Learning Knowledge Graph Embeddings for Natural Language Processing

Muhao Chen
Department of Computer Science
University of California, Los Angeles
Winter 2017

Abstract

Knowledge graph embeddings provide powerful latent semantic representation for the structured knowledge in knowledge graphs, which have been introduced recently. Being different from the already widely-used word embeddings that are conceived from plain text, knowledge graph embeddings enable direct explicit relational inferences among entities via simple calculation of embedding vectors. In particular, they are quite effective at highlighting key concepts underlying sophisticated human languages. Therefore knowledge graph embeddings provide potent tools for modern NLP applications, inasmuch as they well-preserve the multi-faceted knowledge and structures of the knowledge graphs. However, recent research efforts have not progressed much beyond representing simple or multi-mapping relations (e.g. one-to-many, many-to-many) for monolingual knowledge graphs. Many crucial problems, including how to preserve important relational properties, and how to characterize both monolingual and cross-lingual knowledge in multiple language-specific versions of the knowledge bases, still remain largely unsolved. Another pressing challenge is, how to incorporate knowledge graph embeddings into NLP tasks which currently rely on word embeddings or other representation techniques.

In this prospectus, we first propose new models for encoding the multi-faceted knowledge as stated. We start from investigating the approach that captures cross-lingual transitions across difference language-specific versions of embedding spaces, while in each embedding space the monolingual relations are well-preserved. We then study the approach to retain the important relational properties that commonly exist in domain-specific and ontology-level knowledge graphs, including transitivity, symmetry, and hierarchies. After that, we explore how our new embedding models may be used to improve modern NLP tasks, including relation extraction, knowledge alignment, semantic relatedness analysis, and sentiment analysis.

Contents

Chapter 1

Introduction

Nowadays, as computer systems are expected to be more and more intelligent, techniques that help modern applications to understand human languages are in much demand. Amongst all the techniques, the latent semantic models are the most indispensable, for they exploit the latent semantics of lexicons and concepts of human languages, and transform them into tractable and machine-understandable numerical representations. Without which, languages are nothing but combinations of meaningless symbols for the machine. To provide such learning representation, in recent years, embedding models for knowledge graphs have attracted much attention, since they intuitively transform important concepts and entities in human languages into vector representations, and realize relational inferences among them via simple vector calculation. Such novel techniques have effective resolved a few tasks like knowledge graph completion and link prediction, and show the great potential to be incorporated into more natural language processing (NLP) applications.

Knowledge graph embeddings are induced from the multi-faceted and structured information stored in knowledge bases. These knowledge bases, including Wikipedia [Wik16], WordNet [BF13], and ConceptNet [SH13], are modeled as knowledge graphs that store two aspects of knowledge: the *monolingual knowledge* that includes entities and relations recorded in the form of triples, and the *cross-lingual knowledge* that matches the monolingual knowledge among various human languages.

The past half decade of research has paid much attention to learning embeddings for mono-

lingual knowledge, which introduced methods to encode entities in low-dimensional embedding spaces and capture relations as means of translations among entity vectors. Given a triple (h, r, t) where r is the relation between entities h and t, then h and t are represented as two k-dimensional vectors $\mathbf{h}$ and $\mathbf{t}$, respectively. A score function $f_r(\mathbf{h}, \mathbf{t})$ is used to measure the plausibility of (h, r, t), which also implies the transformation $\mathbf{r}$ that characterizes r. Exemplarily, the translation-based model TransE [BUG13] uses the loss function $f_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$ [1], where $\mathbf{r}$ is characterized as a translation vector learnt from the latent connectivity patterns in the knowledge graph. This model provides a flexible way of predicting a missing item in a triple, or verifying the validity of a generated triple. Other works like TransH [WZF14b] and TransR [LLS15], introduce different loss functions to represent the relational translation in other forms, and have achieved promising results in knowledge graph completion and link prediction from text.

However, besides simple and multi-mapping relations, monolingual knowledge can contain other types of complex relations that enforces relational properties such as transitivity and symmetry, as well as hierarchical relations that may be simultaneously endowed transitivity and multi-mapping. Such relations with specific relational properties can exist in large portions in instance-level knowledge graphs, and can constitute major components of domain-specific ontology graphs [AV04, MBS15, SH13]. For example, Freebase contains more than 20% of transitive or symmetrical relations [BEP08]. This amount even rises to 70% in ConceptNet [SH13] and 92% in Yago3 Ontology [MBS15], which respectively contribute at least 26% and 95% hierarchical relations. Common existance of such complex relations makes characterization of monolingual knowledge non-trivial for the embedding models. Failure of preserving these important properties in the capturing the relations will no doubt cause imprecise representation of monolingual knowledge, and further impairs the performance of NLP tasks based-on the embedding techniques. Moreover, not many efforts have been put into the embedding representations of relations with relational properties and hierarchies.

On the other hand, the problem of applying embedding-based techniques on cross-lingual

[1] Hereafter, $\|\cdot\|$ means l_1 or l_2 norm unless explicitly specified.

knowledge remains largely unexplored. Such knowledge, including *inter-lingual links* (ILLs) that match the same entities, and *triple-wise alignment* (TWA) that represents the same relations, is very helpful in aligning and synchronizing different language-specific versions of a knowledge base that evolve independently, as needed to further improve applications built on multilingual knowledge bases, such as Q&A systems, semantic Web, and machine translation. In spite of its importance, this cross-lingual knowledge remains largely unsolved. In fact, in the most successful knowledge base Wikipedia, we find that ILLs cover less than 15% entity alignment.

Undoubtedly, leveraging knowledge graph embeddings to cross-lingual knowledge provides a generic way to help extract and apply such knowledge. However, it is a non-trivial task to find a tractable technique to capture the cross-lingual transitions[2]. Such transitions are more difficult to capture than relational translations for several reasons: (i) a cross-lingual transition has a far larger domain than any monolingual relational translation; (ii) it applies on both entities and relations, which have incoherent vocabularies among different languages; (iii) the known alignment for training such transitions usually accounts for a small percentage of a knowledge base. Moreover, the characterization of monolingual knowledge graph structures has to be well-preserved to ensure the correct representation of the knowledge to be aligned.

Besides, further investigating the potential of applying knowledge graph embeddings in NLP tasks is another urged mission. Though several such tasks have been proposed to be solved using word-embedding-based techniques [NG15, Kim14, TWY14, ZLC15], knowledge graph embeddings are no doubt more powerful in representing the explicit relations of the key language concepts, since their learning process is leveraged to capture latent semantics from structured knowledge. This has been proven in prior work of knowledge graph completion and link prediction [WBY13, WZF14a, WZF14b, LLS15]. After exploring new techniques to represent the multi-faceted mono and multilingual knowledge, we will apply our knowledge graph embeddings into more tasks that previously rely or do not rely on word embeddings, which include document sentiment analysis [Kim14], knowledge alignment [RLN13], unsupervised semantic relatedness

[2]We use the word *transition* here to differentiate from the relational translations among entities in translation-based methods.

analysis [SP14], and relation extraction [LSL16].

We intend to make following contributions in our research:

1. We propose a new model *MTransE* that learns the multilingual knowledge graph structure [CTY16]. MTransE uses a combination of two component models, namely the *knowledge model* and the *alignment model*. The knowledge model is responsible for encoding entities and relations in a language-specific version of knowledge graph. We explore the method that organizes each language-specific version in a separated embedding space. On top of that, the alignment model learns cross-lingual transitions for both entities and relations across different embedding spaces. Therefore we explore the following three representations techniques for cross-lingual alignment: distance-based axis calibration, translation vectors, and linear transformations. We test the effectiveness of different techniques used to capture cross-lingual transitions on large scale trilingual knowledge graphs, and explore how well it preserves monolingual relations like its monolingual counterpart.

2. In our research, we are also investigating another model *On2Vec* that focuses on preserving relational properties and hierarchies of relations in learning process of monolingual knowledge graph embeddings. This model learn role-dedicated mappings for entities placed in different positions of triples in order to preserve relational properties of relations, and strengthen the energy for relation hierarchies in the learning process. Variants of On2Vec can conceived based on different forms of the role-dedicated mapping model and the hierarchy model. We compare this model against state-of-the-art on key tasks of unsupervised relation extraction and link prediction using WordNet and Freebase-based data sets [BUG13], as well as other ontology data sets [MBS15, SH13] to verify its superiority on relations with special properties.

3. We will also apply our new models to support NLP tasks such as document sentiment analysis, cross-lingual knowledge alignment, unsupervised semantic relatedness analysis, and open information extraction, as well as multilingual NLP tasks like cross-lingual entity and

triple alignment. Therefore we will show how our methods of learning knowledge graph embeddings can be useful to help machine process complicated human languages.

The rest of this prospectus is organized as follows. In Chapter 2, we describe the background and related work. Chapter 3 introduces our new approach for learning graph embeddings for the multifaceted knowledge. In Chapter 4, we discuss the results we have already got on some natural-language-related applications, while propose the experimental design of more NLP tasks based on the knowledge graph embedding techniques. Lastly, we outline the research plan in Chapter 5.

Chapter 2

Background and Related Work

In this chapter, we state the background and prior work related to our research. We cover three topics in this chapter: different embedding methods, knowledge graphs, and natural language processing problems that can be solved using embedding-based techniques.

2.1 Embeddings

Embedding-based techniques project discrete concepts or words to a low-dimensional and continuous vector space where co-occurred concepts or words are located close to each other. Compared to conventional discrete representations (e.g., the one-hot encoding [MGS05]), embedding provides more robust representations, particulary for concepts that infrequently appear in corpora, but are with significance of meaning.

In this section, we state the background of embedding-based approaches that are frequently used in NLP tasks. We begin with a brief introduction to word embeddings, then focus on addressing the past advance of knowledge graph embeddings.

2.1.1 Word Embeddings

Word embeddings are the first embedding-based techniques to be well-recognized by the NLP community. A well-known efficient word embedding tool Word2Vec was proposed by [MCC13], where two log-linear models (CBOW and Skip-gram) are proposed to learn the neighboring relation of words in context. Another work that is worth mentioning is GloVe [PSM14], which

utilizes a weighted least squares model that trains on global word-word co-occurrence counts and thus makes efficient use of statistics to reduce embedding space with more meaningful structures. Word embedding models encode syntactic and semantic content implicitly, so that relations among concepts or words can be simply computed as the distances among their embeddings, and can realize so called "linearity" under certain cases. For example, $v('king') - v('man') \approx v('queen') - v('woman')$.[1].

Later work has proven word embeddings to be state-of-the-art feature models in many NLP tasks such as sentiment analysis [Kim14, TWY14], open information extraction [LSL16], and document distance estimation [KSK15].

Several approaches are also proposed to learn cross-lingual transitions among word embeddings separatedly trained on on multilingual parallel text corpora, such as LM [MLS13] and CCA [FD14] which respectively induce cross-lingual transformations among existing monolingual embeddings in forms of linear transformations and canonical component analysis, and OT [XWL15] that induces orthogonal transformations across normalized monolingual embedding spaces.

2.1.2 Knowledge Graph Embeddings

Lately, knowledge graph embedding methods are proposed to learn latent representation from structured corpora. Different from word embeddings, knowledge graph embeddings represent entities or concepts of the graphs as vectors, while the relations among them as different forms of vector calculation that is bound with specific relational semantics. There exists two grand families of knowledge graph embeddings, namely the translation-based and the non-translation-based.

Recently, significant advancement has been made in using the translation-based method to train monolingual knowledge graph embeddings. To characterize a triple (h, r, t), models of this family follow a common assumption $\mathbf{h}_r + \mathbf{r} \approx \mathbf{t}_r$, where $\mathbf{h}_r$ and $\mathbf{t}_r$ are either the original vectors of h and t, or the transformed vectors under a certain transformation w.r.t. relation r. The forerunner TransE [BUG13] sets $\mathbf{h}_r$ and $\mathbf{t}_r$ as the original $\mathbf{h}$ and $\mathbf{t}$, and achieves promising results in handling 1-to-1 relations of entities. Later works improve TransE on multi-mapping relations by

[1] $v(.)$ here means a embedding vector of a given word

introducing relation-specific transformations on entities to obtain different $\mathbf{h}_r$ and $\mathbf{t}_r$, including projections on relation-specific hyperplanes in TransH [WZF14b], linear transformations to heterogeneous relation spaces in TransR [LLS15], dynamic matrices in TransD [JHX15], and other forms [JWL16, NSQ16]. All these variants of TransE specialize entity embeddings for different relations, therefore improving knowledge graph completion on multi-mapping relations at the cost of increased model complexity. Meanwhile translation-based models cooperate well with other models. For example, variants of TransE are combined with word embeddings to help relation extraction from text [WBY13, ZZW15].

In addition to these , there are non-translation-based methods. UM [BWC11] and SE [BGW12] are respectively simplified versions of TransE and TransR. Bilinear model [JRB12] applies a bilinear transformation between $\mathbf{h}$ and $\mathbf{t}$, and HolE [NRP16] defines holographic mapping for relations. These models do not explicitly represent relation in forms of embeddings. SLM [CW08] and NTN [SCM13] adopt neural networks to learn structured data, while TADW [YLZ15] uses random walk on graphs to generate corpora for context-based training. These models are expressive and adaptable for both structured and text corpora, but they are too complex to be incorporated into an architecture supporting multilingual knowledge. Others including neural-based models SLM [CW08] and NTN [SCM13], and random-walk-based model TADW [YLZ15], are expressive and adaptable for both structured and text corpora. These are too complex to be incorporated with other models, including the architecture supporting multilingual knowledge that is to be proposed in our research.

