

How Sensitive Are Cross-Lingual Mappings to Data-Specific Factors?

Nils Landegren

Department of Linguistics

Bachelor's Thesis 15 ECTS credits

Linguistics: Computational Linguistics – Bachelor's Course, LIN621

Bachelor's Programme in Philosophy and Linguistics 180 ECTS credits

Spring semester 2020

Supervisor: Robert Östling

Swedish title: Hur känsliga är tvärspråkiga mappningar för data-specifika faktorer?



Stockholm
University

How Sensitive Are Cross-Lingual Mappings to Data-Specific Factors?

Nils Landegren

Abstract

Vector representations of words, commonly known as word embeddings, aim to capture and quantify similarities between words. A branch of research that has seen much interest is the extension of word embeddings to the cross-lingual setting. The mapping-based approach to learning cross-lingual word embeddings does this by mapping the embedding space of a source language to that of a target language. It is hypothesized that this approach relies on the similarity between source and target embedding spaces. However, it is unclear to what extent this is true. In this work, I investigate this question by looking at factors in the underlying training data that could negatively affect cross-lingual mappings. Experiments were performed on embeddings learned on data varying in domain, quality, and size. Evaluation through bilingual lexicon induction indicates that the mapping-based approach is significantly favoured by larger data sets of high quality, such as Wikipedia data, and that domain difference is mainly detrimental in comparison with in domain mappings of high-quality data. In addition, results from a new set of analogy queries indicate that syntactic relations are more consistently captured by word embeddings than semantic ones, and better predict cross-lingual quality.

Keywords

Cross-lingual mapping, embedding evaluation, data sensitivity, lexicon induction, analogy task

Sammanfattning

Målet med ordvektorer, vanligtvis kallade ordinbäddningar, är att fånga och kvantifiera likheter mellan ord. En forskningsgren som har väckt mycket intresse är den som utvidgar ordinbäddningar till den flerspråkiga miljön. Den s.k. mappningsmetoden gör detta genom att hitta en mappningsmatris som bäst länkar ordrummet hos ett källspråk till ordrummet hos ett målspråk. Det hypotiseras om att denna metods applicerbarhet beror på likheten mellan ordrummen som ska länkas. Det är dock oklart till vilken grad detta är sant. I detta arbete undersöker jag denna fråga genom att titta på faktorer i det underliggande träningsdatat som kan ha negativ effekt på flerspråkiga mappningar. Experiment genomfördes på ordinbäddningar tränade på data från olika domäner, av olika kvalitet, och olika storlekar. Utvärdering genom lexikoninduktion indikerar att flerspråkiga mappningar tar stor fördel av större datamängder av hög kvalitet, så som Wikipedia-data, och att domändifferens utgör en nackdel endast i jämförelse med mappningar inom samma domän av hög kvalitet. Slutligen indikerar resultat från en nyskapad mängd analogifrågor att syntaktiska relationer mellan ord fångas mer konsekvent av ordinbäddningar än semantiska relationer, och är bättre prediktorer av flerspråkig kvalitet.

Nyckelord

Tvärspråkig mappning, inbäddningsutvärdering, datakänslighet, lexikoninduktion, analogitest

Contents

1	Introduction	1
2	Background	3
2.1	Skipgram with negative sampling (SGNS)	3
2.1.1	The classifier	3
2.1.2	The learning	4
2.2	SGNS with subword information	5
2.2.1	N-gram vector representations	5
2.2.2	N-gram classification task	5
2.3	Cross-Lingual Word Embeddings	5
2.3.1	The mapping-based approach	6
2.4	Evaluating word embeddings	7
2.4.1	Word similarity	8
2.4.2	Analogy task	8
2.4.3	Bilingual lexicon induction	8
2.5	Related research	9
3	Purpose and research questions	11
3.1	Purpose	11
3.2	Research questions	11
4	Method	13
4.1	Monolingual data	13
4.1.1	Wikipedia	13
4.1.2	News Crawl	13
4.1.3	Common Crawl	13
4.2	Preprocessing	13
4.2.1	Language identification	14
4.2.2	Deduplication	14
4.2.3	Tokenization	14
4.2.4	Compression	14
4.3	Final data set	14
4.4	Monolingual model	15
4.5	Cross-lingual model	15
4.6	Evaluation methods	16
4.6.1	Analogy task	16
4.6.2	Bilingual lexicon induction	16
4.7	Experimental setup	16
5	Results	18
5.1	Analogy evaluation	18
5.2	In domain mapping	18
5.3	Out-of-domain mapping	18
5.4	Analogy performance as a predictor of BLI performance	18
5.5	Results by research question	20

6	Discussion	23
6.1	Discussion by research question	23
6.2	Discussion of methodology	24
6.2.1	Data and language selection	24
6.2.2	Preprocessing	25
6.2.3	Models	26
6.2.4	Experiments	26
6.3	Discussion of results	27
6.3.1	Main results	27
6.3.2	Expectations	28
6.3.3	Related research	28
6.3.4	Practical applications	30
6.4	Future research	30
7	Conclusions	31
	References	32
A	BLI scores	34

1 Introduction

The idea that meaning could be represented as a vector, a list of numbers each of which referring to some semantic property, and measured in a semantic space dates as far back as the 1950s with work by psychologist Charles E. Osgood, among others. He writes:

The meaning of a concept ... can also be defined ... as that point in the semantic space identified by its coordinates on several factors [semantic dimensions]. In this representation we can ‘see’ the similarity between various concepts on all factors simultaneously in terms of their closeness in the space (Osgood et al. 1957, p 89).

In the same decade, several linguists, sometimes referred to as distributionalists, defined meaning in terms of their distribution in the language:

the linguist’s ‘meaning’ of a morpheme ... is by definition the set of conditional probabilities of its occurrence in context with all other morphemes (Joos 1950, p 708).

From this followed the distributional hypothesis that stated that words (or morphemes etc) that appear in similar contexts will tend to be similar in meaning.

The concept of embedding meanings as vectors in a semantic space, together with the idea of the distributional hypothesis laid the conceptual groundwork for what today is known as vector semantics (Jurafsky 2000, ch 6.2). Vector semantics refers to the kind of computational model that aims to extract contextual information directly from text in order to represent the semantic contents of morphemes, words, sentences and even whole documents, as vectors.

Early on, models represented meanings with vectors containing word-word co-occurrence statistics that were sparse (most dimensions were 0) and as long as entire vocabularies.

Short and dense vectors, more like the embeddings that are widely used today, were first introduced in the form of 300-dimensional vectors computed from term-document matrices through a method called Latent Semantic Analysis (Deerwester et al. 1990).

Later, it was shown that embeddings could be learned through neural models (Bengio et al. 2003), and more recently the simplifying algorithms Skipgram and Contextual Bag of Words (CBOW) were introduced, which could learn embeddings of high quality efficiently through word and context prediction (Mikolov, Chen et al. 2013).

Subsequent work also showed that similarities in the vector spaces of different languages can be taken advantage of to learn a linear mapping between them in order to learn cross-lingual word embeddings (Mikolov, Le et al. 2013). Interestingly, the learned mapping was shown to also be applicable to words not included during training, in effect making it an inductive translation tool.

Since then, the mapping-based approach to learning cross-lingual word embeddings has been improved and generalized upon by Xing et al. (2015) and Artetxe et al. (2016), among others. Recent work has even showed that mappings can be learned from very little supervision (Artetxe et al. 2017) and even no supervision at all (Artetxe et al. 2018; Conneau et al. 2017).

Although much interest and work has been put into the mapping-based approach and cross-lingual word embeddings in general, work that looks at the relationship between the underlying monolingual training data and the quality of the final cross-lingual vector space is not as abundant. Furthermore, the success of the mapping-based approach – both unsupervised and supervised – relies entirely on the similarities between source and target spaces, warranting a closer look at the impact that data-specific factors have on the quality of cross-lingually mapped embedding spaces.

A recent study on the unsupervised mapping-approach of Conneau et al. (2017) showed that the similarity between embedding spaces can be limited by factors in the monolingual training data, such as morphology and domain differences (Søgaard et al. 2018). As far as I know, a similar study for the standard supervised mapping-based approach has not been done. And although the supervised mapping-based approaches are more robust and resistant to differences in embedding spaces, exactly to what extent, is still of interest. If one wants to use this method for learning cross-lingual word embeddings, one would also want to know how much data is needed, how much noise in the data is acceptable, and what kind of domain difference is tolerable, for the method to be viable.

To this end, I evaluate the supervised mapping-based approach of Artetxe et al. (2016), from the standpoint of data-specific factors. In particular, I evaluate the performance of cross-lingual word embeddings learned from data sets differing in size, domain, and quality, on bilingual lexicon induction. To further investigate the relationship between the quality of the monolingual word embeddings and the cross-lingual mapping, I translated the analogy test of Mikolov, Le et al. (2013) to the four languages included in this study. Furthermore, since it has been shown that not all intrinsic tasks (tasks that evaluate the inherent properties of embedding spaces) correlate well with downstream tasks (tasks where embeddings are used as features in a model that is applied to an NLP task; Schnabel et al. 2015; Tsvetkov et al. 2015), I tested the analogy scores of the monolingual embeddings for correlation with the performance of the cross-lingual word embeddings on bilingual lexicon induction.

The structure of this thesis is as follows. In section 2, I give background to the concepts of both monolingual and cross-lingual word embeddings. In section 3, I formulate the purpose of this research, along with several research questions and hypotheses. In section 4, the details of the methodology used in the research is spelled out. In section 5, I present the results of the experiments performed. In section 6, I discuss the results. And finally, in section 7, I summarise the work with some conclusions.

2 Background

In this section I give the theoretical background to the research in this study. I give a short introduction to word embeddings, both monolingual and cross-lingual, by summarizing two popular models in each of the settings.

2.1 Skipgram with negative sampling (SGNS)

Probably the most popular choice in algorithm for learning monolingual word embeddings today, is skipgram with negative sampling (SGNS). The algorithm is one of two variants introduced by Mikolov, Chen et al. (2013), as part of the word2vec software package. The following sections 2.1.1 and 2.1.2 is my adaption of the explanation of SGNS given by Jurafsky (2000, Ch 6.8).

The name that SGNS stands for comes from the type of sequence of text that is used as training samples: the skipgram. A skipgram is a sequence of some units (e.g characters, morphemes, words) where the middle unit has been left out, skipped. Take a sentence $s = w_1, w_2, w_3, w_4, w_5$ for example. Several skipgrams can be extracted from it, depending on the choice of window size. The sub-sequence w_1, w_3 is one such skipgram of s , formed from the window of the first 3 words in s .

The idea of SGNS then, can be summarised as follows. Walk through a large body of text one target word w at a time, extracting the skipgram of each w of some window size. Use these skipgrams as positive training samples, together with some random words as negative samples, and train a binary classifier to be able to predict a given sample as a positive, or negative, sample. In other words, to predict whether the sample is a skipgram where w is the middle word that has been left out.

This task we set for the classifier to perform can be thought of as an artificial task. This is because the idea is not to train and put the classifier to use. The purpose is to learn and use the classifier's parameters (i.e the word embeddings). The classifier is *trained*, to *learn* the word embeddings. This kind of machine learning falls into the category commonly referred to as representation learning. To give a clearer picture of SGNS, I describe in the following the classifier and the learning in some more detail.

