

Embedding Multilingual and Relational Data Using Linear Mappings



A thesis submitted for the degree of
Doctor of Philosophy

by

Xutan Peng

Department of Computer Science

Faculty of Engineering

The University of Sheffield

October 2022

献给我的外公许平华。再次相见前，我会一直想念您。

ACKNOWLEDGEMENTS

To begin with, I would like to thank Mark Stevenson for being a super supportive supervisor. Beyond the one-hour weekly meetings throughout these three years, he devoted much longer time to our supervision relationship, with patience and enthusiasm. Moreover, he not only supported my academic successes, e.g., by offering invaluable research ideas and suggestions (at all levels), but also taught me the proper research practice *by personal example*.

I am ultimately grateful to my PhD Advisor, Aline Villavicencio, for sharing indispensable knowledge, recommending career-advancing opportunities, and providing warm support during the dark moments.

I want to show my sincerest gratitude to Haiping Lu, who played a vital role in my PhD Panel Committee, as well as Nikos Aletras and Loïc Barrault, who kindly aided me as PhD Tutors.

Many thanks to all the lovely people and brilliant minds who have once collaborated or even co-authored with me. Special thanks to Chen Li, who guided me into this fantastic research field and gave generous mentorship. I also want to express my honour of being the colleague of Zheng Li, Chen Luo, Haoming Jiang, Qingyu Yin, Xianfeng Tang, and Bing Yin during the internship within the Amazon QU Team.

Many thanks to Ruizhe Li, Guanyi Chen, Yida Mu, Xiao Li, Rui Mao, Mali Jin, George Chrysostomou, Katerina Margatina, Danae Sanchez Villegas, Xingyi Song, and Haiyang Zhang. I could never finish this long journey without their encouragement. I should extend this deepest gratitude to other friends from the NLP Group, SpandH Group, Machine Learning Group, and SLT-CDT. I will

miss all the time spent and happiness gained in the labs, seminar rooms, pubs, restaurants, and the splendid Peak District.

Many thanks to Yidan Liu, for her love to me all these years, in every possible way. She said I should pursue this degree, and she was right!

Finally, I would like to give a shout-out to my parents. They have always been my most powerful and secure backing.

ABSTRACT

This thesis presents our research on the embedding method, a machine learning technique that encodes real-world signals into high-dimensional vectors. Specifically, we focus on a family of algorithms whose backbone is one simple yet elegant type of topological operation, the *linear mapping*, aka. *linear transformation* or *vector space homomorphism*. Past studies have shown the usefulness of these approaches for modelling complex data, such as lexicons from different languages and networks storing factual relations. However, they also exhibit crucial limitations, including a lack of theoretical justifications, precision drop in challenging setups, and considerable environmental impact during training, among others.

To bridge these gaps, we first identify the unnoticed link between the success of linear Cross-Lingual Word Embedding (CLWE) mappings and the preservation of the implicit analogy relation, using both theoretical and empirical evidence. Next, we propose a post-hoc ℓ_1 -norm rotation step which substantially improves the performance of existing CLWE mappings. Then, beyond solving conventional questions where only modern languages are involved, we extend the application of CLWE mappings to summarising lengthy and opaque historical text. Finally, motivated by the learning procedure of CLWE models, we adopt linear mappings to optimise Knowledge Graph Embeddings (KGEs) iteratively, significantly reducing the carbon footprint required to train the algorithm.

CONTENTS

Acknowledgements	i
Abstract	iii
Acronyms	ix
Language Codes	xi
1 Introduction	1
1.1 Contributions	3
1.2 Thesis Overview	4
1.3 Published Material	8
2 Background	9
2.1 Embedding Cross-Lingual Lexicons	9
2.1.1 Evolution of CLWE Algorithms	10
2.1.2 Debates on Linearity of CLWE Mappings	13
2.1.3 The ℓ_2 Refinement Algorithm	14
2.2 Embedding Real-World Relations	16
2.2.1 Encoding Analogies	16
2.2.2 Encoding Knowledge Graphs	17
2.2.3 Improving Efficiency of Relation Encoding	18
2.3 Summarising Historical Text	19
2.3.1 Text Processing for Historical Languages	20

2.3.2	Cross-Lingual Summarisation and Beyond	21
3	Understanding Linearity of CLWE Mappings	23
3.1	Theoretical Basis	24
3.2	Experiment	27
3.2.1	Indicators	27
3.2.2	Datasets	31
3.2.3	Word Embeddings	34
3.3	Result	34
3.4	Application	39
3.5	Further Implications	41
3.6	Summary and Discussion	43
3.7	Post-Publication Retrospect	44
4	Refining CLWE Mappings via ℓ_1 Norm Optimisation	45
4.1	Methodology	46
4.2	Experimental Setup	49
4.2.1	Datasets	49
4.2.2	Baselines	50
4.2.3	Implementation Details	51
4.3	Results	51
4.3.1	Bilingual Lexicon Induction	52
4.3.2	Natural Language Inference	59
4.4	Summary and Discussion	61
4.5	Post-Publication Retrospect	61
5	Applying CLWE Mappings for Historical Text Summarisation	63
5.1	HISTSUMM Corpus	64
5.1.1	Dataset Construction	64
5.1.2	Dataset Statistics	66
5.1.3	Vicissitudes of News	68
5.2	Methodology	69
5.3	Experimental Setup	72
5.3.1	Training Data	72

5.3.2	Baseline Approaches	73
5.3.3	Model Configurations	73
5.4	Results and Analyses	74
5.4.1	Automatic Evaluation	74
5.4.2	Human Judgement	77
5.4.3	Error Analysis	79
5.5	Summary and Discussion	79
5.6	Post-Publication Retrospect	80
6	Learning KGE Efficiently by Mapping Relational Matrices	81
6.1	Methodology	82
6.1.1	Preliminaries: Segmented Embeddings	82
6.1.2	Efficient KGE Optimisation	84
6.2	Experiment	87
6.2.1	Setups	87
6.2.2	Main Results	89
6.2.3	Ablation Studies	93
6.2.4	Impacts of Dimensionality	95
6.2.5	Interpreting Entity Embeddings	96
6.3	Summary and Discussion	98
6.4	Post-Publication Retrospect	98
7	Conclusions	99
7.1	Summary of Thesis	99
7.2	Evaluation of Thesis Goals	101
7.3	Future Directions	102
	Bibliography	104

ACRONYMS

NLP	Natural Language Processing
CLWE	Cross-Lingual Word Embedding
KGE	Knowledge Graph Embedding
BLI	Bilingual Lexicon Induction
NLI	Natural Language Inference
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
OPA	Orthogonal Procrustes Analysis
ODE	Ordinary Differential Equation
PLM	Pre-trained Language Model
GPU	Graphics Processing Unit
CPU	Central Processing Unit
\mathcal{S}_{LMP}	Score for the linearity of mapping (see Eq. (3.9))
\mathcal{S}_{PAE}	Score for the preservation of analogy encoding (see Eq. (3.10))
xANLG	Cross-lingual analogy dataset (see Chapter 3)

HISTSUMM . . . Historical text summarisation dataset (see Chapter 5)

MRR Mean Reciprocal Rank

ACC Accuracy

INVOLVED LANGUAGES

Family	Branch / Group	Name	Code	
Indo-European	Slavonic	Croatian	HR	
		Polish	PL	
		Russian	RU	
		Slovene	SL	
	Romance	French	FR	
		Italian	IT	
		Spanish	ES	
	Germanic	German	DE	
		English	EN	
	Indo-Aryan	Baltic	Latvian	LV
		Hindi	HI	
Uralic	Finnic	Finnish	FI	
		Estonian	ET	
Sino-Tibetan	Sinitic	Chinese	ZH	
Turkic	Oghuz	Turkish	TR	

INTRODUCTION

Learning encodings for multilingual text and relational data have been two long-standing research goals in the field of Representation Learning (Shannon, 1948; Benenfeld, 1968; Schreiber et al., 1993; Lample and Conneau, 2019; Wang et al., 2021b). The aim of the former is to build a shared space for different languages, so that closely located vectors correspond to semantically related linguistic elements, while the motivation being to mitigate the communication barrier caused by the diversity of human language. The latter takes relation-rich structures (e.g., a network whose nodes are entities and edges are relations) as input and stores the information in high-dimensional vector spaces. Such a relation encoder is an essential component for complex knowledge-based systems.

Conventional methods that perform these two types of embeddings depend upon human input from domain experts and tend to be unreliable in the real world (Hermans, 1996; Gennari et al., 2003). Recently, with the rapid development in the field of Representation Learning, techniques such as Cross-Lingual Word Embedding (CLWE) and Knowledge Graph Embedding (KGE) have emerged. These methods have not only performed remarkably well on mainstream benchmarks, but also reduced the requirement of supervision to the minimum. In addition, some of the most promising CLWE and KGE approaches are centred around the optimisation of linear mappings. This simple linear algebra function could be easily calculated on modern computing devices such as GPUs; therefore, it has become popular means of modelling multilingual and relational

signals (Ruder et al., 2019; Wang et al., 2021b).

Nevertheless, these promising embedding methods are still limited from several perspectives. Firstly, some fundamental hypotheses are not fully understood, thus weakening the models' explainability. For instance, ground-truth CLWE mappings are often assumed to be linear, with both supporting (Mikolov et al., 2013b; Glavaš et al., 2019) and contradicting (Nakashole, 2018; Wang et al., 2021a) experimental results. A more in-depth exploration with theoretical insights is thus needed to settle this debate. Secondly, mapping-based embeddings are not robust enough, e.g., CLWEs may handle challenging setups ineffectively, such as the alignment between polysemous entries (Søgaard et al., 2018). Thirdly, although these embedding algorithms are utilised widely, their potential has not been extensively verified yet. One example is that while CLWEs have successfully connected modern languages (even the low-resource ones), they have not been applied in tackling problems where historical languages are available (cf. § 2.3). Last but not least, even though adopting linear mappings has reduced modelling complexity, many embedding approaches still require massive computational resources, contributing to the emission of greenhouse gases (Strubell et al., 2019).

This thesis, therefore, aims to find answers to the following research questions:

- **When** does the linear mapping make an appropriate approximating function for encoding complex signals such as multilingual lexicons? In particular, can we identify a condition, theoretically and empirically, under which the structures of word embeddings for different languages are similar?
- **How** to improve the embedding precision in difficult scenarios? To achieve this, we must understand what causes model failures through benchmarking tests, case studies, error analyses, etc.
- **Where** can the linear-mapping-based embedding methods be applied beyond existing usages? How to evaluate system performance for new tasks? What if the new tasks introduce new challenges?
- **Why** the embedding approaches are computationally expensive even with the optimisation simplification by the linear mapping? Is it possible to

design algorithms that could further reduce the computational overhead and make the embedding methods “greener”?

- **Whether** the embedding model for one type of signal can motivate that for another?

1.1 Contributions

This thesis presents novel methods, resources, and insights to explore research questions listed in the previous section. Its main contributions are as follows:

- Introduces the previously unnoticed relationship between the linearity of CLWE mappings and the preservation of encoded word analogies, and provides a theoretical analysis of this relationship.
- Describes the construction of a novel cross-lingual analogy test set with five categories of word pairs aligned across twelve diverse languages.
- Provides empirical evidence of our claim on CLWE mapping linearity, and introduces \mathcal{S}_{PAE} to estimate the analogy encoding preservation (and therefore the mapping linearity). We additionally demonstrate that \mathcal{S}_{PAE} can be used as an indicator of the relationship between monolingual word embeddings, independent of trained CLWEs.
- Identifies the sensitivity against outlier data as one cause of modelling failure when CLWEs are designed using the conventional ℓ_2 -norm linear mappings.
- Introduces ℓ_1 -norm loss to the CLWE community for the first time, so as to enhance the robustness of learned mappings.
- Develops a new ℓ_1 norm optimisation scheme for CLWE refinement, which is tested to be effective on word translation and cross-lingual transfer learning benchmarks.
- Proposes a hitherto unexplored and challenging task, namely historical text summarisation.

- Constructs a high-quality summarisation corpus for historical DE and ZH, with modern DE and ZH summaries by experts, to stimulate research in this field.
- Designs a model for historical text summarisation that does not require parallel supervision and provides a validated high-performing baseline for future studies.
- Introduces three novel approaches to substantially reduce computational overhead of embedding large and complex knowledge graphs: full batch learning based on relational matrices, closed-form Orthogonal Procrustes Analysis for KGEs, and non-negative-sampling training.
- Systemically benchmarks the proposed KGE algorithm against thirteen strong baselines on two standard datasets, demonstrating that it retains highly competitive performance with just order-of-minute training time and emissions of less than making two cups of coffee.
- Successfully encodes both entity and relation information in a single vector space for the first time, thereby enriching the expressiveness of entity embeddings and producing new insights into interpretability.

1.2 Thesis Overview

The remaining chapters of this thesis are structured as follows:

Chapter 2 reviews the background literature that is relevant to the remainder of the thesis, including the advances and debates regarding CLWE (especially the linear-mapping-based paradigm), the embeddings for both analogy relation and real-world factual relations, as well as the background for defining a novel cross-lingual transfer learning task, namely historical text summarisation.

Chapter 3 establishes a link between the linearity of CLWE mappings and the preservation of encoded monolingual analogies. This is motivated by the observation that word analogies can be solved via the composition of semantics based

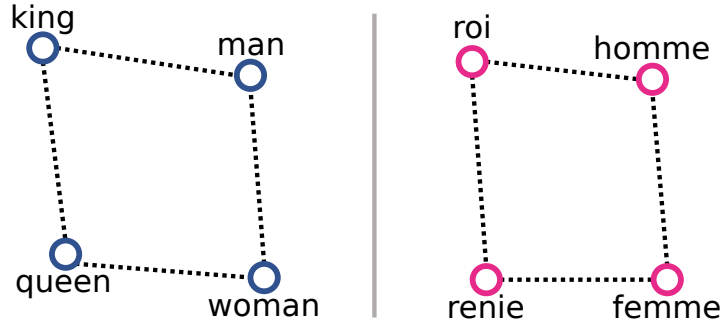


Figure 1.1: Wiki vectors (see § 3.2.3) of English (left) and French (right) analogy word pairs based on Principal Component Analysis (PCA) (Wold et al., 1987). NB: We manually rotate the visualisation to highlight structural similarity.

on vector arithmetic (Mikolov et al., 2013c) and such linguistic regularities might be transferable across languages. More specifically, we notice that if analogies encoded in the embeddings of one language also appear in the embeddings of another, the corresponding multilingual vectors tend to form similar shapes (see Fig. 1.1), suggesting the CLWE mapping between them should be approximately linear. In other words, we suspect that the preservation of analogy encoding indicates the linearity of CLWE mappings.

Our hypothesis is verified both theoretically and empirically. We make a justification that the preservation of analogy encoding should be a *sufficient and necessary* condition for the linearity of CLWE mappings. To provide empirical validation, we first define indicators to qualify the linearity of the ground-truth CLWE mapping (\mathcal{S}_{LMP}) and its preservation of analogy encoding (\mathcal{S}_{PAE}). Next, we build a novel cross-lingual word analogy corpus containing five analogy categories (both semantic and syntactic) for twelve languages that pose pairs of diverse etymological distances. We then benchmark \mathcal{S}_{LMP} and \mathcal{S}_{PAE} on three representative series of word embeddings. In all setups tested, we observe a significant correlation between \mathcal{S}_{LMP} and \mathcal{S}_{PAE} , which provides empirical support for our hypothesis. With this insight, we offer explanations to why the linearity assumption occasionally fails, and consequently, discuss how our research can benefit the development of more effective CLWE algorithms. We also recommend the use of \mathcal{S}_{PAE} to assess mapping linearity in CLWE applications.

Chapter 4 is motivated by the known advantages of ℓ_1 loss (aka. Manhattan distance) over the conventional ℓ_2 loss, as ℓ_1 loss has been mathematically demonstrated to be less affected by outliers (Rousseeuw and Leroy, 1987) and empirically proven useful in computer vision and data mining (Aanæs et al., 2002; De La Torre and Black, 2003; Kwak, 2008). Motivated by this insight, this chapter proposes a simple yet effective post-processing technique to improve the quality of CLWEs: adjust the alignment of *any* cross-lingual vector space to minimise the ℓ_1 loss without violating the orthogonality constraint. Specifically, given existing CLWEs, we bidirectionally retrieve bilingual vectors and optimise their Manhattan distance using a numerical solver. The approach can be applied to any CLWEs, making the post-hoc refinement technique generic and applicable to a wide range of scenarios. We believe this to be the first application of ℓ_1 loss to the CLWE problem.

To demonstrate the effectiveness of our method, we select four state-of-the-art baselines and conduct comprehensive evaluations in both supervised and unsupervised settings. Our experiments involve ten languages from diverse branches/families and embeddings trained on corpora of different domains. In addition to the standard Bilingual Lexicon Induction (BLI) benchmark, we also investigate a downstream task, namely cross-lingual transfer for Natural Language Inference (NLI). In all setups tested, our algorithm significantly improves the performance of strong baselines. Finally, we provide an intuitive visualisation illustrating why ℓ_1 loss is more robust than its ℓ_2 counterpart when refining CLWEs.

Chapter 5 addresses the long-standing need for historical text summarisation through machine summarisation techniques for the first time. We built a high-quality dataset containing historical news articles and corresponding modern summaries. The languages considered are German|DE and Chinese|ZH, mainly due to the following reasons. First, they both have rich textual heritage and accessible (monolingual) training resources for historical and modern language forms. Second, they serve as outstanding representatives of two distinct writing systems (DE for alphabetic and ZH for ideographic languages), and investigating them can lead to generalisable insights for a wide range of other languages. Third, we

have access to linguistic experts in both languages, for composing high-quality gold-standard modern-language summarises for DE and ZH news stories published hundreds of years ago, and for evaluating the output of machine summarisers.

In order to tackle the challenge of a limited amount of resources available for model training (e.g., we have summarisation training data only for the monolingual task with modern languages, and very limited parallel corpora for modern and historical forms of the languages), we propose a transfer-learning-based approach which can be bootstrapped even without cross-lingual supervision. To our knowledge, our work is the first to consider the task of historical text summarisation. As a result, there are no directly relevant methods to compare against. We instead implement two state-of-the-art baselines for standard cross-lingual summarisation, and conduct extensive automatic and human evaluations to show that our proposed method yields better results. Our approach, therefore, provides a strong baseline for future studies on this task to benchmark against.

Chapter 6 is motivated by research in CLWE mappings. To alleviate the computational cost of existing KGE models, we introduce PROCRUSTES, a lightweight, fast, and eco-friendly KGE training technique. PROCRUSTES is built upon three novel techniques. First, to reduce the batch-wise computational overhead, we propose to parallelise batches by grouping tuples according to their relations, which ultimately enables efficient full batch learning. Second, we turn to a closed-form solution for Orthogonal Procrustes Problem to boost the embedding training, which has never been explored in the context of KGEs. Third, to break through the bandwidth bottleneck, our algorithm is allowed to be trained without negative samples.

To verify the effectiveness and efficiency of our proposed method, we benchmark two popular datasets (WN18RR and FB15k-237) against thirteen strong baselines. Experimental results show that PROCRUSTES yields performance competitive with the state-of-the-art while also reducing training time by up to 98.4% and the carbon footprint by up to 99.3%. In addition, we found that our algorithm can produce easily interpretable entity embeddings with richer semantics than previous approaches.

Chapter 7 summarises the highlights of this thesis and discusses potential future work.

1.3 Published Material

Peer-reviewed publications contributing to this thesis:

- Chapter 3: **Xutan Peng**, Mark Stevenson, Chenghua Lin, Chen Li. Understanding Linearity of Cross-Lingual Word Embedding Mappings. *Transactions on Machine Learning Research (TMLR)*, 2022. (Peng et al., 2022)
- Chapter 4: **Xutan Peng**, Chenghua Lin, Mark Stevenson. Cross-Lingual Word Embedding Refinement by ℓ_1 Norm Optimisation. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021. (Peng et al., 2021b)
- Chapter 5: **Xutan Peng**, Yi Zheng, Chenghua Lin, Advait Siddharthan. Summarising Historical Text in Modern Languages. *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021. (Peng et al., 2021c)
- Chapter 6: **Xutan Peng**, Guanyi Chen, Chenghua Lin, Mark Stevenson. Highly Efficient Knowledge Graph Embedding Learning with Orthogonal Procrustes Analysis. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021. (Peng et al., 2021a)

BACKGROUND

This chapter describes a variety of literature closely related to the thesis. First, in § 2.1 we review the research status of CLWE techniques, especially those centred around obtaining the optimal mapping function. We highlight the ongoing debate on whether a CLWE mapping should be assumed linear, which is the key topic in Chapter 3. We also discuss the popular ℓ_2 -norm training objective as it is relevant to the post-hoc CLWE refinement algorithm proposed in Chapter 4. Next, section § 2.2 covers embedding methods for modelling two representative categories of real-world relations: semantic word analogies and the explicit factual triples from knowledge graphs. They are respectively associated with work described in Chapters 3 and 6. Since the major advantage of the algorithm proposed in Chapter 6 is the significantly boosted training speed and reduced environmental impact, in § 2.2.3 we investigate recent advances in making relational embeddings, in particular KGEs, computationally, and therefore also energy, efficient. Finally, § 2.3 presents previous work on modelling historical text using language technologies, the main task in Chapter 5, in addition to work in a related area, cross-lingual summarisation.

2.1 Embedding Cross-Lingual Lexicons

CLWEs encode words from two or more languages in a shared high-dimensional space in which vectors representing lexical items with similar meanings (regard-

less of language) are closely located. Compared with more recently and complex techniques such as cross-lingual Pre-trained Language Models (PLMs), CLWE is orders of magnitude more efficient in terms of training corpora. For example, Kim et al. (2020) show that inadequate monolingual data size (fewer than one million *sentences*) is likely to lead to the collapsed performance of XLM (Lample and Conneau, 2019) even for etymologically close language pairs. Meanwhile, CLWE can easily align word embeddings for languages such as African Amharic and Tigrinya for which only millions of *tokens* are available (Zhang et al., 2020). CLWE training also requires much less computational power, e.g., XLM-R (Conneau et al., 2020) was trained on $500\times$ Tesla V100 GPUs, whereas the training of VecMap (Artetxe et al., 2018) can be finished on a single Titan Xp GPU. Moreover, CLWEs tend to perform better than even the state-of-the-art cross-lingual PLMs on lexical tasks (e.g., BLI and entity linking) (Vulić et al., 2020a), as it is easier to access the cross-lingual lexical knowledge stored in the former (Vulić et al., 2023). As a result, the topic has received significant attention as a promising means to support Natural Language Processing (NLP) for low-resource languages (including ancient languages) and has been used for a range of applications, e.g., Machine Translation (Herold et al., 2021), Sentiment Analysis (Sun et al., 2021), Question Answering (Zhou et al., 2021) and Text Summarisation (Peng et al., 2021c).

2.1.1 Evolution of CLWE Algorithms

Some pioneer studies on constructing language-independent representations focus on abstract linguistic labels (Aone and McKee, 1993). In the machine translation community, the idea of extracting word translation probabilities from parallel sentences is also frequently visited (e.g., Brown et al. (1993) and Och and Ney (2003)). Similarly, the field has witnessed efforts of learning aligned representations using cross-lingual documents, such as Littman et al. (1998) and De Smet et al. (2011). Following this strand, since the popularity of neural-based methods, one paradigm of generating CLWEs is to train shared semantic representations with multilingual texts aligned at sentence or document level (Vulić and Korhonen, 2016; Upadhyay et al., 2016). Although this research direction has been well

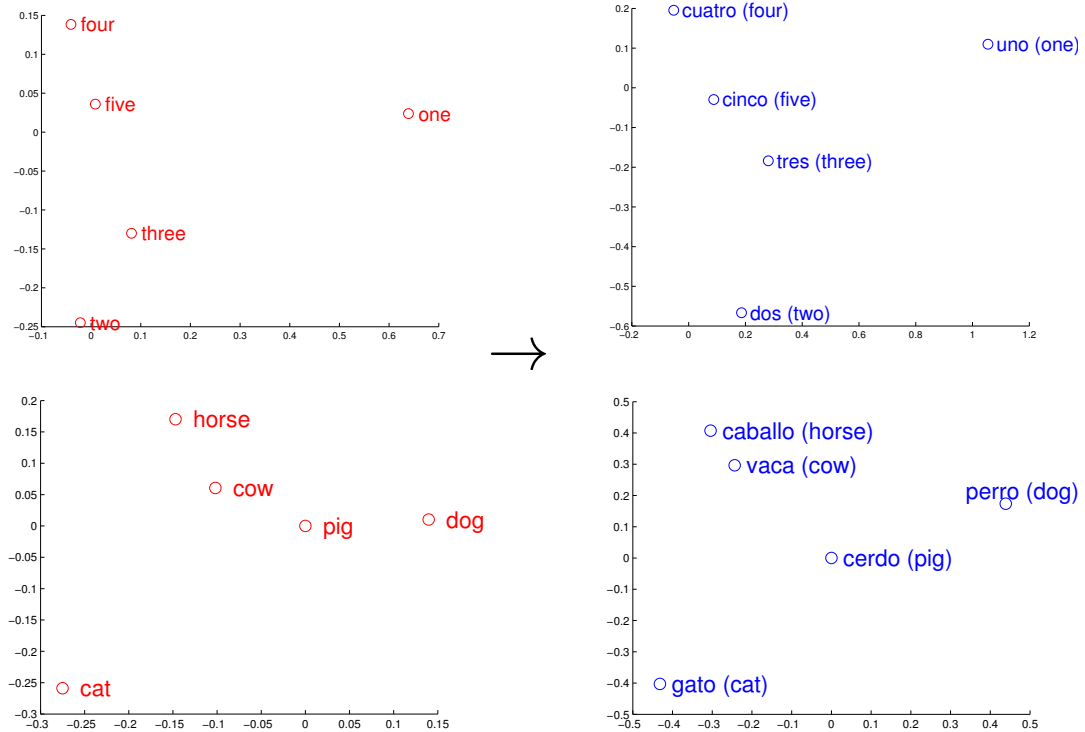


Figure 2.1: Distributed word vector representations of numbers and animals in English (left) and Spanish (right). Taken from Mikolov et al. (2013b).

studied, the parallel setup requirement for model training is expensive, and hence impractical for low-resource languages.

Recent years have seen an increase in interest in projection-based methods. Motivated by the observation that word embeddings for different languages tend to be similar in structure (see Fig. 2.1) (Mikolov et al., 2013b), many researchers have assumed that the mappings between cross-lingual word vectors are linear (Faruqui and Dyer, 2014; Lample et al., 2018b; Li et al., 2021d). Since the input embeddings can be generated independently using monolingual corpora only, projection-based methods reduce the supervision required for training and offer a viable solution for low-resource scenarios.

Xing et al. (2015) showed that the precision of the learned CLWEs can be improved by constraining the mapping function to be orthogonal, which is for-

malised as the so-called ℓ_2 Orthogonal Procrustes Analysis (OPA):

$$\operatorname{argmin}_{\mathbf{M} \in \mathcal{O}} \|\mathbf{A}\mathbf{M} - \mathbf{B}\|_2, \quad (2.1)$$

where \mathbf{M} is the CLWE mapping, \mathcal{O} denotes the orthogonal manifold (aka. the Stiefel manifold (Chu and Trendafilov, 2001)), and \mathbf{A} and \mathbf{B} are matrices composed using vectors from source and target embedding spaces.

While Xing et al. (2015) exploited an approximate and relatively slow gradient-based solver, more recent approaches such as Artetxe et al. (2016) and Smith et al. (2017) introduced an exact closed-form solution for Eq. (2.1). Originally proposed by Schönemann (1966), it utilises Singular Value Decomposition (SVD):

$$\mathbf{M}^* = \mathbf{U}\mathbf{V}^\top, \text{ with } \mathbf{U}\Sigma\mathbf{V}^\top = \text{SVD}(\mathbf{A}^\top\mathbf{B}), \quad (2.2)$$

where \mathbf{M}^* denotes the ℓ_2 -optimal mapping matrix. The efficiency and effectiveness of Eq. (2.2) have led to its application within many other approaches, e.g., Ruder et al. (2018), Joulin et al. (2018) and Glavaš et al. (2019). In particular, PROC-B (Glavaš et al., 2019), a supervised CLWE framework that simply applies multiple iterations of ℓ_2 OPA, has been demonstrated to produce very competitive performance on various benchmark tasks including BLI as well as cross-lingual transfer for NLI and information retrieval.

To achieve more accurate alignment between embedding spaces, effective normalisation techniques are applied (Xing et al., 2015; Artetxe et al., 2016; Ruder et al., 2019). On the one hand, “length standardisation” enforces all word vectors to have the unit length. On the other hand, “mean centring” (for each language) subtracts the average monolingual word vector from all word embeddings, so that this mean vector becomes the origin of the vector space. These steps have the effect of simplifying the mapping from being *affine* (i.e., equivalent to a shifting operation plus a linear mapping) to *linear* by removing the shifting operation.

While the aforementioned approaches still require some weak supervision (i.e., seed dictionaries), there have also been some successful attempts to train CLWEs in a completely unsupervised fashion. For instance, Lample et al. (2018b) proposed a system called MUSE, which bootstraps CLWEs without any bilingual

signal through adversarial learning. VECMAP (Artetxe et al., 2018) applied a self-learning strategy to iteratively compute the optimal mapping and then retrieve bilingual dictionary. Comparing MUSE and VECMAP, the latter tends to be more robust as its similarity-matrix-based heuristic initialisation is more stable in most cases (Glavaš et al., 2019; Ruder et al., 2019). Very recently, some studies bootstrapped unsupervised CLWEs by jointly training word embeddings on concatenated corpora of different languages and achieved good performance (Wang et al., 2020).

2.1.2 Debates on Linearity of CLWE Mappings

Since Mikolov et al. (2013b) discovered that the vectors of word translations exhibit similar structures across different languages, researchers made use of this by assuming that mappings between multilingual embeddings could be modelled using simple linear transformations. Although models based on this assumption have demonstrated strong performance, it has recently been questioned. Researchers have claimed that the structure of multilingual word embeddings may not always be similar. As follows we list recent works that have cast doubt on this linearity assumption and further led researchers to experiment with the use of non-linear mappings. Nakashole and Flauger (2018) and Wang et al. (2021a) pointed out that structural similarities may only hold across particular regions of the embedding spaces rather than over their entirety. Søgaard et al. (2018) examined word vectors trained using different corpora, models and hyper-parameters, and concluded configuration dissimilarity between the monolingual embeddings breaks the assumption that the mapping between them is linear. Patra et al. (2019) investigated various language pairs and discovered that a higher etymological distance is associated with degraded the linearity of CLWE mappings. Vulić et al. (2020b) additionally argued that factors such as limited monolingual resources may also weaken the linearity assumption.

