Revisiting the linearity in cross-lingual embedding mappings: from a perspective of word analogies

Xutan Peng[†], **Chenghua Lin**[†], **Mark Stevenson**[†], **Chen li**[‡]

[†]NLP Group, Department of Computer Science, The University of Sheffield
[‡]School of Computer Science and Engineering, Beihang University
{x.peng, c.lin, mark.stevenson}@sheffield.ac.uk,
lichen@act.buaa.edu.cn

Abstract

Most cross-lingual embedding mapping algorithms assume the optimised transformation functions to be linear. Recent studies showed that on some occasions, learning a linear mapping does not work, indicating that the commonly-used assumption may fail. However, it still remains unclear under which conditions the linearity of cross-lingual embedding mappings holds. In this paper, we rigorously explain that the linearity assumption relies on the consistency of analogical relations encoded by multilingual embeddings. We did extensive experiments to validate this claim. Empirical results based on the analogy completion benchmark and the BLI task demonstrate a strong correlation between whether mappings capture analogical information and are linear.

1 Introduction

Bilingual dictionary is a fundamental component of cross-lingual natural language processing applications. Traditional pipelines for bilingual dictionary construction heavily rely on expert knowledge, which is unpractical in low-resource scenarios. To address this limitation, one line of approaches resort to pre-trained monolingual word embeddings, and project them to a shared embedding space. By retrieving the neighbouring cross-lingual pairs from this space, one can then perform automatic word translation, i.e., Bilingual Lexicon Induction (BLI). As this strand of work can produce high-quality bilingual dictionaries with weak or even no supervision (Artetxe et al., 2018b; Lample et al., 2018a; Ruder et al., 2019), it has been applied to a range of downstream tasks such as unsupervised machine translation (Lample et al., 2018b), zero-shot cross-lingual transfer learning (Hsu et al., 2019), and data augmentation for minority languages (Kumar et al., 2019).

One of the key challenges of embedding based BLI is the design of cross-lingual embedding mapping functions. Motivated by the empirical observation that word embeddings of different languages tend to preserve similar shapes (Mikolov et al., 2013b), a large number of works assume that the mappings between cross-lingual embeddings are linear and hence the learned transformation functions (Faruqui and Dyer, 2014; Xing et al., 2015; Artetxe et al., 2016; Lample et al., 2018a; Ruder et al., 2019). While algorithms based on this linear assumption have produced promising results, it has also been suggested that linear transformation might not always hold between cross-lingual embeddings, leading to the development of works modelling non-linear mappings (Lu et al., 2015) or relaxing the linear assumption (Nakashole, 2018; Patra et al., 2019; Zhang et al., 2019; Lubin et al., 2019). However, the non-linear methods do not seem to be as effective as the linear ones. Yet, all the aforementioned works are purely grounded on empirical observations, and there is little work on providing theoretical insights regarding the legitimacy of the commonly used linearity assumption.

In this paper, we revisit the linearity assumption for cross-lingual mappings of embeddings from a novel departure, where we formally establish the link between the linearity of cross-lingual embedding mappings and word analogies. Our work is motivated by the observation that word analogies, a form of linguistic regularity, allow the composition of semantics via vector arithmetic (Mikolov et al., 2013c; Levy and Goldberg, 2014; Drozd et al., 2016). We hypothesise that such a semantic composition should be transferable across languages, and that the better semantic composition in the monolingual embedding spaces is maintained, the stronger linearity between the cross-lingual embedding mappings. This can also be interpreted as the analogical invariance across multilingual vo-

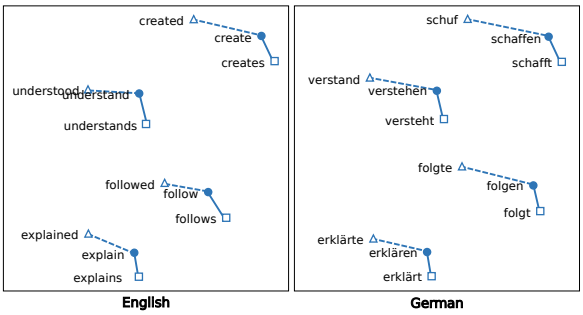


Figure 1: A motivating case for our theory. Word vectors are sampled from the aligned space of English *Word2Vec* and German *Wikipedia2Vec* embeddings in § 4. They are visualized with PCA.

cabularies. For example, Figure 1 shows some analogical relations such as `past-tense` and `plural-verbs`, whose corresponding parallelogram structures are roughly identical in the pre-trained embeddings of English and German.

We verify our claim both theoretically and empirically. First, we prove that the linearity of cross-lingual embedding mappings depends on the preservation of analogical information encoded in monolingual vector spaces. To our knowledge, this is the first attempt to theoretically explore the conditions under which the linearity assumption holds. We also conducted extensive experiments based on analogy reasoning and BLI tasks. Empirical results between English and four other languages (i.e., German, French, Italian, and Polish) indicate a strong correlation between the accuracy of the cross-lingual mapping and performance on the analogy completion task. These empirical results provide support for our claim.

2 Related work

2.1 Linearity assumption for cross-lingual embedding mappings

Mikolov et al. (2013b) first discovered that across the embedding spaces of different languages, the vectors of many word translations share similar structures. Inspired by this observation, a large body of work utilises the geometric similarity as a basic assumption where the designed cross-lingual embedding mapping functions are typically linear (Faruqui and Dyer, 2014; Xing et al., 2015; Artetxe et al., 2016; Ammar et al., 2016). Such a simple assumption turns out to be quite effective, where linear cross-lingual embedding mappings are able to achieve good accuracy with weak or even zero supervision (Smith et al., 2017; Artetxe et al., 2017, 2018a; Lample et al., 2018a).