To train a knowledge graph embedding model, a score function $f_r(\mathbf{h}, \mathbf{t})$ is defined to measure the plausibility of any given triples according to the assumption of the model addressed as above. In Table 2.1 we have listed the score functions of these models. Then the training process minimizes the total loss, which is defined as the sum of scores, via convex optimization algorithms such as stochastic gradient descent (SGD) [WM03], BFGS [LLS15] and AdaGrad [DHS11], except for neural-based methods which are trained using neural or tensor networks. To accelerate in training process, negative sampling [BUG13] are sometimes used in training, especially those translation-

Table 2.1: Score functions of some knowledge graph embedding models. k is the dimensionality of embedding space, $\mathbb{R}$ is the real number field.

Model	Score Function
TransE [BUG13]	$\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_{l_{1/2}}$ $\mathbf{r}\in\mathbb{R}^k$
TransH [WZF14b]	$\left\|(\mathbf{h}-\mathbf{W_r}^\top\mathbf{h}\mathbf{W_r})+\mathbf{r}-(\mathbf{t}-\mathbf{W_r}^\top\mathbf{t}\mathbf{W_r})\right\|_{l_2}$ $\mathbf{W_r},\mathbf{r}\in\mathbb{R}^k$
STransE [NSQ16]	$\|\mathbf{h}\mathbf{M_{r,1}}+\mathbf{r}-\mathbf{t}\mathbf{M_{r,2}}\|_{l_2}$ $\mathbf{r}\in\mathbb{R}^k;\mathbf{M_{r,1}},\mathbf{M_{r,2}}\in\mathbb{R}^{k\times k}$
TransR [LLS15]	$\|\mathbf{h}\mathbf{M_r}+\mathbf{r}-\mathbf{t}\mathbf{M_r}\|_{l_2}$ $\mathbf{r}\in\mathbb{R}^k,\mathbf{M_r}\in\mathbb{R}^{k\times k}$
TransD [JHX15]	$(\mathbf{M_r}(\mathbf{h}+\mathbf{r}-\mathbf{t}))^\top\mathbf{D_r}\mathbf{M_r}(\mathbf{h}+\mathbf{r}-\mathbf{t})$ $\mathbf{M_r}\in\mathbb{R}^{k\times k},\mathbf{D_r}=diag(w_1,w_2,...,w_k)\|_{w_i\in\mathbb{R}}$
UM [BWC11]	$\|\mathbf{h}-\mathbf{t}\|_{l_2}$
SE [BGW12]	$\|\mathbf{h}\mathbf{M_r}-\mathbf{t}\mathbf{M_r}\|_{l_2}$ $\mathbf{M_r}\in\mathbb{R}^{k\times k}$
Bilinear [JRB12]	$\left\|\mathbf{h}^\top\mathbf{M_r}\mathbf{t}\right\|_{l_2}$ $\mathbf{M_r}\in\mathbb{R}^{k\times k}$

based methods. This is realized by randomly corrupting the subject or object of a golden triple (h,r,t) to a corrupted triple (h',r,t'). Thereby the score function is rewritten as the hinge loss as below,

$$max(f_r(\mathbf{h},\mathbf{t})+\gamma-f_r(\mathbf{h}',\mathbf{t}'),0) \tag{2.1}$$

where γ is a positive margin. Usually, negative sampling follows either uniform distribution or Bernoulli distribution to corrupt either h or t, which are so called uniform and Bernoulli negative sampling respectively [WZF14b].

It is noteworthy that the literature has paid attention only to encode monolingual relations of entities within a single language-specific version of the knowledge bases. Indeed, at the best of our knowledge, there is no previous work on learning multilingual knowledge graph embeddings. On the other side, as we previously mentioned, how to preserve relational properties and hierarchies of relations in learning process of the embeddings represents another issue that remains unexplored.

2.2 Knowledge Bases and Knowledge Graphs

Knowledge bases contain descriptions of real-world facts that are modeled as knowledge graphs to store two aspects of knowledge, namely monolingual knowledge and multilingual knowledge.

We have already come across much of monolingual knowledge in the previous sections. In current knowledge bases, such as Wikipedia [Wik16], WordNet [BF13], and ConceptNet [SH13], vast amounts of multilingual knowledge are being created across the multiple language-specific versions of the knowledge base. Such multilingual knowledge, including inter-lingual links (ILLs), and triple-wise alignment (TWA), is very useful in aligning and synchronizing different language-specific versions of a knowledge base that evolve independently, as needed to further improve applications built on multilingual knowledge bases. However, such cross-lingual knowledge is far from complete, while extending it is challenging due to the fact that it is almost not possible for existing corpus to directly provide such knowledge of expertise. Existing approaches, which we will describe in Section 2.3.2 shortly, involve either extensive human involvement or require tuning comprehensive models on information that is external to knowledge graphs. Therefore, we urge to develop an automated and simple method that is able to extend cross-lingual knowledge, just like the embedding-based method we will describe in the next chapter.

From a different perspective, knowledge bases can also be classified as instance-level knowledge graphs that store the relations of specified entities, and ontology-level graphs [NXC16] that store the relations among macro concepts that manage a group of entities. Instance-level knowledge graphs are more scaled size, thus typically contain over millions of entities and thousands of types of relations within each language-specific version [BEP08, LIJ15, Vra12]. However, ontology graphs are sparse and domain-specific knowledge graphs that contain much fewer types of concept relations [MGK14, MBS15]. The majority of relations, especially those in ontology graphs, can be enforced with specific relational properties (e.g., transitivity, symmetry), or hierarchical relations (e.g. taxonomy relations and spatial topological relations [CGW16]). For example, Freebase contains more than 20% of transitive or symmetric relations [BEP08]; ConceptNet [SH13] contains 70% of transitive or symmetric relations, and at least 26% of hierarchical relations; Yago3 Ontology [MBS15] even contains only 17 types of relations (whose statistics we have listed in Table 2.2), while more than 92% of the relations are transitive or symmetric relations, and more than 95% of the relations are hierarchical. Moreover, we can further divide hierarchical relations into refine-

Table 2.2: Number of triples of each relation type in Yago3 Ontology.

Relation	Number	Trans.	Sym.	Hier.
happenedIn	2810			
hasChild	41938			✓
hasAcademicAdvisor	4			✓
livesIn	1600			
isCitizenOf	1197			
isLocatedIn	1549685	✓		✓
wasBornIn	11672			
isMarriedTo	8593		✓	
isLeaderOf	1071			✓
isPoliticianOf	4833			
hasNeighbor	450		✓	
hasCapital	5280			
isConnectedTo	26966	✓	✓	
dealsWith	821		✓	
influences	170			
hasCurrency	4			✓
diedIn	7195			
hasGender	34811			✓
Total num/portion	1699100	92.8%	2.2%	95.8%

ment and coercion relations [CBB06], such that the former divides each coarser concept or entity into more refined ones, and the later is the opposite. Due to that knowledge graphs are widely used in aiding NLP tasks, including those concept-oriented ontology graphs [CHH07, LLY04, NXC16]. Therefore that leads to an urgent mission for developing a better embedding model that preserves relational properties and hierarchies in the learning process.

2.3 Natural Language Processing Tasks

In this section we introduce some NLP tasks we plan to solve with our embedding models.

2.3.1 Relation Extraction

The objective of relation extraction, also known as open information extraction, is to extract or induce new triples from plain text or structured corpora. Since the coverage issue of monolingual knowledge has been widely addressed long ago, parsing-based techniques for completing mono-

lingual knowledge bases have been well studied in the past [CS04, ZSZ05, JZ07, YOM09, SGS11, MAG14]. Recent efforts have been made to conduct unsupervised relation extraction based on the relational inferences of knowledge graph embeddings, whose performance is contingent on well-characterization of monolingual relations. We have seen prominent improvement on this task as methods have been proposed to better handle multi-mapping relations. Therefore as we have mentioned previously, we expect further improvement when newly proposed embedding model can handle other special relations.

On the other hand, supervised approaches construct deep neural networks to extract relations from bags of sentences that mention involved entities [ZLC15, LSL16, NG15]. In order to obtain neural-network-readable signals that reflects the latent semantics of sentences, these supervised approaches use pre-trained word embeddings to transform sentences into tensors. In comparison to traditional parsing-based techniques, neural-based approaches are more flexible as they do not need to pre-define or generate a huge number of parse trees to handle different relations hidden under different sentences, nor do they need labors to verify conflict candidates parsed by different parse trees. Thus we plan to improve the sentence modeling by incorporating knowledge graph embeddings or knowledge-word jointly embeddings [WBY13]. Exploring towards such a direction allows to use knowledge graph embeddings either as the only feature model, or as axillary signals for highlighting concepts in sentences.

2.3.2 Knowledge Alignment

We have known that it is not expected for corpora to directly provide enough alignment information that can be used to exploit cross-lingual knowledge. Some projects though, produce cross-lingual alignment in knowledge bases at the cost of extensive human involvement and designing hand-crafted features dedicated to specific applications. Wikidata [Vra12] and DBpedia [LIJ15] rely on crowdsourcing to create ILLs and relation alignment. YAGO [MBS15] mines association rules on known matches, which combines many confident scores and requires extensively fine tuning. Many other works require sources that are external to the graphs, from well-established schemata or ontologies [NMN11, SAS11, RLN13] to entity descriptions [YST15], which being unavailable

to many knowledge bases such as YAGO, WordNet, and ConceptNet [SH13]. Such approaches also involve complicated model dependencies that are not tractable and reusable. By contrast, embedding-based methods are simple and general, require little human involvement, and generate task-independent features that can contribute to other NLP tasks.

2.3.3 Semantic Relatedness Analysis

Semantic relatedness, or semantic similarity between documents plays an important role in many textual applications such as information retrieval, document classification and clustering, question answering and more. Measurement of semantic relatedness comprises two constituents: an effective representation of documents, and a similarity measure between documents in terms of their respective representations. Difficulty of this task lies in two aspects. Firstly it is entirely unsupervised, which means the representation model has to reflect the distances among document pairs without knowing the ground truth. Secondly, the analysis results are usually measured using its correlation (Pearson or Spearman) against human evaluation of semantic relatedness, which set high requirements to both the document representation techniques and the similarity measure.

Prior work can be classified into two categories. One includes those that aggregate discrete representation of document vocabularies, namely bag-of-words [LNN05], LSA [LNN05], and ESA [GM07]. The other includes those that analyze the relatedness of highlighted concepts in documents based their relations in a knowledge base, such as GED [SP14], ConceptGraphSim [NXC16], and WikiWalk [YRM09]. On the same data set LP50 used by these work, we have already received a better correlation than bag-of-words, LSA, ESA, and GED using annotated Skip-gram (Section 4.3). The next step on this task is to test whether our knowledge graph embeddings can achieve an even better outcome.

Previous works have been done on analyzing cross-lingual semantic relatedness include based on fuzzy logic [HK10], discrete document representation [VSC02], ESA based on encyclopedia knowledge [HM09], and TF-IDF [TC13]. Because our research leverages knowledge graph embeddings to cross-lingual scenarios and enables cross-lingual transitions across language-specific embedding spaces, we have a chance to provide a new method for solving this problem as well, in

the same way in which we analyze the semantic relatedness of monolingual document.

2.3.4 Sentiment Analysis

Sentiment analysis summarizes categorized key opinions or topics in short sentences or in long documents. Therefore, it is sometimes also known as document classification [MY01]. This task has wide applications, including comment rating systems [CGT11], multi-document summarization [GMC00], and taxonomy generation for encyclopedia [CC05]. Most recent work has been using deep neural-networks like convolutional neural networks (CNN) or recurrent neural networks (RNN) to tackle this task [Kim14, TQL15, SPW13], which producing better results than earlier statistical methods like SVM [WM12, MC04]. Similar to supervised relation extraction, neural-based sentiment analysis also adopts word embeddings as the feature model for modeling sentences. We are going to give different insights in using knowledge graph embeddings or graph-text jointly embeddings the primary feature models for modeling documents.

Chapter 3

Learning Knowledge Graph Embeddings

In this chapter, we introduce new approaches for learning multi-facted knowledge graph embeddings.

3.1 Multilingual Knowledge Graph Embeddings

In this section, we propose a multilingual knowledge graph embedding model *MTransE*, that learns the multilingual knowledge graph structure using a combination of two component models, namely the *knowledge model* and the *alignment model*. The knowledge model encodes entities and relations in a language-specific version of knowledge graph. We explore the method that organizes each language-specific version in a separated embedding space, in which MTransE adopts TransE as the knowledge model. On top of that, the alignment model learns cross-lingual transitions for both entities and relations across different embedding spaces, where the following three representations of cross-lingual alignment are considered: distance-based axis calibration, translation vectors, and linear transformations. Thus, we obtain five variants of MTransE based on different loss functions, and identify the best variant by comparing them on cross-lingual alignment tasks using two partially aligned trilingual graphs constructed from Wikipedia triples. We also show that MTransE performs as well as its monolingual counterpart TransE on monolingual tasks.

3.1.1 Modeling

We hereby begin our modeling with the formalization of multilingual knowledge graphs.