These findings motivated work on designing non-linear mapping functions in an effort to improve CLWE performance. For example, Nakashole (2018) and Wang et al. (2021a) relaxed the linearity assumption by combining multiple linear CLWE mappings; Patra et al. (2019) developed a semi-supervised model that

loosened the linearity restriction; Lubin et al. (2019) attempted to reduce the dissimilarity between multilingual embedding manifolds by refining learnt dictionaries; Glavaš and Vulić (2020) first trained a globally optimal linear mapping, then adjusted vector positions to achieve better accuracy; Mohiuddin et al. (2020) used two independently pre-trained auto-encoders to introduce non-linearity to CLWE mappings; Ganesan et al. (2021) obtained inspirations via the back translation paradigm, hence framing CLWE training as to explicitly solve a non-linear and bijective transformation between multilingual word embeddings. Despite these non-linear mappings outperforming their linear counterparts in many setups, in some settings the linear mappings still seem more successful, e.g., the alignment between Portuguese and English word embeddings in Ganesan et al. (2021). Moreover, training non-linear mappings is typically more complex and thus requires more computational resources.

Albeit at the significant recent attention to this problem by the research community, to the best of our knowledge, there has been no in-depth analysis of the conditions for the linearity assumption. The main cause of this research gap is that the majority of previous CLWE work has focused on empirical findings. In Chapter 3, we make the first attempt to explore this direction by providing both theoretical and empirical contributions.

2.1.3 The ℓ_2 Refinement Algorithm

As aforementioned in § 2.1.1 and § 2.1.2, although there are debates around the suitability, or otherwise, of linear mappings, these so-called projection-based approaches are still the most successful CLWE models, as learning mappings between monolingual word vectors requires very little, or even zero, cross-lingual supervision (Lample et al., 2018b; Artetxe et al., 2018; Glavaš et al., 2019).

Mainstream projection-based CLWE models typically identify orthogonal mappings by minimising the topological dissimilarity between source and target embeddings based on ℓ_2 loss (aka. Frobenius loss or squared error) (Glavaš et al., 2019; Ruder et al., 2019). In practice, CLWE models often apply ℓ_2 refinement, a post-processing step shown to improve the quality of the initial alignment (see Ruder et al. (2019) for the survey). Given existing CLWEs $\{\mathbf{X}_{L_A}, \mathbf{X}_{L_B}\}$ for

languages L_A and L_B , bidirectionally one can use approaches such as the classic nearest-neighbour algorithm, the inverted softmax (Smith et al., 2017) and the cross-domain similarity local scaling (CSLS) (Lample et al., 2018b) to retrieve two bilingual dictionaries $D_{L_A \rightarrow L_B}$ and $D_{L_B \rightarrow L_A}$. Note that word pairs in $D_{L_A \rightarrow L_B} \cap D_{L_B \rightarrow L_A}$ are highly reliable, as they form “mutual translations”. Next, one can compose bilingual embedding matrices \mathbf{A} and \mathbf{B} by aligning word vectors (rows) using the above word pairs. Finally, a new orthogonal mapping is learned to fit \mathbf{A} and \mathbf{B} based on least-square regressions, i.e., the optimisation task posed in Eq. (2.1).

Early applications of ℓ_2 refinement applied a *single* iteration, e.g. Vulić and Korhonen (2016). Due to the wide adoption of the closed-form ℓ_2 OPA solution (cf. Eq. (2.2)), recent methods perform multiple iterations. The iterative ℓ_2 refinement strategy is an important component of approaches that bootstrap from small or null training lexicons (Artetxe et al., 2018). However, a single step of refinement is often sufficient to create suitable CLWEs (Lample et al., 2018b; Glavaš et al., 2019).

This learning strategy has two advantages. Besides the fact that adding the orthogonality constraint to the mapping function has been demonstrated to significantly enhance the quality of CLWEs (Xing et al., 2015), the existence of a closed-form solution to the ℓ_2 optima (Schönemann, 1966) greatly simplifies the computation required (Artetxe et al., 2016; Smith et al., 2017).

Despite the popularity, work in various application domains has noted that ℓ_2 loss is not robust to noise and outliers. It is widely known in computer vision that ℓ_2 -loss-based solutions can severely exaggerate noise, leading to inaccurate estimates (Aanæs et al., 2002; De La Torre and Black, 2003). In data mining, PCA using ℓ_2 loss has been shown to be sensitive to the presence of outliers in the input data, degrading the quality of the feature space produced (Kwak, 2008). Previous studies have demonstrated that the processes used to construct monolingual and cross-lingual embeddings may introduce noise (e.g. via reconstruction error (Allen and Hospedales, 2019) and structural variance (Ruder et al., 2019)), making the presence of outliers more likely. Empirical analysis of CLWEs also demonstrates that more distant word pairs (which are more likely to be outliers) have more influence on the behaviour of ℓ_2 loss than closer pairs. This raises the question of the appropriateness of ℓ_2 loss functions for CLWEs.

2.2 Embedding Real-World Relations

2.2.1 Encoding Analogies

In cognitive science, analogy is considered as a central method of human activities, such as perception (Chalmers et al., 1992), memory (Gentner, 1983), communication (Juthe, 2005), problem solving (Holyoak and Thagard, 1994), and decision making (Holland et al., 1987). Therefore, it has received significant focus from the community of machine intelligence research, such as Computer Vision (Mayer, 2009; Johnson et al., 2017) and Robotics (Kiryazov et al., 2007; Cuperman and Verner, 2019). As for NLP, analogy has been employed as a popular intrinsic tool (Ashley, 1988; Miclet et al., 2008), especially since the observation that it can be represented using word embeddings and vector arithmetic (Mikolov et al., 2013c). A popular example based on the analogy “*king is to man as queen is to woman*” shows that the vectors representing the four terms (\mathbf{x}_{king} , \mathbf{x}_{man} , \mathbf{x}_{queen} and \mathbf{x}_{woman}) exhibit the following relation:

$$\mathbf{x}_{king} - \mathbf{x}_{man} \approx \mathbf{x}_{queen} - \mathbf{x}_{woman}. \quad (2.3)$$

Since this discovery, the task of analogy completion has commonly been employed to evaluate the quality of pre-trained word embeddings (Mikolov et al., 2013c; Pennington et al., 2014; Levy and Goldberg, 2014a). This line of research has directly benefited downstream applications (e.g., representation bias removal (Prade and Richard, 2021)) and other relevant domains (e.g., automatic knowledge graph construction (Wang et al., 2021b)). Theoretical analysis has demonstrated a link between embeddings’ analogy encoding and the Pointwise Mutual Information of the training corpus (Arora et al., 2016; Gittens et al., 2017; Allen and Hospedales, 2019; Ethayarajh et al., 2019; Fournier and Dunbar, 2021). Nonetheless, as far as we are aware, the connection between the preservation of analogy encoding and the linearity of CLWE mappings has not been previously investigated.

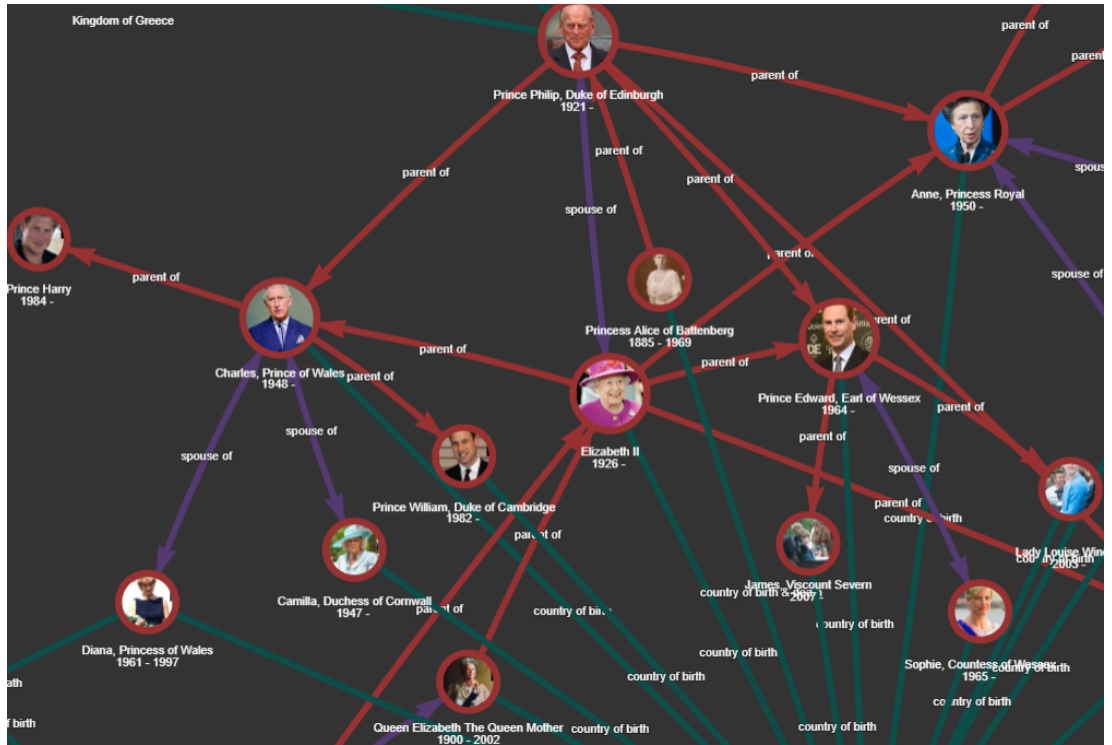


Figure 2.2: A knowledge graph containing the British Royal Family. This screenshot is taken from <https://demo.staple-api.org/> on 8th September, 2022.

2.2.2 Encoding Knowledge Graphs

Knowledge Graphs are core to many NLP tasks and downstream applications, such as E-commerce (Yu et al., 2022a), question answering (Saxena et al., 2020), information management (Li et al., 2020), dialogue agents (He et al., 2017), search engines (Dong et al., 2014) and recommendation systems (Luo et al., 2022). Factual relations stored in a knowledge graph are always in the format of tuples consisting of one head entity, one tail entity (both are nodes in knowledge graphs) and a relation (an edge in knowledge graphs) between them (see an example in Fig. 2.2). KGEs learn representations of relations and entities in a knowledge graph, which are then utilised in downstream tasks like predicting missing relations (Bordes et al., 2013; Sun et al., 2019; Tang et al., 2020).

The application of deep learning has led to a growing body of studies conducted on the matter of training KGEs. Roughly speaking, these KGE methods

fall into two categories: distance-based models and semantic matching models.

The line of researches regarding distance-based models, which measures plausibility of tuples by calculating distance between entities with additive functions, was initialised the KGE technique proposed by Bordes et al. (2013), namely, TransE. After that, a battery of follow-ups have been proposed, including example models like TransH (Wang et al., 2014), TransR (Lin et al., 2015), and TransD (Ji et al., 2015). These algorithms have enhanced ability on modelling complex relations by means of projecting entities into different (more complex) spaces or hyper-planes. More recently, a number of studies attempt to further boost the quality of KGEs through a way of adding orthogonality constraints (Sun et al., 2019; Tang et al., 2020) for maintaining the relation embedding matrix being orthogonal, which is also the paradigm we follow in Chapter 6.

In contrast, semantic matching models measure the plausibility of tuples by computing the similarities between entities with multiplicative functions. Such an similarity function could be realised using, for example, a bilinear function or a neural network. Typical models in this line includes DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), ConvE (Dettmers et al., 2018), TuckER (Balazevic et al., 2019), and QuatE (Zhang et al., 2019).

2.2.3 Improving Efficiency of Relation Encoding

The recent growth in energy requirements for NLP algorithms in general has led to the recognition of the importance of computationally cheap and eco-friendly approaches (Strubell et al., 2019). Regarding the encodings of semantic relations (e.g., word analogies (Ushio et al., 2021)), the increase in computational requirements can, to a large extent, be attributed to the popularity of massive pre-trained language models (e.g., BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020)) that require significant resources to train. A number of solutions have been proposed such as reducing the number of parameters the model contains (Sanh et al., 2019; Bender et al., 2021; Yao et al., 2022).

As for the encoding of factual relations, e.g., KGE, *all* popular approaches described in § 2.2.2 share the same issue of their low speed in both training and inference phases (see Rossi et al. (2021) for a controlled comparison of the efficiency

across different methodologies). In response to this issue, some state-of-the-art KGE algorithms attempted to accelerate their inference speed either through making use of the high-speed of the convolutional neural networks (Dettmers et al., 2018) or through reducing the scale of parameters of the model (Zhang et al., 2019; Zhu et al., 2020b).

In terms of the acceleration of model training, a number of attempts have been conducted in a mostly engineering way. These well-engineered systems adopt linear KGE methods to multi-thread versions in order to make full use of the hardware capacity (Joulin et al., 2017b; Han et al., 2018), which accelerates training time of, for example, TransE, from more than an hour to only a couple of minutes. Nonetheless, this line of work has two major issues. One is that training models faster in this way does not necessarily mean they also emit less, as process scheduling of a multi-thread system can be energy-consuming. The other is that they are all extensions of KGE models based on linear classifiers only (also noting that these models are naturally much faster than other KGE models) without any algorithmic contribution, which leading to the performance of the resulting models limited by the upper bound of models based on linear classifiers (e.g., recent state-of-the-art methods in Tab. 6.2, such as RotH, all belong to other families of KGE approaches).

2.3 Summarising Historical Text

The process of text summarisation is fundamental to research into history, archaeology, and digital humanities (South, 1977). Researchers can better gather and organise information and share knowledge by first identifying the key points in historical documents. However, this can cost a lot of time and effort. On one hand, due to cultural and linguistic variations over time, interpreting historical text can be a challenging and energy-consuming process, even for those with specialist training Gray et al. (2011). To compound this, historical archives can contain narrative documents on a large scale, adding to the workload of manually locating important elements (Gunn, 2011). To reduce these burdens, specialised software has been developed recently, such as MARKUS (Ho and Weerdt, 2014) and DocuSky (Tu et al., 2020). These toolkits aid users in managing and anno-

tating documents but still lack functionalities to automatically process texts at a semantic level. As for the advances in terms of new algorithms, please check § 2.3.1.

Historical text summarisation can be regarded as a special case of cross-lingual summarisation (Leuski et al., 2003; Orăsan and Chiorean, 2008; Cao et al., 2020), a long-standing research topic whereby summaries are generated in a target language from documents in different source languages (more detailed introductions can be found in § 2.3.2). However, historical text summarisation posits some unique challenges. Cross-lingual (i.e., across historical and modern forms of a language) corpora are rather limited (Gray et al., 2011) and therefore historical texts cannot be handled by traditional cross-lingual summarisers, which require cross-lingual supervision or at least large summarisation datasets in both languages (Cao et al., 2020). Further, language use evolves over time, including vocabulary and word spellings and meanings (Gunn, 2011), and historical collections can span hundreds of years. Writing styles also change over time. For instance, while it is common for today’s news stories to present important information in the first few sentences, a pattern exploited by modern news summarisers (See et al., 2017), this was not the norm in older times (White, 1998).

2.3.1 Text Processing for Historical Languages

Early NLP studies for historical documents focus on spelling normalisation (Piotrowski, 2012), machine translation (Oravecz et al., 2010), and sequence labelling applications, e.g., part-of-speech tagging (Rayson et al., 2007) and named entity recognition (Sydow et al., 2011). Since the rise of neural networks, a broader spectrum of applications such as sentiment analysis (Hamilton et al., 2016), information retrieval (Pettersson et al., 2016), and relation extraction (Opitz et al., 2018) have been developed.

Chapter 5 adds to this growing literature in two ways. First, much of the work on historical text processing is focused on English_{|EN}, and work in other languages is still relatively unexplored (Piotrowski, 2012; Rubinstein, 2019). Second, the task of historical text summarisation has never been tackled before, to the best of our knowledge. A lack of non-EN annotated historical resources is a key reason

for the former, and for the latter, resources do not exist in any language. We hope to spur research on historical text summarisation and in particular for non-EN languages through this work.

2.3.2 Cross-Lingual Summarisation and Beyond

The traditional strands of cross-lingual text summarisation systems design pipelines which learn to translate and summarise separately (Leuski et al., 2003; Orăsan and Chiorean, 2008). However, such paradigms suffer from the error propagation problem, i.e., errors produced by upstream modules may accumulate and degrade the output quality (Zhu et al., 2020a). In addition, parallel data to train effective translators is not always accessible (Cao et al., 2020). Recently, end-to-end methods have been applied to alleviate this issue. The main challenge for this research direction is the lack of direct corpora, leading to attempts such as zero-shot learning (Duan et al., 2019), multi-task learning (Zhu et al., 2019), and transfer learning (Cao et al., 2020). Although training requirements have been relaxed by these methods, our extreme setup with summarisation data only available for the target language and very limited parallel data, has never been visited before.

UNDERSTANDING LINEARITY OF CLWE MAPPINGS

The technique of CLWE plays a fundamental role in tackling NLP challenges for low-resource languages. Its dominant approaches assumed that the relationship between embeddings could be represented by a linear mapping, but there has been no exploration of the conditions under which this assumption holds. Such a research gap becomes very critical recently, as it has been evidenced that relaxing mappings to be non-linear can lead to better performance in some cases. We, for the first time, present a theoretical analysis that identifies the preservation of analogies encoded in monolingual word embeddings as a *necessary and sufficient* condition for the ground-truth CLWE mapping between those embeddings to be linear. On a novel cross-lingual analogy dataset that covers five representative analogy categories for twelve distinct languages, we carry out experiments which provide direct empirical support for our theoretical claim. These results offer additional insight into the observations of other researchers and contribute inspiration for the development of more effective cross-lingual representation learning strategies. The resources for this chapter are available at <https://github.com/Pzoom522/xANLG>.

3.1 Theoretical Basis

We denote a ground-truth CLWE mapping as $\mathcal{M} : \mathbf{X} \rightarrow \mathbf{Y}$, where \mathbf{X} and \mathbf{Y} are monolingual word embeddings independently trained for languages L_X and L_Y , respectively.

Proposition. Encoded analogies are preserved during the CLWE mapping $\mathcal{M} \iff \mathcal{M}$ is affine.

Remarks. Following Eq. (2.3), the preservation of analogy encoding under a mapping can be formalised as

$$\mathbf{x}_\alpha - \mathbf{x}_\beta = \mathbf{x}_\gamma - \mathbf{x}_\theta \implies \mathcal{M}(\mathbf{x}_\alpha) - \mathcal{M}(\mathbf{x}_\beta) = \mathcal{M}(\mathbf{x}_\gamma) - \mathcal{M}(\mathbf{x}_\theta), \quad (3.1)$$

where $\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma, \mathbf{x}_\theta \in \mathbf{X}$.

If \mathcal{M} is affine, for d -dimensional monolingual embeddings X we have

$$\mathcal{M}(\mathbf{x}) := M\mathbf{x} + \mathbf{b}, \quad (3.2)$$

where $x \in X$, $M \in \mathbb{R}^{d \times d}$, and $\mathbf{b} \in \mathbb{R}^{d \times 1}$.

Proof for Eq. (3.1) \implies Eq. (3.2). To begin with, by adopting the ‘‘mean centring’’ operation in § 2.1.1, we shift the coordinates of the space of \mathbf{X} , ensuring

$$\mathcal{M}(\vec{0}) = \vec{0}. \quad (3.3)$$

This step greatly simplifies the derivations afterwards, because from now on we just need to demonstrate that \mathcal{M} is a *linear mapping*, i.e., it can be written as $M\mathbf{x}$. By definition, this is equivalent to showing that \mathcal{M} preserves both the operations of addition (aka. additivity) and scalar multiplication (aka. homogeneity).

Additivity can be proved by observing that $(\mathbf{x}_i + \mathbf{x}_j) - \mathbf{x}_j = \mathbf{x}_i - \vec{0}$ and therefore,

$$\begin{aligned} (\mathbf{x}_i + \mathbf{x}_j) - \mathbf{x}_j = \mathbf{x}_i - \vec{0} &\xrightarrow{\text{Eq. (3.1)}} \mathcal{M}(\mathbf{x}_i + \mathbf{x}_j) - \mathcal{M}(\mathbf{x}_j) = \mathcal{M}(\mathbf{x}_i) - \mathcal{M}(\vec{0}) \\ &\xrightarrow{\text{Eq. (3.3)}} \mathcal{M}(\mathbf{x}_i + \mathbf{x}_j) = \mathcal{M}(\mathbf{x}_i) + \mathcal{M}(\mathbf{x}_j). \end{aligned} \quad (3.4)$$

Homogeneity can be proved in four steps.

- **Step 1:** Observe that $\vec{0} - \mathbf{x}_i = -\mathbf{x}_i - \vec{0}$, similar to Eq. (3.4) we can show that

$$\begin{aligned} \vec{0} - \mathbf{x}_i = -\mathbf{x}_i - \vec{0} &\xrightarrow{\text{Eq. (3.1)}} \mathcal{M}(\vec{0}) - \mathcal{M}(\mathbf{x}_i) = \mathcal{M}(-\mathbf{x}_i) - \mathcal{M}(\vec{0}) \\ &\xrightarrow[\times(-1)]{\text{Eq. (3.3)}} \mathcal{M}(\mathbf{x}_i) = -\mathcal{M}(-\mathbf{x}_i). \end{aligned} \quad (3.5)$$

- **Step 2:** Using *mathematical induction*, for arbitrary \mathbf{x}_i , we show that

$$\forall m \in \mathbb{N}^+, \mathcal{M}(m\mathbf{x}_i) = m\mathcal{M}(\mathbf{x}_i) \quad (3.6)$$

holds, where \mathbb{N}^+ is the set of positive natural numbers, as

Base Case: Trivially holds when $m = 1$.

Inductive Step: Assume the inductive hypothesis that $m = k$ ($k \in \mathbb{N}^+$), i.e.,

$$\mathcal{M}(k\mathbf{x}_i) = k\mathcal{M}(\mathbf{x}_i). \quad (3.7)$$

Then, as required, when $m = k + 1$,

$$\begin{aligned} \mathcal{M}((k+1)\mathbf{x}_i) &\xrightarrow{\text{Eq. (3.4)}} \mathcal{M}(k\mathbf{x}_i) + \mathcal{M}(\mathbf{x}_i) \\ &\xrightarrow{\text{Eq. (3.7)}} k\mathcal{M}(\mathbf{x}_i) + \mathcal{M}(\mathbf{x}_i) = (k+1)\mathcal{M}(\mathbf{x}_i). \end{aligned}$$

- **Step 3:** We further justify that

$$\forall n \in \mathbb{N}^+, \mathcal{M}\left(\frac{\mathbf{x}_i}{n}\right) = \frac{\mathcal{M}(\mathbf{x}_i)}{n}, \quad (3.8)$$

as

$$\begin{aligned} \mathcal{M}\left(\frac{\mathbf{x}_i}{n}\right) &= \mathcal{M}\left(\mathbf{x}_i + \left(-\frac{n-1}{n}\mathbf{x}_i\right)\right) \stackrel{\text{Eq. (3.4)}}{=} \mathcal{M}(\mathbf{x}_i) + \mathcal{M}\left(-\frac{n-1}{n}\mathbf{x}_i\right) \\ &\stackrel{\text{Eq. (3.5)}}{=} \mathcal{M}(\mathbf{x}_i) - \mathcal{M}\left(\frac{n-1}{n}\mathbf{x}_i\right) \\ &\stackrel{\text{Eq. (3.6)}}{=} \mathcal{M}(\mathbf{x}_i) - (n-1)\mathcal{M}\left(\frac{\mathbf{x}_i}{n}\right) \end{aligned}$$

directly yields $\mathcal{M}\left(\frac{\mathbf{x}_i}{n}\right) = \frac{\mathcal{M}(\mathbf{x}_i)}{n}$, i.e., Eq. (3.8).

• **Step 4:** Considering the set of rational numbers $\mathbb{Q} = \{0\} \cup \{\pm\frac{m}{n} | \forall m, n\}$, Eqs. (3.3), (3.5), (3.6) and (3.8) jointly justifies the homogeneity of \mathcal{M} for \mathbb{Q} . Because $\mathbb{Q} \subset \mathbb{R}$ is a *dense set*, homogeneity of \mathcal{M} also holds over \mathbb{R} , see Kleiber and Pervin (1969).

Finally, combined with the additivity that has been already justified above, linearity of CLWE mapping \mathcal{M} is proved, i.e., Eq. (3.1) \implies Eq. (3.2). \square

Proof for Eq. (3.2) \implies Eq. (3.1). Justifying this direction is quite straightforward:

$$\begin{aligned} \mathbf{x}_\alpha - \mathbf{x}_\beta = \mathbf{x}_\gamma - \mathbf{x}_\theta &\implies M\mathbf{x}_\alpha - M\mathbf{x}_\beta = M\mathbf{x}_\gamma - M\mathbf{x}_\theta \\ &\implies M\mathbf{x}_\alpha + \mathbf{b} - (M\mathbf{x}_\beta + \mathbf{b}) = M\mathbf{x}_\gamma + \mathbf{b} - (M\mathbf{x}_\theta + \mathbf{b}) \\ &\implies \mathcal{M}(\mathbf{x}_\alpha) - \mathcal{M}(\mathbf{x}_\beta) = \mathcal{M}(\mathbf{x}_\gamma) - \mathcal{M}(\mathbf{x}_\theta). \quad \square \end{aligned}$$

Summarising the proofs for both the forward and reverse directions, we conclude that the proposition holds.

Please note, the high-level assumption of our derivations is that word embedding spaces can be treated as continuous vector spaces, an assumption commonly adopted in previous work, e.g., Levy and Goldberg (2014b), Hashimoto et al. (2016), Zhang et al. (2018), and Ravfogel et al. (2020). Nevertheless, we argue that the inherent discreteness of word embeddings should not be ignored, e.g., an interpolation between two word vectors may not always correspond to an existing word from the vocabulary (Li et al., 2021c). The following sections complement this theoretical insight via experiments which confirm the claim holds empirically.

3.2 Experiment

Our experimental protocol assesses the linearity of the mapping between each pair of pre-trained monolingual word embeddings. We also quantify the extent to which this mapping preserves encoded analogies, i.e., satisfies the condition of Eq. (3.1). We then analyse the correlation between these two indicators. A strong correlation provides evidence to support our theory, and *vice versa*. The indicators used are described in § 3.2.1. Unfortunately, there are no suitable publicly available corpora for our proposed experiments, so we develop a novel word-level analogy test set that is fully parallel across languages, namely xANLG (see § 3.2.2). The pre-trained embeddings used for the tests are described in § 3.2.3.

3.2.1 Indicators

Linearity of CLWE Mapping

Direct measurement of the linearity of a ground-truth CLWE mapping is challenging. One relevant approach is to benchmark the similarity between multilingual word embedding, where the mainstream and state-of-the-art indicators are the so-called spectral-based algorithms (Søgaard et al., 2018; Dubossarsky et al., 2020). However, such methods assume the number of tested vectors to be much larger than the number of dimensions, which does not apply in our scenario (see § 3.2.2). Therefore, we choose to evaluate linearity via the goodness-of-fit of the optimal orthogonal CLWE mapping (cf. Smith et al. (2017)), which is measured as

$$\mathcal{S}_{\text{LMP}} := -\|M^*X - Y\|_F/r \quad \text{with} \quad M^* = \arg \min_M \|MX - Y\|_F \quad (3.9)$$

where $\|\cdot\|_F$ and r denotes the Frobenius norm and the number of X 's rows. To obtain matrices X and Y , from \mathbf{X} and \mathbf{Y} respectively, we first retrieve the vectors corresponding to lexicons of a ground-truth L_X - L_Y dictionary and concatenate them into two matrices. More specifically, if two vectors (represented as rows) share the same index in the two matrices (one for each language), their corresponding words form a translation pair, i.e., the rows of these matrices are

aligned. “Mean centring” is applied to satisfy Eq. (3.3). For fair comparisons across different mapping pairs, in each of X and Y , rows are standardised by scaling the mean Euclidean norm to 1.

Large absolute values of \mathcal{S}_{LMP} mean that the optimal linear mapping is an accurate model of the true relationship between the embeddings, and *vice versa*. \mathcal{S}_{LMP} therefore indicates the degree to which CLWE mappings are linear.

Preservation of Analogy Encoding

To assess how well analogies are preserved across embeddings, we start by probing how analogies are encoded in the monolingual word embeddings. We use the set-based LRCos, the state-of-the-art analogy mining tool for static word embeddings (Drozd et al., 2016).¹ It provides a score in the range of 0 to 1, indicating the correctness of analogy completion in a single language. For the extension in a cross-lingual setup, we further compute the geometric mean:

$$\mathcal{S}_{\text{PAE}} := \sqrt{\text{LRCos}(\mathbf{X}) \times \text{LRCos}(\mathbf{Y})} \quad (3.10)$$

where $\text{LRCos}(\cdot)$ is the accuracy of analogy completion provided by LRCos for embedding \mathbf{X} . To simplify our discussion and analysis from now onward, when performing CLWE mappings, by default we select the monolingual embeddings that best encode analogy, i.e., we restrict $\text{LRCos}(\mathbf{X}) \geq \text{LRCos}(\mathbf{Y})$. $\mathcal{S}_{\text{PAE}} = 1$ indicates all analogies are well encoded in both embeddings, and are preserved by the ground-truth mapping between them. On the other hand, lower \mathcal{S}_{PAE} values indicate deviation from the condition of Eq. (3.1).

Validity of \mathcal{S}_{PAE}

As an aide, we explore the properties of the \mathcal{S}_{PAE} indicator to demonstrate its robustness for the interested reader. The score produced by LRCos is relative to a pre-specified set of *known* analogies. In theory, a low $\text{LRCos}(\mathbf{X})$ score may

¹We have tried alternatives including 3CosAdd (Mikolov et al., 2013a), PairDistance (Levy and Goldberg, 2014a) and 3CosMul (Levy et al., 2015), verifying that they are less accurate than LRCos in most cases. Still, in the experiments they all exhibit similar trends as shown in Tab. 3.2.

not reliably indicate that \mathbf{X} does not encode analogies well since there may be other word pairings within that set that produce higher scores. This naturally raises a question: *does \mathcal{S}_{PAE} really promise the validity as the indicator of analogy encoding preservation?* In other words, it is necessary to investigate whether there exists an *unknown* analogy word pairing encoded by the tested embeddings to an equal or higher degree. If there is, then \mathcal{S}_{PAE} may not reflect the preservation of analogy encoding completely, as unmatched analogy test sets may lead to low LRCos scores even for monolingual embeddings that encode analogies well. We demonstrate that the problem can be considered as an optimal transportation task and \mathcal{S}_{PAE} is guaranteed to be a reliable indicator.

As analysed by Ethayarajh et al. (2019), the degree to which word pairs are encoded as analogies in word embeddings is equivalent to the likelihood that the end points of any two corresponding vector pairs form a high-dimensional coplanar parallelogram. More formally, this task is to identify

$$\mathbf{P}^* = \arg \min_{\mathbf{P}} \sum_{\mathbf{x} \in \mathbf{X}} \mathcal{C}(\mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x})), \quad (3.11)$$

where \mathbf{P} is one possible pairing of vectors in \mathbf{X} and $\mathcal{C}(\cdot)$ is the cost of a given transportation scheme. $\mathcal{T}_{\square}^{\mathbf{P}}(\cdot)$ denotes the corresponding cost-optimal process of moving vectors to satisfy

$$\begin{aligned} \forall \{(\mathbf{x}_{\alpha}, \mathbf{x}_{\beta}), (\mathbf{x}_{\gamma}, \mathbf{x}_{\theta})\} \subseteq \mathbf{P}, \\ \mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\alpha}) - \mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\beta}) = \mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\gamma}) - \mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\theta}), \end{aligned} \quad (3.12)$$

i.e., the end points of $\mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\alpha})$, $\mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\beta})$, $\mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\gamma})$ and $\mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_{\theta})$ form a parallelogram.