Despite its effectiveness, it has been pointed out in a number of works that linear cross-lingual mappings of embeddings are less effective under more-challenging conditions, such as when dealing with rare words, distant language pairs, or embeddings trained on corpora from diverse domains (Søgaard et al., 2018; Ruder et al., 2019; Vulić et al., 2019). Søgaard et al. (2018) examined embeddings trained separately using different corpora, models and parameters and concluded that the dissimilarity in settings will break down invariance in structures, leading to a substantial decline in mapping performance. Patra et al. (2019) investigated various language pairs and discovered that a higher etymological distance weakens cross-lingual linearity. Nakashole and Flauger (2018) claimed that linearity holds on cross-lingual embedding mappings between geometrically-local regions rather than the entire space, which inspired follow-up research to exploit neighbourhood sensitive projections instead of a single linear transformation (Nakashole, 2018). However, their insight is grounded on empirical studies involving a few anchor words, and is not supported by theoretical explanations. In § 5.2.2, we show that the judgement of Nakashole and Flauger (2018) is true given the condition that linearity holds locally rather than globally.

In contrast to works adopting the linearity assumption, there are several attempts to learn non-linear transformations. However, existing works (Mikolov et al., 2013b; Lazaridou et al., 2015; Zhang et al., 2017; Bai et al., 2019) reveal that non-linear functions such as neural networks do not improve performance compared to models based on the linearity assumption; in addition, optimization for non-linear methods can be unstable. Consequently, some researchers have turned to relax the linearity assumption rather than developing full non-linear mapping functions: Zhang et al. (2019) improved the cross-lingual structural similarity by iteratively normalizing monolingual embeddings during training; Patra et al. (2019) developed a semi-supervised model to loosen the restriction of linearity; Lubin et al. (2019) attempted to refine learned dictionaries by reducing the noise of cross-lingual non-linearity.

Nevertheless, all these studies are centered around experimental practice only but lack theoretical insights. It is still unclear under what condition the linearity assumption holds.

2.2 Learning analogies with word vectors

One intriguing phenomenon in word embedding models is that they can reconstruct the so-called *linguistic regularities*: semantics can be composed via vector arithmetic, e.g., the most famous example is "*king - man = queen - woman*" (Mikolov et al., 2013c).

Since this discovery, the community has embraced the analogy completion task (aka. analogy reasoning) as a common standard for evaluating the quality of pre-trained word embeddings (Mikolov et al., 2013c; Pennington et al., 2014; Levy and Goldberg, 2014). In addition,by serving as signals for the inference of semantic hierarchies and relations, the strength of embedding models in restoring analogical information also benefits other applications of representation learning (Bordes et al., 2013; Fu et al., 2014).

To understand why vector offsets mirror word analogies, Pennington et al. (2014) first proposed a widely-cited and intuitive conjecture by probabilistically paraphrasing the vocabulary. Based on the Pointwise Mutual Information (PMI) matrix of word co-occurrences, succeeding work built on this by providing theoretical proofs and empirical evidence (Arora et al., 2016; Gittens et al., 2017; Allen and Hospedales, 2019; Ethayarajh et al., 2019).

One outstanding property of a linear mapping is that any two parallel lines are still parallel after the mapping, and the ratio of their lengths is also preserved. Intuitively speaking, this is mathematically symmetric with the linear nature of monolingual word embeddings which can reflected by completing analogies using vector offset methods. Surprising as it sounds, to the best of our knowledge, this is the first study to associate word analogy with linear embedding mappings.

3 Modeling cross-lingual linearity

Given two parallel word sets S_X and S_Y, let X and Y respectively denote their monolingual embeddings. In this section, we formally prove that the objective cross-lingual embedding mapping $\mathbf{M} : X \to Y$ is linear *iff* the following condition is satisfied:

$$\begin{aligned} &\forall \mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma, \mathbf{x}_\theta \in X, \mathbf{x}_\alpha - \mathbf{x}_\beta = \mathbf{x}_\gamma - \mathbf{x}_\theta \\ \iff &\mathbf{M}(\mathbf{x}_\alpha) - \mathbf{M}(\mathbf{x}_\beta) = \mathbf{M}(\mathbf{x}_\gamma) - \mathbf{M}(\mathbf{x}_\theta). \end{aligned} \tag{1}$$

That is, if an analogical relation exists in embedding space X, then the same relation should also exist in embedding space Y.

3.1 Interpreting the condition

We observe many occasions on which Eq. 1 approximately holds (one of them is shown in Figure 1). These phenomena are not surprising and appear intuitively plausible.

Previous works (see § 2.2) have demonstrated both empirically and theoretically that when analogical relations are encoded in an embedding space, they can be reconstructed via vector arithmetic. Geometrically speaking, we can link the parallelogram structures defined by vector end points with analogical relations existing in the vocabulary (Ethayarajh et al., 2019). Since the task of word translation is to pair semantically equivalent words, up to noise such as corpus bias, lexical variance, etc., $\mathbf{M}$ can be expected to preserve information about analogies across vocabularies. Parallelogram structures are therefore predicted to stay invariant in both embedding spaces.

3.2 Proving the linearity

Starting with a general mapping $\mathbf{M}' : X' \to Y'$, where both X' and Y' are $\mathbb{R}^n$ such that

$$\begin{aligned} &\forall \mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma, \mathbf{x}_\theta \in X', \mathbf{x}_\alpha - \mathbf{x}_\beta = \mathbf{x}_\gamma - \mathbf{x}_\theta \\ \iff &\mathbf{M}'(\mathbf{x}_\alpha) - \mathbf{M}'(\mathbf{x}_\beta) = \mathbf{M}'(\mathbf{x}_\gamma) - \mathbf{M}'(\mathbf{x}_\theta). \end{aligned} \tag{2}$$

Since $\mathbf{M}'$ satisfies Eq. 1, the linearity of $\mathbf{M}$ can be proved as long as $\mathbf{M}'$ is shown to be a linear transformation.

One commonly used pre-processing technique for cross-lingual embedding mapping is *mean centering*, which shifts the coordinate so that its origin coincides with the sum of all monolingual word vectors, i.e., the "semantic sum" of the vocabulary (Artetxe et al., 2016; Lample et al., 2018a; Ruder et al., 2019). As cross-lingual mappings of embeddings aim to project every point to its semantically equivalent position, ideally the "semantic sum" of the source vocabulary should be mapped to that of the target vocabulary, either of which is now $\vec{0}$. To simply the derivation, *w.l.o.g.*, we only consider embeddings normalised via mean centering:

$$\mathbf{M}'(\vec{0}) = \mathbf{M}(\vec{0}) = \vec{0}. \tag{3}$$

By definition, $\mathbf{M}'$ is a linear transformation *iff* it preserves the operations of addition (aka. additivity) and scalar multiplication (aka. homogeneity).

