLS CUIs and filtering disabled concepts from UMLS, which resulted in entity-linked mentions at the sentence-level.

[7] https://lhncbc.nlm.nih.gov/ii/information/MBR/Baselines/2019.html
[8] https://github.com/allenai/scispacy

Model and Data	Original		Filtered	
	AUC	F1	AUC	F1
Amin et al. (2020a)	68.4	64.9	50.8	53.1
Hogan et al. (2021)	82.6	77.6	11.8	19.8

Table 12: State-of-the-art domain-specific Bio-DSRE language models evaluated on the respective datasets before (Original) and after (Filtered) removing test relationships also appearing in the training set.

Named entity recognition (NER) and normalization were two primary sources of errors in biomedical RE, as shown in Kilicoglu et al. (2020). While ScispaCy is reasonably performant among other options for biomedical entity linking, it remains quite noisy in practice; e.g., Vashishth et al. (2021) showed that ScispaCy had only about a 50% accuracy on extracting concepts in benchmark datasets. Despite this being a limitation, using ScispaCy is better than relying on string matching alone (Amin et al., 2020a; Dai et al., 2019; Hogan et al., 2021).

4.4.2 *Knowledge Base*

We use UMLS2019AB[9] as our primary knowledge source and apply a set of rules, resulting in a distilled and carefully reduced version of UMLS2019AB. The UMLS Metathesaurus (Bodenreider, 2004) covers concepts from 222 source vocabularies, thus being the most extensive ontology of biomedical concepts. However, covering all ontologies can be challenging, given the interchangeable nature of the concepts. For example, *programmed cell death 1 ligand 1* is an alias of concept `C1540292` in the HUGO Gene Nomenclature Committee ontology (Povey et al., 2001), and it is an alias of concept `C3272500` in the National Cancer Institute Thesaurus. This makes entity linking more challenging since a surface form can be linked to multiple entity identifiers, and it is easier to have overlaps between training and test sets since the same fact may appear in both with different entity identifiers.

Furthermore, benchmark corpora for biomedical NER (Doğan et al., 2014; Li et al., 2016) and RE (Herrero-Zazo et al., 2013; Krallinger et al., 2017) focus on specific entity types (e. g.diseases, chemicals, proteins), and are usually normalized to a single ontology (Kilicoglu et al., 2020). Following this trend, we also focus on a single vocabulary for Bio-DSRE using SNOMED CT, the most widely used clinical terminology worldwide for documentation and reporting in healthcare (Chang et al., 2020).

Since UMLS classifies each entity in a type taxonomy of Semantic Types (STY) and Semantic Groups (SG) (Figure 33), this allows for narrowing the concepts of interest. Following Chang et al. (2020), we first consider 8 semantic groups in SNOMED CT: Anatomy (ANAT), Chemicals & Drugs (CHEM), Concepts & Ideas (CONC), Devices (DEVI), Disorders (DISO), Phenomena (PHEN), Physiology (PHYS), and Procedures (PROC). We then remove CONC and PHEN as they are far too general to be informative for Bio-DSRE. For a complete list of semantic types covered in MedDistant19, see Table 42. Similarly, each relation is categorized into a type and has a reciprocal relation in UMLS (Table 41), which can result in train-test leakage (Dettmers et al., 2018).

The steps above follow Chang et al. (2020), with the difference that we only consider relations of type *has relationship other than synonymous, narrower, or broader* (RO); this

[9] `https://download.nlm.nih.gov/umls/kss/2019AB/umls-2019AB-full.zip`

Properties	Prior	MD19
approximate entity linking		✓
unique NA sentences		✓
inductive		✓
triples leakage	✓	
NA-type constraint		✓
NA-argument role constraint		✓

Table 13: MedDistant19 (MD19) key data construction properties in comparison with the recent broad-coverage Bio-DSRE works.

Facts	Training	Validation	Testing
Inductive (I)	261,797	48,641	97,861
Transductive (T)	318,524	28,370	56,812

Table 14: The number of raw inductive and transductive SNOMED-KG triples used for alignment with text.

is consistent with prior works in Bio-DSRE. We also exclude uninformative relations, *same_as*, *possibly_equivalent_to*, *associated_with*, *temporally_related_to*, and ignore inverse relations as generally is the case in RE.

In addition, Chang et al. (2020) ensures that the validation and test set do not contain any new entities, making it a transductive learning setting where we assume all test entities are known beforehand. However, we are expected to extract relations between unseen entities in real-world applications of biomedical RE. To support this setup, we derive MedDistant19 using an inductive KG split method proposed by Daza et al. (2021) (see Appendix A in their paper). Table 14 summarizes the statistics of the KGs used for alignment with the text. We use train, validation, and test split ratios of 70%, 10%, and 20%. Relationships are defined between CUIs and have no overlap between training, validation, and test sets.

4.4.2.1 *Knowledge-to-Text Alignment*

We now describe the procedure for searching fact triples to match relational instances in text.

Let $\mathcal{E}$ and $\mathcal{R}$ respectively denote the set of UMLS CUIs and relation types, and let $\mathcal{G} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ denote the set of relationships contained in UMLS. For producing a training-test split, we first create a set $\mathcal{G}^+ \subseteq \mathcal{E} \times \mathcal{E}$ of related entity pairs as:

$$\mathcal{G}^+ = \{(e_i, e_j) \mid \langle e_i, p, e_j \rangle \in \mathcal{G} \vee \langle e_j, p, e_i \rangle \in \mathcal{G}\}$$

Following this, we obtain a set of unrelated entity pairs by corrupting one of the entities in each pair in $\mathcal{G}^+$ and making sure it does not appear in $\mathcal{G}^+$, obtaining a new set $\mathcal{G}^- \subseteq \mathcal{E} \times \mathcal{E}$ of unrelated entities, defined as follows:

$$\begin{aligned}\mathcal{G}^- = &\{(\overline{e_i}, e_j) \mid (e_i, e_j) \in \mathcal{G}^+ \wedge (\overline{e_i}, e_j) \notin \mathcal{G}^+\}\\ &\cup \{(e_i, \overline{e_j}) \mid (e_i, e_j) \in \mathcal{G}^+ \wedge (e_i, \overline{e_j}) \notin \mathcal{G}^+\}\end{aligned}$$

Summary		**Entities**	**Relations**	**STY**	**SG**
		20,256	22	51	6
Split	**Instances**	**Facts**	**Bags**	**Inst. per Bag**	**NA (%)**
Train	450,071	5,455	88,861	5.06	90.0%
Valid	39,434	842	10,475	3.76	91.2%
Test	91,568	1,663	22,606	4.05	91.1%

Table 15: Summary statistics of the MedDistant19 dataset using Inductive SNOMED KG split (Table 14). The number of relations includes the unknown relation type (NA).

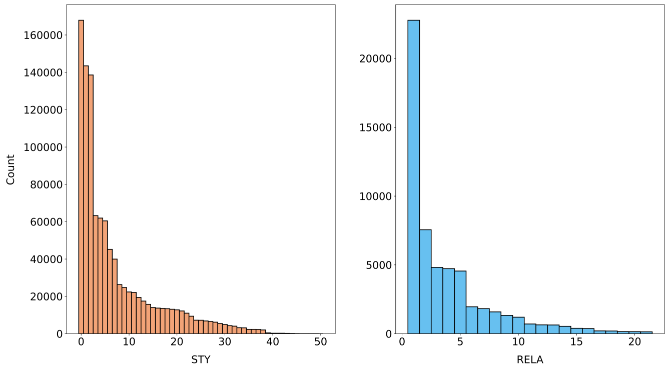


Figure 23: (*Left*) Entity distribution based on Semantic Types. (*Right*) Relations distribution.

During the corruption process, we enforce two constraints: 1) *type constraint* – the two entities appearing in each negative pair in $\mathcal{G}^-$ should belong to an entity type pair from $\mathcal{G}^+$, and 2) *role constraint* – the noisy *head* (*tail*) entity in negative pairs must have appeared in a *head* (*tail*) role from a pair in $\mathcal{G}^+$.

A naive choice for the negative group could be $\mathcal{G}^- = (\mathcal{E} \times \mathcal{E}) - \mathcal{G}^+$, for which the current approach is only a subset; however, enumerating all possible entity pairs can be infeasible if $|\mathcal{E}|$ is high. Furthermore, we do not assume the completeness of UMLS and only derive a *fixed* sub-graph from the 2019 version subject to the to the constraints stated above. This process is similar to Local-Closed World Assumption (LCWA, Dong et al., 2014; Nickel et al., 2016a), in which a KG is assumed to be only locally complete: if we observed a triple for a specific entity $e_i \in \mathcal{E}$, then we assume that any non-existing relationship (e_i, e_j) denotes a false fact and include them in $\mathcal{G}^-$. Therefore, it is likely that if a triple emerges in a new PubMed article such that it violates the negative sampling assumptions, it will be considered a false negative. However, this amount is negligible due to intractable search space that scales with the size of the KG.

For each entity-linked sentence, we only consider those sentences that have SNOMED CT entities and have pairs in $\mathcal{G}^+$ and $\mathcal{G}^-$. Selected positive and negative pairs are mutually exclusive and have no overlap across splits. Since we only consider unique sentences associated with a pair, this makes for unique negative training instances, in contrast to Amin et al. (2020a), who considered generating positive and negative pairs from the same sentence. We define negative examples as relational sentences mentioning argument pairs with *unknown relation type* (`NA`), i.e. there might be a relationship, but the considered set of relations does not cover it. Our design choices are summarized in Table 13.

Model	Bag	Strategy	AUC	F1-micro	F1-macro	P@100	P@200	P@300	P@1k	P@2k
CNN	-	AVG	27.3	33.0	16.1	50.0	46.0	44.0	41.0	33.6
	-	ONE	30.4	36.7	18.2	67.0	58.5	52.6	43.5	34.4
	✓	AVG	30.4	36.2	19.8	70.0	58.0	56.0	46.0	35.5
	✓	ONE	34.6	40.4	17.8	77.0	72.5	67.6	50.0	37.3
	✓	ATT	35.0	40.1	19.8	78.0	73.5	68.6	51.4	36.4
PCNN	-	AVG	27.2	32.4	12.9	54.0	49.5	50.3	40.7	33.2
	-	ONE	29.8	36.7	16.2	66.0	55.5	52.3	44.4	34.2
	✓	AVG	29.6	37.3	20.5	59.0	50.5	50.0	47.0	35.9
	✓	ONE	28.6	36.5	18.1	66.0	65.0	62.0	44.7	33.7
	✓	ATT	32.5	38.2	14.4	71.0	71.0	67.3	49.0	35.2
GRU	-	AVG	42.7	47.4	27.8	78.0	74.0	76.0	59.2	42.7
	-	ONE	46.4	49.3	29.2	86.0	80.5	78.3	61.2	44.9
	✓	AVG	28.6	37.2	17.9	57.0	57.0	56.0	45.3	35.4
	✓	ONE	32.6	40.8	17.7	73.0	70.5	66.3	51.2	37.0
	✓	ATT	36.6	40.9	22.2	77.0	72.0	67.6	51.3	38.7
BERT	-	AVG	**79.8**	**76.1**	**65.3**	95.0	96.0	96.0	**90.2**	67.2
	-	ONE	79.3	**76.1**	64.7	93.0	94.0	94.0	89.2	**67.4**
	✓	AVG	78.3	73.1	51.1	**99.0**	**97.5**	**96.6**	87.8	66.0
	✓	ONE	67.0	55.7	44.4	89.0	90.5	91.0	78.7	57.8
	✓	ATT	64.6	56.4	42.7	89.0	87.5	85.6	75.4	57.9

Table 16: Baseline results for MedDistant19.

We also remove mention-level overlap across the splits and apply type-based mention pruning. Specifically, we pool mentions by type and remove the sentences which have the mention appearing more than 10,000 times. We selected this threshold based on manual inspection of frequent mentions in each semantic type, so the information loss is minimal. At the same time, we still removed generalized mentions such as *disease*, *drugs*, *temperature* etc. We provide a complete list of mentions removed by this step in Table 40. Table 15 shows the final summary of MedDistant19 using an inductive split covering 20,256 entities with 51 types and 343 type pairs. Figure 23 shows entity and relation plots following a long-tail distribution.

4.4.3 *Experiments*

MedDistant19 is released in a format that is compatible with the widely adopted RE framework OpenNRE (Han et al., 2019).[10] To report our results, we use the *corpus-level* Area Under the Precision-Recall (PR) curve (AUC), Micro-F1, Macro-F1, and Precision-at-k (P@k) with $k \in \{100, 200, 300, 1k, 2k\}$, and the *sentence-level* Precision, Recall, and F1. Due to the imbalanced nature of relational instances, following Gao et al. (2021a), we report Macro-F1 values, and following Hogan et al. (2021), we report sentence-level RE results on relationships, including frequent and rare triples.

[10] `https://github.com/suamin/MedDistant19`

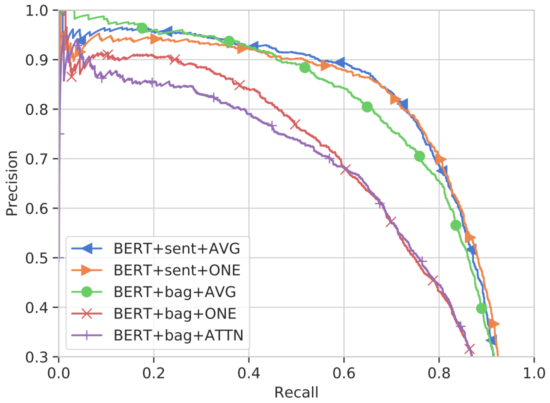


Figure 24: Precision-Recall curves for BERT baselines on MedDistant19.

4.4.4 *Baselines*

Our baseline experiments largely follow the setup of Gao et al. (2021a) with the addition of GRU models.[11] For sentence encoding, we use CNN (Liu et al., 2013), PCNN (Zeng et al., 2015), bidirectional GRU (Hong et al., 2020), and BERT (Devlin et al., 2019a). We use GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013) for CNN/PCNN/GRU models and initialize BERT with BioBERT (Lee et al., 2020).

We trained our models both at *sentence-level* and at *bag-level*. In contrast, prior works only considered bag-level training for Bio-DSRE. The sentence-level setup is similar to standard RE (Wu and He, 2019), with the difference that the evaluation is conducted at the bag-level. We also consider different pooling strategies, namely average (AVG), which averages the representations of sentences in a bag, at least one (ONE, Zeng et al., 2015), which generates relation scores for each sentence in a bag, and then selects the top-scoring sentence, and attention (ATT), which learns an attention mechanism over the sentences within a bag.

4.4.4.1 *Encoder*

Table 16 presents our main results. In all the cases, the BERT sentence encoder performed better than others since pre-trained language models are effective for entity-centric transfer learning (Amin and Neumann, 2021), domain-specific fine-tuning (Amin et al., 2019), and can implicitly store relational knowledge during pre-training (Petroni et al., 2019). This trend is similar to the general domain, and the BERT-based experiments provide consistent baselines lacking in the prior works. Similar to the general domain (Gao et al., 2021a), we find sentence-level training to perform better than the bag-level. However, BERT+bag+AVG had much better precision for the top-scoring triples at the expense of long-tail performance. At the sentence-level, those instances that have been correctly labeled by distant supervision (e. g.Figure 20) provide enough learning signal, given the generalization abilities of LMs. However, the model is supposed to jointly learn from clean and noisy samples in bag-level training, thus limiting its overall performance. But, we do not find this trend for CNN/PCNN. Instead, the bag-level models performed slightly better except for GRU. We further plot Precision-Recall (PR) curves for BERT-based baselines in Figure 24.

[11] `https://github.com/pminervini/meddistant-baselines`

Model	1-1	1-M	M-1
BERT+bag+AVG	**66.6**	**48.3**	**66.6**
BERT+bag+ONE	52.6	33.2	47.1
BERT+bag+ATT	56.4	30.7	26.4

Table 17: Averaged F1-micro score on the relation-specific category for *bag* pooling methods. The categories are defined using the *cardinality* of head and tail SGs.

Model	P	R	F1
All Triples			
BERT+sent+AVG	**0.79**	**0.65**	**0.71**
BERT+bag+AVG	0.72	0.64	0.68
Common Triples			
BERT+sent+AVG	**0.98**	**0.62**	**0.76**
BERT+bag+AVG	0.96	0.60	0.74
Rare Triples			
BERT+sent+AVG	**0.97**	0.70	0.82
BERT+bag+AVG	0.95	**0.73**	**0.83**

Table 18: Sentence-level RE metrics comparing BERT baselines trained at bag and sentence-level with AVG pooling on Rare and Common subsets of MedDistant19. The triples include NA relational instances.

4.4.4.2 *Pooling Strategy*

In all cases, AVG proved to be a better pooling strategy; this finding is consistent with prior works. Both Amin et al. (2020a) and Gao et al. (2021a) found ATT to produce less accurate results with LMs, which we also find to hold true for MedDistant19. To further study the impact of bag-level pooling strategies, we analyze the relation category-specific results. Following Chang et al. (2020), we grouped the relations based on cardinality, where the cardinality is defined as for a given relation type if the set of *head* or *tail* entities belongs to only one semantic group, then it has a cardinality one otherwise, M (many). The results are shown in Table 17 for bag-level BERT-based models with three pooling schemes. On average, models struggled the most with the 1-M category due to a need for more training signals to differentiate between heterogeneous entity types pooled over instances in a bag. While we would expect symmetric performance, to some extent, in 1-M and M-1 categories, the difference highlights that the KB-direction plays a role in Bio-DSRE, which previously has been used to de-noise the training signal (Amin et al., 2020a).

4.4.4.3 *Long-Tail Performance*

Following Hogan et al. (2021), we also perform sentence-level triples evaluation of BERT-based encoders trained at sentence-level and bag-level. The authors divided the triples (including `NA` instances) into two categories: those with 8 or more sentences are defined as *common triples* and others as *rare triples*. Table 18 shows these results. We note that both training strategies performed comparably on rare triples with BERT+sent+AVG more precise than BERT+bag+AVG at the expense of low recall. However, we find a noticeable difference in common triples where BERT+sent+AVG

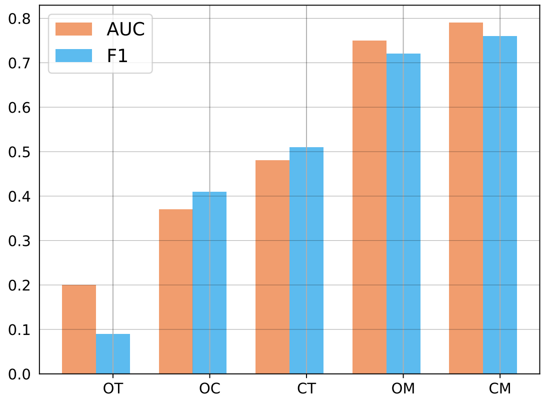


Figure 25: Ablation showing the effect of different text encoding methods, namely Only Type (OT), Only Context (OC), Context + Type (CT), Only Mention (OM), and Context + Mention (CM), with MedDistant19.

performed better. At the bag level, the model can overfit to a certain type and mention heuristics, whereas sentence-level training allows more focus on context. The current state-of-the-art model from Hogan et al. (2021) creates a bag of instances by abstracting entity pairs belonging to the same semantic type pair into a single bag, thus producing heterogeneous bags. Due to such bag creation, it is not suited for sentence-level models.

4.4.5 *Analysis*

Context, Mention, or Type? RE models are known to rely heavily on information from entity mentions, most of which is type information, and existing datasets may leak shallow heuristics via entity mentions that can inflate the prediction results (Peng et al., 2020). To study the importance of mentions, contexts, and entity types in MedDistant19, we take inspiration from (Han et al., 2020; Peng et al., 2020) and conduct an ablation of different text encoding methods. We consider entity mentions with special entity markers (Amin et al., 2020a) as the *Context + Mention* (CM) setting, which is common in RE with LMs. We then remove the context and only use mentions, the *Only Mention* (OM) setting, which reduces to KG-BERT (Yao et al., 2019) for relation prediction. We then only consider the context by replacing subject and object entities with special tokens, resulting in the *Only Context* (OC) setting. Lastly, we consider two type-based (STY) variations as *Only Type* (OT) and *Context + Type* (CT). We train the models at the sentence-level and evaluate them at the bag-level.

We observe in Figure 25 that the CM method had the highest performance, but surprisingly, OM performed quite well. This highlights the ability of LMs to memorize the facts and act as soft KBs (Petroni et al., 2019). This trend is also consistent with general-domain (Peng et al., 2020). The poor performance in the OC setting shows that the model struggles to understand the context, more pronounced in noise-prone distant RE than in supervised RE. Our CT setup can be seen as a sentence-level extrapolation of the AMIL model (Hogan et al., 2021), which struggles to perform better than the baseline (OM). However, comparing OC with CT, it is clear that the model benefits from type information as it can help constrain the space of the relations. Using only

Split	AUC	F1-micro	F1-macro
Inductive (I)	**79.9**	**76.2**	65.4
Transductive (T)	79.6	73.3	**65.9**

Table 19: BERT+sent+AVG performance on corpora created with an inductive and transductive set of triples.

the type information had the least performance as the model fails to disambiguate between different entities belonging to the same type.

Inductive or Transductive? To study the impact of *transductive* and *inductive* splits (Table 14), we created another Bio-DSRE corpus using the transductive train, validation, and test triples. The corpus generated differs from the inductive one, but it can offer insights into the model's ability to handle seen (*transductive*) and unseen (*inductive*) mentions. As shown in Table 19, inductive performance is slightly better than transductive for corpus-level extractions in terms of AUC. However, the F1-macro score is better for transductive. We conclude that the model can learn patterns that exploit mentions and type information to extrapolate to unseen mentions in the inductive setup.

Does Expert Knowledge Help? We now consider several pre-trained LMs with different knowledge capacities specific to biomedical and clinical language understanding to gain insights about the state-of-the-art encoders' performance and effectiveness on the MedDistant19 benchmark.

We use BERT (Devlin et al., 2019a) as the baseline. We next consider only those pre-trained models trained with Masked Language Modeling (MLM) objectives using domain-specific corpora. This includes ClinicalBERT (Alsentzer et al., 2019), BlueBERT (Peng et al., 2019b), BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2021). We categorize these models as non-experts since they are only trained with Masked Language Modeling (MLM) objective.

Secondly, we consider expert models that modify the MLM objective or introduce new pre-training tasks using external knowledge, such as UMLS. MedType (Vashishth et al., 2021), initialized with BioBERT, is pre-trained to predict semantic types. KeBioLM (Yuan et al., 2021), initialized with PubMedBERT, uses relational knowledge by initializing the entity embeddings with TransE (Bordes et al., 2013), improving entity-centric tasks, including RE. UmlsBERT (Michalopoulos et al., 2021), initialized with ClinicalBERT, modifies MLM to mask words belonging to the same CUI and further introduces semantic type embeddings. SapBERT (Liu et al., 2021), initialized with PubMedBERT, introduces a metric learning task for clustering synonyms together in an embedding space.

Table 20 shows the results of these sentence encoders fine-tuned on the MedDistant19 dataset at sentence-level with AVG pooling. Without domain-specific knowledge, BERT performs slightly worse than the lowest-performing biomedical model, highlighting the presence of shallow heuristics in the data common to the general and biomedical domains. While domain-specific pre-training improves the results, similar to Gu et al. (2021), we find clinical LMs underperform on the biomedical RE task. There was no performance gap between BlueBERT, SciBERT, and BioBERT. However, PubMedBERT brought improvement, consistent with Gu et al. (2021).

For expert knowledge-based models, we noted a negative impact on performance. While we would expect type-based models, MedType and UmlsBERT, to bring improvement, their effect can be attributed to overfitting certain types and patterns. KeBioLM,

Encoder	**Knowledge Type**					**AUC**
	Biomedical	*Clinical*	*Type*	*Triples*	*Synonyms*	
Non-Expert Models						
BERT						0.72
ClinicalBERT	✓	✓				0.73
BlueBERT	✓					0.78
SciBERT	✓					0.78
BioBERT	✓					0.79
PubMedBERT	✓					**0.80**
Expert Knowledge Models						
MedType	✓		✓			0.77
KeBioLM	✓			✓		**0.80**
UmlsBERT	✓	✓	✓			0.75
SapBERT	✓				✓	0.78

Table 20: Fine-tuning domain-specific LMs on MedDistant19.

initialized with PubMedBERT, has the same performance despite seeing the triples used in MedDistant19 during pre-training, highlighting the difficulty of the Bio-DSRE. SapBERT, which uses the knowledge of synonyms, also hurt PubMedBERT's performance, suggesting that while synonyms can help in entity linking, RE is a more challenging task in noisy real-world scenarios. With these analyses, we conclude the remainder of **RQ6**.