Therefore, in each language and analogy category of xANLG, we first randomly sample vector pairing samples, leading to $1e5$ different \mathbf{P} (more samples will make the downstream computation overhead unbearable). Next, for each of them, we need to obtain $\mathcal{T}_{\square}^{\mathbf{P}}(\cdot)$ that minimises $\sum_{\mathbf{x} \in \mathbf{X}} \mathcal{C}(\mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}))$ in Eq. (3.11). Our algorithm is explained using the example in Fig. 3.1, where the cardinality of \mathbf{X} and \mathbf{P} is 8 and 4, respectively.

- **Step 1:** Link the end points of the vectors within each word pair, hence

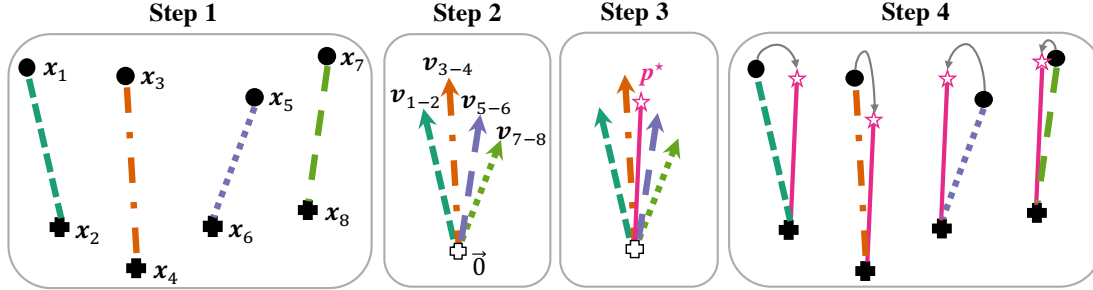


Figure 3.1: An example of solving $\mathcal{T}_{\square}^{\mathbf{P}}(\cdot)$ in Eq. (3.12), with $\mathbf{P} = \{(\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_3, \mathbf{x}_4), (\mathbf{x}_5, \mathbf{x}_6), (\mathbf{x}_7, \mathbf{x}_8)\}$. In the figure we adjust the position of \mathbf{x}_1 , \mathbf{x}_3 , \mathbf{x}_5 and \mathbf{x}_7 in the last step, but it is worth noting that there also exists other feasible $\mathcal{T}_{\square}^{\mathbf{P}}(\cdot)$ given \mathbf{p}^* , e.g., to tune \mathbf{x}_2 , \mathbf{x}_4 , \mathbf{x}_6 and \mathbf{x}_8 instead.

our target is to adjust these end points so that all connecting lines not only have equal length but also remain parallel.

- **Step 2:** For each vector pair $(\mathbf{x}_\alpha, \mathbf{x}_\beta) \in \mathbf{P}$, vectorise its connecting line into an offset vector as $\mathbf{v}_{\alpha-\beta} = \mathbf{x}_\alpha - \mathbf{x}_\beta$.
- **Step 3:** As the start points of all such offset vectors are aggregated at $\vec{0}$, seek a vector \mathbf{p}^* that minimises the total transportation cost between the end point of \mathbf{p}^* and those of all offset vectors (again, note they share a start point at $\vec{0}$).
- **Step 4:** Perform the transportation so that all offset vectors become \mathbf{p}^* , i.e.,

$$\forall (\mathbf{x}_\alpha, \mathbf{x}_\beta) \in \mathbf{P}, \mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_\alpha) - \mathcal{T}_{\square}^{\mathbf{P}}(\mathbf{x}_\beta) = \mathbf{p}^*.$$

In this way, the tuned vector pairs can always form perfect parallelograms. Obviously, as \mathbf{p}^* is at the cost-optimal position (see Step 3), this vector-adjustment scheme is also cost-optimal.

Solving \mathbf{p}^* for high dimensions is non-trivial in real world and is a special case of the NP-hard Facility Location Problem (aka. the P-Median Problem) (Kariv and Hakimi, 1979). We, therefore, use the `scipy.optimize.fmin` implementation of the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) to provide a good-enough solution. To reach convergence, with the mean offset vector as the

initial guess, we set both the absolute errors in parameter and function value between iterations at $1e4$. We experimented with implementing $\mathcal{C}(\cdot)$ using mean Euclidean, Taxicab and Cosine distances respectively. For all analogy categories in all languages, \mathbf{P}^* coincides perfectly with the pre-defined pairing of xANLG. This analysis provides evidence that the situation where *an unknown kind of analogy is better encoded than the ones used* does not occur in practice. \mathcal{S}_{PAE} is thus trustworthy.

3.2.2 Datasets

Calculating the correlation between \mathcal{S}_{LMP} and \mathcal{S}_{PAE} requires a cross-lingual word analogy dataset. This resource would allow us to simultaneously (1) construct two aligned matrices X and Y to check the linearity of CLWE mappings, and (2) obtain the monolingual LRCos scores of both \mathbf{X} and \mathbf{Y} . Three relevant resources were identified, although none of them is suitable for our study.

- Brychcín et al. (2019) described a cross-lingual analogy dataset consisting of word pairs from six closely related European languages, but it has never been made publicly available.
- Ulčar et al. (2020) open-sourced the MCIWAD dataset for nine languages, but the analogy words in different languages are not parallel.²
- Garneau et al. (2021) produced the cross-lingual WiQueen dataset. Unfortunately, a large part of its entries are proper nouns or multi-word terms instead of single-item words, leading to low coverage on the vocabularies of embeddings.

Consequently, we develop xANLG, which we believe to be the first (publicly available) cross-lingual word analogy corpus. For consistency with previous work, xANLG is bootstrapped using established monolingual analogies and cross-lingual dictionaries. xANLG is constructed by starting with a *bilingual* analogy dataset, say, that for L_X and L_Y . Within each analogy category, we first translate word pairs of the L_X analogy corpus into L_Y , using an available L_X - L_Y

²Personal communication with the authors.

dictionary. Next, we check if any translation coincides with its original word pair in L_Y . If it does, such a word pair (in both L_X and L_Y) will be added into the bilingual dataset. This process is repeated for multiple languages to form a cross-lingual corpus.

We use the popular MUSE dictionary (Lample et al., 2018a) which contains a wide range of language pairs. Two existing collections of analogies are utilised:

- **Google Analogy Test Set (GATS)** (Mikolov et al., 2013c), the *de facto* standard benchmark of embedding-based analogy solving. We adopt its extended English version, Bigger Analogy Test Set (BATS) (Gladkova et al., 2016), supplemented with several datasets in other languages inspired by the original GATS: French, Hindi and Polish (Grave et al., 2018), German (Köper et al., 2015) and Spanish (Cardellino, 2019).
- The aforementioned **Multilingual Culture-Independent Word Analogy Datasets (MCIWAD)** (Ulčar et al., 2020).

Due to the differing characteristics of these datasets (e.g., the composition of analogy categories), they are used to produce two separate corpora: $xANLG_G$ and $xANLG_M$. Only categories containing at least 30 word pairs aligned across all languages in the dataset were included. For comparison, 60% of the semantic analogy categories in the commonly used GATS dataset contains fewer than 30 word pairs. The rationale for selecting this value was that it allows a reasonable number of analogy completion questions to be generated, as 30 word pairs can be used to generate as many as 3480 unique analogy completion questions such as “*king:man :: queen:?*”.³ Information in $xANLG_G$ and $xANLG_M$ for the **capital-country** of Hindi was supplemented with manual translations by native speakers. In addition, each analogy included in the data set was checked by at least one fluent speaker of the relevant language to ensure that they are valid.

³For an analogy category with t word pairs, $\binom{t}{2}$ four-item elements can be composed. An arbitrary element, $\alpha:\beta :: \gamma:\theta$, can yield eight analogy completion questions as follows:

$$\begin{array}{cccccc} \alpha:\beta :: \gamma:? & \beta:\alpha :: \theta:? & \gamma:\alpha :: \theta:? & \theta:\beta :: \gamma:? \\ \alpha:\gamma :: \beta:? & \beta:\theta :: \alpha:? & \gamma:\theta :: \alpha:? & \theta:\gamma :: \beta:? \end{array}$$

Hence, $\binom{t}{2} \times 8$ unique questions can be generated.














Category	#	 DE	 EN	 ES	 FR	 HI	 PL	
CAP [†]	31	Budapest Ungarn	Budapest Hungary	Budapest Hungria	Budapest Hongrie	बुडापेस्ट हंगरी	Budapeszt Węgry	
GNDR [†]	30	sohn tochter	son daughter	hijo hija	fil fille	बेटा बेटी	syn córka	
NATL [†]	34	Peru Peruanisch	Peru Peruvian	Perú Peruano	Pérou Péruvien	पेरू पेरू	Peru Peruwiański	
G-PL [‡]	31	kind kinder	child children	niño niños	enfant enfants	बच्चा बच्चे	dziecko dzieci	
Category	#	 EN	 ET	 FI	 HR	 LV	 RU	 SL
ANIM [†]	32	eagle bird	kotkas lind	kotka lintu	orao ptica	ērglis putns	орёл птица	orel ptica
G-PL [‡]	31	machine machines	masin masinad	kone koneet	stroj strojevi	mašīna mašīnas	машина машины	stroj stroji

Table 3.1: Summary of and examples from the XANLG corpus. # denotes the number of cross-lingual analogy word pairs in each language. [†]Semantic: animal-species|ANIM, capital-world|CAP, male-female|GNDR, nation-nationality|NATL. [‡]Syntactic: grammar-plural|G-PL.

The xANLG dataset contains five distinct analogy categories, including both syntactic (morphological) and semantic analogies, and twelve languages from a diverse range of families (see Tab. 3.1). From Indo-European languages, one belongs to the Indo-Aryan branch (Hindi_{|HI}), one to the Baltic branch (Latvian_{|LV}), two to the Germanic branch (English_{|EN}, German_{|DE}), two to the Romance branch (French_{|FR}, Spanish_{|ES}) and four to the Slavonic branch (Croatian_{|HR}, Polish_{|PL}, Russian_{|RU}, Slovene_{|SL}). Two non-Indo-European languages, Estonian_{|ET} and Finnish_{|FI}, both from the Finnic branch of the Uralic family, are also included. In total, they form 15 and 21 language pairs for xANLG_G and xANLG_M, respectively. These pairs span multiple etymological combinations, i.e., intra-language-branch (e.g., ES-FR), inter-language-branch (e.g., DE-RU) and inter-language-family (e.g., HI-ET).

3.2.3 Word Embeddings

To cover the language pairs used in xANLG, we make use of static word embeddings pre-trained on the twelve languages used in the resource. These embeddings consist of three representative open-source series that employ different training corpora, are based on different embedding algorithms, and have different vector dimensions.

- **Wiki**⁴: 300-dimensional, trained on Wikipedia using the Skip-Gram version of FastText (refer to Joulin et al. (2017a) for details).
- **Crawl**⁵: 300-dimensional, trained on CommonCrawl plus Wikipedia using FastText-CBOW.
- **CoNLL**⁶: 100-dimensional, trained on the CoNLL corpus (without lemmatisation) using Word2Vec (Mikolov et al., 2013c).

3.3 Result

Both Spearman’s rank-order (ρ) and Pearson product-moment (r) correlation coefficients are computed to measure the correlation between \mathcal{S}_{LMP} and \mathcal{S}_{PAE} .

⁴<https://fasttext.cc/docs/en/pretrained-vectors.html>

⁵<https://fasttext.cc/docs/en/crawl-vectors.html>

⁶<http://vectors.nlpl.eu/repository/>

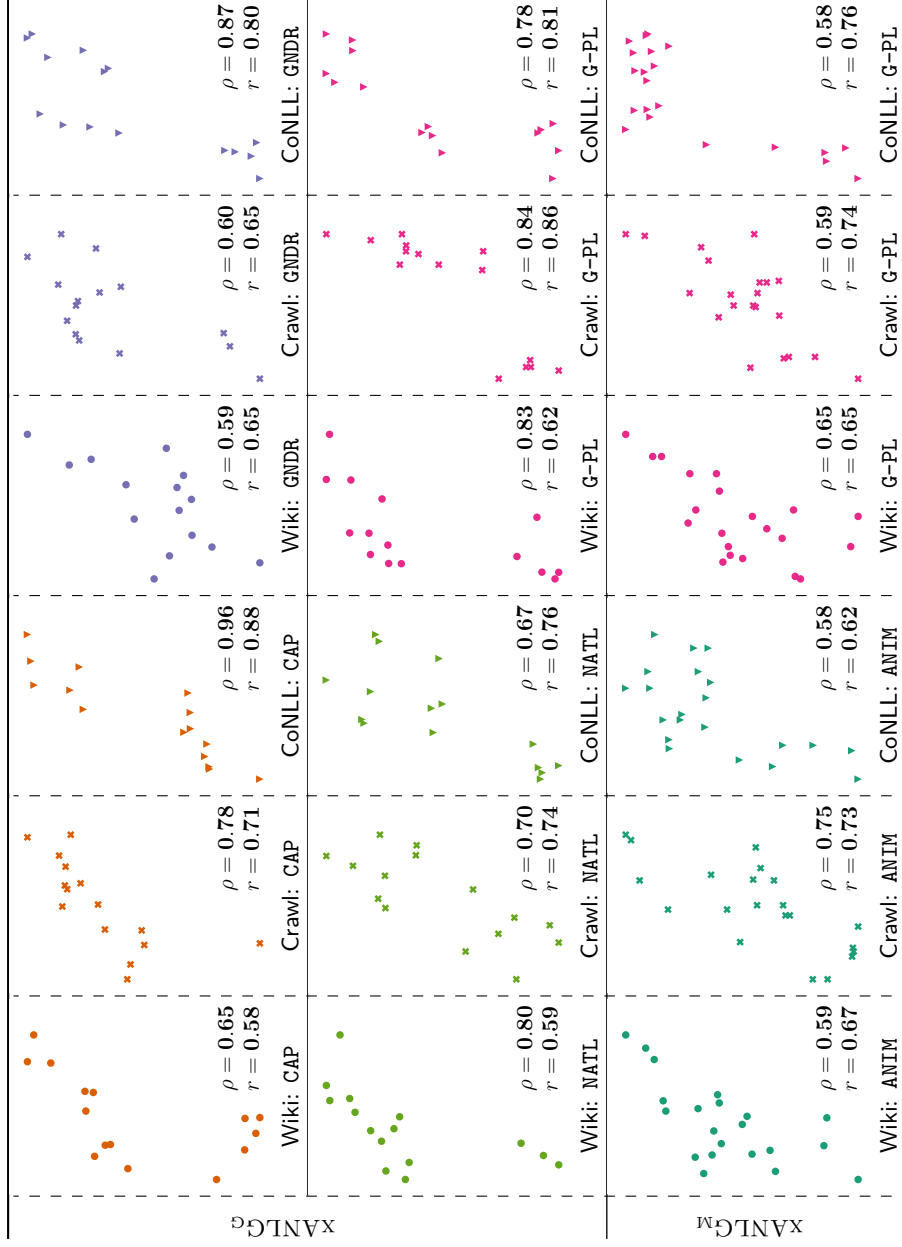


Table 3.2: Correlation coefficients (Spearman’s ρ and Pearson’s r) between S_{LMP} and S_{PAE} . For all groups, we conduct significance tests to estimate the p -value. Empirically, the p -value is always less than $1e-2$ (in most groups it is even less than $1e-3$), indicating a very high confidence level for the experiment results. To facilitate future research and analyses, we present the raw S_{LMP} and LRCos data in Tabs. 3.3 and 3.4, respectively.

XANL _G		EN-DE	EN-ES	EN-FR	EN-HI	EN-PL	DE-ES	DE-FR	DE-HI	DE-PL	ES-FR	ES-HI	ES-PL	FR-HI	FR-PL	HI-PL							
Wiki	CAP	.16	.21	.17	.36	.23	.21	.18	.36	.22	.22	.35	.25	.35	.23	.33							
	GNDR	.32	.42	.39	.26	.35	.48	.40	.41	.36	.39	.43	.38	.30	.40	.42							
	NATL	.18	.16	.15	.14	.20	.19	.19	.33	.21	.16	.30	.21	.14	.20	.32							
Crawl	G-PL	.22	.23	.22	.36	.26	.25	.23	.35	.26	.25	.38	.27	.37	.26	.38							
	CAP	.23	.23	.20	.23	.29	.26	.23	.24	.28	.23	.26	.28	.24	.29	.38							
	GNDR	.57	.58	.59	.56	.54	.65	.66	.57	.59	.64	.56	.57	.56	.57	.58							
CoNLL	NATL	.32	.43	.27	.39	.29	.32	.35	.47	.35	.40	.43	.31	.46	.31	.42							
	G-PL	.35	.24	.33	.48	.29	.33	.37	.44	.42	.33	.47	.33	.48	.42	.51							
	CAP	.31	.58	.32	.55	.39	.58	.32	.56	.38	.59	.66	.59	.56	.40	.55							
XANL _{G^M}	GNDR	.48	.76	.49	.55	.48	.74	.55	.57	.50	.77	.76	.72	.59	.52	.58							
	NATL	.37	.72	.26	.51	.38	.78	.34	.52	.36	.74	.74	.73	.50	.35	.50							
	G-PL	.32	.67	.32	.48	.36	.65	.34	.47	.36	.68	.67	.65	.50	.38	.49							
Wiki	EN	EN	EN	EN	EN	EN	ET	ET	ET	ET	FI	FI	FI	HR	HR	HR							
	ET	FI	HR	LV	RU	SL	FI	HR	LV	RU	SL	HR	LV	RU	SL	SL							
	ANIM	.50	.50	.22	.31	.19	.15	.56	.27	.37	.30	.35	.29	.41	.30	.40	.32	.36	.28	.31	.22	.20	
Crawl	G-PL	.25	.22	.37	.37	.28	.33	.24	.31	.29	.28	.26	.30	.29	.26	.27	.33	.32	.30	.33	.33	.28	.28
	ANIM	.55	.55	.55	.49	.55	.51	.34	.41	.45	.22	.41	.40	.46	.41	.45	.37	.23	.28	.38	.38	.24	.43
	G-PL	.28	.43	.47	.43	.45	.40	.30	.45	.37	.43	.37	.46	.40	.44	.43	.42	.50	.54	.39	.35	.43	.43
CoNLL	ANIM	.54	.54	.99	.55	.50	.53	.29	.74	.46	.37	.43	.87	.51	.38	.46	.64	.77	.98	.42	.36	.41	.41
	ANIM	.54	.54	.99	.55	.50	.53	.29	.74	.46	.37	.43	.87	.51	.38	.46	.64	.77	.98	.42	.36	.41	.41
	G-PL	.45	.40	.52	.42	.40	.42	.37	.77	.41	.41	.40	.81	.37	.36	.39	.84	.66	.77	.36	.40	.38	.38

Table 3.3: Raw S_{LAMP} results (the negative sign is omitted for brevity).

	Wiki				Crawl				CoNLL			
	CAP	GNDR	NATL	G-PL	CAP	GNDR	NATL	G-PL	CAP	GNDR	NATL	G-PL
DE	.68	.25	.21	.23	.47	.48	.79	.77	.65	.43	.41	.55
EN	.94	.33	.94	.58	.57	.67	.76	.94	.87	.57	.79	.61
ES	.45	.13	.35	.13	.40	.57	.68	.87	.13	.07	.07	.17
FR	.92	.27	.76	.13	.65	.50	.85	.87	.48	.14	.24	.35
HI	.29	.30	.42	.07	.58	.59	.59	.32	.32	.37	.31	.16
PL	.16	.21	.26	.10	.29	.55	.82	.84	.45	.45	.38	.52

	Wiki		Crawl		CoNLL	
	ANIM	G-PL	ANIM	G-PL	ANIM	G-PL
EN	.48	.65	.29	.87	.36	.58
ET	.12	.50	.52	1.00	.21	.48
FI	.06	.65	.48	.87	.42	.54
HR	.17	.20	.50	.68	.07	.11
LV	.19	.10	.39	.84	.27	.23
RU	.36	.40	.61	.87	.42	.55
SL	.42	.23	.39	.81	.12	.39

Table 3.4: Raw monolingual LRCos results (upper: xANLG_G; lower: xANLG_M).

Note that, it is not possible to compute the correlations between all pairs due to (1) the number of dimensions varies across embeddings series, and (2) the source and target embeddings have been pre-processed independently for different mappings. Instead, results are grouped by embedding method and analogy category.

Figures in Tab. 3.2 show that a significant positive correlation between \mathcal{S}_{PAE} and \mathcal{S}_{LMP} is observed for all setups. In terms of the Spearman’s ρ , among the 18 groups, 5 exhibit *very strong* correlation ($\rho \geq 0.80$) (with a maximum at 0.96 for CoNLL embeddings on CAP of xANLG_G), 4 show *strong* correlation ($0.80 > \rho \geq 0.70$), and the others have *moderate* correlation ($0.70 > \rho \geq 0.50$) (with a minimum at 0.58: CoNLL embeddings on ANIM and G-PL of xANLG_M). Interestingly, although we do not assume a linear relationship in § 3.1, large values for the Pearson’s r are obtained in practice. To be exact, 4 groups indicate very strong correlation, 6 have strong correlation, while others retain moderate correlation (the minimum r value is 0.58: Wiki embeddings on CAP and G-PL of xANLG_G). These results provide empirical evidence that supplements our

theoretical analysis (§ 3.1) of the relationship between linearity of mappings and analogy preservation.

In addition, we explored whether the analogy type (i.e., semantic or syntactic) affects the correlation. To bootstrap the analysis, for both kinds of correlation coefficients, we divide the 18 experiment groups into two splits, i.e., 12 semantic ones and 6 syntactic ones. After that, we compute a two-treatment ANOVA (Fisher, 1925). For both Spearman’s ρ and Pearson’s r , the results are not significant at $p < 0.1$. Therefore, we conclude that the connection between CLWE mapping linearity and analogy encoding preservation holds across analogy types. We thus recommend testing \mathcal{S}_{PAE} *before* implementing CLWE alignment as an indicator of whether a linear transformation is a good approximation of the ground-truth CLWE mapping (see § 3.4).

Although there are strong correlations between the measures, they are not perfect. We therefore carried out further investigation into the data points in Tab. 3.2 that do not follow the overall trend. Firstly, we identified that some are associated with “crowded” embedding regions, in which the correct answer to an analogy question is not ranked highest by LRCos but the top candidate is a polysemous term (Rogers et al., 2017). One example is the LRCos score of the CAP analogy for PL’s Wiki embeddings, which was underestimated. If we consider the three highest ranked terms, rather than only the top term, then the overall ρ and r of “Wiki: CAP” (the first cell in Tab. 3.2) will increase sharply to 0.79 and 0.76, respectively.

Secondly, we noticed in certain cases the source and target vectors of a word pair are too close (i.e. the distance between them is near zero). This phenomenon introduces noise to the results of analogy metrics such as LRCos (Linzen, 2016; Bolukbasi et al., 2016), and consequently, impact \mathcal{S}_{PAE} . For example, the mean cosine distance between G-PL pairs is smaller in xANLG_M (0.18) than xANLG_G (0.24). Therefore, the \mathcal{S}_{PAE} for G-PL is less reliable for xANLG_M than xANLG_G, leading to a lower correlation.

3.4 Application

As discussed in § 2.1.2, in many scenarios linear CLWE mappings outperform their nonlinear counterparts, while in other setups nonlinear CLWE mappings are more successful. Therefore, an indicator that predicts the relationship between independently pre-trained monolingual word embedding which helps decide whether to use linear or non-linear mappings without training actual CLWEs, would be beneficial. Use of this indicator has the potential to reduce the resources required to find optimal CLWEs (e.g., some recent approaches need several hours of processing on modern GPUs (Peng et al., 2021b; Ormazabal et al., 2021)), with corresponding reductions in carbon footprint.

The proposed \mathcal{S}_{PAE} metric, which can be obtained within several minutes on a single CPU, can be leveraged as such a metric. A high \mathcal{S}_{PAE} score suggests that the linear assumption holds strongly on the ground-truth CLWE mapping, so it is feasible to train a linear CLWE mapping; otherwise, the non-linear approaches are recommended.

To demonstrate this idea in practice, we revisited a systematic evaluation on CLWE models based on linear mappings (Glavaš et al., 2019), which reported Mean Reciprocal Rank (MRR) of five representative linear-mapping-based CLWE approaches on the Word Translation task (the de facto standard for CLWEs). We focus on six language pairs (EN-FI, EN-HR, EN-RU, FI-HR, FI-RU, HR-RU) as they are covered by both xANLG_M and the dataset of Glavaš et al. (2019). Additionally, only Wiki embeddings were involved in the experiments of Glavaš et al. (2019). Thus, for each language pair, we aggregated \mathcal{S}_{PAE} of different analogy categories for Wiki embeddings, then calculated the average, $\bar{\mathcal{S}}_{\text{PAE}}$.

Results are shown in Tab. 3.5, where the Spearman’s ρ between $\bar{\mathcal{S}}_{\text{PAE}}$ and Word Translation performance is highlighted. Strong positive correlations are observed in all setups that were tested. These results demonstrate that $\bar{\mathcal{S}}_{\text{PAE}}$ provides as accurate indication of the real-world performance of linear CLWE mappings, regardless of the language pair, mapping algorithm, or level of supervision (i.e., size of the seed dictionary for training). These results also provide solid support to the main statement of this chapter, i.e., the ground-truth CLWE mapping between monolingual word embeddings is linear *iff* analogies encoded

<i>CLWE method</i>	CCA			PROC			PROC-B			DLV			RCSLS			\bar{S}_{PAE}
	1K	3K	5K	1K	3K	5K	1K	3K	5K	1K	3K	5K	1K	3K	5K	
<i>Seed dict. size</i>	1K	3K	5K	1K	3K	5K	1K	3K	5K	1K	3K	5K	1K	3K	5K	
EN-FI	.26	.35	.38	.27	.37	.40	.36	.38	.38	.27	.37	.40	.31	.40	.44	.41
EN-HR	.22	.30	.33	.23	.31	.34	.30	.34	.30	.23	.31	.33	.27	.36	.38	.32
EN-RU	.34	.43	.45	.35	.45	.46	.42	.45	.45	.35	.44	.47	.40	.49	.51	.46
FI-HR	.17	.26	.29	.19	.27	.29	.26	.29	.26	.18	.27	.29	.21	.30	.32	.23
FI-RU	.21	.31	.34	.23	.31	.34	.32	.33	.32	.23	.31	.34	.26	.34	.38	.33
HR-RU	.26	.35	.37	.27	.35	.37	.35	.37	.35	.26	.35	.37	.29	.38	.40	.26
Spearman's ρ	.83	.82	.86	.83	.84	.88	.83	.86	.88	.84	.84	.87	.87	.88	.90	

Table 3.5: Spearman's ρ between the Word Translation performance (MRR) of linear-mapping-based CLWE methods (from Glavaš et al. (2019); PROC-B's performance with 5K seed dictionary was not available) and the average analogy encoding preservation score (\bar{S}_{PAE}).

in those embeddings are preserved.

3.5 Further Implications

Prior work relevant to the linearity of CLWE mappings has largely been observational (see § 2.1.2). This section sheds new light on these past studies from the novel perspective of word analogies.

Explaining Non-Linearity

We provide three suggested reasons why CLWE mappings are sometimes not approximately linear, all linked with the condition of Eq. (3.1) not being met.

The first may be issues with individual monolingual embeddings (see one such example in the upper part of Fig. 3.2). In particular, popular word embedding algorithms lack the capacity to ensure semantic continuity over the entire embedding space (Linzen, 2016). Hence, vectors for the analogy words may only exhibit local consistency, with Eq. (3.1) breaking down for relatively distant regions. This caused the locality of linearity that has been reported by Nakashole and Flauger (2018), Li et al. (2021d) and Wang et al. (2021a).

The second reason why a CLWE mapping may not be linear is semantic gaps. Despite analogies in our xANLG corpus all are language-agnostic, the analogical relations between words may change or even disappear sometimes. For example, languages pairs may have very different grammars, e.g., Chinese does have the plural morphology (Li and Thompson, 1989), so some types of analogy, e.g. **G-PL** used above, do not hold. Also, analogies may evolve differently across cultures, (see example in the lower part of Fig. 3.2). These two factors go some way to explain why typologically and etymologically distant language pairs tend to have worse alignment (Ruder et al., 2019).

Thirdly, many studies point out that differences in the domain of training data can influence the similarity between multilingual word embeddings (Søgaard et al., 2018; Artetxe et al., 2018). Besides, we argue that due to polysemy, analogies may change from one domain to another. Under such circumstances, Eq. (3.1) is violated and the linear assumption no longer holds.

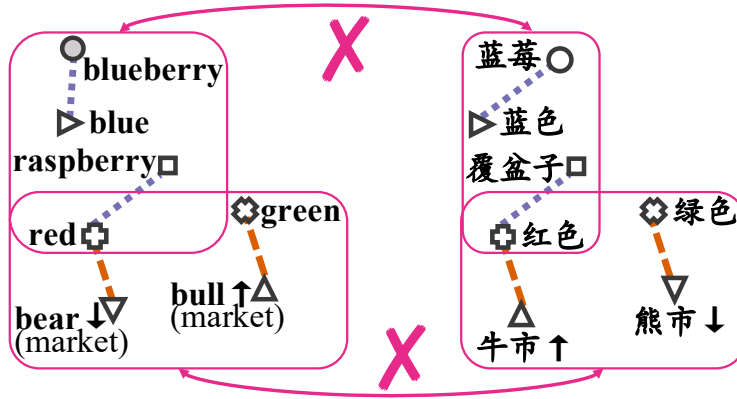


Figure 3.2: Illustration of example scenarios where the CLWE mapping is non-linear. Translations of English (left) and Chinese (right) terms are indicated by shared symbols. **Upper:** The vector for “*blueberry*” (shaded) is ill-positioned in the embedding space, so the condition of Eq. (3.1) is no longer satisfied. **Lower:** In the financial domain some Eastern countries (e.g., China and Japan) traditionally use “*black*” to indicate growth and “*green*” for reduction, while Western countries (e.g., US and UK) assign the opposite meanings to these terms, also not satisfying the condition of Eq. (3.1).

Mitigating Non-Linearity

The proposed analogy-inspired framework justifies the success and failure of the linearity assumption for CLWEs. As discussed earlier, it also suggests a method for indirectly assessing the linearity of a CLWE mapping prior to implementation. Moreover, it offers principled methods for designing more effective CLWE methods. The most straightforward idea is to explicitly use Eq. (3.1) as a training constraint, which has very recently been practised by Garneau et al. (2021).⁷ Based on analogy pairs retrieved from external knowledge bases for different languages, their approach directly learnt to better encode monolingual analogies, particularly those whose vectors are distant in the embedding space. It not only works well on static word embeddings, but also leads to performance gain for large-scale pre-trained cross-lingual language models including the multilingual BERT (Devlin et al., 2019). These results on multiple tasks (e.g., bilingual lexicon induction and cross-lingual sentence retrieval) can be seen as an independent

⁷They cited our earlier preprint as the primary motivation for their approach.

confirmation of this chapter’s main claim and demonstration of its usefulness.

