Expert Systems With Applications 135 (2019) 287–295

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Cross-lingual word analogies using linear transformations between semantic spaces

Tomáš Brychcín[a,*], Stephen Taylor[b], Lukáš Svoboda[b]

[a] NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic
[b] Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

ARTICLE INFO

Article history:
Received 26 November 2018
Revised 22 March 2019
Accepted 9 June 2019
Available online 11 June 2019

Keywords:
Word analogies
Semantic spaces
Linear transformation
Word embeddings
Cross-lingual semantic spaces

ABSTRACT

The ability to represent the meaning of words is one of the core parts of natural language understanding (NLU), with applications ranging across machine translation, summarization, question answering, information retrieval, etc. The need for reasoning in multilingual contexts and transferring knowledge in cross-lingual systems has given rise to cross-lingual semantic spaces, which learn representations of words across different languages.

With growing attention to cross-lingual representations, it has became crucial to investigate proper evaluation schemes. The word-analogy-based evaluation has been one of the most common tools to evaluate linguistic relationships (such as male-female relationships or verb tenses) encoded in monolingual meaning representations. In this paper, we go beyond monolingual representations and generalize the word analogy task across languages to provide a new intrinsic evaluation tool for cross-lingual semantic spaces. Our approach allows examining cross-lingual projections and their impact on different aspects of meaning. It helps to discover potential weaknesses or advantages of cross-lingual methods before they are incorporated into different intelligent systems.

We experiment with six languages within different language families, including English, German, Spanish, Italian, Czech, and Croatian. State-of-the-art monolingual semantic spaces are transformed into a shared space using dictionaries of word translations. We compare several linear transformations and rank them for experiments with monolingual (no transformation), bilingual (one semantic space is transformed to another), and multilingual (all semantic spaces are transformed onto English space) versions of semantic spaces. We show that tested linear transformations preserve relationships between words (word analogies) and lead to impressive results. We achieve average accuracy of 51.1%, 43.1%, and 38.2% for monolingual, bilingual, and multilingual semantic spaces, respectively.

1. Introduction

Word distributional-meaning representations have been the key in recent success in various natural language processing (NLP) tasks. The fundamental assumption (*Distributional Hypothesis*) is that two words are expected to be semantically similar if they occur in similar contexts (they are similarly distributed across the text). This hypothesis was formulated by Harris (1954) several decades ago. Today it is the basis of state-of-the-art distributional semantic models (Bojanowski, Grave, Joulin, & Mikolov, 2017; Mikolov, Chen, Corrado, & Dean, 2013a; Pennington, Socher, & Manning, 2014; Salle, Villavicencio, & Idiart, 2016). These models learn similar semantic vectors for similar words during training. In addition, the vectors capture rich linguistic relationships such as male-female relationships or verb tenses. Such vectors can significantly improve generalization when used as features in various systems, e.g., named entity recognition (Konkol, Brychcín, & Konopík, 2015), sentiment analysis (Hercig, Brychcín, Svoboda, Konkol, & Steinberger, 2016), dialogue act recognition (Brychcín & Král, 2017), etc.

The plain-text corpora are easily available in many languages, yet the manually labeled data (e.g., text annotated with named entities, syntactic dependency trees, etc.) is expensive and mostly available for mainstream languages such as English. Pan and Yang (2010) summarized the transfer learning techniques that can learn to map (to some degree) hand-crafted features from one domain to another. In general, it is difficult to design good features which generalize well across tasks and even more difficult across different languages. These issues have attracted many researches

* Corresponding author.
E-mail addresses: brychcin@kiv.zcu.cz (T. Brychcín), stepheneugenetaylor@gmail.com (S. Taylor), svobikl@kiv.zcu.cz (L. Svoboda).

https://doi.org/10.1016/j.eswa.2019.06.021
0957-4174/© 2019 Elsevier Ltd. All rights reserved.

to move beyond monolingual meaning representations and have given rise to cross-lingual semantic spaces, which learn representations of words across different languages. There are two main implications of this research: a) cross-lingual semantic representation enables reasoning about word meaning in multilingual contexts, which is useful in many applications (cross-lingual information retrieval, machine translation, etc.) and b) it enables transferring of knowledge between languages, especially from resource-rich to poorly-resourced languages.

Several approaches for inducing cross-lingual semantic representation (i.e., unified semantic space for different languages) have been proposed in recent years, each requiring a different form of cross-lingual supervision (Upadhyay, Faruqui, Dyer, & Roth, 2016). They can be roughly divided into three categories according to the level of required alignment: a) document-level alignments (Vulić & Moens, 2016), b) sentence-level alignments (Levy, Søgaard, & Goldberg, 2017), and c) word-level alignments (Mikolov, Le, & Sutskever, 2013b).

We focus on the last case, where a common approach is to train monolingual semantic spaces independently of each other and then to use bilingual dictionaries to transform semantic spaces into a unified space. Most related works rely on linear transformations (Artetxe, Labaka, & Agirre, 2016; Faruqui & Dyer, 2014; Mikolov et al., 2013b) and profit from weak supervision. Vulić and Korhonen (2016) show that bilingual dictionaries with few thousand word pairs are sufficient. Such dictionaries can be easily obtained for most languages. Moreover, the mapping between semantic spaces can be easily extended to a multilingual scenario (more than two languages) (Ammar et al., 2016).

With growing attention to cross-lingual representations, it has became crucial to investigate proper evaluation schemes. Many metrics have already been proposed and they can be roughly divided into *intrinsic* and *extrinsic* evaluation metrics (Schnabel, Labutov, Mimno, & Joachims, 2015). In extrinsic evaluation, word representations are used as input features for a downstream task and we assess the changes in final performance. Cross-lingual applications include, e.g., sentiment analysis (Mogadala & Rettinger, 2016), document classification (Klementiev, Titov, & Bhattarai, 2012), or syntactic dependency parsing (Guo, Che, Yarowsky, Wang, & Liu, 2015). In contrast, intrinsic evaluation provides insights into the quality of representations before they are used in downstream applications. It directly tests syntactic or semantic relationships between words usually by comparison with human similarity judgments (Camacho-Collados, Pilehvar, Collier, & Navigli, 2017; Camacho-Collados, Pilehvar, & Navigli, 2015).