2.1.1 The classifier

To understand the machinery at work here, let's think about what we want the trained classifier to do. The classifier should take a word vector t and a set of context vectors c (the vector representations of the words in the skipgram of t), and if c is a positive sample, return true, otherwise return false. Although the task here is an artificial one, the classifier making correct predictions means it has learned meaningful word embeddings.

To understand how the predictions are made, recall that, according to the distributional hypothesis, two words that often appear in the same context will be similar in meaning. The classifier takes advantage of this idea, and bases its prediction on the similarity of the word vectors. In particular, the higher the similarity between vectors t and c_i , the higher the probability the classifier will output.

From start to finish, the classification can be split into three steps: 1) measure similarity of the vectors, 2) turn the similarity value into a probability value and 3) make a prediction based on a probability threshold. More formally, the steps can be formulated as follows. The similarity of two vectors is measured by taking the dot product between them, with a higher dot product

meaning more similar vectors:

$$\text{similarity}(t, c) \approx t \cdot c \quad (1)$$

Before a prediction can be made the dot product needs to be turned into a probability. A common way to do this is to use the sigmoid function. The sigmoid function takes some number as input and transforms it into a number between 0 and 1:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

The probability of the whole set of context vectors can then be calculated by multiplying the probabilities of each individual vector:

$$P(+|t, c_{1:k}) = \prod_{i=1}^k \frac{1}{1 + e^{-t \cdot c_i}} \quad (3)$$

Where $c_{1:k}$ means each vector in the context set. With the probability value calculated, we can say that any value above the threshold of 0.5 will count as a positive classification.

2.1.2 The learning

Up to now, for the sake of understanding the classifier, we have assumed the parameters of the classifier to be meaningful word embeddings. That is, that the dot product between two word vectors that often appear in the same contexts (have similar skipgrams) is high. However, the point of SGNS is to learn meaningful word embeddings, and before training, the word representations are just randomly initialized vectors, with no meaningful difference in similarity between them. To learn meaningful word vectors, the untrained classifier is given a learning objective and an algorithm to do the learning in each step of classifying.

The learning objective is to maximize the dot product between vectors of similar words, and minimize the dot product between vectors of dissimilar words. This can be defined formally with following objective function:

$$L(\theta) = \sum_{(t,c) \in +} \log P(+|t, c) + \sum_{(t,c) \in -} \log P(-|t, c) \quad (4)$$

Where θ is the parameters of the classifier and where $P(-|t, c) = 1 - P(+|t, c)$. In other words, in maximizing the objective function L , we are trying to find the parameters θ that gives the maximum value of L . To achieve this goal, the learning algorithm Stochastic Gradient Descent (SGD) is applied. In each training instance, with a positive and negative sample in relation to some target word t , the learning algorithm adjusts the vectors towards maximizing the objective function. This is done by using a parameter updating equation. For more details on SGD see for example Jurafsky (2000, Ch.5.4).

To sum up. A classifier, containing parameters in the form of word vectors, together with the means to make predictions using dot product, the sigmoid function and a prediction threshold, is trained towards an objective function. The training of the classifier consists in maximizing the dot product between similar word vectors, and minimizing the dot product between dissimilar word vectors through a learning algorithm. When the training is complete, the parameters of the classifier are retrieved and used as word embeddings.

2.2 SGNS with subword information

The model SGNS described above learns a vector representation for each word in the vocabulary. However, not all words consist of only one morpheme and one meaning. This is especially true for morphologically complex languages like for example Turkish. Turkish is a highly agglutinative language where words can consist of many individual morphemes, each of which carrying their own individual meaning, whether lexical or functional. There have been suggestions to take these more fine-grained meanings below word level into account, and the most successful model is that of Bojanowski et al. (2017)¹. Their model is an extension of the SGNS model and is widely used to learn word embeddings today. There are two major differences between SGNS and this extension.

2.2.1 N-gram vector representations

Firstly, instead of representing a word with only one vector, this model represents words as the sum of the vectors of their n-grams together with a vector of the whole word. To make this more clear take the example of the word *carrot* with n-gram length of 3. First the boundary symbols $<$ and $>$ are added to the word to indicate prefixes and suffixes, resulting in the word $<carrot>$. The word is then split up into n-grams of length 3, resulting in the set of n-grams $<ca, car, arr, rro, rot, ot>$. Each of these n-grams, in addition to the whole word, is represented by its own vector. Summing each of the n-gram vectors is what represents a word in this model. Notice also that because the whole word is also given a vector, the n-grams *car* and *rot* in this example will not be confused for the vectors of the actual words $<car>$ and $<rot>$ were they to appear in the training data.

2.2.2 N-gram classification task

Since the vectors that the classifier now deals with are not on word level, the prediction task is formulated slightly differently. Instead of computing a similarity score using the dot product between a target word vector and a context word vector, Bojanowski et al. (2017) formulates the following scoring function s of some target word t and context word c :

$$s(t, c) = \sum_{g \in G_t} z_g^T v_c \quad (5)$$

Where G is the bag of character n-grams. This simple extension of the SGNS model to subword level makes word embeddings of better quality for languages of higher morphological complexity. Another interesting advantage with this model is that it can compute word embeddings for out-of-vocabulary words as long as the n-grams of the word has been seen during training.

2.3 Cross-Lingual Word Embeddings

In the above section on SGNS I explained one method of learning monolingual word embeddings. In this section I want to give an introduction to what it means to extend the concept of word embeddings to the cross-lingual setting. The question that formulates the central problem of cross-lingual word embeddings is something like: how can we embed words of two or more languages into the same vector space? Another way to put it is to formulate a two-parted

¹The authors supply an open source implementation of the model as a part of the fasttext library (<https://fasttext.cc/>), embeddings learned with this model is thus sometimes referred to as fasttext embeddings

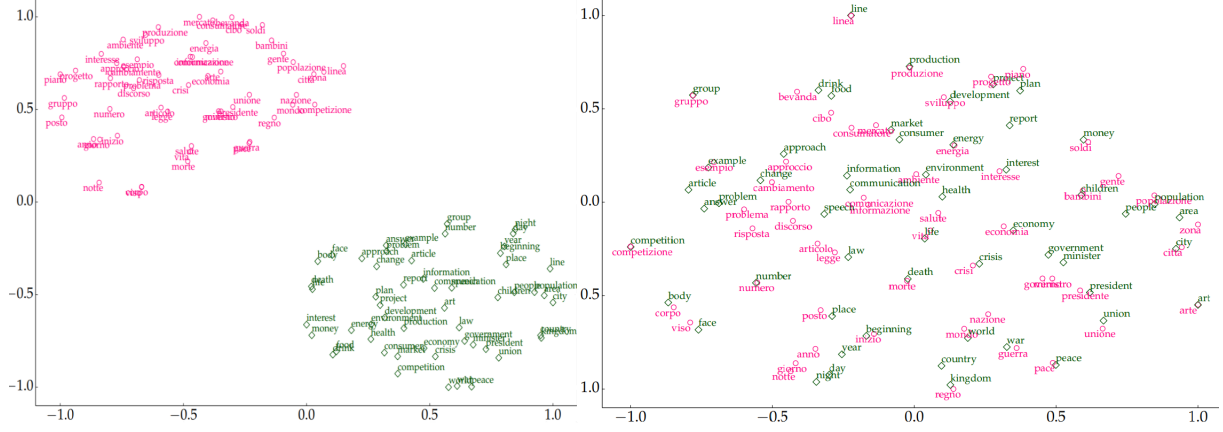


Figure 1: A visualization of two sets of embeddings, Italian (pink) and English (green), before alignment (left) and after alignment (right) (Ruder et al. 2019).

goal. Firstly, we want to maximize the similarity between semantically similar words within one language (the monolingual step). And secondly, we also want to maximize the similarity of semantically similar words across languages (the cross-lingual step). In both these goals maximizing the similarity of words means minimizing the distance between their embeddings. Figure 1 is a visualization of two monolingual word embedding spaces before and after alignment.

There are several approaches to learning cross-lingual word embeddings. Ruder et al. (2019) makes a distinction between approaches in terms of how they divide the task of the monolingual and cross-lingual learning. Mapping based approaches aim to learn a linear transformation between vector spaces, and so neatly separates out the monolingual learning, meaning you can align preexisting monolingual word embeddings. Other approaches, such as pseudo-multilingual corpora-based approaches, formulates the monolingual and cross-lingual learning into the same problem by applying monolingual learning methods (such as SGNS) to pseudo-multilingual corpora. These are corpora consisting of monolingual data that has been manipulated to be representative of more than one language at a time. One way this is done is by randomly replacing words in a corpus with translation equivalents.

As the mapping-based approach is the focus of the evaluation in this work, I will in the following section give a short introduction to this approach. For more examples on alternative approaches, I refer to the survey literature, for example Ruder et al. (2019).

2.3.1 The mapping-based approach

The mapping-based approach to learning cross-lingual word embeddings can be described in the following way. Let X be a matrix representing the vector space of some source language, where X_i denotes the i :th row vector (word embedding) in X . Let Z be a matrix representing, in the same way, some target language, such that the word embedding X_i is aligned with the embedding of its translation in the target language Z_i . What the mapping-based approach aims to do then, is to find a transformation matrix W , such that $WX \approx Z$. When W has been computed, we in effect have a key that projects the vectors of X into the space of Z . There are several more steps, but this is the core machinery of learning cross-lingual word embeddings with the mapping-based approach.

What makes this approach especially interesting is that even though the initial purpose was to learn cross-lingual word embeddings for the embeddings of X and Z , the learned mapping can generate new bilingual word pairs. In particular, the transformation matrix W can be applied

to further word embeddings that were not present in the learning of W , in effect making it an inductive translation tool. Furthermore, the fact that this approach has seen relative success gives insight into the structure of meaning representation of individual languages, and that there are elements of them that are more or less the same, even for distant languages. On the other hand, the success of the approach relies entirely on this assumption of universal similarities in the vector spaces of languages. How far these similarities can realistically be taken advantage of for this kind of purpose is the motivation of much recent research into the topic.

Mikolov, Le et al. (2013) first proposed a method implementing cross-lingual mapping for inducing translations across languages. This was the result of their noticing similarities in the constellations of words in the vector spaces of different languages. In the example they give, they observed that when visualizing vector spaces for English and Spanish with Principal Component Analysis (PCA), the english words 'one' through 'five' were similarly oriented in relation to each other as the words 'uno' through 'cinco' were in spanish. The authors formalize the problem of learning the transformation matrix W as minimizing the sum of squared Euclidean distances:

$$\min_W \sum_{i=1}^n \|Wx_i - z_i\|^2 \quad (6)$$

They then solve this problem using the algorithm SGD. When the optimal transformation matrix W has been found, the authors multiply it by a vector in the source language vector space to approximate it in the target language vector space. They then perform nearest neighbor retrieval on this approximation, using cosine similarity, to produce what should be the correct translation Z_i of X_i .

In more recent work Artetxe et al. (2016) proposed several constraints on this method. In particular, they showed that constraining the transformation matrix W to be orthogonal preserves the monolingual quality of the cross-lingual word embeddings. Furthermore, this orthogonality constraint means that the optimal mapping matrix W can be computed more efficiently with an exact solution, instead of with SGD. This is a known solution to a known problem called the Procrustes Problem, which is solved by obtaining W from taking the singular value decomposition (SVD) of $Z^T X$. SVD is an operation in linear algebra that factorizes some matrix A into the three sub-matrices U, Σ, V^T . In this scenario, we get the mapping matrix W by factorizing $Z^T X$ with SVD and taking the product of V and U^T , i.e. $W = VU^T$. For a better understanding of the Procrustes Problem see for example Gower, Dijksterhuis et al. (2004). Artetxe et al. (2016) also provides a proof of solution in this particular application.