Our study also suggests another unexplored direction: incorporating analogy-based information into non-linear CLWE mappings. Existing work has already introduced non-linearity to CLWE mappings by applying a variety of techniques including directly training non-linear functions (Mohiuddin et al., 2020), tuning linear mappings for outstanding non-isomorphic instances (Glavaš and Vulić, 2020) and learning multiple linear CLWE mappings instead of a single one (Nakashole, 2018; Wang et al., 2021a) (see § 2.1.2). However, there is a lack of theoretical motivation for decisions about how the non-linear mapping should be modelled. Nevertheless, the results presented here suggest that ensembles of linear transformations, covering analogy preserving regions of the embedding space, would make a reasonable approximation of the ground-truth CLWE mappings and that information about analogy preservation could be used to partition embedding spaces into multiple regions, between which independent linear mappings can be learnt. We leave this application as our important future work.

3.6 Summary and Discussion

CLWE bridges the gap between languages and is efficient enough to be applied in situations where limited resources are available, including to endangered languages (Zhang et al., 2020; Ngoc Le and Sadat, 2020). This chapter presented a theoretical analysis of the mechanisms underlying CLWE techniques which has potential to improve these methods. Moreover, the proposed \mathcal{S}_{PAE} metric predicts whether monolingual word embeddings in different languages should be aligned using a linear or non-linear mapping, without actually training the CLWEs. This indicator lowers the computational expense required to identify a suitable mapping approach, thereby reducing the computational power needed and negative environmental effects.

One limitation of CLWEs is the inconvenience of handling multi-word terms. In the future, we will attempt to mitigate this via vector composition (Cordeiro et al., 2016), and expand our experiments on analogies for multi-word expressions. Besides, we plan to enrich xANLG by including new languages (e.g., African ones) and analogy categories (e.g., meronyms) to enable explorations at an even

larger scale.

3.7 Post-Publication Retrospect

Since the work described in this chapter was publicly released, new training paradigms directly motivated by it have been proposed, e.g., Garneau et al. (2021). Beyond building analogy corpora in the cross-lingual setup, very recent work has constructed resources for multi-modal scenarios (Zhang et al., 2023).

REFINING CLWE MAPPINGS VIA ℓ_1 NORM OPTIMISATION

CLWEs encode words from two or more languages in a shared high-dimensional space in which vectors representing words with similar meaning (regardless of language) are closely located. Existing methods for building high-quality CLWEs learn mappings that minimise the ℓ_2 norm loss function. However, this optimisation objective has been demonstrated to be sensitive to outliers. Based on the more robust Manhattan norm (aka. ℓ_1 norm) goodness-of-fit criterion, this chapter proposes a simple post-processing step to improve CLWEs. An advantage of this approach is that it is fully agnostic to the training process of the original CLWEs and can therefore be applied widely. Extensive experiments are performed involving ten diverse languages and embeddings trained on different corpora. Evaluation results based on bilingual lexicon induction and cross-lingual transfer for natural language inference tasks show that the ℓ_1 refinement substantially outperforms four state-of-the-art baselines in both supervised and unsupervised settings. It is therefore recommended that this strategy be adopted as a standard for CLWE methods. The code is available at <https://github.com/Pzoom522/L1-Refinement>.

4.1 Methodology

A common characteristic of CLWE methods that apply the orthogonality constraint is that they optimise using ℓ_2 loss (see § 2.1.3). However, outliers have disproportionate influence in ℓ_2 since the penalty increases quadratically and this can be particularly problematic with noisy data since the solution can “shift” towards them (Rousseeuw and Leroy, 1987). The noise and outliers present in real-world word embeddings may affect the performance of ℓ_2 -loss-based CLWEs.

The ℓ_1 norm cost function is more robust than ℓ_2 loss as it is less affected by outliers (Rousseeuw and Leroy, 1987). Therefore, we propose a refinement algorithm for improving the quality of CLWEs based on ℓ_1 loss. This novel method, which we refer to as ℓ_1 refinement, is generic and can be applied post-hoc to improve the output of existing CLWE models. To our knowledge, the use of alternatives to ℓ_2 -loss-based optimisation has never been explored by the CLWE community.

To begin with, analogous to ℓ_2 OPA (cf. Eq. (2.1)), ℓ_1 OPA can be formally defined and rewritten as

$$\operatorname{argmin}_{\mathbf{M} \in \mathcal{O}} \|\mathbf{AM} - \mathbf{B}\|_1 = \operatorname{argmin}_{\mathbf{M} \in \mathcal{O}} \operatorname{tr}[(\mathbf{AM} - \mathbf{B})^\top \operatorname{sgn}(\mathbf{AM} - \mathbf{B})], \quad (4.1)$$

where $\operatorname{tr}(\cdot)$ returns the matrix trace, $\operatorname{sgn}(\cdot)$ is the signum function, and $\in \mathcal{O}$ denotes that M is subject to the orthogonal constraint. Compared to ℓ_2 OPA which has a closed-form solution, solving Eq. (4.1) is much more challenging due to the discontinuity of $\operatorname{sgn}(\cdot)$. This issue can be addressed by replacing $\operatorname{sgn}(\cdot)$ with $\tanh(\alpha(\cdot))$, a smoothing function parameterised by α , such that

$$\operatorname{argmin}_{\mathbf{M} \in \mathcal{O}} \operatorname{tr}[(\mathbf{AM} - \mathbf{B})^\top \tanh(\alpha(\mathbf{AM} - \mathbf{B}))]. \quad (4.2)$$

Larger values for α lead to closer approximations to $\operatorname{sgn}(\cdot)$ but reduce the smoothing effect. This approach has been used in many applications, such as the activation function of long short-term memory networks (Hochreiter and Schmidhuber, 1997).

However, in practice, we find that Eq. (4.2) remains unsolvable in our case

with standard gradient-based frameworks for two reasons. First, α has to be sufficiently large in order to achieve a good approximation of $\text{sgn}(\cdot)$. Otherwise, relatively small residuals will be down-weighted during fitting and the objective will become biased towards outliers, just similar to ℓ_2 loss. However, satisfying this requirement (i.e., large α) will lead to the activation function $\tanh(\alpha(\cdot))$ becoming easily *saturated*, resulting in an optimisation process that becomes trapped during the early stages. In other words, the optimisation can only reach an unsatisfactory local optimum. Second, the orthogonality constraint (i.e., $\mathbf{M} \in \mathcal{O}$) also makes the optimisation more problematic for these methods.

We address these challenges by adopting the approaches proposed by Trendafilov (2003).¹ This method explicitly encourages the solver to only explore the desired manifold \mathcal{O} thereby reducing the ℓ_1 solver's search space and difficulty of the optimisation problem. We begin by calculating the gradient ∇ w.r.t. the objective in Eq. (4.2) through matrix differentiation:

$$\nabla = \mathbf{A}^\top (\tanh(\mathbf{Z}) + \mathbf{Z} \odot \cosh^{-2}(\mathbf{Z})), \quad (4.3)$$

where $\mathbf{Z} = \alpha(\mathbf{A}\mathbf{M} - \mathbf{B})$ and \odot is the Hadamard product. Next, to find the steepest descent direction while ensuring that any \mathbf{M} produced is orthogonal, we project ∇ onto \mathcal{O} , yielding²

$$\pi_{\mathcal{O}}(\nabla) := \frac{1}{2} \mathbf{M}(\mathbf{M}^\top \nabla - \nabla^\top \mathbf{M}) + (\mathbf{I} - \mathbf{M}\mathbf{M}^\top) \nabla. \quad (4.4)$$

Here \mathbf{I} is an identity matrix with the shape of \mathbf{M} . With Eq. (4.4) defining the optimisation flow, our ℓ_1 loss minimisation problem reduces to an integration problem, as

$$\mathbf{M}^* = \mathbf{M}_0 + \int -\pi_{\mathcal{O}}(\nabla) dt, \quad (4.5)$$

where \mathbf{M}_0 is a proper initial solution of Eq. (4.1) (e.g., ℓ_2 -optimal mapping ob-

¹NB: When borrowing the solution of Trendafilov (2003), we fixed their *weighted matrix* as an identity matrix given the nature of CLWEs. While Trendafilov (2003) proposed to conduct the optimisation from scratch, our algorithm employs the ℓ_2 -norm optima as the starting point. In addition, we adjusted the ODE solver selection, because the ones originally used by Trendafilov (2003) are no longer applicable in the CLWE scenario (due to the large scale of dimensions and data points).

²See Chu and Trendafilov (2001) for derivation details.

Algorithm 1 ℓ_1 refinement

Input: CLWEs $\{\mathbf{X}_{L_A}, \mathbf{X}_{L_B}\}$
Output: updated CLWEs $\{\mathbf{X}_{L_A} \mathbf{M}^*, \mathbf{X}_{L_B}\}$

- 1: $D_{L_A \mapsto L_B} \leftarrow$ build dict via \mathbf{X}_{L_A} and \mathbf{X}_{L_B}
 - 2: $D_{L_B \mapsto L_A} \leftarrow$ build dict via \mathbf{X}_{L_B} and \mathbf{X}_{L_A}
 - 3: $D \leftarrow D_{L_A \mapsto L_B} \cap D_{L_B \mapsto L_A}$
 - 4: $\mathbf{A}, \mathbf{B} \leftarrow$ looks up for D in $\mathbf{X}_{L_A}, \mathbf{X}_{L_B}$
 - 5: perform integration to solve Eq. (4.5) for \mathbf{M}^* , with initial value $\mathbf{M}_0 \leftarrow \mathbf{I}$, until stopping criteria are met
-

tained via Eq. (2.2)).

Empirically, unlike the aforementioned standard gradient-based methods, by following the established policy of Eq. (4.4), we observed that the optimisation process of Eq. (4.5) do not violate the orthogonality restriction or get trapped during early stages. However, this ℓ_1 OPA solver requires extremely small step size to generate reliable solutions (Trendafilov, 2003), making it computationally expensive.³ Therefore, it is impractical to perform ℓ_1 refinement in an iterative fashion like ℓ_2 refinement without significant computational resources.

Previous work has demonstrated that applying the ℓ_1 -loss-based algorithms from a good initial state can speed up the optimisation. For instance, Kwak (2008) found that feature spaces created by ℓ_2 PCA were severely affected by noise. Replacing the cost function with ℓ_1 loss significantly reduced this problem, but required expensive linear programming. To reduce the convergence time, Brooks and Jot (2013) exploited the first principal component from the ℓ_2 solution as an initial guess. Similarly, when reconstructing corrupted pixel matrices, ℓ_2 -loss-based results are far from satisfactory; using ℓ_1 norm estimators can improve the quality, but are too slow to handle large-scale datasets (Aanaes et al., 2002). However, taking the ℓ_2 optima as the starting point allowed less biased reconstructions to be learned in an acceptable time (De La Torre and Black, 2003).

Inspired by these works, we make use of ℓ_1 refinement to carry out post-hoc

³It takes averagely 3 hours and up to 12 hours to perform Eq. (4.5) on an Intel Core i9-9900K CPU. In comparison, the time required to solve Eq. (2.2) in each training loop is less than 1 second and the iterative ℓ_2 -norm-based training takes 1 to 5 hours in total.

enhancement of existing CLWEs. Our full pipeline is described in Algorithm 1 (see § 4.2.3 for implemented configurations). In common with ℓ_2 refinement (cf. § 2.1.3), steps 1-4 bootstrap a synthetic dictionary D and compose bilingual word vector matrices \mathbf{A} and \mathbf{B} which have reliable row-wise correspondence. Taking them as the starting state, in step 5 an identity matrix naturally serves as our initial solution \mathbf{M}_0 .

During the execution of Eq. (4.5), we record ℓ_1 loss per iteration and see if *either* of the following two stopping criteria have been satisfied: (1) the updated ℓ_1 loss exceeds that of the previous iteration; (2) on-the-fly \mathbf{M} has non-negligibly departed from the orthogonal manifold, which can be indicated by the maximum value of the disparity matrix as

$$\max(|\mathbf{M}^\top \mathbf{M} - \mathbf{I}|) > \epsilon, \quad (4.6)$$

where ϵ is a sufficiently small threshold. The resulting \mathbf{M}^* can be used to adjust the word vectors of \mathbf{L}_A and output refined CLWEs.

A significant advantage of our algorithm is its generality: it is fully independent of the method used for creating the original CLWEs and can therefore be used to enhance a wide range of models, both in supervised and unsupervised settings.

4.2 Experimental Setup

4.2.1 Datasets

In order to demonstrate the generality of our proposed method, we conduct experiments using two groups of monolingual word embeddings trained on very different corpora:

- **Wiki-Embs** (Grave et al., 2018): embeddings developed using Wikipedia dumps for a range of ten diverse languages: two Germanic (English_{|EN}, German_{|DE}), two Slavic (Croatian_{|HR}, Russian_{|RU}), three Romance (French_{|FR}, Italian_{|IT}, Spanish_{|ES}) and three non-Indo-European (Finnish_{|FI} from the

Uralic family, Turkish_{|TR} from the Turkic family and Chinese_{|ZH} from the Sino-Tibetan family).

- **News-Embs** (Artetxe et al., 2018): embeddings trained on a multilingual News text collection, i.e., the WaCKy Crawl of {EN, DE, IT}, the Common Crawl of FI, and the WMT News Crawl of ES.

News-Embs are considered to be more challenging for building good quality CLWEs due to the heterogeneous nature of the data, while a considerable portion of the multilingual training corpora for Wiki-Embs are roughly parallel. Following previous studies (Lample et al., 2018b; Artetxe et al., 2018; Zhou et al., 2019a; Glavaš et al., 2019), only the first 200K vocabulary entries are preserved.

4.2.2 Baselines

Glavaš et al. (2019) provided a systematic evaluation for projection-based CLWE models, demonstrating that three methods (i.e., MUSE, VECMAP, and PROC-B) achieve the most competitive performance. A recent algorithm (JA) by Wang et al. (2020) also reported state-of-the-art results. For comprehensive comparison, we therefore use all these four methods as the main baselines for both supervised and unsupervised settings (we directly adopted their official codebases and hyper-parameter configurations):

- **Muse** (Lample et al., 2018b): an *unsupervised* CLWE model based on adversarial learning and iterative ℓ_2 refinement;
- **VecMap** (Artetxe et al., 2018): a robust *unsupervised* framework using a self-learning strategy;
- **Proc-B** (Glavaš et al., 2019): a simple but effective *supervised* approach to creating CLWEs;
- **JA-Muse** and **JA-RCSLS** (Wang et al., 2020): a recently proposed Joint-Align (JA) Framework, which first initialises CLWEs using joint embedding training, followed by vocabularies reallocation. It then utilises off-the-shelf CLWE methods to improve the alignment in both *unsupervised* (**JA-Muse**) and *supervised* (**JA-RCSLS**) settings.

In the original implementations, MUSE, PROC-B and JA were only trained on Wiki-Embs while VECMAP additionally used News-Embs. Although all baselines reported performance for BLI, they used various versions of evaluation sets, hence previous results are not directly comparable with the ones reposted here. More concretely, the testsets for MUSE/JA and VECMAP are two different batches of EN-centric dictionaries, while the testset for PROC-B also supports non-EN translations.

4.2.3 Implementation Details

The CSLS scheme with a neighbourhood size of 10 is adopted to build synthetic dictionaries via the input CLWEs. A variable-coefficient ordinary differential equation (VODE) solver⁴ was implemented for the system described in Eq. (4.5). Suggested by Trendafilov (2003), we set the maximum order at 15, the smoothness coefficient α in Eq. (4.3) at $1e8$, the threshold ϵ in Eq. (4.6) at $1e-5$, and performed the integration with a fixed time interval of $1e-6$. An early-stopping design was adopted to ensure computation completed in a reasonable time: in addition to the two default stopping criteria in § 4.1, integration is terminated if $\int dt$ reaches $5e-3$ (dt is the differentiation term in Eq. (4.5)).

In terms of the tolerance of the VODE solver, we set the absolute tolerance at $1e-7$ and the relative tolerance at $1e-5$, following the established approach of Kulikov (2013). These tolerance settings show good generality empirically and were used for all tested language pairs, datasets, and models in our experiments.

4.3 Results

We evaluate the effectiveness of the proposed ℓ_1 refinement technique on two benchmarks: Bilingual Lexicon Induction (BLI), the *de facto* standard for measuring the quality of CLWEs, and a downstream natural language inference task based on cross-lingual transfer. In addition to comparison against state-of-the-art CLWE models, we also report the performance of the single-iteration ℓ_2 refinement method which follows steps 1-4 of Algorithm 1 then minimises ℓ_2 loss in the

⁴<http://www.netlib.org/ode/vode.f>

final step.

To reduce randomness, we executed each model in each setup three times and the average accuracy (ACC, aka. precision at rank 1) is reported. Following Glavaš et al. (2019), by comparing scores achieved before and after ℓ_1 refinement, statistical significance is indicated via the p -value of two-tailed t-tests with Bonferroni correction (Dror et al., 2018) (note that p -values are not recorded for Tab. 4.2b given the small number of runs).

4.3.1 Bilingual Lexicon Induction

Refining Unsupervised Baselines

Tab. 4.1a follows the main setup of Lample et al. (2018b), who tested six language pairs using Wiki-Embs.⁵ After ℓ_1 refinement, MUSE- ℓ_1 , JA-MUSE- ℓ_1 , and VECMAP- ℓ_1 all significantly ($p < 0.01$) outperform their corresponding base algorithms, with an average 1.1% performance gain over MUSE, 1.1% over JA-MUSE, and 0.5% over VECMAP. To put these improvements in context, Heyman et al. (2019) reported an improvement of 0.4% for VECMAP on same dataset and language pairs.

Our method tends to work better on the more distant language pairs. For instance, for the distant pairs EN- $\{\text{RU}, \text{ZH}\}$, the increments achieved by MUSE- ℓ_1 are 1.6% and 1.3%, respectively; whereas for the close pairs EN- $\{\text{DE}, \text{ES}, \text{FR}\}$ the average gain is a maximum of 0.9%. A similar trend can be observed for JA-MUSE- ℓ_1 and VECMAP- ℓ_1 . (As the VECMAP algorithm always collapses for EN-ZH, no result is reported for this language pair).

Another set of experiments were conducted to evaluate the robustness of our algorithm following the main setup of Artetxe et al. (2018), who tested four language pairs based on the more homogeneous News-Embs. Tab. 4.1b shows that JA-MUSE- ℓ_1 and VECMAP- ℓ_1 consistently improves the original VECMAP with an average gain of 1.2% and 1.0% ($p < 0.01$). Obtaining such substantial improvements over the state-of-the-art is nontrivial. For example, even a recent weakly supervised method by Wang et al. (2019b) is *inferior* to VECMAP by

⁵Note that we are unable to report the result of English to Esperanto as the corresponding dictionary is missing, see <https://git.io/en-eo-dict-issue>.

	EN-DE	EN-ES	EN-FR	EN-RU	EN-ZH
MUSE	74.0	81.7	82.3	44.0	32.5
MUSE- ℓ_2	74.0	82.1	82.6	43.8*	31.9*
MUSE- ℓ_1	75.2	82.6	82.9	45.6*	33.8*
JA-MUSE	74.2	81.4	82.8	45.0	36.1
JA-MUSE- ℓ_2	74.1	81.6	82.7	45.1	36.2
JA-MUSE- ℓ_1	75.4	82.0	83.1	46.3	38.1
VECMAP	75.1	82.3	80.0	49.2	00.0
VECMAP- ℓ_2	74.8	82.3	79.4	48.9	00.0
VECMAP- ℓ_1	75.4	82.9	80.2	49.9	00.0

(a) Wiki-Embs (setup of Lample et al. (2018b)).

	EN-DE	EN-ES	EN-FI	EN-IT
MUSE	00.0	07.1	00.0	09.1
MUSE- ℓ_2	00.0	00.0	00.0	00.0
MUSE- ℓ_1	00.0	00.0	00.0	00.0
JA-MUSE	47.9	48.4	33.0	37.2
JA-MUSE- ℓ_2	47.9	48.6	32.9	37.3
JA-MUSE- ℓ_1	48.8	49.7	35.2	37.7
VECMAP	48.2	48.1	32.6	37.3
VECMAP- ℓ_2	48.1	47.9	32.9	37.1
VECMAP- ℓ_1	49.0	48.9	34.4	37.8

(b) News-Embs (setup of Artetxe et al. (2018)).

Table 4.1: ACC (%) of unsupervised BLI. NB: for EN- $\{\text{RU}, \text{ZH}\}$ we observed one failed run (ACC <10.0%), where we only record the average of successful scores with *.

1.0% average ACC. On the other hand, MUSE fails to produce any analysable result as it always collapses on the more challenging News-Embs. Improvement with ℓ_1 refinement is also larger when language pairs are more distant, e.g., for VECMAP- ℓ_1 the ACC gain on EN-FI is 1.8%, more than double of the gain (0.7%) on the close pairs EN- $\{\text{DE}, \text{IT}\}$ (cf. Tab. 4.1a and above).

We also conduct an ablation study by reporting the performance of ℓ_2 refinement scheme ($\{\text{MUSE}, \text{JA-MUSE}, \text{VECMAP}\}$ - ℓ_2). This observation is in accordance with that of Lample et al. (2018b), who reported that after performing ℓ_2 refinement in the first loop, applying further iterations only produces marginal precision gain, if any.

	EN-DE	EN-FI	EN-FR	EN-HR	EN-IT	EN-RU	EN-TR
JA-RCSLS	50.9	33.9	63.0	29.1	58.3	41.3	29.4
JA-RCSLS- ℓ_2	50.7	33.8	63.0	29.1	58.2	41.3	29.5
JA-RCSLS- ℓ_1	51.6	34.5	63.4	30.4	59.0	41.9	30.2
PROC-B	52.1	36.0	63.3	29.6	60.5	41.9	30.1
PROC-B- ℓ_2	51.8	34.4	63.1	28.2	60.5	39.8	28.0
PROC-B- ℓ_1	52.6	36.3	63.7	30.5	60.5	42.3	30.9

(a) Wiki-Embs (setup of Glavaš et al. (2019)).

	EN-DE	EN-FI	EN-IT
JA-RCSLS	46.8	42.0	37.4
JA-RCSLS- ℓ_2	46.9	42.2	37.5
JA-RCSLS- ℓ_1	48.3	44.6	39.0
PROC-B	47.5	41.4	37.3
PROC-B- ℓ_2	47.1	41.7	37.4
PROC-B- ℓ_1	52.6	43.3	41.1

(b) News-Embs.

Table 4.2: MRR (%) of supervised BLI.

Overall, the ℓ_1 refinement consistently and significantly improve the CLWEs produced by base algorithms, regardless of the embeddings and setups used, thereby demonstrating the effectiveness and robustness of the proposed algorithm.

Refining Supervised Baselines

To test the generalisability of our method, we also applied it on state-of-the-art supervised CLWE models: PROC-B (Glavaš et al., 2019) and JA-RCSLS (Wang et al., 2020). Following the setup of Glavaš et al. (2019), we learn mappings using Wiki-Embs and 1K training splits of their dataset.

Their evaluation code retrieves bilingual word pairs using the classic nearest-neighbour algorithm and outputs the Mean Reciprocal Rank (MRR). As shown in Tab. 4.2a, both JA-RCSLS- ℓ_1 and PROC-B- ℓ_1 outperform the baseline algorithms for all language pairs (with the exception of EN-IT where the score of PROC-B is unchanged) with an average improvement of 0.9% and 0.5%, respectively ($p < 0.01$).

<i>Unsupervised</i>	DE-IT	DE-TR	FI-HR	FI-IT	HR-RU	IT-FR	TR-IT
ICP	44.7	21.5	20.8	26.3	30.9	62.9	24.3
GWA	44.0	10.1	00.9	17.3	00.1	65.5	14.2
MUSE	49.6	23.7	22.8	32.7	00.0	66.2	30.6
MUSE- ℓ_2	50.3	23.9	23.1	32.7	34.9	67.1	30.5*
MUSE- ℓ_1	50.7	26.5	25.4	35.0	37.9	67.6	33.3*
JA-MUSE	50.9	25.6	23.4	34.9	36.9	68.3	34.7
JA-MUSE- ℓ_2	50.9	25.5	23.4	34.7	36.9	68.4	34.7
JA-MUSE- ℓ_1	51.5	28.4	26.1	36.0	37.6	68.7	36.1
VECMAP	49.3	25.3	28.0	35.5	37.6	66.7	33.2
VECMAP- ℓ_2	48.8	25.7	28.5	35.8	38.4	67.0	33.5
VECMAP- ℓ_1	50.1	28.2	30.3	37.1	40.1	67.6	35.9
<i>Supervised</i>							
DLV	42.0	16.7	18.4	24.4	26.4	58.5	20.9
RCSLS	45.3	20.1	21.4	27.2	29.1	63.7	24.6
JA-RSCLS	46.6	20.9	22.1	29.0	29.9	65.2	25.3
JA-RSCLS- ℓ_2	46.4	20.8	22.3	29.0	29.8	65.2	25.3
JA-RSCLS- ℓ_1	47.3	22.2	23.8	30.1	31.2	65.9	26.6
PROC-B	50.7	25.0	26.3	32.8	34.8	66.5	29.8
PROC-B- ℓ_2	50.0	24.1	25.6	31.8	34.3	66.4	29.6
PROC-B- ℓ_1	51.1	25.6	26.9	33.6	35.0	67.4	30.5

Table 4.3: MRR (%) of BLI for non-EN language pairs. MUSE yielded one unsuccessful run for TR-IT, and we only record the average of the two successful scores with *.

JA-RSCLS- ℓ_1 and PROC-B- ℓ_1 were also tested using News-Embs with results shown in Tab. 4.2b⁶. ℓ_1 refinement achieves an impressive improvement for both close (EN- $\{\text{DE}, \text{IT}\}$) and distant (EN-FI) language pairs: average gain of 1.9% and 3.9% respectively and over 5% for EN-DE (PROC-B- ℓ_1) in particular. The ℓ_2 refinement does not benefit the supervised baseline, similar to the lack of improvement observed in the unsupervised setups.

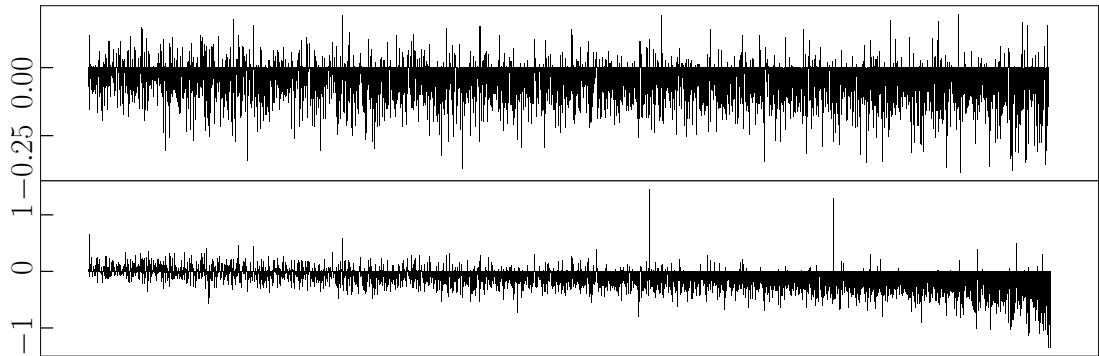
⁶Note that results for EN-ES is not included, as no EN-ES dictionary is provided in Glavaš et al. (2019)’s dataset.

Comparison of Unsupervised and Supervised Settings

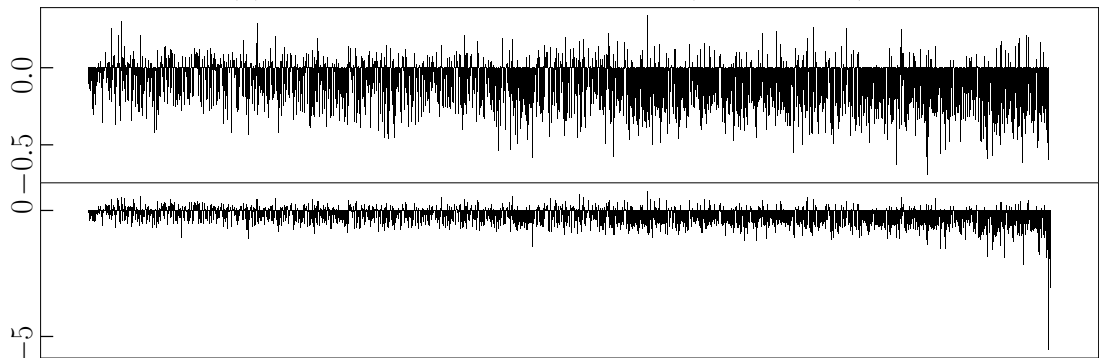
This part provides a comparison of the effectiveness of ℓ_1 refinement in unsupervised and supervised scenarios. Unlike previous experiments where only alignments involving English were investigated, these tests focus on non-EN setups. Glavaš et al. (2019)’s dataset is used to construct seven representative pairs which cover every category of etymological combination, i.e., intra-language-branch {HR–RU, IT–FR}, inter-language-branch {DE–IT}, and inter-language-family {DE–TR, FI–HR, FI–IT, TR–IT}. The 1K training splits are used as seed lexicons in supervised runs. Apart from our main baselines, we further report the results of several other competitive CLWE models: Iterative Closest Point Model (ICP, Hoshen and Wolf, 2018), Gromov-Wasserstein Alignment Model (GWA, Alvarez-Melis and Jaakkola, 2018), Discriminative Latent-Variable Model (DLV, Ruder et al., 2018) and Relaxed CSLS Model (RCSLS, Joulin et al., 2018).

Results shown in Tab. 4.3 demonstrate that the main baselines (MUSE, JA-MUSE, VECMAP, JA-RCSLS, and PROC-B) outperform these other models by a large margin. For all these main baselines, post applying ℓ_1 refinement improves the mapping quality for all language pairs ($p < 0.01$), with average improvements of 1.7%, 1.4%, 1.8%, 1.1%, and 0.8%, respectively. Consistent with findings in the previous experiments, ℓ_2 refinement does not enhance performance. Improvement with ℓ_1 refinement is higher when language pairs are more distant, e.g., for all inter-language-family pairs such as FI–HR and TR–IT, even the minimum improvement of MUSE- ℓ_1 over MUSE is 2.3%.