4.5 CONCLUSION

In this Chapter, we proposed relation-enriched BERT to bag-level MIL (MIL-RBERT) and introduced a simple entity encoding scheme to reduce the noise in distantly supervised biomedical RE. We noted that the position of entities in a sentence and the order in KB encodes the latent direction of relation, which plays an important role in learning under such noise. With a relatively simple data encoding scheme, we showed that it sufficiently reduced noise, alleviating the need for additional tasks, thus addressing **RQ5**.

Following this, we investigated the landscape of biomedical relation extraction benchmarks obtained with distant supervision and found either train-test leakage or coverage limitations, highlighting the need for an accurate broad-coverage benchmark for Bio-DSRE. We alleviated the limitation by utilizing a clinical sub-graph from SNOMED CT for constructing the benchmark and laying out the best practices. We thoroughly evaluated the benchmark with scientific language models, showing promising relational representation capacity, thus addressing **RQ6**.

5 KNOWLEDGE GRAPH COMPLETION IN THE GENERAL AND BIOMEDICAL DOMAIN WITH LOW-RANK BILINEAR POOLING

5.1 INTRODUCTION

A knowledge graph is a large collection of structured data, organized as entities and relations between them, in the form of fact triples <*sub*, rel, *obj*>. However, the usefulness of a KG in the general and biomedical domain is affected primarily by its incompleteness, which we also addressed through corpus-level triples extraction from text for knowledge base enrichment in the last Chapter. Link prediction or knowledge graph completion is a common task in statistical relational learning. It aims to infer missing facts from existing ones by scoring a relation and entities triple to predict its correctness and avoids the costs and time of manually extending knowledge bases. Several KGC models have been proposed, including linear and non-linear approaches, collectively recognized as Knowledge Graph Embedding (KGE) models. A KGE model outputs low-dimensional entity and relation representations that have significant utility for the biomedical domain (Chang et al., 2020). Therefore, we focus on multi-relational knowledge representation here with evaluation in the general and biomedical domains.

Bilinear models have been used in multi-modal learning due to their expressive nature. The fusion of features from different modalities plays a crucial role in the performance of a model. The underlying assumption is that the distributions of features across modalities may vary significantly, and the representation capacity of the fused features may be insufficient, limiting the final prediction performance (Yu et al., 2017). Firstly, we apply the same assumption to knowledge bases by considering that the entities and relations come from different multi-modal distributions, and good fusion between them can complete a KG. However, a significant drawback of using bilinear models is the quadratic growth of parameters, which results in high computational and memory costs and risks overfitting. In multi-modal learning, factorization techniques have been researched to address these challenges (Fukui et al., 2016; Kim et al., 2017; Yu et al., 2017), and constraint-based bilinear maps have become a common standard in KGC (Kazemi and Poole, 2018; Trouillon et al., 2016; Yang et al., 2015). Applying constraints can be seen as hard regularization since it allows for incorporating background knowledge (Kazemi and Poole, 2018) but restricts the learning potential of the model due to limited parameter sharing (Balažević et al., 2019b). We, therefore, focus on a constraint-free and efficient approach using the low-rank factorization of a bilinear model, LowFER. Our work extends the multi-modal factorized bilinear pooling (MFB) model, introduced by Yu et al. (2017), and applies it to the KGC task to address **RQ7** in §5.3 as stated:

> *RQ7: How to represent entity and relation in a parameter efficient way for knowledge graph completion in the general and biomedical domain?*

Secondly, we investigate the theoretical properties of LowFER to understand its representation capacity better. We focus on its expressiveness to correctly model every relation type and to subsume existing models to address **RQ8** in §5.4 as stated:

RQ8: What theoretical insights can be drawn about the expressivity and generalizability of the efficient parameterization?

The contents of this Chapter have appeared in the peer-reviewed article of ***Amin et al. (2020b)*** *and are included here with minor corrections where appropriate. The experimental results on biomedical KBs are only appearing in this dissertation.*

5.2 RELATED WORK

5.2.1 *Non-linear Models for KGC*

KGE models such as ConvE (Dettmers et al., 2018) and HypER (Balažević et al., 2019a) use 2D and 1D convolution on the subject entity and relation representations respectively. Both perform well in practice and are efficient, but the former lacks direct interpretation, whereas the latter is related to tensor factorization. Translational methods (Bordes et al., 2013; Feng et al., 2016; Ji et al., 2015; Lin et al., 2015; Nguyen et al., 2016; Wang et al., 2014) use additive dissimilarity scoring functions, and they differ in terms of the constraints applied to the projection matrices. While interpretable, they are theoretically limited as they have shown to be not *fully expressive* (Kazemi and Poole, 2018; Wang et al., 2018b). There are several other works (Das et al., 2018; Ebisu and Ichise, 2018; Nickel et al., 2016b; Schlichtkrull et al., 2018; Shen et al., 2018; Sun et al., 2019b; Yang et al., 2017), but we will mainly focus on linear models in this Chapter.

5.2.2 *Linear Models for KGC*

All discussed linear models could be seen as decomposing a binary KG tensor using different factorization methods. One way to decompose this tensor is to factorize its slices in the relation dimension with DEDICOMP (Harshman, 1978). RESCAL (Nickel et al., 2011), a relaxed version of DEDICOMP, decomposes using a scoring function that consists of a bilinear product between subject and object entity vectors with a relation-specific matrix. RESCAL tends to overfit due to the quadratic growth of parameters in the number of relations. Others use Canonical Polyadic decomposition (CPD or simply CP) (Harshman and Lundy, 1994; Hitchcock, 1927) to factorize the binary tensor. In CP, each value in the tensor is obtained as a sum of multiple Hadamard products of three vectors, representing the subject, object, and relation. DistMult (Yang et al., 2015), equivalent to INDSCAL (Carroll and Chang, 1970), also uses sum of Hadamard product of three vectors with a diagonal relation matrix, unlike RESCAL, to account for overfitting.

ComplEx (Trouillon and Nickel, 2017; Trouillon et al., 2016) uses complex-valued vectors for entities and relations to explicitly model asymmetric relations. SimplE (Kazemi and Poole, 2018) extends CP by introducing two vectors (*head* and *tail*) for each entity and two for relations (including the inverse). Tucker decomposition (Tucker, 1966) based TuckER (Balažević et al., 2019b) learns a 3D core tensor, which is multiplied with a matrix along each mode to approximate the binary tensor. A key difference between CP-based methods and TuckER is that TuckER learns representations via embeddings and a shared core tensor.

5.3 LOW-RANK KNOWLEDGE GRAPH COMPLETION

5.3.1 *Problem Definition*

Given a set of entities $\mathcal{E}$ and relations $\mathcal{R}$ in a knowledge graph $\mathcal{KG}$, the task of KGC is to assign a score s to a triple (e_s, r, e_o):

$$s = f(e_s, r, e_o)$$

where $e_s \in \mathcal{E}$ is the *subject* entity, $e_o \in \mathcal{E}$ is the *object* entity and $r \in \mathcal{R}$ is the *relation* between them. The scoring function f estimates the general binary tensor $\mathbf{T} \in |\mathcal{E}| \times |\mathcal{R}| \times |\mathcal{E}|$, by assigning a score of 1 to $\mathbf{T}_{ijk}$ if relation r_j exists between entities e_i and e_k, 0 otherwise. The scoring function can be a linear or non-linear model trained to predict the correctness of a triple to belong to the $\mathcal{KG}$.

5.3.2 *Multi-modal Factorized Bilinear Pooling*

Downstream performance for tasks such as visual question answering strongly depends on the multi-modal fusion of features to leverage the heterogeneous data (Liu et al., 2018). Bilinear models are expressive as they allow for pairwise interactions between the feature dimensions but also introduce a considerable number of parameters that lead to high computational and memory costs and the risk of overfitting (Fukui et al., 2016). Substantial research has therefore focused on efficiently computing the bilinear product.

In Multi-modal Compact Bilinear (MCB) pooling (Fukui et al., 2016; Gao et al., 2016), authors employ a sampling-based approximation that uses the property that the tensor sketch projection (Charikar et al., 2004; Pham and Pagh, 2013) of the outer product of two vectors can be represented as their sketches' convolution. Multi-modal Low-rank Bilinear (MLB) pooling (Kim et al., 2017) uses two low-rank projection matrices to transform the features from the original space to a shared space. It is followed by the Hadamard product, which was later generalized by the Multi-modal Factorized Bilinear (MFB) pooling (Yu et al., 2017). In contrast to KGC bilinear models, these models allow for parameter sharing and, generally, are constraint-free. Our work is based on the MFB model but can also be seen as related to Liu et al. (2018), therefore we present it formally next.

Given two feature vectors $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$ and a bilinear map $\mathbf{W} \in \mathbb{R}^{m \times n}$, the bilinear transformation is defined as $z = \mathbf{x}^T\mathbf{W}\mathbf{y} \in \mathbb{R}$. To obtain a vector in $\mathbb{R}^o$, o such maps have to be learned (e.g., in RESCAL, these would be relation-specific matrices), resulting in a large number of parameters. However, $\mathbf{W}$ can be factorized into two low-rank matrices:

$$z = \mathbf{x}^T\mathbf{U}\mathbf{V}^T\mathbf{y} = \mathbf{1}^T(\mathbf{U}^T\mathbf{x} \circ \mathbf{V}^T\mathbf{y})$$

where $\mathbf{U} \in \mathbb{R}^{m \times k}$, $\mathbf{V} \in \mathbb{R}^{n \times k}$, k is the factorization rank, $\circ$ is the Hadamard product of two vectors and $\mathbf{1} \in \mathbb{R}^k$ is a vector of all ones. Therefore, to obtain an output feature vector $\mathbf{z} \in \mathbb{R}^o$, two 3D tensors are required, $\mathbf{W}_x = [\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_o] \xrightarrow{reshape} \mathbf{W}_x'$ and $\mathbf{W}_y = [\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_o] \xrightarrow{reshape} \mathbf{W}_y'$, where $\mathbf{W}_x \in \mathbb{R}^{m \times k \times o}$, $\mathbf{W}_y \in \mathbb{R}^{n \times k \times o}$ are 3D tensors and $\mathbf{W}_x' \in \mathbb{R}^{m \times ko}$, $\mathbf{W}_y' \in \mathbb{R}^{n \times ko}$ are their reshaped 2D matrices respectively. The final (fused) vector $\mathbf{z}$ is then obtained by summing non-overlapping windows of size k over the Hadamard product of projected vectors using $\mathbf{W}_x^{'}$ and $\mathbf{W}_y^{'}$:

$$\mathbf{z} = \text{SumPool}(\mathbf{W}_x^{'T}\mathbf{x} \circ \mathbf{W}_y^{'T}\mathbf{y}, k) \tag{2}$$

At $k = 1$, MFB reduces to MLB, which converges slowly, and MCB requires very high-dimensional vectors to perform well (Yu et al., 2017). Further, MFB significantly lowers the number of parameters with low-rank factorized matrices and leads to better performance.

5.3.3 *LowFER*

Consider that *entities* and *relations* are not intrinsically bound and come from two different modalities, such that *good* fusion between them can potentially result in a knowledge base of fact triples. Entities and relations can be shown to possess certain properties that allow them to function similarly to others within the same modality.

For example, the relation *cause-of* shares inherent properties with the relation *diagnoses*. As such, similar entity pairs can yield similar relations, given appropriate shared properties. Like in multi-modal auditory-visual fusion, where the sound of a roar may better predict a resulting image within the distribution of animals that roar, a relation such as *diagnoses*, can better predict an entity pair within a distribution of (*procedure*, *disease*) entity pairs. In KGC, we assume that the latent decomposition with MFB can help the model capture different aspects of interactions between an entity and a relation, leading to better scoring with the missing entity. We, therefore, apply the **Low**-rank **F**actorization trick of bilinear maps with k-sized non-overlapping summation pooling (cf. §5.3.2) to **E**ntities and **R**elations (LowFER) to study **RQ7**.

More formally, for an entity $e \in \mathcal{E}$, we represent its embedding vector $\mathbf{e}$ of d_e dimension as a look-up from entity embedding matrix $\mathbf{E} \in \mathbb{R}^{n_e \times d_e}$, where $n_e = |\mathcal{E}|$. Similarly, for a relation $r \in \mathcal{R}$, we represent its embedding vector $\mathbf{r}$ of d_r dimension as a look-up from relation embedding $\mathbf{R} \in \mathbb{R}^{n_r \times d_r}$, where $n_r = |\mathcal{R}|$. Then, for a given triple (e_s, r, e_o), we define our scoring function as:

$$f(e_s, r, e_o) := \mathbf{g}(e_s, r) \cdot \mathbf{e}_o = \mathbf{g}(e_s, r)^T \mathbf{e}_o \tag{3}$$

where $\mathbf{g}(.,.) \in \mathbb{R}^{d_e}$ is a vector valued function of the subject entity vector $\mathbf{e}_s$ and the relation vector $\mathbf{r}$, defined from Eq. 2 as:

$$\mathbf{g}(e_s, r) := \text{SumPool}(\mathbf{U}^T\mathbf{e}_s \circ \mathbf{V}^T\mathbf{r}, k) \tag{4}$$

where matrices $\mathbf{U} \in \mathbb{R}^{d_e \times kd_e}$ and $\mathbf{V} \in \mathbb{R}^{d_r \times kd_e}$ represent our model parameters. We can re-write the Eq. 4 more compactly as:

$$\mathbf{g}(e_s, r) = \mathbf{S}^k \text{diag}(\mathbf{U}^T\mathbf{e}_s)\mathbf{V}^T\mathbf{r} \tag{5}$$

where $\text{diag}(\mathbf{U}^T\mathbf{e}_s) \in \mathbb{R}^{kd_e \times kd_e}$ and $\mathbf{S}^k \in \mathbb{R}^{d_e \times kde}$ is a constant matrix[1] such that:

$$\mathbf{S}^k_{i,j} = \begin{cases} 1, & \forall j \in [(i-1)k+1, ik] \\ 0, & \text{otherwise} \end{cases}$$

[1] Note that at $k = 1$, $\mathbf{S}^1 = \mathbf{I}_{d_e}$

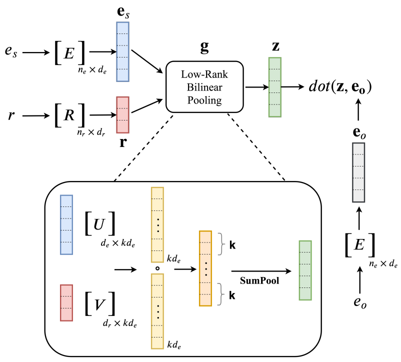


Figure 26: Overview of the LowFER model. For an input tuple (e_s, r) and target entity e_o, we first get entity vectors $\mathbf{e}_s, \mathbf{e}_o \in \mathbb{R}^{d_e}$ from entity embedding matrix $\mathbf{E} \in \mathbb{R}^{n_e \times d_e}$ and relation vector $\mathbf{r} \in \mathbb{R}^{d_r}$ from relation embedding matrix $\mathbf{R} \in \mathbb{R}^{n_r \times d_r}$, where n_e and n_r are number of entities and relations in $\mathcal{KG}$. LowFER projects $\mathbf{e}_s$ and $\mathbf{r}$ into a common space $\mathbb{R}^{kd_e}$ followed by Hadamard product and k-summation pooling, where k is the factorization rank. The output vector $\mathbf{z}$ is then matched against the target entity $\mathbf{e}_o$ to give a final score.

Using this compact notation in Eq. 3, the final scoring function of LowFER is obtained as:

$$f(e_s, r, e_o) = (\mathbf{S}^k \text{diag}(\mathbf{U}^T \mathbf{e}_s) \mathbf{V}^T \mathbf{r})^T \mathbf{e}_o \tag{6}$$

5.3.4 *Training*

To train the LowFER model, we follow the setup of Balažević et al. (2019b). First, we apply a sigmoid non-linearity after Eq. 6 to get the probability $p(y_{(e_s,r,e_o)}) = \sigma(f(e_s, r, e_o))$ of a triple belonging to a $\mathcal{KG}$. Then, for every triple (e_s, r, e_o) in the dataset, a reciprocal relation is added by generating a synthetic example (e_o, r^{-1}, e_s) (Dettmers et al., 2018; Lacroix et al., 2018) to create the training set $\mathcal{D}$. For faster training, Dettmers et al. (2018) introduced 1-N scoring, where each tuple (e_s, r) and (e_o, r^{-1}) is simultaneously scored against all entities $e \in \mathcal{E}$ to predict 1 if $e = e_o$ or e_s respectively and 0 elsewhere (see Trouillon et al. (2017) and Sun et al. (2019b) for other methods to collect negative samples).

The model is trained with binary cross-entropy instead of margin-based ranking loss (Bordes et al., 2013), which is prone to overfitting for knowledge completion (Kazemi and Poole, 2018; Trouillon and Nickel, 2017). For a mini-batch $\mathcal{B}$ of size m drawn from $\mathcal{D}$, we minimize:

$$\min_{\Theta} \frac{1}{m} \sum_{(e,r) \in \mathcal{B}} -\frac{1}{n_e} \sum_{i=1}^{n_e} y_i \log(p(y_{(e,r,e_i)})) + (1 - y_i) \log(1 - p(y_{(e,r,e_i)}))$$

where y_i is a target label for a given entity-relation pair (e, r) for entity e_i, $p(y_{(e,r,e_i)})$ is the model prediction and Θ represents model parameters.

Following Yu et al. (2017), we also apply power normalization $\mathbf{x} \leftarrow \text{sign}(\mathbf{x})|\mathbf{x}|^{0.5}$ and l_2-normalization $\mathbf{x} \leftarrow \mathbf{x}/||\mathbf{x}||_2$ before summation pooling to stabilize the training from large output values as a result of Hadamard product in Eq. 4.

Model	Full Expressibility Bounds
RESCAL (Nickel et al., 2011)	$(d_e, d_r) = (n_e, n_e^2)$
HolE (Nickel et al., 2016b)	$d_e = d_r = 2n_e n_r + 1$
ComplEx (Trouillon et al., 2016)	$d_e = d_r = n_e n_r$
SimplE (Kazemi and Poole, 2018)	$d_e = d_r = \min(n_e n_r, n + 1)$
TuckER (Balažević et al., 2019b)	$(d_e, d_r) = (n_e, n_r)$
LowFER	$(d_e, d_r) = (n_e, n_r)$ for $k = \min(n_e, n_r)$

Table 21: Bounds for *fully expressive* linear models, where n is the number of true facts, and k is the factorization rank. The trivial bound is given by $n_e^2 n_r$.

5.4 THEORETICAL ANALYSIS

5.4.1 *Full Expressibility*

A key theoretical property of knowledge representation models is their ability to be fully expressive, which we define formally as:

Definition 3. *Given a set of entities $\mathcal{E}$, relations $\mathcal{R}$, correct triples $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ and incorrect triples $\mathcal{T}' = \mathcal{E} \times \mathcal{R} \times \mathcal{E} \setminus \mathcal{T}$, then a model $\mathcal{M}$ with scoring function $f(e_s, r, e_o)$ is said to be fully expressive iff it can accurately separate $\mathcal{T}$ from $\mathcal{T}'$ for all $e_s, e_o \in \mathcal{E}$ and $r \in \mathcal{R}$.*

A *fully expressive* model can represent relations of any type, including *symmetric*, *asymmetric*, *reflexive*, and *transitive* among others. Models such as RESCAL, HolE, ComplEx, SimplE, and TuckER have been shown to be *fully expressive* (Balažević et al., 2019b; Kazemi and Poole, 2018; Trouillon and Nickel, 2017; Wang et al., 2018b). On the other hand, DistMult is not *fully expressive* as it enforces symmetric relations only. Further, Wang et al. (2018b) showed that TransE is not *fully expressive*, which was later expanded by Kazemi and Poole (2018), showing that other translational variants, including, FTransE, STransE, FSTransE, TransR, and TransH are likewise not *fully expressive*. By virtue of the universal approximation theorem (Cybenko, 1989; Hornik, 1991), neural networks can be considered *fully expressive* (Kazemi and Poole, 2018). Table 21 summarizes the bounds of linear models that are *fully expressive*. With Proposition 1, we establish that LowFER is *fully expressive* and provide bounds on the entity and relation embedding dimensions and the factorization rank k to partially address **RQ8**.

Proposition 1. *For a set of entities $\mathcal{E}$ and a set of relations $\mathcal{R}$, given any ground truth $\mathcal{T}$, there exists an assignment of values in the LowFER model with entity embeddings of dimension $d_e = |\mathcal{E}|$, relation embeddings of dimension $d_r = |\mathcal{R}|$ and the factorization rank $k = \min(d_e, d_r)$ that makes it fully expressive.*

As a given example, consider a set of entities $\mathcal{E} = \{e_1, e_2, e_3, e_4\}$ and relations $\mathcal{R} = \{r_1, r_2, r_3, r_4\}$ such that r_1 is reflexive, r_2 is symmetric, r_3 is asymmetric, and r_4 is transitive, then for ground truth $\mathcal{T} = \{(e_1, r_1, e_1), (e_1, r_2, e_2)\ , (e_2, r_2, e_1), (e_3, r_3, e_2)\ , (e_4, r_4, e_3), (e_3, r_4, e_1), (e_4, r_4, e_1)\}$ and following the settings in Proposition 1, Figure 27 shows the model parameters **U** and **V** for this toy example. Now, consider the case $k = d_e = n_e$, then **U** copies each entity vector in k-sized slices and **V** buckets target entities per relation such that each source entity is distributed into disjoint sets. Note that reshaping **V** as 3D tensor of size $n_r \times n_e \times n_e$ and transposing the first two dimensions results in binary tensor **T**.

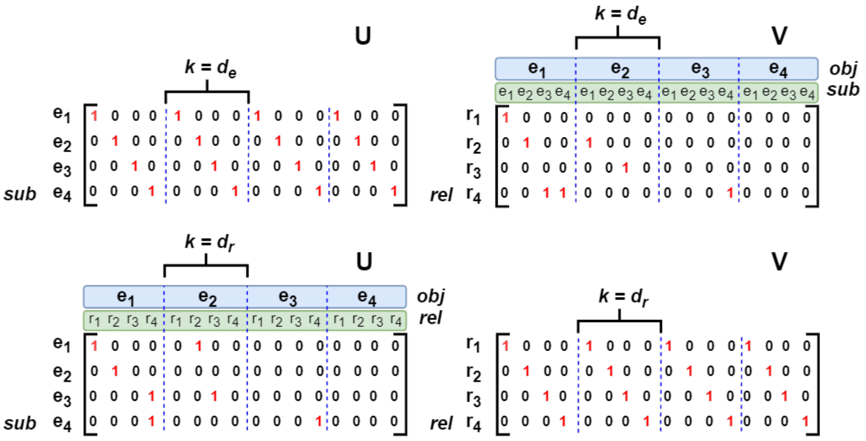

U
k = d_e
e1 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0
e2 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0
e3 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0
sub e4 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1
V
k = d_e
e1 e2 e3 e4 obj
e1 e2 e3 e4 e1 e2 e3 e4 e1 e2 e3 e4 e1 e2 e3 e4 sub
r1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
r2 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0
r3 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
rel r4 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0
U
k = d_r
e1 e2 e3 e4 obj
r1 r2 r3 r4 r1 r2 r3 r4 r1 r2 r3 r4 r1 r2 r3 r4 rel
e1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
e2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
e3 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0
sub e4 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
V
k = d_r
r1 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0
r2 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0
r3 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0
rel r4 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1

Figure 27: LowFER model parameters for a toy dataset under the settings used in Proposition 1. *Top*: For the case when $k = d_e = n_e$. *Bottom*: For the case when $k = d_r = n_r$.

The bounds presented in Table 21 are weak and, in practice, not very useful. They are derived for checking the full expressibility of a model only, which is also referred to as model being *universal* in Wang et al. (2018b), to handle *all-types* of relations with zero error, i.e., perfect reconstruction of the binary tensor $\mathbf{T}$ for a given $\mathcal{KG}$. Since factorization-based methods can be seen as approximating the true binary tensor, more useful bounds can be derived by studying the quality of the approximations for a given accuracy level. The bounds for RESCAL, ComplEx, and HolE are reported from Wang et al. (2018b) while for SimplE (Kazemi and Poole, 2018) and TuckER (Balažević et al., 2019b), from their respective papers.

