EPFL

Thèse n° 7707

Approaching Ontology Alignment through Representation Learning to Bridge the Semantic Gap in Engineering Applications

Présentée le 10 juin 2020

à la Faculté des sciences et techniques de l'ingénieur
Groupe SCI STI DK
Programme doctoral en robotique, contrôle et systèmes intelligents

pour l'obtention du grade de Docteur ès Sciences

par

Prodromos KOLYVAKIS

Acceptée sur proposition du jury

Prof. K. Aminian, président du jury
Prof. D. Kyritsis, directeur de thèse
Dr C. Bekas, rapporteur
Prof. D. Rebholz-Schuhmann, rapporteur
Prof. R. West, rapporteur

École polytechnique fédérale de Lausanne

2020

Ἀριστοτέλης ἐρωτηθεὶς ποῦ
κατοικοῦσιν αἱ μοῦσαι, ἔφη:
“ἐν ταῖς ψυχαῖς τῶν φιλοπόνων”

To my family…

Acknowledgements

The work presented in this thesis constitutes the outcome of a long journey that goes beyond the pages of this manuscript. During this journey, I was fortunate enough to receive endless support and feedback from numerous people. I deeply acknowledge that without them, nothing would be the same. The least thing I can do for them is to open this thesis by thanking them.

First and foremost, I would like to thank my advisor, Prof. Dimitrios Kyritsis, for his guidance and advice. Dimitris has been an admirable supervisor, creative, honest and supportive. His flexibility and kind personality were two important ingredients for making the ICT4SM group such as a warm and pleasant working environment. I was really fortunate enough to have Dimitris as my thesis advisor and I deeply appreciate his belief in me and in my research.

I would also like to thank the members of my thesis committee for putting in the efforts to evaluate the merit of my research work as well as to provide insightful and constructive comments and feedback that shaped this thesis. Specifically, I would like to thank Dr. Costas Bekas, Prof. Dietrich Rebholz-Schuhmann and Prof. Robert West, who acted as my thesis examiners, as well as Prof. Kamiar Aminian, who accepted to act as the thesis jury president.

In the beginning of my PhD journey, I had the unique chance to meet and later work with an amazing person, i.e., Prof. Alexandros Kalousis. From the first time, I understood that apart from a great researcher, Alexandros was a person of high quality and excellent morals. Alexandre, I feel inclined to thank you for everything. In addition, I had the fortune and honor to work with Prof. Barry Smith, whose own work was highly influential in shaping my understanding of ontologies. His lectures, work and our personal communication introduced me to the world of philosophical thinking and made me understand the invaluable importance of the *Ph* in the PhD title. Moreover, I would like to thank my internship supervisor Dr. William Campbell as well as my mentors Dr. Eunah Cho and Stan Peshterliev for making my internship at Amazon Alexa such an unforgettable experience. I would also like to share my gratitude to all the wonderful people that I met thanks to my involvement in the bIoTope research project.

At this point, I would like to share my deepest gratitude to all the people who made Lausanne feel like home. First of all, I would like to thank all the people affiliated or shared connections with the group, i.e., Damiano, Sangje, Gökan, Giannis, Foivos, Christos, Apostolos, Marlène, PA, Sylvia, Corinne and many many more, for everything. Life at the lab would not be the same

without them. I would also like to thank Aggelos, Alexandros, Angeliki (aka Lydia), Costas, Damaskinaki, Dimitris, Eleni, Ergys, Ivi, Lefteris, Manos, Marios, Matt, Mimi, Nathalie, Panagiotis, Panos, Pavlos, Raquel, Sotiris, Stelios, Stella, Sylvie, Thanos, Vaggos and Vivi for being the important ingredient that makes the place you live a lovely place, i.e., friends. Marios and Thanos, who have been close to me even from the first days at NTUA, and are great persons to discuss engineering/scientific stuff and have a great time. Panagioti and Stella, I honestly wish I even knew you during my undergraduate studies; you are unique and amazing persons. Vaggo, I cannot describe you in words; sharing an apartment with you was a once-in-a-lifetime experience. Finally, I want to deeply thank Angelina for her true, honest and endless support all these years as well as for helping me to see the finishing line at the toughest times.

I'm also extremely grateful to all my Greek friends for proving that true friendships can overcome the physical distance barrier and for their continuous support in my endeavour. I want to thank all of of these cherished friends: My childhood friends, aka the "primary school gang". Especially, my friends Thodori, Dimitri, Sotiria, Vasiliki, Mariza, Dimitra, Natalia, Michali, Giorgo and Konstantino. The "kampos gang", i.e., Chabos, Konstantinos, Philippos, Ntinos and Kostas. My friends Evi and Philippo (the Buzz Lightyear). My friends Ioanna, Tzeni, Marilena, Eirini, Afroditi, Vivi, Christina, Alexandra, Vivian and Ilya. My amazing university friends Panos, Spyros, Miltos, Andreas, Tasos, Lia, Theophili, Maria, Anna, Foufoutos and Fotini.

Last but not least, I want to thank my family for their endless love, support, and encouragement during my whole life. My parents, Vasilis and Eleni, for their endless endeavour to provide us with more than I could ever ask for, for believing in me and for teaching us what is important in life and to have faith in human endeavour. My beloved sister who has been always there for me. My grandparents for their unconditional love and for the invaluable life lessons they taught me. Finally, my dearest uncles Michali and Vasili who, although they passed away, will always live inside me; your love for education was highly contagious and I have caught it ...

This research has been supported by grants from the Swiss State Secretariat for Education, Research and Innovation SERI (SERI; contract number 15.0303) through the European Union Horizon 2020 research and innovation programme (grant agreement No 688203), the EU Horizon 2020 research and innovation programme under grant agreement No 723906, and the EU Horizon 2020 research and innovation programme under grant agreement No 825030.

Lausanne, April 12, 2020 P. K.

Abstract

The current information landscape is characterised by a vast amount of relatively semantically homogeneous, when observed in isolation, data silos that are, however, drastically semantically fragmented when considered as a whole. Within each data silo, information can be harvested without the risk of misinterpretation due to conforming to the same ontology that formally defines the types and relations in the application domain. Nonetheless, when data are retrieved from multiple and heterogeneous data silos, special consideration is required to ensure a common and uniform interpretation. Establishing semantic bridges across semantically heterogeneous data silos, i.e., align the corresponding ontologies, becomes, thus, crucial. At the same time, there is an exponential increase in the number of data as well as in the number of heterogeneous data silos. It becomes apparent that the exponentially increasing information landscape prohibits manual curation strategies and illustrates the importance of an automatic computational approach that relies less on human expertise and intervention.

The focal point of this thesis is to build semantic bridges across heterogeneous data silos. We first focus on discovering equivalence relations between entities appearing in the different ontologies used by heterogeneous data silos. To decrease the required human expertise and intervention, we propose to approach the problem of ontology alignment from a representation learning perspective. We demonstrate that by exploiting transfer learning we can overcome the main obstacles, i.e., the small sample size and the serious class imbalance problem, that have been shown to hinder the application of machine learning to the problem. More precisely, our approach is based on embedding ontological terms in a high-dimensional Euclidean space. These terminological embeddings are automatically learned so as they are implicitly tailored to the task of ontology alignment. We compare our proposed methods to state-of-the-art systems based on feature engineering using a plethora of evaluation benchmarks. We present significant performance improvements and we demonstrate the advantages that representation learning brings to the problem of ontology alignment.

Subsequently, we focus on discovering general relations, i.e., not particularly restricted to equivalence relations, existing between entities appearing in the same ontology or knowledge base. This problem is known under the terms knowledge base completion and link prediction. Building on recent research highlighting the advantages of non-Euclidean space, we examine the contribution of geometrical space to the task of knowledge base completion. We focus on the family of translational models that, despite showing a lagging performance on certain

datasets, offer certain advantages with regard to the rules they can effectively represent. We extend these models to the hyperbolic space so as to better reflect the topological properties of knowledge bases. We empirically show, using a variety of link prediction datasets, that hyperbolic space allows to narrow down significantly the performance gap between translational and bilinear models; illustrating that the lagging performance of translational models is not an intrinsic characteristic of them. Another key outcome of this work is to demonstrate a new promising direction for developing models that, although not fully expressive, allow to better represent certain families of rules; opening up for more fine-grained reasoning tasks.

In summary, this thesis proposes new ways to approach the problems of ontology alignment and link prediction in the setting of representation learning. It advances beyond the state-of-the-art methods in a multitude of different ways. It also serves to strengthen our understanding of the role of geometrical space for relation prediction and to illustrate prominent directions for performing more fine-grained reasoning tasks in the embedding space.

Keywords: ontology matching, ontology alignment, sentence embeddings, word embeddings, terminological embeddings, semantic similarity, denoising autoencoder, outlier detection, knowledge base completion, link prediction, knowledge base, ontology, knowledge graph embeddings, hyperbolic embeddings, quasi-chained rules, Poincaré-ball model

Résumé

Le paysage actuel de l'information se caractérise par une grande quantité de silos de données relativement homogènes sémantiquement, lorsqu'ils sont observés de façon isolée, mais qui sont cependant radicalement fragmentés sémantiquement lorsqu'ils sont considérés dans leur ensemble. Au sein de chaque silo de données, les informations peuvent être collectées sans risque d'erreur d'interprétation en raison de leur conformité à une même ontologie qui définit formellement les types et les relations dans le domaine d'application. Néanmoins, lorsque les données sont extraites de silos de données multiples et hétérogènes, une attention particulière est requise pour garantir une interprétation commune et uniforme. L'établissement de ponts sémantiques entre des silos de données sémantiquement hétérogènes, c'est-à-dire l'alignement des ontologies correspondantes, devient donc crucial. Parallèlement, le nombre de données ainsi que le nombre de silos de données hétérogènes augmente exponentiellement. Il apparaît clairement que les stratégies de conservation manuelle ne sont pas adaptées à ce paysage de l'information en croissance exponentielle, et ceci souligne l'importance d'une approche informatique et automatique qui repose moins sur l'expertise et l'intervention humaines.

Le point central de cette thèse est de construire des ponts sémantiques à travers des silos de données hétérogènes. Nous nous concentrons d'abord sur la découverte des relations d'équivalence pouvant exister entre les entités apparaissant dans les différentes ontologies utilisées par des silos de données hétérogènes. Pour diminuer l'expertise et l'intervention humaines requises, nous proposons d'aborder le problème de l'alignement des ontologies dans une perspective d'apprentissage de la représentation. Nous démontrons qu'en exploitant l'apprentissage par transfert, nous pouvons surmonter les principaux obstacles, c'est-à-dire la petite taille de l'échantillon et le grave problème de déséquilibre de classe, qui entravent l'application de l'apprentissage automatique au problème. Plus précisément, notre approche est basée sur le plongement de termes ontologiques dans un espace euclidien de grande dimension. Ces plongements de termes sont automatiquement appris de sorte à ce qu'ils soient implicitement adaptés à la tâche d'alignement des ontologies. Nous comparons nos méthodes proposées à des systèmes de pointe basés sur l'ingénierie des fonctionnalités en utilisant une pléthore de repères d'évaluation. Nous présentons des améliorations de performances significatives et nous démontrons les avantages que l'apprentissage de la représentation apporte au problème de l'alignement des ontologies.

Par la suite, nous nous concentrons sur la découverte de relations générales, c'est-à-dire non particulièrement limitées aux relations d'équivalence, existant entre des entités de la même ontologie ou base de connaissances. Ce problème est connu sous le terme de prévision de liens. En nous appuyant sur des recherches récentes mettant en évidence les avantages de l'espace non-euclidien, nous examinons la contribution de l'espace géométrique à la tâche de prévision de liens. Nous nous concentrons sur la famille des modèles translationnels qui, malgré des performances plus faibles sur certains ensembles de données, offrent certains avantages en ce qui concerne les axiomes qu'ils peuvent modéliser efficacement. Nous étendons ces modèles à l'espace hyperbolique afin de mieux refléter les propriétés topologiques des bases de connaissances. Nous montrons empiriquement, en utilisant une variété de jeux de données de prévision de liens, que l'espace hyperbolique permet de réduire significativement l'écart de performance entre les modèles translationnels et bilinéaires ; illustrant ainsi que le retard de performance des modèles translationnels n'en est pas une caractéristique intrinsèque. Un autre résultat clé de ce travail est de proposer une nouvelle direction prometteuse pour le développement de modèles qui, bien que peu expressifs, permettent de mieux représenter certaines familles de règles ; ouvrant ainsi la voie à des tâches de raisonnement plus fines.

En résumé, cette thèse propose de nouvelles façons d'aborder les problèmes d'alignement d'ontologies et de prévision de liens dans le cadre de l'apprentissage de la représentation. Elle va au-delà des méthodes de pointe d'une multitude de façons différentes. Elle sert également à renforcer notre compréhension du rôle de l'espace géométrique pour la prédiction des relations et à illustrer des directions importantes pour effectuer des tâches de raisonnement plus fines dans l'espace de plongement.

Mots clés : alignement de l'ontologie, plongement de phrases, plongement de mots, plongement de termes, similitude sémantique, auto-encodeurs débruiteurs, détection d'anomalies, prévision de liens, ontologie, base de connaissances, plongement de graphique des connaissances, plongement hyperbolique, axiomes quasi-chaînés, la boule de Poincaré

Contents

List of Figures

List of Tables

List of Acronyms

DAE	Denoising Autoencoder
DESM	Dual Embedding Space Model
DOID	Human Disease Ontology
DRs	Distributed Representations
FMA	Foundational Model of Anatomy
IoT	Internet of Things
KB	Knowledge Base
KBC	Knowledge Base Completion
MA	Adult Mouse Anatomical Dictionary
MeSH	Medical Subject Headings
MRR	Mean Reciprocal Rank
NCI	NCI Thesaurus
NLP	Natural Language Processing
OAEI	Ontology Alignment Evaluation Initiative
QC	Quasi-Chained
SGA	Stochastic Gradient Ascent
SGD	Stochastic Gradient Descent
SNOMED	SNOMED Clinical Terms
Uberon	Uber Anatomy Ontology

1 Introduction

"Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?"

J.R.R. Tolkien, The Hobbit, or There and Back Again

The Information Age has provided a fertile ground for an interconnected world of human beings and things without borders, but certainly not without its share of paradoxes. The quest for information has opened the door for the quest for a systematic way of transforming the information "cacophony into a symphony of meaning" [184] that will allow to harvest this wealth of data and draw inferences out of it. At the time of writing, Gartner estimates that 4.8 billion Internet of Things (IoT) endpoints are available [78], while the total volume of the accumulated digital data, measured in bytes, is claimed to exceed the total number of stars in the observable universe [76, 55]. Nonetheless, this abundance of information came at a cost that hinders both its exploration and exploitation; the sheer semantic inconsistency encountered in the various heterogeneous information sources [112, 197, 62, 225, 228].

Data are not only heterogeneous due to the high variety of data structures and representations they are expressed in, also known as *syntactic heterogeneity* [229, 113], but their heterogeneity also stems from the uncertainty in interpreting them leading to *semantic heterogeneity* [90, 114]. The current information landscape is characterised by a vast amount of relatively semantically homogeneous, when observed in isolation, data silos that are, however, drastically semantically fragmented when considered as a whole [225, 112]. To guarantee a consensual interpretation, the simultaneous transmission of the *meaning* of the various information fields appearing in data seems to be necessary, which, unfortunately, is not an easy task to do [182, 44, 184]. As the American linguist Ray Jackendoff framed it: "meaning is the 'holy grail' not only of linguistics, but also of philosophy, psychology, and neuroscience" [104, p. 288]. Nonetheless, although delivering the precise meaning constitutes a real scientific challenge, constraining the possible interpretations is yet feasible by introducing axioms that help to

avoid invalid entailments [199, 86, 7]. For example, if a *piston pump* is defined to be a subtype of a *pump* while it has been stated that no entity can be both a *liquid* and a *pump*, then it could be safely concluded that a *piston pump* is not a *liquid* [224]. Building on this inferential paradigm, *applied ontologies* aim to provide a rigorous and formal representation of commonsense reality by explicitly specifying the relations existing among entities and constraining the possible interpretations through careful and concise axiomatisation [198, 158, 85].

The pitfalls of semantic heterogeneity become more apparent when the data should be automatically exploited by intelligent artificial agents [188, 243, 225, 96]. For instance, a conversational agent asked to arrange the details of a trip would need to access various heterogeneous data sources to retrieve information about flight schedules, hotel availability, rental car offers, public transport schedules, weather forecast, etc. One of the major challenges that the agent would face is simply not to be *lost in translation*. To provide all the available rental car offers, for example, the agent should be able to find the bearers of knowledge required to infer whether the information fields *automobile* and *car* found in two distinct information sources could be considered as equivalent in the given context without provoking any logical inconsistencies. Similarly to the *pump* example, an ontology containing all the *domain-specific synonyms* of a specific term expressed through equivalence relations could be proven again sufficient for successfully performing this inference. This line of thinking, however, carries the unwarranted assumption of a perfect and complete lexicon of the existing synonymy information. Nevertheless, despite the extend of machine-interpretable factual knowledge available today, it is widely accepted that their coverage is still far from being complete [238].

As shown above, inferring that the terms *automobile* and *car* were synonymous under the given context allows to bypass the heterogeneity gap and fuse different information sources. Therefore, being able to automatically discover relations existing between entities constitutes a central challenge for harvesting the wealth of heterogeneous data available today. The key observation for transforming this computationally intractable problem at first sight into a problem that can be tackled computationally is one prevailing pattern appearing both in nature and language; that a *similar context* between entities is a sign of a possible existence of *common characteristics* between them. For instance, it has been observed in nature that a similar environmental context (e.g., available resources, climate, etc.) between non-coexisting species can provoke the appearance of similar structures [82, 192]. A similar phenomenon is observed in language where words that occur in similar context tend to have similar meaning (e.g., oculist and eye-doctor) [46, 93, 189]. This phenomenon, that became known as *distributional hypothesis* [93], paved the way for harnessing co-occurence statistics between words and phrases, extracted from a plurality of text corpora, as a way to quantify how *semantically* close two distinct words or phrases are [133]. Their *semantic similarity* is measured in terms of how similar is the distribution of the words that surrounds them.

Early attempts of exploiting the distributional hypothesis have focused on manually crafting similarity functions. One classical example is the *Pointwise Mutual Information* [66, p. 28][35] that quantifies the discrepancy between the probabilities of two words to co-occur and to

occur independently. Unfortunately discovering good similarity functions can be highly time consuming. Another line of research has focused on mapping words to high-dimensional vectors in Euclidean Space where the vectorial coordinates are created in such a way so that they correlate with words' occurrence statistics computed on text corpora. In this *bag of words* model [93], the similarity of two distinct words is measured by using more easily defined similarity functions in the Euclidean space (e.g., cosine similarity, distance-based similarity[1], etc.) [136, 92]. It should be noted that part of the difficulty in manually crafting similarity functions has now been transferred in devising good bag of words vectorial representations. Naturally, the bag of words vectorial representations are quite sparse – i.e., many of the vectorial coordinates have really small absolute values; allowing for efficient storage and manipulation. Interestingly, it has been shown that reducing their dimensionality is beneficial for various mainstream tasks [136]. For that reason, various approaches have been proposed in the literature [175, 102, 48] that attempt to reduce the vectors' dimensionality aiming at creating more *dense* word representations and demonstrate better semantics capture capability.

To overcome the burden of manually crafting good dense vectorial representations for words, there has been a recent rise of *distributed representations* (DRs) [18, 149, 150, 176, 179, 5, 50], in which for example words are embedded in a high-dimensional Euclidean space and the word vectorial representations are automatically *learned* in an unsupervised way. The way this works is that the machine learns a mapping from words to high-dimensional vectors which take account of the contexts in which words appear in a plurality of corpora. Vectors of words that appear in the same sorts of context will then be closer together when measured by a similarity function. As above, various similarity functions can be used such as the cosine similarity, the distance-based similarity, etc. That the approach can work without supervision stems from the fact that meaning capture is merely a positive externality of context identification, a task that is not directly related to the meaning discovery task. At the same time, unlike bag of words vectorial representations that neglect the syntactic information present in language, there is a powerful family of DRs that allows to create syntax-aware distributed representations [59, 60, 101, 227]. Suprisingly, DRs have demonstrated unprecedented improvements in various natural language processing tasks [215, 40, 144, 227, 179, 50].

With this in mind, DRs have the potential to bring significant value to the task of automatically discovering relations existing between entities, which can be described by different ontologies. For example, a *terminological embedding* can be assigned to every ontological term in a way that synonymous terms will have a high degree of similarity, as computed by a used similarity function. Additionally, a *relation embedding* and an *entity embedding* can be assigned to every relation and entity, respectively. These relations and entities could appear in different, or even the same, ontologies. These relation and entity embeddings could be constructed in such a way that if a certain relation exists between two entities, then the associated similarity function, that will now operate on three operands, will be expected to have a high value. Learning such embeddings would aid to go beyond the existing available knowledge and identify unknown general relations existing between entities – not particularly restricted to synonymy relations. In this thesis, we focus on devising such *representation learning* architectures that can detect

relations between entities towards bridging the semantic gap that is currently existing.

1.1 Thesis Goals

The primary goal of this thesis is to exploit recent advancements in representation learning to develop relation detection algorithms that discover relations between entities of potentially different ontologies. Our aim is to remove the burden of manually crafting good terminological, entity and relation representations and propose systems that offer substantial performance improvements. Towards this direction, we devise representation learning based methods that exploit semantic information extracted from external resources, including the ontologies themselves, as well as statistical regularities that lay in ontological and knowledge base facts.

We begin by investigating the primary reasons that hindered the application of machine learning to the problem of *ontology alignment*. We identify that two of the prime reasons are (i) the relatively small sample size, i.e., very few ontology alignment scenarios are available to guarantee generalisation, and (ii) the serious class imbalance problem, i.e., the number of true alignments between two ontologies is several orders of magnitude smaller than the number of all possible mappings hindering learning. We design one representation learning based ontology alignment system which by exploiting transfer learning bypasses the aforementioned problems. The system shows state-of-the-art performance demonstrating significant improvements with regard to the recall metric, nevertheless it undergoes a certain amount of degradation in the precision metric. To tackle this shortcoming, we propose an outlier detection mechanism that successfully detects misalignments without significantly harming the recall capability of the system. Finally, we explore different geometrical spaces for embedding the entities and relations existing in ontologies and knowledge bases that have the potential to better reflect their topological properties. To this end, we explore the hyperbolic space and we show that the right choice of geometrical space does not only impact the performance of embedding models for the task of *relation prediction*, but also opens up new opportunities for developing embedding models that allow better representing certain families of ontological axioms. Last but not least, our findings also shed light on understanding which ontologies and knowledge bases mostly benefit from the use of hyperbolic embeddings.

Thesis Statement:

Despite the overwhelm of data, the current information landscape is drastically semantically fragmented hindering both data exploration and exploitation. The exponentially increasing data size deters manual curation strategies and necessitates a rethink in the current computational solutions, motivating an approach that relies less on human expertise and intervention. Learning distributed representations of ontological terms, entities and relations provides a sufficient workforce for automatically building semantic bridges between semantically heterogeneous systems and successfully generalising across a plethora of practical application domains.

1.2 Thesis Contributions

Ontology alignment and knowledge base completion constitute challenging research problems with various prominent applications such as the automatic knowledge exchange and acquisition between intelligent agents [217], acting as support tools for collaborative ontology development initiatives (e.g., the OBO Foundry [201] and the Industrial Ontologies Foundry [103]), information extraction from heterogeneous IoT devices [124], etc. This thesis proposes to leverage semantic information and statistical regularities captured in external corpora, semantic lexicons and the ontologies themselves to learn representations of ontological terms, entities and relations as a key framework to improve the performance of ontology alignment and knowledge base completion systems and to enforce robust domain generalisation. We illustrate the benefits of our proposed methods using a plethora of practical domains and show significant performance improvements over state-of-the-art approaches. Specifically, this thesis makes the following key contributions:

First, we demonstrate one prominent direction to approach the problem of ontology alignment through representation learning. Our approach overcomes the main obstacles, i.e., the small sample size and the serious class imbalance problem, that have been shown to hinder the application of machine learning approaches to the problem. To overcome these problems, we exploit a transfer learning approach that retrofits pre-trained word DRs to the task of semantic similarity using synonymy/antonymy information extracted from semantic lexicons and the ontologies themselves. The method does not exploit information stemming from ground-truth ontology alignments or misalignments for harnessing word DRs tailored to the ontology matching task. We show significant performance improvements against state-of-the-art systems, however at the cost of a certain amount of degradation in the precision metric. Our method also removes the burden of manually crafting appropriate terminological representations as well as semantic similarity functions tailored to the task of ontology alignment.

Second, we investigate the primary reasons behind the aforementioned shortcoming and devise a neural network architecture to tackle this issue. Specifically, we propose an ontology matching system composed of two neural network components that learn terminological embeddings tailored to semantic similarity. The first component discovers a large amount of true alignments between two ontologies but is prone to errors. The second component corrects these errors, i.e., discovers misalignments, by exploiting unsupervised outlier detection. Aiming at increasing further the recall metric, the first component learns *phrase embeddings* exploiting not only semantically similar words but also phrases. We compare our method to state-of-the-art ontology matching systems and show significant performance gains in both precision and recall metrics.

Third, our proposed solution illustrates a way of how we can minimize the number of similarity functions required for aligning two ontologies. Traditionally, ontology matching approaches have been based on feature engineering in order to obtain different measures of similarity

[62]. This plethora of multiple and complementary similarity metrics has introduced various challenges including choosing the most appropriate set of similarity metrics for each task, tuning the various cut-off thresholds used on these metrics, etc. [39]. Unlike in our approach, only one similarity distance is used. Therefore, there is a drastic decrease in the used similarity metrics and thresholds.

Fourth, another contribution of the work presented in this thesis is that of demonstrating that ontology matching can be performed in the absence of ontology's structural information. Specifically, it was an open question whether ontology's structural information is mandatory for performing ontology matching. Our proposed algorithms, presented in Chapter 4, manage to compare favourably against state-of-the-art systems without using any kind of structural information. Our results support that a great ontology matching performance can be achieved even in the absence of any graph-theoretic information.

Fifth, we provide empirical support that external corpora and semantic lexicons provide sufficient information to perform ontology matching Our methods rely on word vectors pre-trained on large external corpora and on synonymy information provided by semantic lexicons also including the ontologies to be matched. Consequently, we can conclude that external corpora and semantic lexicons provide sufficient information to perform ontology matching by only exploiting the ontological terms.

Finally, we show the advantages that non-Euclidean spaces bring to the task of knowledge base completion. To this end, we quantify the contribution of geometrical space to the task of discovering generic relations between entities. The experimental results validate that the right choice of geometrical space is a critical decision that impacts the performance of embedding models for this relation prediction task. Additionally, our results also shed light on understanding which ontologies and knowledge bases mostly benefit from the use of hyperbolic embeddings. Last but not least, our work shows a new promising direction for developing models that, although not fully expressive, allow to better represent certain families of ontological axioms; opening up for more fine-grained reasoning tasks.

1.3 Thesis Organization

This thesis is organized as follows:

- Chapter 2 provides the background and related work on key topics that will be explored and extended in the context of this thesis.
- Chapter 3 introduces a representation learning approach tailored to ontology matching problem. This Chapter describes an ontology matching approach that uses information from ontologies and additional knowledge sources to extract synonymy/antonymy information that is later used to refine pre-trained word vectors so that they are better suited for the ontology matching task.

- Chapter 4 illustrates the benefits of exploiting two neural network components that learn terminological representations tailored to semantic similarity and misalignment detection, respectively. The first component discovers a large amount of true alignments between two ontologies but is prone to errors. The second component corrects these errors.
- Chapter 5 studies the impact of non-Euclidean spaces to the problem of detecting unknown facts in ontologies and knowledge bases. To this end, the hyperbolic space is exploited as a potential candidate for better representing the topological properties of ontologies and knowledge bases.
- Chapter 6 concludes the thesis and presents future research directions.

1.3.1 Bibliographic Notes

This thesis was conducted under the supervision of my advisor, Dimitrios Kyritsis. Chapter 3 is based on a conference paper published in the *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* [120]. Chapter 4 is based on a journal paper published in the *Journal of Biomedical Semantics* in 2018 [121]. Finally, Chapter 5 is based on a conference paper accepted for publication in the *2020 Conference of the European Semantic Web Conference* [123]. A preliminary version of the aforementioned conference paper has appeared as a preprint on arXiv [122].

2 Background & Related Work

"There is a science which studies being qua being, and the properties inherent in it in virtue of its own nature. This science is not the same as any of the so-called particular sciences, for none of the others contemplates being generally qua being; they divide off some portion of it and study the attribute of this portion, as do for example the mathematical sciences."

Aristotle, Metaphysics – Book IV

This chapter provides the background and related work on key topics that will be explored and extended in the context of this thesis. To this end, we briefly introduce ontologies from both a philosophical and an applied perspective and we provide a formal definition of an ontological entity alignment. Next, we cover the background to the distributed representation learning as well as its applications to ontology matching, sentence representation, outlier detection and entity and relation representation on both Euclidean and non-Euclidean spaces.

2.1 Applied Ontologies

Historically, ontology emerged as a fundamental discipline of metaphysics, i.e., the philosophical branch devoted to the study of reality [7]. In that context, ontology became known as the *science of being qua being* concerned with the study of what categories of entities can exist in reality and of the relations that these entities share with each other. Its primary focus lays on identifying the most fundamental features of reality common to all domains. Figure 2.1 shows a part of the *Porphyrian Tree* that illustrates some of the fundamental categories proposed by Aristotle [16]. For instance, the things that exist in reality, according to Aristotle, are divided into *Material* and *Immaterial* substances. An *Animal* is an *Animate Entity* and it can also be safely deduced that it is also a *Material Substance*. Discovering these fundamental categories and the relations that shape them constitutes one of the holy grails of philosophy and its impact on the scientific thought will be immeasurable.

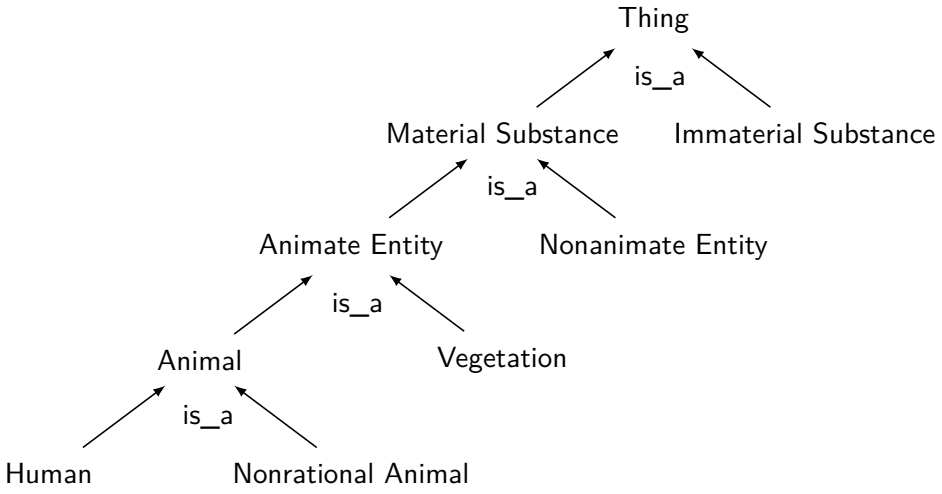

Figure 2.1 – Part of the Porphyrian Tree.

In recent times, the term *ontology* has been extensively used in computer and information science, however, with a different meaning compared to its traditional usage in philosophy. In this setting, ontology refers to a "standardised representational framework" [11, p. xxi] used for describing data and information using a consistent and formally defined set of terms including the relations that associate these terms. Notably, these *applied ontologies* have been proven an important ally in fostering semantic interoperability, i.e., "the ability of two or more systems to exchange information in such a way that the meaning of the information generated by any one system can be automatically interpreted by each receiving system accurately enough to produce results useful to its end users" [11, p. 38]. For the rest of this thesis, when we refer to an *ontology*, we will only mean an *applied ontology*. Based on this agreement, we also omit the word "applied" for brevity.