Although neither of these metrics is perfect, there is considerable interest in evaluating semantic spaces without needing to embed them in a NLP system. Many researchers have argued that analogy is the core of cognition and have tried to address different aspects of meaning by solving word analogy problems (Jurgens, Mohammad, Turney, & Holyoak, 2012; Turney, 2008; Turney, Littman, & Shnayder, 2003). The intrinsic evaluation introduced by Mikolov et al. (2013a), which has been gaining popularity in recent years, tries to address different aspects of meaning by solving word analogy problems. For example, the analogy "*king* is to *queen* as *man* is to *woman*", estimated by the vector equation *king* − *queen* ≈ *man* − *woman*, suggests that word vectors encode information about gender. By designing appropriate analogy questions, we can implicitly test different semantic and syntactic properties of semantic spaces.

Several authors mentioned weaknesses of word-analogy evaluation. Linzen (2016) showed that in some cases the solution is simply a nearest neighbor to the third word in the analogy question. Drozd, Gladkova, and Matsuoka (2016) studied retrieval methods beyond vector differences to solve analogy questions and mentioned inconsistency in results. Despite these weaknesses, word analogies are still one of the most commonly used intrinsic evaluation schemes.

We are particularly concerned with intrinsic evaluations in the cross-lingual environment. Combining distributional information about words in different languages into a unified semantic space (either by mapping or by joint learning) can lose some language-specific properties. On the other hand, Faruqui and Dyer (2014) showed that canonical correlation analysis can even improve the monolingual performance on word similarity tasks by learning from multilingual contexts. Artetxe et al. (2016) have explored how cross-lingual transformations affect the performance of monolingual analogies and have shown that monolingual analogy performances need not suffer from transforming semantic spaces.

In this paper, we propose to evaluate unified semantic spaces using cross-lingual word analogies. For example, the king-queen analogy can be extended by translating the second word pair into Spanish, giving us the vector equation *king* − *queen* ≈ *hombre* − *mujer*. The analogy remains the same, but now it tests the ability to generalize these semantic relationships across both languages. Similarly, the analogy "*walk* is to *walked* as *schwimmen* (German equivalent for *swim*) is to *schwamm* (German equivalent for *swam*)" testifies that cross-lingual word representations encode information about past tense for verbs.

To the best of our knowledge, we are the first to apply this technique of mixed language analogies. In spite of the weaknesses mentioned above, we believe it will be a valuable tool for assessing cross-lingual semantic spaces. We see three main contributions of our work:

- We generalize the word analogy task across languages to provide a new intrinsic evaluation tool for cross-lingual semantic spaces. Our approach allows examining cross-lingual projections and their impact on different aspects of meaning. It helps to discover potential weaknesses or advantages of cross-lingual methods before they are incorporated into different intelligent systems.
- We provide thorough comparison of three different linear transformations on six languages within different language families, including English, German, Spanish, Italian, Czech, and Croatian. We experiment with monolingual (no transformation), bilingual (one semantic space is transformed to another), and multilingual (all semantic spaces are transformed onto English space) versions of semantic spaces. We present very promising results using transformations between any pair of six languages (43.1% accuracy on average). Moreover, the multilingual settings (i.e., all languages are mapped onto English creating unified space for six languages) lead to only small degradation in performance compared to the bilingual case (38.2% accuracy on average).
- We extend available word-analogy corpora for English, German, Spanish, Italian, Czech, and Croatian and select only those analogy types (including both syntactic and semantic questions), which are useful among all these languages. We provide the cross-lingual word analogy corpus publicly available at https://github.com/brychcin/cross-ling-analogies.

This paper is organized as follows. The process of learning cross-lingual word representations via linear transformations is explained in Section 2. We define the cross-lingual word analogy task and introduce the corpus for it in Section 3. The experimental results on six languages are presented and discussed in Section 4. We conclude in Section 5 and offer some directions for future work.

2. Linear transformations between semantic spaces

Given a set of languages $\boldsymbol{L}$, let word $w^a \in \boldsymbol{V}^a$ denote the word in language $a \in \boldsymbol{L}$, where $\boldsymbol{V}^a$ is a vocabulary of that language. Let $S^a : \boldsymbol{V}^a \mapsto \mathbb{R}^d$ be a semantic space for language a, i.e., a function which projects the words w^a into Euclidean space with dimension d. The meaning of the word w^a is represented as a real-valued vector $S^a(w^a)$. We assume the same dimension d for all languages[1]

This paper focuses on linear transformations between semantic spaces. A linear transformation (also called a *linear map*) can be expressed as

$$S^{a\to b}(w^a) = S^a(w^a)\mathbf{T}^{a\to b}, \tag{1}$$

i.e., as a multiplication by a matrix $\mathbf{T}^{a\to b} \in \mathbb{R}^{d\times d}$.

Linear transformation can be used to perform *affine transformations* (e.g., rotation, reflection, translation, scaling, etc.) and other transformations (e.g., column permutation) (Nomizu & Sasaki, 1994)[2]. Composition of such operations is a matrix multiplication, which leads again to a matrix in $\mathbb{R}^{d\times d}$.

For estimating the transformation matrix $\mathbf{T}^{a\to b}$, we use a bilingual dictionary (set of n word pairs) $(w^a, w^b) \in \boldsymbol{D}^{a\to b}$, where $\boldsymbol{D}^{a\to b} \subset \boldsymbol{V}^a \times \boldsymbol{V}^b$ and $|\boldsymbol{D}^{a\to b}| = n$. In our case, we translated the original word forms w^a in language a into language b via *Google translate* (see Section 4). Finally, we use these n aligned word pairs (w^a, w^b) with their corresponding semantic vectors $(S^a(w^a), S^b(w^b))$ to form matrices $\mathbf{X}^a \in \mathbb{R}^{n\times d}$ and $\mathbf{X}^b \in \mathbb{R}^{n\times d}$.

In the following subsections, we discuss three approaches for estimating $\mathbf{T}^{a\to b}$. The optimal transformation matrix with respect to the corresponding criteria is denoted as $\hat{\mathbf{T}}^{a\to b}$.

2.1. Least squares transformation

Following Mikolov et al. (2013b), we can estimate the matrix $\mathbf{T}^{a\to b}$ by minimizing the sum of squared residuals. The optimization problem is given by

$$\hat{\mathbf{T}}^{a\to b} = \underset{\mathbf{T}^{a\to b}}{\operatorname{argmin}} \left\| \mathbf{X}^b - \mathbf{X}^a\mathbf{T}^{a\to b} \right\|_2^2 \tag{2}$$

and can be solved for example by the gradient descent algorithm.