Multilingual Knowledge Graphs

In a knowledge base KB, we use $\mathcal{L}$ to denote the set of languages, and $\mathcal{L}^2$ to denote the 2-combination of $\mathcal{L}$ (i.e., the set of *unordered* language pairs). For a language $L \in \mathcal{L}$, G_L denotes the language-specific knowledge graph of L, and E_L and R_L respectively denote the corresponding vocabularies of entity expression and relation expression. $T = (h, r, t)$ denotes a triple in G_L such that $h, t \in E_L$ and $r \in R_L$. Boldfaced $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$ respectively represent the embedding vectors of head h, relation r, and tail t. For a language pair $(L_1, L_2) \in \mathcal{L}^2$, $\delta(L_1, L_2)$ denotes the alignment set which contains the pairs of triples that have already been aligned between L_1 and L_2. For example, across the languages English and French, we may have $\big($(State of California, capital city, Sacramento), (État de Californie, capitale, Sacramento)$\big) \in \delta$(English, French). The alignment set commonly exists in a small portion in a multilingual knowledge base [Vra12, MBS15, LIJ15], and is one part of knowledge we want to extend.

Our model consists of two components that learn on the two facets of KB: the knowledge model that encodes the entities and relations from each language-specific graph structure, and the alignment model that learns the cross-lingual transitions from the existing alignment. We define a model for each language pair from $\mathcal{L}^2$ that has a non-empty alignment set. Thus, for a KB with more than two languages, a set of models composes the solution. In the following, we use a language pair $(L_i, L_j) \in \mathcal{L}^2$ as an example to describe how we define each component of a model.

Knowledge Model

For each language $L \in \mathcal{L}$, a dedicated k-dimensional embedding space $\mathbb{R}^k_L$ is assigned for vectors of E_L and R_L, where $\mathbb{R}$ is the field of real numbers. We adopt the basic translation-based method of TransE for each involved language, which benefits the cross-lingual tasks by representing entities uniformly under different relations. Therefore its loss function is given as below:

$$S_K = \sum_{L \in \{L_i, L_j\}} \sum_{(h,r,t) \in G_L} \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$$

S_K measures the plausibility of all given triples. By minimizing the loss function, the knowledge model preserves monolingual relations among entities, while also acts as a regularizer for the alignment model. Meanwhile, the knowledge model partitions the knowledge base into disjoint subsets that can be trained in parallel.

Alignment Model

The objective of the alignment model is to construct the transitions between the vector spaces of L_i and L_j. Its loss function is given as below:

$$S_A = \sum_{(T,T') \in \delta(L_i, L_j)} S_a(T, T')$$

Thereof, the alignment score $S_a(T, T')$ iterates through all pairs of aligned triples. Three different techniques to score the alignment are considered: distance-based axis calibration, translation vectors, and linear transformations. Each of them is based on a different assumption, and constitutes different forms of S_a alongside.

Distance-based Axis Calibration. This type of alignment models penalize the alignment based on the distances of cross-lingual counterparts. Either of the following two scorings can be adopted to the model.

$$S_{a_1} = \|\mathbf{h} - \mathbf{h}'\| + \|\mathbf{t} - \mathbf{t}'\|$$

S_{a_1} regulates that correctly aligned multilingual expressions of the same entity tend to have close embedding vectors. Thus by minimizing the loss function that involves S_{a_1} on known pairs of aligned triples, the alignment model adjusts axes of embedding spaces towards the goal of coinciding the vectors of the same entity in different languages.

$$S_{a_2} = \|\mathbf{h} - \mathbf{h}'\| + \|\mathbf{r} - \mathbf{r}'\| + \|\mathbf{t} - \mathbf{t}'\|$$

S_{a_2} overlays the penalty of relation alignment to S_{a_1} to explicitly converge coordinates of the same relation.

The alignment models based on axis calibration assume analogous spatial emergence of items in each language. Therefore, it realizes the cross-lingual transition by carrying forward the vector of a given entity or relation from the space of the original language to that of the other language.

Translation Vectors. This model encodes cross-lingual transitions into vectors. It consolidates alignment into graph structures and characterizes cross-lingual transitions as regular relational translations. Hence S_{a_3} as below is derived.

$$S_{a_3} = \left\|\mathbf{h} + \mathbf{v}_{ij}^{e} - \mathbf{h}'\right\| + \left\|\mathbf{r} + \mathbf{v}_{ij}^{r} - \mathbf{r}'\right\| + \left\|\mathbf{t} + \mathbf{v}_{ij}^{e} - \mathbf{t}'\right\|$$

Thereof $\mathbf{v}_{ij}^{e}$ and $\mathbf{v}_{ij}^{r}$ are respectively deployed as the entity-dedicated and relation-dedicated translation vectors between L_i and L_j, such that we have $\mathbf{e} + \mathbf{v}_{ij}^{e} \approx \mathbf{e}'$ for embedding vectors $\mathbf{e}$, $\mathbf{e}'$ of the same entity e expressed in both languages, and $\mathbf{r} + \mathbf{v}_{ij}^{r} \approx \mathbf{r}'$ for those of the same relation. We deploy two translation vectors instead of one, because there are far more distinct entities than relations, and using one vector easily leads to imbalanced signals from relations.

Such a model obtains a cross-lingual transition of an embedding vector by adding the corresponding translation vector. Moreover, it is easy to see that $\mathbf{v}_{ij}^{e} = -\mathbf{v}_{ji}^{e}$ and $\mathbf{v}_{ij}^{r} = -\mathbf{v}_{ji}^{r}$ hold. Therefore, as we obtain the translation vectors from L_i to L_j, we can always use the same vectors to translate in the opposite direction.

Linear Transformations. The last category of alignment models deduce linear transformations between embedding spaces. S_{a_4} as below learns a $k \times k$ square matrix $\mathbf{M}_{ij}^{e}$ as a linear transformation on entity vectors from L_i to L_j, given k as the dimensionality of the embedding spaces.

$$S_{a_4} = \left\|\mathbf{M}_{ij}^{e}\mathbf{h} - \mathbf{h}'\right\| + \left\|\mathbf{M}_{ij}^{e}\mathbf{t} - \mathbf{t}'\right\|$$

S_{a_5} additionally brings in a second linear transformation $\mathbf{M}_{ij}^{r}$ for relation vectors, which is of the same shape as $\mathbf{M}_{ij}^{e}$. The use of a different matrix is again due to different redundancy of entities and relations.

$$S_{a_5} = \left\|\mathbf{M}_{ij}^{e}\mathbf{h} - \mathbf{h}'\right\| + \left\|\mathbf{M}_{ij}^{r}\mathbf{r} - \mathbf{r}'\right\| + \left\|\mathbf{M}_{ij}^{e}\mathbf{t} - \mathbf{t}'\right\|$$

Unlike axis calibration, linear-transformation-based alignment model treats cross-lingual transitions as the topological transformation of embedding spaces without assuming the similarity of spatial emergence.

The cross-lingual transition of a vector is obtained by applying the corresponding linear transformation. It is noteworthy that, regularization of embedding vectors in the training process (which will be introduced soon after) ensures the invertibility of the linear transformations such that $\mathbf{M}_{ij}^{e\,-1} = \mathbf{M}_{ji}^{e}$ and $\mathbf{M}_{ij}^{r\,-1} = \mathbf{M}_{ji}^{r}$. Thus the transition in the revert direction is always enabled even though the model only learns the transformations of one direction.

Variants of MTransE

Combining the above two component models, MTransE minimizes the following loss function $J = S_K + \alpha S_A$, where α is a hyperparameter that weights S_K and S_A.

As we have given out five variants of the alignment model, each of which correspondingly defines its specific way of computing cross-lingual transitions of embedding vectors. We denote Var_k as the variant of MTransE that adopts the k-th alignment model which employs S_{a_k}. In practice, the searching of a cross-lingual counterpart for a source is always done by querying the nearest neighbor from the result point of the cross-lingual transition. We denote function τ_{ij} that maps a cross-lingual transition of a vector from L_i to L_j, or simply τ in a bilingual context. As stated, the solution in a multi-lingual scenario consists of a set of models of the same variant defined on every language pair in $\mathcal{L}^2$. Table 3.1 summarizes the model complexity, the definition of cross-lingual transitions, and the complexity of searching a cross-lingual counterpart for each variant.

3.1.2 Training

We optimize the loss function using on-line stochastic gradient descent [WM03]. At each step, we update the parameter θ by setting $\theta \leftarrow \theta - \lambda \nabla_\theta J$, where λ is the learning rate. Instead of directly updating J, our implementation optimizes S_K and αS_A alternately. In detail, at each epoch we optimize $\theta \leftarrow \theta - \lambda \nabla_\theta S_K$ and $\theta \leftarrow \theta - \lambda \nabla_\theta \alpha S_A$ in separated groups of steps.

Table 3.1: Summary of model variants.

Var	Model Complexity	Cross-lingual Transition	Search Complexity
Var_1	$\mathrm{O}(n_e kl + n_r kl)$	$\tau_{ij}(\mathbf{e}) = \mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{r}$	$\mathrm{O}(n_e k)$ $\mathrm{O}(n_r k)$
Var_2	$\mathrm{O}(n_e kl + n_r kl)$	$\tau_{ij}(\mathbf{e}) = \mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{r}$	$\mathrm{O}(n_e k)$ $\mathrm{O}(n_r k)$
Var_3	$\mathrm{O}(n_e kl + n_r kl + kl^2)$	$\tau_{ij}(\mathbf{e}) = \mathbf{e} + \mathbf{v}^e_{ij}$ $\tau_{ij}(\mathbf{r}) = \mathbf{r} + \mathbf{v}^r_{ij}$	$\mathrm{O}(n_e k)$ $\mathrm{O}(n_r k)$
Var_4	$\mathrm{O}(n_e kl + n_r kl + 0.5k^2l^2)$	$\tau_{ij}(\mathbf{e}) = \mathbf{M}^e_{ij}\mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{M}^e_{ij}\mathbf{r}$	$\mathrm{O}(n_e k^2 + n_e k)$ $\mathrm{O}(n_r k^2 + n_r k)$
Var_5	$\mathrm{O}(n_e kl + n_r kl + k^2l^2)$	$\tau_{ij}(\mathbf{e}) = \mathbf{M}^e_{ij}\mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{M}^r_{ij}\mathbf{r}$	$\mathrm{O}(n_e k^2 + n_e k)$ $\mathrm{O}(n_r k^2 + n_r k)$

Notation: $\mathbf{e}$ and $\mathbf{r}$ are respectively the vectors of an entity e and a relation r, k is the dimension of the embedding spaces, l is the cardinality of $\mathcal{L}$, n_e and n_r are respectively the number of entities and the number of relations, where $n_e \gg n_r$.

We enforce the constraint that the l_2 norm of any entity embedding vector is 1, thus regularize embedding vectors to a unit spherical surface. This constraint is widely employed in the literature [BUG13, BWU14, JRB12] and has two important effects: (i) it helps avoid the case where the training process trivially minimizes the loss function by shrinking the norm of embedding vectors, and (ii) it implies the invertibility of the linear transformations [XWL15] for Var_4 and Var_5.

We initialize vectors by drawing from a uniform distribution on the unit spherical surface, and initialize matrices using random orthogonal initialization [SMG14]. Negative sampling is not employed in training, in fact that does not noticeably affect the results.

3.2 Property Preserving Knowledge Graph Embeddings

In this section, we briefly introduce another proposed model to learn embeddings that focuses on preserving relational properties, and hierarchies of monolingual relations. In comparison to other translation-based methods, such a model tackling above relations is especially more suitable for encoding ontology graphs. Thus we name this model On2Vec (i.e. ontology to vector).

3.2.1 Modeling

We hereby extend the formalization in Section 3.1.1 to reflect these special relations we are to handle. Now, we only consider any monolingual knowledge graph w.l.o.g. Thus, we omit any language mark L. However, now we extend the vocabulary of relations into $R = R_{tr} \cup R_s \cup R_h \cup R_o$

where R_{tr} is the set of transitive relations, R_s is the set of symmetric relations, R_h is the set of hierarchical relations, and R_o is the set of other simple relations. Thereof, R_{tr} and R_h are not required to be disjoint, while both of them are disjoint with R_o. For transitive relations, that is to say, for each $r \in R_{tr}$, there exists three different entities $e_1, e_2, e_3 \in E$ such that $(e_1, r, e_2), (e_2, r, e_3), (e_1, r, e_3) \in G$. As for symmetric relations, that is to say, for each $r \in R_s$, there exists two different entities $e_1, e_2 \in E$ such that $(e_1, r, e_2), (e_2, r, e_1) \in G$. As for hierarchical relations, we further divide them into $R_h = R_r \cup R_c$ where R_r is the set of refinement relations, R_c is the set of coercion relations. The difference between relations in R_h and multi-mapping relations is that the former consider only the atomic relations (i.e. relations satisfying transitive reduction [CGW16]).

To facilitate the definition of energy functions, we also define a relational operator, namely *refine*:

- Given $r \in R_r$, $e \in E$, define operation $\sigma(e, r) = \{e' | (e, r, e') \in G\}$ as the operation that fetches all the other entities e' that directly applies the refinement relation r to e (which are adjacent nodes to e satisfying transitive reduction).

- Given $r \in R_c$, $e \in E$, define operation $\sigma(e, r) = \{e' | (e', r, e) \in G\}$ as the operation that fetches all the other entities e' that directly applies the coercion relation r to e (which are adjacent nodes to e satisfying transitive reduction).

