The Role of Inflectional Morphology in Co-occurrence Semantic Representations

Alexander Johnson
Fermín Moscoso del Prado Martín

May 2014

Abstract

We investigate the extent to which morphological information can be used to enrich the semantic information contained in word co-occurrence vectors. Traditionally, it has been assumed that the reduction of sparsity provided by stemming or lemmatization can be used to improve word co-occurrence vectors for languages such as English. However, we argue that such an approach could have detrimental effects for morphologically rich languages such as Estonian, in which much semantic information is conveyed by the morphology. To solve this problem, we propose enriching the semantic vectors with morphological information. We compare the effects of lemmatization and morphological enrichment on equivalent corpora of English and Estonian. This provides a new technique for improving semantic vectors in morphologically rich languages.

3

1 Introduction

There are many ways in which a corpus can be preprocessed to improve results. In the present study, we compare the addition of morphological vectors to the preprocessing task of lemmatization. Lemmatization, also sometimes known as stemming, is a way to combine many separate forms or inflections of a word into the vector of its lemma. This reduction of sparsity can improve results.

In English, we use strict syntactic constituent ordering to help convey meaning. In contrast with strict ordering, there are many languages whose morphology can provide similar functionality. Estonian is able to rearrange the constituents in a sentence much more freely than English because of its inflectional paradigm. Estonian has 28 inflectional cases for nouns, whereas in English, there are only two (singular and plural). In circumstances in English where we might use a preposition phrase such as 'up to the' or 'on top of the', Estonian could use the the terminative or adessive case. We use this highly contrastive difference to elucidate the role of morphological information in vector-space models of semantics.

2 Data

The Open Parallel Corpus (OPUS) (http://opus.lingfil.uu.se/) is an online repository that includes translated texts from the Internet. We used the OpenSubtitles2013 corpus, a large collection of movie subtitles. We chose subtitles as our source of data partially because they offer a useful compromise between standardized and formal varieties, machine readability, and descriptive language use. A smaller section of the English corpus was used to make the two corpora of equivalent size, in terms of number of sentences [of sentences][1]. Freely available, open-source corpora use also allows for easily reproducible research and encourages transparency and collaboration.

We then used Helmut Schmid's freely available software TreeTagger to lemmatize the entirety of the corpora. For Estonian, TreeTagger was trained on the Tartu Morphologically disambiguated corpus. For English, TreeTagger was trained on the PENN Treebank and the English morphological database. TreeTagger returns tagged and lemmatized versions of the corpora. Once the corpora had been adjusted to equal size, and lemmatized versions were available, we used a small python script clean the data of punctuation, numbers, and other unwanted characters before calculating statistics of word-context co-occurrence.

3 Method

We choose the 40,000 highest frequency words as terms, and the 2,000 high frequency words as contexts. The thoroughly documented parameter adjustments performed by Bullinaria & Levy (2007) outline through a series of simulations of semantic tasks some of the values which consistently produce the best results. Although there are many types of semantic distance metrics available for use, that positive pointwise mutual information (PPMI) values in conjunction with a cosine distance metric consistently out-perform many of the other options [**?**] on a multitude of semantic tasks. Pointwise mutual information can be defined as Matrix dimensions of 40,000 terms by 2,000 contexts and a context window size of one immediately preceding and following word were used. Then the raw counts of co-occurrence are transformed into PPMI values before semantic distance is determined. (Niwa & Nitta, 94, Church & Hanks, 1990). If x and y represent two terms being compared, then PPMI and cosine similarity can be defined as:

$$ppmi(x, y) = log\frac{p(x, y)}{p(x)p(y)}$$

$$ppmi(x, y) < 0 = 0$$

$$similarity = cos\frac{x \cdot y}{\|x\|_2\|y\|_2}$$

[1] We also obtained the same results by matching by the number of words.

The cosine similarity between the vectors of PPMI values is then used as a measurement of semantic overlap between two given vectors. We compute co-occurrence vectors in three conditions: raw (unlemmatized), lemmatized, and lemmatized with the addition of morphological information. Lemmatization has been shown to reduce vector sparsity, which can lead to more reliable vectors (Rapp, 2003, Baroni et al., 2009, Bullinaria & Levy, 2012). This reduction of sparsity does offer advantages for both languages. However, we argue that for highly inflectional languages, the loss of morphological information could be detrimental, or affect the usefulness of lemmatization. In addition to the semantic vectors, we compute morphological vectors from the lemmatized corpora. Instead of using context words, we used the occurrence of an inflection as a count in a n-dimensional vector where n is equal to the number of possible inflections for that term. This vector represents how often a given term occurs in the corpus as a given form. This morphological vector is then concatenated to the semantic vector in order to regain some of the information that is lost during supposedly lost during lemmatization.