2.4 Evaluating word embeddings

Methods for evaluating word embeddings can be divided into two main categories: intrinsic and extrinsic. Intrinsic tests aim to evaluate inherent properties of the vector representations and the vector spaces, while extrinsic tests consists in some natural language processing (NLP) task, e.g. word translation. Another kind of extrinsic task is the downstream task. Downstream tasks also consist in some NLP task, but more specifically involves inserting the word embeddings as features in a model that is then applied to the task.

Although applying word embeddings as features in NLP models to directly measure their usefulness is considered by many to be the most important type of evaluation, intrinsic evaluation have their advantages in being less computationally expensive and easier to interpret. In the ideal situation, results from an intrinsic test would correlate with performance on extrinsic evaluation, making the intrinsic test a useful first step in quality assessment, and an indication of what intrinsic properties of the word embeddings predict downstream performance.

2.4.1 Word similarity

Testing similarity is the most common evaluation of monolingual word embeddings. In similarity tests, the similarity between embeddings for a pair of words is computed with some measure, such as cosine similarity. This measure is then compared for correlation with a manually annotated gold standard for similarity between the same pair of words. There are several data sets that are commonly used. WordSim-353 is a set of 350 pairs of nouns in English that were ranked in similarity between 0 and 10 by manual annotators (Finkelstein et al. 2001). SimLex-999 is another manually annotated data set with several hundred word pairs in subsets of nouns, verbs and adjectives, focusing primarily on semantic similarity as opposed to semantic relatedness (Hill et al. 2015).

Perhaps more interesting are variations on similarity tests that take contexts into account. Stanford Contextual Word Similarity (SCWS) is a data set constructed by Huang et al. (2012) to evaluate the ability of models to take homonymy and polysemy of word senses into account. The set consists of 2003 pairs of words where each of the words appears in a sentence context. The Word-in-Context (WiC) data set is another data set for contextual similarity that is framed as a set for evaluating binary classification (Pilehvar and Camacho-Collados 2018).

2.4.2 Analogy task

The analogy test comes from an interesting result in recent research on word embeddings. Representations exhibit a certain amount of compositionality in meaning. For example, if you take the vector representation for 'king', you subtract from it the representation for 'man' and add to it the representation for 'woman', with high quality embeddings you should end up with an embedding that is closest to the representation of 'queen'. Analogy tests evaluate the compositionality of the vector representation with analogies such as 'stockholm' - 'sweden' + 'germany' where the vector space should produce 'berlin'. Sets of these analogies in the form of tuples have been created by (Mikolov, Chen et al. 2013; Mikolov, Le et al. 2013). Recently, Gladkova et al. (2016) have created a substantially larger data set for analogy evaluation called The Better Analogy Test Set (BATS) consisting in 99,200 questions in multiple categories.

2.4.3 Bilingual lexicon induction

Bilingual lexicon induction is the most common evaluation for cross-lingual word embeddings. Given a cross-lingual vector space, the evaluation measures how well the vector space links similar words together across languages. More specifically, given a word w_{l1} in the source language, the task is to retrieve the most similar word w_{l2} in the target language, this is called nearest neighbor retrieval. The result is then compared to a gold standard of translation pairs and if the induced pair (w_{l1}, w_{l2}) equals that of the gold standard the answer is correct.

The step of nearest neighbor retrieval presents a problem where some vectors are close in distance to many other vectors, these are called semantic hubs. One way to combat semantic hubs is with a method called cross-domain similarity local scaling (CSLS), proposed by Conneau et al. (2017). CSLS adjust the similarity of two vectors according to their mean similarity to their n nearest vectors. For an exact definition see e.g. Ruder et al. (2019) or Conneau et al. (2017).

To measure the results of BLI, there are several possible metrics that can be used. Simply using accuracy (the percentage of correct answers) is usually not informative enough. A vector space that produces the correct translation as the second most similar word would be incorrect, yet is obviously of better quality than one giving the correct translation as it's 10th suggestion. To account for this, a commonly used metric for evaluating the results of BLI is Precision-at-k

(P@k). P@k deems an answer correct if the correct translation according to the gold standard is in the top k most similar words as suggested by the vector space. Common values used for k are 1, 5 and 10.

Another metric that can be more informative is mean reciprocal rank (MRR). In MRR the score of a translation is relative to the target word's rank in nearest neighbor retrieval. If the correct translation is the second nearest neighbor, its reciprocal rank is 1/2, if its the fifth nearest neighbor its reciprocal rank is 1/5 etc. The total score of the MRR evaluation is computed as the mean of all the individual reciprocal rankings.

BLI as an evaluation task can be thought of both as intrinsic and extrinsic. The task is extrinsic in the sense that it has a practical real world use; a large high quality bilingual lexicon is a useful resource, thus a way to induce them is as well. The intrinsic value of the task on the other hand comes from the initial evidence it gives of how successful the linking between word embeddings across languages are.

2.5 Related research

In this section I summarize some related work on both monolingual and cross-lingual word embedding models, as well as evaluative studies.

The monolingual word embedding model used in this study is that of Bojanowski et al. (2017). The results they reported on similarity evaluation will be compared to the results in this study.

Mikolov, Le et al. (2013) presented the original mapping-based approach for learning cross-lingual word embeddings by a linear transformation between embedding spaces. They evaluated the method on data sets of sizes ranging from 10 million tokens to several billions. They did not, however, look at domain difference or smaller data sets.

As explained in section 2.3, the implementation of Artetxe et al. (2016) is that which is under evaluation in this work. Furthermore a line of research that has received a great amount of attention is one that tries to relax the condition of supervised signals, by utilizing identical words Artetxe et al. (2017), or even fully unsupervised models such as Conneau et al. (2017) and Artetxe et al. (2018).

Recently, Glavas et al. (2019) performed a comprehensive study on cross-lingual word embedding models. They evaluated several models, both supervised and unsupervised, on 28 language pairs. The method of evaluation was BLI together with the cross-lingual downstream tasks document classification, information retrieval, and natural language inference. They used fasttext embeddings pretrained on full Wikipedias of each language. The results they reported on bilingual lexicon induction will be compared to the results in this study. Furthermore, they created evaluation data sets for bilingual lexicon induction which was used in this study.

Finally, I will spell out the work of Søgaard et al. (2018) in some detail, as it is closely related to this work in terms of research questions and experiments. They investigated how certain monolingual word embedding learning scenarios impact the unsupervised mapping-based approach of Conneau et al. (2017). This was motivated by their observations that embedding spaces of different languages tend not to be approximately isomorphic (similar in shape). They further supported these observation by introducing a method for quantifying the similarity between word embeddings based on graph similarity. They found this metric to be a predictor of performance on bilingual lexicon induction.

Their main research questions revolved around three scenarios that could prove challenging for unsupervised cross-lingual mapping, and that could impact similarity between embeddings. Firstly, they looked at the impact of language difference. By mapping between embeddings

of languages differing in morphology they found that some language combinations make for worse similarity scores, and are more difficult for this unsupervised model to map between. For example, the model failed completely for the language pair English (an isolating language with dependent marking) and Finnish (an agglutinative language with mixed marking).

Second, they investigated what the impact of domain difference has on the cross-lingual word embeddings. To do this they mapped all combinations of embeddings learned from the domains EuroParl, Wikipedia, and the EMEA corpus in the medical domain. They found that when domains are different, performance is close to zero on some language pairs.

The third factor they investigated was dimensionality of the word embeddings. The most common is to use embeddings of 300 dimensions and interestingly they found that 40-dimensional word embeddings actually performed better for Estonian, Finnish, and Greek. They hypothesize that monolingual embedding algorithms may over-fit on languages with more unique properties if dimensions are higher.

These results show that there are many factors that can negatively impact the overall similarity between embedding spaces, which in turn significantly limits unsupervised cross-lingual mappings. The research of this work is motivated and formulated in a similar way, but evaluates supervised cross-lingual mappings, and several parts of the findings described above will be discussed in conjunction with the results in the present study. In the next section I formulate several research questions to further guide the research.

3 Purpose and research questions

3.1 Purpose

The purpose of this research is to investigate the ways in which the quality of cross-lingual mappings is contingent on factors in the underlying monolingual training data. By doing this I aim to elucidate data-specific conditions that are favorable – or unfavorable – for the mapping-based approach to learning cross-lingual word embeddings.

3.2 Research questions

In what follows I formulate some research questions and hypotheses to guide the research in this study.

Question 1: How does the amount of training data when learning monolingual word embeddings affect the quality of monolingual word embeddings and of cross-lingual mappings?

Hypothesis 1: Any increase in monolingual training data size means better cross-lingual quality and better monolingual quality. In the case of the mapping-based approach, performance on bilingual lexicon induction has been shown to improve with every increase in data, albeit with diminishing returns on large amounts (Mikolov, Le et al. 2013). In the case of monolingual embeddings however, the model used in this study have been shown to learn word embeddings whose performance on word similarity tasks saturates quickly and deteriorates somewhat on larger data sets Bojanowski et al. (2017).

Hypothesis 2: There is a minimum amount of data that is needed for any gain in cross-lingual mapping quality.

Question 2: How does noise in the training data when learning monolingual word embeddings affect the quality of the monolingual word embeddings, and cross-lingual mappings respectively?

Hypothesis 3: Word embeddings learned from Wikipedia data will achieve the highest scores. Word embeddings learned from the Common crawl corpus will achieve the lowest scores.

Motivation: Wikipedia should contain very little noise, as it is constantly edited. The Common crawl corpus is automatically and indiscriminately collected from the web, and so should contain much more noise, in turn producing embeddings of lower quality. As the News crawl corpus is restricted to news articles, it should be less noisy than the Common crawl set, but noisier than Wikipedia.

Question 3: How does domain difference in the monolingual word embedding training data affect the quality of cross-lingual mappings?

Hypothesis 4: Domain difference is a considerable disadvantage for the mapping-based approach. Motivation: the mapping based approach relies on source and target vector spaces to share some similarity, which in turn relies on the similarity in training data sources. The similarity across domains could be limited for several reasons. If domains are different enough in topic, i.e. use different vocabularies, then some source words will have no target word to align with. Furthermore, if different domains associate different meanings to orthographically

identical words, then some words will not be similarly oriented across vector spaces. Additionally, a similar, albeit unsupervised, model have shown to fail completely under some conditions when mapped across different domains (Søgaard et al. 2018).

Question 4: To what extent is the quality of monolingual word embeddings a predictor of the quality of their cross-lingual mappings?

Hypothesis 5: Analogy test performance correlates well with cross-lingual performance on bilingual lexicon induction, all else being equal. Motivation: as far as I know this correlation has not been tested for before, but I expect that if word embedding quality suffers as a result of noisy data, this will be reflected in the performance of both analogy task and bilingual lexicon induction.

4 Method

In what follows I describe the methodology of the thesis research project. I collected data from three domains and for four languages, described in section 4.1. I then prepared the data with a preprocessing pipeline described in 4.2. In 4.3 the final data set is summarised. To learn monolingual word embeddings I used the model SGNS with subword information outlined in 4.4. The monolingual word embeddings between six language pairs were then cross-lingually mapped, described in 4.5. The two methods used for evaluation is described in section 4.6. Finally, in 4.7 I describe the experimental setup of analogy test and BLI.