Like MTransE, On2Vec tackles relations with relational properties and hierarchies using different model components: the *role-dedicated-mapping model*, and the *hierarchy model*.

Role-dedicated-mapping Model

The reason for which the previous translation-based mappings to fail to preserve relational properties of relations is that those relation-specific entity transformations place entities involved in transitive or symmetric relations into conflict positions.

- Consider $r \in R_{tr}$ and $e_1, e_2, e_3 \in E$ such that $(e_1, r, e_2), (e_2, r, e_3), (e_1, r, e_3) \in G$, where e_1, e_2, and e_3 are mapped to $\mathbf{e_{1r}}$, $\mathbf{e_{2r}}$, and $\mathbf{e_{3r}}$ respectively by the relation-specific entity

transformation w.r.t. r. If $\mathbf{e_{1r}} + \mathbf{r} \approx \mathbf{e_{2r}}$ and $\mathbf{e_{2r}} + \mathbf{r} \approx \mathbf{e_{3r}}$ are learnt for the first and second triples, then there is no way for $\mathbf{e_{1r}} + \mathbf{r} \approx \mathbf{e_{3r}}$ to hold for the third triple, since r is of fixed norm.

- On the other hand, consider $r \in R_s$ and $e_1, e_2 \in E$ such that $(e_1, r, e_2), (e_2, r, e_1) \in G$, where e_1 and e_2 are mapped to $\mathbf{e_{1r}}$ and $\mathbf{e_{2r}}$ respectively by the relation-specific entity transformation w.r.t. r. There is no way for both $\mathbf{e_{1r}} + \mathbf{r} \approx \mathbf{e_{2r}}$ and $\mathbf{e_{2r}} + \mathbf{r} \approx \mathbf{e_{1r}}$ to hold unless we reach a trivial optimization solution such that $\mathbf{r}$ is a zero vector.

Hence, to preserve relational properties, or to solve the conflicts in the above two conditions, role-dedicated-mapping model maps differently the same entities placed at head or tail positions using two role-dedicated mappings respectively. The general form of the role-dedicated-mapping model to score a given triple is as below:

$$S_r = \|f_{1,r}(\mathbf{h}) + \mathbf{r} - f_{2,r}(\mathbf{t})\|$$

where $f_{1,r}$ and $f_{2,r}$ are respectively the head- and tail-dedicated relation-specific transformations. The forms of $f_{1,r}$ and $f_{2,r}$ are decided particularly by what techniques we want to use for encoding entity vectors w.r.t. different relations. For example, we can simply use linear transformations that maps entities to relation-dedicated spaces as follows:

$$S_{r,1} = \|\mathbf{M_{1,r}}\mathbf{h} + \mathbf{r} - \mathbf{M_{2,r}}\mathbf{t}\| \quad s.t. \quad \mathbf{M_{1,r}}, \mathbf{M_{2,r}} \in \mathbb{R}^{k \times k}$$

Or, if relation-dedicated hyperplane is considered instead, then the following scoring can be used:

$$S_{r,2} = \left\|(\mathbf{h} - \mathbf{W_{1,r}^\top}\mathbf{h}\mathbf{W_{1,r}}) + \mathbf{r} - (\mathbf{t} - \mathbf{W_{2,r}^\top}\mathbf{t}\mathbf{W_{2,r}})\right\| \quad s.t. \quad \mathbf{W_{1,r}}, \mathbf{W_{2,r}} \in \mathbb{R}^{k}$$

Other forms of transformations like bilinear transformations and affine transformations may also be used to rewrite S_r, but we plan to limit our exploration on $S_{r,1}$ and $S_{r,2}$ due to that other forms unnecessarily introduce much higher model complexity.

To help more efficient learning, we should also consider negative sampling. Then the complete energy function of the role-dedicated-mapping model is written as the hinge loss as below:

$$S_{rs} = \sum_{(h,r,t)\in G \wedge (h',r,t')\notin G} max(\|f_{1,r}(\mathbf{h}) + \mathbf{r} - f_{2,r}(\mathbf{t})\| - \|f_{1,r}(\mathbf{h}') + \mathbf{r} - f_{2,r}(\mathbf{t}')\| + \gamma_1, 0)$$

Here, (h', r, t') is the corrupted triple from either uniform or bernoulli negative sampling. γ_1 is a positive margin.

Hierarchy Model

The objective of the hierarchy model is to add auxiliary energy to On2Vec to capture relations in hierarchies. Because the hierarchical relations expands to multiple targets in each level. It is very likely that for $e \in E$, the embedding vectors of each $e' \in (e, r)$ tend to separate in the embedding space, therefore impair the relational inferences. To prevent such an issue, this additional energy is dedicated to converge the projected embeddings of every $e' \in (e, r)$.

The energy function of the hierarchy model is defined as:

$$S_h = \sum_{e\in E \wedge r\in R_r \wedge e_1 \in \sigma(e,r)} \omega\left(f_{1,r}(\mathbf{e}) + \mathbf{r}, f_{2,r}(\mathbf{e}_1)\right) + \sum_{e\in E \wedge r\in R_c \wedge e_2 \in \sigma(e,r)} \omega\left(f_{1,r}(\mathbf{e}) - \mathbf{r}, f_{2,r}(\mathbf{e}_2)\right)$$

where ω is a function that monotonically increases w.r.t. the angle of the two input vectors. In practice, we can use dot product, cosine, or $l_{1/2}$-norm as ω. As in the role-dedicated mapping model, we should also consider negative sampling in hierarchy model, therefore rewrite S_h as:

$$\begin{aligned} S_h = &\sum_{e\in E \wedge r\in R_r \wedge e_1 \in \sigma(e,r) \wedge e_1' \notin \sigma(e,r)} max(\omega\left(f_{1,r}(\mathbf{e}) + \mathbf{r}, f_{2,r}(\mathbf{e}_1)\right) + \gamma_2 - \omega\left(f_{1,r}(\mathbf{e}) + \mathbf{r}, f_{2,r}(\mathbf{e}_1')\right), 0) \\ + &\sum_{e\in E \wedge r\in R_c \wedge e_2 \in \sigma(e,r) \wedge e_2' \notin \sigma(e,r)} max(\omega\left(f_{1,r}(\mathbf{e}) - \mathbf{r}, f_{2,r}(\mathbf{e}_2)\right) + \gamma_2 - \omega\left(f_{1,r}(\mathbf{e}) - \mathbf{r}, f_{2,r}(\mathbf{e}_2')\right), 0) \end{aligned}$$

where e_1' and e_2' are negative samples from either uniform or Bernoulli negative sampling, and γ_2 is a positive margin.

3.2.2 Training

The objective of training On2Vec is to minimize the combined score of S_r and S_h. Meanwhile, constraints should be enforced to the norm of all original and projected vectors, so as to prevent embedding vectors from collapsing to a trivial solution in training process, which are listed as below.

$$\|\mathbf{e}\| \leq 1 \wedge \|f_{1,r}(\mathbf{e})\| \leq 1 \wedge \|f_{2,r}(\mathbf{e})\| \leq 1 \wedge \|\mathbf{r}\| \leq 2, \forall e \in E, \forall r \in R$$

The constraints is incorporated as soft-constraints as below.

$$S_c = \sum_{e \in E} max(\|\mathbf{e}\| - 1, 0) + max(\|f_{1,r}(\mathbf{e})\| - 1, 0) + max(\|f_{2,r}(\mathbf{e})\| - 1, 0) + \sum_{r \in R} max(\|\mathbf{r}\| - 2, 0)$$

Thus, to train On2Vec is implemented by optimizing (minimizing) the following joint energy function using batched stochastic gradient descent (SGD),

$$J = S_r + \alpha_1 S_h + \alpha_2 S_c$$

where α_1 and α_2 are both non-negative hyperparameters. So far, we have implemented one variant of On2Vec using TensorFlow [ABC16], where $f_{1,r}$ and $f_{2,r}$ are implemented in the form of linear transformations, and ω is implemented as $l_{1/2}$-norm. Further fine-tuning, implementation of other variants, and further experiments will be completed shortly.

3.3 Joint Embeddings

A word and knowledge graph joint embedding model (or text-graph joint embeddings) is easily created by applying an alignment model between a word embedding model and a knowledge graph embedding model. The advantage of a joint model is that it accepts both signals from knowledge graph structures and plain text contexts, and has a large vocabulary like a word embedding model. Word embeddings are trained on the plain text document D that has a vocabulary V. A sliding window c reads the context of each word in the document D. The alignment model is trained on

the anchor file A that contains entity-word pairs $(w, e) \in A$ such that $w \in V$ and $e \in E$. Suppose the word embedding model is the Skip-gram model in Word2Vec [MCC13] which maximizes the log-linear energy function as below,

$$S_{log} = \sum_{(w,c)\in D} \log p(c|w) = \sum_{(w,c)\in D} (\log e^{v_c \cdot v_w} - \sum_{c' \in D} \log e^{v_{c'} \cdot v_w})$$

where v_w is the embedding vectors of word w, v_c and $v_{c'}$ are embeddings of contexts. The alignment model, which is similar to the one we use for Var_1 of MTransE, calibrates the embedding spaces of words and the knowledge graph by minimizing

$$S_{al} = \sum_{(w,e)\in A} \|v_w - \mathbf{e}\|$$

which leads to the energy function of the joint embedding model

$$J_{joint} = 1 - S_{log} + \alpha_3 J + \alpha_4 S_{al}$$

that is to be minimized in the optimization process. Thereof, J is the energy score of a (monolingual) knowledge graph embedding model, α_3 and α_4 are positive hyperparameters. In some tasks we may consider using the joint embedding model as the feature model. In particular such is the case for supervised relation extraction and sentiment analysis due to that, in these applications we often have a larger vocabulary than knowledge graph embeddings.

Chapter 4

Embedding-based Natural Language Processing

In this chapter, we introduce the four categories of NLP tasks we propose to solve using embedding-based techniques, namely knowledge alignment, relation extraction, semantic relatedness analysis, and sentiment analysis. Thereof, we have already obtained some results in knowledge alignment.

4.1 Knowledge Alignment

The objective of cross-lingual knowledge alignment is to align cross-lingual counterparts in different language-specific versions of knowledge bases. Two levels of knowledge alignment are considered, namely cross-lingual entity matching, and triple-wise alignment verification. In this section we discuss the details of MTransE in handling these two cross-lingual tasks. We also conducted experiments on two monolingual tasks. Furthermore, a *case study* with examples of knowledge alignment is included in the Appendix.

Data Sets. Experimental results on the trilingual data sets WK3l are reported in this section. WK3l contains English (En), French (Fr), and German (De) knowledge graphs under DBpedia's `dbo:Person` domain, where a part of triples are aligned by verifying the ILLs on entities, and multilingual labels of the DBpedia ontology on some relations. The number of entities in each language is adjusted to obtain two data sets. Thereof, for each of the three languages, WK3l-15k matches the number of nodes (about 15,000) with FB15k—the largest monolingual graph used by many recent works [ZZW15, LLS15, JHX15, JWL16], and the number of nodes in WK3l-120k is several times larger. For both data sets, German graphs are sparser than English and French

Table 4.1: Statistics of the WK3l data sets.

Data set	#En triples	#Fr triples	#De triples	#Aligned triples
WK3l-15k	203,502	170,605	145,616	En-Fr:16,470 En-De:37,170
WK3l-120k	1,376,011	767,750	391,108	En-Fr:124,433 En-De:69,413

Table 4.2: Number of extra entity inter-lingual links (ILLs).

Data Set	En-Fr	Fr-En	En-De	De-En
WK3l-15k	3,733	3,815	1,840	1,610
WK3l-120k	42,413	41,513	7,567	5,921

graphs. We also collect extra ILLs that are not covered by the alignment sets for the evaluation of cross-lingual entity matching, whose quantity is shown in Table 4.2. Meanwhile, we also derive another trilingual data set CN3l from the ConceptNet [SH13]. We report the additional results on CN3l in the Appendix, which lead to similar evaluation conclusions.

4.1.1 Cross-lingual Entity Matching

The objective of this task is to match the same entities from different languages in KB. Due to the large candidate space, this task emphasizes more on ranking a set of candidates rather than acquiring the best answer. We perform this task on both data sets to compare five variants of MTransE.

To show the superiority of MTransE, we adapt LM, CCA, and OT (which we have introduced in the related work) to their knowledge graph equivalences.

Evaluation Protocol. Each MTransE variant is trained on a complete data set. LM and CCA are implemented by inducing the corresponding transformations across separately trained knowledge models on monolingual graphs, while using the alignment sets as anchors. Training OT is quite similar to MTransE, we add the process of orthogonalization to the training of the alignment model since the normalization of vectors is already enforced. The extra ILLs are used as ground truth for test. We take these unidirectional links between English-French and English-German, i.e., four directions in total. For each ILL (e, e'), we perform a kNN search from the cross-lingual transition

Table 4.3: Cross-lingual entity matching results.

