Knowledge-Based Systems 187 (2020) 104819

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Linear transformations for cross-lingual semantic textual similarity☆

Tomáš Brychcín

NTIS — New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Czech Republic

HIGHLIGHTS

- Linear transformations project monolingual semantic spaces into a shared space.
- We propose a new transformation outperforming others in the cross-lingual STS task.
- We extend unsupervised STS methods by the word weighting.
- Our approach achieves promising results on several datasets in different languages.

ARTICLE INFO

Article history:
Received 17 October 2018
Received in revised form 3 March 2019
Accepted 26 June 2019
Available online 27 June 2019

Keywords:
Semantic textual similarity
Semantic spaces
Linear transformations
Word embeddings
Cross-lingual semantic spaces

ABSTRACT

Cross-lingual semantic textual similarity systems estimate the degree of the meaning similarity between two sentences, each in a different language. State-of-the-art algorithms usually employ machine translation and combine vast amount of features, making the approach strongly supervised, resource rich, and difficult to use for poorly-resourced languages.

In this paper, we study linear transformations, which project monolingual semantic spaces into a shared space using bilingual dictionaries. We propose a novel transformation, which builds on the best ideas from prior works. We experiment with unsupervised techniques for sentence similarity based only on semantic spaces and we show they can be significantly improved by the word weighting. Our transformation outperforms other methods and together with word weighting leads to very promising results on several datasets in different languages.

1. Introduction

Semantic textual similarity (STS) systems estimate the degree to which two textual fragments (e.g., sentences) are semantically similar to each other. STS systems are usually evaluated by human judgments. The ability to compare two sentences in meaning is one of the core parts of natural language understanding (NLU), with applications ranging across machine translation, summarization, question answering, etc.

SemEval (International Workshop on Semantic Evaluation) has held the STS shared tasks annually since 2012. During this time, many different datasets and methods have been proposed. Early methods focused mainly on surface form of sentences and employed various word matching algorithms [1]. Han et al. [2] added distributional word representations and WordNet, achieving the best performance at SemEval 2013. Word-alignment methods introduced by Sultan et al. [3] yielded the best correlations at SemEval 2014 and 2015. Nowadays, the best performance tends to be obtained by careful feature engineering combining the best approaches from previous years together with deep learning models [4,5].

Lately, research in NLU is moving beyond monolingual meaning comparison. The research is motivated mainly by two factors: (a) cross-lingual semantic similarity metrics enable us to work 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. In the last two years, STS shared tasks [6,7] have been extended by cross-lingual tracks. The best performing systems [5,8] first employ an off-the-shelf machine translation service to translate input sentences into the same language and then apply state-of-the-art monolingual STS models. These highly-tuned approaches rely on manually annotated data, numerous resources, and tools, which significantly limit their applicability for poorly-resourced languages. Unlike the prior works, we study purely unsupervised STS techniques based on word distributional-meaning representations as the only source of information.

The fundamental assumption (*Distributional Hypothesis*) is that two words are expected to be semantically similar if they occur

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.06.027.

E-mail address: brychcin@kiv.zcu.cz.

https://doi.org/10.1016/j.knosys.2019.06.027
0950-7051/

in similar contexts (they are similarly distributed across the text). This hypothesis was formulated by Harris [9] several decades ago. Today it is the basis of state-of-the-art distributional semantic models [10–12]. Unsupervised methods for assembling word representations to estimate textual similarity have been proposed in [8,13,14]. We describe them in detail in Section 2.

Several approaches for inducing cross-lingual word 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 [15]. They can be roughly divided into three categories according to the level of required alignment: (a) document-level alignments [16], (b) sentence-level alignments [17], and (c) word-level alignments [18].

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 [18–21] and profit from weak supervision. Vulić and Korhonen [22] showed that bilingual dictionaries with a 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) [23].

In the very last years, the first attempts to unsupervised bilingual dictionary induction were introduced in [24–26]. These methods exploit the structural similarities across monolingual semantic spaces and automatically infer the cross-lingual mapping. Word translation experiments show that the automatically induced bilingual dictionaries are of high quality.

This paper investigates linear transformations for cross-lingual STS. We see three main contributions of our work:

- We propose a new linear transformation, which outperforms others in the cross-lingual STS task on several datasets.
- We extend previously published methods for unsupervised STS by word weighting. This leads to significantly better results.
- We provide thorough comparison of several linear transformations and several methods for STS.

This paper is organized as follows. In Section 2, we start with description of STS techniques based on combining word representations. The process of learning cross-lingual word representations via linear transformations is explained in Section 3. We propose our transformation in Section 4. We show our experimental results in Section 5 and conclude in Section 6.

2. Semantic textual similarity

Let $w \in \boldsymbol{V}$ denote the word, where $\boldsymbol{V}$ is a vocabulary. Let $\mathcal{S} : \boldsymbol{V} \mapsto \mathbb{R}^d$ be a semantic space, i.e., a function which projects word w into Euclidean space with dimension d. The meaning of the word w is represented as a real-valued vector $\mathbf{v}_w = \mathcal{S}(w)$. We assume *bag-of-words* principle and represent the sentence as a set (bag) $\boldsymbol{s} = \{w \in \boldsymbol{V}\}$, i.e., the word order has no role. Note we allow repetitions of the same word in the sentence (set). Given two sentences $\boldsymbol{s}^x$ and $\boldsymbol{s}^y$, the task is to estimate their semantic similarity $sim(\boldsymbol{s}^x, \boldsymbol{s}^y) \in \mathbb{R}$.