The least squares method also has an analytical solution. By taking the Moore-Penrose pseudo-inverse of $\mathbf{X}^a$, which can be computed using singular value decomposition (SVD) (Campbell & Meyer, 2009), we achieve

$$\hat{\mathbf{T}}^{a\to b} = (\mathbf{X}^{a\top}\mathbf{X}^a)^{-1}\mathbf{X}^{a\top}\mathbf{X}^b. \tag{3}$$

Lazaridou, Dinu, and Baroni (2015) showed that the least squares mapping leads to increasing the *hubness* in the final space, because the set of vectors in $\mathbf{X}^a\hat{\mathbf{T}}^{a\to b}$ has lower variance than in $\mathbf{X}^b$ (points are on average closer to each other).

2.2. Orthogonal transformation

Motivated by inconsistency among the objective functions for learning word representations (based on dot products), the least squares mapping (minimizing Euclidean distances), and word similarity evaluation (based on cosine similarities), Xing, Wang, Liu, and Lin (2015) argued that the transformation matrix in the least squares objective should be orthogonal. For estimating this matrix, they introduced an approximate algorithm composed of gradient descent updates and repeated applications of the SVD. Artetxe et al. (2016) then derived the analytical solution for the orthogonality constraint and showed that this transformation preserves the monolingual performance of the source space.

Orthogonal transformation is the least squares transformation subject to the constraint that the matrix $\mathbf{T}^{a\to b}$ is orthogonal.[3] The optimal transformation matrix is given by

$$\hat{\mathbf{T}}^{a\to b} = \mathbf{V}\mathbf{U}^\top, \tag{4}$$

where matrices $\mathbf{V}$ and $\mathbf{U}$ are obtained using SVD of $\mathbf{X}^{b\top}\mathbf{X}^a$ (i.e., $\mathbf{X}^{b\top}\mathbf{X}^a = \mathbf{U}\Sigma\mathbf{V}^\top$).

2.3. Canonical correlation analysis

Canonical correlation analysis is a way of measuring the linear relationship between two multivariate variables (i.e., vectors). It finds basis vectors for each variable in the pair such that the correlation between the projections of the variables onto these basis vectors is mutually maximized.

Given the sample data $\mathbf{X}^a$ and $\mathbf{X}^b$, at the first step we look for a pair of projection vectors $(\mathbf{c}_1^a \in \mathbb{R}^d, \mathbf{c}_1^b \in \mathbb{R}^d)$ (also called *canonical directions*), whose data projections $(\mathbf{X}^a\mathbf{c}_1^a, \mathbf{X}^b\mathbf{c}_1^b)$ yield the largest Pearson correlation. Once we have the best pair, we ask for the second-best pair. On either side of a and b, we look for $\mathbf{c}_2^a$ and $\mathbf{c}_2^b$ in the subspaces orthogonal to the first canonical directions $\mathbf{c}_1^a$ and $\mathbf{c}_1^b$, respectively, maximizing correlation of data projections. Generally, k-th canonical directions are given by

$$(\mathbf{c}_k^a, \mathbf{c}_k^b) = \underset{\mathbf{c}^a,\mathbf{c}^b}{\operatorname{argmax}}\, \operatorname{cor}(\mathbf{X}^a\mathbf{c}^a, \mathbf{X}^b\mathbf{c}^b), \tag{5}$$

where for each $1 \le i < k$, $(\mathbf{X}^a\mathbf{c}^a) \cdot (\mathbf{X}^a\mathbf{c}_i^a) = 0$ and $(\mathbf{X}^b\mathbf{c}^b) \cdot (\mathbf{X}^b\mathbf{c}_i^b) = 0$. In the end of this process, we have bases of d canonical directions for both sides a and b. We can represent them as a pair of matrices $\mathbf{C}^a \in \mathbb{R}^{d\times d}$ and $\mathbf{C}^b \in \mathbb{R}^{d\times d}$ (each column corresponds to one canonical direction $\mathbf{c}_k^a$ or $\mathbf{c}_k^b$, respectively), which project $\mathbf{X}^a$ and $\mathbf{X}^b$ into a shared space. The exact algorithm for finding these bases is described in (Hardoon, Szedmak, & Shawe-Taylor, 2004).

Faruqui and Dyer (2014) used the canonical correlation analysis for incorporating multilingual contexts into word representations, outperforming the standalone monolingual representations on several intrinsic evaluation metrics. Ammar et al. (2016) extended this work and create a multilingual semantic space for more than fifty languages. Following their approach, the final linear transformation is given by

$$\hat{\mathbf{T}}^{a\to b} = \mathbf{C}^a\mathbf{C}^{b^{-1}}. \tag{6}$$

3. Cross-lingual word analogies

3.1. Definition

The word analogy task consists of questions of the form: word w_1 is to w_2 as word w_3 is to w_4, where the goal is to predict w_4. Basically, the question consists of two pairs of words assuming there is the same relationship in both pairs (e.g., "*Rome* is to *Italy* in the same sense as *Tokyo* is to *Japan*").

The task was originally designed to investigate linear dependencies between words in vector space so that these questions can be answered by simple algebraic operations on corresponding word

[1] Note that all described linear transformations can be easily extended to the general case, where the dimension of two semantic spaces differs.

[2] In the general case, affine transformation is the composition of two functions (a translation and a linear map) represented as $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$. Using so called *augmented matrix* (which extends the dimension by 1), we can rewrite this to $\left|\begin{matrix}\mathbf{y}\\ 1\end{matrix}\right| = \left|\begin{matrix}\mathbf{A} & \mathbf{b}\\ 0\ldots0 & 1\end{matrix}\right| \left|\begin{matrix}\mathbf{x}\\ 1\end{matrix}\right|$, i.e., we can use only matrix multiplication (linear map). In our case, we omit this trick and use only matrix $\mathbf{A}$ similarly to all other prior works on linear transformations for cross-lingual NLP. Moreover, in our experiments (Section 4), we center both source and target semantic spaces towards zero so that no translation is required.

[3] Matrix $\mathbf{A}$ is orthogonal if contains orthonormal rows and columns, i.e., $\mathbf{A}\mathbf{A}^\top = \boldsymbol{I}$. An orthogonal matrix preserves the dot product, i.e., $\mathbf{x} \cdot \mathbf{y} = (\mathbf{A}\mathbf{x}) \cdot (\mathbf{A}\mathbf{y})$, thus the monolingual invariance property.

Table 1
Number of word pairs for each language and each analogy type.

		En	De	Es	It	Cs	Hr
Semantic	Family	24	24	20	20	26	41
	State-currency	29	29	28	29	29	21
	Capital-common-countries	23	23	21	23	23	23
Syntactic	State-adjective	41	41	40	41	41	41
	Adjective-comparative	23	37	5	10	40	77
	Adjective-superlative	20	34	40	29	40	77
	Adjective-opposite	29	29	20	24	27	29
	Noun-plural	112	111	37	36	74	46
	Verb-past-tense	38	40	39	33	95	40

vectors (i.e., the relationship between two words is encoded as a difference of their vectors).