5.4.2 *Relation with TuckER*

Initially, it was shown by Kazemi and Poole (2018) that RESCAL, DistMult, ComplEx, and SimplE belong to a *family of bilinear models* with a different set of constraints. Later, Balažević et al. (2019b) established that TuckER generalizes all of these models as special cases. In this section, we will formulate a relationship between our model and TuckER (Balažević et al., 2019b), followed by relations with the *family of bilinear models* in the next section. This provides a unifying view and shows LowFER's ability to generalize while addressing the remainder of **RQ8**.

TuckER's scoring function is defined as follows (Balažević et al., 2019b):

$$\phi_t(e_s, r, e_o) = \mathcal{W} \times_1 \mathbf{e}_s \times_2 \mathbf{r} \times_3 \mathbf{e}_o \tag{7}$$

where $\mathcal{W} \in \mathbb{R}^{d_e \times d_r \times d_e}$ is the core tensor, $\mathbf{e}_s, \mathbf{e}_o \in \mathbb{R}^{d_e}$ and $\mathbf{r} \in \mathbb{R}^{d_r}$ are subject entity, object entity and the relation vectors respectively. $\times_n$ denotes the tensor product along the n-th mode. First, note that Eq. 5 can be expanded as:

$$\mathbf{S}^k(\mathbf{U}^T\mathbf{e}_s \circ \mathbf{V}^T\mathbf{r}) = \begin{bmatrix} \mathbf{e}_s^T(\sum_{i=1}^{k} \mathbf{u}_i \otimes \mathbf{v}_i)\mathbf{r} \\ \vdots \\ \mathbf{e}_s^T(\sum_{i=(j-1)k+1}^{jk} \mathbf{u}_i \otimes \mathbf{v}_i)\mathbf{r} \\ \vdots \\ \mathbf{e}_s^T(\sum_{i=k(d_e-1)}^{kd_e} \mathbf{u}_i \otimes \mathbf{v}_i)\mathbf{r} \end{bmatrix}$$

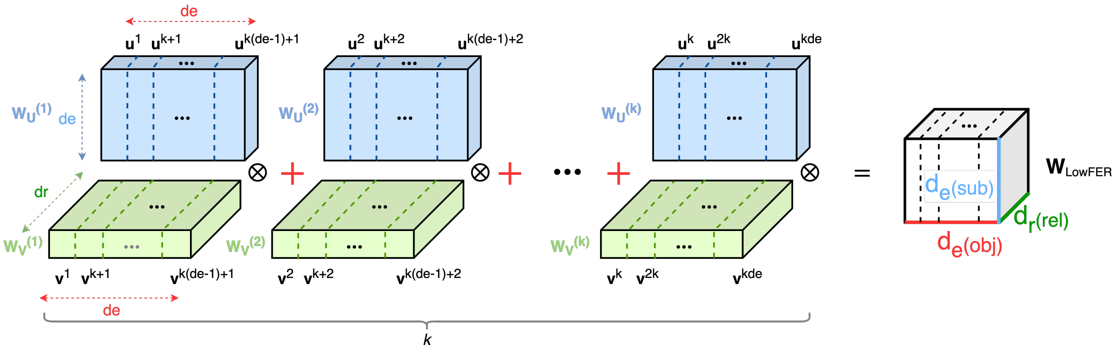


Figure 28: Low-rank approximation of the core tensor $\mathcal{W}$ of TuckER (Balažević et al., 2019b) with LowFER by summing k low-rank 3D tensors, where each tensor is obtained by stacking d_e rank-1 matrices obtained by the outer product of k-apart columns of **U** and **V**.

where $\mathbf{u}_i \in \mathbb{R}^{d_e}$ and $\mathbf{v}_i \in \mathbb{R}^{d_r}$ are column vectors of **U** and **V** respectively and $\otimes$ represents the outer product of two vectors. To take the vectors $\mathbf{e}_s$ and $\mathbf{r}$ out, we realize the above matrix operations differently. We first create k matrices sliced from **U** and **V** each, such that each matrix is formed by choosing all adjacent column vectors that are k distance apart in **U** (and **V**), i.e., for the l-th slice, we have $\mathbf{W}_U^{(l)} = [\mathbf{u}_l, \mathbf{u}_{k+l}, ..., \mathbf{u}_{k(d_e-1)+l}] \in \mathbb{R}^{d_e \times d_e}$ and $\mathbf{W}_V^{(l)} = [\mathbf{v}_l, \mathbf{v}_{k+l}, ..., \mathbf{v}_{k(d_e-1)+l}] \in \mathbb{R}^{d_r \times d_e}$. Taking the column-wise outer product of these sliced matrices forms a 3D tensor in $\mathbb{R}^{d_e \times d_r \times d_e}$. With slight abuse of notation, we also use $\otimes$ to represent this tensor operation. It can be viewed as transforming the matrix obtained by mode-2 Khatri-Rao product into a 3D tensor (Cichocki et al., 2016). Now consider a 3D tensor $\mathbf{W} \in \mathbb{R}^{d_e \times d_r \times d_e}$ as the sum of these k products:

$$\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_U^{(i)} \otimes \mathbf{W}_V^{(i)} \tag{8}$$

Figure 28 shows these operations. With this tensor, the scoring function f in Eq. 6 can be re-written as TuckER's scoring function as follows:

$$\hat{\phi}_t(e_s, r, e_o) = \mathbf{W} \times_1 \mathbf{e}_s \times_2 \mathbf{r} \times_3 \mathbf{e}_o \tag{9}$$

It should be noted that **W** in Eq. 9 is obtained as a summation of k low-rank 3D tensors, each of which is obtained by stacking rank-1 matrices in contrast to TuckER's core tensor $\mathcal{W}$ in Eq. 7, which can be a full rank 3D tensor. Our model can therefore approximate TuckER and can be viewed as a generalization of TuckER (Balažević et al., 2019b). We further show that we can accurately obtain $\mathcal{W}$ with appropriate $\mathbf{W}_U^{(i)}$'s and $\mathbf{W}_V^{(i)}$'s in Eq. 8 through Proposition 2.

Proposition 2. *Given a TuckER model with entity embedding dimension d_e, relation embedding dimension d_r and core tensor $\mathcal{W}$, there exists a LowFER model with $k <= \min(d_e, d_r)$, entity embedding dimension d_e and relation embedding dimension d_r that accurately represents the former.*

LowFER and TuckER parameters grow linearly in the number of entities and relations as $\mathcal{O}(n_e d_e + n_r d_r)$. However, LowFER's shared parameters through decoupled low-rank matrices can control complexity through the factorization rank, making it more flexible, e.g., consider $d = d_e = d_r$, the core tensor $\mathcal{W}$ of TuckER grows as $\mathcal{O}(d^3)$. In contrast, LowFER grows only as $\mathcal{O}(kd^2)$. For example, in Lacroix et al. (2018), authors used $d_e = d_r = 2000$, which would require more than 8 billion parameters to model with TuckER compared to only 4k million for LowFER, with k controlling the growth. More generally, at $k = d_e/2$, LowFER has an equal number of parameters as TuckER; therefore, we expect similar performance at such rank values. In practice, $k = \{1, 10, 30\}$ performs well, thus partially addressing **RQ7**.

5.4.3 *Relation with Family of Bilinear Models*

This section will establish relations between LowFER and other bilinear models. For simplicity, we consider the relation embedding to be a constant matrix $\mathbf{R} = \mathbf{I}_{n_r}$ in all the cases and use $\mathbf{V}$ to model relation parameters. However, the conditions presented here can be extended otherwise with a remark that they are not unique.

5.4.3.1 *RESCAL*

Nickel et al. (2011) define the scoring function as:

$$\phi_r(e_s, r_l, e_o) = \mathbf{e}_s^T \mathbf{W}_l \mathbf{e}_o$$

where $\mathbf{W}_l \in \mathbb{R}^{d_e \times d_e}$ is l-th relation matrix. For LowFER to encode RESCAL with Eq. 6, we set $k = d_e$, $d_r = n_r$ and $\mathbf{U} = [\ \mathbf{I}_{d_e} \mid \mathbf{I}_{d_e} \mid ... \mid \mathbf{I}_{d_e}\ ] \in \mathbb{R}^{d_e \times d_e^2}$ (block matrix partitioned as d_e identity matrices of size $d_e \times d_e$). This is effectively taking a row l from $\mathbf{V} \in \mathbb{R}^{n_r \times {d_e}^2}$, reshaping it to $d_e \times d_e$ matrix and then taking the transpose to get the equivalent $\mathbf{W}_l$ in RESCAL's scoring function.

5.4.3.2 *DistMult*

Yang et al. (2015) define the scoring function as:

$$\phi_d(e_s, r_l, e_o) = \mathbf{e}_s^T \text{diag}(\mathbf{w}_l) \mathbf{e}_o$$

where $\mathbf{w}_l \in \mathbb{R}^{d_e}$ is the vector for l-th relation. For LowFER to encode DistMult with Eq. 6, we set $k = 1$, $d_r = n_r$ and $\mathbf{U} = \mathbf{I}_{d_e}$. This is effectively taking a row l from $\mathbf{V} \in \mathbb{R}^{n_r \times d_e}$ and creating a diagonal matrix of it to get the equivalent $\text{diag}(\mathbf{w}_l)$ in DistMult's scoring function.

5.4.3.3 *SimplE*

Kazemi and Poole (2018) define the scoring function as:

$$\phi_s(e_s, r_l, e_o) = \frac{1}{2}(\mathbf{h}_{e_s}^T \text{diag}(\mathbf{r}_l) \mathbf{t}_{e_o} + \mathbf{h}_{e_o}^T \text{diag}(\mathbf{r}_l^{-1}) \mathbf{t}_{e_s})$$

where $\mathbf{h}_{e_s}, \mathbf{h}_{e_o} \in \mathbb{R}^d$ are subject, object entities head vectors, $\mathbf{t}_{e_s}, \mathbf{t}_{e_o} \in \mathbb{R}^d$ are subject, object entities tail vectors and $\mathbf{r}_l, \mathbf{r}_l^{-1} \in \mathbb{R}^d$ are relation and inverse relation vectors. Let $\hat{\mathbf{e}}_s = [\mathbf{t}_{e_s}; \mathbf{h}_{e_s}] \in \mathbb{R}^{2d}$, $\mathbf{e}_o = [\mathbf{h}_{e_o}; \mathbf{t}_{e_o}] \in \mathbb{R}^{2d}$ and $\hat{\mathbf{r}}_l = [\mathbf{r}_l^{-1}; \mathbf{r}_l] \in \mathbb{R}^{2d}$ then SimplE scoring is equivalent to $\frac{1}{2}\hat{\mathbf{e}}_s^T \text{diag}(\hat{\mathbf{r}}_l) \mathbf{e}_o$, where $\hat{\mathbf{e}}_s$ and $\hat{\mathbf{r}}_l$ are obtained by swapping the head, tail vectors in $\mathbf{e}_s = [\mathbf{h}_{e_s}; \mathbf{t}_{e_s}]$ and relation, inverse relation vectors in $\mathbf{r}_l = [\mathbf{r}_l; \mathbf{r}_l^{-1}]$ respectively. For LowFER to encode SimplE, $\mathbf{U}$ becomes a permutation matrix (ignoring

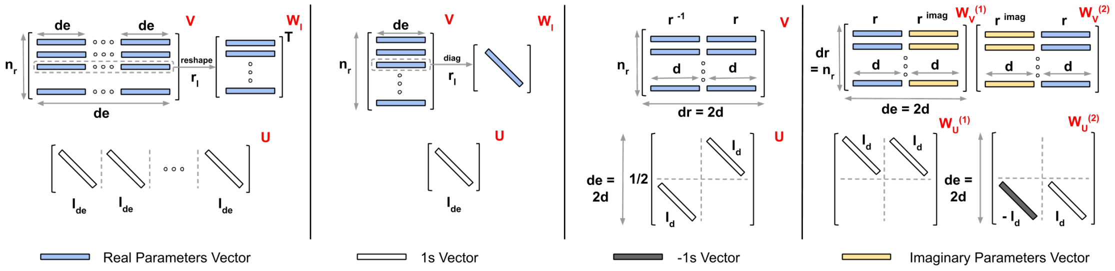


Figure 29: Modeling the family of bilinear models with LowFER, (*from left-to-right*): RESCAL (Nickel et al., 2011), DistMult (Yang et al., 2015), SimplE (Kazemi and Poole, 2018) and ComplEx (Trouillon et al., 2016) (see §5.4.3 for details).

the $\frac{1}{2}$ scaling factor), swapping the first d-half with the second d-half of a given vector in $\mathbb{R}^{2d}$ and l-th row in $\mathbf{V}$ is $\hat{\mathbf{r}}_l$, more specifically, with Eq. 6, we set $k = 1$, $d_e = 2d$, $d_r = n_r$ and $\mathbf{U} \in \mathbb{R}^{2d \times 2d}$ is a block matrix with four partitions such that, $\mathbf{U}_{12} = \mathbf{U}_{21} = \frac{1}{2}\mathbf{I}_d$ and 0s elsewhere.

5.4.3.4 *ComplEx*

Trouillon et al. (2016) define the scoring function as:

$$\begin{aligned}\phi_c(e_s, r_l, e_o) = &\operatorname{Re}(\mathbf{e}_s)^T \operatorname{diag}(\operatorname{Re}(\mathbf{r}_l))\operatorname{Re}(\mathbf{e}_o) + \operatorname{Im}(\mathbf{e}_s)^T \operatorname{diag}(\operatorname{Re}(\mathbf{r}_l))\operatorname{Im}(\mathbf{e}_o)\\ &+ \operatorname{Re}(\mathbf{e}_s)^T \operatorname{diag}(\operatorname{Im}(\mathbf{r}_l))\operatorname{Im}(\mathbf{e}_o) - \operatorname{Im}(\mathbf{e}_s)^T \operatorname{diag}(\operatorname{Im}(\mathbf{r}_l))\operatorname{Re}(\mathbf{e}_o)\end{aligned}$$

where Re(.) and Im(.) represent the real and imaginary parts of a complex vector. Consider $\hat{\mathbf{e}}_s = [\operatorname{Re}(\mathbf{e}_s); \operatorname{Im}(\mathbf{e}_s)] \in \mathbb{R}^{2d}$ and $\hat{\mathbf{e}}_o = [\operatorname{Re}(\mathbf{e}_o); \operatorname{Im}(\mathbf{e}_o)] \in \mathbb{R}^{2d}$ then the ComplEx scoring function can be obtained as $\hat{\mathbf{e}}_s^T \mathbf{W}_l \hat{\mathbf{e}}_o$, where $\mathbf{W}_l \in \mathbb{R}^{2d \times 2d}$ represents the l-th relation matrix such that its diagonal is $[\operatorname{Re}(\mathbf{r}_l); \operatorname{Re}(\mathbf{r}_l)]$, the d offset diagonal is $\operatorname{Im}(\mathbf{r}_l)$ and $-d$ offset diagonal is $-\operatorname{Im}(\mathbf{r}_l)$. For LowFER to encode ComplEx, similar to SimplE, we will use two permutation matrices to obtain the above four terms. That is, in Eq. 9, we have $k = 2$, $d_e = 2d$, $d_r = n_r$, $\mathbf{U} \in \mathbb{R}^{2d \times 4d}$ is such that $\mathbf{W}_U^{(1)}$ is a block matrix with $\mathbf{W}_{U_{11}}^{(1)} = \mathbf{W}_{U_{12}}^{(1)} = \mathbf{I}_d$ and 0 elsewhere. Further, $\mathbf{W}_U^{(2)}$ is also a block matrix with $\mathbf{W}_{U_{21}}^{(2)} = -\mathbf{I}_d$, $\mathbf{W}_{U_{22}}^{(2)} = \mathbf{I}_d$ and 0 elsewhere. Lastly, $\mathbf{V} \in \mathbb{R}^{n_r \times 4d}$ is such that $\mathbf{W}_V^{(1)}$ row l has $[\operatorname{Re}(\mathbf{r}_l); \operatorname{Im}(\mathbf{r}_l)]$ and $\mathbf{W}_V^{(2)}$ row l has $[\operatorname{Im}(\mathbf{r}_l); \operatorname{Re}(\mathbf{r}_l)]$, i.e., $\mathbf{W}_V^{(2)} = \mathbf{W}_V^{(1)}\mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{2d \times 2d}$ is the d-half swapping permutation matrix. Figure 29 demonstrates LowFER parameters for the *family of bilinear models* under the conditions discussed in this section.

5.4.4 *Relation with HypER*

HypER (Balažević et al., 2019a) is a convolutional model based on *hypernetworks* (Ha et al., 2017), where the relation-specific 1D filters are generated by the hypernetwork and convolved with the subject entity vector. Balažević et al. (2019a) showed that it could be understood in terms of tensor factorization up to a non-linearity. With a similar argument, we show that LowFER encodes HypER, bringing it closer to the convolutional approaches.

Dataset	n_e	n_r	n_e/n_r	Training	Validation	Testing
			General KBs			
WN18	40,943	18	2,275	141,442	5,000	5,000
WN18RR	40,943	11	3,722	86,835	3,034	3,134
FB15k	14,951	1,345	11	483,142	50,000	59,071
FB15k-237	14,541	237	61	272,115	17,535	20,466
YAGO10-3	123,182	37	3,329	1,079,040	5,000	5,000
			Biomedical KBs			
UMLS	135	46	4	5,216	625	661
SNOMED CT	293,879	162	1,814	1,965,111	49,103	49,570
SNOMED CT (ES)	137,013	26	5,270	249,110	27,599	55,457

Table 22: Datasets used for KGC experiments, where n_e=number of entities, n_r=number of relations and the entities-to-relations ratio n_e/n_r is approximated to the nearest integer.

HypER scoring function is defined as (Balažević et al., 2019a):

$$\phi_h(e_s, r, e_o) = h(\text{vec}(\mathbf{e}_s * \mathbf{F}_r)\mathbf{W})\mathbf{e}_o \tag{10}$$

where $\mathbf{F}_r = \text{vec}^{-1}(\mathbf{Hr}) \in \mathbb{R}^{n_f \times l_f}$, $\mathbf{H} \in \mathbb{R}^{n_f l_f \times d_r}$ (hypernetwork), $\mathbf{W} \in \mathbb{R}^{n_f l_m \times d_e}$, vec(.) transforms $n \times m$ matrix to nm-sized vector, $\text{vec}^{-1}(.)$ does the reverse operation, $*$ is the convolution operator, $h(.)$ is ReLU non-linearity and n_f, l_f and $l_m = d_e - l_f + 1$ are *number of filters*, *filter length* and *output length* of convolution.

The convolution between a filter and the subject entity embedding can be seen as a matrix multiplication, where the filter is converted to a Toeplitz matrix of size $l_m \times d_e$. With n_f filters, we can realize a 3D tensor of size $n_f \times l_m \times d_e$. Since the filters are generated by the hypernetwork, we have d_r such 3D tensors, resulting in a 4D tensor of size $n_f \times l_m \times d_e \times d_r$ (Balažević et al., 2019a). Without loss of generality, we can view this 4D tensor as a 3D tensor $\mathcal{F} \in \mathbb{R}^{n_f l_m \times d_e \times d_r}$. Taking mode-1 product as $\mathcal{F} \times_1 \mathbf{W}^T$ returns a final tensor $\mathcal{G} \in \mathbb{R}^{d_e \times d_e \times d_r}$. Thus, HypER operations $\text{vec}(\mathbf{e}_s * \mathbf{F}_r)\mathbf{W}$ simplify to $\mathcal{G} \times_3 \mathbf{r} \times_2 \mathbf{e}_s$. At $k = d_e$, with $\mathbf{U} \in \mathbb{R}^{d_e \times d_e^2}$ as block identity matrices (same as in LowFER's relation to RESCAL) and $\mathbf{V} \in \mathbb{R}^{d_r \times de^2}$ set to $\mathbf{G}^T$ ($\mathcal{G}$ viewed as a matrix of size $d_e^2 \times d_r$ and transposed), LowFER's score in Eq. 6 represents HypER, up to the non-linearity.

5.5 EXPERIMENTS

We conducted the experiments on five benchmark datasets in the general domain, WN18 (Bordes et al., 2013), WN18RR (Dettmers et al., 2018), FB15k (Bordes et al., 2013), FB15k-237 (Toutanova et al., 2015) as small-scale KG and YAGO10-3 (Mahdisoltani et al., 2015) as large-scale KG. For biomedical knowledge completion, we consider the UMLS (Kok and Domingos, 2007) as small-scale KG and SNOMED CT (U.S. Edition), extracted from UMLS2019AB, following the steps proposed in Chang et al. (2020) as large-scale KG. We also consider a multilingual, domain-specific, and clinical sub-graph for stroke using SNOMED CT (Spanish Edition) relevant to clinical notes of Table 7 in Chapter 3. See Appendix §C.3 for the dataset details, including best hyperparameters and additional experiments.

		WN18				FB15k			
Linear	**Model**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
No	TransE	0.454	0.089	0.823	0.934	0.380	0.231	0.472	0.641
	Neural LP	0.940	–	–	0.945	0.760	–	–	0.837
	R-GCN	0.819	0.697	0.929	**0.964**	0.696	0.601	0.760	0.842
	ConvE	0.943	0.935	0.946	0.956	0.657	0.558	0.723	0.831
	TorusE	0.947	0.943	0.950	0.954	0.733	0.674	0.771	0.832
	HypER	0.951	0.947	**0.955**	0.958	0.790	0.734	0.829	0.885
Yes	DistMult	0.822	0.728	0.914	0.936	0.654	0.546	0.733	0.824
	HolE	0.938	0.930	0.945	0.949	0.524	0.402	0.613	0.739
	ComplEx	0.941	0.936	0.936	0.947	0.692	0.599	0.759	0.840
	ANALOGY	0.942	0.939	0.944	0.947	0.725	0.646	0.785	0.854
	SimplE	0.942	0.939	0.944	0.947	0.727	0.660	0.773	0.838
	TuckER	**0.953**	**0.949**	**0.955**	0.958	0.795	0.741	0.833	0.892
	LowFER-1	0.949	0.945	0.951	0.956	0.720	0.639	0.774	0.859
	LowFER-10	0.950	0.946	0.952	0.958	0.810	0.760	0.843	0.896
	LowFER-k*	0.950	0.946	0.952	0.958	**0.824**	**0.782**	**0.852**	**0.897**

Table 23: General domain knowledge completion results on WN18 and FB15k.

We implemented LowFER[2] using the open-source code released by TuckER (Balažević et al., 2019b)[3]. We did random search over the embedding dimensions in $\{30, 50, 100, 200, 300\}$ for d_e and d_r. Further, we varied the factorization rank k in $\{1, 5, 10, 30, 50, 100, 150, 200\}$, with $k = 1$ (LowFER-1) and $k = 10$ (LowFER-10) as baselines. For WN18RR and WN18, we found best $d_e = 200$ and $d_r = 30$ with k value of 30 and 10 respectively. For FB15k-237, we found best $d_e = d_r = 200$ at $k = 100$. These embedding dimensions match the best reported in TuckER (Balažević et al., 2019b). However, for FB15k, we found using the configuration of $d_e = 300$ and $d_r = 30$ to be consistently better than $d_e = d_r = 200$. For a fair comparison, we also reported the results for $d_e = d_r = 200$ and the best configuration when $d_e = 200$ and $(d_r, k) \leqslant 200$ (Table 29).

Similar to Balažević et al. (2019b), we used Batch Normalization (Ioffe and Szegedy, 2015) but additionally power normalization and l_2-normalization to stabilize training from large outputs following the Hadamard product in the main scoring function (Yu et al., 2017)[4]. We tested the best-reported hyperparameters of Balažević et al. (2019b) with random search and observed good performance in initial testing. With d_e, d_r and k selected, we used a fixed set of values for the rest of the hyperparameters reported in Balažević et al. (2019b), including learning rate, decay rate, entity embedding dropout, MFB dropout, output dropout and label smoothing (Pereyra et al., 2017; Szegedy et al., 2016) (see Table 43 for the best hyperparameters). We used Adam (Kingma and Ba, 2015a) for optimization. In all the experiments, we trained the models for 500 epochs with batch size 128 and reported the final results on the test set.

[2] https://github.com/suamin/LowFER

[3] https://github.com/ibalazevic/TuckER

[4] We observed no performance degradation by removing these additional normalization techniques but we used it in all the experiments to be consistent with prior work of Yu et al. (2017).