Additivity can be proven by observing that

$$\forall \mathbf{x_i}, \mathbf{x_j} \in X', \mathbf{x_i} - \vec{0} = (\mathbf{x_i} + \mathbf{x_j}) - \mathbf{x_j} \quad (4)$$

always holds, provided Eq. 2 and Eq. 3, consequently

$$\begin{aligned} \text{Eq. 4} &\iff \mathbf{M}'(\mathbf{x_i}) - \mathbf{M}'(\vec{0}) = \mathbf{M}'(\mathbf{x_i} + \mathbf{x_j}) - \mathbf{M}'(\mathbf{x_j}) \\ &\iff \mathbf{M}'(\mathbf{x_i} + \mathbf{x_j}) = \mathbf{M}'(\mathbf{x_i}) + \mathbf{M}'(\mathbf{x_j}). \end{aligned} \quad (5)$$

For homogeneity, to begin with, since

$$\forall \mathbf{x_i} \in X', -\mathbf{x_i} - \vec{0} = \vec{0} - \mathbf{x_i} \quad (6)$$

always holds, similar to Eq. 5 we can show that

$$\begin{aligned} \text{Eq. 6} &\iff \mathbf{M}'(-\mathbf{x_i}) - \mathbf{M}'(\vec{0}) = \mathbf{M}'(\vec{0}) - \mathbf{M}'(\mathbf{x_i}) \\ &\iff \mathbf{M}'(\mathbf{x_i}) = -\mathbf{M}'(-\mathbf{x_i}). \end{aligned} \quad (7)$$

Next, by induction we prove the following lemmas[1]:

Lemma 1. $\forall p \in \mathbb{Z}^+, \mathbf{M}'(p\mathbf{x_i}) = p\mathbf{M}'(\mathbf{x_i})$;

Lemma 2. $\forall q \in \mathbb{Z}^+, \mathbf{M}'(\frac{\mathbf{x_i}}{q}) = \frac{\mathbf{M}'(\mathbf{x_i})}{q}$.

Base case: When $p = q = 1$, both statements trivially hold.
Inductive step: Assume the induction hypothesis that $p = q = t$ $(t \in \mathbb{Z}^+)$ is true, we should verify both statements when $p = q = t+1$.

For Lemma 1, with Eq. 5 we have

$$\mathbf{M}'(p\mathbf{x_i}) = \mathbf{M}'(t\mathbf{x_i}) + \mathbf{M}'(\mathbf{x_i}), \quad (8)$$

so with the induction hypothesis, we get

$$\mathbf{M}'(p\mathbf{x_i}) = t\mathbf{M}'(\mathbf{x_i}) + \mathbf{M}'(\mathbf{x_i}) = p\mathbf{M}'(\mathbf{x_i}), \quad (9)$$

which proves Lemma 1.

As for Lemma 2, the induction hypothesis implies that

$$\mathbf{M}'(\frac{\mathbf{x_i}}{q}) = \mathbf{M}'(\frac{1}{t}\frac{t\mathbf{x_i}}{q}) = \frac{1}{t}\mathbf{M}'(\frac{t\mathbf{x_i}}{t+1}). \quad (10)$$

With Eq. 5 and Eq. 7, algebraically we see that

$$\begin{aligned} \text{Eq. 10} &\iff \mathbf{M}'(\frac{\mathbf{x_i}}{q}) = \frac{1}{t}(\mathbf{M}'(\mathbf{x_i}) + \mathbf{M}'(-\frac{\mathbf{x_i}}{t+1})) \\ &\iff t\mathbf{M}'(\frac{\mathbf{x_i}}{q}) = \mathbf{M}'(\mathbf{x_i}) - \mathbf{M}'(\frac{\mathbf{x_i}}{q}) \\ &\iff \mathbf{M}'(\frac{\mathbf{x_i}}{q}) = \frac{\mathbf{M}'(\mathbf{x_i})}{t+1} = \frac{\mathbf{M}'(\mathbf{x_i})}{q}, \end{aligned} \quad (11)$$

so Lemma 2 is proved as well.

Summarizing Eq. 3, Eq. 7 and lemmas, we justify the homogeneity, and further, the linearity of $\mathbf{M}'$[2]. Therefore, we conclude that $\mathbf{M}$ is linear *iff* it satisfies our proposed condition.

4 Empirical studies

This section reports experiments which validate our theory by demonstrating a strong relationship between our condition and the linearity of cross-lingual embedding mappings[3].

For generality, we represent a range of etymological distances by choosing English as the source language and the following four languages as targets: a Germanic language (German), two Romance languages (Italian and French) and a Slavic language (Polish).

4.1 Setup and data

The extent to which a cross-lingual mapping of embeddings is linear can be quantified as the BLI precision it achieves[4]. However, it is non-trivial to measure to what extent a parallel word set satisfies the proposed condition specified in Eq. 1. One practical solution is to first identify some cross-lingual word pairs with known analogy relations, as we know that the parallelogram structures of monolingual word vectors are related to the analogical relations as discussed in § 3.1. Next, we can perform the analogy completion task based on the analogy word pairs, and the resulting completion accuracy can be used as a measure for indicating how well the analogical relations are preserved in the embeddings.

A thorough search of the relevant resources yields only one possible starting point: the Google Analogy Test Set (GATS) for English (Mikolov et al., 2013c) and its variants for the target languages (Köper et al., 2015; Berardi et al., 2015; Grave et al., 2018). Designed for analogy completion tasks, all these test sets are complied according to similar standards so that they have several types of pre-defined relations in common; besides, many of their word pairs are parallel.

[1] As the major premise, $\forall \mathbf{x_i} \in X'$ is omitted for brevity in both lemmas and their proof.