2.1.1 Ontology Alignment Definition

Before we proceed with the formal definition of an ontological entity alignment, we will introduce the needed formalism. Let O, O' denote two set of terms used in two ontologies and let R be a set of binary relations' symbols. For instance, $=, \neq, is\_a$ can be some of the R set's members. It is of importance to note that ontologies do formally define the set of relations required for describing a certain domain. It should be highlighted that the set R of the binary relations' symbols may or may not have a non-empty intersection with the relations' symbols defined in the two aforementioned ontologies. For example, the set R can contain the symbol for the equivalence relation, i.e., =, although this is not defined in either of the two ontologies. We introduce a set $T = \{(e, r, e') \mid e \in O, e' \in O', r \in R\}$ to denote a set of possible binary relations between O and O' [83]. Moreover, let $f\colon T \rightarrow [0,1] \subset \mathscr{R}$ be a function, called "confidence function", that maps an element of T to a real number v, such that $0 \leq v \leq 1$. The real number v corresponds to the degree of confidence that exists a relation r between e and e' [62].

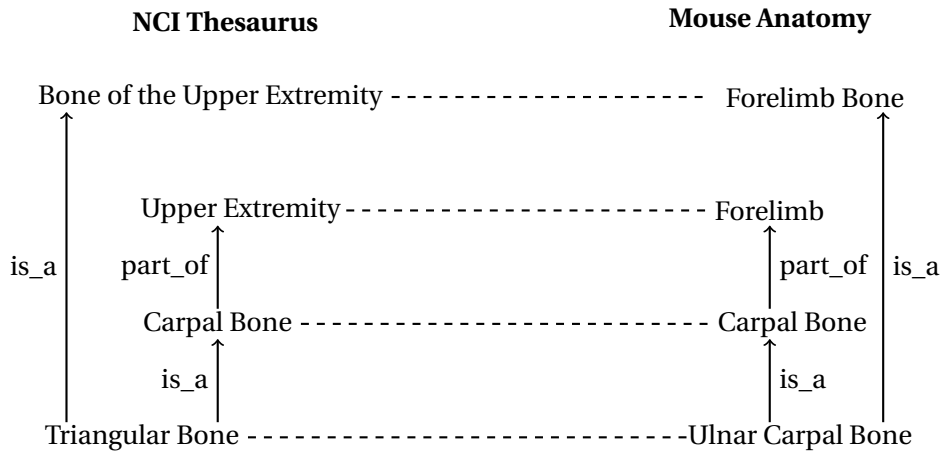

Figure 2.2 – Example of alignments between the NCI Thesaurus and the Mouse Ontology (adapted from [22]). The dashed horizontal lines correspond to equivalence matchings between the NCI Thesaurus and the Mouse Anatomy ontology.

We call a set T of possible relations to be "valid despite integration inconsistency", iff T is satisfiable. As a counterexample, the set $\{(e,=,e'),(e,\neq,e')\}$ corresponds to a non-valid despite integration inconsistency set of relations. It should be noted that in this section we slightly differentiate from the notation used in Description Logics [13], where a relation (Role) between two entities is denoted as: $r(e,e')$. Moreover, it is important to highlight the role of the phrase "despite integration inconsistency" in our definition. The ontology resulting from the integration of two ontologies O and O' via a set of alignments T may lead to semantic inconsistencies [107, 204]. As the focus of ontology alignment lays on the discovery of alignments between two ontologies, we treat the procedure of inconsistency check as a process that starts only after the end of the ontology matching process.[2]

Based on the aforementioned notations and definitions, we will proceed with the formal definition of what an ontological entity alignment is. Let, T be a valid despite integration inconsistency set of relations and f be a confidence function defined over T. Let $(e,r,e') \in T$, we define an ontological entity correspondence between two entities $e \in O$ and $e' \in O'$ as the four-element tuple:

$$cor_r(e,e') = (e,r,e',f(e,r,e')) \tag{2.1}$$

where r is a matching relation between e and e' (e.g., equivalence, subsumption) and $f(e,r,e') \in [0,1]$ is the degree of confidence of the matching relation between e and e'. According to the examples presented in Figure 2.2, (triangular bone, =, ulnar carpal bone, 1.00) and (triangular bone, *is_a*, forelimb bone, 1.00) present one equivalence as well as a subsumption entity correspondence, accordingly.

In Chapter 3, we present an algorithm that discovers many-to-many equivalence correspondences between two ontologies in the sense that we did not exclude the possibility of existence of equivalence correspondences between terms of the same ontology O that map to one or

more synonymous terms of another ontology O'. In Chapter 4, we focus on discovering one-to-one equivalence correspondences between two ontologies. In absence of further relations, the produced set of relations by our proposed algorithms will always correspond to a valid despite integration inconsistency set. Finally, the work presented in Chapter 5 investigates the role of the embedding space in the task of establishing general relations, i.e., not only restricted to equivalence correspondences, between two entities e, e' appearing in the same ontology or knowledge base. This problem is also known in the literature as the knowledge base completion (KBC) or the link prediction problem [168]. Please note that contrary to the definition provided in [121], we did not restrict the two ontologies O, O' to be distinct in our definition in order to provide a unified framework for the discovery of relations between entities appearing in distinct ontologies but also in the same ontology.

2.2 Distributed Representations

Distributed representations are rooted on the information processing theory of *Connectionism* where information processing is performed not in terms of a serial symbol manipulation, as reasoning based on an ontological model would require, but in terms of transformations over "pattern[s] of activity distributed over many computing elements" [100, p. 77] [146]. Distributed representations, whose philosophical origin can even be traced back to Aristotle's work "De memoria et reminiscentia" [206] [10, p. 3], offer a powerful way to approach problems that appear computationally intractable at first sight such as that of computing the semantic similarity between two distinct terms. It should be noted that unlike the ontological representations whose appropriateness is also measured in terms of how well the *a priori* nature of reality is captured [85], the success of distributed representations is mostly measured in a strictly functional way, i.e., whether or not they solve the task at hand in a satisfying way.

As illustrated in Chapter 1, learning distributed representations of ontological terms, entities and relations that are tailored to the problem of ontology alignment and link prediction plays the central role for this thesis. For that reason, the rest of this section presents and discusses state-of-the-art representation learning techniques that, as will be proven in the next chapters, provide a powerful tool for approaching the aforementioned problems. To this end, we begin by studying the problem of learning distributed representations on Riemannian Manifolds. We continue by presenting various state-of-the-art feature engineering approaches for ontology alignment and discuss the early work on applying machine learning and representation learning to the problem. Next, we overview the Siamese CBOW [116] and the Denoising Autoencoder [231] architectures that are exploited and extended in Chapter 4 for deriving terminological representations tailored to semantic similarity. Finally, we overview entity and relation representations learning architectures on both Euclidean and non-Euclidean spaces that are critical for the work presented in Chapter 5, where the contribution of geometrical space for the task of link prediction is explored.

2.2.1 Learning Distributed Representations on Riemannian Manifolds

In this section, we briefly cover some key notions of *Differential Geometry* that will allow us to frame the problem of learning distributed representations on Riemannian manifolds. Going beyond the classical Euclidean setting is the primary focus of Chapter 5 where the hyperbolic space is exploited in an attempt to better represent the topological properties and axioms of knowledge bases and ontologies. It should be noted that this section does not aim to provide an exhaustive and in-depth coverage of this rich field. We welcome the interested readers to refer to [209, 137] for a rigorous treatment of the field.

To begin with, a *n-dimensional manifold* $\mathcal{M}$, conceived as a generalisation to higher dimensions of the notion of the surface in the three-dimensional Euclidean space, is a topological space where around every point belonging in $\mathcal{M}$ we can find a neighbourhood that contains this point and locally resembles the Euclidean space. More precisely, a n-dimensional manifold $\mathcal{M}$ is a topological space having the property that for every $x \in \mathcal{M}$ there is a neighbourhood U such that $x \in U$ and U is homeomorphic to $\mathbb{R}^n$. A *differentiable manifold* (also known as *smooth* or C^∞ manifold) is a manifold that can be locally linearly approximated allowing, thus, to generalise various classical notions of Calculus. For instance, the *tangent space* $T_x\mathcal{M}$ of a smooth manifold at $x \in \mathcal{M}$ generalises the notion of the tangent plane and is defined to be the first order linear approximation of $\mathcal{M}$ around x. The tangent space constitutes a real vector space that intuitively incorporates all the possible directions that are tangentially passing through x.

Building on these definitions, a *Riemannian manifold* $(\mathcal{M}, g)$ is a smooth manifold equipped with a collection $g = \left(g_x\right)_{x \in \mathcal{M}}$ of smoothly varying with respect to x inner products $g_x : T_x\mathcal{M} \times T_x\mathcal{M} \rightarrow \mathbb{R}$ [75]. The aforementioned collection of inner products g, known as *Riemannian metric*, offers a natural way to locally introduce the notions of angle, curve's length, volume, etc. [209, Chapter 9]. Global notions can be defined in a later step by integrating these local contributions. For example, the *geodesic* generalises the notion of a straight line in the Euclidean space and it is defined as the shortest path connecting two points $x, y \in \mathcal{M}$, given by:

$$d(x, y) = \inf_{\gamma} \int_0^1 \sqrt{g_{\gamma(t)}\left(\dot{\gamma}(t), \dot{\gamma}(t)\right)}\, dt \tag{2.2}$$

where $\gamma : [0, 1] \rightarrow \mathcal{M}$ is a piecewise smooth curve such that $\gamma(0) = x$ and $\gamma(1) = y$. It should be noted that the definition of the geodesic makes use of the *local notion* of length $ds = \sqrt{g_{\gamma(t)}\left(\dot{\gamma}(t), \dot{\gamma}(t)\right)}\, dt$.

One fundamental tool in Mathematical Optimisation is that of moving with a constant speed along the direction defined by the derivative of a function. For instance, let $\mathcal{D} = \left\{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\right\}_{i=1}^n$ be a set of observations, $f : \mathcal{X} \rightarrow \mathcal{Y}$ be an approximation function and

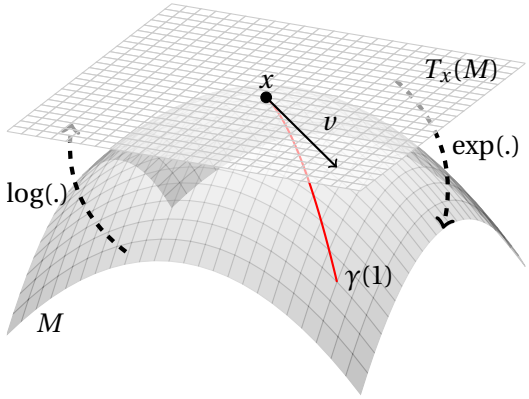

Figure 2.3 – A visualisation of a Riemannian manifold in the 3-dimensional space.

$\mathscr{L}(\cdot,\cdot) \to \mathbb{R}$ be a loss function [195, Chapter 2]. Then, minimising the *empirical risk* on $\mathscr{D}$:

$$\mathscr{R}_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}\left(f(x_i;\boldsymbol{w}), y_i\right) \tag{2.3}$$

based on *Stochastic Gradient Descent* (SGD) requires taking a step in the negative gradient direction as follows, where k denotes the index of the iterative sequence:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla \mathscr{L}\left(f(x_{i_k};\boldsymbol{w}_k), y_{i_k}\right) \tag{2.4}$$

where $1 \le i_k \le n$ is a random index, $k \in \mathbb{N}$ the iteration index and a_k is a positive stepsize [29]. To generalise this notion in the Riemannian setting requires rethinking the SGD update as moving along a geodesic γ starting at $\gamma(0) = \boldsymbol{w}_k$ with velocity $\dot{\gamma}(0) = -\alpha_k \nabla \mathscr{L}\left(f(x_{i_k};\boldsymbol{w}_k), y_{i_k}\right)$. More precisely, the *exponential map* $\exp_x$ at $x \in \mathscr{M}$ is defined to be equal to $\gamma(1)$ where γ is the unique geodesic satisfying $\gamma_v(0) = x$ and $\dot{\gamma}_v(0) = v$, where $v \in T_x\mathscr{M}$ is a *tangent vector* to the manifold at x used as the Riemannian analogue of the gradient [209, p. 333]. The *logarithmic map* $\log_x$ is defined to be the inverse map, i.e., $\log_x = \exp_x^{-1} : \mathscr{M} \to T_x\mathscr{M}$. It is worth noting that the exponential map defines a natural way to project a $v \in T_x\mathscr{M}$ to a point $\exp_x(v) \in \mathscr{M}$ on the manifold. Based on the above, the SGD update defined in Equation (2.4) can be rewritten as follows, where k denotes the index of the iterative sequence:

$$\boldsymbol{w}_{k+1} = \exp_{\boldsymbol{w}_k}\left(-\alpha_k \nabla \mathscr{L}\left(f(x_{i_k};\boldsymbol{w}_k), y_{i_k}\right)\right) \tag{2.5}$$

Figure 2.3 illustrates the above definitions. However, it can be the case that the exponential map is not well-defined for every point for the manifold. Hopefully, the Riemannian manifolds $\mathscr{M}$ whose every geodesic $\gamma : [a,b] \to \mathscr{M}$ can be extended to a geodesic from $\mathbb{R}$ to $\mathscr{M}$ have the property that their exponential map is well-defined on the entire tangent space [209, p. 341]. One classical example of such manifolds, that are known as *geodesically complete manifolds*, is the Poincaré-ball which is studied in Chapter 5.

The distributional hypothesis that was introduced in Chapter 1 provided a fertile ground for the development of a plethora of machine learning-based methods for word and sentence representations [48, 18, 149, 150, 176, 179, 5, 50]. A by now classical example constitutes the work of Mikolov et al. [149] where the *Skip-gram* architecture was proposed. In the rest of this section, we illustrate how this word embedding model can be adapted to the Riemannian setting so as to provide an illustrative example of the definitions provided above. The presentation of this section partially follows the works of [132, 150].

The Skip-gram architecture attributes a high-dimensional Euclidean vector to every word appearing in a fixed vocabulary V. In the Riemannian setting, each words will be assigned to a point in a Riemannian manifold $\mathscr{M}$. The classical training objective pushes the model to predict from a given word its surrounding words in a sentence [150]. More precisely, Skip-gram maximises the following average log probability:

$$\mathscr{L}(\boldsymbol{w}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\leq j\leq c\\ j\neq 0}} \log p(w_{t+j} \mid w_t) \tag{2.6}$$

where $w_i \in \mathscr{M}$ for $1 \leq i \leq |V|$, and c is the *context window* that defines the limits of the considered surrounding words for each word. Please note that the definition of the Skip-gram loss, as defined in Equation (2.6), is rather problematic for words in corpora that do not have enough surrounding words either on their left or on their right side. To simplify the notation, we make the weak assumption that such words are omitted during training. The probability $p(w_j \mid w_t)$ can be formulated using the *softmax function* as follows:

$$p(w_j \mid w_t) = \frac{\exp\big(f(w_j, w_t)\big)}{\sum_{i=1}^{|V|}\exp\big(f(w_i, w_t)\big)} \tag{2.7}$$

where $f : \mathscr{M} \times \mathscr{M} \to \mathbb{R}$ is a similarity function that quantifies the *distributional similarity* of two given words. At this step, it should be noted that we differentiate from the work of Mikolov et al. [150] and we do not introduce either the *hierarchical softmax* [155] or *negative sampling*, since our primary focus is to illustrate the transition from a machine learning model defined in Euclidean space to its analogue in Riemannian space.

To maximize the objective defined in Equation (2.6), we can use *Stochastic Gradient Ascent* (SGA). Let ∇ denote the *Riemannian Gradient* with respect to $\boldsymbol{w}$ and α_k be a positive stepsize sequence. Then, the SGA update can be formulated as:

$$\boldsymbol{w}_{k+1} = \exp_{\boldsymbol{w}_k}\left(\alpha_k \nabla \mathscr{L}\left(\boldsymbol{w}_k\right)\right) \tag{2.8}$$

It should be noted that the exact computation of the exponential map is often difficult, whereas first-order approximations of the exponential map are much more easier to compute [26]. More precisely, we define the retraction $\mathscr{R}_w(v)$ of an exponential map as a map from the tangent space $T_w\mathscr{M}$ to $\mathscr{M}$ having the property $d\left(\mathscr{R}_w(tv), \exp_w(tv)\right) = O(t^2)$. Based on the

definition of retraction, the SGA update can be written as:

$$\boldsymbol{w}_{k+1} = \mathscr{R}_{\boldsymbol{w}_k}(\alpha_k \nabla \mathscr{L}(\boldsymbol{w}_k)) \tag{2.9}$$

In the work of Bonnabel [26], it has been proven that there exist stepsize sequences $(\alpha_k)_{k\geq 0}$ so that the sequence of updates defined in Equation (2.8) converges. Interestingly, it has also been proven that there exist stepsize sequences $(\alpha_k)_{k\geq 0}$ so that the sequence of updates defined in Equation (2.9) converges when the curvature of the Riemannian manifold is non positive [26]. One example of such manifold is the Poincaré-ball which is studied in Chapter 5.

2.2.2 Ontology Matching: From Feature Engineering to Representation Learning

Ontology matching is a rich research field where multiple and complementary approaches have been proposed. In this section, we cover the state-of-the-art feature engineering approaches against which we compare our proposed representation learning based methods that are presented in Chapters 3 and 4. Additionally, we briefly cover key machine learning based algorithms for ontology matching based on binary classification that aided to raise the understanding of the serious class imbalance problem that characterises the problem of ontology matching and hinders learning. Finally, we discuss the early work on approaching the problem of ontology matching through representation learning and motivate the key obstacles of these approaches. At this point, it should be mentioned that ontology matching is a rich and fruitful research field where multiple innovative ideas have been proposed. We welcome the interested reader who seeks additional information to this rich field to refer to [62, 197, 171, 9].

The vast majority of ontology matching research follows the feature engineering approach [234, 38, 117, 106, 64, 163, 88]. Features are generated using a broad range of techniques [9, 92], ranging from the exploitation of terminological information, including structural similarities and logical constraints such as datatype properties, cardinality constraints, etc. Ontology matching is done by acting on the aforementioned features in different ways. Heuristic methods that rely on aggregation functions such as *max, min, average, weighted sum*, etc., to fuse the information found in these features are quite popular [9]. In parallel, they make use of various external semantic lexicons such as Uberon, DOID, Mesh, BioPortal ontologies and Wordnet as a means for incorporating background knowledge useful for discovering semantically similar terms. CroMatcher [88], AML [67, 68] and XMap [53] extract various sophisticated features and use a variety of the aforementioned external domain-specific semantic vocabularies to perform ontology matching. Moreover, LogMap, AML and XMap exploit complete and incomplete reasoning techniques so as to repair incoherent mappings [147]. Unlike the aforementioned approaches, FCA_Map [255, 256] uses Formal Concept Analysis [241] to derive terminological hierarchical structures that are represented as lattices. The matching is performed by aligning the constructed lattices taking into account the lexical and structural information that they incorporate.

Several works exploit supervised machine learning for Ontology Matching [54, 141, 140]. In the work of Mao et al. [141], ontology mapping is explicitly casted as a binary classification problem. The authors generate various domain independent features to describe the characteristics of the entities and train an SVM classifier on a set which provides positive and negative examples of entity alignments. In general, the number of real alignments is orders of magnitude smaller than the number of possible alignments which introduces a serious class imbalance problem [140] hindering learning. Since we only use supervision to refine the word vector representations we avoid altogether the class imbalance problem in the work presented in Chapter 3 and Chapter 4. Representation learning has so far limited impact on ontology matching. To the best of our knowledge, only two approaches, [254] and [244, 205], have explored so far the use of unsupervised deep learning techniques. Both of these approaches use a combination of the class ID, labels, comments, etc. to describe an ontological entity in their algorithms. Zhang et al. [254] are the first ones that investigated the use of word vectors for the problem of ontology matching. They align ontologies based on *word2vec* [149] vectors trained on Wikipedia. They were the first that reported that the general-purpose word vectors were not good candidates for the task of ontology matching. Xiang et al. [244, 205] proposed an entity representation learning algorithm based on Stacked Auto-Encoders [19, 36]. However, training such powerful models with so small training sets is problematic. We overcome both of the aforementioned problems in the works presented in Chapter 3 and Chapter 4 by using a transfer learning approach, known to reduce learning sample complexity [177], which retrofits pre-trained word vectors to a given ontological domain.

2.2.3 Learning Sentence Representations from Labeled Data

In this section, we briefly overview certain neural-based approaches for learning distributed representations exploiting supervision that will be proven useful for the work presented in Chapter 4, where we exploit such neural-based approaches for learning terminological embeddings. To constrain the analysis, we compare neural language models that derive sentence representations of short texts optimized for semantic similarity based on pre-trained word vectors. Nevertheless, we consider in our comparison the Siamese CBOW model [116] since the proposed sentence model in Chapter 4 is highly influenced by it. Likewise, we do not focus on innovative supervised sentence models based on neural networks architectures with more than three layers including [202, 111, 215, 40, 144, 227, 179, 50] and many others. The most similar approach to our extension of Siamese CBOW is the work of Wieting et al. [240]. Wieting et al. address the problem of paraphrase detection where explicit semantic knowledge is also leveraged. Unlike to our approach, a margin-based loss function is used, and negative examples should be sampled at every step introducing an additional computational cost. The most crucial difference is that this model was not explicitly constructed for alleviating the coalescence of semantically similar and semantically associated terms that is discussed in Chapters 3 and 4. Finally, the initial Siamese CBOW model was conceived for learning distributed representations of sentences from unlabeled data. To take advantage of the semantic similarity information already captured in the initial word embeddings, an important

characteristic as demonstrated in various word vectors retrofitting techniques [70, 156, 240], we propose in Chapter 4 to extend the initial model with an knowledge distillation reguralizer [99]. Finally, we further extended the initial softmax setting, with a tempered softmax, with the purpose of enabling the network to capture information hidden in small logit values. For further information regarding the proposed architecture, please refer to the Section 4.2.2.

2.2.4 Autoencoders for Outlier Detection

Neural network applications to the problem of outlier detection have been studied for a long time [242, 143]. Autoencoders seem to be a recent and a very prominent approach to the problem. As has been pointed out in [33], they can be seen as a generalization of the class of linear schemes [94]. Usually, the reconstruction error is used as the outlier score [33]. Recently, Denoising Autoencoders (DAEs) [231] have been used for outlier detection in various applications, such as acoustic novelty detection [142], network's intrusion detection [33], anomalous activities' discovery in video [248]. In Chapter 4, a novel autoencoder-based outlier detection mechanism for discovering ontology misalignements is presented. To the best of our knowledge, this is the first time that the problem of semantic similarity is seen from the viewpoint of outlier detection based on DAEs. Unlike the other approaches, we want to detect outliers in pairs of input. To achieve that, we use the cosine distance over the two produced hidden representations as an outlier score, instead of using the reconstruction error which is customary in the literature. Our motivation is that intrinsic characteristics of the distribution of semantically similar terms are captured in the hidden representation and their cosine distance could serve as an adequate outlier score. Unlike the majority of the aforementioned research work, we do not train end-to-end the DAE but we follow a layer-wise training scheme based on sentence representations produced by our extension of Siamese CBOW. Our impetus is to let the DAE to act on a dataset with significant less noise and bias. Please refer to Section 4.2.3 for concrete details on the proposed outlier detection architecture.

2.2.5 Learning Representations of Entities and Relations Existing in Ontologies & Knowledge Bases

In Chapter 5, we explore the discovery of general relations, i.e., not only equivalence correspondences, between entities appearing in the same ontology or knowledge base. There has been a great line of research dedicated to the task of learning distributed representations for entities and relations in ontologies and knowledge bases. To constrain the analysis, we only consider shallow embedding models that do not exploit deep neural networks or incorporate additional external information beyond the available facts. For an elaborated review of these techniques, please refer to Nickel et al. [168] and Wang et al. [235]. We also exclude from our comparison recent work that explores different types of training regimes such as adversarial training, and/or the inclusion of reciprocal facts [30, 214, 115, 127] to make the analysis less biased to factors that could overshadow the importance of the geometrical space.

In general, the shallow embedding approaches can be divided into two main categories; the translational [27] and the bilinear [167] family of models. In the translational family, the vast majority of models [237, 105, 245, 58] generalise TransE [27], which attempts to model relations as translation operations between the vector representations of the *subject* and *object* entities, as observed in a given fact. In the bilinear family, most of the approaches [249, 169, 221] generalise RESCAL [167] that proposes to model facts through bilinear operations over entity and relations vector representations. In Chapter 5, we focus on the family of translational models, whose performance has been lagging, and propose extensions in the hyperbolic space which by exploiting the topological and the formal properties of KBs bring significant performance improvements.

2.2.6 Hyperbolic Embeddings

In Chapter 5, we explore the contribution of geometrical space for the task of discovering general relations between entities appearing in the same ontology or knowledge base. Specifically, we explore the hyperbolic space as a prominent alternative since it has the potential to better represent the topological properties of ontologies and knowledge bases as it is further explained in Chapter 5. There has been a growing interest in embedding scale-free networks in the hyperbolic space [24, 172]. The majority of these approaches are based on maximum likelihood estimation, that maximises the likelihood of the network's topology given the embedding model [172]. Additionally, Verbeek and Suri [230] investigated the conditions under which general undirected graphs can be embedded in the hyperbolic space with minimum distortion. Hyperbolic geometry was also exploited in various works as a way to exploit hierarchical information and learn more efficient representations [165, 166, 74, 190, 51, 219, 128]. However, this line of work has only focused on single-relational networks.

Recently and in parallel to our work presented in Chapter 5, two other works have explored hyperbolic embeddings for KBs. Contrary to our work where Möbius or Euclidean addition is used as a translational operation, Suzuki et al. [216] exploit vector fields with an attractive point to generalise translation in Riemannian manifolds. Their approach, although promising, shows a degraded performance on commonly used benchmarks. Similarly to our approach, Balažević et al. [14] extend to the hyperbolic space the family of translational models demonstrating significant advancements over state-of-the-art. However, the authors exploit both the hyperbolic as well as the Euclidean space by using the *Möbius Matrix-vector multiplication* and Euclidean scalar biases.[3] Unlike our experimental setup, the authors also include reciprocal facts. Although their approach is beneficial for knowledge base completion, it becomes hard to quantify the contributions of hyperbolic space. This is verified by the fact that their Euclidean model analogue performs in line with their "hybrid" hyperbolic-Euclidean model. Finally, neither of these works studies the types of rules that their proposed models can effectively represent.

3 Word-level Representation Learning for Ontology Alignment

"You shall know a word by the company it keeps!"

John Rupert Firth

While representation learning techniques have shown great promise in application to a number of different Natural Language Processing (NLP) tasks, they have had little impact on the problem of ontology matching. Unlike past work that has focused on feature engineering, we present a novel representation learning approach that overcomes many of the obstacles of previous approaches. Our proposed method, dubbed *DeepAlignment*, refines pre-trained word vectors aiming at deriving terminological representations that are tailored to the ontology matching task. Unlike previous approaches that exploited machine learning, the absence of explicit information relevant to the ontology matching task during the refinement process enables DeepAlignment to overcome the small sample size that characterises the problem of ontology matching. We empirically evaluate our method using standard ontology matching benchmarks. We present significant performance improvements over the current state-of-the-art, demonstrating the advantages that representation learning techniques bring to ontology matching.

3.1 Introduction

Translation across heterogeneous conceptual systems is an important challenge for cognitive science [81, 212]. Ontology Matching constitutes the task of establishing correspondences between semantically related entities from different ontologies, as illustrated in Figure 3.1. Similarly, ontology matching is crucial for accomplishing a mutual understanding across heterogeneous artificial cognitive agents [217]. However, despite the many proposed solutions, it is widely accepted that there is no solution robust enough to deal with the high ontological linguistic variability [196, 197]; hampering, thus, the discovery of shared meanings.

Research in automatic ontology matching has focused on engineering features from termino-

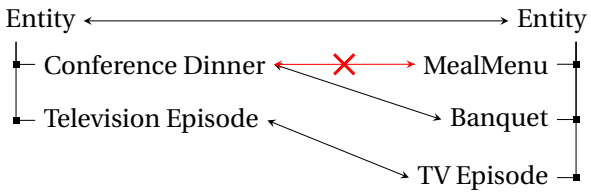

Figure 3.1 – Example of alignments (black lines) and misalignments (red crossed lines) between ontologies.

logical, structural, extensional (ontology instances) and semantic model information extracted from the ontological model. These features are then used to compute ontological entity similarities that will guide the ontology matching. Deriving such features for a given problem is an extremely time consuming task. To make matters worse, these features do not transfer in other domains. As Cheatham and Hitzler [32] have recently shown, the performance of ontology matching based on different textual features varies greatly with the type of ontologies under consideration.

At the same time, machine learning research is characterised by a shift from feature engineering based approaches to feature and representation learning as a result of the performance improvements brought by deep learning methods. A by now classical example is the unsupervised learning of semantic word representations based on the *distributional hypothesis* [93], i.e., the assumption that semantically similar or related words appear in similar contexts [48, 18, 149, 150, 176]. Word vectors have the potential to bring significant value to ontology matching given the fact that a great deal of ontological information comes in textual form.

One drawback of these semantic word embeddings is that they tend to coalesce the notions of *semantic similarity* and *semantic association* [97]. For instance, the word “harness" is highly associated with the word “horse", as they share strong associations, i.e., a harness is often used on horses [136]. From an ontological point of view, however, these types should not be similar. Moreover, as unsupervised learning requires even larger text corpora, the learned vectors tend to bring closer words with similar frequency instead of similar meaning [71]. Clearly, word representations that reflect frequency instead of meaning is an undesired feature if we seek to exploit word vectors for ontology matching; alignment based on such representations will reflect similar frequency instead of similar meaning.

A number of lightweight vector space representation refining techniques were introduced recently in an effort to correct these biases [70, 156]. They use synonymy and antonymy constraints extracted from semantic lexicons to refine the learned word representations and make them better suited for semantic similarity tasks. Such methods are a way to inject domain-specific knowledge to tailor the learned word representations to a given task. As a result, we can exploit the synonymy/antonymy constraints to learn semantic word representations that are better candidates for ontology matching.

In this chapter, we learn representations of ontological entities instead of feature engineering

them. We use the learned representations to compute the entities' semantic distances and to subsequently perform the ontology matching task. In order to represent the ontological entities, we exploit the textual information that accompanies them. We represent words by learning their representations using synonymy and antonymy constraints extracted from general lexical resources and information captured implicitly in ontologies. We cast the problem of ontology matching as an instance of the Stable Marriage problem [73] using the entities semantic distances.

Our approach has a number of advantages. The word embeddings we establish are tailored to the domains and ontologies we want to match. The method relies on a generic unsupervised representation learning solution which is important given the small size of training sets in ontology matching problems. We evaluate our approach on the Conference dataset provided by the Ontology Alignment Evaluation Initiative (OAEI) campaign and on a real world alignment scenario between the Schema.org and the DBpedia ontologies. We compare our method to state-of-the-art ontology matching systems and show significant performance gains on both benchmarks. Our approach demonstrates the advantages that representation learning can bring to the task of ontology matching and shows a novel way to study the problem in the setting of recent advances in NLP.

3.2 Methods

We present an ontology matching approach that uses information from ontologies and additional knowledge sources to extract synonymy/antonymy relations which we use to refine pre-trained word vectors so that they are better suited for the ontology matching task. Since the focus of this chapter lies in detecting equivalence relations between the entities of different ontologies, we only exploit synonymy/antonymy relations. We represent each ontological entity as the bag of words of its textual description, which we complement with the refined word embeddings. We match the entities of two different ontologies by casting the problem of ontology matching as an instance of the Stable Marriage problem. In order to compute the ordering of preferences for each entity, that the Stable Marriage problem requires, we use the entities' pairwise distances. We compute the aforementioned distances using a variant of a document similarity metric.

3.2.1 Preliminaries

In this chapter, we focus on discovering many-to-many equivalence mappings between ontologies. Please refer to Section 2.1.1 for a formal definition of what an entity correspondence is. We will also introduce some additional notation used in this chapter. Let $u_1, u_2 \in \mathbb{R}^d$ be two d-dimensional vectors, we compute their cosine distance as follows: $d(u_1, u_2) = 1 - cos(u_1, u_2)$. For $x \in \mathbb{R}$, we define the *rectifier* activation function as: $\tau(x) = \max(x, 0)$.

3.2.2 Learning Domain Specific Word Vectors

The *counter-fitting* method [156] uses synonymy and antonymy relations extracted from semantic lexicons to refine and adapt pretrained word embeddings for given semantic similarity tasks. We broaden the concept of antonymy relations and allow a larger class of ontology relations to be conceived as antonyms. This allows us to inject domain knowledge encoded in ontologies and produce more appropriate word vectors for the ontology matching task. In the rest of the section we revise the main elements of the counter-fitting method and describe how we can exploit it for learning domain specific word embeddings.