Similar questions can also be designed for cross-lingual cases, i.e., one pair of words is in language a and second is in language b, e.g., "*king* is to *queen* in the same sense as *Bruder* (German equivalent for *brother*) is to *Schwester* (German equivalent for *sister*) or *otec* (Czech equivalent for *father*) is to *matka* (Czech equivalent for *mother*)". In the ideal case, the vector differences should remain the same. In reality, there are several issues including the particular syntax of each language (see Section 4.4).

More formally, we are given a word pair (w_1^a, w_2^a) in language a and a word w_3^b in language b. To find the word w_4^b (related to w_3^b in the same way as w_2^a is related to w_1^a), we first estimate the target vector $\mathbf{v} = S^{a\to b}(w_2^a) - S^{a\to b}(w_1^a) + S^b(w_3^b)$. Then, we go through all words w^b in vocabulary $\boldsymbol{V}^b$ of language b looking for the word most similar to $\mathbf{v}$ according to cosine similarity[4]

$$\hat{w}_4^b = \underset{w^b}{\operatorname{argmax}} \frac{S^b(w^b)\cdot\mathbf{v}}{\left\|S^b(w^b)\right\|_2 \|\mathbf{v}\|_2}. \tag{7}$$

Finally, if $\hat{w}_4^b = w_4^b$, we consider the question is answered correctly. If $a = b$, this becomes the standard monolingual word analogy task as defined in Mikolov et al. (2013a).

3.2. Corpus

We combine and extend available corpora for monolingual word analogies in English (En) (Mikolov et al., 2013a), German (De) (Köper, Scheible, & Schulte im Walde, 2015), Spanish (Es) (Cardellino, 2016), Italian (It) (Berardi, Esuli, & Marcheggiani, 2015), Czech (Cs) (Svoboda & Brychcín, 2016), and Croatian (Hr) (Svoboda & Beliga, 2017). We consider only those analogy types, which exist across all six languages (three semantically oriented and six syntactically oriented analogy types). Table 1 shows the number of word pairs for each analogy type and each language. For all languages, questions composed of single words are taken into account (i.e., no phrases). In the following list we briefly introduce each analogy type and describe the changes and extensions we have made compared with the original corpora:

- *Family*: Family relations based on different gender (male vs. female), e.g., *son* vs. *daughter*.
- *State-currency*: Pairs representing a state and its currency, e.g., *USA* vs. *dollar*. Since this analogy type is not included in the original Czech corpus, we manually translated English word pairs.
- *Capital-common-countries*: Word pairs consist of capital city and the corresponding state, e.g., *Moscow* vs. *Russia*.
- *State-adjective*: Relationship representing the state used as a noun vs. adjective, e.g., *China* vs. *Chinese*. This analogy type is not included in original Czech, Croatian, and Italian corpora. We manually translated English word pairs into these three languages.
- *Adjective-comparative*: Adjectives in basic form and comparative form, e.g., *slow* vs. *slower*. We manually created this part for Spanish as it was not in the original corpus. Note there are very few Spanish and Italian comparatives expressed as a single word.
- *Adjective-superlative*: Adjectives in basic form and superlative form, e.g., *bad* vs. *worst*. Similarly to *adjective-comparative*, we manually created this part for Spanish.
- *Adjective-opposite*: Adjectives in basic form and negation, e.g., *possible* vs. *impossible*.
- *Noun-plural*: Noun in basic form (lemma) and plural form, e.g., *pig* vs. *pigs*.
- *Verb-past-tense*: Verb in infinitive and the past tense (preterite), e.g., *see* vs. *saw*.

4. Experiments

4.1. Settings

Our experiments start with building monolingual semantic spaces for each of tested languages (English, German, Spanish, Italian, Czech, and Croatian). We use character-n-gram-based skip-gram model (Bojanowski et al., 2017), which recently achieved the state-of-the-art performance in the monolingual word analogy task for several languages. For all languages except Croatian, we use word vectors pre-trained on Wikipedia[5]. Relative sizes of Wikipedia corpora are: En 13GB, De 4.3GB, Es 2.5GB, It 2.3GB, Cs 0.6GB, and Hr 0.2GB. The Wikipedia corpus for Croatian yields poor performance, so we combine it with web-crawled texts. We adopted the corpus hrWaC[6] (Šnajder, Padó, & Agić, 2013) and merged it with Croatian Wikipedia. The final Croatian corpus has approximately 1.3 billion tokens. We use settings recommended by Bojanowski et al. (2017), i.e., texts are lowercased, vector dimension is set to $d = 300$, and character n-grams from 3 to 6 characters are used.

Bilingual dictionaries $\boldsymbol{D}^{a\to b}$ between each pair of languages a and b, are created from the n most frequent words in corpus of language a and their translation into language b using Google translate.

We experiment with different global post-processing techniques for semantic spaces, which can significantly boost the final performance in word analogy task (see Section 4.3):

-c Column-wise mean centering (i.e., moving the space towards zero) is a standard step in regression analysis. Artetxe et al. (2016) showed this could lead to improving results of linear mappings.

-u Normalizing word vectors to be unit vectors guarantees that all word pairs in dictionary $\boldsymbol{D}^{a\to b}$ contribute equally to the optimization criteria of linear transformation.

-cu Column-wise mean centering followed by vector normalization.

We always apply the same post-processing for both semantic spaces S^a and S^b in a pair before the linear mapping. We distinguish between two types of cross-lingual semantic spaces:

B Bilingual semantic space is created by linear transformation of S^a onto the space S^b.

[4] In the monolingual case the input question words (i.e., w_1, w_2, and w_3) are discarded during the search as recommended by Mikolov et al. (2013a). In the cross-lingual case this does not make sense because w_1^a and w_2^a are in a different language. Thus we discard only w_3^b from the search.

[5] Semantic spaces for many languages trained on Wikipedia are available to download at https://fasttext.cc.

[6] Available at http://takelab.fer.hr/data.

Table 2
The average accuracies across all combinations of language pairs for different linear transformations and post-processing techniques. The size of bilingual dictionary was set to $n = 20,000$. *No trans.* denotes the monolingual experiments without transforming the spaces.