[2] Pre-trained embeddings are stored as rational numbers, so our proof for homogeneity over $\mathbb{Q}^n$ is sufficient given the scope of this paper. However, since $\mathbb{Q}$ is a *dense set*, homogeneity of $\mathbf{M}'$ holds over $\mathbb{R}^n$ as well. Lack of space forbids detailed treatment of this topic here.

[3] Code and data will be released soon.

[4] Given all embeddings have been normalised with mean centering.

Unfortunately, they are too small to generate reliable results for our experiments (containing between 20 to 70 words per relation). To decide the 300×300 dimensional matrices for projecting embeddings in § 4.1.2, at least 300 linearly independent word vectors are required for each language.

4.1.1 Bootstrapping word sets

Parallel analogical word sets: To address the above issue, the morphological analogies in GATS and its multilingual variants were extended.[5] Inflections and derivations from public lexicons were used: MULTEXT-East (Erjavec et al., 2010) for English and French, DEMorphy (Altinok, 2018) for German, Morph-it (Zanchetta and Baroni, 2005) for Italian and PoliMorf (Woliński et al., 2012) for Polish. To make our data sets parallel, for each analogical relation of each language pair[6], we utilised the corresponding dictionary from the MUSE project[7] to align available entries. During training and testing the cross-lingual mappings of embeddings, we only matched words with the same form and tense to ensure rigour, e.g., a French adverb and an English adjective would not get paired even if they make a sensible translation in the MUSE dictionary. Next, cross-lingual word sets with less than 1.0k English words left were omitted. Finally, for each language pair, we sampled the cross-lingual analogical word sets so that they contain the same amount of English words (see Figure 2) with balanced frequencies. In the "Relation" rows of Table 1, we listed the analogical relations included in the final word sets.

Parallel random word sets: Besides the analogical word sets above, for each language pair, we also randomly sampled cross-lingual word pairs from the MUSE dictionaries. We ensured that these parallel random word sets are comparable to their analogical counterparts in terms of the unique pair amounts and the English token frequencies. However, note that we were unable to enforce each pair in random sets to have the same word class as well, because morphology varies across languages, e.g., in English, there is no morphological equivalent for a Polish *perfective* verb.

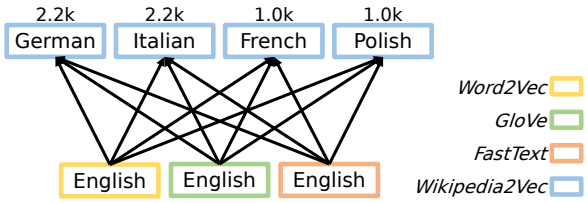


Figure 2: A summary of our experimental setup. Above each target language is the number of unique pairs per parallel word set.

Analogical relations in random word sets are far more complex than the ones in the analogical counterparts, which is similar to the real-world scenarios. Therefore, in our experiments we exploited vectors of the random word sets to baseline the embeddings of the large-scale vocabularies.

4.1.2 Pre-trained embeddings

To support replicability, we downloaded two collections of open-source 300-dimensional embeddings (please refer to their pages for training details): for English, we selected word vectors pre-trained by three representative models (*Word2Vec*, *GloVe* and *FastText*)[8]; for target languages, so as to exclude the potential interference, we chose embeddings pre-trained by another algorithm, *Wikipedia2Vec*[9], and filtered out vectors of entities. For consistency with § 3, we normalised all word vectors by mean centering before experiments. The mapping pairing of all involved embeddings is illustrated in Figure 2.

These word vector models share two advantages: they are all trained on very similar Wikipedia Dumps, and their vocabularies contain all words in the parallel word sets.

4.2 Benchmarks

4.2.1 Testing the condition

The analogy completion task was first run on each monolingual embedding, using GATS[10] and its variants as test sets. We consider four frequently-used metrics: 3CosAdd (Mikolov et al., 2013a), 3CosMul (Levy and Goldberg, 2014), 3CosAvg and LRCos (Drozd et al., 2016). They return the average accuracy rates achieved when completing a given type of analogy with vector arithmetic. Among them, LRCos significantly outperforms others in almost every setting. Therefore, confined to

[5] We attempted to produce new pairs for the semantic analogy classes such as `capital-common-countries` and `family`. However, depending on massive manual annotations, the augmented sets were still not large enough.

[6] We only consider parallel types here: due to morphological divergences, some relations do not appear in both target and source languages, e.g, `adjective-to-adverb` does not exist in German although it exists in English.

[7] `http://bit.ly/2uCKmUh`

[8] `http://bit.ly/2tQKUGe` (model ID ∈ {6, 8, 10})

[9] `http://bit.ly/38S6DMN`

[10] For better compatibility, we adjusted `past-tense` in the original English GATS (see Appendix A for details).

Relation	Plural			Past-Tense			Plural-Verbs			r_s
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	
Mean LRCos	.752	.765	.815	.802	.789	.827	.917	.882	.950	
P@1 (%)	64.3	66.3	64.9	64.8	64.7	66.6	67.6	66.7	67.6	.895
P@5 (%)	80.0	80.9	80.2	80.7	80.6	81.0	81.7	81.8	82.0	.833
P@10 (%)	83.7	84.0	83.5	83.9	83.7	84.6	84.6	85.4	84.9	.706

(a) **German**

Relation	Plural			Past-Tense			Present-Participle			r_s
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	
Mean LRCos	.781	.794	.845	.776	.764	.800	.894	.830	.924	
P@1 (%)	74.1	77.0	76.3	72.5	72.5	73.7	76.0	79.8	79.5	.711
P@5 (%)	87.5	87.5	86.7	85.6	84.5	85.5	90.5	90.6	91.3	.728
P@10 (%)	89.5	89.8	89.1	88.0	87.4	87.8	92.7	92.8	93.1	.733

(b) **Italian**

Relation	Plural			Adjective-to-Adverb			Present-Participle			r_s
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	
Mean LRCos	.849	.863	.919	.573	.530	.625	.807	.750	.835	
P@1 (%)	92.8	89.5	93.9	83.7	84.7	86.0	88.6	96.3	96.0	.617
P@5 (%)	97.9	96.4	98.4	91.5	91.9	92.7	97.4	98.8	98.1	.617
P@10 (%)	98.5	97.0	98.6	93.3	93.4	94.2	98.3	99.0	98.4	.633