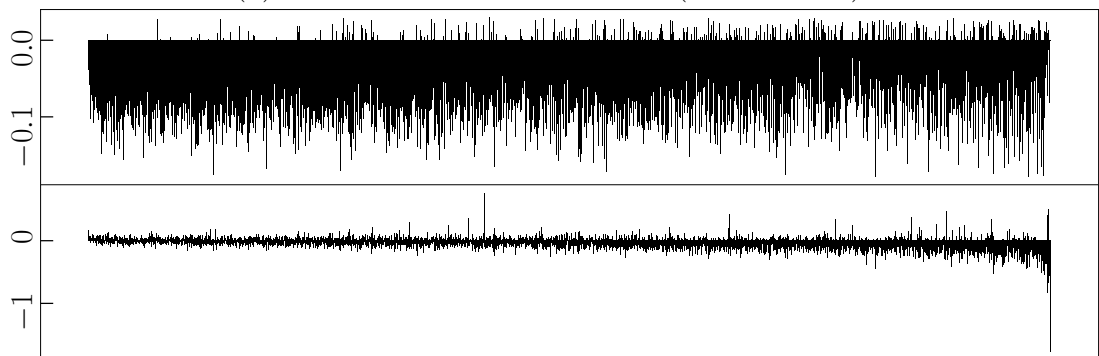
Comparing unsupervised and supervised approaches, it can be observed that MUSE, JA-MUSE and VECMAP achieve higher overall gain with ℓ_1 refinement than JA-RCSLS and PROC-B, where JA-MUSE- ℓ_1 and VECMAP- ℓ_1 give the best overall performance. One possible explanation to this phenomenon is that there is only a single source of possible noise in unsupervised models (i.e. the embedding topology) but for supervised methods noise can also be introduced via the seed lexicons. Consequently unsupervised approaches drive more benefit from ℓ_1 refinement, which reduces the influence of topological outliers in CLWEs.



(a) VECMAP on EN-RU Wiki-Embs (cf. Tab. 4.1a).



(b) PROC-B on EN-FI News-Embs (cf. Tab. 4.2b).



(c) MUSE on IT-FR Wiki-Embs (cf. Tab. 4.3).

Figure 4.1: Changes to $\|\mathbf{AM} - \mathbf{B}\|_2$ after applying ℓ_1 (upper) and ℓ_2 (lower) refinement. Each word pairs is represented by a bar ordered on the x-axis by the distance between them.

Topological Behaviours of ℓ_1 and ℓ_2 Refinements

To validate our assumption that ℓ_2 refinement is more sensitive to outliers while its ℓ_1 counterpart is more robust, we analyse how each refinement strategy changes the distance between bilingual word vector pairs in the synthetic dictionary D (cf. Algorithm 1) constructed from trained CLWE models. Specifically, for each word vector pair we subtract its post-refinement distance from the original distance (i.e., without applying additional ℓ_1 or ℓ_2 refinement step). Fig. 4.1 shows visualisation examples for three algorithms and language pairs, where each bar represents one word pair. It can be observed that ℓ_1 refinement effectively reduces the distance for most word pairs, regardless of their original distance (i.e., indicated by bars with negative values in the figures). The conventional ℓ_2 refinement strategy, in contrast, exhibits very different behaviour and tends to be overly influenced by word pairs with large distance (i.e. by outliers). The reason for this is that the ℓ_2 -norm penalty increases quadratically, causing the solution to put much more weight on optimising distant word pairs (i.e., word pairs on the right end of the X-axis show sharp distance decrements). This observation is in line with Rousseeuw and Leroy (1987) and explains why ℓ_1 loss performs substantially stronger than ℓ_2 loss in the refinement.

Case Study

After aligning EN-RU embeddings with unsupervised MUSE, we measured the distance between vectors corresponding to the ground-truth dictionary of Lample et al. (2018b) (cf. Fig. 4.1a). We then detected large outliers by finding vector pairs whose distance falls above $Q3 + 1.5 \cdot (Q3 - Q1)$, where $Q1$ and $Q3$ respectively denote the lower and upper quartile based on the popular Inter-Quartile Range (Hoaglin et al., 1986).⁷ We found that many of the outliers correspond to polysemous entries, such as {state ($2 \times$ noun meanings and $1 \times$ verb meaning)}, состояние (only means *status*), {type ($2 \times$ nominal meanings and $1 \times$ verb meaning)}, тип (only means *kind*), and {film ($5 \times$ noun meanings)}, фильм (only means

⁷It is worth noting that such an outlier detection process relies on a ground-truth dictionary, which does not exist before the construction of CLWEs in practice. In other words, it is not feasible to directly remove outliers from the learning process as a heuristic.

movie)). We then re-perform ℓ_2 -based mapping after removing these vector pairs, observing that the accuracy jumps to 45.9% (cf. the original ℓ_2 -norm alignment it is 43.8% and after ℓ_1 refinement it is 45.6%, cf. Tab. 4.1). This indicates that although all baselines already make use of preprocessing steps including vector normalization, outlier issues still exist and harms the ℓ_2 norm CLWEs. However, they can be alleviated by the proposed ℓ_1 refinement technique.

Another interesting direction to explore is whether CLWEs can benefit from conducting more ℓ_1 -based optimisation iterations, or in an extreme configuration, replacing all ℓ_2 -based iterations with the ℓ_1 -based ones. Considering the aforesaid low convergence speed, it is unrealistic for us to test all setups. Yet, we installed the ℓ_1 solver on VECMAP and experimented with building EN-RU CLWEs *from scratch*. The training took over four days and yielded an ACC of 51.1%, which is superior to both VECMAP (49.2%) and VECMAP- ℓ_1 (49.9%) (cf. Tab. 4.1a). Such margins once again highlight the usefulness of the proposed ℓ_1 -based CLWE strategy. Future work should thus focus on improving the efficiency of the ℓ_1 -based computation process.

4.3.2 Natural Language Inference

Finally, we experimented with a downstream NLI task in which the aim is to determine whether a “hypothesis” is true (*entailment*), false (*contradiction*) or undetermined (*neutral*), given a “premise”. Higher ACC indicates better encoding of semantics in the tested embeddings. The CLWEs used are those trained with Wiki-Embs for BLI. For MUSE, JA-MUSE and VECMAP, we also obtain CLWEs for EN-TR pair with the same configuration.

Following Glavaš et al. (2019), we first train the Enhanced Sequential Inference Model (Chen et al., 2017) based on the large-scale English MultiNLI corpus (Williams et al., 2018) using vectors of language L_A (EN) from an aligned bilingual embedding space (e.g., EN-DE). Next, we replace the L_A vectors with the vectors of language L_B (e.g., DE), and directly test the trained model on the language L_B portion of the XNLI corpus (Conneau et al., 2018).

Results in Tab. 4.4 show that the CLWEs refined by our algorithm yield the highest ACC for all language pairs in both supervised and unsupervised settings.

<i>Unsupervised</i>	EN-DE	EN-FR	EN-RU	EN-TR
ICP	58.0	51.0	57.2	40.0
GWA	42.7	38.3	37.6	35.9
MUSE	61.1	53.6	36.3	35.9
MUSE- ℓ_2	61.1	53.0	57.3*	48.9*
MUSE- ℓ_1	63.5	55.3	58.9*	52.3*
JA-MUSE	61.3	55.2	58.1	55.0
JA-MUSE- ℓ_2	61.2	55.2	57.6	55.1
JA-MUSE- ℓ_1	62.9	57.9	59.4	57.5
VECMAP	60.4	61.3	58.1	53.4
VECMAP- ℓ_2	60.3	60.6	57.7	53.5
VECMAP- ℓ_1	61.5	63.7	60.1	56.4
<i>Supervised</i>				
RCSLS	37.6	35.7	37.8	38.7
JA-RSCLS	50.2	48.9	51.0	51.7
JA-RSCLS- ℓ_2	50.4	48.6	50.9	51.5
JA-RSCLS- ℓ_1	51.3	50.1	53.2	52.6
PROC-B	61.3	54.3	59.3	56.8
PROC-B- ℓ_2	61.0	54.8	58.9	55.1
PROC-B- ℓ_1	62.1	54.8	60.7	58.2

Table 4.4: ACC (%) of NLI. MUSE yielded one unsuccessful run for EN-RU and EN-TR respectively, which we exclude when calculating the average (with *).

The ℓ_2 refinement, on the contrary, is not beneficial overall. Improvements in cross-lingual transfer for NLI exhibit similar trends to those in the BLI experiments, i.e. greater performance gain for unsupervised methods and more distant language pairs, consistent with previous observations (Glavaš et al., 2019). For instance, MUSE- ℓ_1 JA-MUSE- ℓ_1 and VECMAP- ℓ_1 outperform their baselines by at least 2% in ACC on average ($p < 0.01$), whereas the improvements of JA-RSCLS- ℓ_1 and PROC-B- ℓ_1 over their corresponding base methods are 2% and 2.1% respectively ($p < 0.01$). For both unsupervised and supervised methods, ℓ_1 refinement demonstrates stronger effect for more distant language pairs, e.g., MUSE- ℓ_1 surpasses MUSE by 1.2% for EN-FR, whereas a more impressive 2.7% gain is achieved for EN-TR.

In summary, in addition to improving BLI performance, our ℓ_1 refinement method also produces a significant improvement for a downstream task (NLI),

demonstrating its effectiveness in improving the CLWE quality.

4.4 Summary and Discussion

This work provides an effective post-hoc method to improve CLWEs, advancing the state-of-the-art in both supervised and unsupervised settings. Our comprehensive empirical studies demonstrate that the proposed algorithm can facilitate researches in machine translation, cross-lingual transfer learning, etc. Besides, this chapter introduces and solves an optimisation problem based on an under-explored robust cost function, namely ℓ_1 loss. We believe it could be of interest for the wider community as outlier is a long-standing issue in many artificial intelligence applications.

4.5 Post-Publication Retrospect

The ℓ_1 -based refinement algorithm has been used as a major baseline in other CLWE studies (e.g., Feng et al. (2022)), which renewed the performance records of tasks such as BLI. As for downstream tasks such as cross-lingual NLI, PLMs (e.g., Chi et al. (2021)) have been the dominant approach.

APPLYING CLWE MAPPINGS FOR HISTORICAL TEXT SUMMARISATION

We introduce the task of historical text summarisation, where documents in historical forms of a language are summarised in the corresponding modern language. This is a fundamentally important routine to historians and digital humanities researchers but has never been automated. We compile a high-quality gold-standard text summarisation dataset (for evaluation purposes only), which consists of historical German and Chinese news from hundreds of years ago summarised in modern German or Chinese. Based on cross-lingual transfer learning techniques, we propose a summarisation model that can be trained even with no cross-lingual (historical to modern) parallel data, and further benchmark it against state-of-the-art algorithms. We report automatic and human evaluations that distinguish the historical to modern language summarisation task from standard cross-lingual summarisation (i.e., modern to modern language), highlight the distinctness and value of our dataset, and demonstrate that our transfer learning approach outperforms standard cross-lingual benchmarks on this task. We release our code and data at <https://github.com/Pzoom522/HistSumm>.

5.1 HistSumm Corpus

As covered in Chapters 1 and 2, automatically summarising historical documents in a modern language can reduce the time and efforts needed to access the main points of text written hundreds of years ago, thus directly benefiting researcher in History, Archaeology, Digital Humanities, etc. Note that, the languages of the input and output of this task are the *same* - the primary difference is that they are in modern and historical forms respectively (e.g., modern Chinese and historical Chinese). To bootstrap this research direction, we created the HISTSUMM corpus, which can be used to evaluate historical text summarisation systems.

5.1.1 Dataset Construction

In history and digital humanities research, summarisation is most needed when analysing documentary and narrative text such as news, chronicles, diaries, and memoirs (South, 1977). Therefore, for DE we picked the GerManC dataset (Durrell et al., 2012), which contains Optical Character Recognition (OCR) results of DE newspapers from the years 1650–1800. We randomly selected 100 out of the 383 news stories for manual annotation. For ZH, we chose 『万历邸抄』 (*Wanli Gazette*) as the data source, a collection of news stories from the Wanli period of Ming Dynasty (1573–1620). However, there are no machine-readable versions of Wanli Gazette available; worse still, the calligraphy copies are unrecognisable even for non-expert humans, making the OCR technique inapplicable. Therefore, we performed a thorough literature search on over 200 related academic papers and manually retrieved 100 news texts¹.

The main challenge of historical text summarisation is that the performer needs to be able to both understand stories in the historical language (which can be quite obscure to the general public) and generate well-structured summaries in the respective modern language (which requires good writing skill). To this end, we recruited two experts with degrees in Germanistik and Ancient Chinese Literature, respectively. They were asked to produce summaries in the style of DE MLSUM (Scialom et al., 2020) and ZH LCSTS (Hu et al., 2015), whose news

¹Detailed references are included in the ‘source’ entries of ZH HISTSUMM’s metadata.

DE	Nº34
Story	Jhre Königl. Majest. befinden sich noch vnweit Thorn / ... / dahero zur Erledigung Hoffnung gemacht werden will. <i>(Their Royal Majesties are still not far from Thorn, ... , therefore completion of the hope is desired.)</i>
Summary	Der Krieg zwischen Polen und Schweden dauert an. Von einem Friedensvertrag ist noch nicht der Rede. <i>(The war between Poland and Sweden continues. There is still no talk on the peace treaty.)</i>
ZH	Nº7
Story	有脚夫小民，三四千名集众围绕马监丞衙门，...，冒火突入，捧出敕印。 <i>(Three to four thousand porters gathered around Majiancheng Yamen (a government office), ..., rushed into fire and salvaged the authority's seal.)</i>
Summary	小本生意免税条约未能落实，小商贩被严重剥削，以致百姓聚众闹事并火烧衙门，造成多人伤亡。王炆抢救出公章。 <i>(The tax-exemption act for small businesses was not well implemented and small traders were terribly exploited, leading to riot and arson attack on Yamen with many casualties. Yang Wang salvaged the authority's seal.)</i>

Table 5.1: Examples from our HISTSUMM dataset.

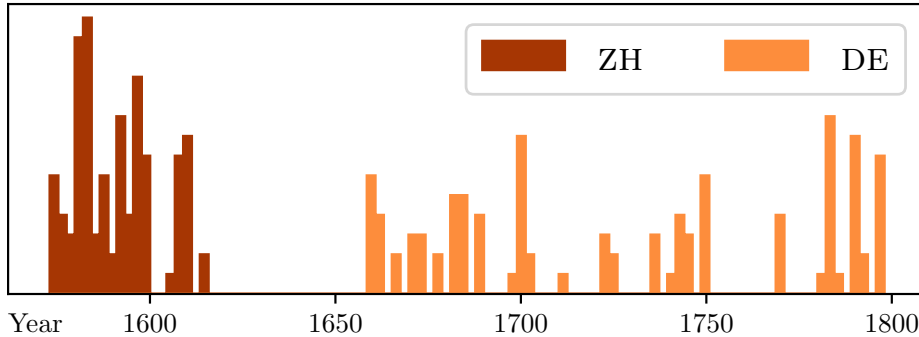


Figure 5.1: Publication time of HISTSUMM stories.

stories and summaries are crawled from the Süddeutsche Zeitung website and posts by professional media on the Sina Weibo platform, respectively. The annotation process turned out to be very effort-intensive: for both languages, the experts spent at least 20 minutes in reading and composing a summary for one single news story (they were paid 30 RMB for each summary). The accomplished corpus of 100 news stories and expert summaries in each language, namely HISTSUMM (see examples in Tab. 5.1), were further examined by six other experts for quality control (see details in § 5.4.2). Microsoft Office Excel was used to assign tasks and collect annotations throughout this study.

5.1.2 Dataset Statistics

Publication time. As visualised in Fig. 5.1, the publication time of DE and ZH HISTSUMM stories exhibits distinguished patterns. Oldness is an important indicator of the domain and linguistic gaps (Gunn, 2011). Considering news in ZH HISTSUMM is on average 137 years older than its DE counterpart, such gaps can be expected to be greater. On the other hand, DE HISTSUMM stories cover a period of 150 years, compared to just 47 years for ZH, indicating the potential for greater linguistic and cultural variation within the DE corpus.

Topic composition. For a high-level view of HISTSUMM’s content, we asked experts to manually classify all news stories into six categories (shown in Fig. 5.2). We see that the topic compositions of DE and ZH HISTSUMM share some similarities. For instance, Military (e.g., battle reports) and Politics (e.g., authorities’

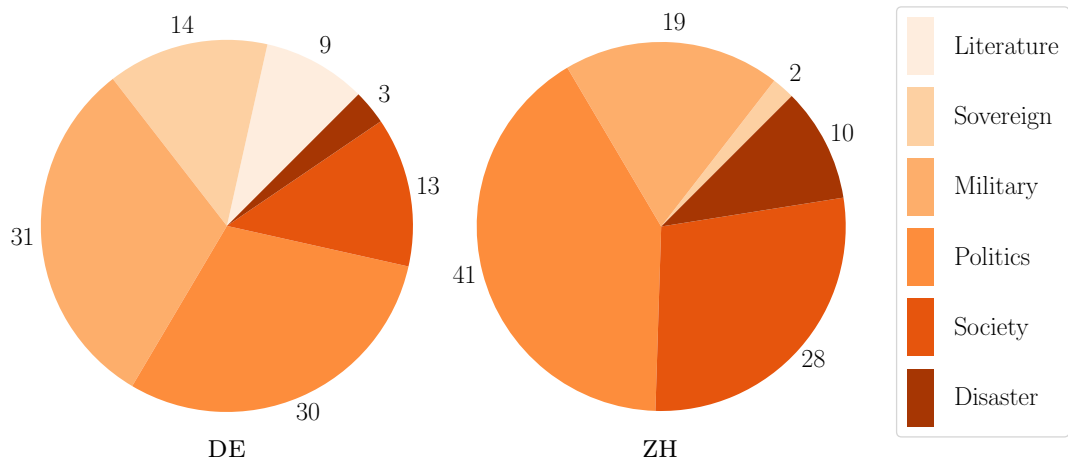


Figure 5.2: Topic composition of HISTSUMM.

	DE (word-level)		ZH (character-level)	
	HISTSUMM	MLSUM	HISTSUMM	LCSTS
L_{story}	268.1	570.6	114.5	102.5
L_{summ}	18.1	30.4	28.2	17.3
CR (%)	6.8	5.3	24.6	16.9

Table 5.2: Comparisons of mean story length (L_{story}), summary length (L_{summ}), and compression rate ($CR = L_{\text{summ}}/L_{\text{story}}$) for summarisation datasets.

policy and personnel changes) together account for more than half the stories in both languages. On the other hand, we also have language-specific observations. 9% DE stories are about Literature (e.g., news about book publications), but this topic is not seen in ZH HISTSUMM. And while 14% DE stories are about Sovereign (e.g., royal families and Holy See), there are only 2 examples in ZH (both about the emperor; we found no record on any religious leader in Wanli Gazette). Also, the topics of Society (e.g., social events and judicial decisions) and Natural Disaster (e.g., earthquakes, droughts, and floods) are more prevalent in the ZH dataset.

Story length. In news summarisation tasks, special attention is paid to the lengths of news stories and summaries (see Tab. 5.2). Comparing DE HISTSUMM with the corresponding modern corpus DE MLSUM, we find that although historical news stories are on average 53% shorter, the overall compression rate (CR s)

is quite similar (6.8% *vs* 5.8%), indicating that key points are summarised to similar extents. Following LCSTS (Hu et al., 2015), the table shows character-level data for ZH, but this is somewhat misleading. While most modern words are double-character, single-character words dominate the historical vocabulary, e.g., the historical word ‘朋’ (*friend*) becomes ‘朋友’ in modern ZH. According to Che et al. (2016), this leads to a character length ratio of approximately 1:1.6 between parallel historical and modern samples. Taking this into account, the *CRs* for the ZH HISTSUMM and LCSTS datasets are also quite similar to each other.

When contrasting DE with ZH (regardless of historical or modern), we notice that the compression rate is quite different. This might reflect stylistic variations with respect to how verbose news reports are in different languages or by different writers.

5.1.3 Vicissitudes of News

Compared with modern news, articles in HISTSUMM reveal several distinct characteristics with respect to writing style, posing new challenges for machine summarisation approaches.

Lexicon. With social and cultural changes over the centuries, lexical pragmatics of both languages have evolved substantially (Gunn, 2011). For DE, some routine concepts from hundreds of years ago are no longer in use today, e.g., the term ‘Brachmonat’ (№41), whose direct translation is *fallow month*, actually refers to *June* as the cultivation of fallow land traditionally begins in that month (Grimm, 1854). We observe a similar phenomenon in ZH HISTSUMM, e.g., ‘贡市’ (№24 and №31) used to refer to markets that were open to foreign merchants, but is no longer in use. For ZH, additionally, we notice that although some historical words are still in use, their semantics have changed over time, e.g., meaning of ‘闻’ has shifted from *hear* to *smell* (№53), and that of ‘走’ has changed from *run* to *walk* (№25).

Syntax. Another aspect of language change is that some historical syntax has been abandoned. Consider ‘daß derselbe noch länger allda/ biß der Frantz. Abgesandter von dannen widerum abreisen möge/ verbleiben soll’ (*the same should still remain there for longer, until the France Ambassador might leave again*) (№33). We find the subordinate clause is inserted within the main clause, whereas in modern DE it should be ‘daß derselbe noch länger allda verbleiben soll, biß der Frantz. Abgesandter von dannen widerum abreisen möge’. For ZH, inversion is common in historical texts but becomes rare in the modern language. For example, sentence ‘王氏之女成仙者’ (*Ms. Wang’s daughter who became a fairy*) (№65) where the attributive adjective is positioned after the head noun, should be ‘王氏之成仙 (的) 女’ according to modern ZH grammars. Also, we observe cases where historical ZH sentences without constituents such as subjects, predicates, objects, prepositions, etc. In these cases, contexts must be utilised to infer corresponding information, e.g., only by adding ‘居正’ (*Juzheng*, a minister’s name) to the context can we interpret the sentence ‘已, 又为私书安之云’ (№20) as ‘after that, (*Juzheng*) wrote a private letter to comfort him’. This adds extra difficulty to the generation of summaries.

Writing style. To inform readers, a popular practice adopted by modern news writers is to introduce key points in the first one or two sentences (White, 1998). Many machine summarisation algorithms leverage this pattern to enhance summarisation quality by incorporating positional signals (Edmundson, 1969; See et al., 2017; Gui et al., 2019). However, this rhetorical technique was not widely used in HISTSUMM, where crucial information may appear in the middle or even the end of stories. For instance, the keyword ‘Türck’ (*Turkish*) (№33) first occurs in the second half of the story; in article №7 of ZH HISTSUMM (see Tab. 5.1), only after reading the last sentence can we know the final outcome (i.e., the authority’s seal had been saved from fire).

5.2 Methodology

Based on the popular cross-lingual transfer learning framework (Li, 2021), we propose a simple historical text summarisation framework (see Fig. 5.3), which

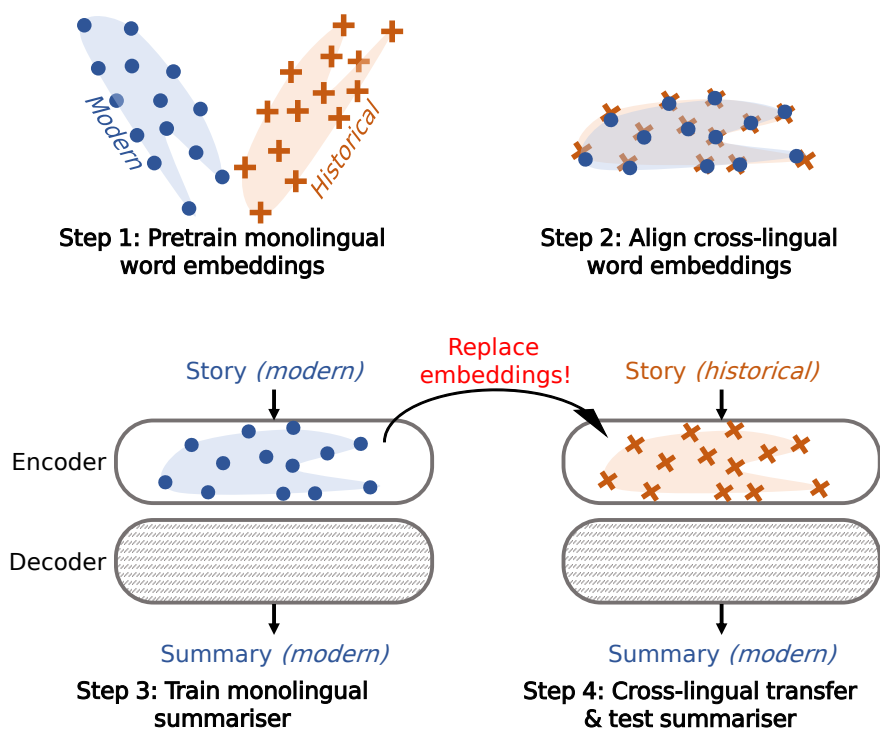


Figure 5.3: Illustration of our proposed framework.

can be trained even without supervision (i.e., parallel historical-modern signals).

Step 1. For both DE and ZH, we begin with respectively training modern and historical monolingual word embeddings. Specially, for DE, following the suggestions of Wang et al. (2019a), we selected subword-based embedding algorithms (e.g., FastText (Joulin et al., 2017a)) as they yield competitive results. In addition to training word embeddings on the raw text, for historical DE we also consider performing text normalisation (NORM) to enhance model performance. This orthographic technique aims to convert words from their historical spellings to modern ones, and has been widely adopted as a standard step by NLP applications for historical alphabetic languages (Bollmann, 2019). Although training a normalisation model in a fully unsupervised setup is not yet realistic, it can get bootstrapped with a single lexicon table to yield satisfactory performance (Ljubešić et al., 2016; Scherrer and Ljubešić, 2016).

For ideographic languages like ZH, word embeddings trained on stroke signals (which is analogous to subword information of alphabetic languages) achieve state-of-the-art performance (Cao et al., 2018), so we utilise them to obtain monolingual vectors. Compared with simplified characters (which dominate our training resources), traditional ones typically provide much richer stroke signals and thus benefit stroke-based embeddings (Chen and Sheng, 2018), e.g., traditional ‘葉’ (*leaf*) contains semantically related components of ‘艹’ (*plant*) and ‘木’ (*wood*), while its simplified version (‘叶’) does not.

Therefore, to improve the model performance we also conduct additional experiments on enhanced corpora which are converted to the traditional glyph using corresponding rules (CONV) (see § 5.3.3 for further details).

Step 2. Next, we respectively build two semantic spaces for DE and ZH, each of which is shared by historical and modern word vectors, using linear CLWE mappings. Given parallel supervision is very limited in real-world scenarios, we mainly consider two bootstrapping strategies: in a fully unsupervised (UspMap) style and through identical lexicon pairs (IdMap). While the former only relies on topological similarities between input vectors, the latter additionally takes advantage of words in the intersected vocabulary as seeds. Although their historical and current meanings can differ (cf. § 5.1.3), in most cases they are similar, providing very weak parallel signals (e.g., ‘Krieg’ (*war*) and ‘Frieden’ (*peace*) are common to historical and modern DE; ‘天’ (*universe*) and ‘人’ (*human*) to historical and modern ZH).

Step 3. In this step, for each of DE and ZH we use a large monolingual modern-language summarisation dataset to train a basic summariser that only takes modern-language inputs. Embedding weights of the encoder are initialised with the *modern* partition of corresponding cross-lingual word vectors in Step 2 and are frozen during the training process, while those of the decoder are randomly initialised and free to update through back-propagation.

Step 4. Upon convergence in the last step, we directly replace the embedding weights of the encoder with the *historical* vectors in the shared vector space,

yielding a new model that can be fed with historical inputs but output modern sentences. This entire process does not require any external parallel supervision.

5.3 Experimental Setup

5.3.1 Training Data

Consistent with § 5.1.1, we selected DE MLSUM and ZH LCSTS as monolingual summarisation training sets. For monolingual corpora for word embedding training, to minimise temporal and domainal variation, we only considered datasets that were similar to articles in MLSUM, LCSTS, and HISTSUMM, i.e. with text from comparable periods and centred around news-related domains.

For modern DE, such resources are easy to access: we directly downloaded the DE News Crawl Corpus released by WMT 2014 workshops (Bojar et al., 2014), which contains shuffled sentences from online news sites. We then conducted tokenisation and removed noise such as emojis and links. For historical DE, besides the already included GerManC corpus, we also saved Deutsches Textarchiv (Nolda, 2019), Mercurius-Baumbank (Ulrike, 2020), and Mannheimer Korpus (Mannheim, 2020) as training data. Articles in these datasets are all relevant to news and have topics such as Society and Politics. Note that we only preserved documents written in 1600 to 1800 to match the publication time of DE HISTSUMM stories (cf. § 5.1.2). Apart from the standard data cleaning procedures (tokenisation and noise removal, as mentioned above), for historical DE corpora we replaced the very common slash symbols (/) with their modern equivalents: commas (,) (Lindemann, 2015). We also lower-cased letters and deleted sentences with less than 10 words, yielding 505K sentences and 12M words in total.

For modern ZH, we further collected news articles in the corpora released by He (2018), Hua et al. (2018), and Xu et al. (2020a) to train better embeddings. For historical ZH, to the best of our knowledge, there is no standalone Ming Dynasty news collection except Wanli Gazette. Therefore, from the resources released by Jiang et al. (2020), we retrieved Ming Dynasty articles belonging to categories² of

²Following the topic taxonomy of Jiang et al. (2020).

Novel, History/Geography, and Military³. Raw historical ZH text does not have punctuation marks, so we first segmented sentences using the Jiayan Toolkit⁴. Although Jiayan supports tokenisation, we skipped this step as the accuracy is unsatisfactory. Given that a considerable amount of historical ZH words only have one character (cf. § 5.1.2 and § 5.1.3), following Li et al. (2018) we simply treated characters as basic units during training. Analogous to historical DE, we removed sentences with less than 10 characters. The remaining corpus has 992k sentences and 28M characters.

5.3.2 Baseline Approaches

In addition to the proposed method, we also consider two strong baselines based on the Cross-lingual Language Modelling paradigm (XLM) (Lample and Conneau, 2019), which has established state-of-the-art performance in the standard cross-lingual summarisation task (Cao et al., 2020). More concretely, for DE and ZH respectively, we pretrain baselines on all available historical and modern corpora using causal language modelling and masked language modelling tasks. Next, they are respectively fine-tuned on modern text summarisation and unsupervised machine translation tasks. The former becomes the (XLM-E2E) baseline, which can be directly executed on HISTSUMM in an end-to-end fashion; the latter (XLM-Pipe) is coupled with the basic summariser for modern inputs in Step 3 of § 5.2 to form a translate-then-summarise pipeline.

5.3.3 Model Configurations

Normalisation and convention. We normalised historical DE text using cSM-Tiser (Ljubešić et al., 2016; Scherrer and Ljubešić, 2016), which is based on character-level statistical machine translation. Following the original papers, we pretrained the normaliser using RIDGES corpus (Odebrecht et al., 2017). As for the ZH character convention, we utilised the popular OpenCC⁵ project which uses a hard-coded lexicon table to convert simplified input characters into their

³Sampling inspection confirmed that their domains are similar to those of Wanli Gazette.

⁴<https://github.com/jiaeyan/Jiayan>

⁵<https://github.com/BYVoid/OpenCC>

traditional forms.