		WN18RR				FB15k-237			
Linear	**Model**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
No	Neural LP	–	–	–	–	0.250	–	–	0.408
	R-GCN	–	–	–	–	0.248	0.151	0.264	0.417
	ConvE	0.430	0.400	0.440	0.520	0.325	0.237	0.356	0.501
	RotatE	–	–	–	–	0.297	0.205	0.328	0.480
	HypER	0.465	0.436	0.477	0.522	0.341	0.252	0.376	0.520
Yes	DistMult	0.430	0.390	0.440	0.490	0.241	0.155	0.263	0.419
	ComplEx	0.440	0.410	0.460	0.510	0.247	0.158	0.275	0.428
	TuckER	**0.470**	**0.443**	**0.482**	**0.526**	0.358	**0.266**	0.394	**0.544**
	LowFER-1	0.454	0.422	0.470	0.515	0.318	0.233	0.348	0.483
	LowFER-10	0.464	0.433	0.477	0.523	0.352	0.261	0.386	0.533
	LowFER-k*	0.465	0.434	0.479	**0.526**	**0.359**	**0.266**	**0.396**	**0.544**

Table 24: General domain knowledge completion results on WN18RR and FB15k-237.

5.5.1 *Results*

General KBs: Table 23 and 24 shows our KGC results in the general domain, where LowFER-1, LowFER-10 and LowFER-k* represent our model for $k = 1$, $k = 10$ and $k =$ best. We choose LowFER-1 and LowFER-10 as baselines. Overall, LowFER reaches competitive performance on all the datasets with state-of-the-art results on FB15k and FB15k-237. On WN18 and WN18RR, TuckER is marginally better than LowFER.

LowFER performs well at low-ranks with significantly less number of parameters compared to other linear models (Table 27). At $k = 1$, it performs better than or on par with non-linear and linear models (including ComplEx and SimplE) except HypER and TuckER. For FB15k-237, LowFER-1 (~3M parameters) outperforms R-GCN, RotatE, DistMult and ComplEx by an average of 5.9% on MRR, and it additionally outperforms convolutional models (ConvE, HypER) at $k = 10$ with only +0.8M parameters. On FB15k, the best reported TuckER model is improved upon, with an absolute +1.9% increase on the toughest Hits@1 metric. This already achieves state-of-the-art with almost half the parameters, ~5.5M, in contrast to TuckER's ~11.3M. On WN18RR and WN18, LowFER-1 outperforms all the models, excluding TuckER and HypER. With LowFER-k*, we just about reach state-of-the-art performance on WN18RR and FB15k-237. On FB15k, we reach new state-of-the-art for ~9.51M parameters with +2.9% and +4.1% improvement on MRR and Hits@1.

The empirical gains can be attributed to LowFER's ability to perform *good* fusion between entities and relations while avoiding overfitting through low-rank matrices remaining parameter efficient, with strong performance even at extreme low-ranks. Further, like TuckER, it allows for parameter sharing through the **U** and **V** matrices, unlike ComplEx and SimplE, which rely only on embedding matrices.

For large-scale KG in the general domain, we report results on YAGO3-10, a subset of YAGO3 (Mahdisoltani et al., 2015), consisting of 123,182 entities and 37 relations such that each entity has at least 10 relations. We used the same best hyperparameters as for WN18RR. Table 25 shows that our model outperforms state-of-the-art models including RotatE and HypER. It is worth noting that LowFER-k* on YAGO3-10 has

Model	MRR	Hits@1	Hits@3	Hits@10
DistMult	0.340	0.240	0.380	0.540
ComplEx	0.360	0.260	0.400	0.550
ConvE	0.440	0.350	0.490	0.620
RotatE	0.495	0.402	0.550	0.670
HypER	0.533	0.455	0.580	0.678
LowFER-k*	**0.537**	**0.457**	**0.583**	**0.688**

Table 25: Large-scale knowledge completion results on YAGO3-10.

Model	MRR	Hits@1	Hits@3	Hits@10
UMLS				
NeuralLP	0.778	0.643	–	0.962
NTP-λ	0.912	0.843	–	**1.000**
MINERVA	0.825	0.728	–	0.968
ComplEx	0.929	0.887	–	0.985
ConvE	**0.940**	**0.902**	–	0.992
LowFER-k*	0.929	0.870	0.985	0.995
SNOMED CT				
TransE	0.346	0.212	–	0.597
ComplEx	**0.461**	0.360	–	**0.652**
DistMult	0.420	0.309	–	0.626
SimplE	0.432	0.337	–	0.615
RotatE	0.317	0.162	–	0.599
LowFER-k∗	0.455	**0.372**	0.494	0.618
SNOMED CT (ES)				
TransE	0.074	0.046	0.0793	0.120
DistMult	0.103	0.067	0.111	0.171
ComplEx	**0.139**	**0.091**	**0.155**	**0.229**
LowFER	0.119	0.087	0.130	0.180

Table 26: Biomedical knowledge completion results with UMLS and SNOMED CT.

only ~26M parameters compared to ~61M parameters of RotatE (Sun et al., 2019b), which also includes their self-adversarial negative sampling.

Biomedical KBs: Table 26 shows our results for all the datasets. We find ComplEx (Trouillon et al., 2016) and LowFER to have comparable performance for UMLS where ConvE performed best. These result discrepancies can partly be attributed to the different subset extractions from the original UMLS (Kok and Domingos, 2007). For SNOMED CT, LowFER had the best performance for the Hits@1 metric at the expense of long-tail performance compared to ComplEx for Hits@10. For domain-specific Spanish KG with SNOMED CT (ES) ComplEx performed relatively better, where we used DGL-KE (Zheng et al., 2020) with default parameters for training the baselines and used our LowFER implementation. For evaluation, LowFER performs 1-N scoring over 137,013 entities where we could only evaluate on 80,000 entities during inference for other models due to memory constraints with the DGL-KE evaluation protocol.

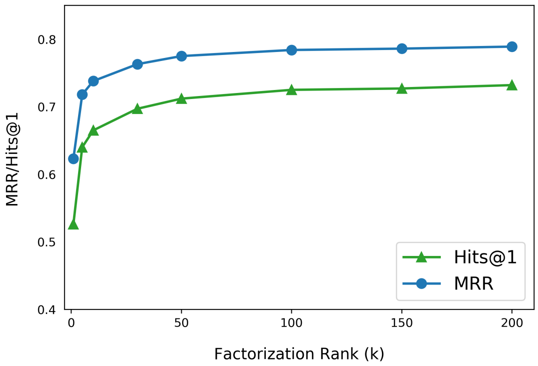


Figure 30: Influence of increasing the LowFER factorization rank on MRR and Hits@1 scores for FB15k.

Model	WN18	FB15k-237	WN18RR	FB15k
ComplEx	16.4	6.0	16.4	6.5
SimplE	16.4	-	16.4	6.5
TuckER	9.4	11.0	9.4	11.3
LowFER-1	8.2	3.0	8.2	4.6
LowFER-10	8.6	3.8	8.6	5.5
LowFER-k*	8.6	11.3	9.6	9.5

Table 27: Comparison between the number of parameters in millions (M) of strong linear models. For LowFER-k*, the k values are 10, 100, 30 and 50 for WN18, FB15k-237, WN18RR and FB15k respectively.

Therefore, the discrepancy in the results may also be due to different implementations and hyperparameter optimization as thoroughly investigated in Ruffinelli et al. (2020).

The overall effectiveness of LowFER on both the general and biomedical domains shows that our model provides a good representation of entities and relations for a given KG to address the remainder of **RQ7**. We supplement further details in Appendix §C.3.3.

5.5.2 *Analysis*

Factorization Rank: From knowledge completion results, we observe that rank plays an important role depending on the entities-to-relations ratio in the dataset. For $d_e = 200$ and $d_r = 30$, we vary k from $\{1, 5, 10, 30, 50, 100, 150, 200\}$ on FB15k and plot the MRR and Hits@1 scores (Figure 30). From $k = 1$ to $k = 5$, the MRR score increases from 0.62 to 0.72 and Hits@1 increases from 0.53 to 0.64. For higher ranks (after 50), the change is minimal. Empirically, the effect of k diminishes as the number of entities per relation becomes larger, e.g., it is ~ 3722 for WN18RR in contrast to ~ 11 for FB15k. We suspect this could be because as $n_e \gg d_e$, most of the knowledge is learned through embedding matrices rather than the model parameters $\mathbf{U}$ and $\mathbf{V}$. To test this, we took a trained LowFER model on the WN18 dataset and added zero mean Gaussian noise with variance in $\{1.0, 1.25, 1.5, 1.75, 2.0\}$ to $\mathbf{U}$ and $\mathbf{V}$ and evaluated on the test set. The MRR score changed from 0.95 to $\{0.92, 0.84, 0.65, 0.42, 0.24\}$ for each level of noise. This shows that, in noisy settings, the embeddings have the potential to capture more knowledge than the shared parameters.

Empirically, we found when $d_e = d_r$, taking $k = d_e/2$ performs nearly the same as TuckER (Balažević et al., 2019b). This can be observed in LowFER-k* for FB15k-

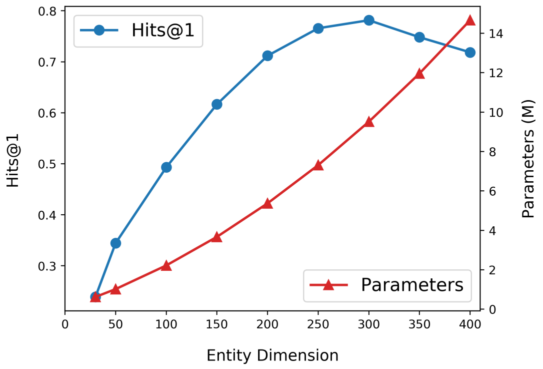


Figure 31: Influence of changing the LowFER entity embedding dimension d_e on Hits@1 metric and growth of parameters in million (M).

k	Params (M)	MRR	Hits@1	Hits@3	Hits@10
1	3.60	0.634	0.538	0.695	0.803
5	3.92	0.720	0.641	0.776	0.860
10	4.33	0.742	0.667	0.790	0.871
30	5.93	0.774	0.709	0.817	0.885
50	7.53	0.776	0.713	0.818	0.886
100	11.53	0.779	0.717	0.821	0.887

Table 28: Knowledge completion results on FB15k with $d_e = d_r = 200$.

237 ($d_e = d_r = 200$, $k = 100$), where our results are almost indistinguishable from TuckER's. This can be expected as the number of parameters in both models is almost the same (~11M). It should be noted that when we train LowFER, we initialize with two i.i.d matrices, which are not shared. This is in contrast to TuckER's core tensor (Eq. 7), allowing us to reach almost the same performance despite less parameter sharing.

Embedding Dimension: The size of entity embedding dimension d_e accounts for the significant number of parameters in LowFER, growing linearly with the number of entities n_e. To study the effect, we trained our models on FB15k, with $d_r = 30$, $k = 50$ constant, and varying de in $\{30, 50, 100, 150, 200, 250, 300, 350, 400\}$. As shown in Figure 31, increasing the entity embedding dimension significantly increases the Hits@1 metric for almost linear growth in the number of parameters. However, it only improves till 300 and starts overfitting afterwards.

In Balažević et al. (2019b), authors reported $d_e = d_r = 200$ as best choice of dimensions for TuckER on FB15k, however, we found using $de = 300$ and $d_r = 30$ better with a lower number of parameters for LowFER. For a fair comparison, we also provide the results for $d_e = d_r = 200$ for k in $\{1, 5, 10, 30, 50, 100\}$ in Table 28. As k increases, we see an improvement over all the metrics. At $k = 100$, where we expected LowFER to match TuckER's performance (MRR=0.795, Hits@1=0.741, ~11 million parameters), it was lower (−1.6% on MRR and −2.4% on Hits@1). In comparison, our model with $d_e = 300$, $d_r = 30$ and $k = 10$ with ~5.6 million parameters only, gives better results than this setting and TuckER. Therefore, at $d_e = d_r = 200$, our model is most likely overfitting.

As noted above it could be that LowFER is overfitting therefore, we did a coarse grid search over the relation embedding dimension in $\{30, 50, 100, 150, 200\}$ and k in $\{1, 5, 10, 30, 50, 100, 150, 200\}$ while keeping $d_e = 200$ fixed. We found $d_r = 50$ at $k = 150$

Model	Params (M)	MRR	Hits@1	Hits@3	Hits@10
TuckER	11.3	0.795	0.741	0.833	0.892
LowFER-k*	10.6	0.795	0.739	0.831	0.891
LowFER-k* + Reg	10.6	0.802	0.749	0.837	0.892

Table 29: Knowledge completion results on FB15k with $d_e = 200$, $d_r = 50$, $k = 150$ and l_2-regularization 0.0005.

	WN18		WN18RR	
	LowFER	TuckER	LowFER	TuckER
also_see	0.638	0.630	0.627	0.614
derivationally_related_form	0.954	0.956	0.957	0.957
has_part	0.944	0.945	0.138	0.129
hypernym	0.961	0.962	0.189	0.189
instance_hypernym	0.986	0.982	0.576	0.591
member_meronym	0.930	0.927	0.155	0.131
member_of_domain_region	0.885	0.885	0.060	0.083
member_of_domain_usage	0.917	0.917	0.025	0.096
similar_to	1.0	1.0	1.0	1.0
synset_domain_topic_of	0.956	0.952	0.494	0.499
verb_group	0.974	0.974	0.974	0.974

Table 30: Relation specific test set results on WN18 and WN18RR with LowFER-k* and best reported TuckER model (Balažević et al., 2019b).

reaches almost the same performance as TuckER with ~10.6M parameters compared to TuckER's ~11.3M parameters. We also experimented with l_2-regularization (Reg) and noted minor improvements, with regularization strength 0.0005. Table 29 summarizes these results. Note that all the experiments reported in the main results were without regularization. In general, we only noticed slight improvements in FB15k with l_2-regularization.

Relation Types: KGC models that can discover relation types automatically without prior knowledge indicate *better* generalization. As shown, and discussed in §5.4, LowFER, among other models (Table 21), can learn to capture all relation types without additional constraints. However, these bounds are loose in practice and require large dimensions, prompting an inspection of their performance on different relation types. In Kazemi and Poole (2018), it was identified that WN18 contains redundant relations, i.e., $\forall e_i, e_j \in \mathcal{E} : (e_i, r_1, e_j) \in \mathcal{T} \Leftrightarrow (e_j, r_2, e_i) \in \mathcal{T}$, such as <*hyponym, hypernym*>, <*meronym, holonym*> etc. To alleviate this, Dettmers et al. (2018) proposed WN18RR with such relations removed since knowledge about one can help infer the knowledge about the other. Table 30 shows the per relation results of LowFER and TuckER on WN18 and WN18RR. We see that performance drops for 7 relations, with an average performance decrease of −70.6% and −69.3% for LowFER and TuckER respectively (with the highest decrease on *member_of_domain_usage* for both). For symmetric relations (such as *derivationally_related_form*), the performance is approximately the same where we observe severe limitations to model asymmetry. We believe this is because LowFER (also TuckER) is constraint-free, and adding certain constraints based on background knowledge is *necessary* to improve the model's accuracy.

SimplE is the only *fully expressive* model that has been formally shown to address these limitations (cf. Proposition 3, 4, and 5 in Kazemi and Poole (2018)). Since LowFER subsumes SimplE, such rules can be studied for extending LowFER to incorporate the background knowledge.

5.6 CONCLUSION

In this Chapter we proposed a simple and parameter-efficient model, LowFER, that performs on par or state-of-the-art in the general and biomedical domain for inferring missing knowledge, thus addressing **RQ7**. It offers a strong baseline to the deep learning based models and raises further interest in studying linear models for KGC.

We showed that LowFER is *fully expressive* and generalizes to other linear models in KGC, providing a unified theoretical view, thus addressing **RQ8**. We also highlighted some limitations concerning gains on harder relations, which still need to be addressed. This shows that the constraint-free and parameter-efficient linear models, which allow for parameter sharing, are better from a modeling perspective but are still similarly limited in learning difficult relations.

6 DISCUSSION AND FUTURE WORK

In this Chapter, we will briefly revisit the research contributions, discuss the work across Chapters and summarize possible future work corresponding to each knowledge acquisition task presented in this dissertation.

6.1 RESEARCH CONTRIBUTIONS

For entity-centric learning, we started by conducting a thorough study of neural methods in Chapter 2 for supervised concept extraction showing the effectiveness of transfer learning with pre-trained multilingual and domain-specific language models which addressed **RQ1** (pp. 15, 17, 23, 30). For unsupervised concept extraction, we presented a Dense Phrase Matching approach which is highly effective when the texts are rich with noun-phrased concepts with a potential to serve in a multilingual setting to address **RQ2** (pp. 15, 25, 29, 30). In Chapter 3, we presented a Transformer based framework for transfer learning research in NER which offers to bridge the gap between growing research in deep transformer models, NER transfer, and domain adaptation addressing **RQ3** (pp. 31, 33, 28, 44). We then applied T2NER to the task of clinical notes de-identification by empirically investigating the *few-shot cross-lingual transfer* property of mBERT and proposed an adaptation strategy that significantly boosts over zero-shot performance while keeping the required size of annotated samples low to address **RQ4** (pp. 32, 39, 41, 43, 44).

For relation-centric learning, we proposed relation-enriched BERT to bag-level multi-instance learning in Chapter 4 and showed that with a KB-sensitive data encoding scheme, it sufficiently reduced distant supervision noise, alleviating the need for additional tasks which addressed **RQ5** (pp. 47, 51, 53, 64). We further investigated the landscape of distantly supervised biomedical relation extraction benchmarks and found either train-test leakage or coverage limitations to propose MedDistant19, which we thoroughly evaluated with scientific language models, showing promising relational representation capacity addressing **RQ6** (pp. 48, 55, 64). In Chapter 5 we proposed a simple and parameter-efficient knowledge graph completion model, LowFER, that performs on par or state-of-the-art in the general and biomedical domain for inferring missing knowledge to address **RQ7** (pp. 65, 68, 73, 79, 82). We showed that LowFER is *fully expressive* and generalizes to other linear models in knowledge graph completion, providing a unified theoretical view that addressed **RQ8** (pp. 66, 70, 71, 82).

6.2 DISCUSSION

Concept extraction is related to biomedical or clinical entity linking, where the first step is to perform named entity recognition and then concept normalization (Neumann et al., 2019). For example, the CANTEMIST (CANcer TExt MIning Shared Task – tumor named entity recognition) 2020 proposes a codes classification task in two stages: NER followed by concept extraction (Miranda-Escalada et al., 2020). The solutions with NER as the first step had consistently better performance than those without, further highlighting that Chapters 2 and 3 are conceptually related. However, we leave such a pipeline approach for future work. Recently, multi-label classification has emerged to address the growing literature on COVID-19 with LitCovid (Chen et al., 2021a) for real-world medical language processing.

While Chapter 3 focused on clinical text de-identification, the methods presented in T2NER generally apply to low-resource medical NER (Crichton et al., 2017), which is a crucial first step for mining triples from the text, the subject of Chapter 4. Furthermore, the sensitive information collected from the text can serve as a PHI-KB itself that can be used for pseudonymization and privacy-enhanced clinical language processing.

The second part of the thesis builds towards knowledge base enrichment with relational learning. The task of mutual learning with relation extraction (Chapter 4) and knowledge graph completion (Chapter 5) has been previously studied (Han et al., 2018a; Toutanova et al., 2015; Weston et al., 2013). The advances in pre-trained language models have shown the presence of relational cues as part of pre-training, which allows PLMs to act as soft KBs (Petroni et al., 2019), thus prompting their use for knowledge extraction.

However, access to an incomplete KB allows learning from statistical correlations to represent entities and relations, as shown in Chapter 5. The two representation learning paradigms from text and KB are complementary but different. When dealing with text, we mainly look at semantics and syntax, while from KB, we primarily learn from the structural components of the underlying knowledge graph. Combining the two is an active research area (Colon-Hernandez et al., 2021).

From a methodological perspective, we consistently found *simple* methods to provide strong empirical gains. Our findings highlight the need for stronger baselines and our limitations show the prevalent challenges of domain-specific NLP.

6.3 FUTURE WORK

Concept Extraction: The methods presented in Chapter 2 can be followed up by joint learning of named entities for concept extraction and normalization. However, required data might be in short supply when dealing with in-domain and low-resource languages (e.g., Estonian or Catalan medical documents). Such deficiencies encourage research for better cross-lingual and cross-domain methods that can be transferred effectively, similar to transfer learning algorithms discussed in Chapter 3. Noting the importance of concept span detection in dense phrase matching, further improvements in this direction are encouraged.

Named Entity Recognition: For T2NER, the following directions can be considered:

- Create benchmark data and compare the transfer learning algorithms (Ramesh Kashyap et al., 2021; Ramponi and Plank, 2020).
- Investigate adding support for traditional few-shot (Huang et al., 2020), nested (Wang et al., 2020b) and document-level (Schweter and Akbik, 2020) NER.
- Assess the framework's performance in terms of speed and efficiency and compare it with other tools[1].

[1] https://github.com/JayYip/bert-multitask-learning

- While the framework focuses on the task of NER, adding related tasks such as relation extraction, entity linking, and question answering might boost performance through multi-task learning.

Furthermore, the results for clinical de-identification show potential for future applications in other low-resource scenarios. A thorough comparative study across domains and languages to qualify the robustness of few-shot transfer should be pursued. Lastly, studying few-shot domain and language adversarial approaches for biomedical NER can be investigated.

Relation Extraction: Compared to methods presented in Chapter 4, a more formal and theoretically ground noise modeling approach could be studied with regularized optimal transport (OT) (Cuturi, 2013), considering a ground metric between textual and KB relation representations. Such an application of OT distances has been shown in T2NER adaptation algorithms (EMD), which can be extrapolated here. Considering the importance and application of tensor factorization (Chapter 5), studying joint factorization approaches with 4th-order tensor decomposition, where one shall consider mutual learning instead of regularized distance with shared tensor, could be explored. Lastly, with the need for emerging multilingual medical applications (Chapter 2 and 3), studying Bio-DSRE for languages other than English could become important, where low-resource conditions are more extreme following the trend in the general domain (Bhartiya, Badola, et al., 2022).

Knowledge Graph Completion: A prominent direction should be comprehensive benchmarking with medical KBs including RepoDB (Brown and Patel, 2017), MSI (Ruiz et al., 2021), Hetionet (Himmelstein and Baranzini, 2015), OpenBioLink (Breit et al., 2020), DRKG (Ioannidis et al., 20 2), and BioKG (Walsh et al., 2020). Studying the impact of regularization schemes as a trade-off between parameter sharing and constraints can be explored. To further enhance the expressivity, a Quaternion (Hamilton, 1866) generalization of Tucker decomposition (Tucker, 1966) can be formulated and further a move towards inductive tensor completion should be investigated.

Part III

APPENDIX

A

FEW-SHOT CROSS-LINGUAL DE-IDENTIFICATION DATASET

In this Appendix, we present additional details about the few-shot cross-lingua de-identification dataset discussed in Chapter 3.

The contents of this Appendix have appeared in ***Amin et al. (2022b)****.*

A.1 GUTTMANN CLINICAL NOTES

In addition to the 7 coarse-grained PHI entities (Stubbs and Uzuner, 2015) discussed in §3.4.3, our dataset contains cross-sentence recurring entities about topics that may be of interest in the clinical domain. These topics are grouped by their potential clinical application areas and are summarized in Table 31.

The label frequency distribution, as noted in Figure 19, is consistent with general characteristics of medical notes, which usually highlight notable events such as symptom onsets, procedures, admissions, transfers, and discharges, in addition to the date of each documentation. As a result, they tend to contain a higher frequency for the DATE PHI, whereas the lower occurrence of the NAME PHI compared to the AGE and LOCATION entities is consistent with how healthcare providers usually communicate patient information.

Medical professionals are trained to refer to patients simply by their age, gender, and the appropriate diagnosis to avoid inadvertently sharing HIPAA-sensitive information, e.g., *"a 60-year-old male with ischemic stroke admitted on [DATE] from [LOCATION] (...)"*. The patient's name may be used at the beginning of a medical note; however, subsequent anaphoric references are often accomplished via pronouns, omitting the NAME entity in the process. In addition, as it is applicable to Spanish medical records, nominative pronouns anaphorically referencing a patient may be omitted as they are grammatical in Spanish. As discussed in the main Chapter, we avoid releasing our dataset due to the presence of real PHI information. However, we will consider replacing the real PHI with synthetic ones, similar to MEDDOCAN, for a possible GDPR-compliant release. Table 32 shows the few-shot training and development corpus statistics.