Let $V = \{v_1, v_2, \dots v_N\}$ be an indexed set of word vectors of size N. The counter-fitting method transforms a pretrained vector set V into a new one $V' = \{v'_1, v'_2, \dots v'_N\}$ based on a set of synonymy and antonymy constraints S and A, respectively. This is done by solving the following non-convex optimization problem:

$$\min_{V'} \kappa_1 AR(V') + \kappa_2 SA(V') + \kappa_3 VSP(V, V')$$

The $AR(V')$ function defined as:

$$AR(V') = \sum_{(u,w)\in A} \tau(1 - d(v'_u, v'_w))$$

is called *antonym repel* and pushes the refined word vectors of “antonymous" words to be away from each other. As we already mentioned, we extend the notion of antonymy relations with respect to its more narrow traditional linguistic definition. We consider that two entities in a given ontology are “antonymous" if they have not been explicitly stated as equivalent, in the sense of a logical assertion or a synonymy relation found in a semantic lexicon.

The $SA(V')$ function defined as:

$$SA(V') = \sum_{(u,w)\in S} d(v'_u, v'_w)$$

is called *synonym attract* and brings closer the transformed word vectors of synonyms. In order to extract synonymy information we search for paraphrases in semantic lexicons. Concretely, let $\omega_1 = \{word_1^1, word_2^1, \dots, word_m^1\}$, $\omega_2 = \{word_1^2, word_2^2, \dots, word_n^2\}$ be the textual information of two entities from different ontologies. If the combination $\{word_i^1, word_j^2\}$ or $\{word_j^2, word_i^1\}$ for some $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ appears as a paraphrase in any semantic lexicon then we add the synonymy information (u, w) in the set S of synonymy constraints.

The $VSP(V, V')$ function defined as:

$$VSP(V, V') = \sum_{i=1}^{N} \sum_{j\in N(i)} \tau(d(v'_i, v'_j) - d(v_i, v_j))$$

forces the refined vector space to reflect the original word-vector distances. $N(i)$ is the set of words that lie within ρ distance from the i-th word vector in the original vector-space. The experiments show that the value of ρ does not affect significantly the performance of the whole algorithm, so for computational efficiency we fix it to $\rho = 0.05$. We minimize the objective function with stochastic gradient descent (SGD). We use as a convergence criterion the norm of the gradient. We continue updating the model until this is smaller than 10^{-5}. In our experiments we typically observe convergence within less than 25 iterations.

3.2.3 Semantic Distance Between Entities

As before, let V' be the refined word vectors and $\omega_1 = \{word_1^1, word_2^1, \ldots, word_m^1\}$, $\omega_2 = \{word_1^2, word_2^2, \ldots, word_n^2\}$ be the textual information that describes two entities from different ontologies. The textual information of an entity can be extracted from different sources such as the entity's name, label, comments, etc. We replace the appearance of a word with its refined word vector. Hence, we end up with two sets of word vectors Q and S, respectively. In order to estimate how semantically similar the corresponding terms of two entities are, we use a semantic distance defined over the entities' terminological representations, i.e., the set of word vectors associated with each entity.

There have been many ways to compute the semantic similarity of two word sets, such as the *Word Moving Distance* [126] and the *Dual Embedding Space Model* (DESM) [159]. We will base our semantic distance δ on a slight variation of the DESM similarity metric. Our metric δ computes the distance of two sets of word vectors Q and S as follows:

$$\delta(Q,S) = \frac{1}{|Q|} \sum_{q_i \in Q} d(q_i, \bar{S}) \tag{3.1}$$

where $\bar{S} = \frac{1}{|S|} \sum_{s_j \in S} \frac{s_j}{\|s_j\|}$ is the normalised average of the word embeddings that constitute the set of words S.

Hence, one of the word vectors' sets is represented by the centroid of its normalized vectors. The overall set-to-set distance δ is the normalized average of the cosine distance d between the computed centroid and the other's set word vectors. A first observation is that the introduced distance is not symmetric. Ideally, we would expect the semantic distance of two word sets to be irrelevant of the order of the inputs. To make it symmetric, we redefine the distance between two sets of word vectors as:

$$dis(\omega_1, \omega_2) = \max(\delta(Q,S), \delta(S,Q)) \tag{3.2}$$

It is important to note that $dis(\omega_1, \omega_2)$ is not a proper distance metric as it does not satisfy the triangle inequality property. Despite this fact, it has proved to work extremely well on all the ontology matching scenarios used for our system evaluation in this chapter.

Algorithm extendMap(e, h, $\mathcal{O}'$, P_e, $i_{e'}$, n, ϵ, r)

Input: source entity: e
hash function from integers to entities: h
subsumption's transitive closure: $\mathcal{O}'$
sorted (increasingly) preference matrix: P_e
index of optimal solution: $i_{e'}$
number of target's ontology entities: n
$\epsilon-$optimality value: ϵ
number of relatives: r
Output: sequence of $\epsilon-$optimal mappings

Initialization: list = $\emptyset$
$opt = P_e[i_{e'}]$
$e' = h(i_{e'})$
for $i = \min(i_{e'} + 1, n)$ to $\min(i_{e'} + r, n)$ **do**
  $tmp = P_e[i]$
  if $abs(opt - tmp) < \epsilon$ **then**
    $e_i = h(i)$
    if $(e_i, e') \in \mathcal{O}'$ or $(e', e_i) \in \mathcal{O}'$ **then**
      list.append($e \rightarrow e_i$)
    end if
  end if
end for

Figure 3.2 – Definition of extendMap algorithm.

3.2.4 Ontology Matching

Similar to the work in [244], we cast the ontology matching problem as an instance of the extension of the Stable Marriage Assignment problem to unequal sets [73, 145]. A Stable Marriage Assignment algorithm computes one-to-one mappings based on a preference $m \times n$ matrix, where m and n is the number of entities in ontologies O and O', respectively. Note that the violation of the triangle inequality by our semantic distance, defined in Equation (3.2), is not an impediment to the Stable Marriage Assignment algorithm [73].

The majority of the ontology matching systems produce equivalence mappings with one-to-one cardinality. Hence, one entity e in ontology O can be mapped to at most one entity in e' in O' and vice versa. According to a recent review [9] only two out of almost twenty ontology matching systems provide solutions to detect many-to-many mappings. However, ontology developers focus on different degrees of granularity, so it is expected that one entity from a given ontology can be aligned to more than one entity in another ontology and vice-verca.

To address this problem, we present the extendMap algorithm, shown in Figure 3.2, that extends the one-to-one mappings of the previous step to many-to-many. The basic idea is that certain alignments that were not contained in the solution of the instance of the Stable

Marriage problem were very close to the optimal alignment, in terms of the semantic distance, and they should also be included in the final alignment set. However, despite the use of refined word vectors, we cannot totally avoid possible misalignments due to the semantic similarity and semantic association coalescence. Therefore, any extension of the original alignment set requires special consideration to avoid the inclusion of misalignments.

A way to circumvent the inclusion of significantly many misalignments is to add the constraint that we can extend an one-to-one mapping to a many-to-many one if and only if every entity that is going to be included in the initial alignment shares a subsumption relation with an entity of the initial mapping. The idea behind this restriction stems from the fact that an equivalence relation can be decomposed into two subsumption relations. Therefore, an entity sharing a subsumption relation with an entity of the initial mapping can be considered a probable candidate for extension since it already satisfies a necessary condition. Below, we give a more formal definition of what we will call an ϵ-optimal mapping between two entities e and e' that belong to two different ontologies O and O', respectively.

Definition 1 *Let $e \leftrightarrow e'$ be a mapping, identified by solving the instance of the Stable Marriage problem, between the entities $e \in O$ and $e' \in O'$, where O and O' are two different ontologies. Let $e \leftrightarrow e''$ be another mapping, where $e'' \in O'$. Given an $\epsilon > 0$, we call the mapping $e \leftrightarrow e''$ ϵ-optimal with respect to the mapping $e \leftrightarrow e'$ if and only if the following two conditions hold:*

- $|dis(\omega_1, \omega_2) - dis(\omega_1, \omega_3)| < \epsilon$, *where $\omega_1, \omega_2, \omega_3$ is the textual information of entities e, e' and e'', respectively.*
- *e' and e'' should share a subsumption relation. Equivalently, there must be a logical assertion stating that either e' is subclass of e'' or e'' is subclass of e'.*

The *subsumption restriction* requires that the extended alignments share a subsumption relation in order to reduce the possibility of matchings between entities that are semantically associated. We iteratively search for ϵ-optimal mappings according to the extendMap algorithm, shown in Figure 3.2, to extend the established one-to-one mappings to many-to-many. For efficiency reasons, we do not check all the entities, but only the r closest entities according to the dis distance. As a final step, we iteratively pass through all the produced alignments and we discard those whose semantic distance is greater than a hyperparameter value $thres$.

Last but not least, it should be noted that another strategy for computing optimal matchings between ontologies is to cast the problem as an instance of the minimum weight graph matching problem [62, p. 191]. Contrary to the Stable Marriage assignment problem where only the relative ordering of the preferences is exploited, the minimum weight graph matching problem exploits the full information provided by the semantic distance. Nonetheless, since we cannot totally alleviate the problem of the semantic similarity and semantic association coalescence, two semantically associated terms can have a really small semantic distance; producing, thus, erroneous mappings. This line of thinking justifies our choice to cast the

ontology matching problem as an instance of the Stable Marriage assignment problem in our attempt to reduce erroneous mappings due to semantic association.

3.3 Results & Discussion

In this section, we present the experiments we performed on the OAEI conference dataset and on a real world alignment scenario between the Schema.org and DBpedia ontologies. One of the main problems that we have encountered with the comparative evaluation of our algorithm is that even though numerous ontology matching algorithms exist, for only a very small portion of them either the respective software or the system's output is publicly available. To the best of our knowledge, among all the systems tested in the conference dataset only AML [38] and LogMap [106] are publicly available. As it happens these are two of the state-of-the-art systems. Moreover, AML offers solutions to detect many-to-many alignments [69] and, thus, constitutes a competitive baseline against which we will compare the performance of our algorithm extendMap which also provides many-to-many alignments.

When training to refine the vector representations an unbalanced proportion of synonymy and antonymy constraints sets can cause problems; the set with the lower cardinality will have limited impact on the final word representations. To overcome this problem, we run an additional step of the counter-fitting procedure using only a small random subset of the supernumerary constraints and all the constraints of the minority set. We randomly undersample the larger set and reduce its cardinality to that of the smaller set. We call this additional step the *recounter-fitting* process. To demonstrate the importance of the recounter-fitting process and test the behavior of the pre-trained word vectors in the absence of synonymy and/or antonymy relations, we have conducted additional experiments which we also present.

In all of our experiments we have applied the counter-fitting process upon the Paragram-SL999 word vectors provided by Wieting et al. [239]. With respect to the textual information extracted for each entity, we have only used the entity's ID (rdf:ID). To estimate the precision, recall and $F1$ measure of all the systems, that we consider for testing, and check for the statistical significance of the results we use an approximate randomization test with 1048576 shuffles, as described in Yeh [250].

3.3.1 Semantic Lexicons

Let $\omega_1 = \{word_1^1, word_2^1, \ldots, word_m^1\}$, $\omega_2 = \{word_1^2, word_2^2, \ldots, word_n^2\}$ be the textual information that accompanies two entities from different ontologies. We extracted the synonymy and antonymy constraints that we used in the experiments from the following semantic lexicons:

WordNet: a well known lexical database for the English language [152]. In our experiments we

did not use WordNet synonyms. Instead, we have included WordNet antonymy pairs together with the "antonymy" relations extracted by the ontologies. The strategy that we have followed in order to create the WordNet's antonymy pairs is to consider as antonyms every two words with antonymous word senses.

PPDB 2.0: the latest release of the Paraphrase Database [174]. We have used this database in two different ways. We have used the largest available single-token terms (XXXL version) in the database and we have extracted the *Equivalence* relations as synonyms, and the *Exclusion* relations as antonyms. Additionally, we have searched the whole XXXL version of PPDB for paraphrases based on the words appeared in two entities from different ontologies. Namely, our strategy was the following: If the pair $(word_i^1, word_j^2)$ or the pair $(word_j^2, word_i^1)$ appeared on the PPDB and their type of relation was not *Exclusion*, we considered it as synonym.

WikiSynonyms: a semantic lexicon which is built by exploiting the Wikipedia redirects to discover terms that are mostly synonymous [42]. In our experiments we have used it only on the Schema.org - DBpedia scenario. Our strategy was the following: we search if there exist synonyms in the WikiSynonyms for the ω_1 and ω_2. If this is the case, we extract them and we stop there. In the opposite case we extract the synonyms for each $word_i^1$ and $word_j^2$.

3.3.2 Hyperparameter Tuning

We tuned the hyperparameters on a set of 100 alignments which we generated by randomly sampling the synonyms and antonyms extracted from WordNet and PPDB. We chose the vocabulary of the 100 alignments so that it is disjoint to the vocabulary that we used in the alignment experiments, described in Section 3.3.3, in order to avoid any information leakage from training to testing. We tuned to maximize the $F1$ measure. In particular, we did a coarse grid search over a parameter space for $\kappa_1, \kappa_2, \kappa_3$, r, ϵ and $thres$. We considered $\kappa_1, \kappa_2 \in [0.35, 0.45]$ and $\kappa_3 \in [0.1, 0.2]$ with common step 0.01, $r \in [1, 10]$ with step 1, $\epsilon \in [0.01, 0.1]$ with step 0.01 and $thres \in [0.3, 0.7]$ with step 0.05. We trained for 25 epochs for each hyperparameter using SGD. The best values were the following: $\kappa_1 = 0.4$, $\kappa_2 = 0.4$, $\kappa_3 = 0.1$, $r = 8$, $\epsilon = 0.07$ and $thres = 0.5$. We used the selected configuration on all the alignment scenarios described below.

3.3.3 Evaluation Benchmarks

One of our evaluation benchmarks comes from the Ontology Alignment Evaluation Initiative (OAEI), which organizes annual campaigns for evaluating ontology matching systems. The external to OAEI evaluation benchmark comes from the provided alignments between the Schema.org and the DBpedia ontologies. We provide some further details for each dataset below:

OAEI Conference Dataset: It contains 7 ontologies addressing the same domain, namely the conference organization. These ontologies are suitable for ontology matching task because of

their heterogeneous character of origin. The overall performance (micro-precision, micro-recall, micro-F1) of the systems is tested upon 21 different test cases. Specifically, we summed up the individual true positives, false positives and false negatives based on the system results for the different ontology matching tasks and, in the next step, we computed the performance metrics. The original reference alignment is not closed under the alignment relation, so the transitive closure should be computed before proceeding on the evaluation of the systems.

Schema.org - DBpedia Alignment: It corresponds to the incomplete mapping between Schema.org[4] and DBpedia[5] ontologies. Schema.org [87] is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond. On the other hand, DBpedia [131] is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. This alignment corresponds to a real case scenario between two of the most widely used ontologies in the web today.

3.3.4 Experimental Results

All the systems presented in the Conference dataset experiments (Table 3.1) fall into the category of feature engineering. CroMatcher [88], AML [38], XMap [52] perform ontology matching based on heuristic methods that rely on aggregation functions. LogMap and LogMapBio [106] use logic-based reasoning over the extracted features and cast the ontology matching to a satisfiability problem.

System	Precision	Recall	Micro-F1
DeepAlignment	0.71	0.80	0.75
CroMatcher	0.76	0.69	0.72
AML	0.79	0.65	0.71
DeepAlignment$_*$	0.68	0.68	0.68
XMap	0.81	0.58	0.67
LogMap	0.79	0.58	0.66
LogMapBio	0.75	0.58	0.65
StringEquiv	0.83	0.50	0.62

Table 3.1 – Results on Conference OAEI dataset. StringEquiv corresponds to ontology matching by simple string equivalence check.

Table 3.1 shows the performance of our algorithm compared to the five top performing systems on the Conference 2016 benchmark, according to the results published in OAEI.[6] It should be noted that, traditionally, the performance of ontology matching systems is evaluated giving equal importance to both precision and recall [62, p. 304]. DeepAligment achieves the highest micro-$F1$ measure and the highest recall. We were able to perform statistical significance test only for the two systems that were publicly available. DeepAlignment is significantly better than both of them with a $p\text{-}value \leq 0.05$. In order to explore the performance effect of the

Lexical Information		Precision	Recall	Micro-F1
Synonyms	Antonyms			
-	-	0.63	0.55	0.59
-	✓	0.67	0.51	0.58
✓	-	0.69	0.72	0.71
✓	Restricted	0.65	0.78	0.71
✓	✓	0.71	0.80	0.75

Table 3.2 – Experiments on Conference OAEI dataset.

many-to-many mappings that DeepAlignment produces we also did experiments where our extendMap algorithm was not used, thus generating only one-to-one alignments. We give these results under the DeepAlignment$_*$ listing. It can be seen that DeepAlignment$_*$ achieves the same level of recall as the state-of-the-art systems and this with no feature engineering. When we compare the performance of DeepAlignment$_*$ and DeepAlignment we see that the use of extendMap generates correct many-to-many alignments and thus it does not produce large numbers of false positives. In any case, however, we retain a small precision which indicates a semantic similarity and semantic association coalescence.

We perform additional experiments to investigate the importance of the counter-fitting step, which are summarized in Table 3.2. In all of these experiments, we have applied the extendMap algorithm. The last row of Table 3.2, corresponds to the best result reported in Table 3.1. The first row gives the results of executing the algorithm without the counter-fitting process, just by providing the Paragram-SL999 word vectors. The results support the importance of the counter-fitting process, which succeeds in tailoring the word embeddings to the ontology matching task. By injecting only antonymy information (second row), we observe an increase in precision, but a decrease in recall. This behavior is due to the fact that the antonym repel factor imposes an orthogonality constraint to the word vectors, leading to higher values of the dis distance. In absence of synonymy information, the majority of words tend to become “antonymous". The third row of Table 3.2 gives the performance when we also include synonyms extracted from PPDB but no antonymy information. We can see that this leads to a large increase in all the recorded performance metrics. Finally, we also include antonymy information only from the Cmt and the Conference ontologies found in the Conference dataset. This has two effects: an increase in recall, but a decrease in precision. This can be explained by the fact that even though all ontologies describe the same domain the description granularity provided by each of them is not capable of giving all the antonymy relations needed to provide more refined alignments.

System	Recall
DeepAlignment$_*$	0.82
LogMap	0.5
AML	0

Table 3.3 – Results on aligning Schema.org and DBpedia ontologies.

Table 3.3 summarizes the results obtained by aligning the Schema.org and DBpedia ontologies. The fact that the alignment is incomplete restricts us on evaluating the performance only in terms of the recall metric. To make the comparison as fair as possible, we did not apply the extendMap algorithm. We should highlight that we have applied the recounter-fitting process because the synonyms that we have extracted from the PPDB and WikiSynonyms were very few compared to the constructed “antonyms". The results of the LogMap system show a quite similar behavior with the experiments conducted in the conference dataset. However the recall of AML is zero. It discovers none of the available alignments even though it manages to recall other quite reasonable matchings, which, however, are not included in the ground truth. According to our understanding, this might be an indication of the absence of domain transferability of the extracted features as well as of the implemented metrics.

Parameters			Recall
Recounter-fitting	Synonyms	Antonyms	
-	-	-	0.71
-	-	✓	0.76
-	✓	-	0.84
-	✓	✓	0.76
✓	✓	Restricted	0.82

Table 3.4 – Experiments on aligning Schema.org and DBpedia ontologies. Restricted indicates that we choose only a small random subset of the antonymy constraints.

We summarize in Table 3.4 the results of the experiments we did on the two domains to study the effect of counter-fitting and recounter-fitting. As we can see, even without the counter-fitting, the semantic embeddings show quite good results. This provides evidence on the importance of using representation learning techniques instead of the classical feature engineering choice. By injecting only antonymy information (second row), we observe a different behavior in the recall metric compared to the one presented in Table 3.2. This can be explained by the fact that while the antonym repel factor imposes an orthogonality constraint, its effect is by no means universal to the whole word vector space. Therefore, a misalignment can be pushed far away leaving the space open for a true alignment to be detected. With the addition of the extracted synonyms, we observe an increase of 0.13 in the recall. However, the insertion of the extracted “antonyms" leads to lower performance. This shows practically the importance of applying the recounter-fitting process that allows both the synonym attract and the antonym repel factors to affect the word vectors.

3.3.5 Further Analysis

DeepAlignment versus initial word vectors: To investigate the impact of the initial pre-trained word vectors on DeepAlignment’s performance, we carried out two additional experiments, this time using a set of word2vec vectors [149], trained on the Google news dataset[7]. We report and compare the obtained results to the ones produced by the use of Paragram-SL999

Counter fitting	Word Vectors	Conference Dataset			Schema.org - DBpedia Dataset
		Precision	Recall	Micro-F1	Recall
-	word2vec	0.64	0.52	0.58	0.74
-	Paragram	0.63	0.55	0.59	0.71
✓	word2vec	0.67	0.75	0.71	0.75
✓	Paragram	0.71	0.80	0.75	0.76

Table 3.5 – Dependency of DeepAlignment's performance on the choice of the initial word vectors. The reported results for the Schema.org - DBpedia scenario were obtained without recounter-fitting.

vectors in Table 3.5. In the absence of counter-fitting, the word2vec vectors achieve better results on the Schema.org - DBpedia scenario, however, they exhibit lower performance on the conference dataset. This observation is in accordance with recent studies [98] which show that different word vectors optimization objectives yield representations tailored to different applications and domains. After the application of the counter-fitting process, the use of Paragram-SL999 vectors leads to a better performance. This fact provides additional evidence that word vectors which reflect semantic similarity are better candidates for being further tailored to the ontology matching task.

DeepAlignment versus lexicons' coverage: The choice and coverage of the different lexical resources may have a determining factor on the performance of DeepAlignment. For that reason, we present in Table 3.6 a set of experiments where we exclude a part of the synonymy/antonymy relations from the various semantic lexicons. For both the matching scenarios, we experimented with excluding all the antonyms from PPDB and WikiSynonyms. For the conference dataset, we additionally experimented with including only a subset of PPDB

Dataset	Experimental Setting	Precision	Recall	Micro-F1
Conference	No antonyms from PPDB & WikiSynonyms	0.67	0.76	0.71
	Only a subset of PPDB synonyms	0.67	0.76	0.71
	All available synonyms/antonyms	0.71	0.80	0.75
Schema.org DBpedia	No antonyms from PPDB & WikiSynonyms	-	0.76	-
	No synonyms from WikiSynonyms	-	0.73	-
	All available synonyms & antonyms	-	0.76	-

Table 3.6 – Dependency of DeepAlignment's performance on the external resources' coverage. The reported results for the Schema.org - DBpedia scenario were obtained without recounter-fitting.

synonyms (50% coverage). Finally, we carried out one experiment where we excluded all the synonymy information extracted from WikiSynonyms for the Schema.org - DBpedia scenario. The resulted performance is presented in the rows 1, 4, 2, 5 of Table 3.6, respectively. The reported results provide evidence that the greater the coverage of synonyms and antonyms, the greater the performance of DeepAlignment will be.

3.4 Conclusions

In this chapter, we propose the refinement of pre-trained word vectors with the purpose of deriving ontological entity descriptions which are tailored to the ontology matching task. The refined word representations are learned so that they incorporate domain knowledge encoded in ontologies as well as knowledge extracted from semantic lexicons. Unlike previous approaches that exploited machine learning, the absence of explicit information relevant to the ontology matching task during the refinement process enables DeepAlignment to overcome the small sample size that characterises the problem of ontology matching. We perform ontology matching by applying the Stable Marriage Assignment algorithm over the entities' pairwise distances. Our experimental results demonstrate significant performance gains over the state-of-the-art and show a novel way to study the problem of ontology matching under the setting of NLP.

4 Phrase-level Representation Learning for Ontology Alignment

"Never ask for the meaning of a word in isolation, but only in the context of a sentence."

Gottlob Frege, Introduction to The Foundations of Arithmetic

In the previous chapter, we presented a representation learning based ontology alignment system which by exploiting transfer learning bypassed many of the obstacles that hindered the application of machine learning to the problem of ontology alignment. The system showed state-of-the-art performance demonstrating significant improvements with regard to the recall metric, however, at the cost of a certain amount of degradation in the precision metric. To tackle this shortcoming, we propose to go beyond the retrofitting of single word embeddings by exploiting a novel mechanism that also allows the retrofitting of phrase embeddings. This enables to make use of all the available paraphrase information and, thus, to better distinguish between true cases of semantic similarity and cases of semantic association. Additionally, we propose an outlier detection mechanism that successfully detects misalignments without significantly harming the recall capability of the system. Our results provide evidence that the approach produces embeddings that are especially well tailored to the ontology matching task, while overcoming the previous obstacles.

4.1 Introduction

Ontologies seek to alleviate the Tower of Babel effect by providing standardized specifications of the intended meanings of the terms used in given domains. Formally, an ontology is "a representational artifact, comprising a taxonomy as proper part, whose representations are intended to designate some combinations of universals, defined classes and certain relations between them" [200]. Ideally, in order to achieve a unique specification for each term, ontologies would be built in such a way as to be non-overlapping in their content. In many cases, however, domains have been represented by multiple ontologies and there thus

arises the task of *ontology matching*, which consists in identifying correspondences among entities (types, classes, relations) across ontologies with overlapping content.

Different ontological representations draw on the different sets of natural language terms used by different groups of human experts [63]. In this way, different and sometimes incommensurable terminologies are used to describe the same entities in reality. This issue, known as the *human idiosyncrasy* problem [200], constitutes the main challenge to discovering equivalence relations between terms in different ontologies.

Ontological terms are typically common nouns or noun phrases. According to whether they do or do not include prepositional clauses [253], the latter may be either composite (for example *Neck of femur*) or simple (for example *First tarsometatarsal joint* or just *Joint*). Such grammatical complexity of ontology terms needs to be taken into account in identifying semantic similarity. But account must be taken also of the ontology's axioms and definitions, and also of the position of the terms in the ontology graph formed when we view these terms as linked together through the *is_a* (subtype), *part_of* and other relations used by the ontology.

The primary challenge to identification of semantic similarity lies in the difficulty we face in distinguishing true cases of similarity from cases where terms are merely "semantically associated"[8]. As a concrete example, the word "harness" is semantically associated with the word "horse" because a harness is often used on horses [136]. Yet the two expressions are not semantically similar. The sorts of large ontologies that are the typical targets of semantic similarity identification contain a huge number of such semantically associated term pairs. This difficulty in distinguishing similarity from semantic association is a well-studied problem in both cognitive science [222] and NLP [118].

Traditionally, feature engineering has been the predominant way to approach the ontology matching problem [197]. In machine learning, a feature is an individual measurable property of a phenomenon in the domain being observed [21]. Here we are interested in features of terms, for instance the number of incoming edges when a term is represented as the vertex of an ontology graph; or a terms's tf-idf value – which is a statistical measure of the frequency of a term's use in a corpus [207]. Feature engineering consists in crafting features of the data that can be used by machine learning algorithms in order to achieve specific tasks. Unfortunately determining which hand-crafted features will be valuable for a given task can be highly time consuming. To make matters worse, as Cheatham and Hitzler have recently shown, the performance of ontology matching based on such engineered features varies greatly with the domain described by the ontologies [32].

As a complement to feature engineering, attempts have been made to develop machine-learning strategies for ontology matching based on binary classification [141]. This means a classifier is trained on a set of alignments between ontologies in which correct and incorrect mappings are identified with the goal of using the trained classifier to predict whether an assertion of semantic equivalence between two terms is or is not true. In general, the number of true alignments between two ontologies is several orders of magnitude smaller than the

number of all possible mappings, and this introduces a serious class imbalance problem [140]. This abundance of negative examples hinders the learning process, as most data mining algorithms assume balanced data sets and so the classifier runs the risk of degenerating into a series of predictions to the effect that every alignment comes to be marked as a misalignment.

Both standard approaches thus fail: feature engineering because of the failure of generalization of the engineered features, and supervised learning because of the class imbalance problem. Our proposal is to address these limitations through the exploitation of unsupervised learning approaches for ontology matching drawing on the recent rise of distributed representations (DRs), in which for example words and sentences are embedded in a high-dimensional Euclidean space [37, 149, 150, 176, 129] in order to provide a means of capturing lexical and sentence meaning in an unsupervised manner. The way this works is that the machine learns a mapping from words to high-dimensional vectors which take account of the contexts in which words appear in a plurality of corpora. Vectors of words that appear in the same sorts of context will then be closer together when measured by a similarity function. That the approach can work without supervision stems from the fact that meaning capture is merely a positive externality of context identification, a task that is unrelated to the meaning discovery task.

Traditionally, corpus driven approaches were based on the *distributional hypothesis*, i.e. the assumption that semantically similar or related words appear in similar contexts [93]. This meant that they tended to learn embeddings that capture both similarity (*horse, stallion*) and relatedness (*horse, harness*) reasonably well, but do very well on neither [118, 97]. In an effort to correct for these biases a number of pre-trained word vector refining techniques were introduced [70, 118, 156]. These techniques are however restricted to retrofitting single words and do not easily generalize to the sorts of nominal phrases that appear in ontologies. Wieting et al. [240, 239] make one step towards addressing the task of tailoring phrase vectors to the achievement of high performance on the semantic similarity task by focusing on the task of paraphrase detection. A paraphrase is a restatement of a given phrase that use different words while preserving meaning. Leveraging what are called universal compositional phrase vectors [153] for the purposes of paraphrase detection provides training data for the task of semantic similarity detection which extends the approach from single words to phrases. Unfortunately, the result still fails as regards the problem of distinguishing semantic similarity and semantic association on rare phrases [240] – constantly appearing on ontologies – which thus again harms performance in ontology matching tasks.

In this work, we tackle the aforementioned challenges and introduce a new framework for representation learning based ontology matching. Our ontology matching algorithm is structured as follows: To represent the nouns and noun-phrases in an ontology, we exploit the context information that accompanies the corresponding expressions when they are used both inside and outside the ontology. More specifically, we create vectors for ontology terms on the basis of information extracted not only from natural language corpora but also from terminological and lexical resources and we join this with information captured both explicitly

and implicitly from the ontologies themselves. Thus we capture contexts in which words are used in definitions and in statements of synonym relations. We also draw inferences from the ontological resources themselves, for example to derive statements of semantic association – the absence of a synonymous statement between two terms with closely similar vectors is taken to imply that as a statement of semantic association obtains between them. We then cast the problem of ontology matching as an instance of the Stable Marriage problem [73] discovering in that way terminological mappings in which there is no pair of terms that would rather be matched to each other than their current matched terms. In order to compute the ordering of preferences for each term, that the Stable Marriage problem requires, we use the terminological representations' pairwise distances. We compute the aforementioned distances using the cosine distance over the phrases representations learned by the phrase retrofitting component. Finally, an outlier detection component sifts through the list of the produced alignments so as to reduce the number of misalignments.

Our main contributions in this chapter are: (i) We demonstrate that word embeddings can be successfully harnessed for ontology matching; a task that requires phrase representations tailored to semantic similarity. This is achieved by showing that knowledge extracted from semantic lexicons and ontologies can be used to inscribe semantic meaning on word vectors. (ii) We additionally show that better results can be achieved on the discrimination task between semantic similarity and semantic association, by casting the problem as an outlier detection. To do so, we present a denoising autoencoder architecture, which implicitly tries to discover a hidden representation tailored to the semantic similarity task. To the best of our knowledge, the overall architecture used for the outlier detection as well as its training procedure is applied for the first time to the problem of discriminating among semantically similar and semantically associated terms. (iii) We use the biomedical domain as our application, due to its importance, its ontological maturity, and to the fact that it constitutes the domain with the larger ontology alignment datasets owing to its high variability in expressing terms. We compare our method to state-of-the-art ontology matching systems and show significant performance gains. Our results demonstrate the advantages that representation learning bring to the problem of ontology matching, shedding light on a new direction for a problem studied for years in the setting of feature engineering.

4.2 Methods

We present a representation learning based ontology matching algorithm that approaches the problem as follows. We propose an ontology matching system that is composed of two neural network components which learn which term alignments correspond to semantic similarity. The first component discovers a large amount of true alignments between two ontologies but is prone to errors. The second component corrects these errors. We present below an overview of the two components.