		-		**-c**		**-u**		**-cu**	
		Acc@1	**Acc@5**	**Acc@1**	**Acc@5**	**Acc@1**	**Acc@5**	**Acc@1**	**Acc@5**
Monoling	No trans.	49.6	63.7	50.1	64.6	50.6	64.6	**51.1**	**65.2**
	M-LS	40.2	55.3	40.3	55.6	41.3	56.5	41.3	56.6
	M-OT	49.6	63.7	50.1	64.6	50.6	64.6	**51.1**	**65.2**
	M-CCA	46.8	61.8	47.6	62.5	47.5	62.4	48.1	63.0
Cross-lingual	B-LS	33.7	51.4	34.3	52.3	33.5	51.1	34.0	52.0
	B-OT	40.1	55.9	40.6	56.6	40.7	56.5	41.2	57.3
	B-CCA	42.3	57.5	42.7	58.2	42.6	57.8	**43.1**	**58.5**
	M-LS	32.2	48.8	32.7	49.3	32.9	49.6	32.5	49.3
	M-OT	37.3	53.7	37.6	54.3	37.8	54.4	**38.2**	**55.0**
	M-CCA	35.3	52.7	36.2	53.8	35.5	52.9	36.0	53.5

M Multilingual semantic space is created by linear transformations of all S^a except English onto the English space (i.e., unified space for all six languages).

We experiment with three techniques for linear mapping (all described in Section 2), namely, least squares transformation (LS), orthogonal transformation (OT), and canonical correlation analysis (CCA). The experiment denoted as B-OT-cu means the bilingual semantic space created by orthogonal transformation with mean centering and unit vectors. M-CCA-c means the multilingual semantic space created by canonical correlation analysis only with mean centering.

4.2. Evaluation

We process the questions and calculate accuracy as defined in Section 3. During the search for an answer we always browse the 300,000 most frequent words in a corresponding language. We calculate the accuracy for each analogy type separately. In prior works on monolingual word analogies, if the question or the correct answer contains an out-of-vocabulary word, it is assumed the question is answered incorrectly. The model we use in our experiments (Bojanowski et al., 2017) is able to estimate the out-of-vocabulary word representations only from the character n-grams (without context). This allows us to process all questions in the cross-lingual analogy corpus.

For each analogy type we process all combinations of pairs between languages a and b (e.g., for the category *family* and the transformation from Czech to German, we have $26 \times 24 = 624$ questions). In the case $a = b$ (i.e., monolingual experiments), we omit the questions composed from two same pairs (e.g., for the category *family* in Italian, we have $20 \times 19 = 380$ questions). The final accuracy is an average over accuracies for individual categories. This is motivated by the fact that for each language and each analogy type, we have a different number of word pairs (see Table 1). By averaging the accuracies each analogy type contributes equally to the final score and the results are comparable across languages. In the following text, Acc@1 denotes the accuracy considering only the most similar word as a correct answer. Acc@5 assumes that the correct answer is in the list of five most similar words. All accuracies are expressed in percentages.

4.3. Global results

Table 2 shows accuracies averaged across all combinations of pairs made of six languages. The columns represent different post-processing techniques and rows different transformations. The upper part of the table shows the monolingual experiments with original spaces without transformation (*No trans.*) compared with the unified multilingual space for all six languages. The orthogonal transformation provides same results as the original semantic space. Canonical correlation analysis leads to slightly lower accuracies and least squares method is worst. The most interesting is the lower part of the table, i.e., cross-lingual experiments, showing the average accuracies over all language pairs, but where source a and target b languages differ $a \neq b$. We can see that canonical correlation analysis performs best for bilingual cases, while orthogonal transformation yields better accuracies in multilingual spaces. In all cases, the mean centering followed by vector normalization led to the best results.

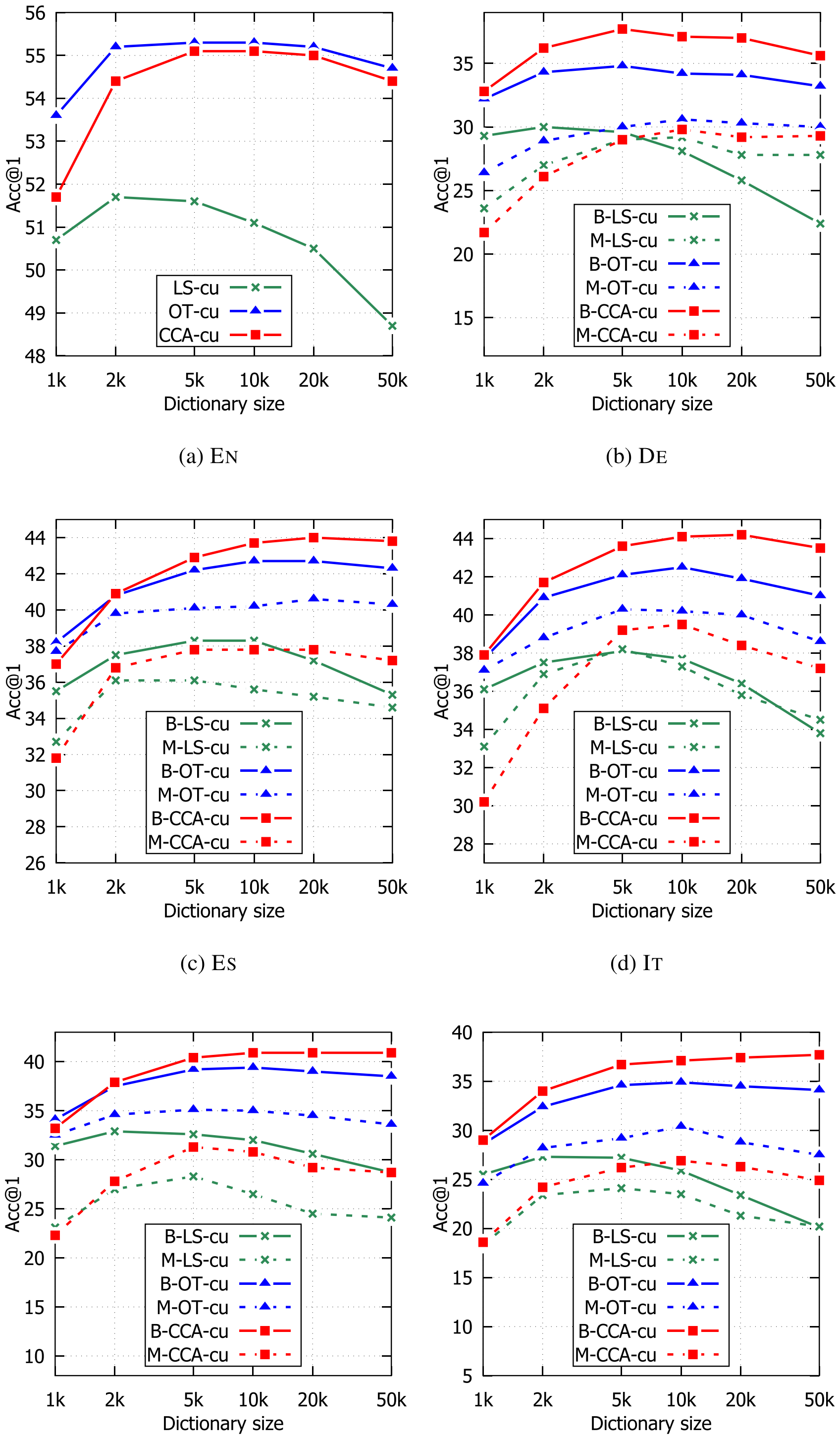
We chose the size of bilingual dictionaries to be $n = 20,000$, because this works best among all languages (see Fig.1). This figure shows the trends for bilingual spaces with varying dictionary size. Accuracies are averaged over all source languages (monolingual spaces, i.e., where $a = b$, are not taken into account). In most cases, the accuracy decreases when $n = 50,000$. We compose the bilingual dictionaries from the most frequent words. The less frequent words in dictionary may have less precise meaning representation, but all of them contribute equally to estimating the linear mapping. We believe that these less frequent words degrade the performance (i.e., more does not necessary mean better). This behavior agrees with the conclusions in (Vulić & Korhonen, 2016). Notably, we are able to achieve very promising results even with very limited dictionaries (i.e., one thousand word pairs).