(c) **French**

Relation	Plural			Adjective-to-Adverb			r_s
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	
Mean LRCos	.756	.769	.819	.440	.407	.480	
P@1 (%)	89.0	85.8	90.7	83.1	83.2	86.7	.771
P@5 (%)	95.2	94.0	96.0	91.1	90.9	92.6	.943
P@10 (%)	96.7	95.5	96.8	92.6	92.5	94.2	.943

(d) **Polish**

Table 1: Evaluation results of analogy and mapping tests. Each target language corresponds to a subtable, while English serves as the common source language. *W2V*, *GV* and *FT* respectively denote English *Word2Vec*, *GloVe* and *FastText* models in § 4.1.2. r_s indicates Spearman's Correlation Coefficients between mean LRCos and BLI precisions (P@1, 5 and 10).

the length, we only focus on LRCos. For the detailed outputs of all four benchmarks, please see Appendix B.

Next, since the condition is centered around the parallelogram structures in both source and target spaces, for each pair of embeddings we calculated the geometric mean of LRCos on the same analogical relation, namely *mean LRCos*: the higher mean LRCos is, the better the parallel analogical word set satisfies the condition, and vice versa.

4.2.2 Testing cross-lingual linearity

Let A and B be two aligned matrices containing vectors of parallel words. Smith et al. (2017) justified that under the pre-condition of linearity, the cross-lingual embedding mapping from A to B should be orthogonal. One significant advantage of explicitly applying the orthogonal constraint is that optimising the transformation matrix M simplifies to an Orthogonal Procrustes problem such that

$$M^\star = \text{argmin}_M \|MA - B\|_F. \quad (12)$$

Previous works (Xing et al., 2015; Smith et al., 2017; Lample et al., 2018a) exploited the fact that this has a closed-form solution based on the singular value decomposition (SVD) of BA^T :

$$M^\star = UV^T, \text{with } U\Sigma V^T = \text{SVD}(BA^T). \quad (13)$$

Therefore, for simplicity and stability, we reflect the linearity of the cross-lingual embedding mappings using the BLI precisions obtained through orthogonal functions: if high-quality word translations can be retrieved, then the ground-truth mappings between embeddings should be almost linear, and vice versa.

One parallel word set was leveraged as both the training and test dictionary in each run. For evaluation results, we report precision at k, i.e. given source words, how often the top k retrieved target words via cosine similarity (cos_sim) include the correct translation ($k \in \{1, 5, 10\}$).

5 Results and analysis

5.1 Does our theory hold?

Table 1 reports the analogy-completion benchmarks and BLI precisions via parallel analogical word sets for all embedding pairs. As mentioned

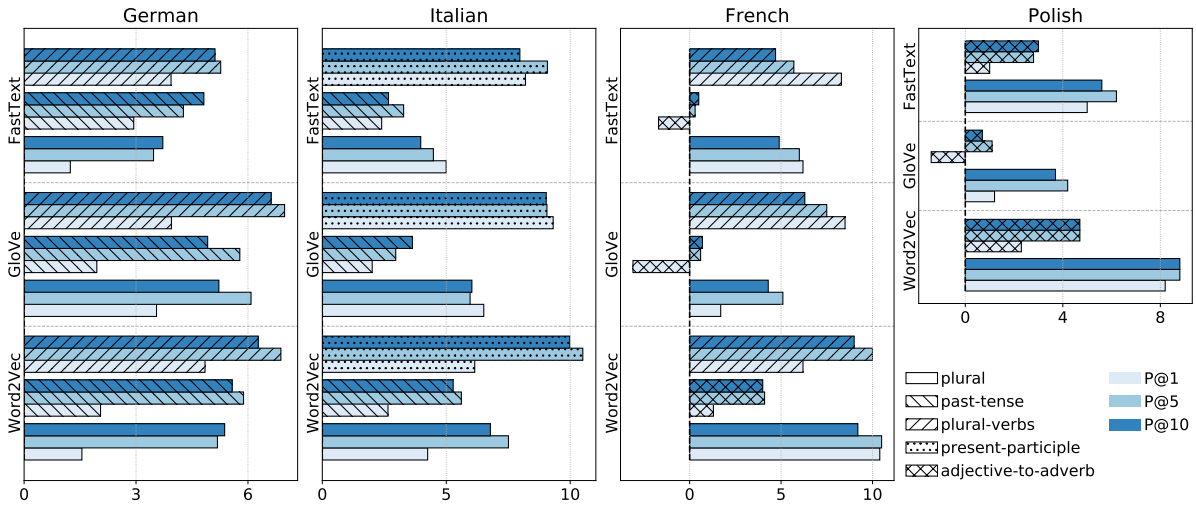


Figure 3: Percentage differences between the BLI precisions via analogical and random groups.

in § 4.1.1, the language pairs have different sizes and word frequencies, so their results are grouped in to separate subtables.

As revealed by the Spearman's Correlation Coefficient in the gray cells (r_s), there is a strong correlation (ranging from 0.617 to 0.943) between our condition and cross-lingual linearity, i.e., the better embeddings can model the analogical relations as parallelogram structures, the more likely they have a linear mapping in between, thereby supporting our theory.

In addition, we can see that *FastText* outperforms the other two models in overall analogy completing accuracy and this is consistent across the subtables. It is reasonable because with sub-words considered, *FastText* can capture more morphological information. On this account, as our condition is better satisfied (mean LRCos is higher), the inducted dictionaries are more precise with *FastText* embeddings.

Therefore, if all analogical information is well learned by word vectors, one can expect the cross-lingual mappings of embeddings to be linear. However, in practice embedding models may not encode every single analogical relation correctly. Ethayarajh et al. (2019) argued that in comparison with words of higher frequencies, rare words are seldom updated during training word embeddings, thus are less likely to get assigned properly. We validate this claim by running the analogy completion benchmark on the 50 English analogical word pairs of the lowest frequencies, finding that for every type of relation the LRCos drops to 0. In this case, as the condition for linearity breaks down, most of these words fail to be aligned with their correct translations in the mapping tests.