We use the ontologies to generate negative training examples that correspond to semanti-

cally associated examples, and additional knowledge sources to extract paraphrases that will correspond to positive examples of semantic similarity. We use these training data to refine pre-trained word vectors so that they are better suited for the semantic similarity task. This task is accomplished by the first component, which we call *phrase retrofitting* component, that retrofits word vectors so that when they are used to represent sentences, the produced sentence embeddings will be tailored to semantic similarity. We represent each ontological term as the bag of words of its textual description[9] which we complement with the refined word embeddings. We construct sentence representations of the terms' textual description by averaging the phrase's aforementioned word vectors. To inscribe semantic similarity onto the sentence embeddings, we construct an optimization criterion which rewards matchings of semantically similar sentence vectors and penalizes matchings of semantically associated ones. Thus the optimization problem adapts word embeddings so that they are more appropriate to the ontology matching task. Nonetheless, one of the prime motivations of our work comes from the observation that although supervision is used to tailor phrase embeddings to the task of semantic similarity, the problem of discriminating semantically similar vs semantically associated terms is not targeted directly. This lack will lead to the presence of a significant number of misalignments, hindering the performance of the algorithm.

For that reason, we further study the discrimination problem in the setting of unsupervised outlier detection. We use the set of sentence representations produced by the phrase retrofitting component to train an denoising autoencoder [231]. The denoising autoencoder (DAE) aims at deriving a hidden representation that captures intrinsic characteristics of the distribution of semantically similar terms. We force the DAE to leverage new sentence representations by learning to reconstruct not only the original sentence but also its paraphrases, thus boosting the semantic similarity information that the new representation brings. Since we are using paraphrases to do so we bring in additional training data, doing essentially data augmentation for the semantically similar part of the problem. The DAE corresponds to our second component which succeeds in discovering misalignments by capturing intrinsic characteristics of semantically similar terms. We match the entities of two different ontologies using the Stable Marriage algorithm over the terminological embeddings' pairwise distances. We compute the aforementioned distances using the cosine distance. Finally, we iteratively pass through all the produced alignments and we discard those that violate a threshold which corresponds to an outlier condition.

4.2.1 Preliminaries

We introduce some additional notation that we will use throughout this chapter. Let $sen_i = \{w_1^i, w_2^i, \ldots, w_m^i\}$ be the phrasal description[9] of a term i represented as a bag of m word vectors. We compute the sentence representation of the entity i, which we denote s_i, by computing the mean of the set sen_i, as per [153]. Let $s_i, s_j \in \mathbb{R}^d$ be two d-dimensional vectors that correspond to two sentence vectors, we compute their cosine distance as follows: $dis(s_i, s_j) = 1 - cos(s_i, s_j)$. In the following, d will denote the dimension of the pre-trained and retrofitted word vectors.

For $x \in \mathbb{R}$, we denote the *rectifier* activation function as: $\tau(x) = \max(x, 0)$, and the *sigmoid* function as: $\sigma(x) = \frac{1}{1+e^{-x}}$.

4.2.2 Building Sentence Representations

In this section, we describe the neural network architecture that will produce sentence embeddings tailored to semantic similarity. Quite recently several works addressed the challenge of directly optimizing word vectors to produce sentence vectors by averaging the bag of the word vectors [240, 116, 98]. The interplay between semantical and physical intuition is that word vectors can be thought as corresponding to the positions of equally weighted masses, where the center of their masses provides information of the mean location of their semantic distribution. Intuitively, the word vectors' "center of the mass" provide a means for measuring where the semantic content primarily "concentrates". Despite the fact that vector addition is insensitive to word order [153], it has been proven that this syntactic agnostic operation provides results that compete favorably with more sophisticated syntax-aware composing operations [98]. We base our phrase retrofitting architecture on an extension of the Siamese CBOW model [116]. The fact that Siamese CBOW provides a native mechanism for discriminating between sentence pairs from different categories explains our choice to build upon this architecture.

Siamese CBOW is a log-linear model aiming at predicting a sentence from its adjacent sentences; addressing the research question whether directly optimizing word vectors for the task of being averaged leads to better suited word vectors for this task compared to word2vec [150]. Let $V = \{v_1, v_2, \dots v_N\}$ be an indexed set of word vectors of size N. The Siamese CBOW model transforms a pre-trained vector set V into a new one, $V' = \{v'_1, v'_2, \dots v'_N\}$, based on two sets of positive, S_i^+, and negative, S_i^-, constraints for a given training sentence s_i. The supervised training criterion in Siamese CBOW rewards co-appearing sentences while penalizing sentences that are unlikely to appear together. Sentence representations are computed by averaging the sentence's constituent word vectors. The reward is given by the pairwise sentence cosine similarity over their learned vectors. Sentences which are likely to appear together should have a high cosine similarity over their learned representations. In the initial paper of Siamese CBOW [116], the set S_i^+ corresponded to sentences appearing next to a given s_i, whereas S_i^- corresponded to sentences that were not observed next to s_i.

Since we want to be able to differentiate between semantically similar and semantically associated sentences we let the sets S_i^+ and S_i^- to be sentences that are semantically similar and semantically associated to a given sentence s_i. In the rest of the section we revise the main elements of the Siamese CBOW architecture and describe the modifications we performed in order to exploit it for learning sentence embeddings that reflect semantic similarity. To take advantage of the semantic similarity information already captured in the initial word vectors, an important characteristic as demonstrated in various word vectors retrofitting techniques [70, 156, 240], we use *knowledge distillation* [99] to penalize large changes in the learned word

vectors with regard to the pre-trained ones.

Our paraphrase retrofitting model retrofits a pre-trained set of word vectors with the purpose of leveraging a new set V' by solving the following optimization problem:

$$\min_{V'} \kappa_S L_S(V') + \kappa_{LD} L_{KD}(V, V'), \tag{4.1}$$

where k_S and k_{LD} are hyperparameters controlling the effect of $L_S(V')$ and $L_{KD}(V, V')$ losses, accordingly. The $L_S(V')$ term is defined as $\frac{1}{N}\sum_{i=1}^{N} L_{S_i}$, where N denotes the number of the training examples. The L_{S_i} term corresponds to categorical cross-entropy loss defined as:

$$L_{S_i} = -\sum_{s_j \in \{S_i^+ \cup S_i^-\}} p(s_i, s_j) \cdot \log(p_\theta(s_i, s_j)), \tag{4.2}$$

where $p(\cdot)$ is the target probability the network should produce, and $p_\theta(\cdot)$ is the prediction it estimates based on parameters θ, using Equation 4.4. The target distribution simply is:

$$p(s_i, s_j) = \begin{cases} \frac{1}{|S^+|}, & \text{if } s_j \in S_i^+ \\ 0, & \text{if } s_j \in S_i^-. \end{cases} \tag{4.3}$$

For instance, if there are two positive and two negative examples, the target distribution is (0.5, 0.5, 0, 0). For a pair of sentences (s_i, s_j), the probability $p_\theta(s_i, s_j)$ is constructed to reflect how likely it is for the sentences to be semantically similar, based on the model parameter θ. The probability $p_\theta(s_i, s_j)$ is computed on the training data set based on the softmax function as follows:

$$p_\theta(s_i, s_j) = \frac{e^{\left(\cos(\mathbf{s_i^\theta}, \mathbf{s_j^\theta})\right)^{1/T}}}{\sum_{s_k \in \{S_i^+ \cup S_i^-\}} e^{\left(\cos(\mathbf{s_i^\theta}, \mathbf{s_k^\theta})\right)^{1/T}}}, \tag{4.4}$$

where s_x^θ denotes the embedding of sentence s_x, based on the model parameter θ. To encourage the network to better discriminate between semantically similar and semantically associated terms, we extend the initial architecture by introducing the parameter T. The parameter T, named *temperature*, is based on the recent work of [99, 135]. Hinton et al. [99] suggest that setting $T > 1$ increases the weight of smaller logit (the inputs of the softmax function) values, enabling the network to capture information hidden in small logit values.

To construct the set S^+, we extract pairs of synonyms from semantic lexicons and the ontologies themselves. To construct the set S^-, we sample a set of semantically associated terms from the ontologies to be matched. Given a sentence s_i, we compute its cosine distance with every term from the two ontologies to be matched, based on the initial pre-trained word vectors. Thereafter, we choose the n terms demonstrating the smaller cosine distance to be the negative examples. To account for that fact that among these n terms there may be a possible alignment, we exclude the n_* closest terms. Equivalently, given the increasingly sorted sequence of the cosine distances, we choose the terms in index positions starting from n_* up to $n + n_*$. For computational efficiency, we carry this process out only once before the

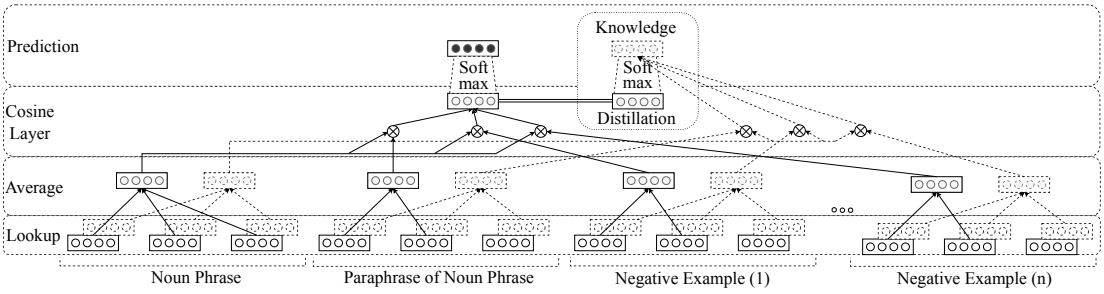

Figure 4.1 – Phrase Retrofitting architecture based on a Siamese CBOW network [116] and Knowledge Distillation [99]. The input projection layer is omitted.

training procedure starts.

Hinton et al. [99] found that using the class probabilities of an already trained network as "soft targets" for another one network constitutes an efficient way of communicating already discovered regularities to the latter network. We exploit, thus, knowledge distillation to emit the original semantic information captured in the pre-trained word vectors to the new ones leveraged by Siamese CBOW. Therefore, we add the Knowledge Distillation loss $L_{KD}(V, V') = \frac{1}{N}\sum_{i=1}^{N} L_{KD_i}$ to the initial Siamese CBOW's loss. The L_{KD_i} term:

$$L_{KD_i} = -\sum_{s_j \in \{S^+ \cup S^-\}} p_{\theta_I}(s_i, s_j) \cdot \log(p_\theta(s_i, s_j)), \tag{4.5}$$

is defined as the categorical cross-entropy between the probabilities obtained with the initial parameters (i.e. θ_I) and the ones with parameters θ.

Based on the observations of Hinton et al. [99], these "soft targets" act as an implicit regularization, guiding the Siamese CBOW's solution closer to the initial word vectors. We would like to highlight that we experimented with various regularizers, such as the ones presented in the works of [70, 239, 156, 34], however, we obtained worse results than the ones reported in our experiments. Figure 4.1 summarizes the overall architecture of our phrase retrofitting model. The dashed rectangles in the *Lookup Layer* correspond to the initial word vectors, which are used to encourage the outputs of the Siamese CBOW network to approximate the outputs produced with the pre-trained ones in every epoch. The word embeddings are averaged in the next layer to produce sentence representations. The cosine similarities between the sentence representations are calculated in the penultimate layer and are used to feed a softmax function so as to produce a final probability distribution. Specifically, we compute the cosine similarity between the sentence representation of the noun phrase and the sentence representations of every positive and negative example of semantic similarity. In the final layer, this probability distribution is used to compute two different categorical cross entropy losses. The left loss encourages the probability distribution values to approximate a target distribution, while the right one penalizes large changes in the learned word vectors with regard to the pre-trained ones. The double horizontal lines in the *Cosine Layer* highlight that these rectangles denote in

fact the same probability distribution, computed in the penultimate layer.

4.2.3 Outlier Detection

The extension of the Siamese CBOW network retrofits pre-trained word vectors to become better suited for constructing sentence embeddings that reflect semantic similarity. Although we sample appropriate negative examples (i.e., semantically associated terms) from the ontologies to be matched, we will never have all the negative examples needed. Moreover, allowing a larger number, n, of negative examples increases the computation needed making it inefficient. We depart from these problems by further casting the problem of discriminating between semantically similar and related terms as an outlier detection. To leverage an additional set of sentence representations more robust to semantic similarity, we use the hidden representation of a Denoising Autoencoder (DAE) [231].

The Siamese CBOW network learns to produce sentence embeddings of ontological terms that are better suited for the task of semantic similarity. We now use the learned sentence vectors to train a DAE. We extend the *standard* architecture of DAEs to reconstruct not only the sentence representation fed as input but also paraphrases of that sentence. Our idea is to improve the sentence representations produced by the Siamese CBOW and make them more robust to paraphrase detection. At the same time, this constitutes an efficient data augmentation technique; very important in problems with relatively small training data sets.

We train the autoencoder once the training of the Siamese CBOW network has been completed. Even if layer-wise training techniques [19] are outweighed nowadays by end-to-end training, we decide to adopt this strategy for two reasons. Firstly, we aim to capture with the DAE intrinsic characteristics of the distribution of the semantically similar terms. DAEs have been proven to really capture characteristics of the data distribution, namely the derivative of the log-density with respect to the input [6]. However, training the DAE on a dataset that does not reflect the true distribution of semantically similar terms introduces surely a barrier to our attempt. Therefore, we leverage in advance sentence representations, through the Siamese CBOW network, more robust to semantic similarity; an action that allows the DAE to act on a dataset with significantly less noise and less bias. Secondly, combining the extended Siamese CBOW architecture together with the DAE and training them end-to-end significantly increases the number of the training parameters. This increase is a clear impediment to a problem lacking an oversupply of training data.

Let $x, y \in \mathbb{R}^d$ be two d-dimensional vectors, representing the sentence vectors of two paraphrases. Our target is not only to reconstruct the sentence representation from a corrupted version of it, but also to reconstruct a paraphrase of the sentence representation based on the partially corrupted one. The corruption procedure, used for regularising the autoencoder, that we followed in our experiments is the following: for each input x, a fixed number of $\lceil \nu d \rceil$ $(0 < \nu < 1)$ components are chosen at random, and their value is forced to 0, while the others are left untouched. The corrupted input $\tilde{x}$ is then mapped, as with the basic autoencoder,

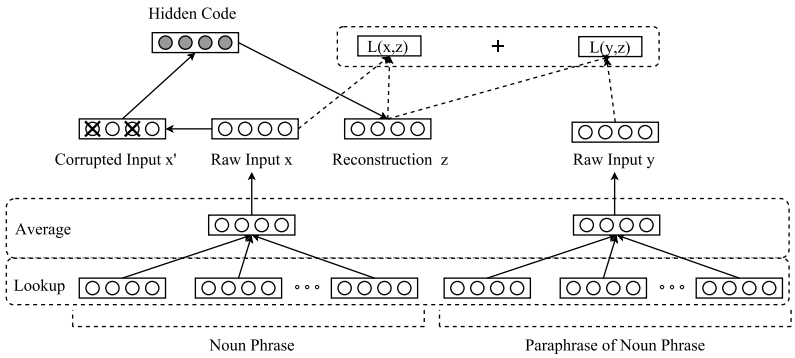

Figure 4.2 – Autoencoder architecture for outlier detection.

to a hidden representation $h = \tau(W\tilde{x} + b)$ from which we reconstruct a $z = \sigma(W'h + b')$. The dimension d_h of the hidden representation $h \in \mathbb{R}^{d_h}$ is treated as a hyperparameter. Similar to the work in [231], the parameters are trained to minimize, over the training set, an average reconstruction error. However, we aim not only to reconstruct the initial sentence but also its paraphrases. For that reason, we use the following reconstruction loss: $L(x,z) + L(y,z) =$

$$\begin{aligned} &= -\sum_{k=1}^{d} [x_k \log z_k + (1 - x_k) \log(1 - z_k)] \\ &\quad -\sum_{k=1}^{d} [y_k \log z_k + (1 - y_k) \log(1 - z_k)]. \end{aligned} \tag{4.6}$$

The x_k, z_k, y_k correspond to the Cartesian coordinates of vectors x, z and y, respectively. The overall process is depicted in Figure 4.2. In this figure, the Lookup and Average layers are similar to the ones depicted in Figure 4.1. A sentence representation x is corrupted to $\tilde{x}$. The autoencoder maps it to h (i.e., the hidden code) and attempts to reconstruct both x and the paraphrase embedding y.

4.2.4 Ontology Matching

The two components that we have presented were build in such a way so that they learn sentence representations which try to disentagle semantic similarity and semantic association. We will now use these representations to solve the ontology matching problem. To align the entities from two different ontologies, we use the extension of the Stable Marriage Assignment problem to unequal sets [73, 145]. This extension of the stable marriage algorithm computes $1-1$ mappings based on a preference $m \times n$ matrix, where m and n is the number of entities in ontologies O and O', respectively. In our setting, a matching is not stable if: (i) there is an element $e_i \in O$ which prefers some given element $e_j \in O'$ over the element to which e_i is already matched, and (ii) e_j also prefers e_i over the element to which e_j is already matched. These properties of a stable matching impose that it does not exist any match (e_i, e_j) by which both e_i and e_j would be individually matched to more similar entities compared to the entities

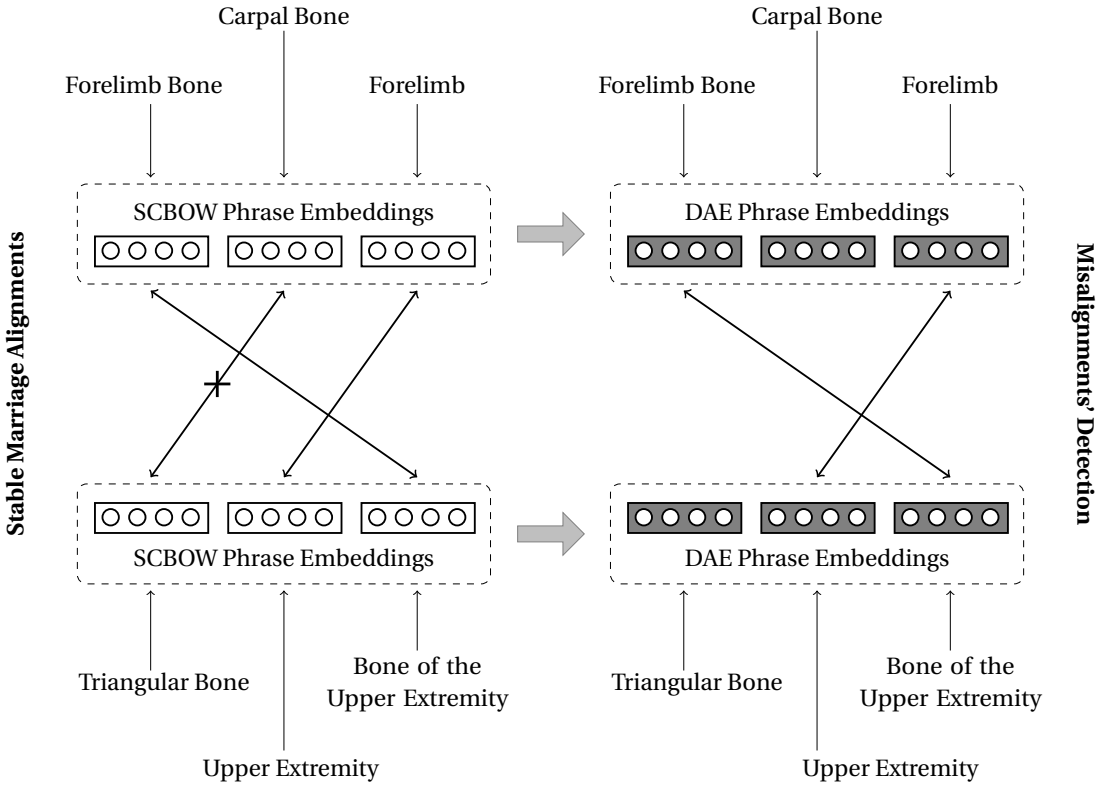

Figure 4.3 – Overall proposed ontology matching architecture.

to which they are currently matched. This leads to a significant reduction in the number of misalignments due to semantic association, provided that the learned representations do reflect the semantic similarity.

The steps of our ontology matching algorithm are the following: We represent each ontological term as the bag of words of its textual description, which we complement with the refined word vectors produced by the phrase retrofitting component. In the next step, we construct phrase embeddings of the terms' textual description[9] by averaging the phrase's word vectors. We cast the problem of ontology matching as an instance of the Stable Marriage problem using the entities' semantic distances. We compute these distances using the cosine distance over the sentences vectors. We iteratively pass through all the produced alignments and we discard those with a cosine distance greater than a certain threshold, t_1. These actions summarize the work of the first component. Note that the violation of the triangle inequality by the cosine distance is not an impediment to the Stable Marriage algorithm [73].

In the next step, we create an additional set of phrase vectors by passing the previously constructed phrase vectors through the DAE architecture. Based on this new embedding's set, we iteratively pass through all the alignments produced in the previous step and we discard those that report a threshold violation. Specifically, we discard those that exhibit a cosine distance, computed over the vectors produced by the DAE, greater than a threshold t_2.

This corresponds to the final step of the outlier detection process as well as of our ontology matching algorithm. The overall ontology matching procedure is illustrated in Figure 4.3.

4.3 Results & Discussion

In this section, we present the experiments we performed on biomedical evaluation benchmarks coming from the Ontology Alignment Evaluation Initiative (OAEI), which organizes annual campaigns for evaluating ontology matching systems. We have chosen the biomedical domain for our evaluation benchmarks owing to its ontological maturity and to the fact that its language use variability is exceptionally high [148]. At the same time, the biomedical domain is characterized by rare words and its natural language content is increasing at an extremely high speed, making hard even for people to manage its rich content [233]. To make matters worse, as it is difficult to learn good word vectors for rare words from only a few examples [194], their generalization on their ontology matching task is questionable. This is a real challenge for domains, such as the biomedical, the industrial, etc, in which existence of words with rare senses is typical. The existence of rare words makes the presence of the phrase retroffiting component crucial to the performance of our ontology alignment framework.

4.3.1 Biomedical Ontologies

We give a brief overview of the four ontologies used in our ontology mapping experiments. Two of them (the Foundational Model of Anatomy and the Adult Mouse anatomical ontologies) are pure anatomical ontologies, while the other two (SNOMED CT and NCI Thesaurus) are broader biomedical ontologies of which anatomy consists a subdomain that they describe [253]. Although more recent versions of these resources are available, we refer to the versions that appear in the Ontology Alignment Evaluation Initiative throughout this work in order to facilitate comparisons across the ontology matching systems.

Foundational Model of Anatomy (FMA): is an evolving ontology that has been under development at the University of Washington since 1994 [187, 170]. Its objective is to conceptualize the phenotypic structure of the human body in a machine readable form.

Adult Mouse Anatomical Dictionary (MA): is a structured controlled vocabulary describing the anatomical structure of the adult mouse [95].

NCI Thesaurus (NCI) provides standard vocabularies for cancer [45] and its anatomy subdomain describes naturally occurring human biological structures, fluids and substances.

SNOMED Clinical Terms (SNOMED): is a systematically organized machine readable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting [56].

4.3.2 Semantic Lexicons

We provide below some details regarding the procedure we followed in order to construct pairs of semantically similar phrases. Let $(word_1^1, word_2^1, \ldots, word_m^1)$, be a term represented as a sequence of m words. The strategy that we have followed in order to create the paraphrases is the following: We considered all the contiguous subsequences of this term. Namely, we considered all the possible contiguous subsequences of the form: $(word_i^1, word_{(i+1)}^1, \ldots, word_j^1)$, $\forall i, j \in \mathbb{N} : 0 \leq i \leq j \leq m$. Based on these contiguous subsequences, we queried the semantic lexicons for paraphrases. Below we give a brief summary of the semantic lexicons that we used in our experiments:

ConceptNet 5: a large semantic graph that describes general human knowledge and how it is expressed in natural language [208]. The scope of ConceptNet includes words and common phrases in any written human language.

BabelNet: a large, wide-coverage multilingual semantic network [160, 161]. BabelNet integrates both lexicographic and encyclopedic knowledge from WordNet and Wikipedia.

WikiSynonyms: a semantic lexicon which is built by exploiting the Wikipedia redirects to discover terms that are mostly synonymous [42].

Apart from the synonymy relations found in these semantic lexicons, we have exploited the fact that in some of the considered ontologies, a type may have one preferred name and some additional paraphrases [253], expressed through multiple rdfs:label relations.

4.3.3 Training

We tuned the hyperparameters on a set of 1000 alignments which we generated by subsampling the SNOMED-NCI ontology matching task.[10] We chose the vocabulary of the 1000 alignments so that it is disjoint to the vocabulary that we used in the alignment experiments, described in the evaluation benchmarks, in order to be sure that there is no information leakage from training to testing. We tuned to maximize the $F1$ measure. We trained with the following hyperparameters: word vector has size (d) 200 and is shared across everywhere. We initialized the word vectors from word vectors pre-trained on a combination of PubMed and PMC texts with texts extracted from a recent English Wikipedia dump [181]. All the initial out-of-vocabulary word vectors are sampled from a normal distribution ($\mu = 0, \sigma^2 = 0.01$). The resulted hyperparameters controlling the effect of retrofitting k_S and knowledge distillation k_{LD} were 10^6 and 10^3, accordingly. The resulted size of the DAE hidden representation (d_h) is 32 and v is set to 0.4. We used $T = 2$ according to a grid search, which also aligns with the authors' recommendations [99]. For the initial sampling of semantically associated terms, we used: $n_* = 2$ and $n = 7$. The best resulted values for the thresholds were the following: $t_1 = t_2 = 0.2$. The phrase retrofitting model was trained over 15 epochs using the Adam optimizer [119] with a learning rate of 0.01 and gradient clipping at 1. The DAE was trained over 15 epochs using the Adadelta optimizer [251] with hyperparameters $\epsilon = 1e{-}8$ and $\rho = 0.95$.

4.3.4 Evaluation Benchmarks

We provide some details regarding the respective size of each ontology matching task in Table 4.1. The reference alignment of the MA - NCI matching scenario is based on the work of Bodenreider et al. [23]. To represent each ontological term for this task, we used the unique rdfs:label that accompanies every type in the ontologies. The alignment scenarios between FMA - NCI and FMA - SNOMED are based on a small fragment of the aforementioned ontologies. The reference alignments of these alignment scenarios are based on the UMLS Metathesaurus [22], which currently consists the most comprehensive effort for integrating independently developed medical thesauri and ontologies. To represent each ontological term for these tasks, we exploited the textual information appearing on the rdf:about tag that accompanies every type in the ontologies. We did not use the rdf:about tag on the MA - NCI matching scenario, since their rdf:about tags provide a language agnostic unique identifier with no direct usable linguistic information. We would like to note that since the Stable Marriage algorithm provides one-to-one correspondences, we have only focused on discovering one-to-one matchings. In addition, a textual preprocessing that we performed led a small number of terms to degenerate into a single common phrase. This preprocessing includes case-folding, tokenization, removal of English stopwords and words coappearing in the vast majority of the terms (for example the word “structure” in SNOMED). Thereafter, we present in Table 4.1 the number of one-to-one types’ equivalences remained after the preprocessing step.

Ontology Matching between:				#Matchings
Ontology I	#Types	Ontology II	#Types	
MA	2744	NCI	3304	1489
FMA	3696	NCI	6488	2504
FMA	10157	SNOMED	13412	7774

Table 4.1 – Respective sizes of the ontology matching tasks.

Last but not least, it is of significant importance to highlight that the reference alignments based on UMLS Metathesaurus will lead to an important number of logical inconsistencies [108, 3]. As our method does not apply reasoning, whether it produces or not incoherence-causing matchings is a completely random process. In our evaluation, we have chosen to also take into account *incoherence-causing* mappings. However, various concerns can be raised about the fairness of comparing against ontology matching systems that make use of automated alignment repair techniques [178, 3]. For instance, the state-of-the-art systems AML [67, 68], LogMap and LogMapBio [106], which are briefly described in the next section, do employ automated alignment repair techniques. Our approach to use the original and incoherent mapping penalizes these systems that perform additional incoherence checks.

Nonetheless, our choice to include inconsistence mappings can be justified in the following way. First, it is a direct consequence of the fact that we approach the problem of ontology

matching from the viewpoint of discovering semantically similar terms. A great number of these inconsistent mappings do correspond to semantically similar terms. Second, we believe that ontology matching can also be used as a curation process during the ontological (re)design phase so as to alleviate the possibility of inappropriate terms' usage. The fact that two distinct truly semantically similar terms from two different ontologies lead to logical inconsistencies during the integration phase can raise an issue for modifying the source ontology [108]. Third, although ontologies constitute a careful attempt to ascribe the intended meaning of a vocabulary used in a target domain, they are error prone as every human artifact. Incoherence check lays on the assumption that both of the ontologies that are going to be matched are indeed error-free representational artifacts. We decided not to make this assumption.

Therefore, we have chosen to treat even the systems that employ automated alignment repair techniques error-prone. For that reason, we considered appropriate to report the performance of the aforementioned systems on the complete reference alignment in the next section. Nevertheless, we refer the reader to the [3] for details on the performance of these systems on incoherence free subsets of the reference alignment set. Under the assumption that the ontologies to be matched are error-free, it can be observed that the automated alignment repair mechanisms of these systems are extremely efficient; a fact that demonstrates the maturity and the robustness of these methods.

4.3.5 Experimental Results

Table 4.2 shows the performance of our algorithm compared to the six top performing systems on the evaluation benchmarks, according to the results published in OAEI Anatomy track (MA - NCI) and in the Large BioMed track (FMA-NCI, FMA-SNOMED).[6] To check for the statistical significance of the results, we used the procedure described in [250]. The systems presented in Table 4.2 starting from the top of the table up to and including LogMapBio fall into the category of feature engineering.[11] CroMatcher [88], AML [67, 68], XMap [53] perform ontology matching based on heuristic methods that rely on aggregation functions. FCA_Map [255, 256] uses Formal Concept Analysis [241] to derive terminological hierarchical structures that are represented as lattices. The matching is performed by aligning the constructed lattices taking into account the lexical and structural information that they incorporate. LogMap and LogMapBio [106] use logic-based reasoning over the extracted features and cast the ontology matching to a satisfiability problem. Some of the systems compute many-to-many alignments between ontologies. For a fair comparison of our system with them, we have also restricted these systems in discovering one-to-one alignments. We excluded the results of XMap for the Large BioMed track, because it uses synonyms extracted by the UMLS Metathesaurus. Systems that use the UMLS Metathesaurus as background knowledge will have a notable advantage since the Large BioMed track's reference alignments are based on it.

We describe in the following the procedure that we followed in order to evaluate the perfor-

System	MA - NCI			FMA-NCI			FMA-SNOMED		
	P	R	F1	P	R	F1	P	R	F1
AML	0.943	0.94	**0.941**	0.908	0.94	0.924	<u>0.938</u>	0.784	0.854
CroMatcher	0.942	0.912	0.927	-	-	-	-	-	-
XMap	0.924	0.877	0.9	-	-	-	-	-	-
FCA_Map	0.922	0.841	0.880	0.89	0.947	0.918	0.918	0.857	0.886
LogMap	0.906	0.850	0.878	0.894	0.930	0.912	0.933	0.721	0.814
LogMapBio	0.875	0.900	0.887	0.88	0.938	0.908	0.93	0.727	0.816
Wieting	0.804	0.879	0.839	0.840	0.857	0.849	0.867	0.851	0.859
Wieting+DAE(O)	0.952	0.871	0.909	0.909	0.851	0.879	0.929	0.832	0.878
SCBOW	0.847	0.917	0.881	0.899	0.895	0.897	0.843	0.866	0.855
SCBOW+DAE(O)	<u>0.968</u>	0.913	0.94	<u>0.976</u>	0.892	**0.932**	0.931	0.856	**0.892**

Table 4.2 – Performance of ontology matching systems across the different matching tasks. Bold and underlined numbers indicate the best $F1$-score and the best precision on each matching task, respectively.

mance of the various ontology matching systems. Since the incoherence-causing mappings were also taken into consideration, all the mappings marked as "?" in the reference alignment were considered as positive. To evaluate the discovery of one-to-one matchings, we clustered all the m-to-n matchings and we counted only once when any of the considered systems discovers any of the m-to-n matchings. Specifically, let $T = \{(e, =, e') \mid e \in O, e' \in O'\}$ be a set of clustered m-to-n matchings. Once an ontology matching system discovers for the first time a $(e, =, e') \in T$, we increase the number of the discovered alignments. However, whenever the same ontology matching system discovers an additional $(e_*, =, e'_*) \in T$, where $(e, =, e') \neq (e_*, =, e'_*)$, we did not take this discovered matching into account. Finally, to evaluate the performance of AML, CroMatcher, XMap, FCA_MAP, LogMap, and LogMapBio, we used the alignments provided by OAEI 2016[6] and applied the procedure described above to get their resulted performance.