4 Analysis

In order to test the reliability of the semantic vectors, we simulate a standard multiple choice Test Of English as a Foreign Language (TOEFL) test (Landauer & Dumais, 1997). We use English and Estonian WordNets (NEED CITATION) to generate sets of semantically similar words (Miller et. al 1990) to use for comparison. One word from the set is randomly selected to represent a related word to one of the top 100 most frequent nouns. We simulate the TOEFL test by randomizing three sets of the pairs, and seeing if the true pairs have the highest semantic similarity (akin to asking a question such as which word is most similar to house: dog, building, hat, man). If the correct pair has the highest similarity, that pair counts as one answer correct on the simulation. We randomize and run the simulation 100 times on those 100 pairs to produce our results.

5 Results

Across the 100 randomizations, we found main effects of both language and condition. English does much better across conditions and is less affected by addition of morphological vectors. Analysis of variance shows significance between Estonian and English ($F[1,594] = 398.3692$, $p<.00001$) as well as condition ($F[2,594] = 317.8981$, $p<.00001$).

In our simulations, the addition of morphological vectors increases the performance by 1.27% for Estonian, and 0.27% for English.

Table 1: TOEFL Performance

Condition	Raw	Lemmatized	Morphology
English	53.94%	61.34%	61.61%
Estonian	48.98%	55.22%	56.49%

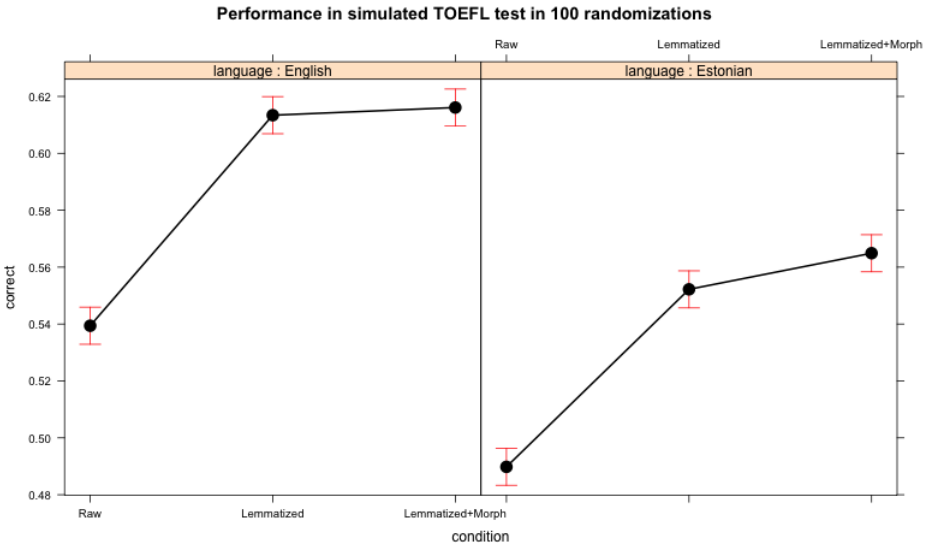

Figure 1. The figures shows performance in three conditions a.) no preprocessing (raw vectors) b.) lemmatized vectors and c.) lemmatized vectors with the addition of morphological vectors in both English and Estonian

6 Conclusions

Within vector space models, there is a diminishing return from increasing the corpus size used to compute the semantic vectors. However, as the corpus size increases, it has been shown that reliability of the vectors to represent semantics in psychologically relevant tasks is also increased. We have chosen to utilize a small corpus in order to test the efficacy of this technique. The addition of morphological vectors is a computationally simple process that capitalizes on the work that is already being done by a lemmatization engine. The incorporation of morphological information that would other wise be discarded has shown to improve the vector's performance on simulated TOEFL tests. The addition of morphological vectors may prove to be a useful tool when used in conjunction with lemmatization. Further investigation is needed to see how effects vary with other preprocessing tasks such as Singular Value Decomposition (SVD), and how these techniques may influence other, higher-order models of distribution based semantics. Highly inflectional languages may benefit more from this

corpora pre-processing task, but more research will need to be done to assess how morphological vectors affect typologically diverse languages (e.g. ones with non-concatenative morphologies).

7 References

Bullinaria, John A., and Joseph P. Levy. Behavior research methods 39.3 (2007): 510-526.
Bullinaria, John A., and Joseph P. Levy. Behavior research methods 44.3 (2012): 890-907.
Church, Kenneth Ward, and Patrick Hanks. Computational linguistics 16.1 (1990): 22-29.
Landauer, Thomas K., and Susan T. Dumais. Psychological review 104.2 (1997): 211.
Manning, Christopher D. Ed. Hinrich Schütze. MIT press, 1999. Miller, George A., et al. International journal of lexicography 3.4 (1990): 235-244.
Niwa, Yoshiki, and Yoshihiko Nitta. Proceedings of the 15th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1994.
OpenSubtitles: https://www.opensubtitles.org
Schmid, Helmut. Proceedings of the ACL SIGDAT-Workshop, 1995.
Tiedemann, Jörg. In Proceedings of the 8th International Conference on Language Resources and Evaluation, (LREC 2012)