This is in line with typical failures when adopting the linearity assumption for cross-lingual embedding mappings: the precisions of BLI for rare lexical items are noticeably lower than those for more frequent words (Ruder et al., 2019); in GAN-based pipelines, expanding training sets by including more low-frequency words may even lead to non-negligible negative effects (Lample et al., 2018a).

5.2 Can our theory improve cross-lingual embedding mappings?

5.2.1 Towards fine-grained mappings

In § 3.1 we interpret parallelogram structures of various shapes and orientations as different analogical relations. Since a word embedding space is de facto discontinuous (Linzen, 2016), across these structures the topology is not consistent. Hence, given real-word vocabularies contain a wide range of analogies, a single linear transformation may not be sufficient enough to map all the diverse structures. This *internal structural inconsistency* serves as another major factor that leads to the unsuccessfully linear mappings. Therefore, we argue that introducing analogy-inspired structural information to learn fine-grained mappings is a more promising framework. To testify its potential, we compare the BLI precisions obtained by linear mappings between embeddings of parallel analogical word sets (analogical groups) and mappings trained in the random settings (random groups)

Figure 3 shows that as expected, precisions via analogical groups remarkably outperform those via the random counterparts in most experiments. In

some specific settings, the differences are greater than 10%. However, in some mappings between `adjective-to-adverb` word sets, the differences are negative, which seems rather wired.

Inspired by Søgaard et al. (2018) and Yang et al. (2019) who report many morphologically related failure cases in BLI, we checked the inducted lexicons and identify the reason: to guarantee rigour when qualifying linearity of analogical groups, we have restricted each pair in the parallel word sets to share the same form and tense, but this constraint was not applied when assessing random groups. As a result, the evaluation settings of analogical groups is naturally more challenging than those of the random groups. One specific example is the English adjective "*passionate*". When its *FastText* vector is projected to French embedding space, the vector of adverb "*passionnément*" ($cos\_sim = .588$) is closer than that of adjective "*passionné*" ($cos\_sim = .517$). In our morphologically stricter dictionaries to evaluate analogical groups, "*passionate-passionnément*" is a wrong pairing, but it can make a correct answer when evaluating random groups.

To check the veracity of our speculation, we relaxed the morphological restrictions by implementing the original MUSE dictionaries, then re-conducted the BLI experiments using analogical groups. This time, all analogical groups obtain higher precisions than random ones (see Appendix C). Especially, as shown in Table 2, even those which have low scores previously can overtake their random counterparts.

		Before			After		
P@		1	5	10	1	5	10
FR	*FT*	-1.7	0.3	0.5	0.4	1.4	1.5
	GV	-3.1	0.6	0.7	0.2	1.2	1.4
PL	*GV*	-1.4	1.1	0.7	0.3	1.9	1.6

Table 2: Percentage differences between the BLI precisions via three analogical groups (`adjective-to-adverb`) and corresponding random groups, before and after relaxing test dictionaries. FR and PL respectively refer to French and Polish as target languages.

Together, we observe that fine-grained analogy-inspired mapping can effectively alleviate the inconsistency among parallelogram structures.

5.2.2 Beyond neighbourhood sensitive mappings

An interesting phenomenon in previous work is mappings between geometrically-local vectors are more likely to have strong linearity (Nakashole and Flauger, 2018). Inspired by this observation, prior to us, Nakashole (2018) also investigated fine-grained cross-lingual embedding mappings. Unlike this paper, her idea is learning to locally transform word vectors from the same embedding neighbourhood, namely *neighbourhood sensitive mappings*. Although this algorithm does bring performance gain, it has not been theoretically explained by the author; in addition, it requires carefully fine-tuning the number of neighbourhoods as a core hyper-parameter.

We justify that in fact, the mechanism of neighbourhood sensitive mappings can be reasoned as a special case of our theory: due to the internal structural inconsistency analysed in § 5.2.1, the spaces surrounding gathering vectors tend to have more uniform structures. Consequently, these vectors by nature better satisfy the condition in § 3, so smaller-scale linear mappings are inclined to be more precise.

With the above insights, we then go beyond this algorithm for analogy sensitive mappings. Ethayarajh et al. (2019) justified that by leveraging the Co-occurrence Shifted PMI matrix of the training corpus, one can group words that belong to the same analogical relation, i.e., their vectors are apt to define a common parallelogram structure. This allows us to automatically reduce the internal structural inconsistency by partitioning a global embedding space into smaller regions, even with no supervision. We leave the exploration of this research direction as our future work.

6 Conclusions

This paper explains the condition for the linearity of cross-lingual embedding mappings. We rigorously prove that linearity holds *iff* analogical relations learned by multilingual embeddings are consistent. Our experiments, involving five languages and various representative embedding models, firmly support the proposed theory. Empirical results also reveal that as the condition for linearity is only satisfied locally rather than globally, there is potential for the development of analogy sensitive mappings which integrate analogical information.

In the future, we want to examine our theory using more analogical relations. Furthermore, we will follow the proposed research direction to improve the performance of cross-lingual embedding mappings.

References

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, pages 223–231, Long Beach, California, USA. PMLR.

Duygu Altinok. 2018. Demorphy, german language morphological analyzer. *CoRR*, abs/1803.00902.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Xuefeng Bai, Hailong Cao, Kehai Chen, and Tiejun Zhao. 2019. A bilingual adversarial autoencoder for unsupervised bilingual lexicon induction. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 27(10):1639–1648.

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *6th Italian Information Retrieval Workshop*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Tomaž Erjavec, Ştefan Bruda, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Peter Holozan, Nancy Ide, Heiki-Jaan Kaalep, Natalia Kotsyba, Csaba Oravecz, Vladimír Petkevič, Greg Priest-Dorman, Igor Shevchenko, Kiril Simov, Lydia Sinapova, Han Steenwijk, Laszlo Tihanyi, Dan Tufiş, and Jean Véronis. 2010. MULTEXT-east free lexicons 4.0. Slovenian language resource repository CLARIN.SI.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209, Baltimore, Maryland. Association for Computational Linguistics.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. 2017. Skip-gram - zipf + uniform = vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5935–5942, Hong Kong, China. Association for Computational Linguistics.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 40–45, London, UK. Association for Computational Linguistics.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872, Florence, Italy. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, Denver, Colorado. Association for Computational Linguistics.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. 2019. Aligning vector-spaces with noisy supervised lexicon. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 460–465, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium. Association for Computational Linguistics.