4.4. Individual results

Table 3 shows accuracies for all language pairs using the best settings (CCA for bilingual cases, OT for multilingual cases, $n = 20,000$, and post-processing -cu) and for both bilingual (B) and multilingual (M) case. Rows represent the source language a and columns the target language b (i.e., given three words w_1^a, w_2^a, and w_3^b, we look for the fourth word w_4^b in column's language).

On the diagonal, we can see the monolingual results; these are the highest accuracies in each column. The highest cross-lingual accuracies are achieved by transforming onto English space (English has by far the highest monolingual accuracy), which supports our choice to use English as a intermediary for multilingual semantic spaces. We believe that English words are easier targets to hit (i.e., to find fourth word in analogy) because they are less inflected, and have fewer variations on the lemma in the same neighborhood of the semantic space. Correspondingly, the high level of inflection in Slavic languages has two consequences: the training data are diluted by the expansion of the vocabulary (both row and column effects) and the search for the final word of the analogy has more nearby alternatives (column effect).

Table 4 shows detailed results for bilingual spaces and for each individual analogy type. Again, rows represent the source



(a) En

(b) De

(c) Es

(d) It

(e) Cs

(f) Hr

Fig. 1. Ranging dictionary size for all languages individually. Accuracies represent the average over all source languages except the one onto which we are transforming. Note for English (En) both cases B and M are equal, because we transform all languages onto English to create multilingual space.

language *a* and columns the target language *b*. The results were achieved using B-CCA-cu transformation with dictionaries of size $n = 20,000$. Each language seems to have strengths and weaknesses.

Interestingly, there are analogies and languages, where bilingual pairs beat monolingual. For example in the *family* analogies (Table 4), English, Spanish, and Italian have the best monolingual results. Most languages profit from having the first two words of the analogy in these languages.

There is not much to say about Tables 4c, 4d, 4g, and 4h; all language pairs simply produce high accuracies. On the contrary, the *state-currency* results (Table 4a) are uniformly poor. One might expect that analogies using the national adjective would work better, because they form a frequent collocation (e.g., *Hun-*

Table 3
Accuracies between all pairs of languages using both bilingual spaces with CCA and multilingual semantic spaces with OT. The size of bilingual dictionaries was set to $n = 20,000$. Post-processing includes mean centering and vector normalization for all cases.

		En		De		Es		It		Cs		Hr	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
En	B-CCA	63.8	77.0	41.3	58.7	45.1	55.8	44.7	59.6	43.9	62.5	41.9	58.7
	M-OT	63.8	77.0	34.5	54.4	41.4	54.2	39.8	56.3	36.3	56.9	31.5	52.6
De	B-CCA	60.8	74.4	46.8	62.6	43.6	56.2	43.8	58.7	42.2	59.9	38.3	56.2
	M-OT	60.8	74.1	46.8	62.6	39.7	51.6	37.6	54.1	33.1	53.2	27.5	48.4
Es	B-CCA	49.2	63.1	35.9	50.0	51.3	62.5	49.7	63.4	36.9	51.9	33.6	49.3
	M-OT	49.9	63.7	29.6	46.3	51.3	62.5	46.8	62.4	32.3	49.1	26.1	44.5
It	B-CCA	50.4	65.5	35.1	50.1	49.8	61.7	52.2	65.4	39.1	54.1	34.7	49.9
	M-OT	50.8	65.9	29.1	46.3	45.9	58.9	52.2	65.4	34.0	50.6	26.8	45.0
Cs	B-CCA	58.9	73.6	36.4	54.3	40.7	54.4	43.1	58.9	50.0	66.1	38.4	55.6
	M-OT	58.0	73.3	31.1	49.9	37.6	51.9	38.5	55.9	50.0	66.1	31.9	50.3
Hr	B-CCA	55.8	72.2	36.0	54.4	40.5	54.9	39.6	56.8	42.3	58.8	42.4	57.8
	M-OT	56.4	72.3	27.2	48.4	38.3	51.8	37.2	54.6	36.7	54.0	42.4	57.8

Table 4
Accuracies (Acc@1) of bilingual semantic spaces using B-CCA-cu for individual analogies.