4.1 Monolingual data

Monolingual word embeddings were learned on several sets of monolingual data, varying in source, language, and size. In particular, four languages were chosen with availability of training and testing data in mind, as well as typological variation: French, Finnish, Russian and Turkish. The language pairs under evaluation were Finnish-French, Finnish-Russian, Russian-French, Turkish-Finnish, Turkish-French, Turkish-Russian. Monolingual data was sourced from Wikipedia, the Common Crawl dataset, and the News Crawl dataset.

4.1.1 Wikipedia

Wikipedia has proved a fantastic resource for much NLP research (Yano and Kang 2016). It sees constant revision and growth, and contains data from many different languages, making it a better natural language representative than many other data sets gathered online. I chose Wikipedia to act as the best case scenario in contrast to the two other more noisy data sets.

I downloaded subsets of the latest (2020) XML dumps for each language, and extracted the raw text using the wikiextractor tool.²

4.1.2 News Crawl

The second source that I collected was the News Crawl data set provided by Workshop on machine translation (WMT) (Barrault et al. 2019). The data set contains monolingual text from online newspapers collected between 2007 and 2019.

4.1.3 Common Crawl

The third source is the Common Crawl data set, also provided by WMT. This corpus contains very large amounts of data, but is collected from online websites with no particular domain in mind, and so is considerably more noisy than the previous two sources. I downloaded subsets of the available raw data from 2016 for the languages Finnish, Turkish, and Russian. For French I downloaded a subset of their deduplicated raw data from 2019.

4.2 Preprocessing

To prepare the monolingual data I performed three steps of preprocessing: language identification, deduplication, and word tokenization.

²<https://github.com/attardi/wikiextractor>

4.2.1 Language identification

To remove text of languages other than the intended one, I used the language identification tool included with the fasttext software package (Grave et al. 2018)³. The tool is a model trained to identify the language of a text sequence given as input. Together with a classification of the language the model predicts for the text, it also reports a confidence value between 0 and 1.

The original authors chose to keep only lines of text in their data that was of length greater than 100 characters, and that were classified as the correct language with at least a confidence degree of 0.8. To motivate my choice of parameters I tested the tool on the Europarl corpus for Finnish and French respectively. In both cases the model classified about 2% of lines with a confidence lower than 0.8 and/or as the wrong language. However, more than 95% of these were of length less than 100 characters. Similarly to the original authors then, I chose to keep only lines that were of length greater than 100 characters and classified as the language in question with a confidence of 0.8 or higher.

4.2.2 Deduplication

Since the data used in this research is collected from online sources, I chose to perform deduplication of the data with the common crawl dedupe program⁴. This can be important for some data sources that are automatically collected from the internet, as they can contain large amounts of boiler-plate text, such as HTML code. In addition, the tool also removes text that is not encoded with UTF-8 encoding.

4.2.3 Tokenization

The data was word tokenized using the Natural Language Toolkit (NLTK) in Python (Bird et al. 2009). In particular, I used the `word_tokenize` method which includes simple punctuation with regular expression as well as models for each language that handles sentence boundaries.

4.2.4 Compression

Since the data used differs in language as well as source, both syntactic and semantic content will vary greatly between sets. To allow for a fair comparison of the language-domain combinations of the data, I chose compressed file size as a metric for determining what data size to compare, the program used was gzip.⁵ Compression has been used before as a way to assess the complexity of a text, especially to measure morphological complexity (Juola 2008). In addition to morphosyntactic differences, factors such as the topic of the text, the style of the author, and noise such as misspelling all contribute to the level of information content. To choose the data sets to use, I extracted subsets of each of the data sets that all equaled the same size when compressed.

4.3 Final data set

The upper limit of data size was determined by the language-corpus combination that had the least amount of data available. This was the Turkish Wikipedia data set, containing 141MB of compressed data. To make sure the amount of information was approximately equal in all

³<https://fasttext.cc/docs/en/language-identification.html#content>

⁴<https://github.com/kpu/preprocess>

⁵<https://www.gnu.org/software/gzip/>

Table 1: Percentages of raw text removed during preprocessing. Language identification also includes removing lines of length < 100 characters.

	Deduplication	Language identification
Wikipedia	1.9%	8%
News Crawl	0.5%	32%
Common Crawl	0.05%	40%

Table 2: Number of word tokens and word types, respectively, for each data set used in monolingual learning. Each data set equals approximately 140MB of compressed data.

	Wikipedia	News crawl	Common crawl
Finnish	51,822,875 / 2,197,079	45,470,474 / 1,814,462	53,479,438 / 2,428,911
French	74,354,535 / 583,611	65,446,007 / 360,796	72,896,275 / 642,490
Russian	49,584,275 / 959,014	44,126,126 / 415,449	49,954,694 / 580,588
Turkish	59,564,288 / 1,188,905	50,848,841 / 708,866	59,742,658 / 1,356,468

languages-corpus sets, subsets of all text files were extracted such that when compressed, they equaled $140\text{MB} \pm 1\text{MB}$ in size. The final monolingual data set thus contained one text file for each corpus-language combination, 12 in total. Table 1 shows how much data was removed, averaged over languages, during deduplication and language identification. It may be noted that somewhat unexpectedly the deduplication removed more data in the ‘less noisy’ data sets. This is most likely due to the fact that a portion of the Common crawl and News crawl data I used had already been deduplicated. Table 2 shows the word token and word type statistics for the final data sets.

4.4 Monolingual model

To learn monolingual word embeddings I used the fasttext model SGNS with subword information (Bojanowski et al. 2017)⁶. I described the model SGNS in section 2.1, and its extension to take subword information into account in 2.2 above. The dimension parameter was set to 300. To allow for reproducible results, I used single-threaded learning. All other parameters were left as default.

4.5 Cross-lingual model

In order to map the monolingual word embeddings and induce the cross-lingual word embeddings (CLWE) I used the open source framework VecMap⁷ (Artetxe et al. 2016). The model is described in section 2.3 above. Since I used the supervised mapping-based approach, a supervision signal in the form of a bilingual lexicon was needed to induce the mapping matrix W . For this purpose I used bilingual lexicons of 5000 bilingual word pairs for each of the language pairs. The data was retrieved from previous work by Glavas et al. 2019, who created the set automatically with google translate by translating the 7000 most frequent words in English to each language.

⁶<https://fasttext.cc/>

⁷<https://github.com/artetxem/vecmap>

4.6 Evaluation methods

4.6.1 Analogy task

To evaluate the monolingual word embeddings, I automatically translated the English analogy test data presented by Mikolov, Le et al. (2013), which contains approximately 30000 analogy queries, of semantic as well as syntactic kind. Each query in the data consists of the 4-tuple of four unique words whose vector representations are evaluated with simple arithmetic operations: $x - y + x_1 \approx y_1$. This simple equation is in essence asking 'what vector y_1 is approximately related to x_1 in the same way that y is related to x '?

The translation was done in Python3 using the google translate API. Each word type in the original document was translated to each of the four languages in question. In the case that a word was translated into more than one word in any of the target languages, each query containing that word was removed from the data set for each of the languages. This resulted in 11207 analogy queries of which 7631 were of semantic kind and 3576 were syntactic.

For clarity I will give a few examples of relation queries included in the set. The semantic category contains relations such as capital-common-countries, where an example is Athens – Greece + Bangkok \approx Thailand. Another is the currency relation, where an example is Argentina – peso + Sweden \approx krona. In the syntactic category there are relations such as the superlative query, e.g. bright – brightest + strange \approx strangest, or the past tense query, e.g. dancing – danced + jumping \approx jumped.

4.6.2 Bilingual lexicon induction

Bilingual lexicon induction (BLI) was performed to evaluate the quality of the cross-lingual word embeddings. The task is to retrieve the word vector w_{l_2} in the target language vector space that is most similar to some word vector w_{l_1} in the source language vector space. If the word of the vector w_{l_2} is a correct translation of the word of w_{l_1} according to a gold standard, the answer is deemed correct. CSLS, which is explained in section 2.4.3, was used for nearest neighbor retrieval.

BLI has the advantages of requiring relatively little evaluation data and is quite computationally inexpensive. Furthermore, the results of orthogonal mapping methods on BLI evaluation have been shown to correlate with downstream tasks such as cross-lingual information retrieval and cross-lingual language inference (Glavas et al. 2019). The data used in the BLI-tests consisted of one test set for each language pair, containing 2000 translation pairs. These sets were also retrieved from Glavas et al. 2019⁸.

Although the most common evaluation metric used along with BLI is precision-@-k, I chose to include mean reciprocal rank (MRR) since it is more informative in tests of smaller data sets. Along with MRR, the percentage of translation pairs that were present in the training data (coverage) and the percentage of correct word translations according to the gold standard (accuracy) is also reported.

4.7 Experimental setup

To motivate the experimental design in this study, I briefly repeat the research questions formulated in section 3.2 above, and present each experiment accordingly.

⁸<https://github.com/codogogo/xling-eval>

Question 1: How does the amount of training data when learning monolingual word embeddings affect the quality of monolingual word embeddings and of cross-lingual mappings?

To answer Question 1 I run each experiment in this setup on data sets of varying size. Each subset is half the size of the previous one and in total there are 11 subsets of each data set presented in table 2 above. This way I hope to reveal the performance trends on analogy task and BLI as a function of data size.

Question 2: How does noise in the training data when learning monolingual word embeddings affect the quality of the monolingual word embeddings, and cross-lingual mappings respectively?

Similar to the strategy of varying data set size, I also chose to vary data set source in each of the experiments to reveal any sensitivities to the quality of the training data. I hope to reveal potential sensitivity to noisy data by using Common crawl embeddings in comparison with the high quality Wikipedia embeddings.

Question 3: How does domain difference in the monolingual word embedding training data affect the quality of cross-lingual mappings?

To investigate the sensitivity of the supervised mapping-based approach to domain difference, I map the in domain pairs of each domain (Wikipedia-Wikipedia etc). I will then compare these results with the three out-of-domain pairs of Wikipedia together with Common crawl, Wikipedia together with News crawl, and Common crawl with News crawl.

Question 4: To what extent is the quality of monolingual word embeddings a predictor of the quality of their cross-lingual mappings?

To investigate the relationship between the quality of the monolingual word embeddings with their subsequent cross-lingually mapped embeddings, I perform a simple linear regression analysis of the analogy task results (independent variable) and the BLI results (dependent variable).

5 Results

To evaluate the mapping-based method of learning CLWEs, three experiments were performed followed by a statistical analysis: (i) an analogy test of monolingual embeddings (ii) an in-domain mapping, (iii) an out-of-domain mapping, and in the analysis (iv) the results of experiment (i) and (ii) were analysed in two regression models. In each of the experiments, tests were performed on embeddings learned from 10 subsets of the monolingual data, each set doubling in size.

5.1 Analogy evaluation

The accuracy of the semantic and syntactic queries, respectively, is presented in figure 2. The embeddings learned from Common crawl data did very poorly, reaching only an accuracy of 3%. All domains performed substantially better on the syntactic queries, with Wikipedia reaching 30%. Since the data set was automatically translated with no manual correction done, it can be expected that some analogy queries may have translated incorrectly. To at least make sure the results were not entirely faulty, I performed the same analogy tests on the English word embeddings learned from the same amount of Wikipedia data. Other than the English word embeddings reaching the higher peak in accuracy of 60%, there was no discernible difference in the results of this test compared to the other four. The English test also showed that highest syntactic accuracy was reached at around 3.5M and 7M tokens. The only difference I found was that in the analogy test of the English word embeddings the accuracy reached the much higher 60% in total.