A.2 ANNOTATION

A.2.1 *Annotators*

Two graduate research assistants completed the annotation of the dataset. Both annotators had at a minimum CEFR[1] B2-C1 Spanish (Castilian) proficiency. One annotator also had clinical experience in the cardiovascular and cerebrovascular specialty, including knowledge of Spanish medical terminology. Neither annotator had formal training in Catalan; both had prior experience working with text data in this language with stroke domain.

[1] https://www.coe.int/en/

Topic Areas	Subcategories
Diagnostics & Treatments	Ischemic vs Hemorrhagic Affected areas and vessels Comorbidities Medication history Associated lifestyle factors Treatments and interventions
Symptoms & Monitoring	Vital signs Lab results and cultures Pain and comfort Bladder and bowel controls
Long-term Care & Discharge Planning	Mobility Cognitive ability Nutrition Psychosocial factors

Table 31: Topics and subcategories in the clinical notes obtained from the Guttmann Institut use to construct the few-shot dataset.

A.2.2 *Guidelines*

The annotation process followed criteria for each entity as described in Stubbs and Uzuner (2015). The 7 entities: AGE, CONTACT, DATE, ID, LOCATION, NAME, and PROFESSION represent a larger granularity of the 18 HIPAA-defined PHI (Stubbs and Uzuner, 2015). We examined the training sets of i2b2 (Stubbs and Uzuner, 2015) and MEDDOCAN (Marimon et al., 2019) and adapted the i2b2 annotation guidelines to create our own annotation guidelines. This step was necessary since we only focused on coarse-grained PHI types compared to fine-grained types considered in these two datasets. The adjusted guidelines utilized in this annotation process are summarized in Table 36.

A.2.3 *Procedure*

Both annotators reviewed and revised their work without discussion or knowledge of the other annotator's work. In cross-revision, the reviewing annotator only made corrections when *labeling inconsistencies* were due to a lack of medical terminology comprehension. During revision, no changes to the original annotator's confidence level rating were made.

Confidence Level: The criteria for the confidence levels are annotator dependent as summarized in Table 35 with examples. PHI has been manually modified from the original data to preserve privacy while maintaining exemplary characteristics for each label entity type.

Skipped Sentences: Each annotator followed an independent set of criteria to exclude sentences from annotation, as demonstrated by examples in Table 34.

Description	Observation
Sentences annotated by annotator (A)	4400
Sentences annotated by annotator (B)	4343
Sentences annotated and revised by (A, B)	**4314**
Agreements	3924
Disagreements	390
Token-level Cohen's Kappa score	0.898
DEVELOPMENT CORPUS	**3924**
w entity mentions	1493
w/o entity mentions	2431
FEW-SHOT TARGET CORPUS	**384**
w entity mentions	369
w/o entity mentions	15

Table 32: Few-shot Cross-lingual De-identification dataset annotation and final statistics.

A.2.4 *Disagreements and Resolution*

An attempt was made to review the 390 sentences where our annotators disagreed to find a resolution. Main sources of disagreement were due to (a) annotation criteria discrepancy, (b) ambiguity between related entities, and (c) annotation errors. After further revision to correct identified errors and clarify ambiguous annotation criteria, an agreement was reached for 384 sentences while 6 sentences were left unannotated due to insufficient context. Confidence levels from both annotators were left unchanged. These sentences constituted our *few-shot target corpus* in the pipeline explained in Figure 18.

Inclusion Criteria Discrepancy: Most disagreements were related to discrepancies in the inclusion of surrounding words such as determiners, punctuation marks, and descriptive phrases. This is prevalent, particularly, in the LOCATION and PROFESSION entities. One annotator considered denoted sentences with these characteristics a lower confidence level of 4 compared to sentences without determiners or punctuation marks surrounding LOCATION tokens. The resolution step changed the annotations to be more consistent with the annotation guidelines described in Table 36.

Ambiguity Between Related Entities: Another source of disagreements in the LOCATION PHI stems from abbreviation usage and confusion with the NAME PHI. In instances where the syntax is ambiguous, annotators may not be able to infer correctly that certain unknown abbreviations are place names. Since it is common that places are named after people's names and vice versa, a lack of contextual information created unresolvable ambiguity regarding the NAME and LOCATION entities. DATE and AGE also demonstrated similar disagreement behavior. In particular, numerical and text expressions involving 'years' may express age or time depending on context.

Annotation Errors: A few disagreements were due to mislabeling or erroneous omissions. There were fewer than 5 such instances in the 390 disagreements. Notable errors were associated with mislabeling proper names that resemble valid named entities. For instance, some assessment tools are named after people or places names e.g. Barcelona Test and Boston (Naming) Test.

PHI	Fine-grained Types
AGE	EDAD_SUJETO_ASISTENCIA
CONTACT	NUMERO_TELEFONO, NUMERO_FAX, CORREO_ELECTRONICO, URL_WEB
DATE	FECHAS
ID	ID_ASEGURAMIENTO, ID_CONTACTO_ASISTENCIAL, NUMERO_BENEF_PLAN_SALUD, IDENTIF_VEHICULOS_NRSERIE_PLACAS, IDENTIF_DISPOSITIVOS_NRSERIE, IDENTIF_BIOMETRICOS, ID_SUJETO_ASISTENCIA, ID_TITULACION_PERSONAL_SANITARIO, ID_EMPLEO_PERSONAL_SANITARIO, OTRO_NUMERO_IDENTIF
LOCATION	HOSPITAL, INSTITUCION, CALLE, TERRITORIO, PAIS, CENTRO_SALUD
NAME	NOMBRE_SUJETO_ASISTENCIA, NOMBRE_PERSONAL_SANITARIO
PROFESSION	PROFESION
OTHER	SEXO_SUJETO_ASISTENCIA, FAMILIARES_SUJETO_ASISTENCIA, OTROS_SUJETO_ASISTENCIA, DIREC_PROT_INTERNET

Table 33: MEDDOCAN PHI (coarse-grained) and fine-grained types (Marimon et al., 2019).

Annotator	Sample sentence	Explanation
A	"Actualmente reside en XXXX-Xxxx Xxxxxxxx, Treballadora Social." [Currently resides in XXXX-Xxxxx Xxxxxxxx, social worker.]	The underlined words are grouped as a single word token. From the context it's clear that 'XXXX' belongs to 'LOCATION' and 'Xxxxx Xxxxxxxx' are 'NAME' entities.
B	"Lmarxa. [sic]" [March or walks]	Annotator does not have enough context to understand this token to annotate.

Table 34: Examples of sentences skipped by annotators and rationales.

A.3 MEDDOCAN NORMALIZATION

The original MEDDOCAN dataset (Marimon et al., 2019) provides document-level de-identification annotations, following 2014 i2b2/UTHealth (Stubbs and Uzuner, 2015), of 1000 clinical notes which are divided into 500, 250, and 250 for training, validation, and testing respectively. It contains 29 fine-grained entity types classified into 8 coarse-grained PHI types as shown in Table 33. Compared to i2b2 (2014), MEDDOCAN has an additional OTHER category which we normalized to O in the BIO schema, resulting in 7 coarse-grained PHI types considered in Chapter 3. We tokenized the 500 training notes resulting in a total of 16,299 sentences. The conversion script is publicly available with T2NER. [2]

[2] https://github.com/suamin/T2NER/blob/master/utils/convert_i2b2style_xml_to_conll.py

Level	Annotator	Criteria	Sample Sentence	Explanation
1	A	Annotator is unable to assign labels due to insufficient contextual information from the given sentence.	“PASe." [PASe.] or [ENTer.]	The token may be an unknown acronym or an oddly typed imperative form of the verb “to enter". Insufficient context.
	B	Annotator is unable to assign labels due to lack of comprehension.	“Allitat, en DDLL.” [n/a]	Annotator did not understand this sentence in Catalan sufficiently to annotate.
2	A	Annotator is unsure about the assigned labels due to contextual ambiguity.	““50 años.” [50 years]	Without any surrounding context, the years can be ‘AGE’ or a temporal expression; annotator thinks it’s most likely to be AGE, but does not feel confident enough to make a determination.
	B	Annotator is unsure about the assigned labels due to lack of medical knowledge or terminology.	“Urocultiu [sic] 13.01: + per *A. baumanii* multiR.” Urine Culture 13.01: + for MDR *A. baumanii*	Annotator omitted this sentence due to uncertainty about the word *A. baumanii*, whether it could be a NAME or a non-labeled entity.
3	A	Annotator is confident about the labels, but some context maybe missing that could change the entity labels.	“712345678)."	This is likely a phone number CONTACT, but may also be an ID entity.
	B	Annotator is confident about the sentence in general, but has some doubt due to presumed lack of specialized knowledge.	“hipoTA [sic] asintomática." [Asymptomatic hypotension.]	Annotator did not specify any label but was unsure whether there was a labeled entity or not.
4	A	Annotator is confident about the labels, but the sentence may have some inconsistencies with the gold standard sentences.	"El marido la vió y llamó a la ambulancia e ingresó en el hospital de Xxxxxxx." [The spouse saw her and called the ambulance and she was admitted to Xxxxxxx hospital.]	Annotator was unsure whether to only annotate ‘Xxxxxxx’ or ‘hospital de Xxxxxxx’ or ‘el hospital de Xxxxxxx’ as LOCATION
	B	Annotator is confident about the labels, but the sentence may have some inconsistencies with the gold standard sample sentences.	“Torna d’Oftalmologia de Xxx Xxxx (Dra. [sic]" [Returns from Xxx Xxxx Ophthalmology (Dra.]	Annotation unsure whether or not to include Ophthalmology as part of ‘LOCATION’
5	A	Annotator is confident and there’s no ambiguity regarding name entities of the labels. This could mean that the sentences have no entities to be annotated or that all the entities needing annotations are consistent with the gold standard sample sentences.	“Cito a control el próximo 25.12.20 y doy pautas a la esposa." [I make a follow-up appointment for the upcoming date 25.12.20 and I give the prescription to the wife.]	It’s clear that ‘25.12.20’ is a DATE PHI.
	B	It is clear to the annotator that the sentence has no entities to be annotated or that the entities are consistent with gold standard annotation. This could be either apparent at first glance or because the sentence has been seen several times before, which increases the annotator’s confidence regarding the assigned label(s).	“Cito a control el próximo 25.12.20 y doy pautas a la esposa." [I make a follow-up appointment for the upcoming date 25.12.20 and I give the prescription to the wife.]	It’s clear that ‘25.12.20’ is a DATE PHI.

Table 35: Confidence level criteria and examples as reported by the two annotators. In instances where PHI entities are utilized in the examples, we replaced the characters with generic alphanumeric characters or with fictitious information (while maintaining the same PHI type).

PHI	Criteria
AGE	Annotate only the numerical part of the expression; include both numerical and word expressions of age (e.g. 36 or thirty-six). Include the words 'years', 'months', and 'days' when they express age. Include expressions that describe an age group e.g. 'adolescent', 'recently born', 'newborn'. Include punctuation associated with age, including separate tokens, e.g. in his/her 30's.
CONTACT	All forms of contact information, e.g. pager, phone numbers, e-mail address. Physical or mailing address is annotated as 'Location' Include punctuation and symbols that occur with contact information, e.g. include all tokens in '(123) 456-789'.
DATE	Include days of the week and months. Include punctuation in all formats. Include the word 'year' and 'month' that is part of a date-time expression, e.g. 'the year 2000'. Include prepositions that are part of a date-time expression, e.g. include the word 'of' in '5th of May'.
ID	Include all identification numbers such as Medical Record Number (MRN), Social Security Number (SSN), Document ID, device lot number, etc. Include any alpha-numeric expressions appearing at the beginning of the document or next to a name that's not formatted as a date. When separated by punctuation, annotate all parts of the expression including punctuation, e.g. include all tokens in '12-34-5678'. Exclude the ID descriptive words and associated punctuation, e.g. exclude 'MRN' and ':' in 'MRN: 1234567'.
LOCATION	Include all place names and all parts of an address: street name, city, state, county, province, region, and country. Include punctuation in the address. Include Zip/postal codes. Include organization names. Include words that specify the location when they appear as part of a 'Location' entity, e.g. include the word 'Center' in 'Social Security Center'.
NAME	Include only the person's names. Include punctuation between first and last names when present Exclude titles and salutations.
PROFESSION	Include all professional titles, e.g. annotate 'MD' in the phrase 'X works as an MD'. Exclude professional titles in name suffixes, e.g. exclude 'MD' in the phrase 'Dr. X Y, MD'. Include professional and occupational descriptions, e.g. annotate 'carpentry in the phrase 'X works in carpentry'. Annotate the entire expression describing a profession, e.g. annotate all tokens in a phrase such as 'worker in a cafeteria'. Exclude workplace names; annotate workplace names as 'Location' instead.

Table 36: Adjusted annotation guidelines with examples for each PHI type.

B

UMLS.V2 AND MEDDISTANT19 BENCHMARKS

In this Appendix, we present additional details for the UMLS.v2 and MedDistant19 benchmarks discussed in Chapter 4.

The contents of §B.2 have appeared in the Appendix of ***Amin et al. (2020a)****. The contents of §B.1 and §B.3 have appeared in the Appendix of* ***Amin et al. (2022a)****.*

B.1 UMLS FILES

In UMLS (Bodenreider, 2004), a concept is provided with a unique identifier called Concept Unique Identifier (CUI), a term status (TS), and whether or not the term is preferred (TTY) in a given vocabulary, e.g., SNOMED CT. The concepts are stored in a file distributed by UMLS called `MRCONSO.RRF`.[1] Each concept further belongs to one or more semantic types (STY), provided in a file called `MRSTY.RRF`, with a type identifier TUI. There are 127 STY[2] in the UMLS2019AB version, which are mapped to 15 semantic groups (SG).[3]. The relationships between the concepts are organized in a multi-relational graph distributed in a file called `MRREL.RRF`[4].

B.2 UMLS.V2 BENCHMARK

Similar to Dai et al. (2019), UMLS[5] (Bodenreider, 2004) was used as our KB and MEDLINE abstracts[6] as our text source. A data summary is shown in Table 37. We approximate the same statistics as reported in Dai et al. (2019) for relations and entities, but it is important to note that the data does not contain the same samples. We divided triples into the train, validation, and test sets, and following Dai et al. (2019) and Weston et al. (2013), we make sure that there are no overlapping textual facts across the splits. Additionally, we added a constraint that there is no sentence-level overlap between the training and held-out sets. We performed group negative sampling, where 20% of the data was reserved for testing, and of the remaining 80%, we used 10% for validation and the rest for training.

Triples	Entities	Relations	Positive Groups	Negative Groups
169,438	27,403	355	92,070	64,448

Table 37: Overall statistics of the UMLS.v2 benchmark.

[1] https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.concept_names_and_sources_file_mr/

[2] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/SemanticTypes_2018AB.txt

[3] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/SemGroups_2018.txt

[4] https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.related_concepts_file_mrrel_rrf/?report=objectonly

[5] We used 2019 release: `umls-2019AB-full`

[6] https://www.nlm.nih.gov/bsd/medline.html

B.2.1 *Knowledge Base*

The fact triples were obtained for English concepts, filtering for RO relation types only (Dai et al., 2019). We collected 9.9M (CUI_head, relation_text, CUI_tail) triples.

B.2.2 *Documents*

From 34.4M abstracts, we extracted 160.4M unique sentences. To perform a fast and scalable search, we used a Trie data structure[7] to index all the textual descriptions of UMLS entities. To obtain a clean set of sentences, we set the minimum and maximum sentence character length to 32 and 256, respectively, and further considered only those sentences where matching entities were mentioned only once. The latter decision was to lower the noise that may come when only one instance of multiple occurrences is marked for a matched entity. With these constraints, the data was reduced to 118.7M matching sentences.

B.2.3 *Groups Linking and Negative Sampling*

Recall the entity groups $\mathcal{G} = \mathcal{G}^+ \cup \mathcal{G}^-$ (§4.3.1). For training with NA relation class, we generated hard negative samples with an open-world assumption (Baldini Soares et al., 2019; Lerer et al., 2019) suited to *bag-level* multiple instance learning (MIL). From 9.9M triples, we removed the relation type and collected 9M CUI groups in the form of (h, t). Since each CUI is linked to more than one textual form, all text combinations for two entities were considered for a given pair, resulting in 531M textual groups $\mathcal{T}$ for the 586 relation types.[8]

Next, for each matched sentence, let $\mathcal{P}_s^2$ denote the size 2 permutations of entities present in the sentence, then $\mathcal{T} \cap \mathcal{P}_s^2$ return groups which are *present in KB* and *have matching evidence* (positive groups, $\mathcal{G}^+$). Simultaneously, with a probability of $\frac{1}{2}$, we removed the h or t entity from this group and replaced it with a novel entity e in the sentence, such that the resulting group (e, t) or (h, e) belonged to $\mathcal{G}^-$. This method resulted in sentences that were seen both for the true triple and the invalid ones. Further, using the constraints that the relation group sizes must be between 10 to 1500, we found 354[9] relation types (approximately the same as Dai et al. (2019)) with 92K positive groups and 2.1M negative groups, which were reduced to 64K by considering a random subset of 70% of the positive groups.

B.2.4 *Bag Composition and Splits*

For bag composition, we created constant-sized bags by randomly under- or over-sampling the sentences in the bag (Han et al., 2019) to avoid larger bias towards common entities (Baldini Soares et al., 2019). The true distribution had a long tail, with more than 50% of the bags having 1 or 2 sentences. We defined a bag to be *uniform* if the special markers represented the same entity in each sentence, either h or t. If the special markers can take on both h or t, we considered that bag to have a *mix* composition. The *k-tag* scheme, on the other hand, naturally generates uniform bags. Further, to

[7] https://github.com/vi3k6i5/flashtext

[8] This step partially leads to the training-test leakage that was identified and addressed in MedDistant19.

[9] 355 including NA relation

Model	Split	Triples	Triples (w/o NA)	Groups	Sentences (Sampled)
k-tag	train	92,972	48,563	92,972	1,487,552
	valid	13,555	8,399	15,963	255,408
	test	33,888	20,988	38,860	621,760
s-tag	train	91,555	47,588	125,852	2,013,632
	valid	13,555	8,399	22,497	359,952
	test	33,888	20,988	55,080	881,280
s-tag+exprels	train	125,155	71,402	125,439	2,007,024
	valid	22,604	16,298	22,607	361,712
	test	55,083	39,282	55,094	881,504

Table 38: Sumary statistics of UMLS.v2 benchmark splits.

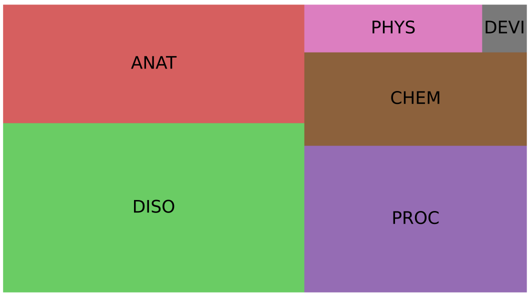


Figure 32: Relative proportions of the entities present in MedDistant19, based on the semantic groups.

support the setting of Wu and He (2019), we followed the *s-tag* scheme and expanded the relations by adding a suffix to denote the directions as $r(e_1, e_2)$ or $r(e_2, e_1)$, with the exception of the NA class, resulting in 709 classes. For fair comparisons with *k-tag*, we generated uniform bags with *s-tag* as well, by keeping e_1 and e_2 the same per bag. Due to these bag composition and class expansion (in one setting, *exprels*) differences, we generated three different splits, supporting each scheme, with the same test sets in cases where the classes were not expanded, and a different test set when the classes are expanded. Table 38 shows the statistics for these splits.

B.3 MEDDISTANT19 BENCHMARK

This section presents additional details about MedDistant19, including the final set of relations considered (with their inverses obtained from the UMLS) and a complete list of semantic types (STY). Since, in relation extraction (RE), we are not interested in bidirectional extractions; therefore it is sufficient to only model one direction. For more discussion on the relations in UMLS, including transitive closures, see §3.1 in Chang et al. (2020). We used UMLS2019AB to be consistent with the related works. The final set of relations considered in MedDistant19 is presented in Table 41. Note that we only considered relations belonging to the *RO* (*has a relationship other than synonymous, narrower, or broader*) type, which is consistent with UMLS.v2. This consideration ignores relations such as *isa*, which defines hierarchy among relations.

B.3.1 *Semantic Groups and Semantic Types*

As we noted in Figure 23, entities and relations follow a long-tail distribution. This has a significant impact on the quality of the dataset created. For example, in the general domain, the standard benchmark, NYT10 (Riedel et al., 2010), has more than half of the positive instances belonging to one relation type `/location/location/contains`. Figure 32 shows the relative proportions of the semantic groups in MedDistant19.

Further, we used an inductive split set with 70%, 10%, and 20% proportions of train, validation, and test splits for constructing MedDistant19. Below are example instances from the benchmark in OpenNRE (Han et al., 2019) format:

```
{
  "text": "In one patient who showed an increase of plasma
     prolactin level , associated with low testosterone and
     LH , a microadenoma of the pituitary gland ( prolactinoma )
     was detected .",
  "h": {
    "id": "C0032005", "pos": [130, 145],
    "name": "pituitary gland"
  },
  "t": {
    "id": "C0033375", "pos": [148, 160],
    "name": "prolactinoma"
  },
  "relation": "finding_site_of"
}

/----------------------------------------------------------/

{
  "text": "Severe heart disease may result in cardiac cirrhosis
     in the elderly , with ascites and hepatomegaly .",
  "h": {
    "id": "C0018799", "pos": [7, 20],
    "name": "heart disease"
  },
  "t": {
    "id": "C0085699", "pos": [35, 52],
    "name": "cardiac cirrhosis"
  },
  "relation": "cause_of"
}

/----------------------------------------------------------/

{
  "text": "Complications closely associated to the osteosynthesis
     appeared only in instable fractures ( 7 % ) .",
  "h": {
    "id": "C0016658", "pos": [81, 90],
    "name": "fractures"
  },
  "t": {
    "id": "C0016642", "pos": [40, 54],
    "name": "osteosynthesis"
   },
   "relation":
   "direct_morphology_of"
}

/----------------------------------------------------------/

{
  "text": "Gluten proteins , the culprits in celiac disease
     ( CD ) , show striking similarities in primary
     structure with human salivary proline-rich proteins
     ( PRPs ) .",
  "h": {
    "id": "C2362561", "pos": [0, 15],
    "name": "Gluten proteins"
  },
  "t": {
    "id": "C0007570", "pos": [34, 48],
    "name": "celiac disease"
  },
  "relation":
  "causative_agent_of"
}

/----------------------------------------------------------/

{
  "text": "Postherpetic neuralgia is an unfortunate aftermath
     of shingles , and is most likely to develop , and most
     persistent , in elderly patients .",
  "h": {
    "id": "C0032768", "pos": [0, 22],
    "name": "Postherpetic neuralgia"
  },
  "t": {
    "id": "C0019360", "pos": [54, 62],
    "name": "shingles"
  },
  "relation": "occurs_after"
}
```

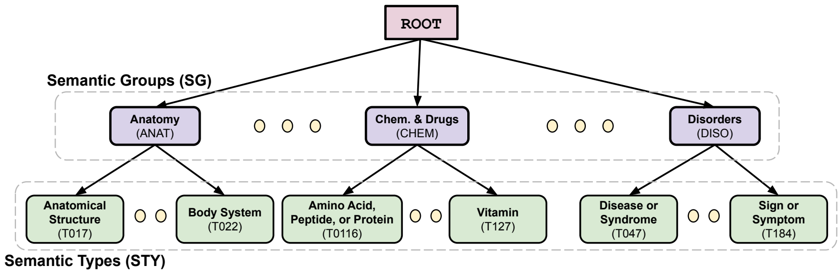


Figure 33: Type Hierarchy: each concept in the UMLS is classified under a type taxonomy. The *coarse-grained* and *fine-grained* entity types are referred to as **Semantic Group (SG)** and **Semantic Type (STY)** respectively.

B.3.2 *UMLS License Agreement*

To use the MedDistant19 benchmark, the user must have signed the UMLS agreement[10]. The UMLS agreement requires those who use the UMLS (Bodenreider, 2004) to file a brief report once a year to summarize their use of the UMLS. It also requires acknowledging that the UMLS contains copyrighted material and that those copyright restrictions are respected. The UMLS agreement requires users to agree to obtain agreements for EACH copyrighted source before its use within a commercial or production application.