Ndapa Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 221–227, Melbourne, Australia. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted

softmax. In *International Conference on Learning Representations*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4398–4409, Hong Kong, China. Association for Computational Linguistics.

Marcin Woliński, Marcin Miłkowski, Maciej Ogrodniczuk, and Adam Przepiórkowski. 2012. PoliMorf: a (not so) new open morphological dictionary for polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 860–864, Istanbul, Turkey. European Language Resources Association (ELRA).

Chao Xing, 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*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Pengcheng Yang, Fuli Luo, Peng Chen, Tianyu Liu, and Xu Sun. 2019. MAAM: A morphology-aware alignment model for unsupervised bilingual lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3190–3196, Florence, Italy. Association for Computational Linguistics.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. In *Corpus Linguistics*.

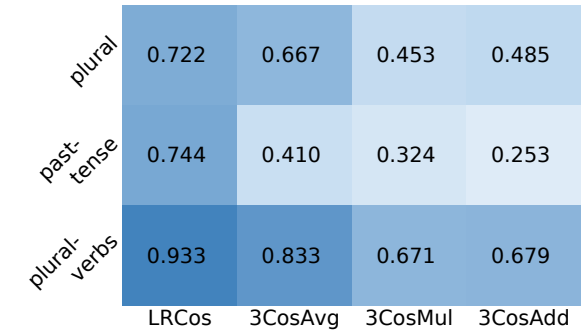
Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3180–3189, Florence, Italy. Association for Computational Linguistics.

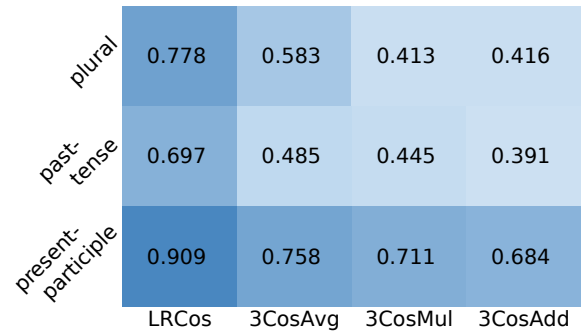
A Adjusting English GATS

Unlike in the multilingual variants, in the original copy of English GATS `past-tense` is remarkably disparate, which is quite puzzling: it is formed by word pairs like "*dancing-danced*" rather than "*dance-danced*". To solve the incompatibility caused, in English `past-tense` we replaced all present-participle verbs with their stems.

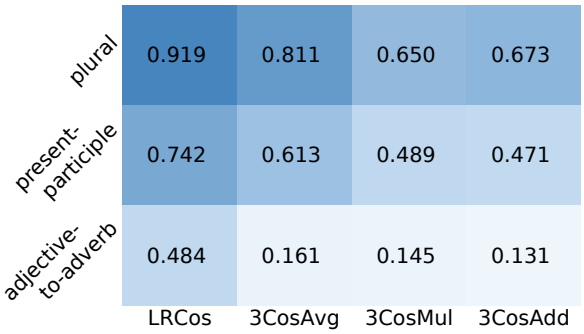
B Comprehensive outputs in analogy completion tests