5.2 In domain mapping

The results from the BLI tests of the in-domain mappings are presented in figure 3. Each of the domains saw a steady increase in MRR as data size increased. At each subset starting from the fourth (440K tokens), Wikipedia achieved the highest MRR score. The scores of News crawl and Common crawl were similar, Common crawl performing somewhat better on the two largest sets. See appendix A for BLI scores of individual language pairs.

5.3 Out-of-domain mapping

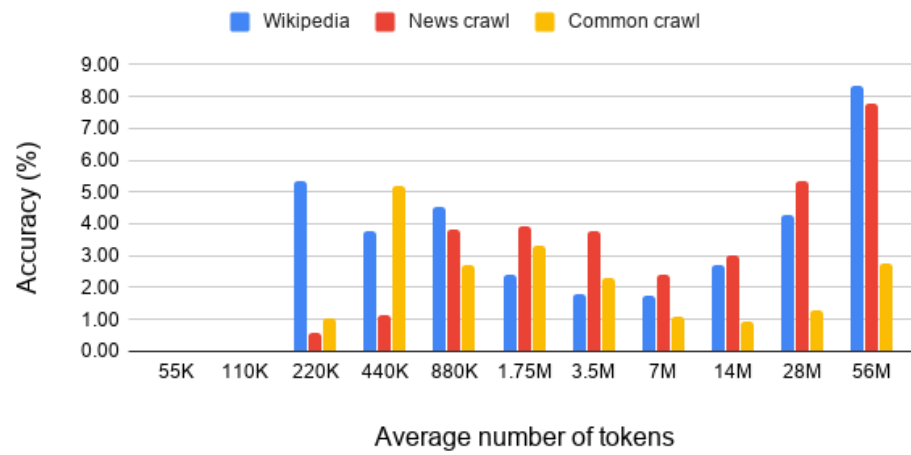
The results from the BLI tests of the out-of-domain mappings are presented in figure 4. Each of the domain pairs had a similar increasing trend as those of the in-domain pairs Common-Common and News-News. The Wikipedia-Common pair reached a slightly higher MRR than the two others in the final data set.

5.4 Analogy performance as a predictor of BLI performance

To test whether analogy test scores can be taken as indicators of performance on BLI, I performed two regression analyses for the semantic and syntactic analogy scores separately, since they showed very different results. The models are visualized in figures 5 and 6. The semantic analogy scores did not explain any significant amount of variation in the BLI scores, $R^2=0.2175$, $F(1, 9)=2.501$, $p=0.1482$. The syntactic analogy scores explained a significant amount of the variation in BLI scores, $R^2=0.803$, $F(1, 9)=36.7$, $p=0.0001887$.

Semantic accuracy

Analogy test



Syntactic accuracy

Analogy test

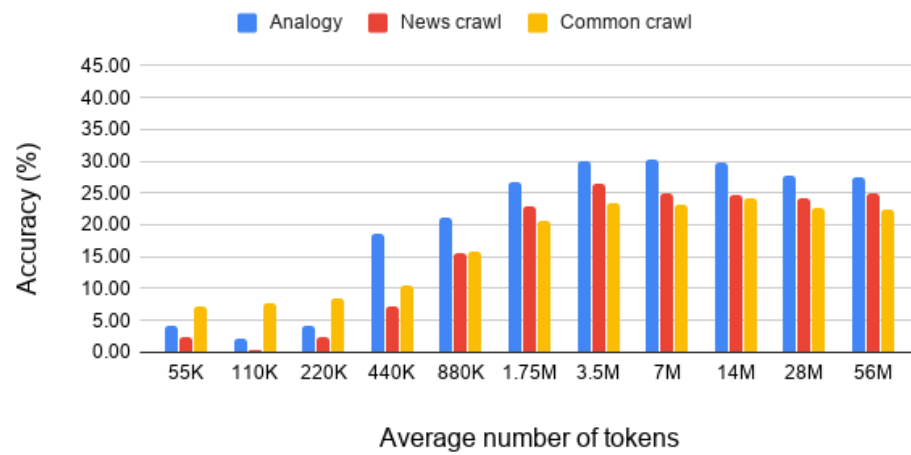


Figure 2: Semantic (top) and syntactic (bottom) accuracy of analogy test.

In domain

Bilingual lexicon induction

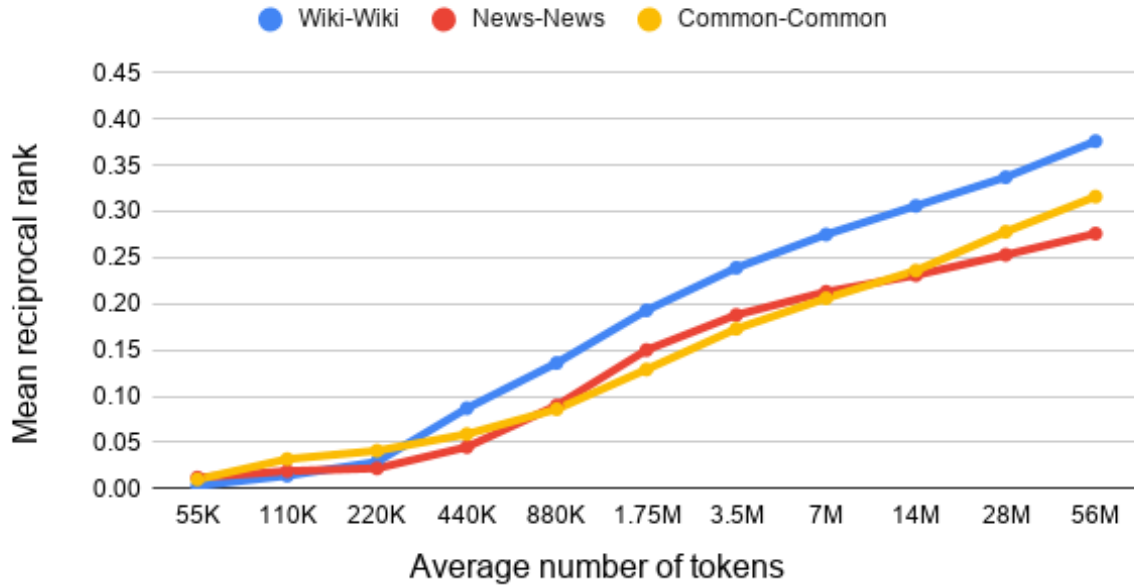


Figure 3: BLI performance of cross-lingual mappings between Wikipedia embeddings (Wiki-Wiki), News crawl embeddings (News-News), and Common crawl embeddings (Common-Common) averaged across languages.

5.5 Results by research question

1) Data size. In general, the monolingual word embeddings did not improve on the analogy test for every increase in data. On the syntactic analogies, around 25% accuracy was reached on only 3.5M tokens, after which accuracy diminished slightly as data size increased. The semantic analogy test saw very mixed results on the smaller data sets, with an increasing trend from the data size of 7M tokens.

Although the coverage of all domains in the analogy test increased steadily, Wikipedia increased significantly faster. Common crawl and News Crawl reached 66% and 62% coverage respectively, on all of the 56M tokens, while Wikipedia reached 72% on the data set of 28M tokens, and reached 83% coverage on the largest data set.

The BLI test of the cross-lingually mapped embeddings did show an increase in MRR with every increase in data. The largest early jump in MRR were between the data sizes of 220K tokens and 440K tokens. For the in-domain mapping of Wikipedia this increase was from 0.02 to 0.09. Furthermore, although only MRR is reported here, accuracy was also recorded. Out of the total 36 runs, in 30 runs accuracy was 0% at the second smallest data set of 110K tokens of data, and in 15 runs accuracy was 0% at the third smallest data set of 220K tokens.

2) Data quality/noise. The difference in level of noise between the three domains showed in the results, although the difference between the News crawl and Common crawl sets was not considerable. The coverage increased steadily with News crawl and Common crawl ending evenly at 62% and 66% respectively. Wikipedia saw a much greater coverage of 83% in the final data size. To note is that all domains performed much better on the syntactic analogies than on the semantic analogies. In particular the Common crawl did very poorly reaching only 3% accuracy on the semantic analogies on the largest data set. The BLI tests also showed a

Out-of-domain

Bilingual lexicon induction

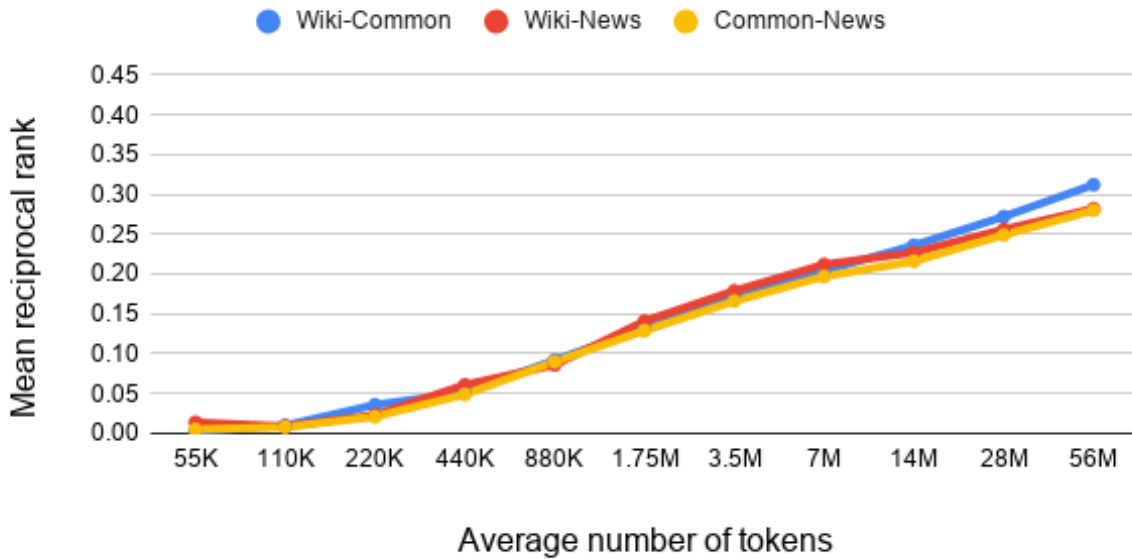


Figure 4: BLI performance of cross-lingual mappings between Wikipedia and Common crawl embeddings (Wiki-Common), Wikipedia and News crawl embeddings (Wiki-News), and between Common crawl and News crawl embeddings (Common-News), averaged across languages.

difference in quality between the domains, at least in terms of Wikipedia being much better. For example, the in-domain mappings of Wikipedia reached an MRR of 0.27 on 7M tokens, while the in-domain mapping of News crawl needed 56M tokens to reach the same MRR.

3) Domain difference. The results of the out-of-domain BLI test shows scores very similar to those of the in-domain mappings of News crawl and Common crawl. In fact, on the largest data set it was the News crawl in-domain mapping that was the lowest at the MRR of 0.27.

4) Analogy/BLI correlation. The two regression analyses performed showed that semantic analogy performance did not predict the BLI performance of the CLWEs. The syntactic analogy performance however, did.