Word embedding. As discussed in § 5.2, when training DE and ZH monolingual embeddings, we respectively ran subword-based FastText (Joulin et al., 2017a) and stroke-based Cw2Vec (Cao et al., 2018). For both languages, we set the dimension at 100 and learned embeddings for all available tokens (i.e., `minCount` = 1). Other hyperparameters followed the default configurations. After training, we preserved the most frequent 50K tokens in each vocabulary (NB: historical ZH only has 13K unique tokens). To obtain aligned spaces for modern and historical vectors, we then utilised the robust VecMap framework (Artetxe et al., 2018) with its original settings.

Summarisation model. We implemented our main model based on the robust Pointer-Generator Network (See et al., 2017), which is a hybrid framework for extractive (to copy source expressions via pointing) and abstractive (to produce novel words) summarisation models. After setting up the encoder and decoder (cf. in Step 3 of § 5.2), we started training with the default configurations. As for the two baselines which are quite heavyweight (XLM (Lample and Conneau, 2019) is based on BERT (Devlin et al., 2019) and has 250M valid parameters), we trained them from scratch with FP16 precision due to moderate computational power access. All other hyperparameter values followed the official XLM settings. To ensure the baselines can yield their highest possible performance, we trained them on the enhanced corpora, i.e., normalised DE (`NORM`) and converted ZH (`CONV`).

5.4 Results and Analyses

5.4.1 Automatic Evaluation

We assessed all models with the standard ROUGE metric (Lin, 2004), reporting F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. Following Hu et al. (2015), the ROUGE score of ZH outputs are calculated on character-level.

As shown in Tab. 5.3, for DE, our proposed methods are comparable to the baseline approaches or outperform the baselines by small amounts; for ZH, our

DE	ROUGE-1	ROUGE-2	ROUGE-L
XLM-Pipe	12.72	2.88	10.67
XLM-E2E	13.48	3.27	11.25
UspMap	13.36	3.02	11.28
UspMap+NORM	13.78	3.59	11.60
IdMap	13.45	3.10	11.38
IdMap+NORM	14.37	3.30	12.14
ZH			
XLM-Pipe	10.91	2.96	9.83
XLM-E2E	12.67	3.86	11.02
UspMap	13.09	4.25	11.31
UspMap+CONV	16.38	6.06	14.00
IdMap	18.38	7.05	15.89
IdMap+CONV	19.22	7.42	16.52

Table 5.3: ROUGE F1 scores (%) on HISTSUMM.

EN→ZH	ROUGE-1	ROUGE-2	ROUGE-L
XLM-Pipe	14.93	4.14	12.62
XLM-E2E	18.02	5.10	15.39
UspMap	11.43	1.27	10.07
IdMap	12.06	1.72	10.93
ZH→EN			
XLM-Pipe	9.08	3.29	7.43
XLM-E2E	12.97	4.31	10.95
UspMap	5.15	0.84	2.42
IdMap	5.98	1.33	2.90

Table 5.4: ROUGE F1 scores (%) of *standard* cross-lingual summarisation. Following Cao et al. (2020), for monolingual pretraining, we used corpora in § 5.3.3 (57M sentences) for modern ZH and annotated Gigaword (Napoles et al., 2012) (183M sentences) for EN; for summarisation training, we used LCSTS for EN→ZH and CNN/DM dataset (Hermann et al., 2015) for ZH→EN; for testing, we used the data released by Zhu et al. (2019).

models are superior by large margins. Given that XLM-based models require a lot more training resources than our model, we consider this a positive result. For comparison of the strengths and weaknesses of the models, we show their performance for a modern cross-lingual summarisation task in Tab. 5.4. To heighten

the contrast we chose two languages (ZH and EN) from different families and with minimal overlap of vocabulary. As shown in Tab. 5.4, the XLM-based models outperform our method on this modern language cross-lingual summarisation task by large margins.

The difference in the performance of models on the modern and historical summarisation tasks illustrate key differences in the tasks and also some of the shortcomings of the models. Firstly, the great temporal gap (up to 400 years for DE and 600 years for ZH) between our historical and modern data hurts the XLM paradigm, which relies heavily on the similarity between corpora (Kim et al., 2020). In addition, Kim et al. (2020) also show that inadequate monolingual data size (less than 1M sentences) is likely to lead to unsatisfactory performance of XLM, even for etymologically close language pairs such as EN-DE. In our experiments we only have 505K and 992K sentences for historical DE and ZH (cf. § 5.3.1). On the other hand, considering the negative influence of the error-propagation issue (cf. § 2.3.2), the poor performance of `XLM-Pipe` is not surprising and is in line with observations of Cao et al. (2020) and Zhu et al. (2020a). Our model instead makes use of cross-lingual embeddings, including bootstrapping from identical lexicon pairs. This approach helps overcome data sparsity issues for the historical summarisation tasks and is also successful at leveraging the similarities in the language pairs. However, its performance drops when the two languages are as far apart as EN and ZH.

When analysing the ablation results of the proposed method, on DE and ZH we found different trends. For DE, scores achieved by all the four setups show minor variance. To be specific, models bootstrapped with identical word pairs outperformed the unsupervised ones, and models trained on normalised data yielded stronger performance. Among all tested versions, `UspMap+NORM` got the best score in ROUGE-2 and `IdMap+NORM` led in ROUGE-1 and ROUGE-L, indicating that the normalisation enhancement does benefit DE historical text summarisation models. For ZH, as predicted, with richer glyph information encoded, the stroke-based embedding method can better learn word semantics. We find that `UspMap+CONV` outperforms `UspMap` and `IdMap+CONV` outperforms `IdMap`. Adding identical words during mapping initialisation brings substantial benefits too: 3.58% and 2.52% ROUGE-L improvement for `IdMap` over `UspMap` and

IdMap+CONV over UspMap+CONV, respectively.

5.4.2 Human Judgement

To gain further insights, we invited six experts to conduct human evaluations. Like the annotators in § 5.1.1, they also held degrees in Germanistik or Ancient Chinese Literature. Beyond the standard dimensions of summarisation evaluation (Informativeness, Conciseness, and Fluency), we added ‘Currentness’ as the fourth, which focuses on measuring ‘to what extent a summary follows current rather than early linguistic styles’. We used a five-point Likert scale, with 1 for worst and 5 for best. For each language, experts were only asked to rate the gold-standard human summary and the summaries generated by the XLM-E2E baseline and the best two setups in § 5.4.1. For each of the 100 news stories in each language, 3 experts independently each rated the three model outputs and the human summary. They were paid 5 RMB for each sample.

The final results are given in Tab. 5.5. When comparing different systems, we report statistical significance as the p -value of two-tailed t-tests with Bonferroni correction (Dror et al., 2018). We found that in all aspects the scores for the gold-standard summaries were always above 4 points, indicating the high quality of the gold-standard summaries. Across both languages, our models outperform the baseline for informativeness and conciseness ($p < 0.01$) and achieve comparable levels of fluency and currentness. Summaries generated by XLM-E2E were slightly more fluent than our approach for both DE and ZH ($p < 0.05$), indicating that the baseline has merit with respect to its language modelling abilities. However, it tended to make errors in understanding historical inputs and locating key points; e.g. the human reference for ZH article №57 is focused on the commander’s decision of bursting the river to beat the rebel army (‘宁夏之役中，魏学曾为了击溃叛乱部落，决定决河灌城’), but XLM-E2E summarises it as 黄河大堤水，比塔顶还高几丈’ (*the surface of the river is several feet higher than the tower top*), which is fluent but irrelevant.

As for different setups of the proposed algorithm, for DE, in dimensions of Informativeness, Conciseness and Fluency, the performance of UspMap+Norm and IdMap+NORM was almost equally good. The improvement from utilising identical

DE	Informativeness	Conciseness	Fluency	Currentness
Expert	4.85 (.08)	5.00 (.00)	4.94 (.03)	4.99 (.00)
XLM-E2E	2.26 (.20)	2.35 (.24)	3.34 (.19)	3.67 (.23)
UspMap+NORM	2.51 (.18)	2.53 (.22)	3.28 (.22)	3.64 (.24)
IdMap+NORM	2.52 (.18)	2.54 (.20)	3.32 (.28)	3.72 (.24)
ZH				
Expert	4.72 (.10)	4.98 (.01)	4.97 (.02)	4.90 (.04)
XLM-E2E	2.18 (.23)	2.21 (.27)	2.80 (.22)	2.53 (.23)
IdMap	2.39 (.19)	2.49 (.26)	2.66 (.25)	2.50 (.23)
IdMap+CONV	2.37 (.21)	2.57 (.28)	2.78 (.24)	2.59 (.25)

Table 5.5: Average human ratings on HISTSUMM (variance is in parentheses).

word pairs for CLWE mapping seems more evident for Currentness, i.e., the average score was 0.08 higher ($p < 0.05$). For ZH, while IdMap and IdMap+CONV achieved close Informativeness scores, the latter outperforms the former in other three aspects by 0.08, 0.12, and 0.09 respectively ($p < 0.01$). This observation indicates that when the lexical encoding is improved with enriched stroke-level information, the model is less likely to include redundant information in the summaries (i.e., conciseness score is higher), and the produced sentences are more fluent in terms of modern ZH grammars.

5.4.3 Error Analysis

We further analysed model inputs with the lowest scores in § 5.4.2, and found that they were mostly for stories whose content was dissimilar to *any* sample in modern training sets. For instance, five ZH texts in HISTSUMM are on themes not seen in modern news (i.e., witchcraft (№65), monsters (№35 and №46), and abnormal astromancy (№8 and №28)). On these texts, even the best-performing IdMap+CONV model outputs a large number of [UNK] tokens and can merely achieve average Informativeness, Conciseness, Fluency, and Correctness scores of 1.41, 1.67, 1.83, and 1.60 respectively, which are significantly below its overall results in Tab. 5.5. This reveals the current system’s shortcoming when processing inputs with theme-level *zero-shot* patterns. This issue is typically ignored in the cross-lingual summarisation literature due to the rarity of such cases in modern language tasks. However, we argue that a key contribution of our proposed task and dataset is that they together indicate new improvement directions beyond standard cross-lingual summarisation studies, such as the challenges of zero-shot generalisation and historical linguistic gaps (cf. § 5.1.3).

5.5 Summary and Discussion

We attack the problem of automatically summarising historical texts presented in its modern language variant, which is highly relevant for digital humanities in general and historical research in particular. As the first of this research strand, we carefully curated a high-quality corpus of historical news. This dataset in-

volves two structurally different languages, German and Chinese, and can serve as (part of a) standard for future studies. Moreover, we developed a pipeline using cross-lingual transfer learning. We also designed methods to further deal with the challenges of historical texts, including spelling variation, language change, and writing style shift over time. The work in this chapter not only offers bootstrapping resources (baselines, benchmarks, etc.) for a new task, but also demonstrates the applications of CLWE mappings beyond modern languages.

As analysed in § 5.4, despite our proposed CLWE-based method can generate good modern-language summaries for a few historical story samples, the overall scores of both automatic and human evaluations are pretty low (especially in the Chinese setups). Therefore, although the tested machinery can be used as an interesting demonstration that inspires follow-up studies on the task of historical text summarisation, we believe it is not ready to be employed as a practical tool so far. To improve the generation quality of CLWE-based algorithms, one potential direction is to include the tokenisation adaptation step (Pfeiffer et al., 2021) in the pipeline. As for the PLM-based schemes, to tackle the lexicon mismatch issue during transfer learning (see § 5.4.1), the idea of initialising their embeddings with CLWEs (Minixhofer et al., 2022) is also worth a visit.

5.6 Post-Publication Retrospect

Our work has motivated new efforts on CLWE algorithms (e.g., Sannigrahi and Read (2022)), cross-lingual summarisation systems (e.g., Jiang et al. (2022)), and Digital Humanities (e.g., Domingo and Casacuberta (2022)). Nevertheless, historical text summarisation is still a very challenging task that has no practical solution.

LEARNING KGE EFFICIENTLY BY MAPPING RELATIONAL MATRICES

KGEs have been intensively explored in recent years due to their promise for a wide range of applications. However, existing studies focus on improving the final model performance without acknowledging the computational cost of the proposed approaches, in terms of execution time and environmental impact. This chapter proposes a simple yet effective KGE framework which can reduce the training time and carbon footprint by orders of magnitudes compared with state-of-the-art approaches, while producing competitive performance. We highlight three technical innovations: full batch learning via relational matrices, closed-form Orthogonal Procrustes Analysis for KGEs, and non-negative-sampling training. In addition, as the first KGE method whose entity embeddings also store full relation information, our trained models encode rich semantics and are highly interpretable. Comprehensive experiments and ablation studies involving 13 strong baselines and two standard datasets verify the effectiveness and efficiency of our algorithm. The code is available at <https://github.com/Pzoom522/Procrustes-KGE>.

6.1 Methodology

We propose a highly efficient and lightweight method for training KGEs called PROCRUSTES¹, which is more efficient in terms of time consumption and CO₂ emissions than previous counterparts by orders of magnitude while retaining strong performance. This is achieved by introducing three novel optimisation strategies, namely, relational mini-batch, closed-form Orthogonal Procrustes Analysis, and non-negative sampling training.

6.1.1 Preliminaries: Segmented Embeddings

Our proposed PROCRUSTES model is built upon *segmented embeddings*, a technique which has been leveraged by a number of promising recent approaches to KGE learning (e.g., RotatE (Sun et al., 2019), SEEK (Xu et al., 2020b), and OTE (Tang et al., 2020)). In contrast to conventional methods for KGEs where each entity only corresponds to one single vector, algorithms adopting segmented embeddings explicitly divide the entity representation space into multiple independent sub-spaces. During training each entity is encoded as a concatenation of decoupled sub-vectors (i.e., different segments, and hence the name). For example, as shown in Fig. 6.1, to encode a graph with 7 entities, the embedding of the t th entity is the row-wise concatenation of its d/d_s sub-vectors (i.e., $e_{t,1} \hat{\wedge} e_{t,2} \hat{\wedge} \dots \hat{\wedge} e_{t,d/d_s}$), where d and d_s denote the dimensions of entity vectors and sub-vectors, respectively. Employing segmented embeddings permits parallel processing of the structurally separated sub-spaces, and hence significantly boosts the overall training speed. Furthermore, segmented embeddings can also enhance the overall expressiveness of our model, while substantially reducing the dimension of matrix calculations. We provide detailed discussion on the empirical influence of segmented embedding setups in § 6.2.4.

¹Following the tradition of naming a KGE method (e.g., TransE, ComplEx, and RotatE), we capitalised the letter “E” here.

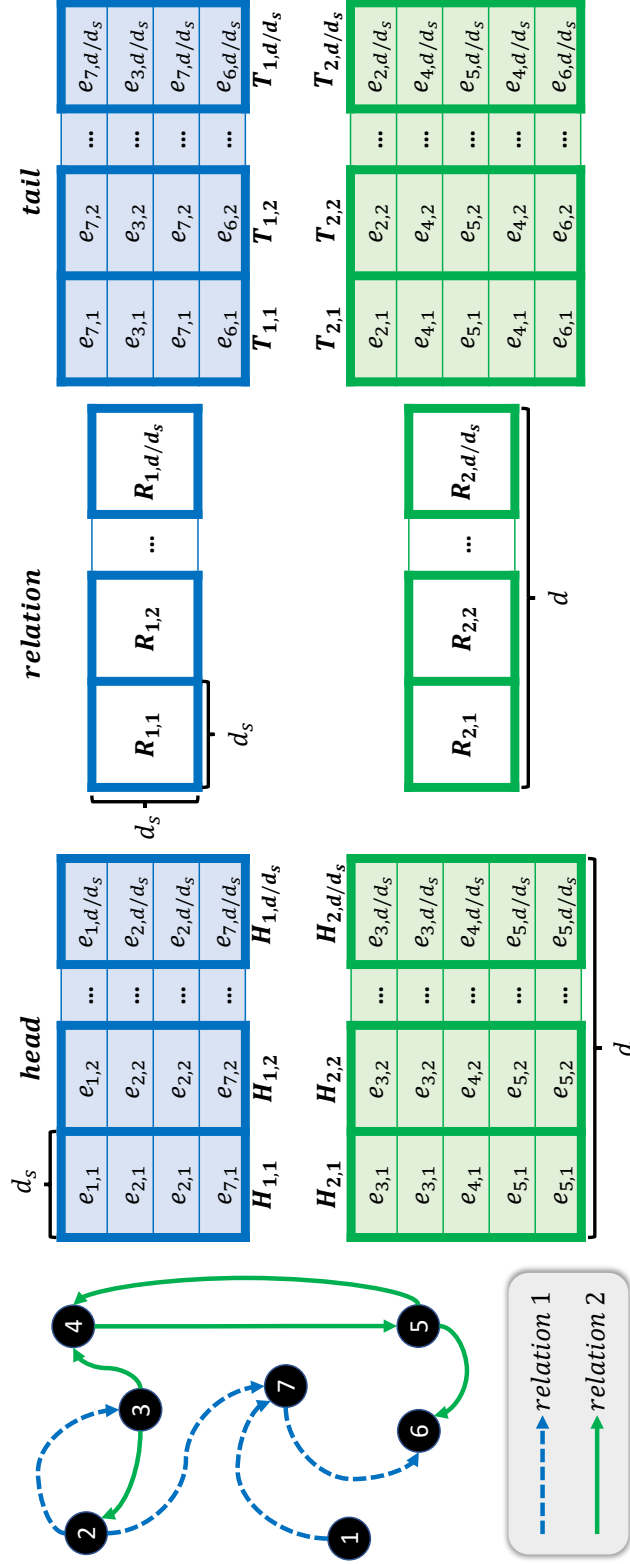


Figure 6.1: The by-relation partitioning architecture of PROGRUSTEs for a toy graph (left). Matrices involved in the computation of Eq. (6.1) are divided into two relational matrices: the **upper** is for *relation 1* (**dashed**) and the **lower** is for *relation 2* (**solid**).

6.1.2 Efficient KGE Optimisation

Full Batch Learning via Relational Matrices

Segmented embeddings can speed up training process by parallelising tuple-wise computation. In this section, we propose a full batch learning technique via relational matrices, which can optimise batch-wise computation to further reduce training time. This idea is motivated by the observation that previous neural KGE frameworks all perform training based on random batches constructed from tuples consisting of different types of relations (Bordes et al., 2013; Trouillon et al., 2016; Schlichtkrull et al., 2018; Chami et al., 2020; Huang et al., 2022). Such a training paradigm is based on random batches which, although straightforward to implement, is difficult to parallelise. This is due to the nature of computer process scheduling: during the interval between a process reading and updating the relation embeddings, they are likely to be modified by other processes, leading to synchronisation errors and consequently result in unintended data corruption, degraded optimisation, or even convergence issues.

To tackle this challenge, we propose to construct batches by grouping tuples which contain *the same relations*. The advantage of this novel strategy is two-fold. For one thing, it naturally reduces the original tuple-level computation to simple matrix-level arithmetic. For another and more importantly, we can then easily ensure that the embedding of each relation is only accessible by *one single process*. Such a training strategy completely avoids the data corruption issue. In addition, it makes the employment of the full batch learning technique (each batch covers all training samples) possible, which offers a robust solution for parallelising the KGEs training process and hence can greatly enhance the training speed. To the best of our knowledge, this approach has never been explored by the KGE community.

As illustrated in Fig. 6.1, we first separate the embedding space into segments (cf. § 6.1.1) and arrange batches based on relations. After that, for each training step, the workflow of PROCRUSTES is essentially decomposed into $m \times d/d_s$ parallel optimisation processes, where m is the number of relation types. Let i and j denote the indices of relation types and sub-spaces, respectively, then the column-wise concatenations of the j th sub-vectors of all tuples of i th relations

can be symbolised as $H_{i,j}$ (for head entities) and $T_{i,j}$ (for tail entities). Similarly, $R_{i,j}$ denotes the corresponding relation embedding matrix in the j th sub-space. The final objective function of PROCRUSTES becomes

$$\mathcal{L} = \sum_{i=1}^m \sum_{j=1}^{d/d_s} \|H_{i,j}R_{i,j} - T_{i,j}\|_2. \quad (6.1)$$

Orthogonal Procrustes Analysis

Our key optimisation objective, as formulated in Eq. (6.1), is to minimise the Euclidean distance between the head and tail matrices for each parallel process. In addition, following Sun et al. (2019) and Tang et al. (2020), we restrict the relation embedding matrix $R_{i,j}$ to be orthogonal throughout model training, which has been shown effective in improving KGE quality. Previous KGE models use different approaches to impose orthogonality. For instance, RotatE (Sun et al., 2019) takes advantage of a corollary of Euler’s identity and defines its relation embedding as

$$R_{i,j} = \begin{bmatrix} \cos \theta_{i,j} & \sin \theta_{i,j} \\ -\sin \theta_{i,j} & \cos \theta_{i,j} \end{bmatrix}, \quad (6.2)$$

which is controlled by a learnable parameter $\theta_{i,j}$. Although Eq. (6.2) holds orthogonality and retains simplicity, it is essentially a special case of segmented embedding where d_s equals 2. As a result, $R_{i,j}$ is always two-dimensional, which greatly limits the modelling capacity (see § 6.2.4 for discussion on the impact of dimensionality). To overcome this limitation, OTE (Tang et al., 2020) explicitly orthogonalises $R_{i,j}$ using the Gram-Schmidt algorithm per back-propagation step. However, while this scheme works well for a wide range of d_s (i.e., the dimension for the sub-vector), similar to RotatE, OTE finds a good model solution based on gradient descent, which is computationally very expensive.

We address the computational issue by proposing a highly efficient method utilising the proposed parallelism of full batch learning. With full batch learning, comparing with existing methods which deal with heterogeneous relations, PROCRUSTES only needs to optimise one single $R_{i,j}$ in each process, which becomes a simple constrained matrix regression task. More importantly, we can directly use

Eq. (2.2), the *closed-form* solution which has been widely adopted in the CLWE community (cf. § 2.1). During each iteration, PROCRUSTES can directly find the *globally* optimal embedding for each relation given the current entity embeddings by applying Eq. (2.2). Then, based on the calculated \mathcal{L} , PROCRUSTES updates entity embeddings through the back propagation mechanism (NB: the relation embeddings do not require gradients here). This process is repeated until convergence. As the optimisation of relation embeddings can be done almost instantly per iteration thanks to the closed-form Eq. (2.2), PROCRUSTES is significantly (orders of magnitude) faster than RotatE and OTE. In addition, compared with entity embeddings of all other KGE models which are updated separately with relation embedding, entity embeddings trained by PROCRUSTES can be used to restore relation embeddings directly (via Eq. (2.2)). In other words, PROCRUSTES can encode richer information in the entity space than its counterparts (see § 6.2.5).

Further Optimisation Schemes

As recently surveyed by Ruffinelli et al. (2020), existing KGE methods employ negative sampling as a standard technique for reducing training time, where update is performed only on a subset of parameters by calculating loss based on the generated negative samples. With our proposed closed-form solution (i.e., Eq. (2.2)), computing gradients to update embeddings is no longer an efficiency bottleneck for PROCRUSTES. Instead, the speed bottleneck turns out to be the extra bandwidth being occupied due to the added negative samples. Therefore, for PROCRUSTES, we do not employ negative sampling but rather update all embeddings during each round of back propagation with positive samples only, in order to further optimise the training speed.

We also discovered that if we do not apply any additional conditions during training, PROCRUSTES tends to fall into a trivial optimum after several updates, i.e., $\mathcal{L} = 0$, with all values in $H_{i,j}$, $T_{i,j}$ and $R_{i,j}$ being zero. In other words, the model collapses with nothing encoded at all. This is somewhat unsurprising as such trivial optima often yields large gradient and leads to this behaviour (Zhou et al., 2019b). To mitigate this degeneration issue, inspired by the geometric

	FB15k-237	WN18RR
Entities	14,541	40,943
Relations	237	11
Train samples	272,115	86,835
Validate samples	17,535	3,034
Test samples	20,466	3,134

Table 6.1: Basic statistics of the two benchmark datasets.

meaning of orthogonal $R_{i,j}$ (i.e., to rotate $H_{i,j}$ towards $T_{i,j}$ around the coordinate origin, without changing vector length) and popular practice in the CLWE studies (Artetxe et al., 2018; Vulić et al., 2019), we propose to constrain all entities to a high-dimensional *hypersphere* by performing two spherisation steps in every epoch. The first technique, namely *length normalisation*, ensures the row-wise Euclidean norm of $H_{i,j}$ and $T_{i,j}$ to always be one. This helps PROCRUSTES avoid the trivial optimum as $H_{i,j}$, $T_{i,j}$ and $R_{i,j}$ cannot be zero-matrices with a positive norm. The second operation is *centring*, which respectively translates $H_{i,j}$ and $T_{i,j}$ so that the column-wise sum of each matrix becomes a zero vector (note that each row denotes a sub-vector of an entity), otherwise, the relations may be modelled as affine transformations instead (similar to the scenario discussed in § 3.1) Our experiments validated that employing these two simple constraints effectively alleviates the trivial optimum issue (see § 6.2).

6.2 Experiment

6.2.1 Setups

We assess the performance of PROCRUSTES on the task of multi-relational link prediction, which is the *de facto* standard of KGE evaluation.

Datasets. In this study, following previous works (e.g., baselines in Tab. 6.2), we employ two benchmark datasets for link prediction: FB15K-237 (Toutanova and Chen, 2015), which consists of sub-graphs extracted from Freebase, and contains no inverse relations; and (2) WN18RR (Dettmers et al., 2018), which is

extracted from WordNet. Tab. 6.1 shows descriptive statistics for these two datasets, indicating that FB15K-237 is larger in size and has more types relations while WN18RR has more entities. We use the same training, validating, and testing splits as past studies.

Evaluation metrics. Consistent with Sun et al. (2019) and Tang et al. (2020), we report Hit Ratio with cut-off values $n = 1, 3, 10$ (i.e., H1, H3, and H10) and Mean Reciprocal Rank (MRR). Additionally, as to efficiency, we report the time cost and CO₂ emissions for each model, i.e., from the beginning of training until convergence.

Baselines. We compare PROCRUSTES to not only classical neural graph embedding methods, including TransE (Bordes et al., 2013), DistMulti (Yang et al., 2015), and ComplEx (Trouillon et al., 2016), but also embedding techniques recently reporting state-of-the-art performance on either WN18RR or FB15k-237, including R-GCN (Schlichtkrull et al., 2018), ConvE (Dettmers et al., 2018), A2N (Bansal et al., 2019), RotatE (Sun et al., 2019), SACN (Shang et al., 2019), TuckER (Balazevic et al., 2019), QuatE (Zhang et al., 2019), InteractE (Vashishth et al., 2020), OTE (Tang et al., 2020), and RotH (Chami et al., 2020). For all these baselines, we use the official code and published hyper-parameters to facilitate reproducibility.

Implementation details. All experiments are conducted on a workstation with one NVIDIA GTX 1080 Ti GPU and one Intel Core i9-9900K CPU, which is widely applicable to moderate industrial/academic environments. We use the Experiment Impact Tracker (Henderson et al., 2020) to benchmark the time and carbon footprint of training. To reduce measurement error, in each setup we fix the random seeds, run PROCRUSTES and all baselines for three times and reported the average.

The key hyper-parameters of our model is d and d_s , which are respectively set at 2K and 20 for both datasets. The detailed selection process is described in § 6.2.4. We train each model for a maximum of 2K epochs and check if the validation MRR stops increasing every 100 epochs after 100 epochs. For WN18RR

and FB15k-237 respectively, we report the best hyperparameters as fixed learning rates of 0.001 and 0.05 (Adam optimiser), and stopping epochs of 1K and 200.

6.2.2 Main Results

Tab. 6.2 reports the results of both our PROCRUSTES and all other 13 baselines on both WN18RR and FB15k-237 datasets. We analyse these results from two dimensions:

- **Effectiveness:** the model performance on link prediction task (MRR is our main indicator);
- **Efficiency:** system training time and carbon footprint (i.e., CO₂ emissions).

Regarding the performance on WN18RR, we found that PROCRUSTES performs as good as or even better than previous state-of-the-art approaches. To be concrete, out of all 13 baselines, it beats 11 in H10, (at least) 9 in H3 and 8 in MRR. The models outperformed by PROCRUSTES include not only all methods prior to 2019, but also several approaches published in 2019 or even 2020. Notably, when compared with the RotatE and OTE, two highly competitive methods which have similar architectures to PROCRUSTES (i.e., with segmented embeddings and orthogonal constraints), our PROCRUSTES can learn KGEs with higher quality (i.e., 0.014 and 0.005 higher in MRR, respectively). This evidences the effectiveness of the proposed approaches in § 6.1 in modelling knowledge tuples.

While PROCRUSTES achieves very competitive performance, it requires significantly less time for training: it converges in merely **14 minutes**, more than 100 times faster than strong-performing counterparts such as SACN. Moreover, it is very environmentally friendly: from bootstrapping to convergence, PROCRUSTES only emits **37g** of CO₂, which is even less than making two cups of coffee². On the contrary, the baselines emit on average 1469g and up to 5342g CO₂: the latter is even roughly equal to the carbon footprint of a coach ride from Los Angeles to San Diego³.

²<https://tinyurl.com/coffee-co2>

³<https://tinyurl.com/GHG-report-2019>







	WN18RR					FB15k-237						
	MRR	H1	H3	H10			MRR	H1	H3	H10		
TransE (2013)	.226	-	-	.501	85	367	.294	-	-	.465	96	370
DistMult (2015)	.430	.390	.440	.490	79	309	.241	.155	.263	.419	91	350
ComplEx (2016)	.440	.410	.460	.510	130	493	.247	.158	.275	.428	121	534
R-GCN (2018)	.417	.387	.442	.476	138	572	.248	.151	.264	.417	152	598
ConVE (2018)	.430	.400	.440	.520	840	3702	.325	.237	.356	.501	1007	4053
A2N (2019)	.450	.420	.460	.510	203	758	.317	.232	.348	.486	229	751
SACN (2019)	.470	.430	.480	.540	1539	5342	.352	.261	.385	.536	1128	4589
Tucker (2019)	.470	.443	.482	.526	173	686	.358	.266	.392	.544	184	704
Quate (2019)	.488	.438	.508	.582	176	880	.348	.248	.382	.550	180	945
InteractE (2020)	.463	.430	-	.528	254	1152	.354	.263	-	.535	267	1173
Roth (2020)	.496	.449	.514	.586	192	903	.344	.246	.380	.535	207	1120
RotatE (2019)	.439	.390	.456	.527	255	823	.297	.205	.328	.480	343	1006
OTE (2020)	.448	.402	.465	.531	304	1008	.309	.213	.337	.483	320	1144
PROCRUSTES (ours)	.453	.408	.491	.549	14	37	.295	.241	.310	.433	9	42
w/ NS (ours)	.457	.411	.494	.551	44	124	.302	.245	.333	.465	37	159
w/ TB (ours)	.468	.417	.498	.557	92	268	.326	.247	.354	.492	56	243
w/ NS+TB (ours)	.474	.421	.502	.569	131	346	.345	.249	.379	.541	85	285

Table 6.2: Model effectiveness and efficiency on link prediction benchmarks. : training time (*minutes*); : carbon dioxide production (*grams*). NS: negative sampling; TB: traditional batch. The performance results of baselines are coloured **heavily** and **lightly** if they are below those of PROCRUSTES and “w/ NS+TB”, respectively. State-of-the-art scores are in **bold**. Following Balazevic et al. (2019) and Zhang et al. (2019), for fair comparison, both RotatE and OTE results are reported with conventional negative sampling rather than the self-adversarial one.