Data Set	WK3l-15k								WK3l-120k			
Aligned Languages	En-Fr		Fr-En		En-De		De-En		En-Fr	Fr-En	En-De	De-En
Metric	Hits@10	Mean	Hits@10	Mean	Hits@10	Mean	Hits@10	Mean	Hits@10	Hits@10	Hits@10	Hits@10
LM	12.31	3621.17	10.42	3660.98	22.17	5891.13	15.21	6114.08	11.74	14.26	24.52	13.58
CCA	20.78	3094.25	19.44	3017.90	26.46	5550.89	22.30	5855.61	19.47	12.85	25.54	20.39
OT	44.97	508.39	40.92	461.18	44.47	155.47	49.24	145.47	38.91	37.19	38.85	34.21
Var_1	51.05	470.29	46.64	436.47	48.67	146.13	50.60	167.02	38.58	36.52	42.06	47.79
Var_2	45.25	570.72	41.74	565.38	46.27	168.33	49.00	211.94	31.88	30.84	41.22	40.39
Var_3	38.64	587.46	36.44	464.64	50.82	125.15	52.16	151.84	38.26	36.45	50.48	52.24
Var_4	59.24	**190.26**	**57.48**	**199.64**	**66.25**	**74.62**	**68.53**	**86.33**	**48.66**	47.43	57.56	63.49
Var_5	**59.52**	191.36	57.07	204.45	60.25	99.48	66.03	95.79	45.65	**47.48**	**64.22**	**67.85**

point of e (i.e., $\tau(\mathbf{e})$) and record the rank of $\mathbf{e}'$. Following the convention, we aggregate two metrics over all test cases, i.e., the proportion of ranks no larger than 10 $Hits@10$ (in percentage), and the mean rank $Mean$. Thereof, we prefer higher $Hits@10$ and lower $Mean$ that indicate a better outcome.

For training, we select the learning rate λ among $\{0.001, 0.01, 0.1\}$, α among $\{1, 2.5, 5, 7.5\}$, l_1 or l_2 norm in loss functions, and dimensionality k among $\{50, 75, 100, 125\}$. The best configuration on WK3l-15k is $\lambda = 0.01$, $\alpha = 5$, $k = 75$, l_1 norm for Var_1, Var_2, LM, and CCA, l_2 norm for other variants and OT. While the best configuration on WK3l-120k is $\lambda = 0.01$, $\alpha = 5$, $k = 100$, and l_2 norm for all models. The training on both data sets takes 400 epochs.

Results. We report $Hits@10$ and $Mean$ for WK3l-15k, and $Hits@10$ for WK3l-120k, on the four involved directions of cross-lingual matching in Table 4.3. As expected, without joint adapting the monolingual vector spaces with alignment, LM and CCA are largely outperformed by the rest. While the orthogonality constraint being too strong to be enforced in these cases, OT performs at most closely to the simplest cases of MTransE. For MTransE, Var_4 and Var_5 outperform the other three variants under all settings. The fairly close results obtained by these two variants indicate that the interference caused by learning an additional relation-dedicated transformation in Var_5 is negligible to the entity-dedicated transformation. Correspondingly, we believe that the reason for Var_3 to be outperformed by Var_4 and Var_5 is that it fails to differentiate well the over-frequent cross-lingual alignment from regular relations. Therefore, the characterization for cross-lingual alignment is negatively affected by the learning process for monolingual relations in a visible degree. Axis calibration appears to be unstable on this task. We hypothesize that this simple technique is affected by two factors: coherence between language-specific versions, and density

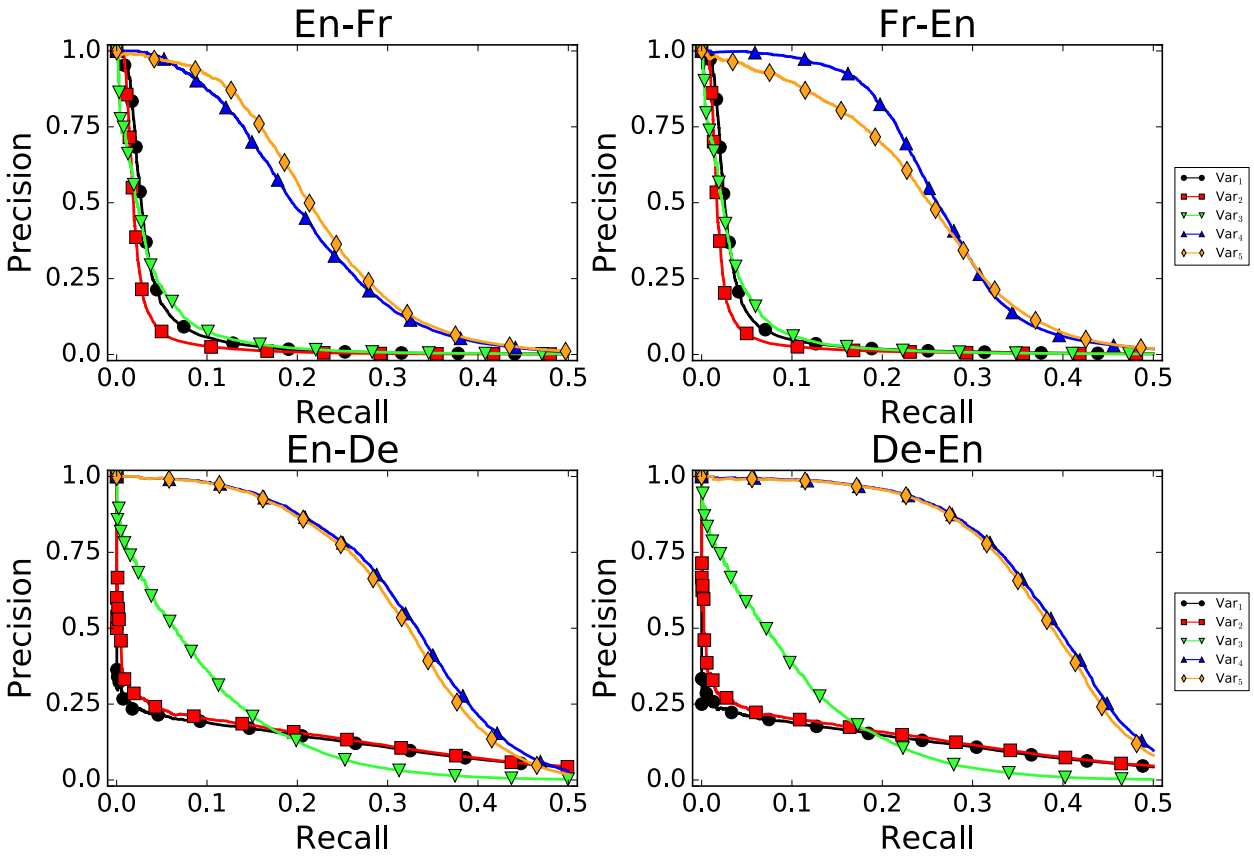

Figure 4.1: Precision-recall curves for cross-lingual entity matching on WK3l-15k.

of the graphs. Var_2 is always outperformed by Var_1 due to the negative effect of the calibration based on relations. We believe this is because multi-mapping relations are not so well-captured by TransE as explained in [WZF14b], and this disturbs the calibration of the entire embedding spaces. Although Var_1 still precedes Var_3 on entity matching between English and French graphs in WK3l-15k, coherence somewhat drops alongside when scaling up to the larger data set so as to hinder the calibration. The German graphs are sparse, thus should have set a barrier for precisely constructing embedding vectors and hindered calibration on the other side. Therefore Var_1 still performs closely to Var_3 in the English-German task on WK3l-15k and English-French task on WK3l-120k, but is preceded by Var_3 in the last setting. In general, the variants that use linear transformations are the most desired. This conclusion is supported by their promising outcome on this task, and it is also reflected in the precision-recall curves shown in Figure 4.1.

Table 4.4: Accuracy of TWA verification (%).

Data Set	WK3l-15k		WK3l-120k	
Languages	En&Fr	En&De	En&Fr	En&De
LM	52.23	63.61	59.98	59.98
CCA	52.28	66.49	65.89	61.01
OT	93.20	87.97	88.65	85.24
Var_1	93.25	91.24	91.27	91.35
Var_2	90.24	86.59	89.36	86.29
Var_3	90.38	84.24	87.99	87.04
Var_4	94.58	**95.03**	**93.48**	93.06
Var_5	**94.90**	94.95	92.63	**93.66**

Table 4.5: Results of tail prediction (*Hits*@10).

Data Set	WK3l-15k		WK3l-120k	
Language	En	Fr	En	Fr
TransE	**42.19**	25.06	36.78	25.38
Var_1	40.37	23.45	39.09	25.52
Var_2	40.80	24.77	36.02	21.13
Var_3	40.97	22.26	35.99	19.69
Var_4	41.03	25.46	**39.64**	**25.59**
Var_5	41.79	**25.77**	38.35	24.68

Table 4.6: Results of relation prediction (*Hits*@10).

Data Set	WK3l-15k		WK3l-120k	
Language	En	Fr	En	Fr
TransE	61.79	62.55	60.06	65.29
Var_1	60.18	60.73	**61.75**	65.46
Var_2	54.33	62.98	61.11	61.47
Var_3	58.32	59.44	60.14	48.06
Var_4	63.74	**64.77**	60.26	**67.64**
Var_5	**64.79**	63.71	60.77	66.86

4.1.2 Triple-wise Alignment Verification

This task is to verify whether a given pair of aligned triples are truly cross-lingual counterparts. It produces a classifier that helps with verifying candidates of triple matching [NMN11, RLN13].

Evaluation Protocol. We create positive cases by isolating 20% of the alignment set. Similar to [SCM13], we randomly corrupt positive cases to generate negative cases. In detail, given a pair of

correctly aligned triples (T, T'), it is corrupted by (i) randomly replacing one of the six elements in the two triples with another element from the same language, or (ii) randomly substituting either T or T' with another triple from the same language. Cases (i) and (ii) respectively contribute negative cases that are as many as 100% and 50% of positive cases. We use 10-fold cross-validation on these cases to train and evaluate the classifier.

We use a simple threshold-based classifier similar to the widely-used ones for triple classification [SCM13, WZF14b, LLS15]. For a given pair of aligned triples $(T, T') = \big((h, r, t), (h', r', t')\big)$, the dissimilarity function is defined as $f_d(T, T') = \|\tau(\mathbf{h}) - \mathbf{h}'\|_2 + \|\tau(\mathbf{r}) - \mathbf{r}'\|_2 + \|\tau(\mathbf{t}) - \mathbf{t}'\|_2$. The classifier finds a threshold σ such that $f_d < \sigma$ implies positive, otherwise negative. The value of σ is determined by maximizing the accuracy for each fold on the training set. Such a simple classification rule adequately relies on how precisely each model represents cross-lingual transitions for both entities and relations.

We carry forward the corresponding configuration from the last experiment, just to show the performance of each variant under controlled variables.

Results. Table 4.4 shows the mean accuracy, with a standard deviation below 0.009 in cross-validation for all settings. Thus, the results are statistically sufficient to reflect the performance of classifiers. Note that the results appear to be better than those of the previous task since this is a binary classification problem. Intuitively, the linear-transformation-based MTransE perform steadily and take the lead on all settings. We also observe that Var_5, though learns an additional relation-dedicated transformation, still performs considerably close to Var_4 (the difference is at most 0.85%). The simple Var_1 is the runner-up, and is between 1.65% and 3.79% to the optimal solutions. However the relation-dedicated calibration in Var_2 causes a notable setback (4.12%$\sim$8.44% from the optimal). The performance of Var_3 falls behind slightly more than Var_2 (4.52%$\sim$10.79% from the optimal) due to the failure in distinguishing cross-lingual alignment from regular relations. Meanwhile, we single out the accuracy on the portion of negative cases where only the relation is corrupted for English-French in WK3l-15k. The five variants receive 97.73%, 93.78%, 82.34%, 98.57%, and 98.54%, respectively. The close accuracy of Var_4 and

Var_5 indicates that the only transformation learnt from entities in Var_4 is enough to substitute the relation-dedicated transformation in Var_5 for discriminating relation alignment, while learning the additional transformation in Var_5 does not notably interfere the original one. However, it applies differently to axis calibration since Var_2 does not improve but actually impairs the cross-lingual transitions for relations. For the same reasons as above, LM and CCA do not match with MTransE in this experiment as well, while OT perform closely to some variants of MTransE, but is still left behind by Var_4 and Var_5.

4.1.3 Monolingual Tasks

The above experiments have shown the strong capability of MTransE in handling cross-lingual tasks. We next report the results on comparing MTransE with its monolingual counterpart TransE on two monolingual tasks introduced in the literature [BUG13, BGW14], namely tail prediction (predicting t given h and r) and relation prediction (predicting r given h and t), using the English and French versions of our data sets. Like previous works [BUG13, WZF14b, JWL16], for each language version, 10% triples are selected as the test set, and the remaining becomes the training set. Each MTransE variant is trained upon both language versions of the training set for the knowledge model, while the intersection between the alignment set and the training set is used for the alignment models. TransE is trained on either language version of the training set. Again, we use the configuration from the previous experiment.

Results. The results for $Hits@10$ are reported in Tables 4.5 and 4.6. They imply that MTransE preserves well the characterization of monolingual knowledge. For each setting, Var_1, Var_4, and Var_5 perform at least as well as TransE, and some even outperforms TransE under certain settings. This signifies that the alignment model does not interfere much with the knowledge model in characterizing monolingual relations, but might have actually strengthened it since coherent portions of knowledge are unified by the alignment model. Since such coherence is currently not measured, this question is left as a future work. The other question that deserves further attention is, how other knowledge models involving relation-specific entity transformations [WZF14b, LLS15, JHX15, JWL16, NSQ16] may influence monolingual and cross-lingual tasks.