BLI MRR against semantic analogy accuracy

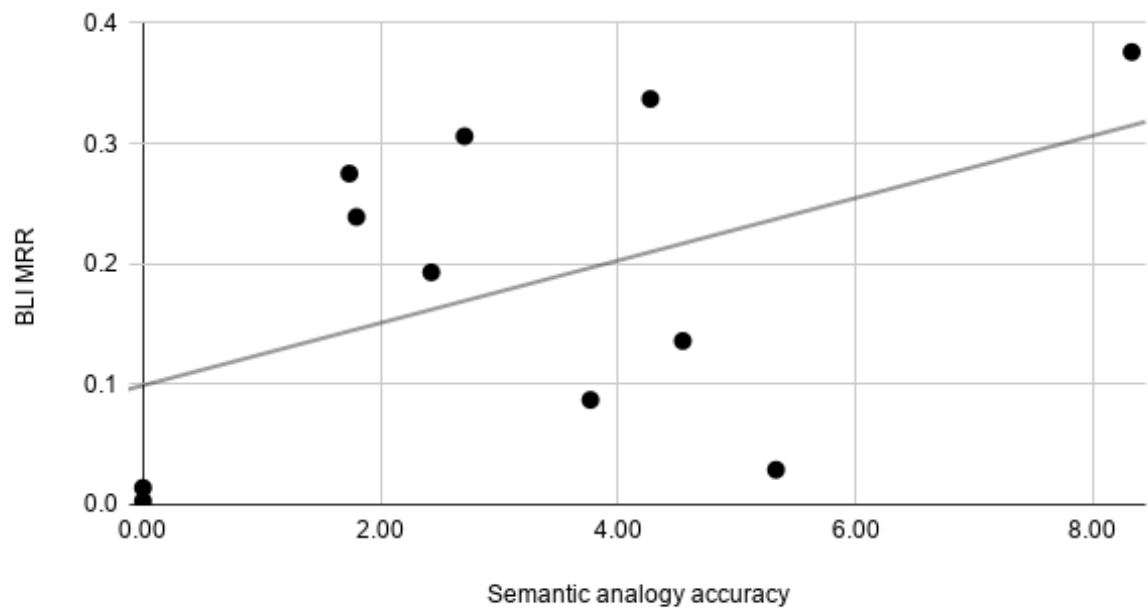


Figure 5: Regression model of the semantic analogy scores of Wikipedia embeddings (independent variable), and the BLI scores of the Wikipedia in-domain CLWEs (dependent variable)

BLI MRR against syntactic analogy accuracy

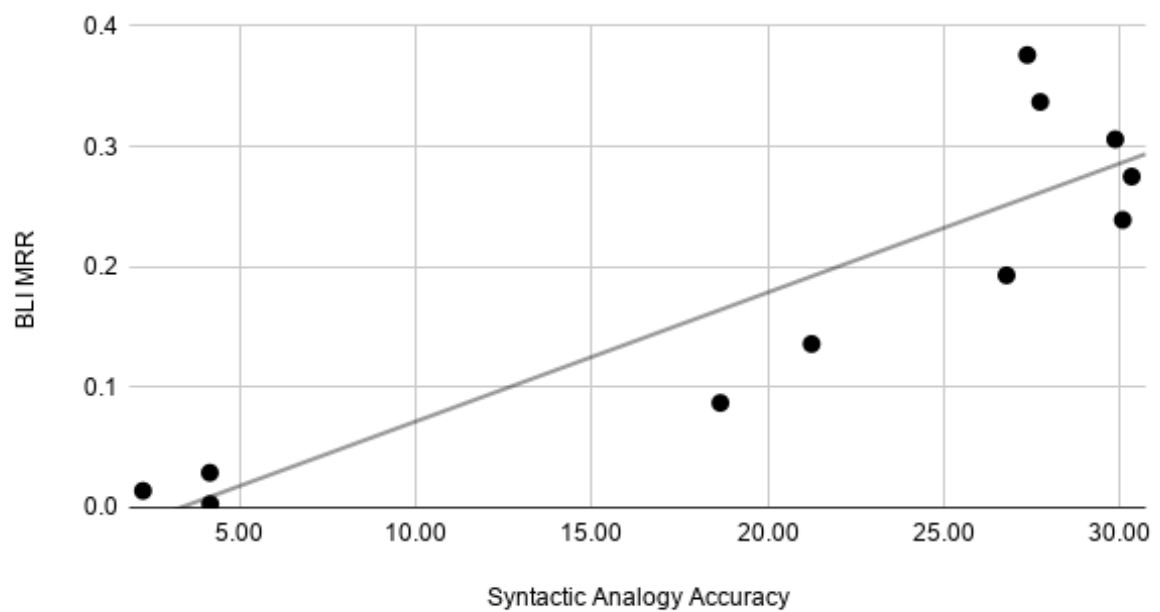


Figure 6: Regression model of the syntactic analogy scores of Wikipedia embeddings (independent variable), and the BLI scores of the Wikipedia in-domain CLWEs (dependent variable)

6 Discussion

To place the results of this study into a broader perspective, I discuss in the following the research questions and hypotheses as well as the methodology and results of the research.

6.1 Discussion by research question

In the following I discuss some potential answers that the results presented in section 5 give to the research questions, and to what extent the hypotheses received support from them. I will in the following sections also discuss to what extent these answers hold in the general sense.

Question 1: How does the amount of training data when learning monolingual word embeddings affect the quality of monolingual word embeddings and of cross-lingual mappings?

In the general case of machine learning models, more training data usually means better results. From the results in this study, it can be concluded that the mapping-based approach to learning CLWE:s prefers monolingual word embeddings learned from large amounts of monolingual training data, judging by the steady increase in MRR with data size increase in both figure 3 and 4. Interestingly, from the results of the monolingual evaluation it seems that larger sets of training data is instead more likely to mean worse performance, this is most pronounced in the syntactic analogy results in figure 2. This is discussed further in the later section 6.3.

Hypothesis 1: Any increase in monolingual training data size means better cross-lingual quality, and better monolingual quality.

The above answer gives support to hypothesis 1, although it appears that analogy evaluation may not properly reflect this in the monolingual setting, especially syntactic evaluation. The syntactic analogy results showed that more training data does not necessarily mean better performance, which can be seen in 2.

Hypothesis 2: There is a minimum amount of data that is needed for any gain in cross-lingual mapping quality.

The results of the BLI tests (figure 3 and figure 4) and syntactic analogy test (bottom of figure 2) indicate that somewhere around the average of 440K tokens is where the minimum lies. As the results of the semantic analogy test (top of figure 2) was more inconsistent, it is difficult to pinpoint a minimum, although one answer is that for performance to show a consistent increase, a minimum of 7M tokens is required.

Question 2: How does noise in the training data when learning monolingual word embeddings affect the quality of the monolingual word embeddings, and cross-lingual mappings respectively?

It is expected that any strings that do not correspond to an expression in the natural language will act as noise during training. This could be misspelling, incorrect grammar, or boiler plate text such as HTML code, tags, links etc. As expected, noise in the training data did have a negative effect on the embedding quality in terms of analogy test and BLI performance. This was most apparent in the BLI test where the in domain mappings of Wikipedia, seen in figure 3, needed

substantially less data than the other two domains to reach the same score.

Hypothesis 3: Word embeddings learned from Wikipedia data will achieve the highest scores. Word embeddings learned from the Common crawl corpus will achieve the lowest scores.

Hypothesis 3 received partial support from the results in this study. Wikipedia data, which can be considered high quality monolingual data, substantially outperformed the two other domains. However, there was no great difference between the domains News crawl and Common crawl.

Question 3: How does domain difference in the monolingual word embedding training data affect the quality of cross-lingual mappings?

The results of the out-of-domain BLI experiment, presented in section 5.3 and visualized in figure 4, when compared to the in domain results, presented in section 5.2 and visualized in figure 3 indicate that domain difference in general has some negative effect, but that the effect is not very large. Furthermore, it was interesting to see that each of the out-of-domain mappings varied less in performance than the in domain mappings, almost as if averaging out the performance.

Hypothesis 4: Domain difference is a considerable disadvantage for the mapping-based approach to learning CLWE:s.

Hypothesis 4 is somewhat supported by comparing the out-of-domain results against the in domain results. Only in the case where the training data is from Wikipedia did the in domain mappings outperform each the out-of-domain mappings by a large margin.

Question 4: To what extent is the quality of monolingual word embeddings a predictor of the quality of their cross-lingual mappings?

Two regression models were presented in section 5.4, and visualized in figure 5 and 6. These models indicate that syntactic (but not semantic) analogy performance is a good predictor of BLI performance.

Hypothesis 5: Analogy test performance correlates well with BLI performance all else being equal.

Hypothesis 5 received partial support, since only syntactic analogy performance showed a reasonably linear relationship with BLI performance.

6.2 Discussion of methodology

6.2.1 Data and language selection

When choosing languages for a study the ideal situation would be to include every language there is. Of course this is not feasible for the scope of a single study. It is important however, to reach for this ideal and choose your languages to be as representative of the variation of natural language as possible. Many studies in the field of word embeddings have failed so far to be representative in this way, with many languages coming from the same language families, and where English is one of the languages in most, or all, language pairs. This is perhaps due to the

lack of large amounts of monolingual training data for many languages.

With this in mind, I chose the languages Finnish, French, Russian, and Turkish. The reasoning for the choice of languages came down to the following factors. Firstly, these are four languages from different language families, except for the Indo-European French and Russian, although these are at least from separate sub-families, which means bias because of language similarity due to family relation should be lower. Secondly, although language similarity was not a research question in this study, I wanted my language selection to be representative of morphological variation (dependent and mixed marking, agglutinative and fusional morphology, number of cases), something that has been shown to challenge the unsupervised mapping-based approach (Søgaard et al. 2018). Thirdly, availability of evaluation data (bilingual lexicons). And finally, other languages that have more available monolingual data were not prioritized because much larger data sets would mean computation times exceeding the scope of this study.

The sources of monolingual training data used in this study was Wikipedia, News crawl, and Common crawl. These were chosen for the purpose to explore the effect of domain difference and data quality on the end result. A problem in this study is that no measure was taken to confirm the level of quality attributed to the data sources. For example one measure that could have been used would be compressing large amounts of data in one language from all three domains and observing how efficiently they were compressed, lower efficiency meaning more noise in the data.

6.2.2 Preprocessing

In the preparation of the monolingual training data I used a pipeline of several preprocessing steps. I chose to perform the first three in line with previous work (Grave et al. 2018). Language identification, deduplication, and tokenization provide better training data and do not cost too much time to perform.

The fourth step of processing was to establish a way to compare data sizes fairly between languages and domains. The number of sentences of a data set is often used as a metric for fairly comparing size. I did not use this as a measure because I used sources which could potentially differ largely in sentence length (which was confirmed by the removal of a large amount of short lines during preprocessing).

I chose instead to use compressed file size as a metric for comparison. In performing this step, my reasoning was that any difference in propositional content between domains, as well as morphosyntactic differences between languages would be minimized.

The implementation of this could have been done better. Firstly, I chose megabytes as the metric for comparing file size. A more accurate way would have been to compare on the level of single bytes, especially on the very small subsets. This was due to a lack of programming knowledge. Secondly, when extracting subsets I did not make sure that every subset was of approximately the same compressed file size, although I manually checked for some of the subsets for each domain and language combination. I did not find any subsets differing more than 1MB in compressed size (which is still a lot for the smaller subsets).