B.3.3 *Risks*

While MedDistant19 does not have direct risk, we provide the dataset while asking users to respect the UMLS license before downloading it. This user agreement is needed to use our benchmark and to respect the source ontologies licenses. We provide this with the hope to accelerate reproducible research in Bio-DSRE by having ready-to-use corpora, with only the condition that the user has obtained the license. We provide users with this note and hope this will be respected. However, there is a risk that users may download the data and re-distribute it without respecting the UMLS license. In case of such exploitation, we will add the UMLS authentication layer to protect data where the user will be required to provide a UMLS API key, which will be validated, and only then will the data be allowed to be downloaded.

B.3.4 *Limitations*

We provide several limitations of our work as presented in its current form. MedDistant19 aims to introduce a new benchmark with good practices. However, it is still limited in the scope of ontologies considered. It also has a limited subset of relation types provided by UMLS. For example, the current benchmark does not include an important relation *may_treat* because it is outside SNOMED CT. Since MedDistant19 is focused on SNOMED CT, it lacks coverage of important protein-protein interactions, drug side-effects, and relations involving genes as provided by RxNorm, Gene Ontology, etc.

MedDistant19 is automatically created and susceptible to noise and thus needs to be approached carefully as a potential source for biomedical knowledge. While the

[10] https://uts.nlm.nih.gov/license.html

Encoder	Bag Size	Batch Size	Embedding
CNN+sent+AVG	-	128	biowordvec
CNN+sent+ONE	-	128	biowordvec
CNN+bag+AVG	8	128	GloVe
CNN+bag+ONE	16	256	GloVe
CNN+bag+ATT	8	256	GloVe
PCNN+sent+AVG	-	128	biowordvec
PCNN+sent+ONE	-	128	biowordvec
PCNN+bag+AVG	4	128	GloVe
PCNN+bag+ONE	8	128	GloVe
PCNN+bag+ATT	8	128	GloVe
GRU+sent+AVG	-	128	biowordvec
GRU+sent+ONE	-	128	biowordvec
GRU+bag+AVG	8	128	biow2v
GRU+bag+ONE	16	256	GloVe
GRU+bag+ATT	16	128	GloVe

Table 39: Best hyperparameters for CNN, PCNN, and GRU sentence encoders.

dataset was not created to represent *true* biomedical knowledge, it has the potential to be treated as a reliable reference.

B.3.5 *Experimental Setup and Hyperparameters*

We followed the experimental setup of Gao et al. (2021a) for BERT-based experiments. Specifically, we used batch size 64, with a learning rate of 2e-5, maximum sequence length 128, and bag size 4. We used a single NVIDIA Tesla V100-32GB for BERT-based experiments. Each experiment took about 1.5 hrs, with half an hour per epoch. We also attempted to perform a grid search for BERT experiments, but it was too expensive to continue; therefore, we abandoned those jobs. Since we only used the `base` models, they amount to 110 million parameters. During fine-tuning, we did not freeze any parts of the model.

For CNN and PCNN, we performed grid search with Adam (Kingma and Ba, 2015b) optimizer using learning rate 0.001 for 20 epochs with batch size $\in \{128, 256\}$, bag size $\in \{4, 8, 16, 32\}$, 200-d word embeddings $\in$ {Word2Vec (Mikolov et al., 2013)[11], GloVe (Pennington et al., 2014)}, and with (test-time) pooling $\in$ {ONE, AVG} when using sentence-level training and pooling in {ONE, AVG, ATT} when using bag-level training. We ran this task on a cluster with support for array jobs. These amounted to over 700 experiments and took 3 days. We fixed other hyperparameters from literature (Han et al., 2018a), with position dimension set to 5, kernel size set to 3, and dropout set to 0.5. These are also the default in OpenNRE (Han et al., 2019). The hyperparameters that had the most influence were batch size, bag size, and pre-trained word embeddings. All the experiments reported in MedDistant19 were with a single run.

For sentence tokenization with SciSpacy, it took 9hrs with 32 CPUs (4GB each) and a batch size of 1024 to extract 151M sentences. Further, the SciSpacy entity linking job took about half TB of RAM with 72 CPUs (6GB each) with a batch size of 4096 with 40hrs of run-time to link 145M unique sentences.

[11] We used domain-specific word embeddings *biowordvec* and *biow2v* similar to Marchesin and Silvello (2022).

Semantic Type	**10k-20k**	**20k-30k**	**⩾ 30k**
Body Part, Organ, or Organ Component	*bladder, heart, retinal, lungs, spinal, kidneys, colon*	*eyes, lung, kidney, intestinal*	*liver, brain*
Organism Function	*death*	*period, blood pressure*	-
Body Location or Region	*head*	-	-
Therapeutic or Preventive Procedure	*injection, prevention, chemotherapy, application* *resection, infusion, treatments, therapeutic* *surgical treatment, CT, surgical, transplantation*	*stimulation, delivery*	*intervention, procedure, removal, operation*
Neoplastic Process	*cancer*	-	*tumor, tumors*
Disease or Syndrome	*obesity, disorder, disorders*	*diseases, stroke*	*disease, infection, condition, hypertension*
Laboratory Procedure	*test, erythrocytes*	-	*cells*
Diagnostic Procedure	*US, biopsy, ultrasound*	*MRI*	-
Finding	*lesion, interaction, mass, difficulty, dependent*	*abnormal*	*presence, positive, negative, severe, lesions*
Hormone	*insulin*	-	-
Biologically Active Substance	*amino acids, glucose, ATP*	*protein, proteins*	
Pharmacologic Substance	*medication*	-	*drugs, drug*
Injury or Poisoning	*strains*	*injury, exposure*	*damage*
Tissue	*tissue, bone marrow, tissues*	-	-
Organism Attribute	*male*	-	*temperature, age*
Immunologic Factor	*antibody, antibodies*	-	-
Health Care Activity	*investigations*	*examination*	*assessment*
Body Substance	*plasma, blood, skin*	-	-
Body System	-	*cardiovascular*	-
Mental Process	-	-	*concentrations, concentration*
Congenital Abnormality	-	*abnormalities*	-

Table 40: Semantic types affected by type-based mention pruning with removed mentions placed in their respective frequency bins.

B.4 DISCUSSION

In the biomedical domain, health experts are often concerned with a particular type of interaction, for example, drug-target and gene-disease. However, the number of ontologies is constantly growing (222 in UMLS2019AB), thus a growing need for a more general-purpose relation extraction benchmark. Broad-coverage benchmarks exist for biomedical entity linking, such as MedMentions (Mohan and Li, 2018), but they still lack many important concepts involved in relational learning. The research community has come up with several RE benchmarks (see Table 10), but the challenge remains as new entities and relations emerge with the constant growth of biomedical literature. Hence, constructing a broad benchmark for biomedical RE is challenging due to domain requirements; nonetheless, having an accurate benchmark could offer utility for future research.

Further, the train-test overlap highlights the need to systematically assess the proposed benchmarks for inconsistencies that can overestimate the model performance. Similar assessments have shown in QA generalization where train-test overlap inflates the model performance (Liu et al., 2022). Related to RE generalization, Rosenman et al. (2020) exposed shallow heuristics while Taillé et al. (2021) showed that neural RE models could retain triples, primarily due to type hints. MedDistant19 partially addresses these issues by an inductive setup that can offer insights into the generalization trend in biomedical RE using unseen entities.

Relation	**Inverse Relation**
finding_site_of	*has_finding_site*
associated_morphology_of	*has_associated_morphology*
method_of	*has_method*
interprets	*is_interpreted_by*
direct_procedure_site_of	*has_direct_procedure_site*
causative_agent_of	*has_causative_agent*
active_ingredient_of	*has_active_ingredient*
interpretation_of	*has_interpretation*
component_of	*has_component*
indirect_procedure_site_of	*has_indirect_procedure_site*
direct_morphology_of	*has_direct_morphology*
cause_of	*due_to*
direct_substance_of	*has_direct_substance*
uses_device	*device_used_by*
focus_of	*has_focus*
direct_device_of	*has_direct_device*
procedure_site_of	*has_procedure_site*
uses_substance	*substance_used_by*
associated_finding_of	*has_associated_finding*
occurs_after	*occurs_before*
is_modification_of	*has_modification*

Table 41: *(Left)* 21 relations included in MedDistant19, excluding NA relation. *(Right)* For completeness, we also include their inverse relations.

SG	TUI	Semantic Type
ANAT	T017	Anatomical Structure
	T029	Body Location or Region
	T023	Body Part, Organ, or Organ Component
	T030	Body Space or Junction
	T031	Body Substance
	T022	Body System
	T021	Fully Formed Anatomical Structure
	T024	Tissue
CHEM	T116	Amino Acid, Peptide, or Protein
	T195	Antibiotic
	T123	Biologically Active Substance
	T103	Chemical
	T200	Clinical Drug
	T196	Element, Ion, or Isotope
	T126	Enzyme
	T131	Hazardous or Poisonous Substance
	T125	Hormone
	T129	Immunologic Factor
	T130	Indicator, Reagent, or Diagnostic Aid
	T197	Inorganic Chemical
	T114	Nucleic Acid, Nucleoside, or Nucleotide
	T109	Organic Chemical
	T121	Pharmacologic Substance
	T192	Receptor
	T127	Vitamin
DEVI	T074	Medical Device
	T075	Research Device
DISO	T020	Acquired Abnormality
	T190	Anatomical Abnormality
	T049	Cell or Molecular Dysfunction
	T019	Congenital Abnormality
	T047	Disease or Syndrome
	T033	Finding
	T037	Injury or Poisoning
	T048	Mental or Behavioral Dysfunction
	T191	Neoplastic Process
	T046	Pathologic Function
	T184	Sign or Symptom
PHYS	T201	Clinical Attribute
	T041	Mental Process
	T032	Organism Attribute
	T040	Organism Function
	T042	Organ or Tissue Function
	T039	Physiologic Function
PROC	T060	Diagnostic Procedure
	T065	Educational Activity
	T058	Health Care Activity
	T059	Laboratory Procedure
	T063	Molecular Biology Research Technique
	T062	Research Activity
	T061	Therapeutic or Preventive Procedure

Table 42: 51 semantic types (STY) along with their TUIs and semantic groups (SG) covered in MedDistant19.

C LOWFER PROOFS AND EXPERIMENTAL DETAILS

In this Appendix, we provide proofs and additional experimental details for LowFER discussed in Chapter 5.

The contents of this Appendix have appeared in ***Amin et al. (2020b)*** *with the exception of biomedical KBs.*

C.1 PROOFS

C.1.1 *Proposition 1*

Proof. First, we will prove the case for $k = d_e$, with the proof for the case $k = d_r$ following a similar argument. For both cases, we represent entity embedding vector as $\mathbf{e}_i \in \{0, 1\}^{|\mathcal{E}|}$, such that only i-th element is 1, and similarly, relation embedding vector as $\mathbf{r}_j \in \{0, 1\}^{|\mathcal{R}|}$, such that only j-th element is 1. We represent with $\mathbf{U} \in \mathbb{R}^{d_e \times k d_e}$ and $\mathbf{V} \in \mathbb{R}^{d_r \times k d_e}$ the model parameters, then, given any triple $(e_i, r_j, e_l) \in \mathcal{T}$ with indices (i, j, l), such that $1 \leqslant i, l \leqslant |\mathcal{E}|$ and $1 \leqslant j \leqslant |\mathcal{R}|$:

For $k = d_e$: We let $\mathbf{U}_{mn} = 1$ for $n = m + (o - 1)d_e$, for all m in $\{1, ..., d_e\}$ and for all o in $\{1, ..., k\}$ and 0 otherwise. Further, let $\mathbf{V}_{pq} = 1$ for $p = j$ and $q = (l - 1)d_e + i$ and 0 otherwise. Applying $\mathbf{g}(e_i, r_j)$ and taking dot product of the resultant vector with $\mathbf{e}_l$ (Eq. 6) perfectly represents the ground truth as 1. Also, for any triple in $\mathcal{T}'$, a score of 0 is assigned.

For $k = d_r$: We let $\mathbf{U}_{mn} = 1$ for $m = i$ and $n = (l - 1)d_e + j$ and 0 otherwise. Further, let $\mathbf{V}_{pq} = 1$ for $q = p + (o - 1)d_e$, for all p in $\{1, ..., d_r\}$ and for all o in $\{1, .., k\}$ and 0 otherwise. Rest of the argument follows the same as for $k = d_e$. □

C.1.2 *Proposition 2*

Proof. From Eq. 8 and 9, observe that the m-th slice of the core tensor $\mathcal{W}$ on object dimension is approximated by adding k rank-1 matrices, each of which is a cross product between m-th column in $\mathbf{W}_U^{(l)}$ and m-th column in $\mathbf{W}_V^{(l)}$, for all l in $\{1, ..., d_e\}$. Each slice of the core tensor $\mathcal{W}$ on object dimension has a maximum rank $\min(d_e, d_r)$ and from Singular Value Decomposition (SVD), there exists n ($\leqslant \min(d_e, d_r)$) scaled left singular and scaled right singular vectors whose sum of the cross products is equal to the slice. By choosing these scaled left singular vectors, scaled right singular vectors and zero vectors (in case the rank of the corresponding slice is less than the maximum rank of any such slice) as columns for matrices $\mathbf{W}_U^{(l)}$, $\mathbf{W}_V^{(l)}$, for all l in $\{1, ..., d_e\}$, the core tensor $\mathcal{W}$ is obtained from Eq. 8 with $k \leqslant \min(d_e, d_r)$. □

C.1.3 *Proposition 3*

RotatE (Sun et al., 2019b), a state-of-the-art dissimilarity-based model alleviates the issues of TransE by learning counterclockwise rotations in the complex space. For a triple (h, r, t), RotatE models the tail entity as $\mathbf{t} = \mathbf{h} \circ \mathbf{r}$, where $\mathbf{h}, \mathbf{t} \in \mathbb{C}^d$ are head and tail embeddings and $\mathbf{r} \in \mathbb{C}^d$ is the relation embedding with a restriction on the element-wise modulus, $|r_i| = 1$. Therefore, it only affects the phases of the entity embeddings in the complex vector space. Sun et al. (2019b) showed that it can learn *symmetric*, *assymmetric*, *inverse* and *composition* relations (cf. Lemma 1, 2, 3) and degenerates to TransE (cf. Theorem 4). However, we note that RotatE is also not *fully expressive* due to its inability to model the transitive relations in the general case, i.e., irrespective of the size of the embedding dimension.

Proposition 3. *RotatE is not fully expressive due to a limitation on transitive relations.*

Proof. Consider $\{e_1, e_2, e_3\} = \Delta \subset \mathcal{E}$ and $r \in \mathcal{R}$ be a transitive relation on Δ such that $r(e_1, e_2), r(e_2, e_3)$ and $r(e_1, e_3)$ belong to the ground truth. Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{r} \in \mathbb{C}^d$ be the embedding vectors for RotatE. Let us assume that $r(e_1, e_2)$ and $r(e_2, e_3)$ hold with RotatE, then we get $\mathbf{e}_2 = \mathbf{r} \circ \mathbf{e}_1$ and $\mathbf{e}_3 = \mathbf{r} \circ \mathbf{e}_2$. From definition of transitive relation we know that $r(e_1, e_2) \wedge r(e_2, e_3) \implies r(e_1, e_3)$, here we obtain $\mathbf{e}_3 = \mathbf{r} \circ \mathbf{r} \circ \mathbf{e}_1$. Therefore for $r(e_1, e_3)$ to hold with RotatE, we must have $\mathbf{r} \circ \mathbf{r} = \mathbf{r} \implies \mathbf{r} = \mathbf{1}$, which in turn suggest $\mathbf{e}_1 = \mathbf{e}_2 = \mathbf{e}_3$ but e_1, e_2, e_3 are distinct entities. More concretely, this condition requires that for all elements of relation embedding r_i, $\cos(\theta_{r,i}) + i\sin(\theta_{r,i})$ should match $\cos(2\theta_{r,i}) + i\sin(2\theta_{r,i})$, which is only possible when $\theta_{r,i} \in \{0, 2\pi\}$, effectively no rotation. □

C.2 SCORING SUBSUMPTIONS

Here we briefly summarize the subsumption findings of related works. We only discuss the published findings and refrain from any implied results.

First, Hayashi and Shimbo (2017) showed the equivalence between ComplEx and HolE up to a constant factor using Parseval's theorem[1], which was also discussed in Trouillon and Nickel (2017). Then, the key contributions came from the work of Wang et al. (2018b), who showed that RESCAL subsumes TransE, ComplEx, HolE, and DistMult by the arguments of ranking tensor. Kazemi and Poole (2018) presented a unified understanding of RESCAL, DistMult, ComplEx, and SimplE as the *family of bilinear models* under different constraints on the bilinear map. In contrast to the black box 2D-convolution based ConvE model, HypER (Balažević et al., 2019a) showed that 1D-convolution with *hypernetworks* (Ha et al., 2017) come close to well-established factorization based methods up to a non-linearity. Balažević et al. (2019b) showed that with certain constraints on the core tensor of the Tucker decomposition (Tucker, 1966), it could subsume the *family of bilinear models*.

In §5.4.2, we showed that LowFER subsumes TuckER and can be seen as providing a low-rank approximation of the core tensor[2] with accurate representation under certain conditions (Proposition 2). We also showed that LowFER could subsume the *family of*

[1] For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it states that $\mathbf{x}^T\mathbf{y} = \frac{1}{d}\mathcal{F}(\mathbf{x})^T\overline{\mathcal{F}(\mathbf{y})}$, where $\mathcal{F} : \mathbb{R}^d \to \mathbb{C}^d$ is the discrete Fourier transform (DFT).

[2] The rank of a tensor is the minimal number of rank-1 tensors that yield it in a linear combination. It is known that the tensor rank is NP-hard to compute, and for a 3rd-order tensor $n \times m \times k$, it can be more than $\min(n, m, k)$ but no more than $\min(nm, nk, mk)$ (Miettinen, 2011). Whereas, the n-rank of a tensor

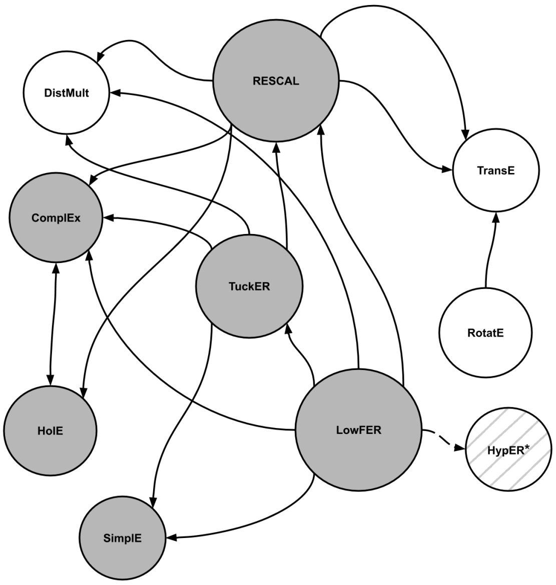


Figure 34: Subsumption map of KGC models for known relationships: each node represents a model, where a directed edge shows that the parent node has shown to subsume the child under some conditions. The dotted line shows that the relationship is not general enough, where the grey nodes represent *fully expressive* models, the white nodes represent the models that have shown to be not *fully expressive*, and the dashed ones are ones where this property is not known. The size of a node is relative to the number of outgoing edges. * HypER (Balažević et al., 2019a) has shown to be related to factorization-based methods up to a non-linearity, but the authors did not specify any explicit modeling subsumption of other models.

bilinear models in §5.4.3 and HypER up to a non-linearity in §5.4.4. Figure 34[3] provides a network style map for the models discussed here.

C.3 EXPERIMENTS

In this section, we will present the evaluation metrics' details, choice of hyperparameters and report additional experiments.

C.3.1 *Evaluation Metrics*

In §5.5 we reported the results with standard metrics of Mean Reciprocal Rank (MRR) and Hits@k for $k \in \{1, 3, 10\}$. In general, for each test triple (e_s, r, e_o), we score all the triples (e_s, r, e) for all $e \in \mathcal{E}$. We then compute the inverse rank of the true triple and average them over all the examples. However, Bordes et al. (2013) identified an issue with this evaluation and introduced *filtered* MRR, where we only consider triples of the form $\{(e_s, r, e) \mid \forall e \in \mathcal{E} \text{ s.t. } (e_s, r, e) \notin \text{train} \cup \text{valid} \cup \text{test}\}$ during evaluation. We, therefore, reported *filtered* MRR for all the experiments. The Hits@k metric computes the percentage of test triples whose ranking is less than or equal to k.

$\mathcal{W}$ is the dimension of the vector space spanned by the n-mode vectors, which are the columns of the matrix unfolding $\mathbf{W}_{(n)}$ (De Lathauwer et al., 2000).

[3] https://bit.ly/3k641Ba

Dataset	lr	dr	d_e	d_r	k	dE	dMFB	dOut	ls
WN18	0.005	0.995	200	30	10	0.2	0.1	0.2	0.1
WN18RR	0.01	1.0	200	30	30	0.2	0.2	0.3	0.1
FB15k	0.003	0.99	300	30	50	0.2	0.2	0.3	0.0
FB15k-237	0.0005	1.0	200	200	100	0.3	0.4	0.5	0.1
YAGO10-3	0.01	1.0	200	30	30	0.2	0.2	0.3	0.1
UMLS	0.001	1.0	200	200	100	0.3	0.4	0.5	0.1
SNOMED CT	0.0005	1.0	200	200	100	0.3	0.4	0.5	0.1
SNOMED CT (ES)	0.0005	1.0	256	32	12	0.3	0.4	0.5	0.1

Table 43: Best performing hyper-parameter values for LowFER, where lr=learning rate, dr=decay rate, d_e=entity embedding dimension, d_r=relation embedding dimension, k=LowFER factorization rank, dE=entity embedding dropout, dMFB=MFB dropout, dOut=output dropout and ls=label smoothing. Please note that dE, dMFB and dOut are the same as d#1, d#2 and d#3 as in TuckER (see Appendix A in Balažević et al. (2019b)) respectively.

Dataset	MRR	Hits@1	Hits@3	Hits@10
FB15k-237 ↓	0.345	0.256	0.378	0.526
FB15k ↓	0.818	0.771	0.850	0.898
WN18RR ↓	0.457	0.429	0.469	0.511
WN18	0.950	0.946	0.952	0.957

Table 44: KGC results with LowFER-k* and additional `tanh` non-linearity. The ↓ shows that the performance went down compared to the linear counterparts reported in Tables 23 and 24.

C.3.2 *LowFER with Non-linearity*

Similar to Kim et al. (2017), here we perform a simple ablation study by adding non-linearity to the LowFER scoring function as follows:

$$\bar{f}(e_s, r, e_o) = (\sigma(\mathbf{S}^k \mathrm{diag}(\mathbf{U}^T \mathbf{e}_s)\mathbf{V}^T \mathbf{r}))^T \mathbf{e}_o$$

where we use hyperbolic tangent $\sigma = \mathtt{tanh}$ non-linearity. Applying a non-linear activation function can be seen as increasing the representation capacity of the model, but Table 44 shows that the general performance of LowFER goes down.

C.3.3 *Models Comparison*

In §5.5, we compared LowFER with non-linear models including ConvE (Dettmers et al., 2018), R-GCN (Schlichtkrull et al., 2018), Neural LP (Yang et al., 2017), RotatE (Sun et al., 2019b)[4], TransE (Bordes et al., 2013), TorusE (Ebisu and Ichise, 2018) and HypER (Balažević et al., 2019a). In linear models, we compared against DistMult (Yang et al., 2015), HolE (Nickel et al., 2016b), ComplEx (Trouillon et al., 2016), ANALOGY (Liu et al., 2017a), SimplE (Kazemi and Poole, 2018) and state-of-the-art TuckER (Balažević et al., 2019b) model. Results for the Canonical Tensor Decomposition (Lacroix et al., 2018) were not included due to the uncommon choice of extremely large embedding dimensions of $d_e = d_r = 2000$.

Additional models that were not reported in the main results due to partial results but were still outperformed by LowFER include M-Walk (Shen et al., 2018) with their

[4] Where we reported their results in Table 24 without the self-adversarial negative sampling. For a fair comparison, see Appendix H in their paper.

reported metrics of MRR=0.437, Hits@1=0.414, and Hits@3=0.445 on WN18RR and MINERVA (Das et al., 2018) with Hits@10=0.456 on FB15k-237.

The results in Table 23 and 24 for all the models were taken from Balažević et al. (2019a) and Balažević et al. (2019b) respectively. Lastly, to perform per relations comparisons, we trained the TuckER models with the best-reported configurations in Balažević et al. (2019b) for WN18 and WN18RR.

For YAGO10-3, results for DistMult, ComplEx, and ConvE were taken from Dettmers et al. (2018) and for RotatE (Sun et al., 2019b) (with self-adversarial negative sampling) and HypER (Balažević et al., 2019a) were taken from respective papers.