Table 4.7: Results of relation extraction.

Data Set	WN18		FB15k	
Metric	*Hits*@10	*Mean*	*Hits*@10	*Mean*
UM [BGW12]	35.3	315	6.3	1074
RESCAL [NTK11]	37.2	1163	44.1	683
SE [BWC11]	80.5	985	39.8	162
SME [BGW12]	74.1	533	40.8	154
Bilinear [JRB12]	81.6	456	33.1	164
TransE [BUG13]	89.2	251	47.1	125
TransH [WZF14b]	82.3	388	64.4	87
TransR [LLS15]	92.0	225	68.7	**77**
TransD [JHX15]	92.5	212	77.3	91
On2Vec	**94.46**	**64**	**80.9**	80.8

4.2 Relation Extraction

The objective of relation extraction is to extend the monolingual knowledge graphs with new relations that are mined from either structured knowledge or plain text. When applying embedding-based techniques, typically the former is unsupervised, while the latter is unsupervised.

4.2.1 Unsupervised Relation Extraction

Because knowledge graph embeddings enable direct relational inferences across entities, unsupervised relation extraction is easily realized via vector operations and kNN search, just like we have seen in Section 4.1.3. Such a method has a limitation in the coverage of knowledge it can extend, since it does not mine new entities or relations from resources that are external to current knowledge graphs. Therefore, unsupervised relation extraction is usually precise only under small recall-rate. However, unsupervised relation extraction allows predicting missing heads, missing tails as well as missing relations for triples, therefore is very efficient and is widely used in many recent works [BUG13, WZF14b, LLS15, JWL16, NSQ16, BGW12].

To quickly compare with related works on this task, we directly apply the first implemented variant of On2Vec on the relation extraction tasks on WN18 and FB15k data sets [LLS15], where the former contains 22% transitive or symmetric relations and 16% hierarchical relations, and the latter contains 17% relations with relational properties and 32% with hierarchies. This task uses

held-out evaluation, and the partitioning of the data sets are already given in [WZF14b]. WN18 contains 40,943 entities and 18 types of relations, 141,442 triples are used as the training set and 5,000 triples are used as the test set. FB15k contains 14,951 entities and 1,345 types of relations, while 483,142 triples are used as the training set, and 59,071 triples are used as the test set. For WN18, the training parameters are configured as learning rate $\lambda = 0.001$, batch size $B = 1200$, $k = 50$, $\alpha_1 = 0.2$, $\alpha_2 = 0.1$, l_1-norm, $\gamma_1 = 3$, $\gamma_2 = 3$. For FB15k, the training parameters are configured as learning rate $\lambda = 0.001$, batch size $B = 4800$, $k = 100$, $\alpha_1 = 0.5$, $\alpha_2 = 0.5$, l_1-norm, $\gamma_1 = 1$, $\gamma_2 = 1$. Training process is limited to 500 epochs like TransR. The model is trained on each training set, and given a case (h, r, t) from the test set, $Hits@10$ and $Mean$ are recorded for $\mathbf{r}$ and $f_{2,r}(\mathbf{t}) - f_{1,r}(\mathbf{h})$. In table 4.7, we report the comparison of the current variant of On2Vec against other approaches. And we do see a slight improvement of On2Vec on the state-of-the-art.

On the other side, since the Freebase and WordNet data sets only contain a small portion of those special relations, we expect On2Vec to receive further advantage on Yago Ontology (whose statistics are reported in Table 2.2) where the majority of relations are with relational properties or hierarchies, therefore test the effectiveness of On2Vec on ontology extraction. We are also going to test other On2Vec variants with different implementation of $f_{1,r}$, $f_{2,r}$ and ω functions.

4.2.2 Supervised Relation Extraction

Supervised relation extraction mines relations between entities using deep neural networks such as CNN [ZLC15, NG15, LSL16] and RNN [TQL15]. A typical CNN-based learning structure is shown as Figure 4.2. Training corpora are organized as entity pairs, the bag of sentences that mentions each entity pair, and the label of the relation between such each entity pair. A pre-trained feature model is used to transform each bag of sentences and the pair of entities it mentions into a tensor. Along with other signals, such as distant supervision [ZLC15] and selective attention [LSL16], the tensor is fed into several convolution layers of different kernel sizes separately, which coincide after a max-pooling layer, then the last soft-max layer produces the relation as a class label. The entire training process is realized by back-propagation.

Although this task produces structured knowledge, the feature model of previous work is large-

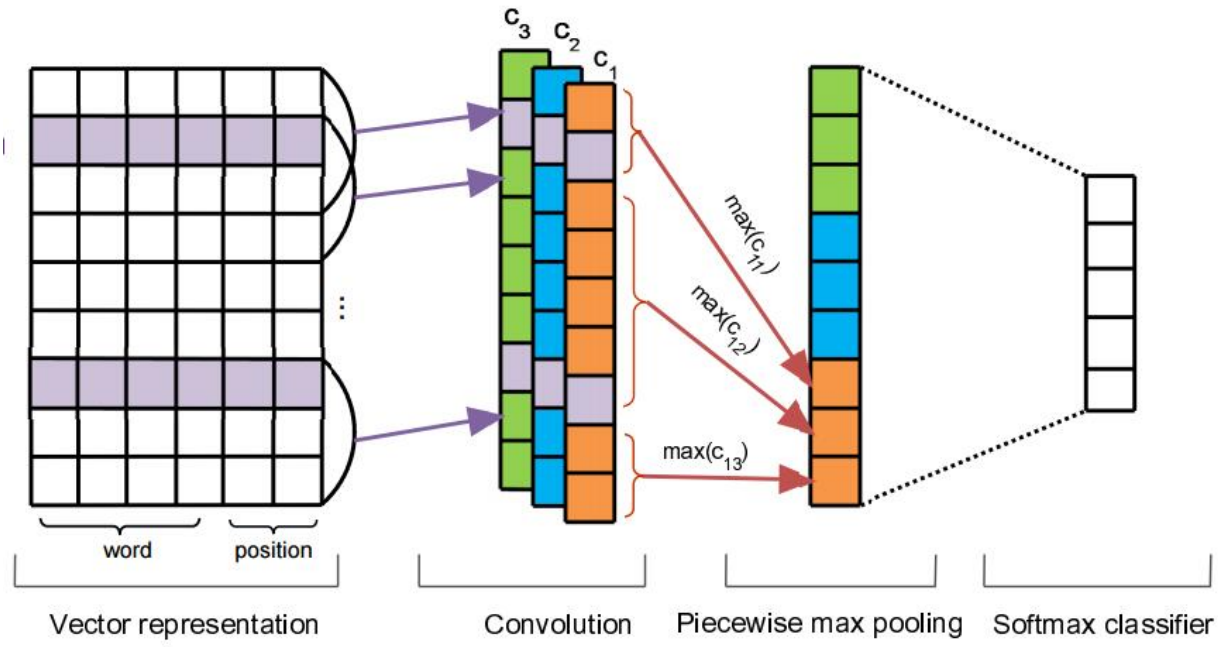

Figure 4.2: Supervised relation extraction learning using CNN.

ly limited to word embeddings, which is weak since it lacks the important signal from knowledge structures themselves. Therefore, we seek to substitute the feature model using a word-graph joint embedding model. Along with other signals including distant supervision and selective attention, the new representation of bags of sentences is expected to be improved as the latent semantics of knowledge graph structures is involved.

4.3 Semantic Relatedness Analysis

In this section, we address the plan of using embedding-based techniques on unsupervised document semantic relatedness analysis.

We evaluate embedding-based unsupervised semantic relatedness analysis based on the LP50 data set [LNN05]. In LP50 there are 50 short documents (51 to 126 words each) from the Australian Broadcasting Corporation's news mail service, i.e. 1225 document pairs. 83 university students were asked to rate the relatedness of the document pairs from scale 1 (not related) to 5 (very related). Each document pair is rated by 8 to 12 different people. We denote d as the distri-

Table 4.8: Results of related work and our preliminary method on LP50. We also compute the human upperbound by randomly separating the human scorers into two equal-numbered groups, which gets 0.77-0.80 of Pearson correlation coefficient.

Method	Pearson correlation coefficient
Takelab [SGK12]	0.08
DKPro [BBG12]	0.21
Bag-of-words [LNN05]	0.1-0.5
ESA [BZG11]	0.46-0.59
LSA [LNN05]	0.60
GED unweighted [SP14]	0.61
GED weighted [SP14]	0.63
ConceptGraphSim [NXC16]	0.73
WikiWalk [YRM09]	0.74
Annotated Skip-gram	0.66

bution of semantic distances of the document pairs, each of which is obtained by normalizing the averaged human-ratings between 0 and 1.

Like some of the related work [SP14, NXC16], we use TagMe [FS12] to highlight all Wikipedia entities appearing in the LP50 documents. Then we used the normalized mean embedding vectors of highlighted entities to obtain the vector representation of each document. Therefore, each predicted distance of the documents are simply calculating by the cosine distance of their vector representation, whose distribution is denoted as $\hat{d}$. The Pearson correlation coefficient to be used to evaluate the prediction result is defined as:

$$\rho_{d,\hat{d}} = \frac{E((d - \mu_d)(\hat{d} - \mu_{\hat{d}}))}{\sigma_d \cdot \sigma_{\hat{d}}}$$

where E stands for the expectation, μ stands for the mean, and σ stands for the standard deviation.

Since prior work does not use an embedding-based technique on this task, we first create a baseline solution using annotated Skip-gram trained on the entire Wikipedia dump. To effectively represent the documents, we want the embedding model to recognize as many highlighted entities as possible in its vocabulary. Therefore, instead of directly using the ordinary technique to recognize phrases based on frequent bi-grams or tri-grams in the corpus, we first use prefix-

tree-aided maximum matching [Wei73] to annotate entities appearing in the Wikipedia articles as unique IDs. We also substitute links that appear in the corpus to their referred entity names. The Skip-gram model is trained on the annotated corpus with the dimensionality as 300, minimum frequency as 10, length of context as 20, and number of epochs as 5. Surprisingly, unlike most of the related work that conceives semantic representation from the Wikipedia graph structures [BZG11, SP14, YRM09], the Skip-gram model that is learnt from annotated plain text has already performed better than some recent work (though is still outperformed by another few). Meanwhile, we have also developed the entire workflow for analyzing document similarity analysis based on different latent representations of documents. Therefore the other results will be shortly obtained as we apply the newly-developed embedding models.

For the next step on this task, we have fetched and cleaned the subgraph that contains all the triples related to the highlighted entities from Yago3. We plan to train the On2Vec on the Yago3 subgraph and the On2Vec and Skip-gram joint embeddings on the Yago3 subgraph and annotated Wikipedia dump, so as to see if knowledge-graph-augmented latent semantics can improve the task of unsupervised monolingual semantic relatedness analysis.

Substituting the above embedding models with MTransE directly enables the relatedness analysis on multilingual documents such as the Miller-Charles data sets [HM09], where encyclopedia concepts of different languages are paired and rated for relatedness using a similar 5-point scale. The additional difficulty introduced by this task is how to highlight key concepts in articles of different languages, which we plan to use the same prefix-tree-aided maximum matching approach for annotating entities on Wikipedia dumps.

4.4 Sentiment Analysis

The last proposed task is training sentiment analysis model using convolutional neural networks. The network structure will be the same as the one in Section 4.2.2, which is also trained as a classifier. Different from related work in Section 2.3.4 which all rely on pre-trained word embeddings for sentence modeling, we are using On2Vec in this task for knowledge graph augmented sentence

Table 4.9: Data sets for sentiment analysis. For each data set, whether to use the specified test set size, or 10-fold cross validation (CV) is as summarized in [Kim14].

Data set	#classes	average length	#sentences	#vocabulary	test set
Movie Review [PL05]	2	20	10662	18765	CV
Stanford Treebank [SPW13]	5	18	11855	17836	2210
Subjectivity [PL04]	2	23	21323	10000	CV
TREC [LR02]	6	10	5952	9592	500
Customer Review [HL04]	2	19	3375	5340	CV

representation. Two approaches are considered, i.e. tensor representation from text and knowledge joint embeddings (Section 4.2.2), or aggregated vector representation on highlighted entities (Section 4.3).

The experiment will be conducted on five comprehensive sentence sets on different topic domains shown in Table 4.9, and results will be compared against the other approaches summarized in [Kim14].

Chapter 5

Research Plan

Our future research focuses on two aspects, i.e. 1) continue developing the approaches for learning latent representation on multi-faceted relational knowledge, and learning joint with word contexts; 2) solve various natural language processing tasks based on the learnt latent representation. We propose the following research plan for the next 24 months. The schedule is tentative and content orders are subject to change.

- ***Improving Property-preserving Embeddings***. In *Winter* and *Spring 2017* we will focus on implementing several variants of On2Vec obtained by adopting different role-dedicated mapping models and hierarchy models. This stage of work also includes the release of a cleaned Yago-based data set, which being different from the two common data sets FB15k (Freebase-based) and WN18 (WordNet-based), contains a majority of triples with relational properties and hierarchical relations. The experiments of unsupervised relation extraction and link prediction will be extended on the Freebase-based and WordNet-based data sets to show the ability of On2Vec variants in encoding simple relations in ordinary knowledge graphs. The same study will also be conducted on the Yago-based data set to determine the performance on handling special graphs in Ontology graphs.