In addition to this, it was quite unexpected to see that 140MB of News crawl data contained a fair bit less amount of word tokens compared to the other two sources. For example, the 140MB of French Wikipedia data used contained about 74M tokens, while the 140MB of French News crawl data used contained about 65M tokens. This would be less surprising if it was true about the Common crawl data. I am not sure what the cause of this discrepancy is. It could be due to difference in topic, for example it may be that news articles in general contain longer words, which would lower the amount of tokens.

6.2.3 Models

I chose the model SGNS with subword information to learn the monolingual word embeddings used in this study. As far as I know, this is no controversial choice, as most studies do. The model learns high quality word embeddings quickly. Perhaps there were parameters that could have produced even better embeddings, but the parameters were the same across each experiment within this study at the least.

The choice of evaluating the supervised mapping-based approach of Artetxe et al. (2016) could perhaps be questioned when there are similar but semi-unsupervised (Artetxe et al. 2017) and even fully unsupervised mapping-based approaches (Artetxe et al. 2018), (Conneau et al. 2017) to inducing CLWE:s. I argue that this choice is warranted for a number of reasons.

Firstly, the practical superiority of unsupervised mapping-based models is questionable at best. The intended purpose of these models is resource low languages where the supervision signal of a bilingual dictionary does not exist and can not easily be induced. There are two reasons for questioning this. Firstly, it is becoming more and more realistic to meet the requirement of a few thousand translated word pairs.

Secondly it has been shown that the mapping-based approach to learning CLWE:s prefers word embeddings learned on large sets of monolingual training data, in this study and others (Søgaard et al. 2018). The problem is that these two scenarios tend to co-occur; if a language does not have enough data to induce a bilingual lexicon of a few thousand word pairs, then probably it does not have enough monolingual training data to properly take advantage of the mapping-based approach in the first place.

Besides the questionable applicability of the unsupervised mapping-based approaches, it has also been shown that they can fail completely as a result of difference in morphology in languages and domain difference in training data (Søgaard et al. 2018). Thus, if the unsupervised mapping-based approaches can not perform better than supervised ones, there seems to be no real reason to prefer them.

6.2.4 Experiments

I used analogy test as an evaluation method for interpreting monolingual word embedding quality. The fact that embedding spaces can encode relations between words as the distances between their embeddings makes this an interesting evaluation method. There are other monolingual evaluation methods to choose from. The scope of this study did not allow for more than one however, and analogy evaluation has been frequently used, and so has many benchmarks for comparison.

When it comes to the application of the method, there were several problems that render the reliability of the results questionable. Firstly, the fact that the words in the evaluation data were automatically translated means that any potential errors went unchecked. It would have been preferred to check the translations for errors with native speakers of each language.

Secondly, there exists alternative data sets to choose from. Gladkova et al. (2016) has been recent work done on unlocking larger data sets for the analogy task. Their data set contains many more categories and also takes synonymy into account by considering more than one correct answer in their gold standard.

To measure the quality of the cross-lingually mapped word embeddings, I chose BLI as the evaluation method. I argue that BLI is a good choice for evaluating cross-lingual quality. Firstly because word translation is always a useful application, and secondly because it is highly interpretable, since it tests that the most similar words across languages are oriented closely together, as one would expect. In a way, the method is both extrinsic and intrinsic. Furthermore, BLI has

been shown to correlate well with the downstream tasks cross-lingual information retrieval and cross-lingual language inference (Glavas et al. 2019).

With these being said, arguably there is still room for interpretability in the evaluation in this work. For example, Søgaard et al. (2018) uses a graph similarity measure to directly quantify the similarity of source and target spaces. A metric like this could perhaps have made the results more reliable in interpretation.

6.3 Discussion of results

6.3.1 Main results

Perhaps the most interesting part of the results in this study is the large contrast between the BLI results and the analogy task results. The BLI scores showed a very steady increase with more training data, every increase in data size meant some increase in performance, and the differences between domains were not very large. On the other hand, the scores of the analogy task experiment were not as consistent. In general, they did not show a steady increase with more training data, with some increases in training data even meaning worse performance. There are several potential explanations for the contrast between the two experiments.

Firstly, the most obvious reason is the sub-optimal evaluation data in the analogy task experiment. The data was automatically translated without correction of any bad translations. Furthermore, the requirement of one-to-one word translations across all four languages meant that two thirds of the queries were excluded, which meant less variation in type of word relation queried for. For example, the semantic queries predominantly consisted of country-city relation queries. The words in these queries can be expected to have lower frequency. This being said, the overall trends did not seem to be a result of bad translations, because tests on untranslated queries on English embeddings gave similar results.

Another factor to consider is the fundamental differences in the evaluation methods. Although both of the methods are types of similarity tests, the queries are in essence different, and the BLI query is arguably of less complexity. The type of query asked in BLI is a simple nearest neighbor question: what embedding y is closest to embedding x ? On the other hand, the query asked in analogy test is instead something like: what embedding y_1 is oriented in relation to x_1 in the same way that y_2 is oriented in relation to x_2 ? I think that at least some of the contrast in the results can be expected to stem from this difference.

Besides the contrast in the results between experiments, there was also an interesting contrast between the two categories in the analogy experiment. The accuracy trend of the syntactic analogy experiment was especially interesting. It quickly jumped to almost 20% on only 440K tokens of training data, then seemed to be saturated on the also relatively small amount of 3.5M tokens and deteriorated from there. Again, the evaluation data was not the best, and the overall accuracy most likely suffered from this, but the saturating trend of the syntactic analogies was also the case on English word embeddings, thus bad translations appears not to be the cause.

Instead I suspect that the cause lies somewhere in the model architecture of SGNS with subword information. This trend has been seen before in similarity evaluation (figure 1 in Bojanowski et al. 2017). In their experiments it can be seen that other word embedding models that are restricted to word level embeddings do not saturate in this way. The specific cause for this is however unclear to me.

While the syntactic analogy performance was decent, the semantic analogies did quite poorly, reaching at the most 8%. I suspect that this was due to the evaluation data, which mostly contained country-capital relation queries. It can be expected that the proper names of cities and

countries do not have a very high frequency of occurrence. This means that there simply isn't enough training samples for the relation to be encoded accurately enough.

I suspect that the two largest factors to consider when comparing results of BLI and analogy tasks is word frequency in the evaluation data, and type of relation queried for in the analogy test. Since it seems that different relations are captured by embeddings to different degrees, not all can be expected to correlate well with BLI performance.

6.3.2 Expectations

Besides the results discussed so far, there were several things about the results that did meet expectations. Firstly, the fact that none of the CLWE:s reached competitive BLI scores comes down to the relatively small amount of training data that was used. Furthermore it was expected that Wikipedia embeddings would be of better quality than those of other domains. Although it was somewhat unexpected that the in domain BLI experiment of the Wikipedia embeddings would outperform the other mappings as much as they did.

On the other hand, it was also expected that domain difference would worsen performance more than what was actually the case. There was no large difference between the out-of-domain mappings and the in domain mappings of News crawl and Common crawl. It could be the case however that these differences accumulates with larger data sizes than was used in this study. Additionally, one can expect that as the topic changes to be more different between domains the larger the effect is. There are of course other domains that make for larger differences than the ones included in this study. An advantage for this study would have been to investigate this by for example quantifying similarity between embeddings, as was done in the work of Sogaard et al. (2018).

I also expected that data quality would have more of an impact than what it seems like it did. I expected the Common crawl embeddings would be substantially outperformed by both of the other domains. However there was no real observable difference in BLI performance between the News crawl and Common crawl embeddings, and only a small difference in the analogy task performance. One thing to keep in mind here is that the News crawl data sets contained somewhat less linguistic data in terms of word tokens.

6.3.3 Related research

Glavas et al. (2019) used a setup very similar to the one in this study. In comparison with their benchmarks on the six language pairs in this study, I achieved higher MRR on five out of six, with the sixth being almost exactly the same score. This is quite surprising when considering the fact that the mapping-based approach benefits a fair amount from larger sets of training data, and the fact that Glavas et al. (2019) indeed used embeddings learned on much larger data sets. Take the language pair Russian-French for example. They achieved an MRR of 0.470 on this language pair with embeddings learned on full Wikipedias for each language – which is more than 1B tokens for French, and more than 800M for Russian. Meanwhile I achieved an MRR of 0.466, basically the same score, on embeddings learned on only 74M tokens of French Wikipedia and 49M tokens of Russian Wikipedia. This means that there is either a mistake somewhere in my experiment or in theirs, or on the other hand that there is some difference in my setup that optimizes the mappings, or the BLI by quite a large amount. Although the experimental setup in this study is quite similar to the one in Glavas et al. (2019), there are a few things that differ, which could potentially explain the discrepancy in the results.

Firstly, I used CSLS retrieval in the BLI, which they did not. CSLS is used to avoid the negative impact of semantic hubs in the embedding space. Semantic hubs are word embeddings

that are close in distance to very many other word embeddings (Radovanovic et al. 2010). The use of CSLS in my experiment probably offers some explanation to the discrepancy, as it has been reported to consistently outperform cosine similarity when used in BLI (Ruder et al. 2019; Søgaard et al. 2018), but it is perhaps not the whole cause. For an exact definition of CSLS see Conneau et al. (2017) or Ruder et al. (2019).

Another difference is the trimming down of vocabularies. It is common practice to trim down word embedding vocabularies to the 200k most common words, which they did but I did not. I did not trim the vocabularies for the single reason that for most data sizes in this study, the vocabularies did not reach this number in any case. However, for the largest data sets, this means that some of the vocabularies were close to or even more than 1M word embeddings in size. Although this difference in setup probably will have affected the results, it is unclear to me whether it is good or bad for the overall performance.

A third difference is another variation of the training data that could potentially affect mapping performance. This has to do with how similar in size the training data sets are for the source embedding space and target embedding space respectively. It is unclear whether a large difference in training data size is better or worse for performance. In my setup, each of the mappings were between embedding spaces that were learned on very similar amounts of training data. In their study however, since they used embeddings learned on full Wikipedias for each of the languages, some of the source and target spaces was learned on training data of very differing sizes. Take the language pair of Finnish-French for example. The full Wikipedia of French is something like 1B tokens, while for Finnish it is closer to 100M tokens. It could be that training data size difference between spaces makes for a great disadvantage, and could be another potential explanation of the high MRR scores in this study, since training data size difference was very small in my experimental setup. Although training data size difference has not been investigated much thus far, I think the relatively high scores in this study in comparison to previous research warrants a closer look at just how large this effect can be.

Søgaard et al. (2018) investigated the impact of language difference as well as domain difference on the unsupervised mapping-based approach of Conneau et al. (2017). In comparison with their reported effect of language difference, the results in this study showed no obvious effect when it comes to dependency marking, morphological type (agglutinative, fusional etc), or number of cases. And although language similarity was not a main research question in this study, I include the BLI scores of the in domain Wikipedia embeddings, for each of the individual language pairs in table 3 in appendix A, for future reference on this topic.

In comparison with their results on domain difference, they observed a much larger negative effect of domain difference on BLI performance than what was reported in this study. This can be explained by several factors. First and most obvious, is the difference in model architecture. The unsupervised model of Conneau et al. (2017) uses adversarial training to find an initial mapping matrix. It can be expected that this technique is a lot more sensitive to differences in embedding spaces. In addition to this, they used other domains that could perhaps give rise to larger differences, for example they used a medical domain corpus.