	En	De	Es	It	Cs	Hr
En	68.8	52.4	85.4	76.0	41.2	47.2
De	65.5	48.0	76.3	66.5	35.9	40.9
Es	70.6	49.0	86.8	74.5	43.1	45.2
It	65.4	45.6	81.8	72.9	39.2	45.2
Cs	61.5	38.6	74.0	65.0	35.6	42.0
Hr	57.4	33.3	62.8	60.0	32.6	37.1

(a) Family

	En	De	Es	It	Cs	Hr
En	11.1	7.4	3.9	4.4	2.1	5.3
De	5.8	6.7	1.5	3.2	1.5	3.4
Es	6.5	3.7	2.8	3.4	1.8	1.4
It	6.3	4.9	2.8	3.7	3.0	3.1
Cs	3.4	2.7	1.7	2.5	1.0	1.6
Hr	5.3	5.7	1.5	1.6	1.5	4.3

(b) State-currency

	En	De	Es	It	Cs	Hr
En	95.3	81.7	86.7	86.8	48.8	53.3
De	91.9	82.6	85.9	89.2	55.0	49.3
Es	93.8	82.8	83.3	84.5	54.7	47.8
It	93.6	83.2	85.7	88.9	54.3	53.1
Cs	91.1	77.9	79.5	80.0	44.9	43.9
Hr	71.1	55.0	64.4	55.6	25.9	32.2

(c) Capital-common-countries

	En	De	Es	It	Cs	Hr
En	91.2	58.7	90.2	91.8	86.3	88.0
De	91.1	75.9	86.0	92.8	73.5	80.2
Es	90.5	71.5	87.4	94.1	83.8	83.6
It	90.5	61.5	89.8	89.1	90.1	85.2
Cs	88.5	44.6	86.6	90.4	92.7	80.2
Hr	86.6	66.8	82.0	85.4	82.7	86.5

(d) State-adjective

	En	De	Es	It	Cs	Hr
En	78.5	55.1	1.7	10.0	34.5	31.6
De	68.4	59.1	2.2	7.3	17.4	16.8
Es	34.8	29.2	25.0	12.0	4.5	9.6
It	41.3	31.9	8.0	13.3	4.3	5.5
Cs	76.4	49.7	2.0	15.3	48.4	33.1
Hr	67.6	45.3	8.3	15.6	32.6	32.2

(e) Adjective-comparative

	En	De	Es	It	Cs	Hr
En	68.9	15.3	11.5	20.2	12.0	19.0
De	63.8	32.9	12.5	20.1	15.6	19.8
Es	4.6	0.4	32.8	37.3	0.0	0.2
It	5.9	0.5	24.9	62.1	0.1	0.2
Cs	54.6	21.8	4.9	24.3	28.5	17.3
Hr	57.2	21.6	7.5	10.8	21.6	29.2

(f) Adjective-superlative

	En	De	Es	It	Cs	Hr
En	51.4	39.4	40.7	38.2	79.8	49.8
De	49.5	33.5	42.9	37.9	78.7	47.4
Es	46.2	37.2	40.3	37.3	76.7	47.2
It	49.6	38.9	43.3	38.9	79.2	47.4
Cs	49.6	35.6	33.9	34.3	78.9	41.0
Hr	46.6	40.2	41.4	36.4	77.8	51.9

(g) Adjective-opposite

	En	De	Es	It	Cs	Hr
En	66.8	48.2	67.6	45.0	32.6	40.7
De	66.1	49.0	65.2	41.8	33.2	40.4
Es	68.9	48.6	71.7	55.4	32.1	45.0
It	68.6	48.4	72.5	52.6	33.3	42.8
Cs	62.2	43.8	61.7	36.2	39.4	31.3
Hr	66.8	47.6	63.2	41.8	32.9	44.2

(h) Noun-plural

	En	De	Es	It	Cs	Hr
En	42.2	13.1	18.6	29.7	58.0	41.8
De	45.1	33.4	20.3	35.1	69.0	46.8
Es	27.0	0.8	31.8	48.6	35.2	22.4
It	32.2	1.4	39.3	48.7	48.8	29.8
Cs	42.5	13.1	22.4	40.0	80.8	55.6
Hr	43.6	8.7	33.1	48.9	73.4	63.6

(i) Verb-past-tense

garian forint), but those analogies also perform poorly (for En → En we achieved 11.1%).

In Tables 4e and 4f, comparative and superlative adjectives, both Romance languages (Spanish and Italian) are the anomalies. Both languages form the comparative with an adjective clitic, and both use surrounding syntax to distinguish between comparative and superlative. This syntactic dependency is sufficient to make them outliers.

In *verb-past-tense* (Table 4i), German is an outlier. Monolingually it works fairly well, but it frequently misses with other languages. It turns out that the cosine similarity spread and variance is greater for the German vector offsets. For all languages except English and German, the infinitive form (the first element of the word pair) is distinctively marked. In English and German, it can be confused with other forms of the verb and with nouns. Perhaps, this problem is more evident for German, where the first words in pairs may be displaced depending on the relative frequencies of the other senses. So in a monolingual German analogy, the first words of the two pairs are displaced depending on the relative frequencies of the other senses. Since the German pairs are not semantically similar to other languages, it is not surprising for the bilingual analogies to fail. This effect probably also accounts for the numbers for En for this analogy.

5. Summary

5.1. Conclusion

This paper investigated cross-lingual meaning representations which serve as key features in cross-lingual systems (e.g., cross-lingual information retrieval, machine translation, etc.). The meaning representation which generalizes across different languages needs to be precisely evaluated in order to see the impact of cross-lingual projections on different aspects of meaning. With growing attention to cross-lingual systems, it has became crucial to investigate proper evaluation schemes for cross-lingual semantic spaces before they are embedded into a final system.

For that purpose, we extended the word-analogy evaluation scheme onto cross-lingual environment and prepared the corpus for it. The new cross-lingual word analogy corpus is publicly available for the research community. To the best of our knowledge, we are first to evaluate word analogies across languages.

We experimented with six languages (namely, English, German, Spanish, Italian, Czech, and Croatian) within different language families. We explored linear transformations (including least squares method, orthogonal transformation, and canonical correlation analysis) to build bilingual (two languages) and multilingual (more than two languages) semantics spaces and ranked them according to their accuracy in searching word analogies.

Canonical correlation analysis was proved to perform best for bilingual semantic spaces (top-1 accuracy 43.1% and top-5 accuracy 58.5%). On the other hand, orthogonal transformation performed better in multilingual environment where all six languages were represented in a unified space (top-1 accuracy 38.2% and top-5 accuracy 55.0%).

The most important finding is that the created cross-lingual semantic spaces preserve important linguistics relationships between words (word analogies) even across conceptually different languages, with accuracy not far bellow their monolingual counterparts (top-1 accuracy 51.1% and top-5 accuracy 65.2%). We believe our approach will help researches with building and evaluating methods for meaning representation required in various cross-lingual systems, including for example, sentiment analysers, fact-checking and fake-news detection systems, cross-lingual information retrieval systems, etc.

5.2. Future work

As a future work we see the potential mainly in the following research directions. First of all, we plan to experiment and extend the corpus with additional languages within different language families, especially non-European languages, such as Chinese, Japanese, Hindustani, Arabic, Malay, Bengali, etc. In addition, we would like to explore another analogy types, including both syntactic and semantic aspects of words or phrases.