- ***Joint Embeddings***. The text-graph joint embeddings will be implemented and tuned in *Fall 2017* such that future work will be able to use it as the feature model for modeling sentences and documents. The best variant of On2Vec, TransR and TransE will be adopted as the

knowledge model component for training the joint embeddings, and Skip-gram or Glove will be adopted as the text model. For corpora, this model will be using on annotated Wikipedia dump (Section 4.3) plus a Yago-based ontology graph that shares high coverage of the entity vocabularies with the annotated Wikipedia dump.

- ***Supervised Relation Extraction***. The research on supervised relation extraction from plain text will start from *Summer 2017*. A neural-based relation extractor will be implemented and experimentally evaluated on Google's knowledge graph project during the author's summer cooperation with Google Search Team. In detail, this version of model uses knowledge-graph-augmented vector representation to model Wikipedia's entity short descriptions along with other (optional) signals including distant supervision and selective attention. Held-out evaluation will be conducted to reflect the effectiveness of the extractor, while the Search Team will also use our model to extract non-preexisting relational data. Starting from *Winter 2018* to *Spring 2018* the extractor will be further tested on extracting relations from Wikipedia articles when the text-graph joint embeddings are incorporated.

- ***Sentiment Analysis***. Sentiment analysis experiments will be conducted in *Summer 2017* to classify Wikipedia articles with missing domains, which is a part of the project assigned to the author by the Google Search Team. In *Summer 2017* we will first use the same classifier structure (with modified layers) developed for supervised relation extraction. Since this task is going to be extended to different domain-specific learning resources as stated in Section 4.4, this work will be continued in *Spring 2018 to Fall 2018*, during which time we plan to implement the classifer using other neural-based models as well.

- ***Semantic Relatedness Analysis***. For monolingual semantic relatedness analysis, as stated in Section 4.3, we have developed the workflow for the analysis, which has already been used to evaluate the annotated Skip-gram-based approach. Therefore, during the time scale from *Fall 2017* to *Winter 2018*, more results are expected as we incorporate newly-developed embedding models. We postpone the work on cross-lingual semantic relatedness analysis to

Winter 2019 since it is expected to follow the public release of MTransE and corresponding learning resources, and requires cultivating rated multilingual documented sets for evaluating this task.

- ***Cross-lingual Tasks***. We have completed the development of MTransE variants and evaluated their performance on knowledge alignment. Currently we are on our way in publishing this work and releasing the corresponding implementation and learning resources. We plan to improve MTransE by incorporating knowledge models with relation-specific entity transformations as well as alignment models with other characterization techniques after *Winter 2019*.

Bibliography

[ABC16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. "TensorFlow: a system for large-scale machine learning." In *OSDI*, pp. 265–283. USENIX Association, 2016.

[AV04] Grigoris Antoniou and Frank Van Harmelen. "Web ontology language: Owl." In *Handbook on ontologies*, pp. 67–92. Springer, 2004.

[BBG12] Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. "Ukp: Computing semantic textual similarity by combining multiple content similarity measures." In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pp. 435–440. ACL, 2012.

[BEP08] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge." In *SIGMOD*, pp. 1247–1250. ACM, 2008.

[BF13] Francis Bond and Ryan Foster. "Linking and Extending an Open Multilingual Wordnet." In *ACL*, pp. 1352–1362, 2013.

[BGW12] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. "Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing." In *AISTATS*, pp. 127–135, 2012.

[BGW14] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. "A semantic matching energy function for learning with multi-relational data." *Machine Learning*, **94**(2):233–259, 2014.

[BUG13] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. "Translating embeddings for modeling multi-relational data." In *NIPS*, pp. 2787–2795, 2013.

[BWC11] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. "Learning Structured Embeddings of Knowledge Bases." In *AAAI*, pp. 301–306, 2011.

[BWU14] Antoine Bordes, Jason Weston, and Nicolas Usunier. "Open question answering with weakly supervised embedding models." In *ECML-PKDD*, pp. 165–180, 2014.

[BZG11] Daniel Bär, Torsten Zesch, and Iryna Gurevych. "A Reflective View on Text Similarity." In *RANLP*, pp. 515–520, 2011.

[CBB06] Elena Camossi, Michela Bertolotto, and Elisa Bertino. "A multigranular object-oriented framework supporting spatio-temporal granularity conversions." *International Journal of Geographical Information Science*, **20**(05):511–534, 2006.

[CC05] Shui-Lung Chuang and Lee-Feng Chien. "Taxonomy generation for text segments: A practical web-based approach." *ACM Transactions on Information Systems (TOIS)*, **23**(4):363–396, 2005.

[CGT11] Bee-Chung Chen, Jian Guo, Belle Tseng, and Jie Yang. "User reputation in a comment rating environment." In *KDD*, pp. 159–167. ACM, 2011.

[CGW16] Muhao Chen, Shi Gao, and X Sean Wang. "Converting spatiotemporal data Among heterogeneous granularity systems." In *FUZZ-IEEE*, pp. 984–992. IEEE, 2016.

[CHH07] Philipp Cimiano, Peter Haase, Matthias Herold, Matthias Mantel, and Paul Buitelaar. "LexOnto: A model for ontology lexicons for ontology-based NLP." In *ISWC*, 2007.

[CS04] Aron Culotta and Jeffrey Sorensen. "Dependency tree kernels for relation extraction." In *ACL*, p. 423, 2004.

[CTY16] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. "Multi-lingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment." *arXiv preprint arXiv:1611.03954*, 2016.

[CW08] Ronan Collobert and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." In *ICML*, pp. 160–167, 2008.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *JMLR*, **12**(Jul):2121–2159, 2011.

[FD14] Manaal Faruqui and Chris Dyer. "Improving vector space word representations using multilingual correlation." *EACL*, 2014.

[FS12] Paolo Ferragina and Ugo Scaiella. "Fast and accurate annotation of short texts with wikipedia pages." *IEEE software*, **29**(1):70–75, 2012.

[GM07] Evgeniy Gabrilovich and Shaul Markovitch. "Computing semantic relatedness using wikipedia-based explicit semantic analysis." In *IJCAI*, volume 7, pp. 1606–1611, 2007.

[GMC00] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. "Multi-document summarization by sentence extraction." In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization-Volume 4*, pp. 40–48. ACL, 2000.

[HK10] Hsun-Hui Huang and Yau-Hwang Kuo. "Cross-lingual document representation and semantic similarity measure: a fuzzy set and rough set based approach." *IEEE Transactions on Fuzzy Systems*, **18**(6):1098–1111, 2010.

[HL04] Minqing Hu and Bing Liu. "Mining and summarizing customer reviews." In *KDD*, pp. 168–177, 2004.

[HM09] Samer Hassan and Rada Mihalcea. "Cross-lingual semantic relatedness using encyclopedic knowledge." In *EMNLP*, pp. 1192–1201, 2009.

[JHX15] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. "Knowledge graph embedding via dynamic mapping matrix." In *ACL*, pp. 687–696, 2015.

[JRB12] Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. "A latent factor model for highly multi-relational data." In *NIPS*, pp. 3167–3175, 2012.

[JWL16] Yantao Jia, Yuanzhuo Wang, Hailun Lin, Xiaolong Jin, and Xueqi Cheng. "Locally Adaptive Translation for Knowledge Graph Embedding." In *AAAI*, 2016.

[JZ07] Jing Jiang and ChengXiang Zhai. "A Systematic Exploration of the Feature Space for Relation Extraction." In *NAACL HLT*, pp. 113–120, 2007.

[Kim14] Yoon Kim. "Convolutional neural networks for sentence classification." *EMNLP*, 2014.

[KSK15] Matt J Kusner, Yu Sun, Nicholas I Kolkin, Kilian Q Weinberger, et al. "From Word Embeddings To Document Distances." In *ICML*, volume 15, pp. 957–966, 2015.

[LIJ15] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia." *Semantic Web*, **6**(2):167–195, 2015.

[LLS15] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. "Learning Entity and Relation Embeddings for Knowledge Graph Completion." In *AAAI*, pp. 2181–2187, 2015.

[LLY04] Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. "An effective approach to document retrieval via utilizing WordNet and recognizing phrases." In *SIGIR*, pp. 266–272. ACM, 2004.

[LNN05] Michael D Lee, DAniel J Navarro, and Hannah Nikkerud. "An empirical evaluation of models of text document similarity." In *CogSci*, volume 27, 2005.

[LR02] Xin Li and Dan Roth. "Learning question classifiers." In *ACL*, pp. 1–7, 2002.

[LSL16] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. "Neural relation extraction with selective attention over instances." In *ACL*, volume 1, pp. 2124–2133, 2016.

[MAG14] Hamid Mousavi, Maurizio Atzori, Shi Gao, and Carlo Zaniolo. "Text-Mining, Structured Queries, and Knowledge Management on Web Document Corpora." *SIGMOD Record*, **43**(3):48–54, 2014.

[MBS15] Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. "Yago3: A knowledge base from multilingual Wikipedias." In *CIDR*, 2015.

[MC04] Tony Mullen and Nigel Collier. "Sentiment Analysis using Support Vector Machines with Diverse Information Sources." In *EMNLP*, volume 4, pp. 412–418, 2004.

[MCC13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." *ICLR*, 2013.

[MGK14] Hamid Mousavi, Shi Gao, Deirdre Kerr, Markus Iseli, and Carlo Zaniolo. "Mining Semantics Structures from Syntactic Structures in Web Document Corpora." *International Journal of Semantic Computing*, **8**(04):461–489, 2014.

[MGS05] Simone Marinai, Marco Gori, and Giovanni Soda. "Artificial neural networks for document analysis and recognition." *IEEE TPAMI*, **27**(1):23–35, 2005.

[MLS13] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. "Exploiting similarities among languages for machine translation." *arXiv*, 2013.

[MY01] Larry M Manevitz and Malik Yousef. "One-class SVMs for document classification." *JMLR*, **2**(Dec):139–154, 2001.

[NG15] Thien Huu Nguyen and Ralph Grishman. "Relation extraction: Perspective from convolutional neural networks." In *NAACL-HLT*, pp. 39–48, 2015.

[NMN11] Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen, and Juliana Freire. "Multilingual schema matching for Wikipedia infoboxes." *PVLDB*, **5**(2):133–144, 2011.

[NRP16] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. "Holographic Embeddings of Knowledge Graphs." In *AAAI*, 2016.

[NSQ16] Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. "STransE: a novel embedding model of entities and relationships in knowledge bases." In *NAACL HLT*, pp. 460–466, 2016.

[NTK11] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. "A three-way model for collective learning on multi-relational data." In *ICML*, pp. 809–816, 2011.

[NXC16] Yuan Ni, Qiong Kai Xu, Feng Cao, Yosi Mass, Dafna Sheinwald, Hui Jia Zhu, and Shao Sheng Cao. "Semantic documents relatedness using concept graph representation." In *WSDM*, pp. 635–644. ACM, 2016.

[PL04] Bo Pang and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." In *ACL*, 2004.

[PL05] Bo Pang and Lillian Lee. "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales." In *ACL*, pp. 115–124, 2005.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global Vectors for Word Representation." In *EMNLP*, volume 14, pp. 1532–1543, 2014.

[RLN13] Daniel Rinser, Dustin Lange, Felix Naumann, and Gerhard Weikum. "Cross-lingual entity matching and infobox alignment in Wikipedia." *Information Systems*, **38**(6):887–907, 2013.

[SAS11] Fabian M Suchanek, Serge Abiteboul, Pierre Senellart, and Tom Mitchell. "Paris: Probabilistic alignment of relations, instances, and schema." *PVLDB*, **5**(3):157–168, 2011.

[SCM13] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. "Reasoning with neural tensor networks for knowledge base completion." In *NIPS*, pp. 926–934, 2013.

[SGK12] Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. "Takelab: Systems for measuring semantic text similarity." In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pp. 441–448. ACL, 2012.

[SGS11] Ang Sun, Ralph Grishman, and Satoshi Sekine. "Semi-supervised relation extraction with large-scale word clustering." In *ACL*, pp. 521–529, 2011.

[SH13] Robert Speer and Catherine Havasi. "ConceptNet 5: A large semantic network for relational knowledge." *The Peoples Web Meets NLP*, pp. 161–176, 2013.

[SMG14] Andrew M Saxe, James L McClelland, and Surya Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks." *ICLR*, 2014.

[SP14] Michael Schuhmacher and Simone Paolo Ponzetto. "Knowledge-based graph document modeling." In *WSDM*, pp. 543–552. ACM, 2014.

[SPW13] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." In *EMNLP*, volume 1631, 2013.

[TC13] Khoi-Nguyen Tran and Peter Christen. "Identifying multilingual Wikipedia articles based on cross language similarity and activity." In *CIKM*, pp. 1485–1488, 2013.

[TQL15] Duyu Tang, Bing Qin, and Ting Liu. "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification." In *EMNLP*, pp. 1422–1432, 2015.

[TWY14] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." In *ACL*, pp. 1555–1565, 2014.

[Vra12] Denny Vrandečić. "Wikidata: A new platform for collaborative data collection." In *WWW*, pp. 1063–1064, 2012.

[VSC02] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. "Inferring a semantic representation of text via cross-language correlation analysis." In *NIPS*, volume 1, p. 4, 2002.

[WBY13] Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. "Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction." In *EMNLP*, pp. 1366–1371, 2013.