To explore the performance details of our algorithm, we report in Table 4.2 its performance results with and without outlier detection. Moreover, we included experiments in which instead of training word embeddings based on our extension of the Siamese CBOW, we have used the optimization criterion presented in [239] to produce an alternative set of word vectors. As before, we present experiments on which we exclude our outlier detection mechanism and experiments on which we allow it.[12] We present these experiments under the listings: SCBOW, SCBOW+DAE(O), Wieting, Wieting+DAE(O), accordingly.

SCBOW+DAE(O) is the top performing algorithm in two of the three ontology mappings tasks (FMA-NCI, FMA-SNOMED); in these two its F1 score is significantly better than that of all the other algorithms. In MA-NCA its F1 score is similar to AML, the best system there, but the performance difference is statistically significant. At the same time, SCBOW+DAE(O) achieves the highest precision on two out of three ontology matching tasks. In terms of recall, SCBOW+DAE(O) demonstrates lower performance in the ontology matching tasks.

However, we would like to note that we have not used any semantic lexicons specific to the biomedical domains compared to the other systems. For instance, AML uses three sources of biomedical background knowledge to extract synonyms. Specifically, it exploits the Uber Anatomy Ontology (Uberon), the Human Disease Ontology (DOID), and the Medical Subject Headings (MeSH). Hence, our reported recall can be explained due to the lower coverage of biomedical terminology in the semantic lexicons that we have used. Our motivation for relying only on domain-agnostic semantic lexicons[13] stems from the fact that our intention is to create an ontology matching algorithm applicable to many domains. The success of these general semantic lexicons for such a rich in terminology domain, provides additional evidence that the proposed methodology may also generalize to other domains. However, further experimentation is needed to verify the adequacy and appropriateness of these semantic lexicons to other domains. It is among our future directions to test the applicability of our proposed algorithm to other domains.

Comparing the recall of SCBOW and SCBOW+ DAE(O), we see that the incorporation of the DAE produces sentence embeddings that are tailored to the semantic similarity task.[14] The small precision of SCBOW, in all experiments, indicates a semantic similarity and semantic association coalescence. Considering both the precision and the recall metric, we can observe that the outlier detection mechanism identifies misalignments while preserving most of the true alignments. This fact provides empirical support on the necessity of the outlier detection. To validate the importance of our phrase retrofitting component, we further analyze the behavior of aligning ontologies based on the word embedding produced by running the procedure described in [239] (listed as Wieting). As we can see SCBOW achieves statistically significant higher recall than Wieting in all our experiments and in two of the three cases statistically significant greater precision. This behavior indicates the superiority of SCBOW in injecting semantic similarity to word embeddings as well as to produce word vectors tailored to the ontology matching task. We further extended the Wieting experiment by applying our outlier detection mechanism trained on the word vectors produced by the procedure described in [239]. It can be seen that this extension leads to the same effects as the ones summarized in the SCBOW - SCBOW+DAE(O) comparison. These results give evidence that our DAE-based outlier detection component constitutes a mechanism applicable to various sentence embeddings' producing architectures.

4.3.6 Ablation Study

In this section, an ablation study is carried out to investigate the necessity of each of the described components, as well as their effect on the ontology matching performance. Figure 4.4 shows a feature ablation study of our method. In Table 4.3, we give the descriptions of the experiments. We conducted experiments on which the phrase retrofitting component was not used, hence the ontology matching task was only performed based on the pre-trained word vectors. Moreover, we have experimented on performing the ontology matching task with the features generated by the DAE. Our prime motivation was to test whether the features

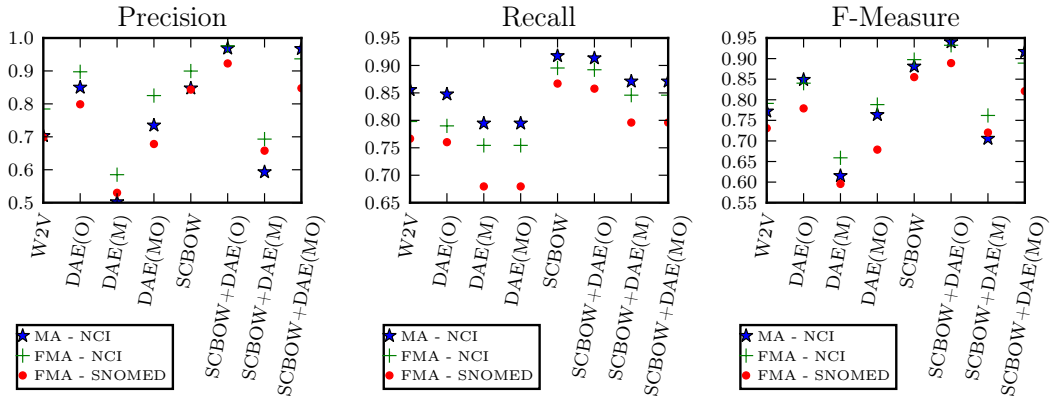

Figure 4.4 – Feature ablation study of our proposed approach across all the experimental ontology matching tasks.

produced by the DAE could be used to compute the cosine distances needed for estimating the preference matrix used by the Stable Marriage's algorithm. Hence, we differentiate in this subsection and we allow the DAE features to be used for Matching and/or Outlier Detection.

To begin with, it can be observed that all the performance metrics' figures undergo the same qualitative behavior. This result demonstrates that our algorithm exhibits a consistent behavior under the ablation study across all the experiments, which constitutes an important factor for inducing conclusions from experiments. The experiment W2V gives the results of executing the algorithm without the phrase retrofitting process, just by providing the pre-trained word vectors [181]. The performance of W2V in terms of Precision/Recall is systematically lower compared to all cases in which the initial word2vec vectors are retrofitted. These results support the importance of the phrase retrofitting process (experiments of which are presented under the listing SCBOW in Figure 4.4), which succeeds in tailoring the word embeddings to the ontology matching task. The pre-trained word vectors, even though they were trained on PubMed and PMC texts, retain small precision and recall. This fact indicates a semantic similarity and semantic association coalescence and sheds light on the importance of the

Experiment's Code:	Phrase Retrofitting	DAE Features:	
		Matching	Outlier Detection
W2V	-	-	-
DAE(O)	-	-	✓
DAE(M)	-	✓	-
DAE(MO)	-	✓	✓
SCBOW	✓	-	-
SCBOW+DAE(O)	✓	-	✓
SCBOW+DAE(M)	✓	✓	-
SCBOW+DAE(MO)	✓	✓	✓

Table 4.3 – Ablation study experiment's listings.

Terminology to be matched	Matching based on SCBOW	Matching based on word2vec
MA-NCI		
gastrointestinal tract	digestive system	respiratory tract
tarsal joint	carpal tarsal bone	metacarpo phalangeal joint
thyroid gland epithelial tissue	thyroid gland medulla	prostate gland epithelium
FMA-NCI		
cardiac muscle tissue	heart muscle	muscle tissue
set of carpal bones	carpus bone	sacral bone
white matter of telencephalon	brain white matter	white matter
FMA-SNOMED		
zone of ligament of ankle joint	accessory ligament of ankle joint	entire ligament of elbow joint
muscle of anterior compartment of leg	compartment of lower leg	entire interosseus muscle of hand
dartos muscle	dartos layer of scrotum	tendon of psoas muscle

Table 4.4 – Sample misalignments produced by aligning ontologies using either SCBOW or word2vec vectors.

retrofitting procedure.

Training the DAE on the pre-trained word vectors - DAE(O) - adds a significant performance gain on precision, which witnesses the effectiveness of the architecture for outlier detection. However, DAE(O)'s precision is almost the same as the one presented in the SCBOW experiment. Only when the phrase retrofitting component is combined with the DAE for outlier detection - SCBOW+DAE(O) - we manage to surpass the aforementioned precision value and achieve our best $F1$-score. Finally, our experiments on aligning ontologies by only using the DAE features demonstrate that these features are inadequate for this task. One prime explanation of this behavior is that DAE features are only exposed to synonymy information. At the same time, the dimensionality reduction of DAE features may lead them to lose a lot of valuable information captured in them for discriminating between semantically similar and semantically associated terms. Note also that the preference matrix required by the Stable Marriage solution requires each term of an ontology O to be compared across all the possible terms of another ontology O'. Thereafter, the vectors based on which the preference matrix will be computed need to capture the needed information adequate for discriminating between semantically similar and semantic associated terms.

4.3.7 Error Analysis

Recent studies provide evidence that different sentence representations objectives yield different intended representation preferable for different intended applications [98]. Moreover, our results reported in Table 4.2 on aligning ontologies with word vectors trained based on the method presented in [239] provide further evidence in the same direction. In Table 4.4, we demonstrate a sample of misalignments produced by aligning ontologies using the Stable Marriage's solution based on a preference matrix computed either on SCBOW or word2vec vectors. It can be seen that the SCBOW misalignments demonstrate even a better spatial

consistency compared to the word2vec misalignments. This result combined with high $F1$-score reported in the SCBOW results in Table 4.4 show that ontological knowledge can be an important ally in the task of harnessing terminological embeddings tailored to semantic similarity. Moreover, this error analysis provides additional support for the significance of retrofitting general-purpose word embeddings before being applied in a domain-specific setting. It can be observed that general-purpose word vectors capture both similarity and relatedness reasonably well, but neither perfectly as it has been already observed in various works [118, 97].

4.3.8 Runtime Analysis

In this section, we report the runtimes of our ontology matching algorithm for the different matching scenarios. Since our method – SCBOW+DAE(O) – consists of three major steps, we present in Table 4.5 the time devoted to each of them as well as their sum. In brief, the steps of our algorithm are the following: the training of the phrase retrofitting component (Step 1), the solution to the stable marriage assignment problem (Step 2), and finally the training of the DAE-based outlier detection mechanism (Step 3). All the reported experiments were performed on a desktop computer with an Intel® Core™ i7-6800K (3.60GHz) processor with 32GB RAM and two NVIDIA® GeForce® GTX™ 1080 (8GB) graphic cards. The implementation was done in Python using Theano [20, 17].

Matching Task	Running Time (seconds)			
	Step 1	Step 2	Step 3	Total
MA - NCI	337	34	36	407
FMA - NCI	490	82	40	612
FMA - SNOMED	609	490	41	1140

Table 4.5 – Runtimes of the steps in the proposed algorithm.

As it can be seen on Table 4.5, the majority of the time is allotted to the training of the phrase retrofitting framework. In addition, it can be observed that the training overhead of the outlier detection mechanism is significantly smaller compared to the other steps. However, one important tendency can be observed in the FMA - SNOMED matching scenario. Specifically, the runtime of the second step has considerably increased and is comparable to the runtime of the first step. This can be explained by the worst-case time complexity of the McVitie and Wilson's algorithm [145], that has been used, which is $\mathcal{O}(n^2)$. Moreover, the computation of the preference matrix required for the defining the stable marriage assignment problem's instance has worst-case time complexity $\mathcal{O}(n^2)$. At the same time, the space complexity of the second step is $\mathcal{O}(n^2)$, since it requires the storage of the preference matrices. On the contrary, various techniques [130, 47] and frameworks [20, 17, 1, 173] have been proposed and implemented for distributing the training and inference task of DRs. Although our implementation exploits these highly optimized frameworks for DRs, the choice of using the McVitie and Wilson's

System	Training Data	MA - NCI			FMA - NCI			FMA - SNOMED		
		P	R	F1	P	R	F1	P	R	F1
SCBOW	SL	0.845	0.911	0.877	0.897	0.840	0.868	0.795	0.773	0.784
SCBOW	SL + AS	0.847	0.917	0.881	0.899	0.895	0.897	0.843	0.866	0.855
SCBOW + DAE(O)	SL	0.946	0.905	0.925	0.972	0.830	0.895	0.912	0.759	0.829
SCBOW + DAE(O)	SL + AS	0.968	0.913	0.94	0.976	0.892	0.932	0.931	0.856	0.892

Table 4.6 – Proposed algorithm's performance in relation to the used synonymy information sources. SL denotes the setting where the used synonyms only come from ConceptNet 5, BabelNet, and WikiSynonyms, whereas AS denotes the setting where the additional synonyms found in the ontologies to be matched have also been used.

algorithm introduces a significant performance barrier for aligning larger ontologies than the ones considered in our experiments. However, it was recently shown that a relationship exists between the class of computing greedy weighted matching problems and the stable marriage problems [139]. The authors exploit this strong relationship to design scalable parallel implementations for solving large instances of the stable marriage problems. It is among our future work to test the effectiveness of those implementations as well as to experiment with different graph matching algorithms that will offer better time and space complexity.

4.3.9 Importance of the Ontology Extracted Synonyms

As described in Section 4.3.2, apart from the synonymy information extracted from ConceptNet 5, BabelNet, and WikiSynonyms, we have exploited the fact that, in some of the considered ontologies, a type may have one preferred name and some additional paraphrases expressed through multiple rdfs:label relations. In this section, we provide an additional set of experiments that aims to measure the importance of these extracted synonyms. This extracted synonymy information constitutes the 0.008%, 0.26%, 0.65% of the training data used in the MA - NCI, FMA - NCI, FMA - SNOMED matching scenarios, respectively. The high variance in their contribution to the training data provide us a means for partially evaluating the correlation between the relative change in the training data and the F1-score.

In Table 4.6, we compare the performance of SCBOW and SCBOW+DAE(O) trained with only the available information from the semantic lexicons, with that presented in Table 4.2 where all the the synonymy information was available. It can be observed that the additional synonymy information affects positively both SCBOW and SCBOW+DAE(O). To better illustrate this correlation, we present in Figure 4.5 how the relative change in the training data is reflected to the relative difference in the performance of our algorithm.

It transpires that the F1-score's relative change monotonically increases with the relative difference in the available data. This behavior constitutes a consistency check for our proposed method, since it aligns with our intuition that increasing the synonymy information leads to producing terminological embeddings more robust to semantic similarity. Regarding

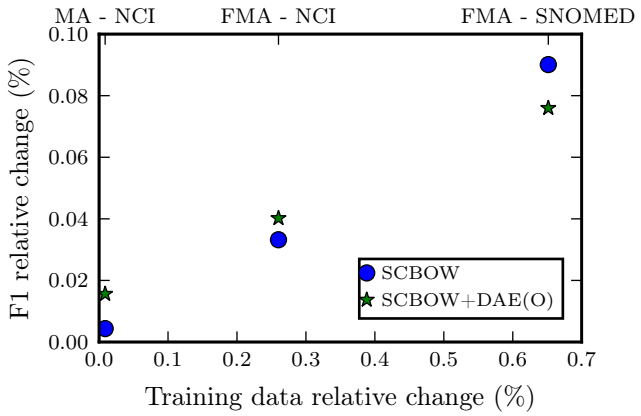

Figure 4.5 – Correlation between the relative change in training data's size and $F1$-score.

the additional benefit that this additional synonymy information brings, a maximum gain of 0.07 in the F1-score is observed across all the matching scenarios. This fact provides supplementary empirical support on the adequacy of the used general semantic lexicons as a means of providing the semantic similarity training data needed by our method. Although this additional synonymy information is important for comparing favorably with the state-of-the-art systems, it does not constitute a catalytic factor for the method's success.

Nonetheless, further experimentation is needed to verify the adequacy of these general semantic lexicons as well as to investigate the correlation between the training data size and the proposed method's performance. We leave for future work the further experimentation with supplementary matching scenarios, different training data sizes and synonymy information sources.

4.3.10 Threshold Sensitivity Analysis

In this section, we perform a sensitivity analysis for the thresholds t_1 and t_2. These thresholds constitute a means for quantifying if two terms are semantically similar or semantically associated. It is worth noting that the tuning of these thresholds can be decoupled. Equivalently, the t_1 threshold can be tuned to optimize the performance of SCBOW, and based on the resulted value the tuning of t_2 can be performed so as to optimize the performance of the outlier detection mechanism. Figure 4.6 shows a threshold sensitivity analysis of our method. For exploring the effect of t_1, we present on the left sub-figure of Figure 4.6 the performance of SCBOW for all the different matching scenarios when varying the value of threshold t_1 between 0 and 1.0. Similarly, the right sub-figure of Figure 4.6 shows the performance of SCBOW+DAE(O) when t_1 is set to 0.2 and the value of t_2 varies in $[0, 1.0]$.

To begin with, it can be seen that both of the threshold sensitivity analysis' figures undergo analogous qualitative behavior across the different ontology matching tasks. At the same time,

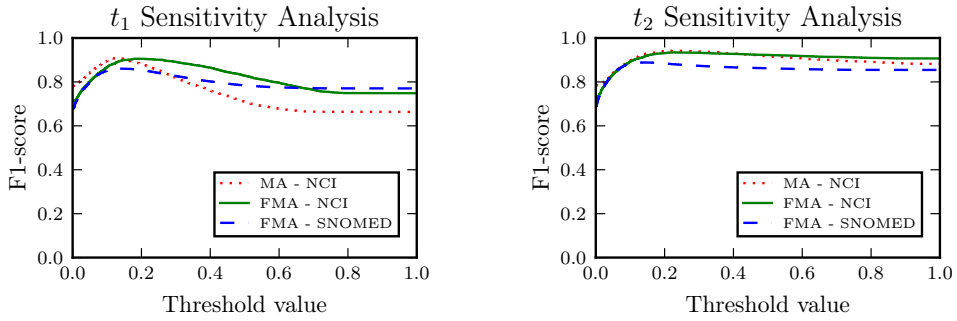

Figure 4.6 – Sensitivity analysis of the proposed algorithm's performance with different threshold values.

it is observed that the performance (F1-score) monotonically increases when the value of t_1 varies between 0 and approximately 0.2. In the t_1 sub-figure, the performance monotonically decreases with $t_1 \in [0.2, 0.6]$ and reaches an asymptotic value at about 0.6. In the case of t_2, although the performance decreases when the value of t_2 exceeds 0.2, the rate of the decrease is significantly lower compared to the rate of decrease of t_1.

It can be seen that although further tuning and experimentation with the values of t_1, t_2 can give better results for each ontology matching task, the values that resulted from the hyperparameter tuning (described in Section 4.3.3) are significantly close to the optimal ones. Moreover, it can be concluded that t_1 values greater than 0.2 have a greater negative impact on the performance compared to the performance drop when t_2 exceeds 0.2. Finally, it should be highlighted that apart from the hyperparameter tuning, no additional direct supervision based on the ground truth alignments is used by our method when we align the ontologies of the considered matching scenarios.

4.3.11 Implications & Limitations

Traditionally, ontology matching approaches have been based on feature engineering in order to obtain different measures of similarity [62]. This plethora of multiple and complementary similarity metrics has introduced various challenges including choosing the most appropriate set of similarity metrics for each task, tuning the various cut-off thresholds used on these metrics, etc. [39]. As a solution to these challenges, various sophisticated solutions have been proposed such as automating the configuration selection process by applying machine learning algorithms on a set of features extracted from the ontologies [39]. Unlike in our approach, only one similarity distance is used; the cosine distance upon the learned features of the phrase retrofitting and the DAE framework. Therefore, there is a drastic decrease in the used similarity metrics and thresholds.

At the same time, it was an open question whether ontology's structural information is really required for performing ontology matching. Our proposed algorithm manages to compare

favorably against state-of-the-art systems without using any kind of structural information. Our results support that a great ontology matching performance can be achieved even in the absence of any graph-theoretic information. However, we avoid to conclude that structural information is not necessary. We leave for future work the investigation of how the ontology's structural information can be exploited in the frame of DRs. Similarly, our method relies on word vectors pre-trained on large external corpora and on synonymy information provided by semantic lexicons also including the ontologies to be matched. Consequently, we can make the conclusion that external corpora and semantic lexicons provide sufficient information to perform ontology matching by only exploiting the ontologies' terms.

Nonetheless, our approach has also certain shortcomings. To begin with, our proposed algorithm is restricted on discovering one-to-one correspondences between two ontologies. At the same time, the use of the McVitie and Wilson's algorithm in our current implementation introduces a significant performance barrier for aligning lager ontologies than the ones considered in our experiments. Although our experimental results demonstrated that high precision can be achieved without using the OWL's reasoning capabilities, our recall remains lower compared to the state-of-the-art systems across all the ontology matching tasks. Taking into account the results presented in Section 4.3.9, it may be concluded that more synonymy information is required to be extracted from supplementary semantic lexicons so as to increase this performance metric. This observation introduces another one weakness of our algorithm; that of closely depending on available external corpora and semantic lexicons. All the aforementioned open questions and shortcomings demonstrate various interesting and important directions for our future work and investigation.

4.4 Conclusions

In this chapter, we address the problem of ontology matching from a representation learning perspective. We propose the refinement of pre-trained word vectors so that when they are used to represent ontological terms, the produced terminological embeddings will be tailored to the ontology matching task. The retrofitted word vectors are learned so that they incorporate domain knowledge encoded in ontologies and semantic lexicons. We cast the problem of ontology matching as an instance of the Stable Marriage problem using the terminological vectors' distances to compute the preference matrix. We compute the aforementioned distances using the cosine distance over the terminological vectors learned by our proposed phrase retrofitting process. Finally, an outlier detection component, based on a denoising autoencoder, sifts through the list of the produced alignments so as to reduce the number of misalignments. Our experimental results demonstrate significant performance gains over the state-of-the-art and indicate a new pathway for ontology matching; a problem which has been traditionally studied under the setting of feature engineering.

5 The Role of Geometrical Space for Link Prediction

"There is no branch of mathematics, however abstract, which may not some day be applied to phenomena of the real world."

Nikolai Ivanovich Lobachevsky

The primary focus of the previous two chapters was on discovering equivalence relations between ontological terms. In this chapter, we study the problem of discovering general relations i.e., not particularly restricted to equivalence relations, between entities appearing in the same ontology or knowledge base, known as *link prediction* or *knowledge base completion*. This is of significant importance since both ontologies and knowledge bases contain a plethora of different relations such as the subsumption or the mereology relation. Building on recent research highlighting the advantages of non-Euclidean space, we examine the contribution of geometrical space to the task of knowledge base completion. We focus on the family of translational models whose performance has been lagging. We extend these models to the hyperbolic space so as to better reflect the topological properties of knowledge bases. We investigate the type of regularities that our model, dubbed *HyperKG*, can capture and show that it is a prominent candidate for effectively representing a subset of Datalog rules. We empirically show, using a variety of link prediction datasets, that hyperbolic space allows to narrow down significantly the performance gap between translational and bilinear models and effectively represent certain types of rules.

5.1 Introduction

Learning in the presence of structured information is an important challenge for artificial intelligence [157, 186, 79]. Knowledge Bases (KBs) such as WordNet [151], Freebase [25], YAGO [213] and DBpedia [131] constitute valuable such resources needed for a plethora of practical applications, including question answering and information extraction. However, despite their formidable number of facts, it is widely accepted that their coverage is still far from

being complete [203, 238]. This shortcoming has opened the door for a number of studies addressing the problem of automatic knowledge base completion (KBC) or link prediction [168]. The impetus of these studies arises from the hypothesis that statistical regularities lay in KB facts, which when correctly exploited can result in the discovery of missing true facts [246]. Building on the great generalisation capability of distributed representations, a great line of research [167, 27, 249, 169, 221] has focused on learning KB vector space embeddings as a way of predicting the plausibility of a fact.

An intrinsic characteristic of knowledge graphs is that they present power-law (or scale-free) degree distributions as many other networks [65, 211]. In an attempt of understanding scale-free networks' properties, various generative models have been proposed such as the models of Barabási and Albert [15] and Van Der Hofstad [226]. Interestingly, Krioukov et al. [125] have shown that scale-free networks naturally emerge in the hyperbolic space. Recently, the hyperbolic geometry was exploited in various works [165, 166, 74, 190] as a means to provide high-quality embeddings for hierarchical structures. Hyperbolic space has the potential to bring significant value in the task of KBC since it offers a natural way to take the KB's topological information into account. Furthermore, many of the relations appearing in KBs lead to hierarchical and hierarchical-like structures [134].

At the same time, the expressiveness of various KB embedding models has been recently examined in terms of their ability to express any ground truth of facts [115, 236]. Moreover, Gutiérrez-Basulto and Schockaert [89] have proceeded one step further and investigated the compatibility between ontological axioms and different types of KB embeddings. Specifically, the authors have proved that a certain family of rules, i.e., the quasi-chained rules which form a subset of Datalog rules [2], can be exactly represented by a KB embedding model whose relations are modelled as convex regions; ensuring, thus, logical consistency in the facts induced by this KB embedding model. In the light of this result, it seems important that the appropriateness of a KB embedding model should not only be measured in terms of fully expressiveness but also in terms of the rules that it can model.

In this chapter, we explore geometrical spaces having the potential to better represent KBs' topological properties and rules and examine the performance implications on KBC. We focus on the family of translational models [27] that attempt to model the statistical regularities as vector translations between entities' vector representations, and whose performance has been lagging. We extend the translational models by learning embeddings of KB entities and relations in the Poincaré-ball model of hyperbolic geometry. We do so by learning compositional vector representations [153] of the entities appearing in a given fact based on translations. The implausibility of a fact is measured in terms of the hyperbolic distance between the compositional vector representations of its entities and the learned relation vector. We prove that the relation regions captured by our proposed model are convex. Our model becomes, thus, a prominent candidate for representing effectively quasi-chained rules.

Among our contributions is the proposal of a novel KB embedding model as well as a reg-

ularisation scheme on the Poincaré-ball model, whose effectiveness we prove empirically. Furthermore, we prove that translational models do not suffer from the restrictions identified by Kazemi and Poole [115] in the case where a fact is considered valid when its implausibility score is below a certain non-zero threshold. We evaluate our approach on various benchmark datasets and our experimental results show that our work makes a big step towards (i) closing the performance gap between translational and bilinear models and (ii) enhancing our understanding of which KBs mostly benefit from exploiting hyperbolic embeddings. Last but not least, our work demonstrates that the choice of geometrical space plays a significant role for KBC and illustrates the importance of taking both the topological and the formal properties of KBs into account.

5.2 Methods

We present an extension of translational models to the Poincaré-ball model of hyperbolic geometry. We represent both entities and relations as points in a high dimensional hyperbolic space. We construct a composite vectorial representation of the entities appearing in a given fact by exploiting translational and permutational operations. We measure the implausibility of a given fact through the hyperbolic distance between the composite vectorial representation of the entities and the vectorial representation of the relation. In addition, we propose a novel regularisation scheme on the Poincaré-ball model. We also resolve certain misconceptions regarding the expressivity of translational models. Finally, we demonstrate that our proposed model is a prominent candidate for effectively representing quasi-chained rules.

5.2.1 Preliminaries

We introduce some definitions and additional notation that we will use throughout this chapter. We denote the vector concatenation operation by the symbol $\oplus$ and the inner product by $\langle \cdot, \cdot \rangle$. We define the *rectifier* activation function as: $[\cdot]_+ := \max(\cdot, 0)$.

Quasi-chained Rules. Let $\mathbf{E}, \mathbf{N}$ and $\mathbf{V}$ be disjoint sets of *entities, (labelled) nulls* and *variables*, respectively.[15] Let $\mathbf{R}$ be the set of relation symbols. A *term* t is an element in $\mathbf{E} \cup \mathbf{N} \cup \mathbf{V}$; an *atom* α is an expression of the form $R(t_1, t_2)$, where R is a *relation* between the terms t_1, t_2. Let $\mathsf{terms}(\alpha) := \{t_1, t_2\}$; $\mathsf{vars}(\alpha) := \mathsf{terms}(\alpha) \cap \mathbf{V}$ and B_n for $n \geq 0$, H_k for $k \geq 1$ be atoms with terms in $\mathbf{E} \cup \mathbf{V}$. Additionally, let $X_j \in \mathbf{V}$ for $j \geq 1$. A *quasi-chained (QC) rule* σ [89] is an expression of the form:

$$B_1 \wedge \ldots \wedge B_n \rightarrow \exists X_1, \ldots, X_j . H_1 \wedge \ldots \wedge H_k, \tag{5.1}$$

where for all $i : 1 \leq i \leq n$

$$|(\mathsf{vars}(B_1) \cup \ldots \cup \mathsf{vars}(B_{i-1})) \cap \mathsf{vars}(B_i)| \leq 1$$

The QC rules constitute a subset of Datalog rules. A *database* D is a finite set of *facts*, i.e., a

set of atoms with terms in **E**. A *knowledge base (KB)* $\mathcal{K}$ consists of a pair (Σ, D) where Σ is an ontology whose axioms are QC rules and D a database. It should be noted that no constraint is imposed on the number of available axioms in the ontology. The ontology could be minimal in the sense of only defining the relation symbols. However, any type of rule, whether it is the product of the ontological design or results from formalising a statistical regularity, should belong to the family of QC rules. The Gene Ontology [12] constitutes one notable example of an ontology that exhibits QC rules.

Circular Permutation Matrices. An orthogonal matrix is defined as a real square matrix whose columns and rows are orthogonal unit vectors (i.e., orthonormal vectors), i.e.,

$$Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I \tag{5.2}$$

where I is the identity matrix. Orthogonal matrices preserve the vector inner product and, thus, they also preserve the Euclidean norms. Let $1 \leq i < n$, we define the *circular permutation matrix* Π_i to be the orthogonal $n \times n$ matrix that is associated with the following circular permutation of a n-dimensional vector $\boldsymbol{x}$:

$$\begin{pmatrix} x_1 & \cdots & x_{n-i} & x_{n-i+1} & \cdots & x_n \\ x_{i+1} & \cdots & x_n & x_1 & \cdots & x_i \end{pmatrix} \tag{5.3}$$

where x_i is the ith coordinate of $\boldsymbol{x}$ and i controls the number of $n-i$ successive circular shifts.

Hyperbolic Space. In this work, we exploit the Poincaré-ball model of the hyperbolic geometry. The Poincaré-ball model is the Riemannian manifold $\mathbb{P}^n = (\mathbb{B}^n, d_p)$, where $\mathbb{B}^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < 1\}$ and d_p is the distance function:

$$\begin{aligned} d_p(\boldsymbol{u}, \boldsymbol{v}) &= \operatorname{acosh}(1 + 2\delta(\boldsymbol{u}, \boldsymbol{v})) \\ \delta(\boldsymbol{u}, \boldsymbol{v}) &= \frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)} \end{aligned} \tag{5.4}$$

The Poincaré-ball model presents a group-like structure when it is equipped with the *Möbius addition* [223, 185], defined by:

$$\boldsymbol{u} \boxplus \boldsymbol{v} := \frac{(1 + 2\langle \boldsymbol{u}, \boldsymbol{v} \rangle + \|\boldsymbol{v}\|^2)\boldsymbol{u} + (1 - \|\boldsymbol{u}\|^2)\boldsymbol{v}}{1 + 2\langle \boldsymbol{u}, \boldsymbol{v} \rangle + \|\boldsymbol{u}\|^2 \|\boldsymbol{v}\|^2} \tag{5.5}$$

The isometries of $(\mathbb{B}^n, d_p)$ can be expressed as a composition of a left gyrotranslation with an orthogonal transformation restricted to $\mathbb{B}^n$, where the *left gyrotranslation* is defined as $L_u : v \mapsto u \boxplus v$ [4, 185]. Therefore, circular permutations constitute zero-left gyrotranslation isometries of the Poincaré-ball model.

5.2.2 Hyperbolic Knowledge Graph Embeddings

The database of a KB consists of a set of facts in the form of $R(subject, object)$. We will learn hyperbolic embeddings of entities and relations such that valid facts will have a lower implausibility score than the invalid ones. To learn such representations, we extend the work of Bordes et al. [27] by defining a translation-based model in the hyperbolic space; embedding, thus, both entities and relations in the same space.