A problem here is that this consideration is left unconfirmed because of the fact that I did not use any method for directly measuring the differences in embedding spaces. Thus, it could be that embedding spaces in this study were quite different, but that this did not have much impact on the results. Or, on the other hand, embedding spaces may have actually been quite similar. It would of course have been preferable to have included some measurement of embedding space difference.

6.3.4 Practical applications

The experiments of this study were performed in order to investigate the advantages and disadvantages that certain factors in the training data lends to the mapping-based approach to learning CLWEs. Looking at the results of these experiments, there are a few practical applications that one can keep in mind when applying the model.

Overall, it can be said that the mapping-based approach in its supervised form is a very robust model for learning CLWEs. The model prefers large amounts of data, but can be expected to pick up in performance relatively quickly. Noisy data does disadvantage the model and high quality data like that of Wikipedia can reduce the amount of training data needed by a substantial amount. Domain difference does not seem to have such a large impact as it does on unsupervised mapping-based approaches.

When it comes to predicting the quality of mapped CLWEs, syntactic analogy performance can be used as an indicator. That is, there seems to be a reasonably linear relation between monolingual quality and cross-lingual quality of word embeddings, at least in terms of analogy task performance and BLI performance. The extent to which correlation test results mirrors this relationship depends largely on what analogy relations is queried for and ultimately on the word frequency of the words in the evaluation data.

Since the results of the analogy experiments in this study supports the notion that the successful encoding of word relations in embedding spaces depends largely on the frequency of occurrence of the given words, I would propose to investigate this by performing experiments on varying sizes of training data using a more comprehensive set of relation queries such as that of Gladkova et al. (2016).

6.4 Future research

Previous research has investigated how well different properties of word embeddings (dimensions, model, parameters etc.) capture different word relations (Gladkova et al. 2016). The results on the analogy task in this study indicate that the performance as a function of data size depends largely on what kind of relation is being tested for. It would be interesting to further investigate this to see whether some word relations are better captured at specific sizes of training data.

In the cross-lingual setting, I would in the future like to investigate how training data size difference between source and target embedding spaces affect the quality of cross-lingual mappings. Most studies do not take this into account when applying the mapping-based approach. If training data size difference turns out to negatively affect cross-lingual mapping quality, then this would warrant an additional preprocessing step of controlling for size in future studies.

7 Conclusions

To summarize this study I again revisit the research questions with reliability and generalizability and the above discussion in mind, concluded with a brief summary.

Question 1: How does the amount of training data when learning monolingual word embeddings affect the quality of monolingual word embeddings and of cross-lingual mappings?

In the case of cross-lingual mapping quality, this study confirmed that performance on bilingual lexicon induction depends on having embeddings learned on large amounts of monolingual training data. On the other hand, the analogy task results did not provide support for the hypothesis that more data equals better monolingual quality.

Question 2: How does noise in the training data when learning monolingual word embeddings affect the quality of the monolingual word embeddings, and cross-lingual mappings respectively?

Monolingual training data of lower quality has a negative effect on both monolingual quality, in terms of analogy task performance, and cross-lingual quality, in terms of BLI performance. This means that lower quality data may need to be of many times larger size to perform on par with high quality data, such as Wikipedia text.

Question 3: How does domain difference in the monolingual word embedding training data affect the quality of cross-lingual mappings?

This study showed that domain difference has a negative effect on the quality of cross-lingual mappings, in terms of out-of-domain mappings performing worse than in domain mappings on the BLI task.

Question 4: To what extent is the quality of monolingual word embeddings a predictor of the quality of their cross-lingual mappings?

This study supported the idea that the quality of monolingual word embeddings, in terms of analogy task performance, predict the quality of their cross-lingual mappings. However, this depends on what relation is queried for. In particular, this study showed that syntactic analogy performance, as opposed to semantic analogy performance, are better predictors of BLI performance.

In conclusion, this study showed that the mapping-based approach to learning cross-lingual word embeddings is a robust method, but that there are several properties of the monolingual training data that needs to be taken into account when applying it. In addition, syntactic analogy task performance was shown to be a good predictor of cross-lingual quality. Besides the factors investigated in this study, training data size difference between between source and target embedding spaces is also of potential interest for future work.

References

- Artetxe, Mikel, Gorka Labaka and Eneko Agirre (July 2018). ‘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)*. Melbourne, Australia: Association for Computational Linguistics, pp. 789–798. DOI: [10.18653/v1/P18-1073](https://doi.org/10.18653/v1/P18-1073). URL: <https://www.aclweb.org/anthology/P18-1073>.
- (July 2017). ‘Learning bilingual word embeddings with (almost) no bilingual data’. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 451–462. DOI: [10.18653/v1/P17-1042](https://doi.org/10.18653/v1/P17-1042). URL: <https://www.aclweb.org/anthology/P17-1042>.
- (Nov. 2016). ‘Learning principled bilingual mappings of word embeddings while preserving monolingual invariance’. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2289–2294. DOI: [10.18653/v1/D16-1250](https://doi.org/10.18653/v1/D16-1250). URL: <https://www.aclweb.org/anthology/D16-1250>.
- Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post and Marcos Zampieri (Aug. 2019). ‘Findings of the 2019 Conference on Machine Translation (WMT19)’. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pp. 1–61. DOI: [10.18653/v1/W19-5301](https://doi.org/10.18653/v1/W19-5301). URL: <https://www.aclweb.org/anthology/W19-5301>.
- Bengio, Yoshua, Réjean Ducharme, Pascal Vincent and Christian Jauvin (2003). ‘A neural probabilistic language model’. In: *Journal of machine learning research* 3.Feb, pp. 1137–1155.
- Bird, Steven, Ewan Klein and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. ‘O’Reilly Media, Inc.’
- Bojanowski, Piotr, Edouard Grave, Armand Joulin and Tomas Mikolov (2017). ‘Enriching word vectors with subword information’. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.
- Conneau, Alexis, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer and Hervé Jégou (2017). ‘Word translation without parallel data’. In: *arXiv preprint arXiv:1710.04087*.
- Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer and Richard Harshman (1990). ‘Indexing by latent semantic analysis’. In: *Journal of the American society for information science* 41.6, pp. 391–407.
- Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman and Eytan Ruppín (2001). ‘Placing search in context: The concept revisited’. In: *Proceedings of the 10th international conference on World Wide Web*, pp. 406–414.
- Gladkova, Anna, Aleksandr Drozd and Satoshi Matsuoka (2016). ‘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*, pp. 8–15.
- Glavas, Goran, Robert Litschko, Sebastian Ruder and Ivan Vulic (2019). ‘How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions’. In: *arXiv preprint arXiv:1902.00508*.
- Gower, John C, Garnt B Dijkstra et al. (2004). *Procrustes problems*. Vol. 30. Oxford University Press on Demand.

- Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin and Tomas Mikolov (2018). ‘Learning word vectors for 157 languages’. In: *arXiv preprint arXiv:1802.06893*.
- Hill, Felix, Roi Reichart and Anna Korhonen (2015). ‘Simlex-999: Evaluating semantic models with (genuine) similarity estimation’. In: *Computational Linguistics* 41.4, pp. 665–695.
- Huang, Eric H, Richard Socher, Christopher D Manning and Andrew Y Ng (2012). ‘Improving word representations via global context and multiple word prototypes’. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pp. 873–882.
- Joos, Martin (1950). ‘Description of language design’. In: *The Journal of the Acoustical Society of America* 22.6, pp. 701–707.
- Juola, Patrick (2008). ‘Assessing linguistic complexity’. In: *Language complexity: Typology, contact, change*, pp. 89–108.
- Jurafsky, Dan (2000). *Speech & language processing*. Pearson Education India.
- Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean (2013). ‘Efficient estimation of word representations in vector space’. In: *arXiv preprint arXiv:1301.3781*.
- Mikolov, Tomas, Quoc V Le and Ilya Sutskever (2013). ‘Exploiting similarities among languages for machine translation’. In: *arXiv preprint arXiv:1309.4168*.
- Osgood, Charles Egerton, George J Suci and Percy H Tannenbaum (1957). *The measurement of meaning*. 47. University of Illinois press.
- Pilehvar, Mohammad Taher and Jose Camacho-Collados (2018). ‘Wic: the word-in-context dataset for evaluating context-sensitive meaning representations’. In: *arXiv preprint arXiv:-1808.09121*.
- Radovanovic, Milos, Alexandros Nanopoulos and Mirjana Ivanovic (2010). ‘Hubs in space: Popular nearest neighbors in high-dimensional data’. In: *Journal of Machine Learning Research* 11.sept, pp. 2487–2531.
- Ruder, Sebastian, Ivan Vulić and Anders Søgaard (2019). ‘A survey of cross-lingual word embedding models’. In: *Journal of Artificial Intelligence Research* 65, pp. 569–631.
- Schnabel, Tobias, Igor Labutov, David Mimno and Thorsten Joachims (2015). ‘Evaluation methods for unsupervised word embeddings’. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 298–307.
- Søgaard, Anders, Sebastian Ruder and Ivan Vulić (2018). ‘On the limitations of unsupervised bilingual dictionary induction’. In: *arXiv preprint arXiv:1805.03620*.
- Tsvetkov, Yulia, Manaal Faruqui, Wang Ling, Guillaume Lample and Chris Dyer (2015). ‘Evaluation of word vector representations by subspace alignment’. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2049–2054.
- Xing, Chao, Dong Wang, Chao Liu and Yiye Lin (2015). ‘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*, pp. 1006–1011.
- Yano, Tae and Moonyoung Kang (2016). *Taking advantage of wikipedia in natural language processing*. Tech. rep. Technical report, Carnegie Mellon University Language Technologies Institute.

Appendix A BLI scores

Table 3: Mean reciprocal rank (MRR) and accuracy (ACC) of each language pair of the in domain mappings of Wikipedia on bilingual lexicon induction.

FI-FR		FI-RU		RU-FR		TR-FI		TR-FR		TR-RU	
MRR	ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR	ACC
0.404	32.36	0.355	27.65	0.466	39.24	0.308	23.07	0.414	33.16	0.312	23.54
0.356	27.75	0.321	24.9	0.421	33.82	0.275	20.26	0.365	28.26	0.285	21.72
0.31	23.21	0.291	22.28	0.385	30.76	0.257	19.22	0.321	24.47	0.271	20.38
0.266	19.58	0.265	20.19	0.354	28.1	0.226	16.11	0.29	22.18	0.25	18.55
0.23	16.73	0.231	17.01	0.323	25.93	0.203	14.76	0.242	17.96	0.207	14.14
0.2	14.15	0.186	12.33	0.232	16.92	0.179	12.41	0.19	13	0.173	12.07
0.119	7.47	0.148	9.58	0.164	11.75	0.14	9.81	0.123	7.84	0.124	8
0.091	5.49	0.085	3.54	0.086	5.24	0.096	5.52	0.075	4.84	0.088	4.7
0.02	0	0.012	0	0.023	1.33	0.045	0	0.047	1.54	0.027	0
0.001	0	0.002	0	0.014	0	0.04	3.85	0.028	1.79	0.001	0
0.002	0	0.002	0	0.007	0	0.002	0	0.003	0	0.003	0

Stockholm University
SE-106 91 Stockholm, Sweden
Telephone +46 (0)8 16 20 00
<https://www.su.se/>



Stockholm
University