For biomedical knowledge graph completion, results for UMLS were reported from Lin et al. (2018), for SNOMED CT from Chang et al. (2020), and for SNOMED CT (ES) are our own runs.

BIBLIOGRAPHY

[1] Asma Ben Abacha and Pierre Zweigenbaum. "Automatic extraction of semantic relations between medical entities: a rule based approach." In: *Journal of biomedical semantics* 2.5 (2011), pp. 1–11.

[2] Oshin Agarwal, Yinfei Yang, Byron C Wallace, and Ani Nenkova. "Interpretability analysis for named entity recognition to understand system predictions and how they can improve." In: *ArXiv preprint* abs/2004.04564 (2020). URL: https://arxiv.org/abs/2004.04564.

[3] Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. "Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus." In: *Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) Held in Conjunction with LREC*. Citeseer. 2012, pp. 49–54.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. "Publicly Available Clinical BERT Embeddings." In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://aclanthology.org/W19-1909.

[5] Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. "TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 1558–1569. DOI: 10.18653/v1/2020.acl-main.142. URL: https://aclanthology.org/2020.acl-main.142.

[6] Christoph Alt, Marc Hübner, and Leonhard Hennig. "Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 1388–1398. DOI: 10.18653/v1/P19-1134. URL: https://aclanthology.org/P19-1134.

[7] Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva, and Günter Neumann. "A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction." In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, 2020, pp. 187–194. DOI: 10.18653/v1/2020.bionlp-1.20. URL: https://aclanthology.org/2020.bionlp-1.20.

[8] Saadullah Amin, Pasquale Minervini, David Chang, Pontus Stenetorp, and Günter Neumann. "MedDistant19: Towards an Accurate Benchmark for Broad-Coverage Biomedical Relation Extraction." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 2259–2277. URL: https://aclanthology.org/2022.coling-1.198/.

[9] Saadullah Amin and Günter Neumann. "T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 2021, pp. 212–220. URL: https://aclanthology.org/2021.eacl-demos.25.

[10] Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. "MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT." In: *Proceedings of the 20th Conference and Labs of the Evaluation Forum (Working Notes)*. Lugano, Switzerland: CEUR Workshop Proceedings, 2019, pp. 1–15. URL: http://ceur-ws.org/Vol-2380/paper_67.pdf.

[11] Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Wixted, Alejandro García-Rudolph, Catalina Martínez-Costa, and Günter Neumann. "Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts." In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 200–211. DOI: 10.18653/v1/2022.bionlp-1.20. URL: https://aclanthology.org/2022.bionlp-1.20.

[12] Saadullah Amin, Stalin Varanasi, Katherine Ann Dunfield, and Günter Neumann. "LowFER: Low-rank Bilinear Pooling for Link Prediction." In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 257–268. URL: http://proceedings.mlr.press/v119/amin20a.html.

[13] Reinald Kim Amplayo, Kyungjae Lee, Jinyeong Yeo, and Seung-won Hwang. "Translations as Additional Contexts for Sentence Classification." In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. Ed. by Jérôme Lang. ijcai.org, 2018, pp. 3955–3961. DOI: 10.24963/ijcai.2018/550. URL: https://doi.org/10.24963/ijcai.2018/550.

[14] Sophia Ananiadou and John McNaught. "Text mining for biology and biomedicine." In: (2006).

[15] Martín Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks." In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 214–223. URL: http://proceedings.mlr.press/v70/arjovsky17a.html.

[16] Alan R Aronson and François-Michel Lang. “An overview of MetaMap: historical perspective and recent advances.” In: *Journal of the American Medical Informatics Association* 17.3 (2010), pp. 229–236.

[17] Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. “Generalisation in named entity recognition: A quantitative analysis.” In: *Computer Speech & Language* 44 (2017), pp. 61–83.

[18] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1409.0473.

[19] Ivana Balažević, Carl Allen, and Timothy M Hospedales. “Hypernetwork knowledge graph embeddings.” In: *International Conference on Artificial Neural Networks*. Springer. 2019, pp. 553–565.

[20] Ivana Balažević, Carl Allen, and Timothy Hospedales. “TuckER: Tensor Factorization for Knowledge Graph Completion.” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 5185–5194. DOI: 10.18653/v1/D19-1522. URL: https://aclanthology.org/D19-1522.

[21] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. “Matching the Blanks: Distributional Similarity for Relation Learning.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 2895–2905. DOI: 10.18653/v1/P19-1279. URL: https://aclanthology.org/P19-1279.

[22] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. “Multi-label classification of patient notes: case study on ICD code assignment.” In: *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[23] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: A Pretrained Language Model for Scientific Text.” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3615–3620. DOI: 10.18653/v1/D19-1371. URL: https://aclanthology.org/D19-1371.

[24] Asma Ben Abacha and Pierre Zweigenbaum. “Automatic extraction of semantic relations between medical entities: a rule based approach.” In: *Journal of biomedical semantics* 2.5 (2011), pp. 1–11.

[25] Bettina Bert, Antje Dörendahl, Nora Leich, Julia Vietze, Matthias Steinfath, Justyna Chmielewska, Andreas Hensel, Barbara Grune, and Gilbert Schönfelder. “Rethinking 3R strategies: Digging deeper into AnimalTestInfo promotes transparency in in vivo biomedical research.” In: *PLoS biology* 15.12 (2017), e2003217.

[26] Abhyuday Bhartiya, Kartikeya Badola, et al. “DiS-ReX: A Multilingual Dataset for Distantly Supervised Relation Extraction.” In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2022, pp. 849–863.

[27] Olivier Bodenreider. “The unified medical language system (UMLS): integrating biomedical terminology.” In: *Nucleic acids research* 32.suppl_1 (2004), pp. D267–D270.

[28] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. “Freebase: a collaboratively created graph database for structuring human knowledge.” In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008, pp. 1247–1250.

[29] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. “Translating Embeddings for Modeling Multi-relational Data.” In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger. 2013, pp. 2787–2795. URL: https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

[30] Rabia Bounaama and Abderrahim Amine. “Tlemcen University at CELF eHealth 2018 Team techno: Multilingual Information Extraction-ICD10 coding.” In: CLEF. 2018.

[31] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. “Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research.” In: *BMC bioinformatics* 16.1 (2015), pp. 1–17.

[32] Anna Breit, Simon Ott, Asan Agibetov, and Matthias Samwald. “OpenBioLink: a benchmarking framework for large-scale biomedical link prediction.” In: *Bioinformatics* 36.13 (2020), pp. 4097–4098.

[33] Adam S Brown and Chirag J Patel. “A standard database for drug repositioning.” In: *Scientific data* 4.1 (2017), pp. 1–7.

[34] Kathi Canese and Sarah Weis. “PubMed: the bibliographic database.” In: *The NCBI Handbook* 2 (2013), p. 1.

[35] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. “Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss.” In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 1565–1576. URL: https://proceedings.neurips.cc/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html.

[36] Yu Cao, Meng Fang, Baosheng Yu, and Joey Tianyi Zhou. "Unsupervised Domain Adaptation on Reading Comprehension." In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 7480–7487. URL: https://aaai.org/ojs/index.php/AAAI/article/view/6245.

[37] J Douglas Carroll and Jih-Jie Chang. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition." In: *Psychometrika* 35.3 (1970), pp. 283–319.

[38] David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Andrew Taylor. "Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings." In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, 2020, pp. 167–176. DOI: 10.18653/v1/2020.bionlp-1.18. URL: https://aclanthology.org/2020.bionlp-1.18.

[39] Moses Charikar, Kevin Chen, and Martin Farach-Colton. "Finding frequent items in data streams." In: *Theoretical Computer Science* 312.1 (2004), pp. 3–15.

[40] Qingyu Chen, Alexis Allot, and Zhiyong Lu. "LitCovid: an open database of COVID-19 literature." In: *Nucleic acids research* 49.D1 (2021), pp. D1534–D1540.

[41] Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. "AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 743–753. DOI: 10.18653/v1/2021.acl-long.61. URL: https://aclanthology.org/2021.acl-long.61.

[42] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. "Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification." In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 557–570. DOI: 10.1162/tacl_a_00039. URL: https://aclanthology.org/Q18-1039.

[43] Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, Christian Lovis, et al. "Use and understanding of anonymization and de-identification in the biomedical literature: scoping review." In: *Journal of medical Internet research* 21.5 (2019), e13484.

[44] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, Danilo P Mandic, et al. "Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions." In: *Foundations and Trends® in Machine Learning* 9.4-5 (2016), pp. 249–429.

[45] Jacob Cohen. "A coefficient of agreement for nominal scales." In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[46] Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. "Combining pre-trained language models and structured knowledge." In: *ArXiv preprint* abs/2101.12294 (2021). URL: https://arxiv.org/abs/2101.12294.

[47] Alexis Conneau and Guillaume Lample. "Cross-lingual Language Model Pretraining." In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 7057–7067. URL: https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

[48] Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll. "Automatic Code Assignment to Medical Text." In: *Biological, translational, and clinical language processing*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 129–136. URL: https://aclanthology.org/W07-1017.

[49] Mark Craven, Johan Kumlien, et al. "Constructing biological knowledge bases by extracting information from text sources." In: *ISMB*. Vol. 1999. 1999, pp. 77–86.

[50] Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. "A neural network multi-task learning approach to biomedical named entity recognition." In: *BMC bioinformatics* 18.1 (2017), pp. 1–14.

[51] Silviu Cucerzan. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data." In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 708–716. URL: https://aclanthology.org/D07-1074.

[52] Aron Culotta and Jeffrey Sorensen. "Dependency Tree Kernels for Relation Extraction." In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barcelona, Spain, 2004, pp. 423–429. DOI: 10.3115/1218955.1219009. URL: https://aclanthology.org/P04-1054.

[53] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger. 2013, pp. 2292–2300. URL: https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

[54] George Cybenko. "Approximation by superpositions of a sigmoidal function." In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[55] Qin Dai, Naoya Inoue, Paul Reisert, Ryo Takahashi, and Kentaro Inui. "Distantly Supervised Biomedical Knowledge Acquisition via Knowledge Graph Based Attention." In: *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 1–10. DOI: 10.18653/v1/W19-2601. URL: https://aclanthology.org/W19-2601.

[56] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. "Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning." In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=Syg-YfWCW.

[57] Hal Daumé III. "Frustratingly Easy Domain Adaptation." In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 256–263. URL: https://aclanthology.org/P07-1033.

[58] Daniel Daza, Michael Cochez, and Paul Groth. "Inductive Entity Representations from Text via Link Prediction." In: *Proceedings of The Web Conference 2021*. 2021. DOI: 10.1145/3442381.3450141.

[59] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "A multilinear singular value decomposition." In: *SIAM journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[60] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. "De-identification of patient notes with recurrent neural networks." In: *Journal of the American Medical Informatics Association* 24.3 (2017), pp. 596–606.

[61] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. "Convolutional 2D Knowledge Graph Embeddings." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 1811–1818. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17366.

[62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[63] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[64] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. "NCBI disease corpus: a resource for disease name recognition and concept normalization." In: *Journal of biomedical informatics* 47 (2014), pp. 1–10.

[65] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. "Knowledge vault: a web-scale approach to probabilistic knowledge fusion." In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. Ed. by Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani. ACM, 2014, pp. 601–610. DOI: 10.1145/2623330.2623623. URL: https://doi.org/10.1145/2623330.2623623.

[66] Kevin Donnelly et al. "SNOMED-CT: The advanced terminology and coding system for eHealth." In: *Studies in health technology and informatics* 121 (2006), p. 279.

[67] Takuma Ebisu and Ryutaro Ichise. "TorusE: Knowledge Graph Embedding on a Lie Group." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 1819–1826. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16227.

[68] Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. *The Anonymisation Decision-Making Framework. UKAN*. 2016.

[69] Benjamin Farber, Dayne Freitag, Nizar Habash, and Owen Rambow. "Improving NER in Arabic Using a Morphological Tagger." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/625_paper.pdf.

[70] Jun Feng, Minlie Huang, Mingdong Wang, Mantong Zhou, Yu Hao, and Xiaoyan Zhu. "Knowledge graph embedding by flexible translation." In: *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*. 2016.

[71] Rémi Flicoteaux. "ECSTRA-APHP@ CLEF eHealth2018-task 1: ICD10 Code Extraction from Death Certificates." In: CLEF. 2018.

[72] Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. "Adversarial Learning of Privacy-Preserving Text Representations for De-Identification of Medical Records." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 5829–5839. DOI: 10.18653/v1/P19-1584. URL: https://aclanthology.org/P19-1584.

[73] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 457–468. DOI: 10.18653/v1/D16-1044. URL: https://aclanthology.org/D16-1044.

[74] Yaroslav Ganin and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." In: *International conference on machine learning*. PMLR. 2015, pp. 1180–1189.

[75] Tianyu Gao, Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyu Xie, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. "Manual Evaluation Matters: Reviewing Test Protocols of Distantly Supervised Relation Extraction." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 1306–1318. DOI: 10.18653/v1/2021.findings-acl.112. URL: https://aclanthology.org/2021.findings-acl.112.

[76] Tianyu Gao, Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyu Xie, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. "Manual Evaluation Matters: Reviewing Test Protocols of Distantly Supervised Relation Extraction." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 1306–1318. DOI: 10.18653/v1/2021.findings-acl.112. URL: https://aclanthology.org/2021.findings-acl.112.

[77] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. "Compact Bilinear Pooling." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 317–326. DOI: 10.1109/CVPR.2016.41. URL: https://doi.org/10.1109/CVPR.2016.41.

[78] Julien Gobeill and Patrick Ruch. "Instance-based learning for ICD10 categorization." In: CLEF. 2018.

[79] Yoav Goldberg. "Assessing BERT's Syntactic Abilities." In: *ArXiv preprint* abs/1901.05287 (2019). URL: https://arxiv.org/abs/1901.05287.

[80] Yves Grandvalet and Yoshua Bengio. "Semi-supervised Learning by Entropy Minimization." In: *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*. 2004, pp. 529–536. URL: https://proceedings.neurips.cc/paper/2004/hash/96f2b50b5d3613adf9c27049b2a888c7-Abstract.html.

[81] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. "Learning Word Vectors for 157 Languages." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. URL: https://aclanthology.org/L18-1550.

[82] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-specific language model pretraining for biomedical natural language processing." In: *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1 (2021), pp. 1–23.

[83] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. "Improved Training of Wasserstein GANs." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 5767–5777. URL: https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

[84] Patrick P Gunn, Allen M Fremont, Melissa Bottrell, Lisa R Shugarman, Jolene Galegher, and Tora Bikson. "The health insurance portability and accountability act privacy rule: a practical guide for researchers." In: *Medical care* (2004), pp. 321–327.

[85] David Ha, Andrew M. Dai, and Quoc V. Le. "HyperNetworks." In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=rkpACe1lx.

[86] William Rowan Hamilton. *Elements of quaternions*. London: Longmans, Green, & Company, 1866.

[87] Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. "More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction." In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2020, pp. 745–758. URL: https://aclanthology.org/2020.aacl-main.75.

[88] Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. "OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 169–174. DOI: 10.18653/v1/D19-3029. URL: https://aclanthology.org/D19-3029.

[89] Xu Han, Zhiyuan Liu, and Maosong Sun. "Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 4832–4839. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16691.

[90] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. "FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4803–4809. DOI: 10.18653/v1/D18-1514. URL: https://aclanthology.org/D18-1514.

[91] Richard A Harshman. "Models for analysis of asymmetrical relationships among N objects or stimuli." In: *First Joint Meeting of the Psychometric Society and the Society of Mathematical Psychology, Hamilton, Ontario, 1978*. 1978.

[92] Richard A Harshman and Margaret E Lundy. "PARAFAC: Parallel factor analysis." In: *Computational Statistics & Data Analysis* 18.1 (1994), pp. 39–72.

[93] Tzvika Hartman, Michael D Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, et al. "Customization scenarios for de-identification of clinical notes." In: *BMC medical informatics and decision making* 20.1 (2020), pp. 1–9.

[94] Katsuhiko Hayashi and Masashi Shimbo. "On the Equivalence of Holographic and Complex Embeddings for Link Prediction." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 554–559. DOI: 10.18653/v1/P17-2088. URL: https://aclanthology.org/P17-2088.

[95] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. "The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions." In: *Journal of biomedical informatics* 46.5 (2013), pp. 914–920.

[96] Daniel S Himmelstein and Sergio E Baranzini. "Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes." In: *PLoS computational biology* 11.7 (2015), e1004259.

[97] Frank L Hitchcock. "The expression of a tensor or a polyadic as a sum of products." In: *Journal of Mathematics and Physics* 6.1-4 (1927), pp. 164–189.

[98] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. "Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 541–550. URL: https://aclanthology.org/P11-1055.

[99] William P Hogan, Molly Huang, Yannis Katsis, Tyler Baldwin, Ho-Cheol Kim, Yoshiki Baeza, Andrew Bartko, and Chun-Nan Hsu. "Abstractified Multi-instance Learning (AMIL) for Biomedical Relation Extraction." In: *3rd Conference on Automated Knowledge Base Construction*. 2021.

[100] L. Hong, J. Lin, S. Li, F. Wan, H. Yang, T. Jiang, D. Zhao, and J. Zeng. "A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories." In: *Nature Machine Intelligence* 2 (2020), pp. 347–355. DOI: 10.1038/s42256-020-0189-y. URL: https://www.nature.com/articles/s42256-020-0189-y.

[101] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear. 2017.

[102] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks." In: *Neural networks* 4.2 (1991), pp. 251–257.

[103] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. "Few-Shot Named Entity Recognition: A Comprehensive Study." In: *ArXiv preprint* abs/2012.14978 (2020). URL: https://arxiv.org/abs/2012.14978.

[104] Vassilis N Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xiangxiang Zeng, and George Karypis. "Drkg-drug repurposing knowledge graph for covid-19." In: *ArXiv preprint* abs/2010 (20 2). URL: https://arxiv.org/abs/2010.

[105] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, 2015, pp. 448–456. URL: http://proceedings.mlr.press/v37/ioffe15.html.

[106] Julia Ive, Natalia Viani, David Chandran, André Bittar, and Sumithra Velupillai. "KCL-Health-NLP@ CLEF eHealth 2018 Task 1: ICD-10 Coding of French and Italian Death Certificates with Character-Level Convolutional Neural Networks." In: *19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018, Avignon, France, 10 September 2018 through 14 September 2018*. Vol. 2125. CEUR-WS. 2018.

[107] Serena Jeblee, Akshay Budhkar, Saša Milic, Jeff Pinto, Chloé Pou-Prom, Krishnapriya Vishnubhotla, Graeme Hirst, and Frank Rudzicz. "Toronto CL at CLEF 2018 eHealth Task 1: Multi-lingual ICD-10 Coding using an Ensemble of Recurrent and Convolutional Neural Networks." In: CLEF. 2018.

[108] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. "Knowledge Graph Embedding via Dynamic Mapping Matrix." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015, pp. 687–696. DOI: 10.3115/v1/P15-1067. URL: https://aclanthology.org/P15-1067.

[109] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. "A survey on knowledge graphs: Representation, acquisition, and applications." In: *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[110] Chen Jia, Xiaobo Liang, and Yue Zhang. "Cross-Domain NER using Cross-Domain Language Modeling." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 2464–2474. DOI: 10.18653/v1/P19-1236. URL: https://aclanthology.org/P19-1236.

[111] Chen Jia and Yue Zhang. "Multi-Cell Compositional LSTM for NER Domain Adaptation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 5906–5917. DOI: 10.18653/v1/2020.acl-main.524. URL: https://aclanthology.org/2020.acl-main.524.

[112] Junguang Jiang, Bo Fu, and Mingsheng Long. *Transfer-Learning-library*. https://github.com/thuml/Transfer-Learning-Library. 2020.

[113] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. "MIMIC-III, a freely accessible critical care database." In: *Scientific data* 3.1 (2016), pp. 1–9.

[114] Jeff Johnson, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs." In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.

[115] Rie Johnson and Tong Zhang. "Deep Pyramid Convolutional Neural Networks for Text Categorization." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 562–570. DOI: 10.18653/v1/P17-1052. URL: https://aclanthology.org/P17-1052.

[116] Seyed Mehran Kazemi and David Poole. "SimplE Embedding for Link Prediction in Knowledge Graphs." In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett. 2018, pp. 4289–4300. URL: https://proceedings.neurips.cc/paper/2018/hash/b2ab001909a8a6f04b51920306046ce5-Abstract.html.

[117] Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, and João Palotti. "Overview of the CLEF eHealth Evaluation Lab 2019." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Lecture Notes in Computer Science*. Ed. by Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, et al. Berlin Heidelberg, Germany: Springer, 2019.

[118] Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, et al. "Overview of the CLEF eHealth evaluation lab 2019." In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2019, pp. 322–339.

[119] Phillip Keung, Yichao Lu, and Vikas Bhardwaj. "Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 1355–1360. DOI: 10.18653/v1/D19-1138. URL: https://aclanthology.org/D19-1138.

[120] Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. "Don't Use English Dev: On the Zero-Shot Cross-Lingual Evaluation of Contextual Embeddings." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 549–554. DOI: 10.18653/v1/2020.emnlp-main.40. URL: https://aclanthology.org/2020.emnlp-main.40.

[121] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C Rindflesch. "Constructing a semantic predication gold standard from the biomedical literature." In: *BMC bioinformatics* 12.1 (2011), pp. 1–17.

[122] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. "Broad-coverage biomedical relation extraction with SemRep." In: *BMC bioinformatics* 21 (2020), pp. 1–28.

[123] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. "Hadamard Product for Low-rank Bilinear Pooling." In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=r1rhWnZkg.

[124] Yoon Kim. "Convolutional Neural Networks for Sentence Classification." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1746–1751. DOI: 10.3115/v1/D14-1181. URL: https://aclanthology.org/D14-1181.

[125] Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. “New Transfer Learning Techniques for Disparate Label Sets.” In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015, pp. 473–482. DOI: 10.3115/v1/P15-1046. URL: https://aclanthology.org/P15-1046.

[126] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization.” In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[127] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization.” In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[128] Stanley Kok and Pedro M. Domingos. “Statistical predicate invention.” In: *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*. Ed. by Zoubin Ghahramani. Vol. 227. ACM International Conference Proceeding Series. ACM, 2007, pp. 433–440. DOI: 10.1145/1273496.1273551. URL: https://doi.org/10.1145/1273496.1273551.

[129] Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, and Ander Intxaurrondo. “Overview of the BioCreative VI chemical-protein interaction Track.” In: *Proceedings of the sixth BioCreative challenge evaluation workshop*. Vol. 1. 2017, pp. 141–146.

[130] Jayant Krishnamurthy and Tom M. Mitchell. “Learning a Compositional Semantics for Freebase with an Open Predicate Vocabulary.” In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 257–270. DOI: 10.1162/tacl_a_00137. URL: https://aclanthology.org/Q15-1019.

[131] Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. “Canonical Tensor Decomposition for Knowledge Base Completion.” In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2869–2878. URL: http://proceedings.mlr.press/v80/lacroix18a.html.

[132] John D Lafferty, Andrew McCallum, and Fernando CN Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.” In: *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, pp. 282–289.

[133] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. “Neural Architectures for Named Entity Recognition.” In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 260–270. DOI: 10.18653/v1/N16-1030. URL: https://aclanthology.org/N16-1030.

[134] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. “Transfer Learning for Named-Entity Recognition with Neural Networks.” In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. URL: https://aclanthology.org/L18-1708.

[135] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. “BioBERT: pre-trained biomedical language representation model for biomedical text mining.” In: *ArXiv preprint* abs/1901.08746 (2019). URL: https://arxiv.org/abs/1901.08746.

[136] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining.” In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[137] Adam Lerer, Ledell Wu, Jiajun Shen, Timothée Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. “Pytorch-BigGraph: A Large Scale Graph Embedding System.” In: *Proceedings of Machine Learning and Systems 2019, MLSys 2019, Stanford, CA, USA, March 31 - April 2, 2019*. Ed. by Ameet Talwalkar, Virginia Smith, and Matei Zaharia. mlsys.org, 2019. URL: https://proceedings.mlsys.org/book/282.pdf.

[138] Guoliang Li, Dong Deng, and Jianhua Feng. “Faerie: efficient filtering algorithms for approximate dictionary-based entity extraction.” In: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011*. Ed. by Timos K. Sellis, Renée J. Miller, Anastasios Kementsietsidis, and Yannis Velegrakis. ACM, 2011, pp. 529–540. DOI: 10.1145/1989323.1989379. URL: https://doi.org/10.1145/1989323.1989379.