[Wei73] Peter Weiner. "Linear pattern matching algorithms." In *IEEE Conference Record of 14th Annual Symposium on Switching and Automata Theory*, pp. 1–11. IEEE, 1973.

[Wik16] Wikipedia, 2016. https://www.wikipedia.org/.

[WM03] D Randall Wilson and Tony R Martinez. "The general inefficiency of batch training for gradient descent learning." *Neural Networks*, **16**(10):1429–1451, 2003.

[WM12] Sida Wang and Christopher D Manning. "Baselines and bigrams: Simple, good sentiment and topic classification." In *ACL*, pp. 90–94, 2012.

[WZF14a] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. "Knowledge Graph and Text Jointly Embedding." In *EMNLP*, pp. 1591–1601. ACL, 2014.

[WZF14b] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. "Knowledge Graph Embedding by Translating on Hyperplanes." In *AAAI*, pp. 1112–1119, 2014.

[XWL15] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. "Normalized word embedding and orthogonal transform for bilingual word translation." In *NAACL HLT*, pp. 1006–1011, 2015.

[YLZ15] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Chang. "Network Representation Learning with Rich Text Information." In *IJCAI*, pp. 2111–2117, 2015.

[YOM09] Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang, and Mitsuru Ishizuka. "Unsupervised relation extraction by mining Wikipedia texts using information from the web." In *ACL*, pp. 1021–1029, 2009.

[YRM09] Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. "WikiWalk: random walks on Wikipedia for semantic relatedness." In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pp. 41–49. ACL, 2009.

[YST15] Yang Yang, Yizhou Sun, Jie Tang, Bo Ma, and Juanzi Li. "Entity matching across heterogeneous sources." In *KDD*, pp. 1395–1404, 2015.

[ZLC15] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks." In *EMNLP*, pp. 1753–1762, 2015.

[ZSZ05] Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. "Exploring various knowledge in relation extraction." In *ACL*, pp. 427–434, 2005.

[ZZW15] Huaping Zhong, Jianwen Zhang, Zhen Wang, Hai Wan, and Zheng Chen. "Aligning knowledge and text embeddings by entity descriptions." In *EMNLP*, pp. 267–272, 2015.

Chapter 6

Appendix

6.1 Examples of Knowledge Alignment

We have already shown the effectiveness of MTransE in aligning cross-lingual knowledge, especially the linear-transformation-based variants Var_4 and Var_5. Now we discuss several examples to reveal insights on how our methods may be used in cross-lingual knowledge augmentation.

Table 6.1: Examples of cross-lingual entity matching.

Entity	Target	Candidates (in ascending order of rank by Euclidean distance)
Barack Obama	French	**Barack Obama**, *George Bush*, *Jimmy Carter*, George Kalkoa
	German	**Barack Obama**, *Bill Clinton*, *George h. w. Bush*, Hamid Karzai
Paris	French	**Paris**, *Amsterdam*, **à Paris**, *Manchester*, De Smet
	German	**Paris**, *Languedoc*, *Constantine*, *Saint-maurice*, *Nancy*
California	French	*San Francisco*, *Los Angeles*, *Santa Monica*, **Californie**
	German	**Kalifornien**, *Los Angeles*, *Palm Springs*, *Santa Monica*
rock music	French	*post-punk*, **rock alternatif**, *smooth jazz*, *soul jazz*
	German	**rockmusik**, *soul*, *death metal*, *dance-pop*

We start with the search of cross-lingual counterparts of entities and relations. We choose an entity (or relation) in English and then show the nearest candidates in French and German, respectively. These candidates are listed by decreasing values of the Euclidean distance between their vectors in the target language space and the result point of cross-lingual transition. Several examples are shown in Table 6.1 and Table 6.2. In all tables of this subsection, we mark the exact answers as **boldfaced**, and the conceptually close ones as *italic*. For example, in Table 6.1, besides

Table 6.2: Examples of cross-lingual relation matching.

Relation	Target	Candidates (in ascending order of rank by Euclidean distance)
capital	French	**capitale**, *territoire*, pays accrèditant, lieu de veneration
	German	**hauptstadt**, *hauptort*, *gründungsort*, *city*
nationality	French	**nationalié**, **pays de naissance**, *domicile*, *résidence*
	German	**nationalität**, **nation**, letzter start, *sterbeort*
language	French	**langue**, réalisations, lieu deces, *nationalitè*
	German	**sprache**, **originalsprache**, **lang**, *land*
nickname	French	**surnom**, descendant, texte, *nom de ring*
	German	**spitzname**, originaltitel, *names*, **alternativnamen**

Table 6.3: Examples of cross-lingual triple completion.

Query	Target	Candidates (in ascending order of rank)
(Adam Lambert, genre, ?*t*)	French	*musique indèpendante*, **musique alternative**, ode, **glam rock**
	German	**popmusik**, **dance-pop**, no wave, *soul*
(Ronaldinho, position, ?*t*)	French	**milieu offensif**, **attaquant**, *quarterback*, *latèral gauche*
	German	**stürmer**, *linker flügel*, **angriffsspieler**, *rechter flgel*
(Italy, ?*r*, Rome)	French	**capitale**, **plus grande ville**, **chef-lieu**, garnison
	German	**hauptstadt**, **hauptort**, verwaltungssitz, stadion
(Barack Obama, ?*r*, George Bush)	French	*ministre-prèsident*, **prèdècesseur**, *premier ministre*, *prèsident du conseil*
	German	**vorgänger**, **vorgängerin**, besetzung, lied
(?*h*, instrument, guitar)	French	**Brant Bjork**, **Chris Garneau**, *David Draiman*, **Ian Mackaye**
	German	**Phil Manzanera**, *Styles P.*, *Tina Charles*, **Luke Bryan**

boldfacing the exactly correct answers for Barack Obama and Paris, we consider those who have also been U.S. presidents as conceptually close to Barack Obama, and European cities other than Paris as conceptually close to Paris. Also, in Table 6.2, those French and German relations that have the meaning of settlements of significance are considered as conceptually close to *capital*.

We then move on to the more complicated cross-lingual triple completion task. We construct queries by replacing one element in an English triple with a question mark, for which we seek for answers in another language. Our methods need to transfer the remaining elements to the space of the target language and pick the best answer for the missing element. Table 6.3 shows some query answers. It is noteworthy that the basic queries are already useful for aided cross-lingual

Table 6.4: Statistics of the CN3l data set.

Type of triples	En triples	Fr triples	De triples	Aligned triples
Number of triples	47,696	18,624	25,560	En-Fr:3,668 En-De:8,588
Type of ILLs	En-Fr	Fr-En	En-De	De-En
Number of ILLs	2,154	2,146	3,485	3,813

Table 6.5: Cross-lingual entity matching (CN3l).

Languages	En-Fr		Fr-En		En-De		De-En	
Metric	Hits@10	Mean	Hits@10	Mean	Hits@10	Mean	Hits@10	Mean
LM	25.45	1302.83	20.16	1884.70	30.12	954.71	18.04	1487.90
CCA	27.96	1204.91	26.40	1740.83	28.76	1176.09	25.30	1834.21
OT	68.43	42.30	67.06	33.86	72.34	74.98	69.47	44.38
Var_1	61.37	55.16	69.27	33.60	63.06	74.54	63.56	79.79
Var_2	44.06	226.63	57.15	95.13	49.07	219.97	49.15	214.58
Var_3	73.73	29.34	77.02	14.82	70.55	50.83	70.96	47.99
Var_4	**86.83**	**16.64**	**80.62**	7.86	88.89	**7.16**	**95.67**	**1.47**
Var_5	86.21	16.99	80.19	**7.34**	**89.19**	8.27	95.53	1.63

Table 6.6: Accuracy of triple-wise alignment verification (%).

Languages	En&Fr	En&De
LM	60.53	51.55
CCA	81.57	79.54
OT	93.01	87.59
Var_1	93.92	91.89
Var_2	87.30	82.70
Var_3	88.95	84.80
Var_4	**97.46**	**96.63**
Var_5	97.18	95.42

augmentation of knowledge. However, developing a joint model to support complex queries on multilingual knowledge graphs based on MTransE generated features appears to be a promising future work to support Q&A on multilingual knowledge bases.

Figure 6.1 shows the PCA projection of the *same* six English entities in their original English space and in French space after transformation. We can observe that the vectors of English entities show certain structures, where the U.S. cities are grouped together and other countries' cities are well separated. After transformation into French space, these English entities not only keep their original spatial emergence, but also are close to their corresponding entities in French. This illustrates the transformation preserves mono-lingual structure and also it is able to capture cross-lingual information. We believe this example illustrates the good performance we have demonstrated in cross-lingual tasks including cross-lingual entity matching and triple-wise alignment verification.

6.2 Additional Experimental Results

We derive another data set CN3l from the MIT ConceptNet to evaluate MTransE, whose statistics are shown in Table 6.4. Though being a smaller data set than WK3l-15k, knowledge graphs in CN3l are highly sparse. Thereof, each language version of CN3l contains around 7,500 n-

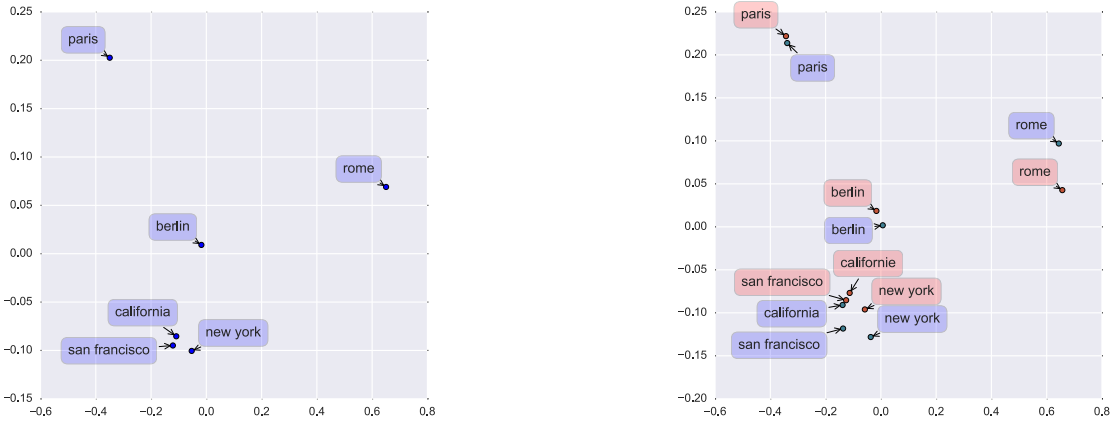

Figure 6.1: Visualization of the result of Var_4 for the same six English entities in their original space (left) and in French space after being transformed (right). English entities are rendered in blue, and the corresponding French entities are in light ruby.

odes and less than 41 types of relations. The alignment sets are created based on the relation `TranslationOf` of the ConceptNet. In this section we report the results of the two cross-lingual tasks on the CN3l data set for MTransE as well as all baselines. Basically, these results lead to similar conclusions as we have on WK3l.

Evaluation Protocol. The metrics and evaluation procedures are the same as those on WK3l. We select λ among $\{0.001, 0.01, 0.05\}$, α among $\{1, 2.5, 5, 7.5\}$, l_1 or l_2 norm in loss functions, and dimensionality k among $\{25, 50, 75\}$. Optimal parameters are configured as $\lambda = 0.001$, $\alpha = 2.5$, $k = 50$, and l_1 norm for all models. To perform the evaluation under controlled variables, we again use one configuration on each model respectively in the two experiments. Since the data set is smaller, training is limited to 200 epochs.

Results of Cross-lingual Entity Matching. The results are reported in Table 6.5. For the baselines, LM and CCA are again left far behind for being disjointedly trained. OT, however, takes the position ahead of Var_1. This is likely because the knowledge graphs in CN3l are highly sparse, therefore fewer interference of monolingual relations among entities makes the orthogonality constraint easier to fulfill. Albeit, in all settings, OT is still largely outperformed by Var_4 and Var_5, which receives amazingly good results, thus steadily being the optimal solutions. Interestingly,

Var_3 now ranks right behind the linear-transformation-based variants in most settings. This is quite reasonable because the cross-lingual transitions, which are regarded as a type of relation by Var_3 in the graphs, are now way less frequent in CN3l than they are in the much denser and more heterogeneous WK3l. Thus, this explains why it performs better than Var_1. For the same reason as we stated in Section 4.1.1, Var_2 is placed at last of the five MTransE variants in matching cross-lingual entities.

Results of Triple-wise Alignment Verification. The results shown in Table 6.6 reflects the same conclusions of the experiment performed on WK3l that, the linear-transformation-based variants takes the lead ahead of the rest MTransE variants and the baselines. While Var_1, despite being the simplest, takes the second place with a satisfying accuracy in triple-wise alignment verification as well. The relation-dedicated calibration still causes interference in the optimization process of Var_2, therefore leads to a 4%$\sim$9% drop of accuracy from Var_1. Var_3 performs slightly better than Var_3. On triple-wise alignment verification on CN3l, we receive exactly the same placement for evaluating the five MTransE variants. Meanwhile, for the baselines, OT is slightly worse than Var_1, CCA also receives acceptable accuracy which is however worse than all MTransE variants, while the accuracy of LM is just slightly better than random guessing.

Above all, the results in CN3l indicates that MTransE also works promisingly on very sparse multilingual graphs, while the linear-transformation-based variants are the best representation techniques.