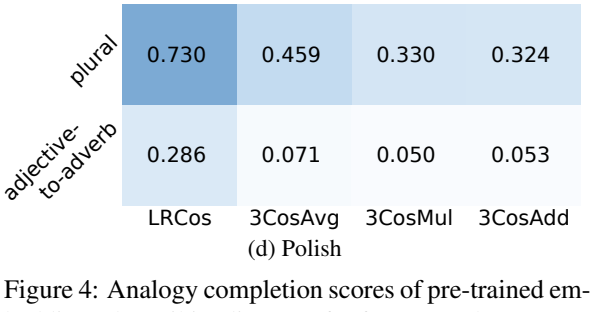
(a) German



(b) Italian

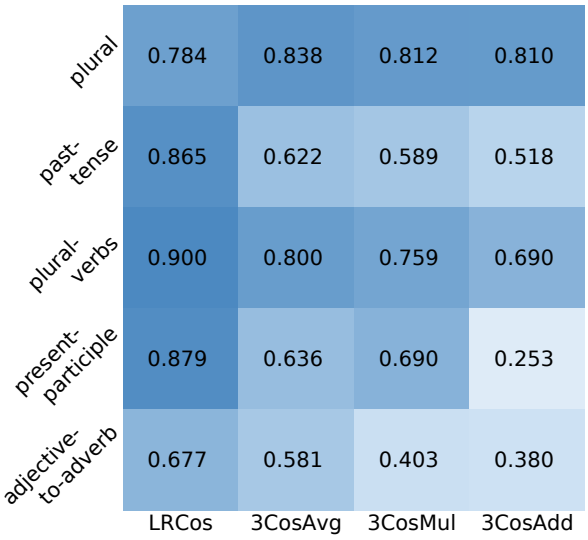


(c) French

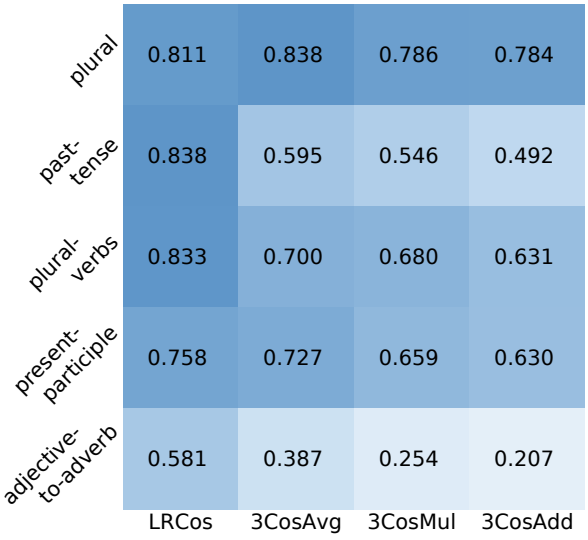


(d) Polish

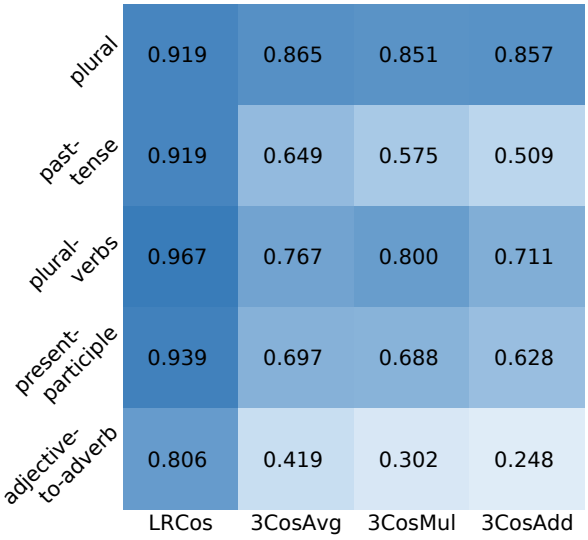
Figure 4: Analogy completion scores of pre-trained embeddings (by Wikipedia2Vec) for four target languages.



(a) Word2Vec



(b) GloVe



(c) FastText

Figure 5: Analogy completion scores of pre-trained embeddings (by three different models) for English.

C BLI evaluations without morphological restrictions

	Analogical Groups									Random Groups		
	Plural			Past-Tense			Plural-Verbs					
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*
P@1 (%)	71.2	72.8	71.5	69.6	72.5	72.3	75.2	72.5	75.5	62.8	62.8	63.7
P@5 (%)	83.7	84.8	83.9	84.0	84.6	85.3	88.3	86.7	87.9	74.8	74.8	76.3
P@10 (%)	86.7	86.5	86.5	87.4	88.4	88.2	90.7	89.9	91.0	78.3	78.8	79.8

(a) **German**

	Analogical Groups									Random Groups		
	Plural			Past-Tense			Present-Participle					
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*
P@1 (%)	77.5	79.4	78.9	76.5	75.2	78.0	79.8	81.9	83.4	69.8	70.5	71.3
P@5 (%)	89.2	89.0	88.3	88.8	87.2	89.1	92.5	92.0	93.0	80.0	81.5	82.2
P@10 (%)	91.1	91.1	90.5	90.9	89.9	90.9	94.5	94.0	94.7	82.2	83.8	85.1

(b) **Italian**

	Analogical Groups									Random Groups		
	Plural			Adjective-to-Adverb			Present-Participle					
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*
P@1 (%)	94.7	92.0	95.7	85.4	88.0	88.1	90.4	96.5	96.4	82.4	87.8	87.7
P@5 (%)	98.8	97.3	99.0	92.9	92.5	93.8	97.9	98.8	98.4	87.4	91.3	92.4
P@10 (%)	99.2	97.8	99.3	94.4	94.1	95.2	98.6	99.0	98.6	89.3	92.7	93.7

(c) **French**

	Analogical Groups						Random Groups		
	Plural			Adjective-to-Adverb					
Source Model	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*	*W2V*	*GV*	*FT*
P@1 (%)	91.7	88.9	93.1	84.3	84.9	87.7	80.8	84.6	85.7
P@5 (%)	97.0	95.9	97.3	91.9	91.7	93.2	86.4	89.9	89.8
P@10 (%)	98.1	97.2	98.1	93.1	93.4	94.7	87.9	91.8	91.2

(d) **Polish**

Table 3: BLI precisions via random and analogical groups. All mappings are evaluated using the original MUSE dictionaries (without morphological restrictions).

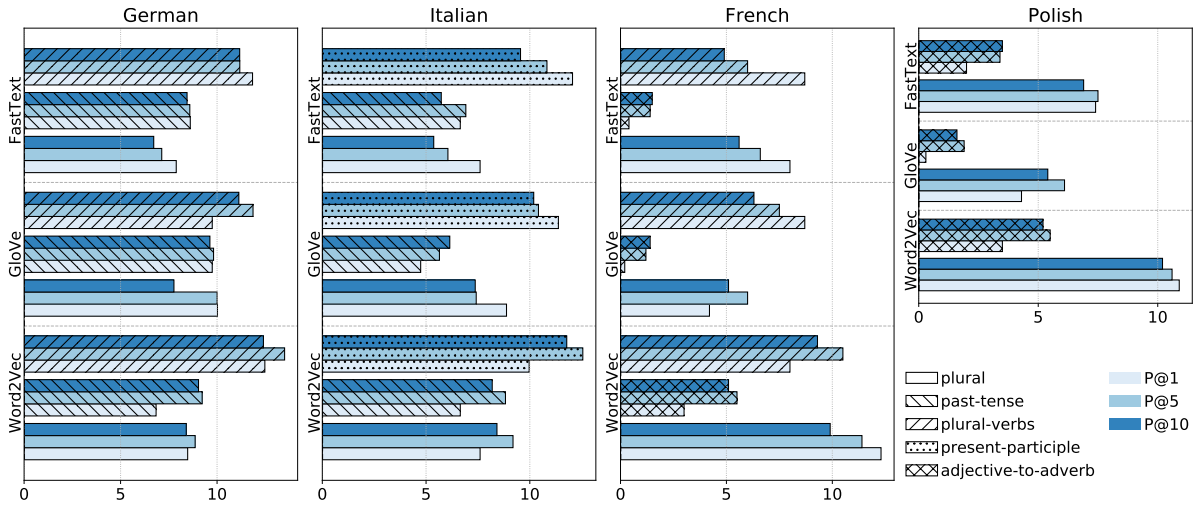


Figure 6: Percentage differences between the BLI precisions via analogical and random groups, based on the data in Table 3. Note that all values in this figure are positive and are no less than their counterparts in Figure 3.