Let $\boldsymbol{s}, \boldsymbol{r}, \boldsymbol{o} \in \mathbb{B}^n$ be the hyperbolic embeddings of the *subject, relation and object*, respectively, appearing in the $R(subject, object)$ fact. We define a *term embedding* as a function $\xi : \mathbb{B}^n \times \mathbb{B}^n \to \mathbb{B}^n$, that creates a composite vector representation for the pair $(subject, object)$. Since our motivation is to generalise the translation models to the hyperbolic space, a natural way to define the term embeddings is by using the Möbius addition. However, we found out empirically that the normal addition in the Euclidean space generalises better than the Möbius addition. We provide a possible explanation for this behaviour in an ablation study presented in the Results & Analysis section. To introduce non-commutativity in the term composition function, we use a circular permutation matrix to project the object embeddings. Non-commutativity is important because it allows to model asymmetric relations with compositional representations [169]. Therefore, we define the term embedding as: $\boldsymbol{s} + \Pi_\beta \boldsymbol{o}$, where β is a hyperparameter controlling the number of successive circular shifts. To enforce the term embeddings to stay in the Poincaré-ball, we constrain all the entity embeddings to have a Euclidean norm less than 0.5. Namely, $\|\boldsymbol{e}\| < 0.5$ and $\|\boldsymbol{r}\| < 1.0$ for all entity and relation vectors, respectively. It should be noted that the entities' norm constraints do not restrict term embeddings to span the Poincaré-ball. We define the implausibility score as the hyperbolic distance between the term and the relation embeddings. Specifically, the implausibility score of a fact is defined as:

$$f_R(s, o) = d_p(\boldsymbol{s} + \Pi_\beta \boldsymbol{o}, \boldsymbol{r}) \tag{5.6}$$

Figure 5.1 provides an illustration of the HyperKG model in $\mathbb{P}^2$. We follow previous work [27] to minimise the following hinge loss function:

$$\mathscr{L} = \sum_{\substack{R(s,o)\sim P, \\ R'(s',o')\sim N}} \left[\gamma + f_R(s, o) - f_{R'}(s', o')\right]_+ \tag{5.7}$$

where P is the training set consisting of valid facts, N is a set of corrupted facts. To create the corrupted facts, we experimented with two strategies. We replaced randomly either the subject or the object of a valid fact with a random entity (but not both at the same time). We denote with $\#_{negs_E}$ the number of negative examples. Furthermore, we experimented with replacing randomly the relation while retaining intact the entities of a valid fact. We denote with $\#_{negs_R}$ the number of "relation-corrupted" negative examples. We employ the "*Bernoulli*" sampling method to generate incorrect facts [237, 105, 246].

As pointed out in different studies [27, 49, 127], regularisation techniques are really beneficial

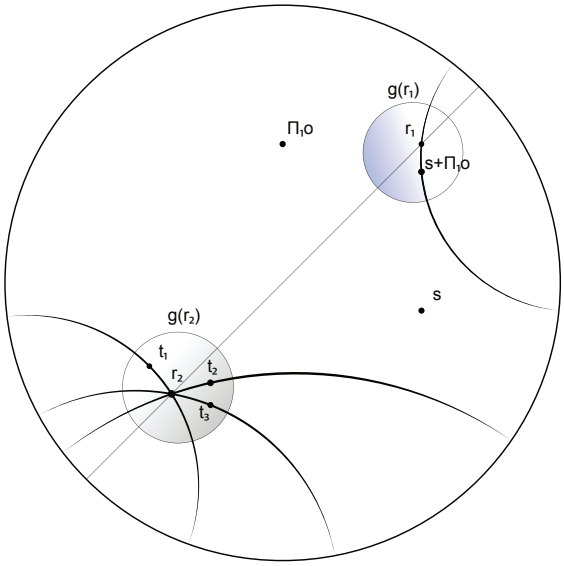

Figure 5.1 – A visualisation of HyperKG model in the $\mathbb{P}^2$ space. The geodesics of the disk model are circles perpendicular to its boundary. The zero-curvature geodesic passing from the origin corresponds to the line $\epsilon : y - x = 0$ in the Euclidean plane. Reflections over the line ϵ are equivalent to Π_1 permutations in the plane. $s, \Pi_1 o, s + \Pi_1 o$ are the subject vector, the permuted object vector and the composite term vector, respectively. $g(r_1), g(r_2)$ denote the geometric loci of term vectors satisfying relations R_1, R_2, with relation vectors $r1, r2$. t_1, t_2, t_3 are valid term vectors for the relation R_2.

for the task of KBC. Nonetheless, very few of the classical regularisation methods are directly applicable or easily generalisable in the Poincaré-ball model of hyperbolic space. For instance, the ℓ_2 regularisation constraint imposes vectors to stay close to the origin, which can lead to underflows. The same holds for dropout [210], when a rather large dropout rate was used.[16] In our experiments, we noticed a tendency of the vectors to stay close to the origin. Imposing a constraint to the vectors to stay away from the origin stabilised the training procedure and increased the model's generalisation capability. It should be noted that as the points in the Poincaré-ball approach the ball's boundary their distance $d_p(\boldsymbol{u}, \boldsymbol{v})$ approaches $d_p(\boldsymbol{u}, \mathbf{0}) + d_p(\mathbf{0}, \boldsymbol{v})$, which is analogous to the fact that in a tree the shortest path between two siblings is the path through their parent [190]. Building on this observation, our regulariser further imposes this "tree-like" property. Additionally, since the volume in hyperbolic space grows exponentially, our regulariser implicitly penalises crowding. Let $\Theta := \{\boldsymbol{e}_i\}_{i=1}^{|\mathbf{E}|} \bigcup \{\boldsymbol{r}_i\}_{i=1}^{|\mathbf{R}|}$ be the set of all entity and relation vectors, where $|\mathbf{E}|, |\mathbf{R}|$ denote the cardinalities of the sets $\mathbf{E}, \mathbf{R}$, respectively. $\mathscr{R}(\Theta)$ defines our proposed regularisation loss function:

$$\mathscr{R}(\Theta) = \sum_{i=1}^{|\mathbf{E}|+|\mathbf{R}|} (1 - \| \boldsymbol{\theta}_i \|^2) \tag{5.8}$$

The overall embedding loss is now defined as $\mathscr{L}'(\Theta) = \mathscr{L}(\Theta) + \lambda \mathscr{R}(\Theta)$, where λ is a hyper-

parameter controlling the regularisation effect. We define $a_i := 0.5$, if θ_i corresponds to an entity vector and $a_i := 1.0$, otherwise. To minimise $\mathscr{L}'(\Theta)$, we solve the following optimisation problem:

$$\Theta' \leftarrow \arg\min_{\Theta} \mathscr{L}'(\Theta) \qquad \text{s.t. } \forall \boldsymbol{\theta}_i \in \Theta : \|\boldsymbol{\theta}_i\| < a_i. \tag{5.9}$$

To solve Equation (5.9), we follow Nickel and Kiela [165] and use Riemannian SGD (RSGD; 26). In RSGD, the parameter updates are of the form:

$$\boldsymbol{\theta}_{t+1} = \mathfrak{R}_{\boldsymbol{\theta}_t}\left(-\eta \nabla_R \mathscr{L}'(\boldsymbol{\theta}_t)\right)$$

where $\mathfrak{R}_{\boldsymbol{\theta}_t}$ denotes the retraction onto the open d-dimensional unit ball at $\boldsymbol{\theta}_t$ and η denotes the learning rate. The Riemannian gradient of $\mathscr{L}'(\boldsymbol{\theta})$ is denoted by $\nabla_R \in \mathscr{T}_\theta \mathbb{B}$. The Riemannian gradient can be computed as $\nabla_R = \frac{(1-\|\boldsymbol{\theta}_t\|^2)^2}{4}\nabla_E$, where ∇_E denotes the Euclidean gradient of $\mathscr{L}'(\boldsymbol{\theta})$. Similarly to Nickel and Kiela [165], we use the following retraction operation $\mathfrak{R}_{\boldsymbol{\theta}}(\boldsymbol{v}) = \boldsymbol{\theta} + \boldsymbol{v}$.

To constrain the embeddings to remain within the Poincaré ball and respect the additional constraints, we use the following projection:

$$\text{proj}(\boldsymbol{\theta}, a) = \begin{cases} a\boldsymbol{\theta} / (\|\boldsymbol{\theta}\| + \varepsilon) & \text{if } \|\boldsymbol{\theta}\| \geq a \\ \boldsymbol{\theta} & \text{otherwise}, \end{cases} \tag{5.10}$$

where ε is a small constant to ensure numerical stability. In all experiments we used $\varepsilon = 10^{-5}$. Let a be the constraint imposed on vector $\boldsymbol{\theta}$, the full update for a single embedding is then of the form:

$$\boldsymbol{\theta}_{t+1} \leftarrow \text{proj}\left(\boldsymbol{\theta}_t - \eta \frac{(1-\|\boldsymbol{\theta}_t\|^2)^2}{4}\nabla_E, a\right). \tag{5.11}$$

We initialise the embeddings using the Xavier initialization scheme [80], where we use Equation (5.10) for projecting the vectors whose norms violate the imposed constraints.

5.2.3 Convex Relation Spaces

In this section, we investigate the type of rules that HyperKG can model. Recently, Wang et al. [236] proved that the bilinear models are universal, i.e., they can represent every possible fact given that the dimensionality of the vectors is sufficient. The authors have also shown that the TransE model is not universal. In parallel, Kazemi and Poole [115] have shown that the FTransE model [72], which is the most general translational model proposed in the literature, imposes some severe restrictions on the types of relations the translational models can represent. In the core of their proof lies the assumption that the implausibility score defined by the FTransE model approaches zero for all given valid facts. Nonetheless, this condition is less likely to be met from an optimisation perspective [245].

Additionally, Gutiérrez-Basulto and Schockaert [89] studied the types of regularities that

KB embedding methods can capture. To allow for a formal characterisation, the authors considered hard thresholds λ_R such that a fact $R(s,o)$ is considered valid iff $s_R(\boldsymbol{s},\boldsymbol{o}) \leq \lambda_R$, where $s_R(.,.)$ is the implausibility score. It should be highlighted that KB embeddings are often learned based on a maximum-margin loss function, which ideally leads to hard-threshold separation. The vector space representation of a given relation R can then be viewed as a region $n(R)$ in $\mathbb{R}^{2n}$, defined as follows:

$$n(R) = \{\boldsymbol{s} \oplus \boldsymbol{o} \,|\, s_R(\boldsymbol{s},\boldsymbol{o}) \leq \lambda_R\} \tag{5.12}$$

Based on this view of the relation space, the authors prove that although bilinear models are fully expressive, they impose constraints on the type of rules they can learn. Specifically, let $R_1(X,Y) \rightarrow S(X,Y)$, $R_2(X,Y) \rightarrow S(X,Y)$ be two valid rules. The bilinear models impose either that $R_1(X,Y) \rightarrow R_2(X,Y)$ or $R_2(X,Y) \rightarrow R_1(X,Y)$; introducing, thus, a number of restrictions on the type of subsumption hierarchies they can model. Gutiérrez-Basulto and Schockaert [89], additionally, prove that there exists a KB embedding model with convex relation regions that can correctly represent knowledge bases whose axioms belong to the family of QC rules. Equivalently, any inductive reasoning made by the aforementioned KB embedding model would be logically consistent and deductively closed with respect to the ontological rules. It can be easily verified that the relation regions of TransE [27] are indeed convex. This result is in accordance with the results of Wang et al. [236]; TransE is not fully expressive. However, it could be a prominent candidate for representing QC rules consistently. Nonetheless, this result seems to be in conflict with the results of Kazemi and Poole [115]. Let $s_R^{TE}(s,o)$ be the implausibility score of TransE, we demystify this seeming inconsistency by proving the following lemma:

Lemma 1 *The restrictions proved by Kazemi and Poole [115] do not apply to the TransE model when a fact is considered valid iff $s_R^{TE}(s,o) \leq \lambda_R$ for sufficient $\lambda_R > 0$.*

We prove Lemma 1 in Appendix A.1, by constructing counterexamples for each one of the restrictions. Since the restrictions can be lifted for the TransE model, we can safely conclude that they are not, in general, valid for all its generalisations. In parallel, we built upon the formal characterisation of relations regions, defined in Equation (5.12) and we prove that the relation regions captured by HyperKG are indeed convex. Specifically, we prove:

Proposition 1 *The geometric locus of the term vectors, in the form of $\boldsymbol{s} + \Pi_\beta \boldsymbol{o}$, that satisfy the equation $d_p(\boldsymbol{s} + \Pi_\beta \boldsymbol{o}, \boldsymbol{r}) \leq \lambda_R$ for some $\lambda_R > 0$ corresponds to a d-dimensional closed ball in the Euclidean space. Let $\rho = \frac{\cosh(\lambda_R)-1}{2}(1 - \|\boldsymbol{r}\|^2)$, the geometric locus can be written as*

$$\left|\left| \boldsymbol{s} + \Pi_\beta \boldsymbol{o} - \frac{\boldsymbol{r}}{\rho+1} \right|\right|^2 \leq \frac{\rho}{\rho+1} + \frac{\|\boldsymbol{r}\|^2}{(\rho+1)^2} - \frac{\|\boldsymbol{r}\|^2}{\rho+1}, \tag{5.13}$$

where the ball's radius is guaranteed to be strictly greater than zero.

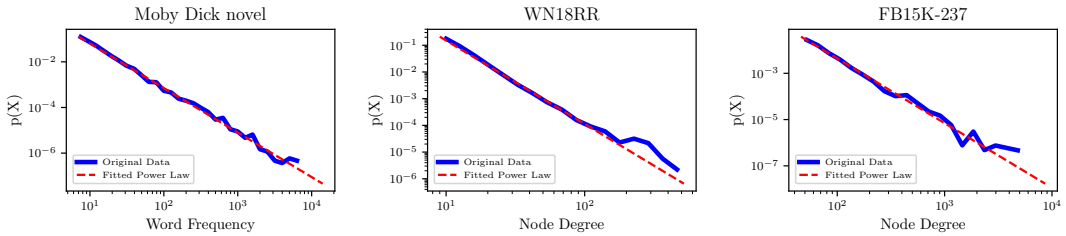

Figure 5.2 – A visualisation of the probability density functions using a histogram with log-log axes.

The proof of Proposition 1 can also be found in the Appendix A.1. By exploiting the triangle inequality, we can easily verify that the relation regions captured by HyperKG are indeed convex. Figure 5.1 provides an illustration of the geometric loci captured by HyperKG in $\mathbb{B}^2$. This result shows that HyperKG constitutes another one prominent embedding model for effectively representing QC rules.

5.3 Results & Discussion

We evaluate our HyperKG model on the task of KBC using two sets of experiments. We conduct experiments on the WN18RR [49] and FB15k-237 [220] datasets. We also construct two datasets whose statistical regularities can be expressed as QC rules to test our model's performance in their presence. WN18RR and FB15k-237 constitute refined subsets of WN18 and FB15K that were introduced by Bordes et al. [27]. Toutanova and Chen [220] identified that WN18 and FB15K contained a lot of reversible relations, enabling, thus, various KB embedding models to generalise easily. Exploiting this fact, Dettmers et al. [49] obtained state-of-the-art results only by using a simple reversal rule. WN18RR and FB15k-237 were carefully created to alleviate this leakage of information.

To test whether the scale-free distribution provides a reasonable means for modelling topological properties of knowledge graphs, we investigate the degree distributions of WN18RR and FB15k-237. Similarly to Steyvers and Tenenbaum [211], we treat the knowledge graphs as undirected networks. We also compare against the distribution of the frequency of word usage in the English language; a phenomenon that is known to follow a power-law distribution [257]. To do so, we used the frequency of word usage in Herman Melville's novel "Moby Dick" [162]. We followed the procedure described by Alstott et al. [8]. In Figure 5.2, we show our analysis where we demonstrate on a histogram with log-log axes the probability density function with regard to the observed property for each dataset, including the fitted power-law distribution. It can be seen that the power-law distribution provides a reasonable means for also describing the degree distribution of KBs; justifying the work of Steyvers and Tenenbaum [211]. The fluctuations in the cases of WN18RR and FB15k-237 could be explained by the fact that the datasets are subsets of more complete KBs; a fact that introduces noise which in turn can

explain deviations from the perfection of a theoretical distribution [8].

5.3.1 Evaluation Benchmarks

To test our model's performance on capturing QC rules, we extract from Wikidata [232, 61] two subsets of facts that satisfy the following rules:

(a) $is\_a(X,Y) \wedge part\_of(Y,Z) \rightarrow part\_of(X,Z)$

(b) $part\_of(X,Y) \wedge is\_a(Y,Z) \rightarrow part\_of(X,Z)$

The relations is_a, $part\_of$ correspond to the subsumption and the mereology relation, respectively, which are two of the most common relations encountered in KBs [193]. Recent studies have noted that many real world KB relations have very few facts [247], raising the importance of generalising with limited number of facts. To test our model in the presence of sparse long-tail relations, we kept the created datasets sufficiently small. For each type of the aforementioned rules, we extract 200 facts that satisfy them from Wikidata. We construct two datasets that we dub WD and WD_{++}. The dataset WD contains only the facts that satisfy rule (**a**). WD_{++} extends WD by also including the facts satisfying rule (**b**). The evaluation protocol was the following: For every dataset, we split all the facts randomly in train (80%), validation (10%), and test (10%) set, such that the validation and test sets only contain a subset of the rules' consequents in the form of $part\_of(X,Z)$. Table 5.1 provides details regarding the respective size of each dataset.

Dataset	| E |	| R |	#Train	#Valid	#Test
WN18RR	40,943	11	86,835	3,034	3,134
FB15k-237	14,541	237	272,115	17,535	20,466
WD	418	2	550	25	25
WD_{++}	763	2	1,120	40	40

Table 5.1 – Statistics of the experimental datasets.

5.3.2 Evaluation Protocol & Implementation Details

In the KBC task the models are evaluated based on their capability to answer queries such as $R(subject, \mathbf{?})$ and $R(\mathbf{?}, object)$ [27]; predicting, thus, the missing entity. Specifically, all the possible corruptions are obtained by replacing either the *subject* or the *object* and the entities are ranked based on the values of the implausibility score. The models should assign lower implausibility scores to valid facts and higher scores to implausible ones. We use the "**Filtered**" setting protocol [27], i.e., not taking any corrupted facts that exist in KB into account. We employ two common evaluation metrics: Mean Reciprocal Rank (MRR), and Hits@10 (i.e., the proportion of the valid/test triples ranking in top 10 predictions). Higher MRR or higher Hits@10 indicate better performance.

Dataset	Model	$\#_{negs_E}$	$\#_{negs_R}$	η	λ	n	γ,	β
WN18RR	HyperKG	10	0	0.01	0.8	100	1.0	$\lfloor\frac{n}{2}\rfloor$
WN18RR	HyperKG (Möbius addition)	10	0	0.01	-	100	1.0	$\lfloor\frac{n}{2}\rfloor$
WN18RR	HyperKG (no regularisation)	10	0	0.01	0.0	100	1.0	$\lfloor\frac{n}{2}\rfloor$
FB15k-237	HyperKG	5	0	0.01	0.2	100	0.5	$\lfloor\frac{n}{2}\rfloor$
FB15k-237	HyperKG (Möbius addition)	5	0	0.01	-	100	0.5	$\lfloor\frac{n}{2}\rfloor$
FB15k-237	HyperKG (no regularisation)	5	0	0.01	0.0	100	0.5	$\lfloor\frac{n}{2}\rfloor$
WD	HyperKG	1	1	0.8	0	100	7	$\lfloor\frac{n}{2}\rfloor$
WD_{++}	HyperKG	1	1	0.1	0	100	7	$\lfloor\frac{n}{2}\rfloor$

Table 5.2 – HyperKG's hyperparameters used across the different experiments.

The reported results are given for the best set of hyperparameters evaluated on the validation set using grid search. Varying the batch size had no effect on the performance. Therefore, we divided every epoch into 10 mini-batches. The hyperparameter search space was the following: $\#_{negs_E} \in \{1,2,3,4,5,8,10,12,15\}$, $\#_{negs_R} \in \{0,1,2\}$, $\eta \in \{0.8,0.5,0.2,0.1,0.05,0.01,0.005\}$, $\beta \in \{\lfloor\frac{3n}{4}\rfloor, \lfloor\frac{n}{2}\rfloor, \lfloor\frac{n}{4}\rfloor, 0\}$, $\gamma \in \{7.0,5.0,2.0,1.5,1.0,0.8,0.5,0.2,0.1\}$, the embeddings' dimension $n \in \{40, 100, 200\}$, and $\lambda \in \{2.0,1.5,1.0,0.8,0.6,0.4,0.2,0.1,0.0\}$. We used early stopping based on the validation's set filtered MRR performance, computed every 50 epochs with a maximum number of 2000 epochs. We report in Table 5.2 the best hyperparameters for our HyperKG model that were used across the different experiments. For WD and WD_{++}, we did not use the "*Bernoulli*" sampling method, but instead we corrupted the subject and object of a fact with equal probability.

For the experiments on the WD and WD_{++} datasets, we used the public available implementations of TransE [27] and ComplEx [221] provided in the OpenKE framework [91]. The reported results are given for the best set of hyperparameters evaluated on the validation set using grid search. We divided every epoch into 64 mini-batches.

The hyperparameter search space for TransE was the following: the dimensionality of embeddings $n \in \{50,100\}$, SGD learning rate $\in \{0.0001,0.0005,0.001,0.005\}$, l_1-norm or l_2-norm, and margin $\gamma \in \{1,3,5,7\}$. The highest MRR scores were achieved when using l_1-norm, learning rate at 0.005, $\gamma = 7$ and $n = 50$ for both WD and WD_{++}.

The hyperparameter search space for ComplEx was the following: $n \in \{50,100\}$, $\lambda \in \{0.1,0.03, 0.01,0.003,0.001,0.0003,0.0\}$, $\alpha_0 \in \{1.0,0.5,0.2,0.1,0.05,0.02,0.01\}$, $\eta \in \{1,2,5,10\}$ where n the dimensionality of embeddings, λ the L_2 regularisation parameter, α_0 the AdaGrad's initial learning rate, and η the number of negative examples generated per positive training triple. For WD, the highest MRR score was achieved using a learning rate of 0.05, $\lambda = 0.1$, $\eta = 5$ and $n = 50$. For WD_{++}, the highest MRR score was achieved using a learning rate of 0.05, $\lambda = 0.1$, $\eta = 5$ and $n = 100$.

Method	Type	WN18RR		FB15k-237	
		MRR	H@10	MRR	H@10
DISTMULT [249] [⋆]	Bilinear	0.43	49	0.24	41
ComplEx [221] [⋆]	Bilinear	0.44	51	0.24	42
TransE [27] [⋆]	Translational	0.22	50	0.29	46
HyperKG (Möbius addition)	Translational	0.30	44	0.19	32
HyperKG (no regularisation)	Translational	0.30	46	0.25	41
HyperKG	Translational	0.41	50	0.28	45

Table 5.3 – Experimental results on WN18RR and FB15k-237 test sets. MRR and H@10 denote the mean reciprocal rank and Hits@10 (in %), respectively. [⋆]: Results are taken from Nguyen et al. [164].

5.3.3 Results & Analysis

Table 5.3 compares the experimental results of our HyperKG model with previous published results on WN18RR and FB15k-237 datasets. We have experimentally validated that both datasets present power-law degree distributions. Additionally, WN18RR contains more hierarchical-like relations compared to FB15k-237 [14]. We compare against the shallow KB embedding models DISTMULT [249], ComplEx [221] and TransE [27], which constitute important representatives of bilinear and translational models. We exclude from our comparison recent work that explores different types of training regimes such as adversarial training, the inclusion of reciprocal facts and/or multiple geometrical spaces [30, 214, 115, 127, 14] to make the analysis less biased to factors that could overshadow the importance of the embedding space. We give the results of our algorithm under the HyperKG listing.

Although HyperKG belongs to the translational family of KB embedding models, it achieves comparable performance to the other models on the WN18RR dataset. When we compare the performance of HyperKG and TransE, we see that HyperKG achieves almost the double MRR score. This consequently shows that the lower MRR performance of TransE is not an intrinsic characteristic of the translational models, but a restriction that can be lifted by the right choice of geometrical space. With regard to Hits@10 on WN18RR, HyperKG exhibits slightly lower performance compared to ComplEx. On the FB15k-237 dataset, however, HyperKG and TransE demonstrate almost the same behaviour outperforming DISTMULT and ComplEx in both metrics. Since the performance gap between TransE and HyperKG is small, we hypothesise that this is due to a less fine-grained hyperparameter tuning. Overall, the hyperbolic space appears to be more beneficial for datasets that contain many hierarchical-like relations such as WN18RR, without a significant performance degradation in the other case.

We also report in Table 5.3 two additional experiments where we explore the performance boost that our regularisation scheme brings as well as the behaviour of HyperKG when the Möbius addition is used instead of the Euclidean one. In the experiment where the Möbius addition was used, we removed the constraint for the entity vectors to have a norm less than 0.5. Although the Möbius addition is non-commutative, we found beneficial to keep the

Method	WD		WD_{++}	
	MRR	H@10	MRR	H@10
ComplEx	0.92	98	0.81	92
TransE	0.88	96	0.89	98
HyperKG	0.98	98	0.93	98

Table 5.4 – Experimental results on WD and WD_{++} test sets. MRR and H@10 denote the mean reciprocal rank and Hits@10 (in %), respectively.

permutation matrix. Nonetheless, we do not use our regularisation scheme. Therefore, the implausibility score is $d_p(\boldsymbol{s} \boxplus \Pi_\beta \boldsymbol{o}, \boldsymbol{r})$. To investigate the effect of our proposed regularisation scheme, we show results where our regularisation scheme, defined in Equation (5.8), is not used, keeping, however, the rest of the architecture the same. Comparing the performance of the HyperKG variation using the Möbius addition against the performance of the HyperKG without regularisation, we can observe that we can achieve better results by using the Euclidean addition. This can be explained as follows. Generally, there is no unique and universal geometrical space adequate for every dataset [84]. To recover Euclidean Space from the Poincaré-ball model equipped with the Möbius addition, the ball's radius should grow to infinity [223]. Instead, by using the Euclidean addition and since the hyperbolic metric is locally Euclidean, HyperKG can model facts for which the Euclidean Space is more appropriate by learning to retain small distances. Last but not least, we can observe that our proposed regularisation scheme is beneficial in terms of both MRR and Hits@10 on both datasets.

Table 5.4 reports the results on the WD and WD_{++} datasets. We compare HyperKG performance against that of TransE and ComplEx. It can be observed that none of the models manages to totally capture the statistical regularities of these datasets. All the models undergo similar Hits@10 performance on both datasets. HyperKG and TransE, that both have convex relation spaces, outperform ComplEx on both datasets in terms of MRR and Hits@10. Furthermore, the translational models show a relatively steady performance compared to ComplEx, whose performance deteriorates in the presence of the two rules appearing in WD_{++}. Our results point to a promising direction for developing less expressive KB embedding models which can, however, better represent certain types of rules.

5.4 Conclusions

In this chapter, we examined the importance of the geometrical space for the task of KBC. We showed that the lagging performance of translational models compared to the bilinear ones is not an intrinsic characteristic of them but a restriction that can be lifted in the hyperbolic space. Our results validated that the right choice of geometrical space is a critical decision that impacts the performance of KB embedding models. Our findings also shed light on understanding which KBs mostly benefit from the use of hyperbolic embeddings. Moreover, we demonstrated a new promising direction for developing models that, although not fully

expressive, allow to better represent certain families of rules; opening up for more fine-grained reasoning tasks.

6 Conclusion

MR. MARTIN: *I have a little girl, my little daughter, she lives with me, dear lady. She is two years old, has a white eye and a red eye, she is very pretty, and her name is Alice, dear lady.*
MRS. MARTIN: *What a bizarre coincidence! I, too, have a little girl. She is two years old, has a white eye and a red eye, she is very pretty, and her name is Alice, too, dear sir!*
MR. MARTIN: *How curious it is and what a coincidence! And bizarre! Perhaps they are the same, dear lady!*

Eugene Ionesco, The Bald Soprano

Metaphorically speaking, the current information landscape, as discussed in Chapter 1, can be thought of consisting of a vast number of complexes of *semantically connected* islands [112]. Within each complex of islands, information can be harvested without the risk of misinterpretation. Nonetheless, when a *message* is traversed between two such complexes of islands, special consideration is required in order to ensure that it is interpreted in a common and uniform way. Establishing semantic bridges across these heterogeneous complexes of islands becomes, thus, crucial for accomplishing a mutual understanding. In addition, the information landscape is characterised by an exponential increase in the number of islands that are emerging on the *information ocean* as well as in the number of new complexes that are being formed. It becomes, thus, apparent that the exponentially increasing information landscape prohibits manual curation strategies and illustrates the importance of an automatic computational approach that relies less on human expertise and intervention.

The goal of this thesis has been to demonstrate that learning distributed representations of ontological terms, entities and relations provides a sufficient workforce for automatically building semantic bridges between semantically heterogeneous complexes of islands and successfully generalising across a plethora of practical application domains. Figure 6.1 illustrates the idea by showing a part of the Porphyrian Tree, presented in Figure 2.1 of Chapter 2, where learned representations are assigned to every term and relation that appear in it. For brevity and ease of human-readability, the entity embeddings were omitted. Making the liaison with the *fast and slow thinking* [110], the distributed representations are exploited to

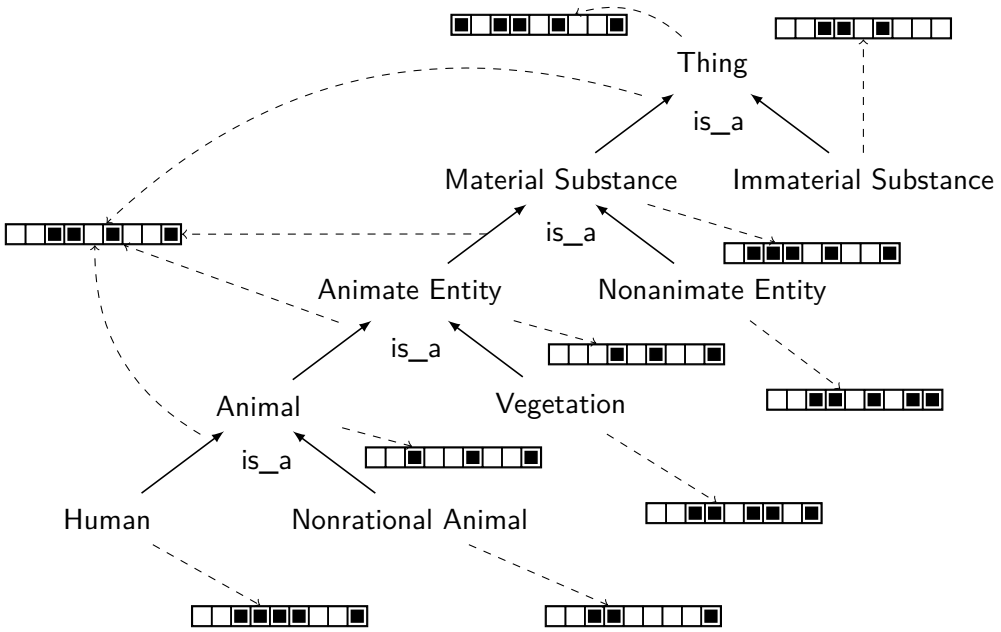

Figure 6.1 – Part of Porphyrian Tree extended with distributed representations.

provide *fast* approximate answers, whereas ontologies provide the needed mechanism for *slow* reasoning [154]. In this thesis, we demonstrated that this *fast and approximate thinking* can be harnessed to discover equivalence relations between terms appearing in different ontologies. The novel proposed approaches have been shown to present significant performance improvements over the current state-of-the-art. Furthermore, we showed that the geometrical space over which the representations are learned affects drastically the performance of link prediction and we illustrated the importance for evaluating the models in terms of the families of rules that they can capture. In the rest of this chapter, we summarise the key contributions of this thesis and discuss interesting future directions that could address open challenges related to the challenging problems of ontology alignment and link prediction.

6.1 Summary of contributions

In the following, we summarise the contributions of this thesis.

In Chapter 3, we demonstrated that we can approach the problem of ontology alignment through representation learning. Our key proposal was to refine pre-trained word vectors aiming at deriving terminological representations that are tailored to the ontology matching task. We demonstrated that by exploiting transfer learning we can overcome the main obstacles, i.e., the small sample size and the serious class imbalance problem, that have been shown to hinder the application of machine learning to the problem. We empirically evaluated our algorithm using a standard ontology matching benchmark as well as a real world alignment scenario between Schema.org and the DBpedia ontologies. We compared our method against

state-of-the-art ontology matching systems based on feature engineering and our method showed significant performance gains on both benchmarks, however at the cost of a certain amount of degradation in the precision metric. Our experiments on the real world scenario also illustrated the fact that hand-crafted features and similarity metrics can indeed fail to generalise in specific domains. Interestingly, even without retrofitting, pre-trained word vectors achieved high recall on the Schema.org - DBpedia ontology matching scenario. Additionally, our ablation study provided empirical evidence that the initial choice of pre-trained word vectors affects the final performance and that choosing initial word vectors that are already tailored to semantic similarity can lead to a performance boost. Finally, our ablation study provided empirical support that (i) the initial choice of the semantic lexicons affects the performance and that (ii) higher coverage of synonymy and antonymy information has a positive effect on the final performance.