[139] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. “BioCreative V CDR task corpus: a resource for chemical disease relation extraction.” In: *Database* 2016 (2016).

[140] Bill Yuchen Lin and Wei Lu. “Neural Adaptation Layers for Cross-domain Named Entity Recognition.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2012–2022. DOI: 10.18653/v1/D18-1226. URL: https://aclanthology.org/D18-1226.

[141] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. “Focal Loss for Dense Object Detection.” In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324. URL: https://doi.org/10.1109/ICCV.2017.324.

[142] Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. "Modeling Relation Paths for Representation Learning of Knowledge Bases." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 705–714. DOI: 10.18653/v1/D15-1082. URL: https://aclanthology.org/D15-1082.

[143] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. "Neural Relation Extraction with Selective Attention over Instances." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 2124–2133. DOI: 10.18653/v1/P16-1200. URL: https://aclanthology.org/P16-1200.

[144] Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. "A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 799–809. DOI: 10.18653/v1/P18-1074. URL: https://aclanthology.org/P18-1074.

[145] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. "Anonymisation Models for Text Data: State of the art, Challenges and Future Directions." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 4188–4203. DOI: 10.18653/v1/2021.acl-long.323. URL: https://aclanthology.org/2021.acl-long.323.

[146] ChunYang Liu, WenBo Sun, WenHan Chao, and Wanxiang Che. "Convolution neural network for relation extraction." In: *International Conference on Advanced Data Mining and Applications*. Springer. 2013, pp. 231–242.

[147] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. "Self-Alignment Pretraining for Biomedical Entity Representations." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 4228–4238. DOI: 10.18653/v1/2021.naacl-main.334. URL: https://aclanthology.org/2021.naacl-main.334.

[148] Hanxiao Liu, Yuexin Wu, and Yiming Yang. "Analogical Inference for Multi-relational Embeddings." In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 2168–2178. URL: http://proceedings.mlr.press/v70/liu17d.html.

[149] Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. "Challenges in Generalization in Open Domain Question Answering." In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 2014–2029. DOI: 10.18653/v1/2022.findings-naacl.155. URL: https://aclanthology.org/2022.findings-naacl.155.

[150] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." In: *ArXiv preprint* abs/1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[151] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. "De-identification of clinical notes via recurrent neural network and conditional random field." In: *Journal of biomedical informatics* 75 (2017), S34–S42.

[152] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. "Efficient Low-rank Multimodal Fusion With Modality-Specific Factors." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2247–2256. DOI: 10.18653/v1/P18-1209. URL: https://aclanthology.org/P18-1209.

[153] Edward Loper and Steven Bird. "NLTK: The Natural Language Toolkit." In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 63–70. DOI: 10.3115/1118108.1118117. URL: https://aclanthology.org/W02-0109.

[154] Daniel Loureiro and Alípio Mário Jorge. "Medlinker: Medical entity linking with neural representations and dictionary matching." In: *European Conference on Information Retrieval*. Springer. 2020, pp. 230–237.

[155] Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. "Learning with Noise: Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrix." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 430–439. DOI: 10.18653/v1/P17-1040. URL: https://aclanthology.org/P17-1040.

[156] Yuan Luo, Özlem Uzuner, and Peter Szolovits. "Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations." In: *Briefings in bioinformatics* 18.1 (2017), pp. 160–178.

[157] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. "YAGO3: A Knowledge Base from Multilingual Wikipedias." In: *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. www.cidrdb.org, 2015. URL: http://cidrdb.org/cidr2015/Papers/CIDR15_Paper1.pdf.

[158] S. Marchesin and G. Silvello. "TBGA: a large-scale Gene-Disease Association dataset for Biomedical Relation Extraction." In: *BMC Bioinformatics* 23.1 (2022), p. 111. DOI: 10.1186/s12859-022-04646-6. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04646-6.

[159] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. “Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results.” In: *IberLEF@ SEPLN*. 2019, pp. 618–638.

[160] Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. “Automatic de-identification of textual documents in the electronic health record: a review of recent research.” In: *BMC medical research methodology* 10.1 (2010), pp. 1–16.

[161] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. “UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus.” In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 1744–1753. DOI: 10.18653/v1/2021.naacl-main.139. URL: https://aclanthology.org/2021.naacl-main.139.

[162] Pauli Miettinen. “Boolean tensor factorizations.” In: *2011 IEEE 11th International Conference on Data Mining*. IEEE. 2011, pp. 447–456.

[163] Rada Mihalcea and Paul Tarau. “TextRank: Bringing Order into Text.” In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 404–411. URL: https://aclanthology.org/W04-3252.

[164] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. “Distributed Representations of Words and Phrases and their Compositionality.” In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger. 2013, pp. 3111–3119. URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[165] George A. Miller. “WordNet: A Lexical Database for English.” In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. 1992. URL: https://aclanthology.org/H92-1116.

[166] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. “Distant supervision for relation extraction without labeled data.” In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 1003–1011. URL: https://aclanthology.org/P09-1113.

[167] Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. “Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results.” In: *IberLEF@ SEPLN* (2020), pp. 303–323.

[168] Sunil Mohan and Donghui Li. “MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts.” In: *Automated Knowledge Base Construction (AKBC)*. 2018.

[169] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. “ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing.” In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 319–327. DOI: 10.18653/v1/W19-5034. URL: https://aclanthology.org/W19-5034.

[170] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. “Clinical natural language processing in languages other than english: opportunities and challenges.” In: *Journal of biomedical semantics* 9.1 (2018), pp. 1–13.

[171] Mariana L Neves, Daniel Butzke, Antje Dörendahl, Nora Leich, Benedikt Hummel, Gilbert Schönfelder, and Barbara Grune. “Overview of the CLEF eHealth 2019 Multilingual Information Extraction.” In: *CLEF (Working Notes)*. 2019.

[172] Mariana Neves, Daniel Butzke, Antje Dörendahl, Nora Leich, Benedikt Hummel, Gilbert Schönfelder, and Barbara Grune. “Overview of the CLEF eHealth 2019 Multilingual Information Extraction.” In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Lecture Notes in Computer Science*. Ed. by Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, et al. Berlin Heidelberg, Germany: Springer, 2019.

[173] Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. “STransE: a novel embedding model of entities and relationships in knowledge bases.” In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 460–466. DOI: 10.18653/v1/N16-1054. URL: https://aclanthology.org/N16-1054.

[174] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. “A Review of Relational Machine Learning for Knowledge Graphs.” In: *Proc. IEEE* 104.1 (2016), pp. 11–33.

[175] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. “Holographic Embeddings of Knowledge Graphs.” In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, 2016, pp. 1955–1961. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12484.

[176] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. "A Three-Way Model for Collective Learning on Multi-Relational Data." In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. Ed. by Lise Getoor and Tobias Scheffer. Omnipress, 2011, pp. 809–816. URL: https://icml.cc/2011/papers/438_icmlpaper.pdf.

[177] Naoaki Okazaki and Jun'ichi Tsujii. "Simple and Efficient Algorithm for Approximate Dictionary Matching." In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, 2010, pp. 851–859. URL: https://aclanthology.org/C10-1096.

[178] World Health Organization. *International statistical classification of diseases and related health problems*. Vol. 1. World Health Organization, 2004.

[179] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[180] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. "Cross-lingual Name Tagging and Linking for 282 Languages." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1946–1958. DOI: 10.18653/v1/P17-1178. URL: https://aclanthology.org/P17-1178.

[181] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 8024–8035. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[182] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python." In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[183] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. "Learning from Context or Names? An Empirical Study on Neural Relation Extraction." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 3661–3672. DOI: 10.18653/v1/2020.emnlp-main.298. URL: https://aclanthology.org/2020.emnlp-main.298.

[184] Nanyun Peng and Mark Dredze. "Multi-task Domain Adaptation for Sequence Tagging." In: *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 91–100. DOI: 10.18653/v1/W17-2612. URL: https://aclanthology.org/W17-2612.

[185] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. "Moment Matching for Multi-Source Domain Adaptation." In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 1406–1415. DOI: 10.1109/ICCV.2019.00149. URL: https://doi.org/10.1109/ICCV.2019.00149.

[186] Yifan Peng, Chih-Hsuan Wei, and Zhiyong Lu. "Improving chemical disease relation extraction with rich features and weakly labeled data." In: *Journal of cheminformatics* 8.1 (2016), pp. 1–12.

[187] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets." In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 58–65. DOI: 10.18653/v1/W19-5006. URL: https://aclanthology.org/W19-5006.

[188] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

[189] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. "Regularizing neural networks by penalizing confident output distributions." In: *ArXiv preprint* abs/1701.06548 (2017). URL: https://arxiv.org/abs/1701.06548.

[190] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

[191] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 2463–2473. DOI: 10.18653/v1/D19-1250. URL: https://aclanthology.org/D19-1250.

[192] Ninh Pham and Rasmus Pagh. "Fast and scalable polynomial kernels via explicit feature maps." In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. Ed. by Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy. ACM, 2013, pp. 239–247. DOI: 10.1145/2487575.2487591. URL: https://doi.org/10.1145/2487575.2487591.

[193] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. "The DisGeNET knowledge platform for disease genomics: 2019 update." In: *Nucleic acids research* 48.D1 (2020), pp. D845–D855.

[194] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

[195] S Povey, R Lovering, E Bruford, M Wright, M Lush, and H Wain. "The HUGO Gene Nomenclature Committee (HGNC)." In: *Hum Genet* 109.6 (2001), pp. 678–680. DOI: 10.1007/s00439-001-0615-0. URL: http://www.ncbi.nlm.nih.gov/pubmed/11810281.

[196] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained models for natural language processing: A survey." In: *Science China Technological Sciences* 63.10 (2020), pp. 1872–1897.

[197] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. *Improving language understanding with unsupervised learning*. Tech. rep. Technical report, OpenAI, 2018.

[198] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." In: *OpenAI blog* 1.8 (2019), p. 9.

[199] Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. "Few-Shot Question Answering by Pretraining Span Selection." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 3066–3079. DOI: 10.18653/v1/2021.acl-long.239. URL: https://aclanthology.org/2021.acl-long.239.

[200] Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. "Domain Divergences: A Survey and Empirical Analysis." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 1830–1849. DOI: 10.18653/v1/2021.naacl-main.147. URL: https://aclanthology.org/2021.naacl-main.147.

[201] Alan Ramponi and Barbara Plank. "Neural Unsupervised Domain Adaptation in NLP—A Survey." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 6838–6855. DOI: 10.18653/v1/2020.coling-main.603. URL: https://aclanthology.org/2020.coling-main.603.

[202] Lev Ratinov and Dan Roth. "Design Challenges and Misconceptions in Named Entity Recognition." In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 147–155. URL: https://aclanthology.org/W09-1119.

[203] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. "YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames." In: *International semantic web conference*. Springer. 2016, pp. 177–185.

[204] Protection Regulation. "Regulation (EU) 2016/679 of the European Parliament and of the Council." In: *Regulation (eu)* 679 (2016), p. 2016.

[205] Marek Rei. "Semi-supervised Multitask Learning for Sequence Labeling." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 2121–2130. DOI: 10.18653/v1/P17-1194. URL: https://aclanthology.org/P17-1194.

[206] Nils Reimers and Iryna Gurevych. "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 338–348. DOI: 10.18653/v1/D17-1035. URL: https://aclanthology.org/D17-1035.

[207] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

[208] Sebastian Riedel, Limin Yao, and Andrew McCallum. "Modeling relations and their mentions without labeled text." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2010, pp. 148–163.

[209] Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. "Modeling Missing Data in Distant Supervision for Information Extraction." In: *Transactions of the Association for Computational Linguistics* 1 (2013), pp. 367–378. DOI: 10.1162/tacl_a_00234. URL: https://aclanthology.org/Q13-1030.

[210] Roland Roller and Mark Stevenson. "Self-supervised relation extraction using UMLS." In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2014, pp. 116–127.

[211] Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. "Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 3702–3710. DOI: 10.18653/v1/2020.emnlp-main.302. URL: https://aclanthology.org/2020.emnlp-main.302.

[212] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. "You CAN Teach an Old Dog New Tricks! On Training Knowledge Graph Embeddings." In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: https://openreview.net/forum?id=BkxSmlBFvr.

[213] Camilo Ruiz, Marinka Zitnik, and Jure Leskovec. "Identification of disease treatment mechanisms through the multiscale interactome." In: *Nature communications* 12.1 (2021), pp. 1–15.

[214] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. "Semi-Supervised Domain Adaptation via Minimax Entropy." In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 8049–8057. DOI: 10.1109/ICCV.2019.00814. URL: https://doi.org/10.1109/ICCV.2019.00814.

[215] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. "Maximum Classifier Discrepancy for Unsupervised Domain Adaptation." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 3723–3732. DOI: 10.1109/CVPR.2018.00392. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Saito_Maximum_Classifier_Discrepancy_CVPR_2018_paper.html.

[216] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 507–513.

[217] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. "Modeling relational data with graph convolutional networks." In: *European Semantic Web Conference*. Springer. 2018, pp. 593–607.

[218] Stefan Schweter and Alan Akbik. "FLERT: Document-Level Features for Named Entity Recognition." In: *ArXiv preprint* abs/2011.06993 (2020). URL: https://arxiv.org/abs/2011.06993.

[219] Isabel Segura-Bedmar, Paloma Martínez Fernández, and Daniel Sánchez Cisneros. "The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts." In: (2011).

[220] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, 2013, pp. 341–350. URL: https://aclanthology.org/S13-2056.

[221] Jurica Ševa, Mario Sänger, and Ulf Leser. "WBI at CLEF eHealth 2018 Task 1: Language-independent ICD-10 coding using multi-lingual embeddings and recurrent neural networks." In: CLEF. 2018.

[222] Yelong Shen, Jianshu Chen, Po-Sen Huang, Yuqing Guo, and Jianfeng Gao. "M-Walk: Learning to Walk over Graphs using Monte Carlo Tree Search." In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett. 2018, pp. 6787–6798. URL: https://proceedings.neurips.cc/paper/2018/hash/c6f798b844366ccd65d99bc7f31e0e02-Abstract.html.

[223] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis." In: *IEEE journal of biomedical and health informatics* 22.5 (2017), pp. 1589–1604.

[224] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis." In: *IEEE journal of biomedical and health informatics* 22.5 (2018), pp. 1589–1604.

[225] Carlos N Silla and Alex A Freitas. "A survey of hierarchical classification across different application domains." In: *Data Mining and Knowledge Discovery* 22.1-2 (2011), pp. 31–72.

[226] Luca Soldaini and Nazli Goharian. "Quickumls: a fast, unsupervised approach for medical concept extraction." In: *MedIR workshop, sigir*. 2016, pp. 1–4.

[227] Amber Stubbs, Michele Filannino, and Özlem Uzuner. "De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1." In: *Journal of biomedical informatics* 75 (2017), S4–S18.

[228] Amber Stubbs and Özlem Uzuner. "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus." In: *Journal of biomedical informatics* 58 (2015), S20–S29.

[229] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. "ERNIE: Enhanced representation through knowledge integration." In: *ArXiv preprint* abs/1904.09223 (2019). URL: https://arxiv.org/abs/1904.09223.

[230] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. "RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space." In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=HkgEQnRqYQ.

[231] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. "Multi-instance Multi-label Learning for Relation Extraction." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 455–465. URL: https://aclanthology.org/D12-1042.

[232] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. "Rethinking the Inception Architecture for Computer Vision." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308. URL: https://doi.org/10.1109/CVPR.2016.308.

[233] Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. "Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 10438–10449. DOI: 10.18653/v1/2021.emnlp-main.816. URL: https://aclanthology.org/2021.emnlp-main.816.

[234] Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. "Reducing Wrong Labels in Distant Supervision for Relation Extraction." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 721–729. URL: https://aclanthology.org/P12-1076.

[235] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 4593–4601. DOI: 10.18653/v1/P19-1452. URL: https://aclanthology.org/P19-1452.

[236] Erik F. Tjong Kim Sang. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition." In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002. URL: https://aclanthology.org/W02-2024.

[237] Erik F. Tjong Kim Sang. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition." In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002. URL: https://aclanthology.org/W02-2024.

[238] Erik F. Tjong Kim Sang. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition." In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002. URL: https://aclanthology.org/W02-2024.

[239] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. "Representing Text for Joint Embedding of Text and Knowledge Bases." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1499–1509. DOI: 10.18653/v1/D15-1174. URL: https://aclanthology.org/D15-1174.

[240] Théo Trouillon, Christopher R Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. "Knowledge graph completion via complex tensor factorization." In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 4735–4772.

[241] Théo Trouillon and Maximilian Nickel. "Complex and holographic embeddings of knowledge graphs: a comparison." In: *ArXiv preprint* abs/1707.01475 (2017). URL: https://arxiv.org/abs/1707.01475.

[242] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. "Complex Embeddings for Simple Link Prediction." In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 2071–2080. URL: http://proceedings.mlr.press/v48/trouillon16.html.

[243] Ledyard R Tucker. "Some mathematical notes on three-mode factor analysis." In: *Psychometrika* 31.3 (1966), pp. 279–311.

[244] Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. "The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships." In: *Journal of biomedical informatics* 45.5 (2012), pp. 879–884.

[245] Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P Rosé. "Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets." In: *Journal of Biomedical Informatics* 121 (2021), p. 103880.

[246] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All you Need." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[247] Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H Nilsson. "Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial." In: *International journal of medical informatics* 78.12 (2009), e19–e26.

[248] Brian Walsh, Sameh K. Mohamed, and Vít Novácek. "BioKG: A Knowledge Graph for Relational Learning On Biological Data." In: *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. Ed. by Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux. ACM, 2020, pp. 3173–3180. DOI: 10.1145/3340531.3412776. URL: https://doi.org/10.1145/3340531.3412776.

[249] Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro. "Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 8476–8488. DOI: 10.18653/v1/2020.acl-main.750. URL: https://aclanthology.org/2020.acl-main.750.

[250] Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. "Pyramid: A Layered Model for Nested Named Entity Recognition." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 5918–5928. DOI: 10.18653/v1/2020.acl-main.525. URL: https://aclanthology.org/2020.acl-main.525.

[251] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. "Revisiting multiple instance neural networks." In: *Pattern Recognition* 74 (2018), pp. 15–24.

[252] Yanjie Wang, Rainer Gemulla, and Hui Li. "On Multi-Relational Link Prediction With Bilinear Models." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 4227–4234. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16900.

[253] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. "Knowledge Graph Embedding by Translating on Hyperplanes." In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*. Ed. by Carla E. Brodley and Peter Stone. AAAI Press, 2014, pp. 1112–1119. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531.

[254] Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. "Label-Aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1–15. DOI: 10.18653/v1/N18-1001. URL: https://aclanthology.org/N18-1001.

[255] Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. "Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework." In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: https://openreview.net/forum?id=S1l-C0NtwS.

[256] Gerhard Weikum, Xin Luna Dong, Simon Razniewski, Fabian Suchanek, et al. "Machine knowledge: Creation and curation of comprehensive knowledge bases." In: *Foundations and Trends® in Databases* 10.2-4 (2021), pp. 108–490.

[257] Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. "Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013, pp. 1366–1371. URL: https://aclanthology.org/D13-1136.

[258] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. "DrugBank 5.0: a major update to the DrugBank database for 2018." In: *Nucleic acids research* 46.D1 (2018), pp. D1074–D1082.

[259] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

[260] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. "Scalable Zero-shot Entity Linking with Dense Entity Retrieval." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 6397–6407. DOI: 10.18653/v1/2020.emnlp-main.519. URL: https://aclanthology.org/2020.emnlp-main.519.

[261] Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jianguang Lou. "UniTrans : Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by Christian Bessiere. ijcai.org, 2020, pp. 3926–3932. DOI: 10.24963/ijcai.2020/543. URL: https://doi.org/10.24963/ijcai.2020/543.

[262] Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. "Enhanced meta-learning for cross-lingual named entity recognition with minimal resources." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 9274–9281.

[263] Shanchan Wu and Yifan He. "Enriching Pre-trained Language Model with Entity Information for Relation Classification." In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. Ed. by Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu. ACM, 2019, pp. 2361–2364. DOI: 10.1145/3357384.3358119. URL: https://doi.org/10.1145/3357384.3358119.

[264] Shijie Wu and Mark Dredze. "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077.

[265] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." In: *ArXiv preprint* abs/1609.08144 (2016). URL: https://arxiv.org/abs/1609.08144.

[266] Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. "Neural Cross-Lingual Named Entity Recognition with Minimal Resources." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 369–379. DOI: 10.18653/v1/D18-1034. URL: https://aclanthology.org/D18-1034.

[267] R. Xing, J. Luo, and T. Song. "BioRel: towards large-scale biomedical relation extraction." In: *BMC Bioinformatics* 21-S.16 (2020), p. 543. DOI: 10.1186/s12859-020-03889-5. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03889-5.

[268] Rong Xu and QuanQiu Wang. "Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature." In: *Journal of biomedical informatics* 51 (2014), pp. 191–199.

[269] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. "Embedding Entities and Relations for Learning and Inference in Knowledge Bases." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6575.

[270] Fan Yang, Zhilin Yang, and William W. Cohen. "Differentiable Learning of Logical Rules for Knowledge Base Reasoning." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 2319–2328. URL: https://proceedings.neurips.cc/paper/2017/hash/0e55666a4ad822e0e34299df3591d979-Abstract.html.

[271] Jie Yang, Shuailong Liang, and Yue Zhang. "Design Challenges and Misconceptions in Neural Sequence Labeling." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 3879–3889. URL: https://aclanthology.org/C18-1327.

[272] Jie Yang and Yue Zhang. "NCRF++: An Open-source Neural Sequence Labeling Toolkit." In: *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 74–79. DOI: 10.18653/v1/P18-4013. URL: https://aclanthology.org/P18-4013.

[273] Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. "A study of deep learning methods for de-identification of clinical notes in cross-institute settings." In: *BMC Medical Informatics and Decision Making* 19.5 (2019), pp. 1–9.

[274] Yi Yang and Arzoo Katiyar. "Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 6365–6375. DOI: 10.18653/v1/2020.emnlp-main.516. URL: https://aclanthology.org/2020.emnlp-main.516.

[275] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. "Hierarchical Attention Networks for Document Classification." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 1480–1489. DOI: 10.18653/v1/N16-1174. URL: https://aclanthology.org/N16-1174.

[276] Liang Yao, Chengsheng Mao, and Yuan Luo. "KG-BERT: BERT for knowledge graph completion." In: *ArXiv preprint* abs/1909.03193 (2019). URL: https://arxiv.org/abs/1909.03193.

[277] Andrew Yates and Nazli Goharian. "ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites." In: *European Conference on Information Retrieval*. Springer. 2013, pp. 816–819.

[278] Vithya Yogarajan, Michael Mayo, and Bernhard Pfahringer. "A survey of automatic de-identification of longitudinal clinical narratives." In: *ArXiv preprint* abs/1810.06765 (2018). URL: https://arxiv.org/abs/1810.06765.

[279] Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. "Learning and evaluating general linguistic intelligence." In: *ArXiv preprint* abs/1901.11373 (2019). URL: https://arxiv.org/abs/1901.11373.

[280] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering." In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 1839–1848. DOI: 10.1109/ICCV.2017.202. URL: https://doi.org/10.1109/ICCV.2017.202.

[281] Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. "Improving Biomedical Pretrained Language Models with Knowledge." In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, 2021, pp. 180–190. DOI: 10.18653/v1/2021.bionlp-1.20. URL: https://aclanthology.org/2021.bionlp-1.20.

[282] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1753–1762. DOI: 10.18653/v1/D15-1203. URL: https://aclanthology.org/D15-1203.

[283] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. "Relation Classification via Convolutional Deep Neural Network." In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014, pp. 2335–2344. URL: https://aclanthology.org/C14-1220.

[284] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. "Character-level Convolutional Networks for Text Classification." In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett. 2015, pp. 649–657. URL: https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

[285] Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. "DGL-KE: Training Knowledge Graph Embeddings at Scale." In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, 739–748.

[286] Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. "Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3461–3471. DOI: 10.18653/v1/P19-1336. URL: https://aclanthology.org/P19-1336.

[287] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. "Domain Adaptive Ensemble Learning." In: *ArXiv preprint* abs/2003.07325 (2020). URL: https://arxiv.org/abs/2003.07325.

[288] Xiaohua Zhou, Xiaodan Zhang, and Xiaohua Hu. "MaxMatcher: Biological concept extraction using approximate dictionary lookup." In: *Pacific RIM international conference on artificial intelligence*. Springer. 2006, pp. 1145–1149.