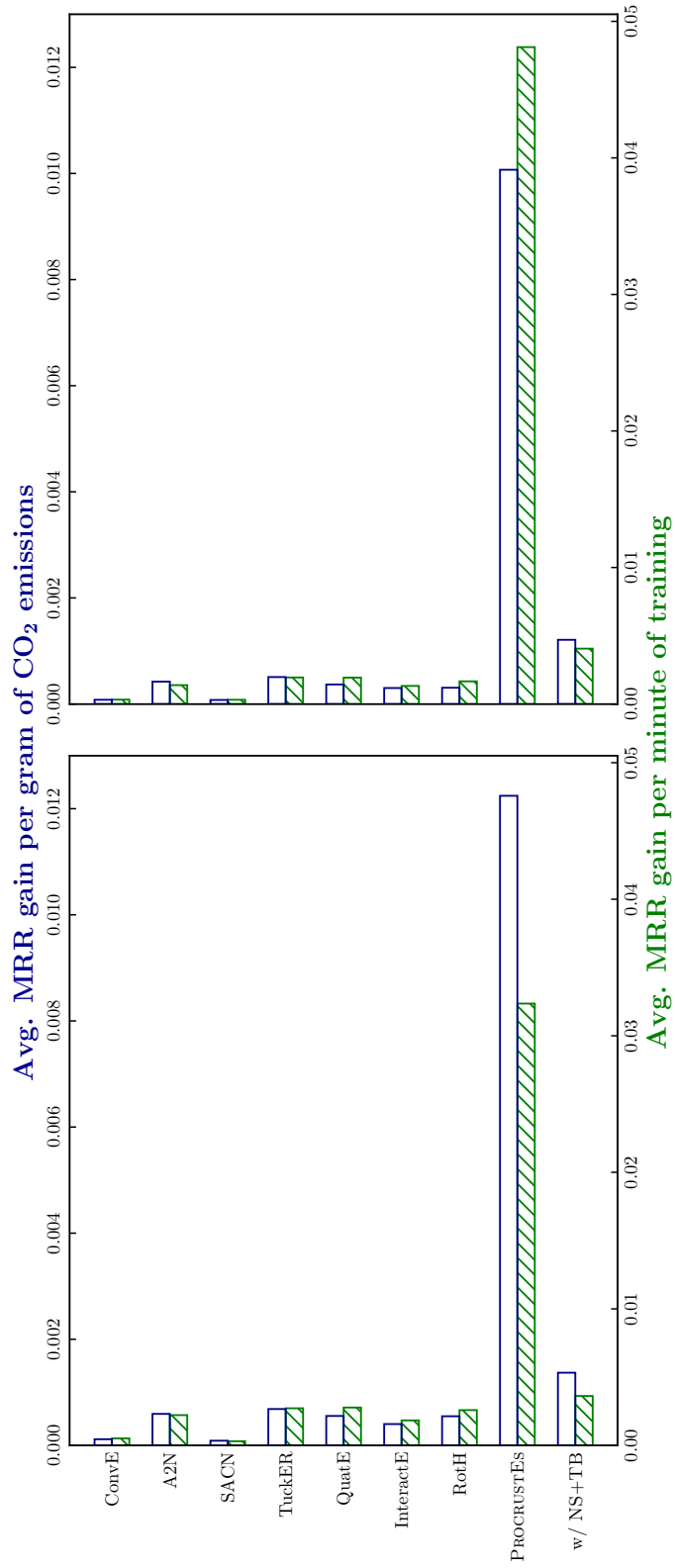


Figure 6.2: Unified effectiveness-efficiency comparison between most competitive KGE models in Tab. 6.2. The left and right sub-figures are respectively for WN18RR and FB15k-237.

As for the testing results on FB15k-237, we found that although PROCRUSTES seems less outstanding (we investigate the reasons in § 6.2.3), it still outperforms at least 7 more complex baselines in H1 and almost all models prior to 2019 in MRR. Furthermore, similar to the observation on WN18RR, it demonstrates great advantage in terms of efficiency. While all baselines need 91 to 1128 minutes to coverage with 350g to 4589g CO₂ produced, PROCRUSTES can learn embeddings of similar quality in just **9 minutes** and with **42g emissions**. By employing both traditional batch and negative sampling, we show that PROCRUSTES can achieve near-state-of-the-art performance on both datasets. We discuss this in detail in § 6.2.3.

To provide a unified comparisons between PROCRUSTES and the most strong-performing baselines on both effectiveness and efficiency, we further investigate the following question: How much performance gain can we obtain by spending *unit* time on training or making *unit* emissions? We did analysis by calculating MRR/(training time) and MRR/(carbon footprint) and the results are presented in Fig. 6.2. It is obvious that among all competitive KGE models, PROCRUSTES is the most economic algorithm in terms of performance-cost trade-off: it is *more than 20 times* more efficient than any past works, in terms of both performance per unit training time and per unit CO₂ emissions.

We also investigate baseline performance with a shorter training schedule. From scratch, we train RotH, the best performing algorithm on WN18RR, and stop the experiment when MRR reaches the performance of PROCRUSTES. On WN18RR, RotH takes 50 minutes ($3.6\times$ PROCRUSTES) and emits 211g CO₂ ($5.7\times$ PROCRUSTES); on FB15k-237 RotH takes 45 minutes ($5.0\times$ PROCRUSTES) and emits 218g CO₂ ($5.2\times$ PROCRUSTES). These results once again highlight the efficiency superiority of our approach.

To further ascertain the efficiency advantage of PROCRUSTES by ruling out factors such as numbers of all epochs, we pick four frameworks with strongest MRR performance and estimate their bandwidth during training, as illustrated in Fig. 6.3.

We can see that although some baselines have been engineered for enhanced computational efficiency, e.g., by default RotH creates 24 threads for multiprocessing, on both datasets they still substantially underperform PROCRUSTES with

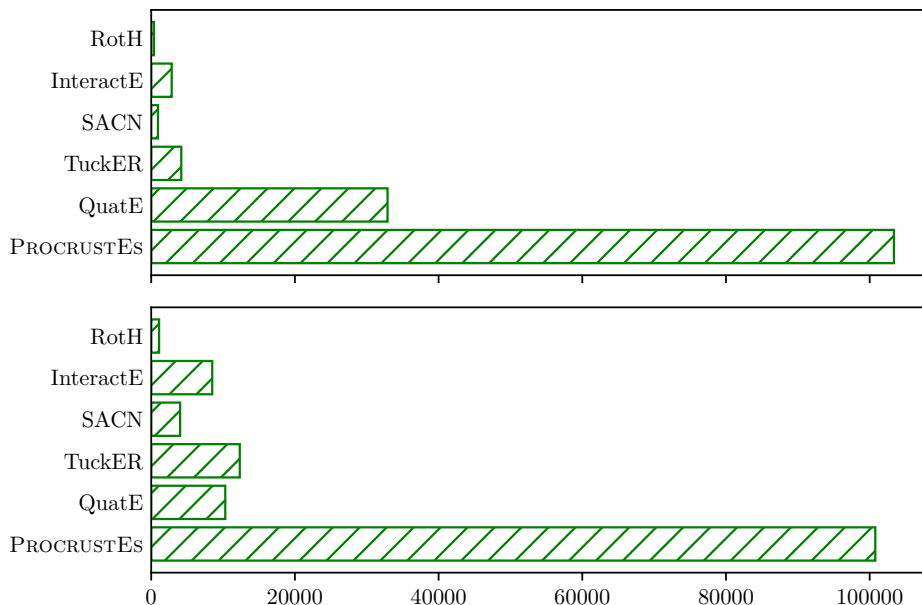


Figure 6.3: Comparison of bandwidth (number of processed samples per second). The upper and the lower are respectively for WN18RR and FB15k-237.

huge margins in terms of bandwidth.

6.2.3 Ablation Studies

To better understand the performance difference of PROCRUSTES on WN18RR and FB15k-237, we dive deeply into the dataset statistics in Tab. 6.1. Goyal et al. (2017) and Hoffer et al. (2017) found that although full batch learning can boost training speed and may benefit performance, when the data distribution is too sparse, it may be trapped into sharp minimum. As the average number of samples linked to each relation is significantly smaller for FB15k-237 than for WN18RR (1148 *vs* 7894), the distribution of the former is likely to be more sparse and the generalisability of PROCRUSTES may thus be harmed. For another, FB15k-237 has finer-grained relation types (237 *vs.* 11 of WN18RR), so intuitively the likelihood of tuples sharing similar relations rises. However, as PROCRUSTES omits negative sampling to trade for speed, sometimes it may be less discriminative for look-alike tuples.

To validate the above hypotheses, we additionally conduct ablation studies by

switching back to traditional batch mode and/or adding negative sampling modules⁴. Configurations where the closed-form optimisation, Eq. (2.2), is replaced by gradient descent, are omitted since the resulting architecture is very similar to OTE. As shown in the lower section of Tab. 6.2, both using either traditional or negative sampling (i.e., w/ NS and w/ TB) can improve the performance of PROCRUSTES for all metrics. For example, on WN18RR our approach (w/ NS+TB) outperforms most baselines and is close to the performance of QuatE and RotH, but thanks to the Orthogonal Procrustes Analysis, the computational cost of our approach is significantly less. Compared to WN18RR, the gain of our model on FB15k-237 by adopting negative sampling and traditional batch is even more significant, achieving near-state-of-the-art performance (i.e., compared to TuckER, the MRR is only 1.3% less with merely 4.9% of the computational time). These observations verify our aforementioned hypotheses. We also found out that traditional batch is more effective than negative sampling for PROCRUSTES in terms of improving model performance. On the other hand, however, adding these two techniques can reduce the original efficiency of PROCRUSTES to some extent.

Nevertheless, as Eq. (2.2) is not only fast but also energy-saving (as only basic matrix arithmetic on GPUs is involved), even PROCRUSTES with the “w/ NS+TB” configuration preserves great advantage in training time and carbon footprint. Moreover, it achieves near-state-of-the-art effectiveness on both datasets (cf. Tab. 6.2) and still exceeds strong baselines in training efficiency with large margins (cf. Fig. 6.2). One interesting observation is that, while the training time of RotH is merely $1.47\times$ of that of PROCRUSTES (w/ NS+TB), their emission levels are drastically different. This is because RotH implements 24-thread multiprocessing by default while our approach creates only one process. Within similar training time, methods like RotH will thus consume a lot more power and emit a lot more CO₂. Therefore, for effectiveness-intensive applications, we recommend training PROCRUSTES in transitional batches with negative sampling, as it can then yield cutting-edge performance without losing its eco-friendly fashion.

⁴Following Sun et al. (2019), we set the batch size at 1024 and the negative sample size at 128.

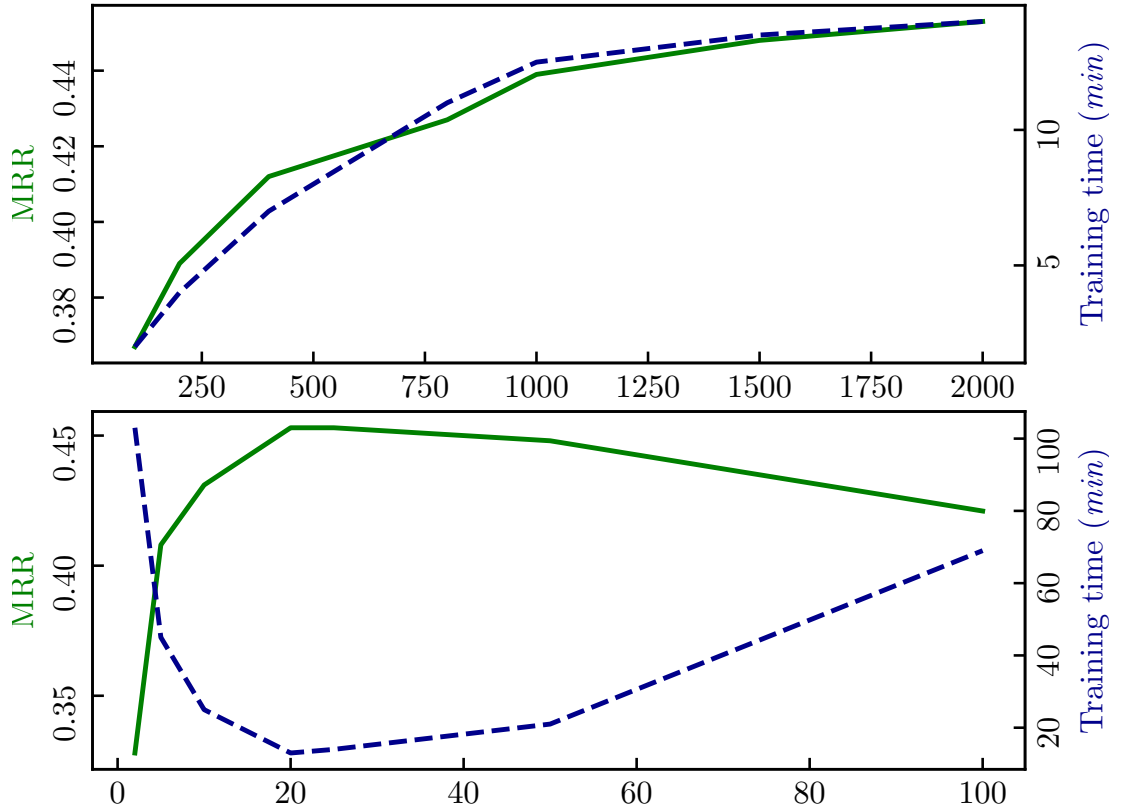


Figure 6.4: With different d (upper) and d_s (lower), the training time and convergence MRR of PROCRUSTES on WN18RR (results on FB15k-237 exhibit similar trends). X-axes denote dimensionality.

6.2.4 Impacts of Dimensionality

When configuring PROCRUSTES, we set d_s at 20 following the recommendation of Tang et al. (2020). As for d , intuitively the larger it is, the more capacious PROCRUSTES will be (which was later verified empirically), so we decide 2K as the best setting given the 11GB graphics memory limit of our hardware.

To further examine how the selection of these two dimensional hyperparameters influence the effectiveness and efficiency of PROCRUSTES, we first trained PROCRUSTES with $d \in \{100, 200, 400, 800, 1K, 1.5K, 2K\}$ and plotted results based on the validation set, as shown in Fig. 6.4. It is evident that with the increase of d , the model performance (indicated by MRR) grows but the training time also rises. Observing the curvature of training time almost saturates when

$d \geq 1K$, Then, for the dimension of sub-embeddings, we fixed d at 2K and enumerated $d_s \in \{2, 5, 10, 20, 25, 50, 100\}$. For algorithm performance, the pattern we witnessed is on par with that reported by Tang et al. (2020), i.e., before d_s reaches 20 or 25 the effectiveness jumps rapidly, but after that the model slowly degrades, as the learning capacity of the network reduces. Coincidentally, the training speed also climbs its peak when d_s is 20, making it indisputably become our optimal choice.

6.2.5 Interpreting Entity Embeddings

Building on the fact that PROCRUSTES fuses entity information and relation information (in other words, for a specific entity, the information of the entity itself and of its corresponding relations is encoded in a single vector), the location of a entity is more expressive and, thus, the related entity embedding is more interpretable. Picking up on that, we do visualisation study on the trained entity embeddings. To this end, we conduct dimension reduction on the embeddings using PCA, which reduces the dimensionality of an entity embedding from 2K to three⁵. Fig. 6.4 shows the visualisation result, from which we see a diagram with 6 “arms”. This is far distinct from the distributional topology of conventional semantic representations, e.g., word embeddings (Mikolov et al., 2013c).

In Fig. 6.4, we also list the representative entities that fall in some clusters on each arm. Each cluster is referred by an ID (from A1 to F2). When we zoom into this list, we observe something interesting:

- **First**, entities on the same arm are semantically similar, or, in other words, these entities belong to the same category. Concretely, entities on arm A are locations, those on arm B are biochemical terms, and those on arm C are military related entities. Entities on arm D, E, and F consists of entities refer to concepts of law, botany, and occupation, respectively.
- **Second**, significant differences exist between each cluster/position on a arm. One example is that, for arm A, A1 are entities for cities, such as *Stuttgart*, *Houston*, *Nanning*; A2 is about entities for rivers, mountains,

⁵We disable axes and grids for visualisation’s clarity.

- A1** chittagong, cartagena, pittsburgh_of_the_south, le_havre, nanning, stuttgart, kolkata, houston, windy_city, ...
- A2** yellowstone_river, atlas_mountains, san_fernando_valley, sambre_river, Nile_river, susquehanna_river, rhine_river, ...
- A3** sudan, balkanshe_alps, east_malaysia, lower_egypt, kalimantan, turkistan, tobago, lowlands_of_scotland, sicily, ...
- B1** mefoxin, metharbital, valium, amobarbital, procaine, nitrostat, tenormin, minor_tranquillizer, cancer_drug, ...
- B2** epinephrine, steroid_hormone, internal_secretion, alkaloid, gallamine, prolactin, luteinizing_hormone, ...
- C1** military_formation, retreat, tactics, strategic_warning, peacekeeping_operation, unauthorized_absence, ...
- C2** commando, sailor_boy, outpost, saddam's_martyrs, military_advisor, battlewagon, commander, ...
- D** plaintiff, remittance, franchise, summons, false_pretens, suspect, amnesty, legal_principle, disclaimer, affidavit, ...
- E** genus_ambrosia, gloxinia, saintpaulia, genus_cestrum, genus_eriophyllum, valerianella, genus_chrysopsis, ...
- F** moneyer, teacher, researcher, president, prime_minister, wheeler_dealer, house_servant, victualler, burglar, ...

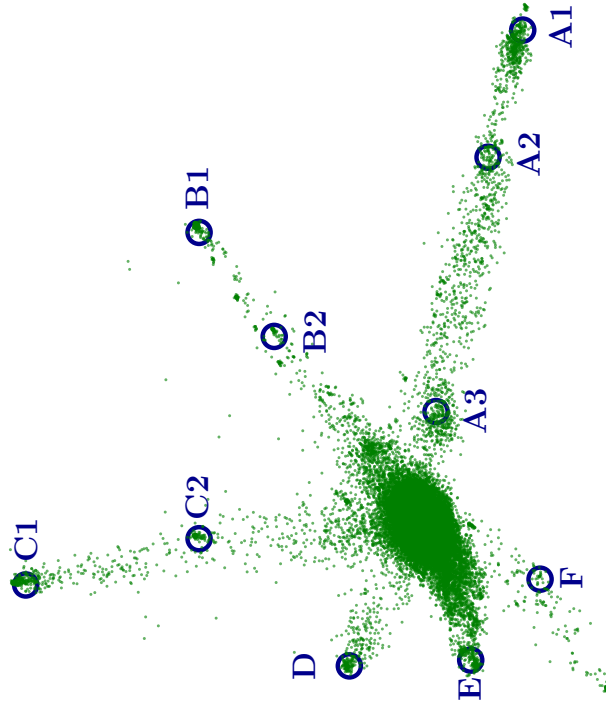


Figure 6.5: 3D PCA visualisation of PROCRUSTES entity embeddings for WN18RR.

etc.; and A3 contains entities referring to countries or regions. Similarly, while B1 mainly consists of medicine names, entities in B2 obviously relate to chemical terms.

- **Third**, PROCRUSTES can also put the “nick name” of a entity into the correct corresponding cluster. For example, *Windy City* (i.e., Chicago) and *Pittsburgh of the South* (i.e., Birmingham) were successfully recognised as names for cities.

6.3 Summary and Discussion

We provided a efficient KGE training framework in this chapter. As having been proven in experiments, it emits less greenhouse gases and therefore, has less negative environmental repercussions than any other KGE approaches. Additionally, rather than encoding relations in their own vector space, our method projects them in the space of entities implicitly, which improves both expressiveness and interpretability of the learned embeddings.

6.4 Post-Publication Retrospect

Besides serving as a major baseline in a list of follow-up studies (e.g., Li et al. (2021b), Li et al. (2022), and Wang et al. (2022d)), ideas of PROCRUSTES have directly contributed to the development of more advanced KGE methods (e.g., Li et al. (2021a) and Wang et al. (2022a)). Beyond the construction of KGEs, new KGE evaluation metrics were also motivated by PROCRUSTES, e.g., Bastos et al. (2023).

CONCLUSIONS

This thesis describes research on linear-mapping-based encoding methods for multilingual and relational signals, namely CLWE and KGE. The main outputs include state-of-the-art algorithms, new resources such as corpora and metrics, as well as novel insights. This chapter summarises the key contributions throughout the thesis and suggests exciting directions for future exploration.

7.1 Summary of Thesis

Chapter 2 began with introducing cross-lingual embedding methods, among which CLWE is one of the most commonly used. Specifically, it reviewed the debate on the linearity assumption made by some previous work on CLWE and the wide adoption of ℓ_2 refinement, which are relevant to Chapters 3 and 4. Next, the chapter described studies on two representative categories of relational encodings: analogies learned by word vectors and factual information stored by KGEs. To motivate Chapter 6, we analysed the efficiency limitations of existing methods. Lastly, we discussed literature on applying NLP techniques to historical text and the research topic of cross-lingual summarisation, which are both related to the task of historical text summarisation in Chapter 5.

Chapter 3 made the first attempt to explore the conditions under which CLWE mappings are linear. Theoretically, we show that this widely-adopted assumption holds *iff* the analogies encoded are preserved across embeddings for different

languages. We describe the construction of a novel cross-lingual word analogy dataset for a diverse range of languages and analogy categories and propose indicators to quantify linearity and analogy preservation. Experimental results on three distinct embedding series firmly support our hypothesis. We also demonstrate how our insight into the connection between linearity and analogy preservation can be used to better understand past observations about the limitations of linear CLWE mappings, particularly when they are ineffective. Our findings regarding the preservation of analogy encoding provide a test that can be applied to determine the likely success of any attempt to create linear mappings between multilingual embeddings.

Chapter 4 proposed a generic post-processing technique to enhance CLWE performance based on optimising ℓ_1 loss. This algorithm is motivated by successful applications in other research fields (e.g. computer vision and data mining) which exploit the ℓ_1 norm cost function since it has been shown to be more robust to noisy data than the commonly-adopted ℓ_2 loss. The approach was evaluated using ten diverse languages and word embeddings from different domains on the popular BLI benchmark, as well as a downstream task of cross-lingual transfer for NLI. Results demonstrated that our algorithm can significantly improve the quality of CLWEs in both supervised and unsupervised setups. It is therefore recommended that this straightforward technique be applied to improve performance of CLWEs.

Chapter 5 introduced the new task of summarising historical documents in modern languages, a previously unexplored but important application of cross-lingual summarisation that can support historians and digital humanities researchers. To facilitate future research on this topic, we constructed the first summarisation corpus for historical news in DE and ZH with the support of linguistic experts. We also proposed a transfer learning method that makes effective use of similarities between languages and therefore requires limited or even zero parallel supervision. Our automatic and human evaluations demonstrated the strengths of our method over state-of-the-art baselines. This is the first study of automated historical text summarisation.

Chapter 6 was motivated by the closed-form Procrustes Analysis widely used in the field of CLWE. In this chapter, we proposed a novel KGE training frame-

work, namely PROCRUSTES, which is eco-friendly, time-efficient and can yield very competitive or even near-state-of-the-art performance. Experiments show that our method is valuable especially considering its significant and substantial reduction on training time and carbon footprint.

7.2 Evaluation of Thesis Goals

This thesis answers the five research questions raised in Chapter 1.

- **When does the linear mapping make an appropriate approximating function for encoding complex signals such as multilingual lexicons?**

This question is answered in Chapter 3. Using theoretical and empirical evidence, we justified that the similarity between multilingual word embeddings (i.e., CLWE mapping is linear) *iff* analogy encodings in monolingual embeddings are preserved.

- **How to improve the embedding precision in difficult scenarios?**

This question is answered in Chapter 4. We discovered that the conventional ℓ_2 refinement tends to be sensitive against outliers, commonly seen in the alignment between polysemous or rare words. We, therefore, proposed an optimisation algorithm based on the ℓ_1 loss that substantially outperforms strong baselines in extensive experiments.

- **Where can the linear-mapping-based embedding methods be applied beyond existing usages?**

This question is answered in Chapter 5. Beyond the conventional modern-modern alignment, we introduced CLWE technique to align modern words with historical ones and designed workarounds to new challenges. We composed the first historical text summarisation dataset to verify the model's effectiveness.

- **Why the embedding approaches are computationally expensive even with the optimisation simplification by the linear mapping?**

This question is answered in Chapter 6. We identified sample-centred parameter updating, negative sampling, and random batch training as the efficiency bottlenecks in KGE training, which motivated us to propose PROCRUSTES that speeds up KGE training by orders of magnitude.

- **Whether the embedding model for one type of signal can motivate that for another?**

This question is answered in both Chapters 3 and 6. On the one hand, we show that the linearity of CLWE mapping is related to the encoding of the analogy relation. On the other hand, the optimisation approach of CLWE mappings can motivate the development of KGE.

7.3 Future Directions

Firstly, the main insight of Chapter 3, i.e., the relationship between analogy encoding and cross-lingual representation learning, can motivate the design of new training objectives (e.g., to complete multilingual analogies) that improve cross-lingual models. Besides, as discussed in § 3.4, it can also be applied to evaluate the goodness of cross-lingual representation systems in the follow-up studies. **Secondly**, the strategy of replacing ℓ_2 -based objectives with ℓ_1 -based ones in Chapter 4 can be extended to cross-lingual PLMs, especially when more efficient solvers are available. **Thirdly**, as analysed in § 5.5, future works can employ more advanced PLM techniques to attack the historical text summarisation problem proposed in Chapter 4, so as to develop more practical tools for the research of Digital Humanities. **Lastly**, it is interesting to explore whether KGEs bootstrapped by the lightweight PROCRUSTES (proposed in Chapter 6) can be further enhanced by more effective but slower methods (e.g., OTE), so as to boost the trade off between efficiency and performance.

In addition to these concrete ideas, there are also several more general but very exciting avenues for future research, including:

- **Pre-trained cross-lingual and relational language modelling.** Although large-scale pre-trained language models require massive resources during training, they have proved remarkably successful on many tasks,

demonstrating impressive performance and outstanding scalability (Bender et al., 2021). Very recently, attempts have been made to extend PLMs by enhancing them with knowledge graphs (Petroni et al., 2019; Yu et al., 2022b) or/and training them in multilingual setups (Conneau et al., 2020; Zhou et al., 2022). However, most of these studies only consider single-hop triples or fail to incorporate cross-lingual signals into factual reasoning. This thesis has demonstrated the existence of strong connections between cross-lingual and relational encodings, which could contribute to the development of language models that can use both signals simultaneously.

- **Multi-modal cross-lingual and relational encoding.** While this thesis considers setups with only text and graph data, adding new modalities will likely improve the model’s performance. For instance, video information can be language-independent, which may help the model understand multilingual text. Likewise, modern Knowledge Graphs often contain numerous images, which can benefit relational embeddings. Very recent methods have achieved early-stage success (Zheng et al., 2021; Singh et al., 2022), but to our knowledge, there still lacks unified models that encode both multilingual and relational signals with multi-modal input.
- **Temporal signal processing.** Experiments in Chapter 5 highlight how temporal shifts in data lead to challenges when training cross-lingual models. Similarly, relational information may also be dynamic, adding difficulties to relational encodings. Therefore, it is useful to explore how to mitigate the gaps when training embeddings for data from a large period and how to keep already trained embeddings up to date. Recent studies on temporal modelling methods have demonstrated the great potential of this direction, but they also report challenges such as inferior precision against strong static counterparts (Sadeghian et al., 2021; Wang et al., 2022b,c).

BIBLIOGRAPHY

- Henrik Aanæs, Rune Fisker, Kalle Åström, and Jens Michael Carstensen. Robust factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1215–1225, 2002. ISSN 0162-8828. doi: 10.1109/TPAMI.2002.1033213.
- Carl Allen and Timothy M. Hospedales. Analogies explained: Towards understanding word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 223–231. PMLR, 2019. URL <http://proceedings.mlr.press/v97/allen19a.html>.
- David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL <https://www.aclweb.org/anthology/D18-1214>.
- Chinatsu Aone and Douglas McKee. A language-independent anaphora resolution system for understanding multilingual texts. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 156–163, Columbus, Ohio, USA, June 1993. Association for Computational Linguistics. doi: 10.3115/981574.981595. URL <https://aclanthology.org/P93-1021>.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A

- latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl.a.00106. URL <https://www.aclweb.org/anthology/Q16-1028>.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1250. URL <https://www.aclweb.org/anthology/D16-1250>.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL <https://www.aclweb.org/anthology/P18-1073>.
- Kevin D. Ashley. *Arguing by Analogy in Law: A Case-Based Model*, pages 205–224. Springer Netherlands, Dordrecht, 1988. ISBN 978-94-015-7811-0. doi: 10.1007/978-94-015-7811-0_10. URL https://doi.org/10.1007/978-94-015-7811-0_10.
- Ivana Balazevic, Carl Allen, and Timothy Hospedales. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1522. URL <https://www.aclweb.org/anthology/D19-1522>.
- Trapit Bansal, Da-Cheng Juan, Sujith Ravi, and Andrew McCallum. A2N: Attending to neighbors for knowledge graph inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4387–4392, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1431. URL <https://www.aclweb.org/anthology/P19-1431>.

- Anson Bastos, Kuldeep Singh, Abhishek Nadgeri, Johannes Hoffart, Toyotaro Suzumura, and Manish Singh. Can persistent homology provide an efficient alternative for evaluation of knowledge graph completion methods? In *Proceedings of the ACM Web Conference 2023*, WWW '23. Association for Computing Machinery, 2023.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL <https://doi.org/10.1145/3442188.3445922>.
- Alan R Benenfeld. Generation and encoding of the project intrex augmented catalog data base. *Clinic on Library Applications of Data Processing (6th: 1968)*, 1968.
- Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3302. URL <https://www.aclweb.org/anthology/W14-3302>.
- Marcel Bollmann. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1389. URL <https://www.aclweb.org/anthology/N19-1389>.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on*

- Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.
- Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 2787–2795, Lake Tahoe, Nevada, United States, 2013. URL <https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html>.
- J. Paul Brooks and Sapan Jot. pcaL1 : An implementation in r of three methods for ℓ_1 -norm principal component analysis. In *Optimization Online preprint*, 2013.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL <https://aclanthology.org/J93-2003>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- Tomáš Brychcín, Stephen Taylor, and Lukáš Svoboda. Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Systems with Applications*, 135:287 – 295, 2019. ISSN 0957-4174. doi: <https://doi.org/>

10.1016/j.eswa.2019.06.021. URL <http://www.sciencedirect.com/science/article/pii/S0957417419304191>.

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. Cw2Vec: Learning Chinese word embeddings with stroke n-grams. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*. AAAI, 2018.