In Chapter 4, we addressed the shortcomings of the previous proposed method. We proposed to go beyond the retrofitting of single word embeddings by exploiting a novel architecture that allowed tailoring phrase embeddings to semantic similarity. Since not all ontological terms consist of common nouns or short noun phrases, the newly proposed retrofitting architecture enabled to make use of all the available paraphrase information and, thus, to better distinguish between true cases of semantic similarity and cases of semantic association. Additionally, we proposed a novel outlier detection mechanism that successfully detected misalignments without significantly harming the recall capability of the system. To evaluate the performance of our proposed algorithm, we used the biomedical domain as our application, due to its importance, its ontological maturity, and to the fact that it constitutes the domain with the larger ontology alignment datasets owing to its high variability in expressing terms. We compared our method to state-of-the-art ontology matching systems based on feature engineering and showed significant performance gains. Specifically, our algorithm was the top performing algorithm in two of the three ontology mappings tasks while the performance difference in the third ontology matching scenario was really small.

In an extensive ablation study, we empirically validated the effectiveness as well as the importance of the novel phrase retrofitting architecture and the outlier detection mechanism. Furthermore, we illustrated again the importance of retrofitting by demonstrating that the initial pre-trained word vectors do not achieve good performance despite being trained on biomedical corpora. One of the key outcomes of this work was to show that general semantic lexicons can be an adequate source of synonymy information even for the biomedical domain which is characterised by a high degree of linguistic variability. Furthermore, our error analysis of sampled misalignments showed that the retrofitted word vectors demonstrate an even better spatial consistency compared to the pre-trained word vectors; providing additional support to the importance of tailoring the initial representations to semantic similarity. A key outcome of this work was to show that we can drastically decrease the number of similarity functions, and, thus, the number of cut-off thresholds, required for aligning two ontologies. Additionally, our results demonstrated that a great ontology matching performance can be achieved even in the absence of any graph-theoretic information; answering, thus, an open question in the

field of ontology matching. Specifically, our work provided empirical evidence that external corpora and semantic lexicons provide sufficient information to perform ontology matching by only exploiting the ontologies' terms.

Last but not least, in Chapter 5, we studied the problem of discovering general relations between entities appearing in the same ontology or knowledge base. We did so by learning entity and relation embedding by exploiting statistical regularities laying in the ontological or KB facts. We began by clarifying certain misconceptions regarding the expressiveness of the family of translation models. In the next, building on recent research highlighting the advantages of non-Euclidean space, we examined the contribution of geometrical space to the task of knowledge base completion. Despite the fact that the family of translational models has certain advantages with regard to the rules it can effectively represent, recent work has shown that it demonstrates worse performance compared to the bilinear family of shallow knowledge graph embeddings. Our work focused on examining whether the lower performance of translational models in certain datasets is an intrinsic characteristic of them or a restriction that can be lifted by the right choice of geometrical space. We chose the hyperbolic space since recent work had illustrated its advantages for harnessing high quality embeddings for hierarchical and/or scale-free graphs; properties that also appear in ontologies and knowledge bases. We evaluated our method using a variety of link prediction datasets and our experimental results showed that the hyperbolic space allows to narrow down significantly the performance gap between translational and bilinear models; illustrating that the lagging performance of translational models is not an intrinsic characteristic of them.

Another key outcome of our work was to demonstrate that the appropriateness of a KB embedding model should not only be measured in terms of fully expressiveness but also in terms of the rules that it can model. Our experimental results validated the superior performance of translational models against that of the bilinear ones on datasets containing facts that satisfy quasi-chained rules. Therefore, our work pointed to a new promising direction for developing models that, although not fully expressive, allow to better represent certain families of rules; opening up for more fine-grained reasoning tasks. Another contribution of our work was the proof that our proposed translational model in the hyperbolic space is also a prominent candidate for representing effectively quasi-chained rules. Finally, among our contributions was the proposal of a novel KB embedding model as well as a regularisation scheme on the Poincaré-ball model whose effectiveness we proved empirically.

6.2 Future Directions

The results, presented in this thesis, indicate several interesting directions for future research. The last section of this thesis is devoted to discuss some unaddressed or open issues and propose directions for future work.

6.2.1 Ontology Alignment Performance versus Semantic Lexicons

In Chapters 3 and 4, it was demonstrated that we can harness terminological embeddings tailored to ontology alignment by extracting synonymy information from semantic lexicons and the ontologies themselves. It was shown that there exist publicly available semantic lexicons that can provide both domain-neutral as well as domain-specific synonymy information. Furthemore, the results presented in Section 4.3.9 provided supplementary empirical support on the adequacy of the used general semantic lexicons to provide the required synonymy information to train our proposed methods. In accordance with our intuition, the results presented in Sections 3.3.5 and 4.3.9 provided evidence that the greater the coverage of synonyms, the greater the performance of our proposed algorithms will be. It is worth noting that the vast majority of the state-of-the-art systems based on feature engineering also exploit information extracted from semantic lexicons. Nonetheless, despite the fact that our work has illustrated that the number of used similarity metrics can be drastically reduced and the features can be learned instead of being engineered, the problem of identifying the appropriate semantic lexicons remains open.

In the future, it would be interesting to gain further insight into which existing semantic lexicons are more appropriate for the target application domain. For instance, a recommendation system could be constructed to propose, in advance, which is the most appropriate combination of semantic lexicons to be used for a specific ontology matching scenario. One of the main obstacles that such an approach could face is the small sample size that characterises the problem of ontology matching. However, as this thesis illustrated, quite often such problems can be circumvented by the use of a transfer learning approach that successfully generalises to the task at hand. Moreover, considering the example of WikiSynonyms [42] that exploits the Wikipedia redirects to discover terms that are mostly synonymous, another approach could be to devise a robust method to extract synonymous domain-specific terms in an automatic way. The induction of domain-specific lexicons constitutes an active field of research [218, 174, 252] and further research on this regard could also be beneficial for ontology alignment based on evidence provided by the results reported in this thesis.

6.2.2 Further Embedding Models & Spaces

Recent work has demonstrated that Transformer-based architectures [227] can bring significant performance improvements in various NLP tasks [50, 183, 41]. It is interesting to observe that our proposed approaches presented in Chapters 3 and 4 and the aforementioned Transformer-based architectures share one important common characteristic; both of them fine-tune task-agnostic representations, learned in a self-supervised way, to the task at hand. Therefore, it would be a prominent future direction to experiment with this new family of architectures. It is important to note that the architectures presented in Sections 3.2.2 and 4.2.2 can easily be extended to this setting. Moreover, the outlier detection mechanism presented in Section 4.2.3 can also be easily adapted by replacing the Siamese CBOW representations with

the representations computed by a Transformer-based architecture. One possible drawback could be the time increase in computing the pairwise distances between terms, however, there are already existing architectures [191] that drastically decrease the computation time and with future research on this direction, we believe that this shortcoming could be eliminated.

There are also interesting challenges for link prediction. One important direction is to investigate different embedding spaces and which types of rules these can effectively represent. For instance, recent work has explored the usage of quantum embeddings [77] that learn entity and relation embeddings in a way that logical operation can be directly performed over these embeddings. Additionally, it is of equal importance to understand whether SGD-based optimisation algorithms can indeed discover KB embeddings that are consistent according to a specific family of rules. If so, one intriguing question is whether certain regularisation schemes have an implicit bias to discover such solutions. Answers to these questions could be of significant help in speeding up the inference time and fostering interpretability of KB embedding models.

6.2.3 Discovering General Relations between Entities of Distinct Ontologies

The focus of Chapters 3 and 4 has laid in devising terminological embeddings tailored to semantic similarity. This, in turn, enabled the discovery of equivalence relations between entities appearing in different ontologies. Furthermore, the work presented in Chapter 5 focused on learning entity and relation representations in a way that these reflect statistical regularities occurring in a specific ontology or knowledge base. This made possible the discovery of general relations, i.e., not restricted to equivalence relations, between entities of the same ontology or knowledge base. One prominent direction is to extend our work and allow the discovery of general relations between entities appearing in different ontologies. This would be of significant importance since both ontologies and knowledge bases contain a plethora of different relations, e.g., the subsumption, the mereology relation, etc., and the ontological alignments should by no means be restricted to equivalence relations.

One way to approach this problem would be to learn joint representations of terms and entities by coupling the individual losses of the two different tasks and introducing two hyper-parameters to control the effect of the two losses. This could allow the structural information stemming from the the statistical regularities captured in the graph to *flow into* the joint entity-terminological embeddings. This could also help to shed more light on the importance of the structural information for ontology matching. As it was mentioned in Section 4.3.11, our results support that a great ontology matching performance can be achieved even in the absence of any graph-theoretic information. However, the question whether the structural information can be beneficial remains open. From another point of view, the synonymy information that will be captured in the joint entity-terminological embeddings would allow the discovery of general relations between entities from distinct ontologies and knowledge bases, since, now, different ontologies are embedded in the same space and the problem of

predicting general relations can be seen as a link prediction one. Another one approach could be to exploit Transformer-based architectures and treat the triples appearing in ontologies as simple sentences in the form of *subject, predicate, object*. Likewise, the synonymy information can also be treated in a similar setting, e.g., $term\_x$, $is\_synonymous\_to$, $term\_y$. Recent work [28, 180] has started exploring such architectures for jointly embedding sentences and information coming from knowledge bases and this could be a prominent approach for discovering general relations.

6.2.4 A Communicational Approach for Ontology Alignment

As illustrated in the epigraph of this chapter, quite often, humans achieve a mutual understanding through multi-step communication interactions. Nonetheless, the problem of ontology alignment has been traditionally approached in a more static way. Recent work [138, 109] has made the first steps towards incorporating multi-step interactions during the ontology alignment process. However, the usage of representation learning has not yet been explored by this line of research. One interesting approach could be to exploit the recent work in the field of conversational Artificial Intelligence [43, 31] and cast the problem as that of achieving semantic coordination through multi-step communication interactions between two agents whose application domain is described by different ontologies. In that setting, ontological reasoning could be exploited as a source of argumentation and alignments will be accepted if and only if both agents agree that the resulted alignments *respect their common understanding of the world*. One of the possible advantages of this approach is that of the interpretability of the results, especially, in the case where the agents' arguments avoid the phenomenon of language degeneration by respecting the main structural levels of natural language. Another advantage is that the overall alignment procedure could potentially allow for end-to-end differentiation, which is not the case when we cast ontology alignment as an instance of the Stable Marriage problem; opening up, thus, for an end-to-end differentiable approach.

A Appendix

A.1 Omitted Proofs

Proof of Lemma 1: We begin by introducing the TransE model [27]. In the TransE model, the entities and the relations are represented as vectors in the Euclidean space. Let, $\boldsymbol{s}, \boldsymbol{r}, \boldsymbol{o} \in \mathbb{R}^d$ denote the subject, relation and the object embeddings, respectively. The implausibility score for a fact $R(s,o)$ is defined as $||\boldsymbol{s}+\boldsymbol{r}-\boldsymbol{o}||$, where $||.||$ denotes either the ℓ_1 or the ℓ_2 norm. Let P define a set of valid facts. In the following we introduce some additional definitions needed for the introduction of the restrictions.

- A relation r is **reflexive** on a set E of entities if $(e,r,e) \in P$ for all entities $e \in E$.
- A relation r is **symmetric** on a set E of entities if $(e_1,r,e_2) \in P \iff (e_2,r,e_1) \in P$ for all pairs of entities $e_1, e_2 \in E$.
- A relation r is **transitive** on a set E of entities if $(e_1,r,e_2) \in P \wedge (e_2,r,e_3) \in P \Rightarrow (e_1,r,e_3) \in P$ for all $e_1, e_2, e_3 \in E$.

In the following, we list the restrictions mentioned in Kazemi and Poole [115].

- R1 : If a relation r is reflexive on $\Delta \subset E$, r must also be symmetric on Δ.
- R2 : If r is reflexive on $\Delta \subset E$, r must also be transitive on Δ.
- R3 : If entity e_1 has relation r with every entity in $\Delta \subset E$ and entity e_2 has relation r with one of the entities in Δ, then e_2 must have the relation r with every entity in Δ.

Let $n, m \in \mathbb{N}$, $i, j \in \mathbb{R}$ and $a \in \mathbb{R}_+^*$. Let $\boldsymbol{v} = (v_1, v_2, \ldots, v_m) \in \mathbb{R}^m$ and $\boldsymbol{u} \in \mathbb{R}^n$. We denote with $(v_1, v_2, \ldots, v_m; \boldsymbol{u})$ the concatenation of vectors $\boldsymbol{v}$ and $\boldsymbol{u}$. Let $\mathbf{0}_n \in \mathbb{R}^n$ be the zero n-dimensional

vector. For each restriction, we consider a minimum valid set of instances that could satisfy the restriction and we construct a counterexample that satisfies restriction's conditions but not the conclusion. In the following, we assume that ||.|| denotes the ℓ_2 norm. It can be easily verified that these counterexamples also apply, with no modification, when the ℓ_1 norm is used.

R1 : This restriction translates to:

$$\left.\begin{array}{ll} \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_1}\| & \leq a \\ \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_2}\| & \leq a \\ \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_2}\| & \leq a \end{array}\right\} \Rightarrow \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_1}\| \leq a \tag{A.1}$$

Let $n \geq 1$, $\boldsymbol{r} = (a;\boldsymbol{0}_{n-1})$, $\boldsymbol{e_1} = (i-a;\boldsymbol{0}_{n-1})$ and $\boldsymbol{e_2} = (i+a;\boldsymbol{0}_{n-1})$, then:

$$\begin{array}{l} \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_1}\| = \|((i+2a-(i-a));\boldsymbol{0}_{n-1})\| \Rightarrow \\ \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_1}\| = \sqrt{3}a > a \end{array} \tag{A.2}$$

R2 : This restriction translates to:

$$\left.\begin{array}{ll} \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_1}\| & \leq a \\ \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_2}\| & \leq a \\ \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_3}\| & \leq a \\ \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_2}\| & \leq a \\ \|\boldsymbol{e_3}+\boldsymbol{r}-\boldsymbol{e_3}\| & \leq a \end{array}\right\} \Rightarrow \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_3}\| \leq a \tag{A.3}$$

Let $n \geq 1$, $\boldsymbol{r} = (a;\boldsymbol{0}_{n-1})$, $\boldsymbol{e_1} = (i-a;\boldsymbol{0}_{n-1})$, $\boldsymbol{e_2} = (i+a;\boldsymbol{0}_{n-1})$ and $\boldsymbol{e_3} = (i+3a;\boldsymbol{0}_{n-1})$, then:

$$\begin{array}{l} \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_3}\| = \|((i-(i+3a));\boldsymbol{0}_{n-1})\| \Rightarrow \\ \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_3}\| = \sqrt{3}a > a \end{array} \tag{A.4}$$

R3 : This restriction translates to:

$$\left.\begin{array}{ll} \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_1}\| & \leq a \\ \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_2}\| & \leq a \\ \|\boldsymbol{e_1}+\boldsymbol{r}-\boldsymbol{e_3}\| & \leq a \\ \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_3}\| & \leq a \end{array}\right\} \Rightarrow \begin{array}{c} \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_2}\| \leq a \\ \wedge \\ \|\boldsymbol{e_2}+\boldsymbol{r}-\boldsymbol{e_1}\| \leq a \end{array} \tag{A.5}$$

Let $n \geq 2$, $\boldsymbol{r} = (a; \mathbf{0}_{n-1})$, $\boldsymbol{e_1} = (i; \mathbf{0}_{n-1})$, $\boldsymbol{e_2} = (i + \frac{3a}{2}, \frac{a}{2}; \mathbf{0}_{n-2})$ and $\boldsymbol{e_3} = (i + 2a; \mathbf{0}_{n-1})$, then:

$$\begin{aligned} \|\boldsymbol{e_2} + \boldsymbol{r} - \boldsymbol{e_1}\| &= \|(i + \frac{3a}{2} + a - i, \frac{a}{2}; \mathbf{0}_{n-2})\| \Rightarrow \\ \|\boldsymbol{e_2} + \boldsymbol{r} - \boldsymbol{e_1}\| &= \frac{\sqrt{26}}{2} a > a \end{aligned} \tag{A.6}$$

This ends our proof.

Proof of Proposition 1: Let $\|\boldsymbol{s} + \Pi_\beta \boldsymbol{o}\| < 1$, $\|\boldsymbol{r}\| < 1$ and $\lambda_R > 0$, we investigate the type of the geometric locus of the term vectors in the form of $\boldsymbol{s} + \Pi_\beta \boldsymbol{o}$ that satisfy the following equation:

$$d_p(\boldsymbol{s} + \Pi_\beta \boldsymbol{o}, \boldsymbol{r}) \leq \lambda_R \tag{A.7}$$

To simplify the notation, we denote $\boldsymbol{x} := \boldsymbol{s} + \Pi_\beta \boldsymbol{o}$.

$$\begin{aligned} d_p(\boldsymbol{x}, \boldsymbol{r}) &\leq \lambda_R & \iff \\ 1 + 2\delta(\boldsymbol{x}, \boldsymbol{r}) &\leq \cosh(\lambda_R) & \iff \\ \delta(\boldsymbol{x}, \boldsymbol{r}) &\leq \frac{\cosh(\lambda_R) - 1}{2} \end{aligned} \tag{A.8}$$

Let $\alpha = (\cosh(\lambda_R) - 1)/2$. We should note that $\alpha > 0$, since $\forall x \in \mathbb{R}^* : cosh(x) > 1$. Then, we have:

$$\frac{\|\boldsymbol{x} - \boldsymbol{r}\|^2}{(1 - \|\boldsymbol{x}\|^2)(1 - \|\boldsymbol{r}\|^2)} \leq a \tag{A.9}$$

Be setting $\rho = a(1 - \|\boldsymbol{r}\|^2)$, the inequality A.9 becomes:

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{r}\|^2 &\leq \rho(1 - \boldsymbol{x}\|^2) \iff \\ (\rho + 1)\|\boldsymbol{x}\|^2 - 2 * \boldsymbol{x}\boldsymbol{r} + \|\boldsymbol{r}\|^2 &\leq \rho \iff \\ \|\boldsymbol{x}\|^2 - 2 * \boldsymbol{x}\frac{\boldsymbol{r}}{\rho + 1} + \frac{\|\boldsymbol{r}\|^2}{\rho + 1} &\leq \frac{\rho}{\rho + 1} \iff \\ \|\boldsymbol{x} - \frac{\boldsymbol{r}}{\rho + 1}\|^2 &\leq \frac{\rho}{\rho + 1} + \frac{\|\boldsymbol{r}\|^2}{(\rho + 1)^2} - \frac{\|\boldsymbol{r}\|^2}{\rho + 1} \end{aligned} \tag{A.10}$$

We prove in the following that:

$$\frac{\rho}{\rho + 1} + \frac{\|\boldsymbol{r}\|^2}{(\rho + 1)^2} - \frac{\|\boldsymbol{r}\|^2}{\rho + 1} > 0. \tag{A.11}$$

First, we note that since $\|\boldsymbol{r}\| < 1$, we also have that $\rho > 0$ based on the fact that $\alpha > 0$ and

$1 - \|\boldsymbol{r}\|^2 > 0$. Then, we have:

$$\begin{aligned}
&\frac{\rho}{\rho+1} + \frac{\|\boldsymbol{r}\|^2}{(\rho+1)^2} - \frac{\|\boldsymbol{r}\|^2}{\rho+1} = \\
&= \frac{1}{\rho+1}\left(\rho + \frac{\|\boldsymbol{r}\|^2}{\rho+1} - \|\boldsymbol{r}\|^2\right) \\
&= \frac{1}{\rho+1}\left(\rho + \frac{1-\rho-1}{\rho+1}\|\boldsymbol{r}\|^2\right) \\
&= \frac{\rho}{\rho+1}\left(1 - \frac{1}{\rho+1}\|\boldsymbol{r}\|^2\right)
\end{aligned}$$

We observe that $\frac{\rho}{\rho+1} > 0$, hence, it is sufficient to check whether $1 - \frac{1}{\rho+1}\|\boldsymbol{r}\|^2 > 0$. We note that since $\|\boldsymbol{r}\| < 1$ and $\rho > 0$, we have $\frac{\|\boldsymbol{r}\|^2}{\rho+1} < \frac{1}{\rho+1}$. However, $\frac{1}{\rho+1} < 1$. This concludes our proof.

It should be noted that since no specific property of the non-commutative composite vector representation of the pair $(\boldsymbol{s}, \boldsymbol{o})$ was used in the proof above (the proof works for a general $\boldsymbol{x} \in \mathbb{B}^n$), this Proposition also demonstrates a general property of the Poincaré-ball model: its hyperbolic balls correspond to Euclidean balls of different centers and radii.

Notes

1. Let (X, d_X) be a metric space and $x, y \in X$, the *distance-based similarity*, $sim_X(x, y)$, is defined as $sim_X(x, y) = \frac{1}{1+d_X(x,y)}$ [92, p. 46].

2. We provide further justification for this choice in Section 4.3.4.

3. The matrix, used in Möbius multiplication, and the biases are defined on Euclidean space and are learned through Euclidean SGD.

4. https://github.com/schemaorg/schemaorg/blob/sdo-callisto/data/releases/3.2/schema.ttl

5. http://downloads.dbpedia.org/2014/dbpedia_2014.owl.bz2

6. http://oaei.ontologymatching.org/2016/

7. https://code.google.com/p/word2vec

8. This term is known in the NLP community as "conceptually associated". We have chosen to depart from the standard terminology for reasons summarized in [11, p. 7].

9. We provide further details on the textual information used in our experiments in Section 4.3.4.

10. These are available on OAEI's 2016 Large BioMed Track.

11. For a detailed overview and comparison of the systems please refer to [57].

12. We have also performed hyperparameter tuning in the SNOMED-NCI matching task and the resulted hypeparameters were the same as the ones reported in [239].

13. Except for the synonymy information found in some ontologies and is expressed through multiple labels (rdfs:label) for a given type.

14. All the experiments are statistically significant with a p-value ≤ 0.05.

15. Only existential variables can be mapped to labelled nulls.

16. In our experiments, we noticed that a rather small dropout rate had no effect on the model's generalisation capability.

Bibliography

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016.*, pages 265–283, 2016. (Cited on page 54)

[2] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of databases: the logical level.* Addison-Wesley Longman Publishing Co., Inc., 1995. (Cited on page 60)

[3] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, E. Jiménez-Ruiz, E. Kuss, P. Lambrix, H. Leopold, H. Li, C. Meilicke, S. Montanelli, C. Pesquita, T. Saveta, P. Shvaiko, A. Splendiani, H. Stuckenschmidt, K. Todorov, C. T. dos Santos, and O. Zamazal. Results of the ontology alignment evaluation initiative 2016. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18, 2016.*, volume 1766, pages 73–129. RWTH, 2016. (Cited on pages 48 and 49)

[4] L. V. Ahlfors. Invariant operators and integral representations in hyperbolic space. *Mathematica Scandinavica*, 36(1):27–43, 1975. (Cited on page 62)

[5] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1139. (Cited on pages 3 and 15)

[6] G. Alain and Y. Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014. (Cited on page 43)

[7] M. B. Almeida. Revisiting ontologies: A necessary clarification. *Journal of the American Society for Information Science and Technology*, 64(8):1682–1693, 2013. doi: 10.1002/

asi.22861. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22861. (Cited on pages 2 and 9)

[8] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014. (Cited on pages 67 and 68)

[9] S. Anam, Y. S. Kim, B. H. Kang, and Q. Liu. Review of ontology matching approaches and challenges. *International journal of Computer Science and Network Solutions*, 3(3): 1–27, 2015. (Cited on pages 16 and 26)

[10] J. Anderson, E. Rosenfeld, and A. Pellionisz. Neurocomputing 2: Directions for research, 1990. URL https://books.google.ch/books?id=AjpcRQAACAAJ. (Cited on page 12)

[11] R. Arp, B. Smith, and A. D. Spear. *Building ontologies with basic formal ontology*. The MIT Press, Cambridge, Mass., 2015. ISBN 0262527812, 9780262527811. (Cited on pages 10 and 85)

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000. (Cited on page 62)

[13] F. Baader. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, New York, NY, USA, 2003. (Cited on page 11)

[14] I. Balažević, C. Allen, and T. Hospedales. Multi-relational poincaré graph embeddings. In *Advances in Neural Information Processing Systems*, 2019. (Cited on pages 19 and 70)

[15] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286 (5439):509–512, 1999. (Cited on page 60)

[16] J. Barnes. *Porphyry Introduction*. Clarendon later ancient philosophers. Clarendon Press, 2006. ISBN 9780199288694. URL https://books.google.ch/books?id=RFentSvPG3sC. (Cited on page 9)

[17] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012. (Cited on page 54)

[18] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003. (Cited on pages 3, 15, and 22)

[19] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153, 2007. (Cited on pages 17 and 43)

[20] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Austin, TX, June 2010. (Cited on page 54)

[21] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, Oxford, 1995. (Cited on page 36)

[22] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. (Cited on pages xi, 11, and 48)

[23] O. Bodenreider, T. F. Hayamizu, M. Ringwald, S. de Coronado, and S. Zhang. Of mice and men: Aligning mouse and human anatomies. In *AMIA 2005, American Medical Informatics Association Annual Symposium, Washington, DC, USA, October 22-26, 2005*, 2005. URL http://knowledge.amia.org/amia-55142-a2005a-1.613296/t-001-1.616182/f-001-1.616183/a-012-1.616655/a-013-1.616652. (Cited on page 48)

[24] M. Boguná, F. Papadopoulos, and D. Krioukov. Sustaining the internet with hyperbolic mapping. *Nature communications*, 1:62, 2010. (Cited on page 19)

[25] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008. (Cited on page 59)

[26] S. Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Trans. Automat. Contr.*, 58(9):2217–2229, 2013. (Cited on pages 15, 16, and 65)

[27] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013. (Cited on pages 19, 60, 63, 66, 67, 68, 69, 70, and 81)

[28] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1470. URL https://www.aclweb.org/anthology/P19-1470. (Cited on page 79)

[29] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173. (Cited on page 14)

[30] L. Cai and W. Y. Wang. KBGAN: Adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1470–1480, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1133. URL https://www.aclweb.org/anthology/N18-1133. (Cited on pages 18 and 70)

[31] K. Cao, A. Lazaridou, M. Lanctot, J. Z. Leibo, K. Tuyls, and S. Clark. Emergent communication through negotiation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk6WhagRW. (Cited on page 79)

[32] M. Cheatham and P. Hitzler. String similarity metrics for ontology alignment. In *International Semantic Web Conference*, pages 294–309, Heidelberg (DE), 2013. Springer. (Cited on pages 22 and 36)

[33] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 90–98. SIAM, 2017. (Cited on page 18)

[34] M. Chen. Efficient vector representation for documents through corruption. *CoRR*, abs/1707.02377, 2017. URL http://arxiv.org/abs/1707.02377. (Cited on page 42)

[35] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada, June 1989. Association for Computational Linguistics. doi: 10.3115/981623.981633. (Cited on page 2)

[36] A. Coates, H. Lee, and A. Y. Ng. An analysis of single-layer networks in unsupervised feature learning. *Ann Arbor*, 1001(48109):2, 2010. (Cited on page 17)

[37] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12 (Aug):2493–2537, 2011. (Cited on page 37)

[38] I. F. Cruz, F. P. Antonelli, and C. Stroe. Agreementmaker: efficient matching for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment*, 2(2):1586–1589, 2009. (Cited on pages 16, 28, and 30)

[39] I. F. Cruz, A. Fabiani, F. Caimi, C. Stroe, and M. Palmonari. Automatic configuration selection using ontology matching task profiling. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, editors, *The Semantic Web: Research and Applications*, pages 179–194, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-30284-8. (Cited on pages 6 and 57)

[40] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3079–3087. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.pdf. (Cited on pages 3 and 17)

[41] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://www.aclweb.org/anthology/P19-1285. (Cited on page 77)

[42] W. Dakka and P. G. Ipeirotis. Automatic extraction of useful facet hierarchies from text databases. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08, pages 466–475, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-1-4244-1836-7. doi: 10.1109/ICDE.2008.4497455. URL https://doi.org/10.1109/ICDE.2008.4497455. (Cited on pages 29, 47, and 77)

[43] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979, 2017. (Cited on page 79)

[44] D. Davidson. Truth and meaning. *Synthese*, 17(1):304–323, 1967. doi: 10.1007/BF00485035. (Cited on page 1)

[45] S. de Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, and L. W. Wright. NCI thesaurus: Using science-based terminology to integrate cancer research results. In *MEDINFO 2004 - Proceedings of the 11th World Congress on Medical Informatics, San Francisco, California, USA, September 7-11, 2004*, pages 33–37, 2004. doi: 10.3233/978-1-60750-949-3-33. URL https://doi.org/10.3233/978-1-60750-949-3-33. (Cited on page 46)

[46] F. de Saussure and W. Baskin. *Course in General Linguistics: Translated by Wade Baskin. Edited by Perry Meisel and Haun Saussy*. Columbia University Press, 2011. URL http://www.jstor.org/stable/10.7312/saus15726. (Cited on page 2)

[47] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng. Large scale distributed deep networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1223–1231, USA, 2012. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999134.2999271. (Cited on page 54)

[48] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990. (Cited on pages 3, 15, and 22)

[49] T. Dettmers, P. Minervini, P. Stenetorp, and S. 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*, pages 1811–1818, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17366. (Cited on pages 63 and 67)

[50] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423. (Cited on pages 3, 15, 17, and 77)

[51] B. Dhingra, C. Shallue, M. Norouzi, A. Dai, and G. Dahl. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-1708. URL http://aclweb.org/anthology/W18-1708. (Cited on page 19)

[52] W. E. Djeddi and M. T. Khadir. Xmap: a novel structural approach for alignment of owl-full ontologies. In *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, pages 368–373. IEEE, 2010. (Cited on page 30)

[53] W. E. Djeddi and M. T. Khadir. A novel approach using context-based measure for matching large scale ontologies. In L. Bellatreche and M. K. Mohania, editors, *Data Warehousing and Knowledge Discovery*, pages 320–331, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10160-6. (Cited on pages 16 and 49)

[54] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. *Ontology Matching: A Machine Learning Approach*, pages 385–403. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-24750-0. doi: 10.1007/978-3-540-24750-0_19. URL https://doi.org/10.1007/978-3-540-24750-0_19. (Cited on page 17)

[55] Domo. Data Never Sleeps 7.0. https://www.domo.com/learn/data-never-sleeps-7, 2019. [Online; accessed 27-September-2019]. (Cited on page 1)

[56] K. Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006. (Cited on page 46)

[57] Z. Dragisic, V. Ivanova, H. Li, and P. Lambrix. Experiences from the anatomy track in the ontology alignment evaluation initiative. *Journal of biomedical semantics*, 8(1):56, 2017. (Cited on page 85)

[58] T. Ebisu and R. Ichise. Toruse: Knowledge graph embedding on a lie group. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. (Cited on page 19)

[59] J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179 – 211, 1990. ISSN 0364-0213. doi: https://doi.org/10.1016/0364-0213(90)90002-E. URL http://www.sciencedirect.com/science/article/pii/036402139090002E. (Cited on page 3)

[60] J. L. Elman. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582, 2009. doi: 10.1111/j.1551-6709.2009.01023.x. (Cited on page 3)

[61] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić. Introducing wikidata to the linked data web. In *International Semantic Web Conference*, pages 50–65. Springer, 2014. (Cited on page 68)

[62] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013. (Cited on pages 1, 6, 10, 16, 27, 30, and 57)

[63] P. Faber Benítez. The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, 1:107–134, 2009. (Cited on page 36)

[64] M. Fahad, N. Moalla, and A. Bouras. Detection and resolution of semantic inconsistency and redundancy in an automatic ontology merging system. *Journal of Intelligent Information Systems*, 39(2):535–557, 2012. (Cited on page 16)

[65] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999. (Cited on page 60)

[66] R. M. Fano. *Transmission of information : a statistical theory of communications*. M.I.T. Press & Wiley, New York, 1961. (Cited on page 2)

[67] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. Cruz, and F. Couto. The agreementmakerlight ontology matching system. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8185 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 527–541, 11 2013. ISBN 9783642410291. (Cited on pages 16, 48, and 49)

[68] D. Faria, C. Pesquita, E. Santos, I. F. Cruz, and F. M. Couto. Agreementmakerlight 2.0: Towards efficient large-scale ontology matching. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, pages 457–460, Aachen, Germany, Germany, 2014. CEUR-WS.org. URL http://dl.acm.org/citation.cfm?id=2878453.2878568. (Cited on pages 16, 48, and 49)

[69] D. Faria, C. Martins, A. Nanavaty, D. Oliveira, B. Sowkarthiga, A. Taheri, C. Pesquita, F. M. Couto, and I. F. Cruz. Aml results for oaei 2015. In *OM*, pages 116–123, 2015. (Cited on page 28)