Brychcín and Svoboda [8] showed that inverse-document-frequency (IDF) weighting can boost STS performance. Inspired by their approach, we assume not all words in vocabulary $\boldsymbol{V}$ are of the same importance. We represent this importance by IDF weight $\lambda_w = idf(w)$. The importance of sentence $\boldsymbol{s}$ is then represented as a sum of corresponding word weights $\lambda_{\boldsymbol{s}} = \sum_{w \in \boldsymbol{s}} \lambda_w$.

In the following text we describe three STS approaches, which rely only on the word meaning representations, and we extend them to incorporate word weights λ_w. For the original version of STS algorithms, we consider uniform weighting of words, i.e., $\lambda_w = 1$ for all $w \in \boldsymbol{V}$.

2.1. Linear combination

Following *Frege's principle of compositionality* [27], which states that the meaning of a complex expression is determined as a composition of its parts (i.e., words), we represent the meaning of the sentence as a linear combination of word vectors

$$\mathbf{v}_{\boldsymbol{s}} = \frac{\sum_{w \in \boldsymbol{s}} \lambda_w \mathbf{v}_w}{\lambda_{\boldsymbol{s}}}. \tag{1}$$

Brychcín and Svoboda [8] showed that this approach leads to very good representation of short sentences. The similarity between sentences is then calculated as a cosine of angle between sentence vectors

$$sim(\boldsymbol{s}^x, \boldsymbol{s}^y) = cos(\mathbf{v}_{\boldsymbol{s}^x}, \mathbf{v}_{\boldsymbol{s}^y}). \tag{2}$$

2.2. Principal angles

Mu et al. [13] observed that the most information about words in a sentence is encoded in a low-rank subspace. Consequently, similar sentences should have similar subspaces. Using *Principal Component Analysis*, a technique for dimensionality reduction, we can find the linear subspace with most of the variance in word vectors.

Let $\mathbf{W} \in \mathbb{R}^{d \times |\boldsymbol{s}|}$ denote the sentence matrix with column vectors given by $\lambda_w \mathbf{v}_w$ for all $w \in \boldsymbol{s}$. Using *Singular Value Decomposition* (SVD) we decompose the matrix into $\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$. The matrix $\mathbf{U}_r$ is obtained by truncating the matrix $\mathbf{U}$ to keep only first r principal components. Finally, the similarity between two sentences $\boldsymbol{s}^x$ and $\boldsymbol{s}^y$ is defined as L2-norm of the singular values between corresponding subspaces $\mathbf{U}_r^x$ and $\mathbf{U}_r^y$

$$sim(\boldsymbol{s}^x, \boldsymbol{s}^y) = \sqrt{\sum_{i=1}^{r} \sigma_i^2}, \tag{3}$$

where σ_i denote the ith singular value of matrix $\mathbf{U}_r^{x\top}\mathbf{U}_r^y$. Note the singular values are in fact the cosines of the principal angles and can be obtained by SVD of $\mathbf{U}_r^{x\top}\mathbf{U}_r^y$ from the diagonal matrix $\boldsymbol{\Sigma}$.

According to Mu et al. [13], this method outperforms linear combination of word vectors on several datasets.

2.3. Optimal matching

The method presented in [3] has been very successful in STS tasks in the last years. This method finds and aligns the words that have similar meaning and similar role in the input sentences. The method is considered to be unsupervised in the sense it does not require sentence similarity judgments, however it relies on various language-specific tools (e.g., named entity recognition, dependency parsing, etc.).

Following this approach, Glavaš et al. [14] introduced the unsupervised word alignment method, which relies only on word meaning representations. Let $\boldsymbol{M} \subset \boldsymbol{s}^x \times \boldsymbol{s}^y$ denote the matching for input sentences $\boldsymbol{s}^x$ and $\boldsymbol{s}^y$ consisting of aligned word pairs $(w^x, w^y) \in \boldsymbol{M}$. Every word in both sentences can be at most in one pair. In fact we have a complete bipartite graph with nodes represented by words in input sentences, where the weight for an edge is the cosine similarity $\delta_{w^x, w^y} = cos\left(\mathbf{v}_{w^x}, \mathbf{v}_{w^y}\right)$. The task is to find an optimal matching (alignment) $\hat{\boldsymbol{M}}$ with the highest sum of cosine similarities between words in pairs. It can be found by so called *Hungarian Method* [28]. Finally, we estimate the matching score for $\boldsymbol{s}^x$ as

$$\delta_{\boldsymbol{s}^x} = \frac{1}{\lambda_{\boldsymbol{s}^x}} \sum_{(w^x, w^y) \in \hat{\boldsymbol{M}}} \lambda_{w^x} \delta_{w^x, w^y}. \tag{4}$$