Another interesting direction for future work is to compare different architectures for cross-lingual semantic spaces requiring different level of supervision, i.e., based on sentence or document-level alignments (Levy et al., 2017; Vulić & Moens, 2016). It would be very interesting to assess how different forms of alignment affect different aspects of meaning representation.

In this work we rely on character-n-gram-based skip-gram approach (Bojanowski et al., 2017) for building monolingual semantic spaces. Different approaches (Mikolov et al., 2013a; Pennington et al., 2014; Salle et al., 2016) can have different properties when used in cross-lingual settings. To compare and evaluate more approaches is definitely worthwhile.

Last but not least, we would like to experiment with other techniques for searching word analogies (beyond simple vector arithmetic operations) such as the one presented in (Levy & Goldberg, 2014) or (Gittens, Achlioptas, & Mahoney, 2017). Other methods could be less sensitive to inaccuracies caused by cross-lingual projections and thus can lead to better results.

Disclosure of conflict of interest

None.

Credit authorship contribution statement

Tomáš Brychcín: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing. **Stephen Taylor:** Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing. **Lukáš Svoboda:** Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing.

Acknowledgments

This work was supported from ERDF “Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)” (No.: CZ.02.1.01/0.0/0.0/17_048/0007267) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures”. Lastly, we would like to thank the anonymous reviewers for their insightful feedback.

References

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., & Smith, N. A. (2016). Massively multilingual word embeddings. *CoRR, abs/1602.01925*.

Artetxe, M., Labaka, G., & Agirre, E. (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* (pp. 2289–2294). Austin, Texas: Association for Computational Linguistics.

Berardi, G., Esuli, A., & Marcheggiani, D. (2015). Word embeddings go to Italy: A comparison of models and training datasets. In *Proceedings of the 6th italian information retrieval workshop, cagliari, italy, may 25–26, 2015*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Brychcín, T., & Král, P. (2017). Unsupervised dialogue act induction using gaussian mixtures. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (pp. 485–490). Association for Computational Linguistics.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., & Navigli, R. (2017). SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 15–26). Vancouver, Canada: Association for Computational Linguistics.

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2015). A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (pp. 1–7). Beijing, China: Association for Computational Linguistics.

Campbell, S. L., & Meyer, C. D. (2009). *Generalized inverses of linear transformations*. Society for Industrial and Applied Mathematics. doi:10.1137/1.9780898719048.

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings. http://www.crscardellino.me/SBWCE/.

Drozd, A., Gladkova, A., & Matsuoka, S. (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* (pp. 3519–3530). Osaka, Japan: The COLING 2016 Organizing Committee.

Faruqui, M., & Dyer, C. (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* (pp. 462–471). Gothenburg, Sweden: Association for Computational Linguistics.

Gittens, A., Achlioptas, D., & Mahoney, M. W. (2017). Skip-gram – zipf + uniform = vector additivity. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 69–76). Association for Computational Linguistics. doi:10.18653/v1/P17-1007.

Guo, J., Che, W., Yarowsky, D., Wang, H., & Liu, T. (2015). Cross-lingual dependency parsing based on distributed representations. 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)* (pp. 1234–1244). Beijing, China: Association for Computational Linguistics.

Hardoon, D. R., Szedmak, S. R., & Shawe-Taylor, J. R. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation, 16*(12), 2639–2664.

Harris, Z. (1954). Distributional structure. *Word, 10*(23), 146–162.

Hercig, T., Brychcín, T., Svoboda, L., Konkol, M., & Steinberger, J. (2016). Unsupervised methods to improve aspect-based sentiment analysis in Czech. *Computación y Sistemas, 20*(3), 365–375. doi:10.13053/CyS-20-3-2469.

Jurgens, D., Mohammad, S., Turney, P., & Holyoak, K. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The first joint conference on lexical and computational semantics – volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (SemEval 2012)* (pp. 356–364). Montréal, Canada: Association for Computational Linguistics.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of coling 2012* (pp. 1459–1474). Mumbai, India: The COLING 2012 Organizing Committee.

Konkol, M., Brychcín, T., & Konopík, M. (2015). Latent semantics in named entity recognition. *Expert Systems with Applications, 42*(7), 3470–3479. doi:10.1016/j.eswa.2014.12.015.

Köper, M., Scheible, C., & Schulte im Walde, S. (2015). Multilingual reliability and “semantic” structure of continuous word spaces. In *Proceedings of the 11th international conference on computational semantics* (pp. 40–45). London, UK: Association for Computational Linguistics.

Lazaridou, A., Dinu, G., & Baroni, M. (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)* (pp. 270–280). Beijing, China: Association for Computational Linguistics.

Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning* (pp. 171–180). Association for Computational Linguistics. doi:10.3115/v1/W14-1618.

Levy, O., Søgaard, A., & Goldberg, Y. (2017). A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (pp. 765–774). Valencia, Spain: Association for Computational Linguistics.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. In *In proceedings of the first workshop on evaluating vector space representations for nlp* (pp. 13–18). Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781*.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR, abs/1309.4168*.

Mogadala, A., & Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 692–702). San Diego, California: Association for Computational Linguistics.

Nomizu, K., & Sasaki, T. (1994). Affine differential geometry: Geometry of affine immersions. *Cambridge Tracts in Mathematics*. Cambridge University Press.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. doi:10.1109/TKDE.2009.191.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Salle, A., Villavicencio, A., & Idiart, M. (2016). Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 419–424). Berlin, Germany: Association for Computational Linguistics.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 298–307). Lisbon, Portugal: Association for Computational Linguistics.

Svoboda, L., & Beliga, S. (2017). Evaluation of Croatian word embeddings. *CoRR, abs/1711.01804*.

Svoboda, L., & Brychcín, T. (2016). New word analogy corpus for exploring embeddings of Czech words. *CoRR, abs/1608.00789*.

Turney, P. D. (2008). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research, 33*, 615–655. doi:10.1613/jair.2693.

Turney, P. D., Littman, M. L., & Shnayder, J. B. V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *In proceedings of the international conference on recent advances in natural language processing* (pp. 482–489).

Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1661–1670). Berlin, Germany: Association for Computational Linguistics.

Šnajder, J., Padó, S., & Agić, v. (2013). Building and evaluating a distributional memory for Croatian. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 784–789). Sofia, Bulgaria: Association for Computational Linguistics.

Vulić, I., & Korhonen, A. (2016). On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 247–257). Berlin, Germany: Association for Computational Linguistics.

Vulić, I., & Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research, 55*, 953–994.

Xing, C., Wang, D., Liu, C., & Lin, Y. (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). Denver, Colorado: Association for Computational Linguistics.