[70] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, pages 1606–1615, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N15-1184. (Cited on pages 18, 22, 37, 40, and 42)

[71] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proc. of the 1st Workshop on Evaluating Vector Space Representations for NLP*, 2016. URL http://aclweb.org/anthology/W/W16/W16-2506.pdf. (Cited on page 22)

[72] J. Feng, M. Huang, M. Wang, M. Zhou, Y. Hao, and X. Zhu. Knowledge graph embedding by flexible translation. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'16, pages 557–560. AAAI Press, 2016. URL http://dl.acm.org/citation.cfm?id=3032027.3032102. (Cited on page 65)

[73] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. (Cited on pages 23, 26, 38, 44, and 45)

[74] O. Ganea, G. Becigneul, and T. Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1646–1655, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/ganea18a.html. (Cited on pages 19 and 60)

[75] O. Ganea, G. Becigneul, and T. Hofmann. Hyperbolic neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5345–5355. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7780-hyperbolic-neural-networks.pdf. (Cited on page 13)

[76] J. Gantz and D. Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012):1–16, 2012. (Cited on page 1)

[77] D. Garg, S. Ikbal, S. K. Srivastava, H. Vishwakarma, H. Karanam, and L. V. Subramaniam. Quantum embedding of knowledge for reasoning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5595–5605. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8797-quantum-embedding-of-knowledge-for-reasoning.pdf. (Cited on page 78)

[78] Gartner. Gartner says 5.8 billion enterprise and automotive iot endpoints will be in use in 2020. Available: https://www.gartner.com/en/newsroom/press-releases/2019-08-29-gartner-says-5-8-billion-enterprise-and-automotive-io, 2019. [Online; accessed 27-September-2019]. (Cited on page 1)

[79] L. Getoor and B. Taskar. *Introduction to statistical relational learning*, volume 1. MIT press Cambridge, 2007. (Cited on page 59)

[80] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/glorot10a.html. (Cited on page 65)

[81] R. L. Goldstone and B. J. Rogosky. Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84(3):295–320, 2002. (Cited on page 21)

[82] S. J. Gould. *The panda's thumb: More reflections in natural history*. WW Norton & company, 1992. (Cited on page 2)

[83] A. Groß, C. Pruski, and E. Rahm. Evolution of biomedical ontologies and mappings: Overview of recent approaches. *Computational and structural biotechnology journal*, 14:333–340, 2016. (Cited on page 10)

[84] A. Gu, F. Sala, B. Gunel, and C. Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2018. (Cited on page 71)

[85] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43(5):625 – 640, 1995. ISSN 1071-5819. doi: https://doi.org/10.1006/ijhc.1995.1066. URL http://www.sciencedirect.com/science/article/pii/S107158198571066X. (Cited on pages 2 and 12)

[86] N. Guarino. Formal ontology in information systems. In *Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*, pages 3–15, Amsterdam, The Netherlands, The Netherlands, 1998. IOS Press. ISBN 9051993994. (Cited on page 2)

[87] R. V. Guha, D. Brickley, and S. MacBeth. Schema.org: Evolution of structured data on the web. *Queue*, 13(9):10–37, Nov. 2015. ISSN 1542-7730. doi: 10.1145/2857274.2857276. URL https://doi.org/10.1145/2857274.2857276. (Cited on page 30)

[88] M. Gulić, B. Vrdoljak, and M. Banek. Cromatcher: An ontology matching system based on automated weighted aggregation and iterative final alignment. *Web Semantics: Science, Services and Agents on the World Wide Web*, 41:50–71, 2016. (Cited on pages 16, 30, and 49)

[89] V. Gutiérrez-Basulto and S. Schockaert. From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules. In *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2018. (Cited on pages 60, 61, 65, and 66)

[90] A. Halevy. Why your data won't mix. *Queue*, 3(8):50–58, Oct. 2005. ISSN 1542-7730. doi: 10.1145/1103822.1103836. URL http://doi.acm.org/10.1145/1103822.1103836. (Cited on page 1)

[91] X. Han, S. Cao, X. Lv, Y. Lin, Z. Liu, M. Sun, and J. Li. Openke: An open toolkit for knowledge embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 139–144. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-2024. (Cited on page 69)

[92] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8 (1):1–254, 2015. (Cited on pages 3, 16, and 85)

[93] Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. (Cited on pages 2, 3, 22, and 37)

[94] S. Hawkins, H. He, G. Williams, and R. Baxter. Outlier detection using replicator neural networks. In Y. Kambayashi, W. Winiwarter, and M. Arikawa, editors, *Data Warehousing and Knowledge Discovery*, pages 170–180, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-46145-6. (Cited on page 18)

[95] T. F. Hayamizu, M. Mangan, J. P. Corradi, J. A. Kadin, and M. Ringwald. The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Genome biology*, 6(3): R29, 2005. (Cited on page 46)

[96] J. Heflin and J. Hendler. Semantic interoperability on the web. Technical report, MARYLAND UNIV COLLEGE PARK DEPT OF COMPUTER SCIENCE, 2000. (Cited on page 2)

[97] F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015. URL http://www.aclweb.org/anthology/J15-4004. (Cited on pages 22, 37, and 54)

[98] F. Hill, K. Cho, and A. Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377. Association for Computational Linguistics, 2016. URL http://www.aclweb.org/anthology/N16-1162. (Cited on pages 33, 40, and 53)

[99] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. (Cited on pages xi, 18, 40, 41, 42, and 47)

[100] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. *Distributed Representations*, page 77–109. MIT Press, Cambridge, MA, USA, 1986. ISBN 026268053X. (Cited on page 12)

[101] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, Nov. 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735. (Cited on page 3)

[102] H. Hotelling. Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology*, 24:417–441, 1933. (Cited on page 3)

[103] IOF. Industrial Ontologies Foundry. https://www.industrialontologies.org/, 2020. [Online; accessed 10-February-2020]. (Cited on page 5)

[104] R. Jackendoff and R. S. Jackendoff. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, USA, 2002. (Cited on page 1)

[105] G. Ji, S. He, L. Xu, K. Liu, and J. 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)*, pages 687–696. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1067. URL http://aclweb.org/anthology/P15-1067. (Cited on pages 19 and 63)

[106] E. Jiménez-Ruiz and B. Cuenca Grau. Logmap: Logic-based and scalable ontology matching. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist, editors, *The Semantic Web – ISWC 2011*, pages 273–288, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25073-6. (Cited on pages 16, 28, 30, 48, and 49)

[107] E. Jiménez-Ruiz, B. C. Grau, I. Horrocks, and R. Berlanga. Ontology integration using mappings: Towards getting the right logical consequences. In *European Semantic Web Conference*, pages 173–187, Heidelberg (DE), 2009. Springer. (Cited on page 11)

[108] E. Jiménez-Ruiz, B. C. Grau, I. Horrocks, and R. Berlanga. Logic-based assessment of the compatibility of umls ontology sources. *Journal of biomedical semantics*, 2(1):S2, 2011. (Cited on pages 48 and 49)

[109] E. Jiménez-Ruiz, T. R. Payne, A. Solimando, and V. Tamma. Limiting logical violations in ontology alignment through negotiation. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'16, page 217–226. AAAI Press, 2016. (Cited on page 79)

[110] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631 0374275637. URL https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7. (Cited on page 73)

[111] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665. Association for Computational Linguistics, 2014. URL http://www.aclweb.org/anthology/P14-1062. (Cited on page 17)

[112] Y. Kalfoglou and Y. Kalfoglou. *Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2009. ISBN 160566894X, 9781605668949. (Cited on pages 1 and 73)

[113] M. Karpathiotakis. Just-in-time analytics over heterogeneous data and hardware. page 201, 2017. doi: 10.5075/epfl-thesis-8077. URL http://infoscience.epfl.ch/record/232585. (Cited on page 1)

[114] V. Kashyap and A. Sheth. *Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies*. 1998. (Cited on page 1)

[115] S. M. Kazemi and D. Poole. Simple embedding for link prediction in knowledge graphs. In *Advances in neural information processing systems*, pages 4284–4295, 2018. (Cited on pages 18, 60, 61, 65, 66, 70, and 81)

[116] T. Kenter, A. Borisov, and M. de Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 941–951, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1089. (Cited on pages xi, 12, 17, 40, and 42)

[117] M. Khadir, A. Djeddai, and W. Djeddi. Xmap++: A novel semantic approach for alignment of owl-full ontologies based on semantic relationship using wordnet. In *Innovation in Information & Communication Technology (ISIICT), 2011 Fourth International Symposium on*, pages 13–18. IEEE, 2011. (Cited on page 16)

[118] D. Kiela, F. Hill, and S. Clark. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1242. URL http://www.aclweb.org/anthology/D15-1242. (Cited on pages 36, 37, and 54)

[119] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 47)

[120] P. Kolyvakis, A. Kalousis, and D. Kiritsis. DeepAlignment: Unsupervised ontology matching with refined word vectors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 787–798, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1072. (Cited on page 7)

[121] P. Kolyvakis, A. Kalousis, B. Smith, and D. Kiritsis. Biomedical ontology alignment: an approach based on representation learning. *Journal of Biomedical Semantics*, 9(1):21, 2018. ISSN 2041-1480. doi: 10.1186/s13326-018-0187-8. URL https://doi.org/10.1186/s13326-018-0187-8. (Cited on pages 7 and 12)

[122] P. Kolyvakis, A. Kalousis, and D. Kiritsis. HyperKG: Hyperbolic knowledge graph embeddings for knowledge base completion. *arXiv preprint arXiv:1908.04895*, 2019. (Cited on page 7)

[123] P. Kolyvakis, A. Kalousis, and D. Kiritsis. Hyperbolic knowledge graph embeddings for knowledge base completion. In *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Greece, May 31 - June 4, 2020, Proceedings*, Lecture Notes in Computer Science. Springer, 2020. (Cited on page 7)

[124] K. Kotis, A. Katasonov, and J. Leino. Aligning smart and control entities in the iot. In S. Andreev, S. Balandin, and Y. Koucheryavy, editors, *Internet of Things, Smart Spaces, and Next Generation Networking*, pages 39–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-32686-8. (Cited on page 5)

[125] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá. Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82:036106, Sep 2010. doi: 10.1103/PhysRevE.82.036106. URL https://link.aps.org/doi/10.1103/PhysRevE.82.036106. (Cited on page 60)

[126] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 957–966, 2015. (Cited on page 25)

[127] T. Lacroix, N. Usunier, and G. Obozinski. Canonical tensor decomposition for knowledge base completion. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2863–2872, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/lacroix18a.html. (Cited on pages 18, 63, and 70)

[128] M. Le, S. Roller, L. Papaxanthos, D. Kiela, and M. Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3231–3241, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1313. URL https://www.aclweb.org/anthology/P19-1313. (Cited on page 19)

[129] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/le14.html. (Cited on page 37)

[130] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 265–272, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104516. (Cited on page 54)

[131] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. (Cited on pages 30 and 59)

[132] M. Leimeister and B. J. Wilson. Skip-gram word embeddings in hyperbolic space. *CoRR*, abs/1809.01498, 2018. URL http://arxiv.org/abs/1809.01498. (Cited on page 15)

[133] A. Lenci. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20:1–31, 01 2008. (Cited on page 2)

[134] M. Li, Y. Jia, Y. Wang, J. Li, and X. Cheng. Hierarchy-based link prediction in knowledge graphs. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 77–78, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4144-8. doi: 10.1145/2872518.2889387. URL https://doi.org/10.1145/2872518.2889387. (Cited on page 60)

[135] Z. Li and D. Hoiem. Learning without forgetting. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 614–629, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0. (Cited on page 41)

[136] C. Lofi. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. *Database Society of Japan (DBSJ) Journal*, 14(1):1–9, 03/2016 2016. ISSN ISSN 2189-0390. (Cited on pages 3, 22, and 36)

[137] W. T. Loring. *An introduction to manifolds*. Springer New York, 2008. (Cited on page 13)

[138] P. Maio and N. Silva. An extensible argument-based ontology matching negotiation approach. *Science of Computer Programming*, 95:3 – 25, 2014. ISSN 0167-6423. doi: https://doi.org/10.1016/j.scico.2014.01.011. URL http://www.sciencedirect.com/science/article/pii/S0167642314000264. Special Issue on Systems Development by Means of Semantic Technologies. (Cited on page 79)

[139] F. Manne, M. Naim, H. Lerring, and M. Halappanavar. On stable marriages and greedy matchings. In *2016 Proceedings of the Seventh SIAM Workshop on Combinatorial Scientific Computing*, pages 92–101. SIAM, 2016. (Cited on page 55)

[140] M. Mao, Y. Peng, and M. Spring. Ontology mapping: As a binary classification problem. In *Fourth International Conference on Semantics, Knowledge and Grid, SKG '08, Beijing, China, December 3-5, 2008*, pages 20–25, 2008. doi: 10.1109/SKG.2008.101. URL https://doi.org/10.1109/SKG.2008.101. (Cited on pages 17 and 37)

[141] M. Mao, Y. Peng, and M. Spring. Ontology mapping: As a binary classification problem. *Concurr. Comput. : Pract. Exper.*, 23(9):1010–1025, June 2011. ISSN 1532-0626. doi: 10.1002/cpe.1633. URL http://dx.doi.org/10.1002/cpe.1633. (Cited on pages 17 and 36)

[142] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller. A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1996–2000. IEEE, 2015. (Cited on page 18)

[143] M. Markou and S. Singh. Novelty detection: a review—part 2: neural network based approaches. *Signal Processing*, 83(12):2499 – 2521, 2003. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2003.07.019. URL http://www.sciencedirect.com/science/article/pii/S0165168403002032. (Cited on page 18)

[144] B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors.pdf. (Cited on pages 3 and 17)

[145] D. McVitie and L. B. Wilson. Stable marriage assignment for unequal sets. *BIT Numerical Mathematics*, 10(3):295–309, 1970. (Cited on pages 26, 44, and 54)

[146] D. A. Medler. A brief history of connectionism. *Neural Computing Surveys*, 1:61–101, 1998. (Cited on page 12)

[147] C. Meilicke. *Alignment incoherence in ontology matching*. PhD thesis, University of Mannheim, 2011. URL https://ub-madoc.bib.uni-mannheim.de/29351. (Cited on page 16)

[148] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144, 2008. (Cited on page 46)

[149] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *International Conference on Learning Representations*, 2013. URL http://arxiv.org/abs/1301.3781. (Cited on pages 3, 15, 17, 22, 32, and 37)

[150] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. 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.*, pages 3111–3119, 2013. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality. (Cited on pages 3, 15, 22, 37, and 40)

[151] G. Miller. *WordNet: An electronic lexical database*. MIT press, 1998. (Cited on page 59)

[152] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995. (Cited on page 28)

[153] J. Mitchell and M. Lapata. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P08-1028. (Cited on pages 37, 39, 40, and 60)

[154] S. Mittal, A. Joshi, and T. Finin. Thinking, fast and slow: Combining vector spaces and knowledge graphs. *arXiv preprint arXiv:1708.03310*, 2017. (Cited on page 74)

[155] F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer, 2005. (Cited on page 15)

[156] N. Mrkšić, D. Ó Séaghdha, B. Thomson, M. Gašić, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-1018. (Cited on pages 18, 22, 24, 37, 40, and 42)

[157] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994. (Cited on page 59)

[158] K. Munn and B. Smith. *Applied Ontology: An Introduction*. Frankfurt: ontos, 2008. (Cited on page 2)

[159] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana. Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 83–84. International World Wide Web Conferences Steering Committee, 2016. (Cited on page 25)

[160] R. Navigli and S. P. Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-1023. (Cited on page 47)

[161] R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. (Cited on page 47)

[162] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005. (Cited on page 67)

[163] D. H. Ngo and Z. Bellahsene. Yam++:(not) yet another matcher for ontology matching task. In *BDA'2012: 28e journées Bases de Données Avancées*, pages N–A, 2012. (Cited on page 16)

[164] D. Q. Nguyen, T. D. Nguyen, D. Q. Nguyen, and D. Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of*

the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 327–333. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2053. URL http://aclweb.org/anthology/N18-2053. (Cited on pages xiii and 70)

[165] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347, 2017. (Cited on pages 19, 60, and 65)

[166] M. Nickel and D. Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3779–3788, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/nickel18a.html. (Cited on pages 19 and 60)

[167] M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 809–816, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104584. (Cited on pages 19 and 60)

[168] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016. (Cited on pages 12, 18, and 60)

[169] M. Nickel, L. Rosasco, and T. Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1955–1961. AAAI Press, 2016. URL http://dl.acm.org/citation.cfm?id=3016100.3016172. (Cited on pages 19, 60, and 63)

[170] N. F. Noy, M. A. Musen, J. L. Mejino, and C. Rosse. Pushing the envelope: challenges in a frame-based representation of human anatomy. *Data & Knowledge Engineering*, 48(3): 335–359, 2004. (Cited on page 46)

[171] L. Otero-Cerdeira, F. J. Rodríguez-Martínez, and A. Gómez-Rodríguez. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949–971, 2015. (Cited on page 16)

[172] F. Papadopoulos, R. Aldecoa, and D. Krioukov. Network geometry inference using common neighbors. *Physical Review E*, 92(2):022807, 2015. (Cited on page 19)

[173] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. (Cited on page 54)

[174] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings,

and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-2070. (Cited on pages 29 and 77)

[175] K. Pearson. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901. doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720. (Cited on page 3)

[176] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1162. (Cited on pages 3, 15, 22, and 37)

[177] A. Pentina and S. Ben-David. Multi-task and lifelong learning of kernels. In K. Chaudhuri, C. Gentile, and S. Zilles, editors, *Algorithmic Learning Theory*, pages 194–208, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24486-0. (Cited on page 17)

[178] C. Pesquita, D. Faria, E. Santos, and F. M. Couto. To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013.*, pages 13–24, 2013. URL http://ceur-ws.org/Vol-1111/om2013_Tpaper2.pdf. (Cited on page 48)

[179] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. 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)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202. (Cited on pages 3, 15, and 17)

[180] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. 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)*, pages 2463–2473, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://www.aclweb.org/anthology/D19-1250. (Cited on page 79)

[181] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM)*, pages 39–44, 2013. (Cited on pages 47 and 52)

[182] W. V. O. Quine. Two dogmas of empiricism. *Philosophical Review*, 60(1):20–43, 1951. doi: 10.2307/2266637. (Cited on page 1)

[183] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018. (Cited on page 77)

[184] M. A. L. Ralph, E. Jefferies, K. Patterson, and T. T. Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42, 2017. (Cited on page 1)

[185] T. M. Rassias and T. Suksumran. An inequality related to möbius transformations. *arXiv preprint arXiv:1902.05003*, 2019. (Cited on page 62)

[186] M. Richardson and P. Domingos. Markov logic networks. *Machine learning*, 62(1-2): 107–136, 2006. (Cited on page 59)

[187] C. Rosse and J. L. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500, 2003. (Cited on page 46)

[188] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009. ISBN 0136042597, 9780136042594. (Cited on page 2)

[189] M. Sahlgren. The distributional hypothesis. *Italian Journal of Disability Studies*, 20: 33–53, 2008. (Cited on page 2)

[190] F. Sala, C. De Sa, A. Gu, and C. Re. Representation tradeoffs for hyperbolic embeddings. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4460–4469, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/sala18a.html. (Cited on pages 19, 60, and 64)

[191] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*, 2019. (Cited on page 78)

[192] D. Schluter. Ecology and the origin of species. *Trends in Ecology & Evolution*, 16(7): 372–380, 2001. ISSN 0169-5347. doi: https://doi.org/10.1016/S0169-5347(01)02198-X. URL http://www.sciencedirect.com/science/article/pii/S016953470102198X. (Cited on page 2)

[193] U. Schwarz and B. Smith. Ontological relations. *Applied Ontology. An Introduction*, 219: 234, 2008. (Cited on page 68)

[194] I. Sergienya and H. Schütze. Learning better embeddings for rare words using distributional representations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 280–285, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1033. URL http://www.aclweb.org/anthology/D15-1033. (Cited on page 46)

[195] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135. (Cited on page 14)

[196] P. Shvaiko and J. Euzenat. Ten challenges for ontology matching. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1164–1182. Springer, 2008. (Cited on page 21)

[197] P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2013. (Cited on pages 1, 16, 21, and 36)

[198] B. Smith. Ontology. In L. Floridi, editor, *Blackwell Guide to the Philosophy of Computing and Information*, pages 155–166. Oxford: Blackwell, 2003. (Cited on page 2)

[199] B. Smith and K. Mulligan. Framework for formal ontology. *Topoi*, 2(1):73–85, Jun 1983. ISSN 1572-8749. doi: 10.1007/BF00139703. URL https://doi.org/10.1007/BF00139703. (Cited on page 2)

[200] B. Smith, W. Kusnierczyk, D. Schober, and W. Ceusters. Towards a reference terminology for ontology research and development in the biomedical domain. In *KR-MED 2006, Formal Biomedical Knowledge Representation, Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation: "Biomedical Ontology in Action" (KR-MED 2006), Collocated with the 4th International Conference on Formal Ontology in Information Systems (FOIS-2006), Baltimore, Maryland, USA, November 8, 2006*, volume 2006, pages 57–66, 2006. (Cited on pages 35 and 36)

[201] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, S. Lewis, and T. O. Consortium. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007. ISSN 1546-1696. doi: 10.1038/nbt1346. URL https://doi.org/10.1038/nbt1346. (Cited on page 5)

[202] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 151–161, 2011. URL http://www.aclweb.org/anthology/D11-1014. (Cited on page 17)

[203] R. Socher, D. Chen, C. D. Manning, and A. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934, 2013. (Cited on page 60)

[204] A. Solimando, E. Jiménez-Ruiz, and G. Guerrini. Detecting and correcting conservativity principle violations in ontology-to-ontology mappings. In *International Semantic Web Conference*, pages 1–16. Springer International Publishing Switzerland, 2014. (Cited on page 11)

[205] S. Song, X. Zhang, and G. Qin. Multi-domain ontology mapping based on semantics. *Cluster Computing*, 20(4):3379–3391, 2017. (Cited on page 17)

[206] R. Sorabji. *Aristotle on memory*. Duckworth, London, second edition. edition, 2004. ISBN 0715632396. (Cited on page 12)

[207] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. (Cited on page 36)

[208] R. Speer and C. Havasi. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3679–3686, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/1072.html. (Cited on page 47)

[209] M. D. Spivak. *A comprehensive introduction to differential geometry*, volume I. Publish or perish, 3rd edition, 1999. (Cited on pages 13 and 14)

[210] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html. (Cited on page 64)

[211] M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005. (Cited on pages 60 and 67)

[212] A. Stolk, L. Verhagen, and I. Toni. Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, 20(3):180–191, 2016. (Cited on page 21)

[213] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007. (Cited on page 59)

[214] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkgEQnRqYQ. (Cited on pages 18 and 70)

[215] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf. (Cited on pages 3 and 17)

[216] A. Suzuki, Y. Enokida, and K. Yamanishi. Riemannian transe: Multi-relational graph embedding in non-euclidean space, 2019. URL https://openreview.net/forum?id=r1xRW3A9YX. (Cited on page 19)

[217] J. M. Taylor. Mapping human understanding to robotic perception. *Procedia Computer Science*, 56:514–519, 2015. (Cited on pages 5 and 21)

[218] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 214–221. Association for Computational Linguistics, July 2002. doi: 10.3115/1118693.1118721. URL https://www.aclweb.org/anthology/W02-1028. (Cited on page 77)

[219] A. Tifrea, G. Becigneul, and O.-E. Ganea. Poincare glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Ske5r3AqK7. (Cited on page 19)

[220] K. Toutanova and D. Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66. Association for Computational Linguistics, 2015. doi: 10.18653/v1/W15-4007. URL http://aclweb.org/anthology/W15-4007. (Cited on page 67)

[221] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080, 2016. (Cited on pages 19, 60, 69, and 70)

[222] A. Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977. (Cited on page 36)

[223] A. A. Ungar. *Beyond the Einstein addition law and its gyroscopic Thomas precession: The theory of gyrogroups and gyrovector spaces*, volume 117. Springer Science & Business Media, 2012. (Cited on pages 62 and 71)

[224] M. Uschold. Where are the semantics in the semantic web? *Ai Magazine*, 24(3):25–25, 2003. (Cited on page 2)

[225] M. Uschold and M. Gruninger. Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, 33(4):58–64, Dec. 2004. ISSN 0163-5808. doi: 10.1145/1041410.1041420. URL http://doi.acm.org/10.1145/1041410.1041420. (Cited on pages 1 and 2)

[226] R. Van Der Hofstad. Random graphs and complex networks. *Available on http://www. win. tue. nl/rhofstad/NotesRGCN. pdf*, 11, 2009. (Cited on page 60)

[227] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. (Cited on pages 3, 17, and 77)

[228] K. H. Veltman. Syntactic and semantic interoperability: new approaches to knowledge and the semantic web. *New Review of Information Networking*, 7(1):159–183, 2001. (Cited on page 1)

[229] V. Ventrone. Semantic heterogeneity as a result of domain evolution. *ACM SIGMOD Record*, 20(4):16–20, 1991. (Cited on page 1)

[230] K. Verbeek and S. Suri. Metric embedding, hyperbolic space, and social networks. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 501. ACM, 2014. (Cited on page 19)

[231] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294. URL http://doi.acm.org/10.1145/1390156.1390294. (Cited on pages 12, 18, 39, 43, and 44)

[232] D. Vrandečić and M. Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, Sept. 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL http://doi.acm.org/10.1145/2629489. (Cited on page 68)

[233] C. Wang, L. Cao, and B. Zhou. Medical synonym extraction with concept space models. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 989–995, Buenos Aires, Argentina, 2015. AAAI Press. URL http://dl.acm.org/citation.cfm?id=2832249.2832386. (Cited on page 46)

[234] P. Wang and B. Xu. Lily: Ontology alignment results for oaei 2008. In *Proceedings of the 3rd International Conference on Ontology Matching-Volume 431*, pages 167–175. CEUR-WS. org, 2008. (Cited on page 16)

[235] Q. Wang, Z. Mao, B. Wang, and L. Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017. (Cited on page 18)

[236] Y. Wang, R. Gemulla, and H. Li. On multi-relational link prediction with bilinear models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. (Cited on pages 60, 65, and 66)

[237] Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. 2014. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531. (Cited on pages 19 and 63)

[238] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd international conference on World wide web*, pages 515–526. ACM, 2014. (Cited on pages 2 and 60)

[239] J. Wieting, M. Bansal, K. Gimpel, K. Livescu, and D. Roth. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358, 2015. ISSN 2307-387X. (Cited on pages 28, 37, 42, 50, 51, 53, and 85)

[240] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. Towards universal paraphrastic sentence embeddings. In *Proceedings of International Conference on Learning Representations*, 2016. (Cited on pages 17, 18, 37, and 40)

[241] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In *Ordered sets*, pages 445–470. Springer, Dordrecht, 1982. (Cited on pages 16 and 49)

[242] G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu. A comparative study of rnn for outlier detection in data mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 709–712, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1754-4. URL http://dl.acm.org/citation.cfm?id=844380.844788. (Cited on page 18)

[243] M. Wooldridge. Intelligent agents. *Multiagent systems*, 35(4):51, 1999. (Cited on page 2)

[244] C. Xiang, T. Jiang, B. Chang, and Z. Sui. Ersom: A structural ontology matching approach using automatically learned entity representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2419–2429, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/D15-1289. (Cited on pages 17 and 26)

[245] H. Xiao, M. Huang, and X. Zhu. From one point to a manifold: Knowledge graph embedding for precise link prediction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1315–1321. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3060621.3060804. (Cited on pages 19 and 65)

[246] Q. Xie, X. Ma, Z. Dai, and E. Hovy. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–962. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1088. URL http://aclweb.org/anthology/P17-1088. (Cited on pages 60 and 63)

[247] W. Xiong, M. Yu, S. Chang, X. Guo, and W. Y. Wang. One-shot relational learning for knowledge graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1980–1990, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1223. (Cited on page 68)

[248] D. Xu, Y. Yan, E. Ricci, and N. Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156(Supplement C):117 – 127, 2017. ISSN 1077-3142. URL http://www.sciencedirect.com/science/article/pii/S1077314216301618. Image and Video Understanding in Big Data. (Cited on page 18)

[249] B. Yang, W. tau Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations*, volume abs/1412.6575, 2015. (Cited on pages 19, 60, and 70)

[250] A. Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. doi: 10.3115/992730.992783. URL https://doi.org/10.3115/992730.992783. (Cited on pages 28 and 49)

[251] M. D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012. URL http://arxiv.org/abs/1212.5701. (Cited on page 47)

[252] M. Zhang, Y. Liu, H. Luan, and M. Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL https://www.aclweb.org/anthology/P17-1179. (Cited on page 77)

[253] S. Zhang and O. Bodenreider. Experience in aligning anatomical ontologies. *International journal on Semantic Web and information systems*, 3(2):1, 2007. (Cited on pages 36, 46, and 47)

[254] Y. Zhang, X. Wang, S. Lai, S. He, K. Liu, J. Zhao, and X. Lv. Ontology matching with word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 34–45. Springer, Berlin, 2014. (Cited on page 17)

[255] M. Zhao and S. Zhang. Identifying and validating ontology mappings by formal concept analysis. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18, 2016.*, pages 61–72, 2016. URL http://ceur-ws.org/Vol-1766/om2016_Tpaper6.pdf. (Cited on pages 16 and 49)

[256] M. Zhao, S. Zhang, W. Li, and G. Chen. Matching biomedical ontologies based on formal concept analysis. *Journal of biomedical semantics*, 9(1):11, 2018. (Cited on pages 16 and 49)

[257] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949. (Cited on page 67)

Prodromos **Kolyvakis**

Ph.D Student in Robotics, Control & Intelligent Systems
Place de la Riponne 4, 1005, Lausanne, VD, Switzerland
(+41) 76-690-5989 | prokolyvakis@gmail.com | prokolyvakis | prodromoskolyvakis

"Rise above oneself and grasp the world." – Archimedes

Interests

Applying deep learning for NLP applications in the presence of structured, semi-structured and unstructured data.

Experience

École Polytechnique Fédérale de Lausanne — *Lausanne, Switzerland*

Doctoral Assistant — *Feb. 2016 - May 2020*

- Working on Ontology Matching & Knowledge Base Completion exploiting representation learning techniques.

Research Assistant in the H2020 bIoTope project — *Feb. 2016 - May 2019*

- Working on techniques for establishing semantic interoperability in IoT devices.

Amazon Alexa AI — *Cambridge, Massachusetts*

Applied Scientist Intern — *May 2019 - August. 2019*

Education

École Polytechnique Fédérale de Lausanne (EPFL) — *Lausanne, Switzerland*

Ph.D. Student in Robotics, Control and Intelligent Systems — *Feb. 2016 - May 2020*

- Thesis: Approaching Ontology Alignment through Representation Learning to Bridge the Semantic Gap in Engineering Applications
- Advisor: Prof. Dimitris Kiritsis

National Technical University of Athens (NTUA) — *Athens, Greece*

Dipl.-Ing. in Electrical and Computer Engineering — *Sep. 2009 - Nov. 2014*

- Major: Computer Science – Minor: Control Theory, Robotics and Applied Math
- Grade: 8.66/10 "Very Good" – Top 10%
- Thesis: Study of a novel automatic writer identification method and application to important ancient documents

Honors & Awards

2018	**Program Committee Member**, 14th International Workshop on Ontology Matching	*Lausanne, Vaud*
2014	**Ranked in Top 6%,** IEEEXtreme Programming Competition 8.0	*Athens, Greece*
2009	**Excellence Prize**, by the Greek Youth Learning – "Morfosis Neolaias" – foundation	*Ioannina, Greece*

Selected Publications

- Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. DeepAlignment: Unsupervised Ontology Matching with Refined Word Vectors. In *NAACL 2018, Volume 1 (Long Papers)*, 2018.
- Prodromos Kolyvakis, Alexandros Kalousis, Barry Smith, and Dimitris Kiritsis. Biomedical ontology alignment: an approach based on representation learning. *Journal of Biomedical Semantics*, 9(1):21, Aug 2018.
- Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. Hyperbolic knowledge graph embeddings for knowledge base completion. In *ESWC 2020*.

Skills

Programming & Scripting	Python, C, C++, R, Javascript, Haskell, SQL, SPARQL, Bash Shell, LaTeX
Frameworks & Technologies	Pandas, NumPy, SciPy, scikit-learn, TensorFlow, Keras, PyTorch, NLTK, Apache Spark, Node JS, Amazon EC2
Continuous Integration	Git, Docker, Jenkins
Databases	MySQL, Blazegraph, MongoDB, Neo4j
Languages	Greek, English, French