Yue Cao, Hui Liu, and Xiaojun Wan. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.554. URL <https://www.aclweb.org/anthology/2020.acl-main.554>.

Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, August 2019. URL <https://crscardellino.github.io/SBWCE/>.

David J. Chalmers, Robert M. French, and Douglas R. Hofstadter. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4 (3):185–211, 1992. doi: 10.1080/09528139208953747.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6901–6914, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.617. URL <https://www.aclweb.org/anthology/2020.acl-main.617>.

Chao Che, Wenwen Guo, and Jianxin Zhang. Sentence alignment method based on maximum entropy model using anchor sentences. In Maosong Sun, Xuanjing Huang, Hongfei Lin, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 76–85, Cham, 2016. Springer International Publishing. ISBN 978-3-319-47674-2.

- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1152. URL <https://www.aclweb.org/anthology/P17-1152>.
- Wenfan Chen and Weiguo Sheng. A hybrid learning scheme for Chinese word embedding. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 84–90, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3011. URL <https://www.aclweb.org/anthology/W18-3011>.
- Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL <https://aclanthology.org/2021.naacl-main.280>.
- Moody T Chu and Nickolay T Trendafilov. The orthogonally constrained regression revisited. *Journal of Computational and Graphical Statistics*, pages 746–771, 2001.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL <https://www.aclweb.org/anthology/D18-1269>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettle-

- moyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1221–1225, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL <https://aclanthology.org/L16-1194>.
- Dan Cuperman and Igor M. Verner. Fostering analogical reasoning through creating robotic models of biological systems. *Journal of Science Education and Technology*, 28(2):pp. 90–103, 2019. ISSN 10590145, 15731839. URL <https://www.jstor.org/stable/48699947>.
- Fernando De La Torre and Michael J Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 2003.
- Wim De Smet, Jie Tang, and Marie-Francine Moens. Knowledge transfer across multilingual corpora via latent topics. In Joshua Zhexue Huang, Longbing Cao, and Jaideep Srivastava, editors, *Advances in Knowledge Discovery and Data Mining*, pages 549–560, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-20841-6.
- Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press, 2018. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17366>.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://www.aclweb.org/anthology/N19-1423>.
- Miguel Domingo and Francisco Casacuberta. An interactive machine translation framework for modernizing the language of historical documents. In Armando J. Pinho, Petia Georgieva, Luís F. Teixeira, and Joan Andreu Sánchez, editors, *Pattern Recognition and Image Analysis*, pages 41–53, Cham, 2022. Springer International Publishing. ISBN 978-3-031-04881-4.
- Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmam, Shaohua Sun, and Wei Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani, editors, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610. ACM, 2014. doi: 10.1145/2623330.2623623. URL <https://doi.org/10.1145/2623330.2623623>.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/P18-1128>.
- Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL <https://www.aclweb.org/anthology/C16-1332>.

- Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1305. URL <https://www.aclweb.org/anthology/P19-1305>.
- Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2377–2390, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.186. URL <https://aclanthology.org/2020.emnlp-main.186>.
- Martin Durrell, Paul Bennett, Silke Scheible, and Richard J. Whitt. GermanC, 2012. URL <http://hdl.handle.net/20.500.12024/2544>. Oxford Text Archive.
- Harold P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285, April 1969. ISSN 0004-5411. doi: 10.1145/321510.321519. URL <https://doi.org/10.1145/321510.321519>.
- Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1315. URL <https://www.aclweb.org/anthology/P19-1315>.
- Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL <https://www.aclweb.org/anthology/E14-1049>.

- Zihao Feng, Hailong Cao, Tiejun Zhao, Weixuan Wang, and Wei Peng. Cross-lingual feature extraction from monolingual corpora for low-resource unsupervised bilingual lexicon induction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5278–5287, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.469>.
- Ronald A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925. URL <http://psychclassics.yorku.ca/Fisher/Methods/>.
- Louis Fournier and Ewan Dunbar. Paraphrases do not explain word analogies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2129–2134, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.182. URL <https://aclanthology.org/2021.eacl-main.182>.
- Ashwinkumar Ganesan, Francis Ferraro, and Tim Oates. Learning a reversible embedding mapping using bi-directional manifold alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3132–3139, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.276. URL <https://aclanthology.org/2021.findings-acl.276>.
- Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulić, and Anders Søgaard. Analogy training multilingual encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12884–12892, May 2021. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17524>.
- John H Gennari, Mark A Musen, Ray W Ferguson, William E Grosso, Monica Crubézy, Henrik Eriksson, Natalya F Noy, and Samson W Tu. The evolution of protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies*, 58(1):89–123, 2003.
- Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. ISSN 0364-0213. doi: <https://doi.org/>

10.1016/S0364-0213(83)80009-3. URL <https://www.sciencedirect.com/science/article/pii/S0364021383800093>.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. Skip-Gram - Zipf + Uniform = Vector Additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1007. URL <https://www.aclweb.org/anthology/P17-1007>.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2002. URL <https://aclanthology.org/N16-2002>.

Goran Glavaš and Ivan Vulić. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.675. URL <https://aclanthology.org/2020.acl-main.675>.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL <https://www.aclweb.org/anthology/P19-1070>.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training Imagenet in 1 hour. In *International Conference on Learning Representations*, 2017.

- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://www.aclweb.org/anthology/L18-1550>.
- Russell D Gray, Quentin D Atkinson, and Simon J Greenhill. Language evolution and human history: what a difference a date makes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2011.
- Jacob Grimm. *Deutsches Wörterbuch*. Hirzel, 1854.
- Min Gui, Junfeng Tian, Rui Wang, and Zhenglu Yang. Attention optimization for abstractive document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1222–1228, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1117. URL <https://www.aclweb.org/anthology/D19-1117>.
- Simon Gunn. *Research Methods for History*. Research Methods for the Arts and Humanities Series. Edinburgh University Press, 2011. ISBN 9780748654048. URL <https://books.google.co.uk/books?id=HZJvAAAAQBAJ>.
- William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1057. URL <https://www.aclweb.org/anthology/D16-1057>.
- Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. OpenKE: An open toolkit for knowledge embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 139–144, Brussels, Belgium, 2018. Association

for Computational Linguistics. doi: 10.18653/v1/D18-2024. URL <https://www.aclweb.org/anthology/D18-2024>.

Tatsunori B. Hashimoto, David Alvarez-Melis, and Tommi S. Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286, 2016. doi: 10.1162/tacl.a_00098. URL <https://aclanthology.org/Q16-1020>.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1162. URL <https://www.aclweb.org/anthology/P17-1162>.

Zhengfang He. Chinese short text summarization dataset, 2018. URL <https://tinyurl.com/yy5sheep>.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning, 2020.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc., 2015. URL <http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf>.

Theo Hermans. Norms and the determination of translation: A theoretical framework. In *Translation, Power, Subversion*. Multilingual Matters, 1996.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. Data filtering using cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–172, Online, June 2021.

- Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.15. URL <https://aclanthology.org/2021.naacl-main.15>.
- Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1890–1902, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1188. URL <https://www.aclweb.org/anthology/N19-1188>.
- Hou Ieong Brent Ho and Hilde De Weerd. MARKUS. Text Analysis and Reading Platform, 2014. URL <http://dh.chinese-empires.eu/beta/>. Funded by the European Research Council and the Digging into Data Challenge.
- David C Hoaglin, Boris Iglewicz, and John W Tukey. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 1986.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.
- Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1731–1741, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/a5e0ff62be0b08456fc7f1e88812af3d-Abstract.html>.
- John H. Holland, Keith J. Holyoak, Richard E. Nisbett, Paul R. Thagard, and Stephen W. Smoliar. Induction: Processes of inference, learning, and discovery. *IEEE Expert*, 2(3):92–93, 1987. doi: 10.1109/MEX.1987.4307100.

- Keith J. Holyoak and Paul Thagard. *Mental Leaps: Analogy in Creative Thought*. The MIT Press, 12 1994. ISBN 9780262275620. doi: 10.7551/mitpress/4549.001.0001. URL <https://doi.org/10.7551/mitpress/4549.001.0001>.
- Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1043. URL <https://www.aclweb.org/anthology/D18-1043>.
- Baotian Hu, Qingcai Chen, and Fangze Zhu. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1229. URL <https://www.aclweb.org/anthology/D15-1229>.
- Lifeng Hua, Xiaojun Wan, and Lei Li. Overview of the NLPCC 2017 shared task: Single document summarization. In Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 942–947, Cham, 2018. Springer International Publishing.
- Hongren Huang, Chen Li, Xutan Peng, Lifang He, Shu Guo, Hao Peng, Lihong Wang, and Jianxin Li. Cross-knowledge-graph entity alignment via relation prediction. *Knowledge-Based Systems*, 240:107813, 2022. ISSN 0950-7051. doi: <https://doi.org/10.1016/j.knosys.2021.107813>. URL <https://www.sciencedirect.com/science/article/pii/S095070512101011X>.
- Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1067. URL <https://www.aclweb.org/anthology/P15-1067>.

- Shuyu Jiang, Dengbiao Tu, Xingshu Chen, Rui Tang, Wenxian Wang, and Haizhou Wang. Cptgraphsum: Let key clues guide the cross-lingual abstractive summarization. *arXiv preprint arXiv:2203.02797*, 2022.
- Yan-ting Jiang, Yu-ting Pan, and Le Yang. A research on verbal classifiers collocation in pre-modern Chinese based on statistics and word embedding. In *Journal of Xihua University (Philosophy & Social Sciences)*, 2020. URL https://github.com/JiangYanting/Pre-modern_Chinese_corpus_dataset.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017a. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/E17-2068>.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Maximilian Nickel, and Tomáš Mikolov. Fast linear model for knowledge graph embeddings. In *6th Workshop on Automated Knowledge Base Construction, AKBC@NIPS 2017*. OpenReview.net, 2017b.
- Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1330. URL <https://www.aclweb.org/anthology/D18-1330>.
- André Juthe. Argument by analogy. *Argumentation*, 19(1):1–27, 2005. doi: 10.1007/s10503-005-2314-9.

- O. Kariv and S. L. Hakimi. An algorithmic approach to network location problems. II: The P-Medians. *SIAM Journal on Applied Mathematics*, 37(3):539–560, 1979. doi: 10.1137/0137041. URL <https://doi.org/10.1137/0137041>.
- Yunsu Kim, Miguel Graça, and Hermann Ney. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL <https://www.aclweb.org/anthology/2020.eamt-1.5>.
- Kiril Kiryazov, Georgi Petkov, Maurice Grinberg, Boicho Kokinov, and Christian Balkenius. The interplay of analogy-making with active vision and motor control in anticipatory robots. In Martin V. Butz, Olivier Sigaud, Giovanni Pezzulo, and Gianluca Baldassarre, editors, *Anticipatory Behavior in Adaptive Learning Systems*, pages 233–253, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74262-3.
- Martin Kleiber and W. J. Pervin. A generalized Banach-Mazur theorem. *Bulletin of the Australian Mathematical Society*, 1(2):169–173, 1969. doi: 10.1017/S0004972700041411.
- Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. Multilingual reliability and “semantic” structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 40–45, London, UK, April 2015. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W15-0105>.
- Gennady Yu Kulikov. Cheap global error estimation in some Runge–Kutta pairs. *IMA Journal of Numerical Analysis*, 2013.
- Nojun Kwak. Principal component analysis based on ℓ_1 -norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.
- Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018a. URL <https://openreview.net/forum?id=rkYTTf-AZ>.
- Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018b. URL <https://openreview.net/forum?id=H196sainb>.
- Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. Cross-lingual c*st*rd: English access to Hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3): 245–269, 2003.
- Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan, June 2014a. Association for Computational Linguistics. doi: 10.3115/v1/W14-1618. URL <https://www.aclweb.org/anthology/W14-1618>.
- Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014b. URL <https://proceedings.neurips.cc/paper/2014/file/feab05aa91085b7a8012516bc3533958-Paper.pdf>.
- Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. doi: 10.1162/tacl_a.00134. URL <https://aclanthology.org/Q15-1016>.
- Charles N Li and Sandra A Thompson. *Mandarin Chinese: A functional reference grammar*, volume 3. Univ of California Press, 1989.

- Chen Li, Xutan Peng, Shanghang Zhang, Hao Peng, Philip S. Yu, Min He, Linfeng Du, and Lihong Wang. Modeling relation paths for knowledge base completion via joint adversarial training. *Knowledge-Based Systems*, 201-202:105865, 2020. ISSN 0950-7051. doi: <https://doi.org/10.1016/j.knosys.2020.105865>. URL <https://www.sciencedirect.com/science/article/pii/S0950705120302264>.
- Chen Li, Xutan Peng, Yuhang Niu, Shanghang Zhang, Hao Peng, Chuan Zhou, and Jianxin Li. Learning graph attention-aware knowledge graph embedding. *Neurocomputing*, 461:516–529, 2021a. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2021.01.139>. URL <https://www.sciencedirect.com/science/article/pii/S0925231221010961>.
- Ren Li, Yanan Cao, Qiannan Zhu, Xiaoxue Li, and Fang Fang. Is there more pattern in knowledge graph? exploring proximity pattern for knowledge graph embedding. *arXiv preprint arXiv:2110.00720*, 2021b.
- Ren Li, Yanan Cao, Qiannan Zhu, Guanqun Bi, Fang Fang, Yi Liu, and Qian Li. How does knowledge graph embedding extrapolate to unseen data: A semantic evidence view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5781–5791, Jun. 2022. doi: 10.1609/aaai.v36i5.20521. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20521>.
- Ruizhe Li, Xutan Peng, Chenghua Lin, Wenge Rong, and Zhigang Chen. On the low-density latent regions of vae-based language models. In Luca Bertinetto, João F. Henriques, Samuel Albanie, Michela Paganini, and Gül Varol, editors, *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, volume 148 of *Proceedings of Machine Learning Research*, pages 343–357. PMLR, 11 Dec 2021c. URL <https://proceedings.mlr.press/v148/li21a.html>.
- Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/P18-2023>.

- Yuling Li, Kui Yu, and Yuhong Zhang. Learning cross-lingual mappings in imperfectly isomorphic embedding spaces. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2630–2642, 2021d. doi: 10.1109/TASLP.2021.3097935.
- Zheng Li. *Neural Knowledge Transfer for Low-Source Sentiment Analysis: Cross-Domain, Cross-Task & Cross-Lingual*. Hong Kong University of Science and Technology (Hong Kong), 2021.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W04-1013>.
- Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2181–2187. AAAI Press, 2015. URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9571>.
- Mary Lindemann. A Ruler’s Consort in Early Modern Germany: Aemilia Juliana of Schwarzburg-Rudolstadt. *German History*, 33(2):291–292, 04 2015. ISSN 0266-3554. doi: 10.1093/gerhis/ghv012. URL <https://doi.org/10.1093/gerhis/ghv012>.
- Tal Linzen. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2503. URL <https://www.aclweb.org/anthology/W16-2503>.
- Michael L. Littman, Susan T. Dumais, and Thomas K. Landauer. *Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing*, pages 51–62. Springer US, Boston, MA, 1998. ISBN 978-1-4615-5661-9. doi: 10.1007/

978-1-4615-5661-9_5. URL https://doi.org/10.1007/978-1-4615-5661-9_5.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 146–155, 2016.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. Aligning vector-spaces with noisy supervised lexicon. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 460–465, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1045. URL <https://www.aclweb.org/anthology/N19-1045>.

Chen Luo, William Headden, Neela Avudaiappan, Haoming Jiang, Tianyu Cao, Qingyu Yin, Yifan Gao, Zheng Li, Rahul Goutam, Haiyang Zhang, and Bing Yin. Query attribute recommendation at amazon search. In *Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22*, page 506–508, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3547395. URL <https://doi.org/10.1145/3523227.3547395>.

IDS Mannheim. Mannheimer Korpus Historischer Zeitungen und Zeitschriften (Version 1.0), 2020. URL <https://doi.org/10.34644/laudatio-dev-miUsD3MB7CArCQ9C6Cu1>. Institut für Deutsche Sprache Mannheim.

Richard E. Mayer. *Multimedia Learning*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511811678.

Laurent Miclet, Sabri Bayoudh, and Arnaud Delhay. Analogical dissimilarity. *Journal of Artificial Intelligence Research*, 32(1):793–824, aug 2008. ISSN 1076-9757.

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013a. URL <https://openreview.net/forum?id=idpCd0WtqXd60>.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013b. URL <http://arxiv.org/abs/1309.4168>.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, pages 3111–3119, USA, 2013c. Curran Associates Inc. URL <http://dl.acm.org/citation.cfm?id=2999792.2999959>.
- Benjamin Minixhofer, Fabian Paischer, and Navid Rekasaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.293. URL <https://aclanthology.org/2022.naacl-main.293>.
- Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. LNMap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2712–2723, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.215. URL <https://aclanthology.org/2020.emnlp-main.215>.
- Ndapa Nakashole. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium, October-

November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1047. URL <https://www.aclweb.org/anthology/D18-1047>.

Ndapa Nakashole and Raphael Flauger. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 221–227, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2036. URL <https://www.aclweb.org/anthology/P18-2036>.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction, AKBC-WEKEX '12*, page 95–100, USA, 2012. Association for Computational Linguistics.

J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308. URL <https://doi.org/10.1093/comjnl/7.4.308>.

Tan Ngoc Le and Fatiha Sadat. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.410. URL <https://aclanthology.org/2020.coling-main.410>.

Andreas Nolda. Deutsches Textarchiv (1600–1900), 2019. URL <http://hdl.handle.net/21.11120/0000-0005-0ABA-F>. Berlin-Brandenburg Academy of Sciences and Humanities (BBAW).

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 03 2003. ISSN 0891-2017. doi: 10.1162/089120103321337421. URL <https://doi.org/10.1162/089120103321337421>.

- Carolin Odebrecht, Malte Belz, Amir Zeldes, Anke Lüdeling, and Thomas Krause. Ridges herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation*, 51(3):695–725, 2017.
- Juri Opitz, Leo Born, and Vivi Nastase. Induction of a large-scale knowledge graph from the Regesta Imperii. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 159–168, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W18-4518>.
- Constantin Orăsan and Oana Andreea Chiorean. Evaluation of a cross-lingual Romanian-English multi-document summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, May 2008.
- Csaba Oravecz, Bálint Sass, and Eszter Simon. Semi-automatic normalization of old Hungarian codices. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, pages 55–59, 2010.
- Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. Beyond offline mapping: Learning cross-lingual word embeddings through context anchoring. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6479–6489, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.506. URL <https://aclanthology.org/2021.acl-long.506>.
- Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1018. URL <https://www.aclweb.org/anthology/P19-1018>.

- Xutan Peng, Guanyi Chen, Chenghua Lin, and Mark Stevenson. Highly efficient knowledge graph embedding learning with Orthogonal Procrustes Analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2364–2375, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.187. URL <https://aclanthology.org/2021.naacl-main.187>.
- Xutan Peng, Chenghua Lin, and Mark Stevenson. Cross-lingual word embedding refinement by ℓ_1 norm optimisation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2690–2701, Online, June 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.214. URL <https://aclanthology.org/2021.naacl-main.214>.
- Xutan Peng, Yi Zheng, Chenghua Lin, and Advait Siddharthan. Summarising historical text in modern languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3123–3142, Online, April 2021c. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2021.eacl-main.273>.
- Xutan Peng, Mark Stevenson, Chenghua Lin, and Chen Li. Understanding linearity of cross-lingual word embedding mappings. *Transactions on Machine Learning Research*, 2022. URL <https://openreview.net/forum?id=8HuyXvbvqX>.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL <https://www.aclweb.org/anthology/D14-1162>.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Lan-*

- guage Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL <https://aclanthology.org/D19-1250>.
- Eva Pettersson, J. Lindström, B. Jacobsson, and Rosemarie Fiebranz. Histsearch - implementation and evaluation of a web-based tool for automatic information extraction from historical text. In *HistoInformatics@DH*, 2016.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL <https://aclanthology.org/2021.emnlp-main.800>.
- Michael Piotrowski. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157, 2012.
- Henri Prade and Gilles Richard. Analogical proportions: Why they are useful in ai. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4568–4576. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/621. URL <https://doi.org/10.24963/ijcai.2021/621>. Survey Track.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL <https://aclanthology.org/2020.acl-main.647>.
- Paul Rayson, Dawn E Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the bard: Evaluating the accuracy of a modern POS tagger

- on early modern english corpora. In *Proceedings of the Corpus Linguistics conference: CL2007*, 2007.
- Anna Rogers, Aleksandr Drozd, and Bofang Li. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-1017. URL <https://aclanthology.org/S17-1017>.
- Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Martinato, and Paolo Merialdo. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data*, 15(2), January 2021. ISSN 1556-4681. doi: 10.1145/3424672. URL <https://doi.org/10.1145/3424672>.
- Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., USA, 1987. ISBN 0471852333.
- Aynat Rubinstein. Historical corpora meet the digital humanities: the Jerusalem Corpus of emergent modern Hebrew. *Language Resources and Evaluation*, 53(4):807–835, 2019.
- Sebastian Ruder, Ryan Cotterell, Yova Kementchedjheva, and Anders Søgaard. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 458–468, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1042. URL <https://www.aclweb.org/anthology/D18-1042>.
- Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1): 569–630, May 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL <https://doi.org/10.1613/jair.1.11640>.
- Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *8th International*

- Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=BkxSmlBFvr>.
- Ali Sadeghian, Mohammadreza Armandpour, Anthony Colas, and Daisy Zhe Wang. Chronor: Rotation based temporal knowledge graph embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6471–6479, May 2021. doi: 10.1609/aaai.v35i7.16802. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16802>.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*, 2019.
- Sonal Sannigrahi and Jesse Read. Isomorphic cross-lingual embeddings for low-resource languages. In *Workshop on Representation Learning for NLP*, 2022.
- Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.412. URL <https://www.aclweb.org/anthology/2020.acl-main.412>.
- Yves Scherrer and Nikola Ljubešić. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 248–255, 2016.
- Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*. Springer, 2018.
- Guus Schreiber, Bob Wielinga, and Joost Breuker. *KADS: A principled approach to knowledge-based system development*, volume 11. Academic Press, 1993.

- Peter Schönemann. A generalized solution of the Orthogonal Procrustes Problem. *Psychometrika*, 1966. URL <https://EconPapers.repec.org/RePEc:spr:psycho:v:31:y:1966:i:1:p:1-10>.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. MLSUM: The multilingual summarization corpus, 2020.
- Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL <https://www.aclweb.org/anthology/P17-1099>.
- Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3060–3067. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33013060. URL <https://doi.org/10.1609/aaai.v33i01.33013060>.
- Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June 2022.
- Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=r1Aab85gg>.

- Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL <https://www.aclweb.org/anthology/P18-1072>.
- Stanley A. South. *Method and Theory in Historical Archeology*. Institute for Research on Poverty Monograph Series. Academic Press, 1977. ISBN 9780126557503. URL <https://books.google.co.uk/books?id=ktmjAAAAIAAJ>.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL <https://www.aclweb.org/anthology/P19-1355>.
- Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.204. URL <https://aclanthology.org/2021.eacl-main.204>.
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=HkgEQnRqYQ>.
- Marcin Sydow, Krzysztof Ciesielski, and Jakub Wajda. Introducing diversity to log-based query suggestions to deal with underspecified user queries. In *International Joint Conferences on Security and Intelligent Information Systems*, pages 251–264. Springer, 2011.

- Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2713–2722, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.241. URL <https://www.aclweb.org/anthology/2020.acl-main.241>.
- Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4007. URL <https://www.aclweb.org/anthology/W15-4007>.
- Nickolay T. Trendafilov. On the ℓ_1 Procrustes problem. *Future Generation Computer Systems*, 2003. ISSN 0167-739X. doi: [https://doi.org/10.1016/S0167-739X\(03\)00043-8](https://doi.org/10.1016/S0167-739X(03)00043-8). URL <http://www.sciencedirect.com/science/article/pii/S0167739X03000438>. Selected papers on Theoretical and Computational Aspects of Structural Dynamical Systems in Linear Algebra and Control.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2071–2080. JMLR.org, 2016.
- Hsieh-Chang Tu, Jieh Hsiang, I-Mei Hung, and Chijui Hu. Docusky, a personal digital humanities platform for scholars. *Journal of Chinese History*, 4(2): 564–580, 2020. doi: 10.1017/jch.2020.28.
- Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. Multilingual culture-independent word analogy datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.501>.

- Demske Ulrike. Mercurius-Baumbank (Version 1.1), 2020. URL <https://doi.org/10.34644/laudatio-dev-VyQiCnMB7CArCQ9CjF30>. Universität Potsdam.
- Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1157. URL <https://www.aclweb.org/anthology/P16-1157>.
- Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.280. URL <https://aclanthology.org/2021.acl-long.280>.
- Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, Nilesh Agrawal, and Partha Talukdar. Interact: Improving convolution-based knowledge graph embeddings by increasing feature interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- Ivan Vulić and Anna Korhonen. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1024. URL <https://www.aclweb.org/anthology/P16-1024>.
- Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China, 2019. Asso-

ciation for Computational Linguistics. doi: 10.18653/v1/D19-1449. URL <https://www.aclweb.org/anthology/D19-1449>.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.586. URL <https://aclanthology.org/2020.emnlp-main.586>.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.257. URL <https://aclanthology.org/2020.emnlp-main.257>.

Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti, and Anna Korhonen. Probing cross-lingual lexical knowledge from multilingual sentence encoders. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2023.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. Evaluating word embedding models: methods and experimental results. *AP-SIPA Transactions on Signal and Information Processing*, 8:e19, 2019a. doi: 10.1017/ATSIP.2019.12.

Haozhou Wang, James Henderson, and Paola Merlo. Weakly-supervised concept-based adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4419–4430, Hong Kong, China, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1450. URL <https://www.aclweb.org/anthology/D19-1450>.

- Haozhou Wang, James Henderson, and Paola Merlo. Multi-adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 463–472, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.39. URL <https://aclanthology.org/2021.naacl-main.39>.
- Kai Wang, Yu Liu, and Quan Z. Sheng. Swift and sure: Hardness-aware contrastive learning for low-dimensional knowledge graph embeddings. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 838–849, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3511927. URL <https://doi.org/10.1145/3485447.3511927>.
- Meihong Wang, Linling Qiu, and Xiaoli Wang. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3), 2021b. ISSN 2073-8994. doi: 10.3390/sym13030485. URL <https://www.mdpi.com/2073-8994/13/3/485>.
- Ruijie Wang, Zheng Li, Danqing Zhang, Qingyu Yin, Tong Zhao, Bing Yin, and Tarek Abdelzaher. Rete: Retrieval-enhanced temporal event forecasting on unified query product evolutionary graph. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 462–472, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3511974. URL <https://doi.org/10.1145/3485447.3511974>.
- Ruijie Wang, zheng li, Dachun Sun, Shengzhong Liu, Jinning Li, Bing Yin, and Tarek Abdelzaher. Learning to sample and aggregate: Few-shot reasoning over temporal knowledge graphs. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022c. URL <https://openreview.net/forum?id=1LmgISIDZJ>.
- Zhe Wang, Xiaomei Li, and Zhongwen Guo. Bi2e: Bidirectional knowledge graph embeddings based on subject-object feature spaces. In Mohamed Sellami, Paolo Ceravolo, Hajo A. Reijers, Walid Gaaloul, and Hervé Panetto, editors, *Cooperative Information Systems*, pages 3–18, Cham, 2022d. Springer International Publishing. ISBN 978-3-031-17834-4.

- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1112–1119. AAAI Press, 2014. URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531>.
- Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. Cross-lingual alignment vs joint training: A comparative study and A simple unified framework. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=S11-C0NtwS>.
- Peter R White. *Telling media tales: The news story as rhetoric*. Department of Linguistics, Faculty of Arts, University of Sydney, 1998.
- Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL <https://www.aclweb.org/anthology/N18-1101>.
- Svante Wold, Kim Esbensen, and Paul Geladi. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37 – 52, 1987. ISSN 0169-7439. doi: [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9). URL <http://www.sciencedirect.com/science/article/pii/0169743987800849>. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.
- Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver,

- Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL <https://www.aclweb.org/anthology/N15-1104>.
- Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, Yiming Cui, Cong Yu, Qianqian Dong, Yin Tian, Dian Yu, Bo Shi, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaowei Hua Liu, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark, 2020a.
- Wentao Xu, Shun Zheng, Liang He, Bin Shao, Jian Yin, and Tie-Yan Liu. SEEK: Segmented embedding of knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3897, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.358. URL <https://www.aclweb.org/anthology/2020.acl-main.358>.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 2015*. URL <http://arxiv.org/abs/1412.6575>.
- Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. NLP from scratch without large-scale pretraining: A simple and efficient framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25438–25451. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/yao22c.html>.
- Changlong Yu, Weiqi Wang, Xin Liu, Jiabin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. Folkscope: Intention knowledge graph construction for discovering e-commerce commonsense. *arXiv preprint arXiv:2211.08316*, 2022a.

- Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Computer Survey*, jan 2022b. ISSN 0360-0300. doi: 10.1145/3512467. URL <https://doi.org/10.1145/3512467>. Just Accepted.
- Mozhi Zhang, Yoshinari Fujinuma, and Jordan Boyd-Graber. Exploiting cross-lingual subword similarities in low-resource document classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9547–9554, 2020.
- Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. Multimodal analogical reasoning over knowledge graphs. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=NRHajbzg8y0P>.
- Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, pages 2731–2741, Vancouver, BC, Canada, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/d961e9f236177d65d21100592edb0769-Abstract.html>.
- Yi Zhang, Jie Lu, Feng Liu, Qian Liu, Alan Porter, Hongshu Chen, and Guangquan Zhang. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. *Journal of Informetrics*, 12(4):1099–1117, 2018. ISSN 1751-1577. doi: <https://doi.org/10.1016/j.joi.2018.09.004>. URL <https://www.sciencedirect.com/science/article/pii/S1751157718300257>.
- Wenfeng Zheng, Lirong Yin, Xiaobing Chen, Zhiyang Ma, Shan Liu, and Bo Yang. Knowledge base graph embedding module design for visual question answering model. *Pattern Recognition*, 120:108153, 2021. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2021.108153>. URL <https://www.sciencedirect.com/science/article/pii/S003132032100340X>.

- Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. Density matching for bilingual word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1588–1598, Minneapolis, Minnesota, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1161. URL <https://www.aclweb.org/anthology/N19-1161>.
- Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou, and Tuo Zhao. Toward understanding the importance of noise in training neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7594–7602. PMLR, 2019b. URL <http://proceedings.mlr.press/v97/zhou19d.html>.
- Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. Prix-LM: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5412–5424, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.371. URL <https://aclanthology.org/2022.acl-long.371>.
- Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.465. URL <https://aclanthology.org/2021.naacl-main.465>.
- Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Pro-*

cessing (EMNLP-IJCNLP), pages 3054–3064, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1302. URL <https://www.aclweb.org/anthology/D19-1302>.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.121. URL <https://www.aclweb.org/anthology/2020.acl-main.121>.

Yushan Zhu, Wen Zhang, Hui Chen, Xu Cheng, Wei Zhang, and Huajun Chen. DistilE: Distilling knowledge graph embeddings for faster and cheaper reasoning, 2020b.