Word Embeddings in Hebrew: Initial Results

Oded Avraham and **Yoav Goldberg**

Bar Ilan University

Abstract

Word embeddings algorithms got considerable attention in the past few years. However, they were all applied to English, a language with very limited morphology. We present our initial results with inferring word embeddings on Hebrew, a language with a much richer inflectional morphology system. The embeddings seem to provide a mix of semantic and morphological properties. Using lemmatization helps direct the resulting similarities away from the morphological similarity and towards semantic similarity. We are currently looking at improving the control over the aspects of the resulting similarities by investigating more refined and task-directed lemmatization.

Introduction

Extracting word similarities from text is a big challenge that has been extensively studied. In past few years, many algorithms had shown impressive results, most of them were evaluated on English text. Our purpose is to run an existing state-of-art word similarity algorithm on Hebrew text, evaluate its performance and try to improve it by applying adjustments. We have good reasons to believe that an algorithm which works well on English, will require adjustments to work well on Hebrew too. Hebrew is a morphologically rich language, with a structure that is quite different from English. The special feature in Hebrew which we expect to affect the algorithm the most is the rich templatic morphology: Hebrew words follow a complex morphological structure, which is based on a root and a template system. Word forms can encode gender, number, person and tense, and also construct-state (סמיכות). For example, the nouns חתול, חתולה, חתולים, חתולי all inflections of the base form חתול. The verbs קפץ, תקפוץ, קפוץ, יקפוץ, קופצים are all inflections of the base form קפץ. The adjectives כחול, כחולים, כחולת are all inflections of the base form כחול.

The interplay of morphological inflections and semantics in word embeddings

The algorithm we use is word2vec[1], which learns continuous distributed vector representations of words in a corpus. Having the trained vectors, we look for words which represented by similar vectors. After running the algorithm on the entire Hebrew Wikipedia, we've detected some interesting phenomena:

- Noun similarities tend to agree on gender, number and possessive form, for example: out of the top-10 similarities were found for the noun רחובותיה, only two had different possessive form and only one had different gender. The rest matched the exact morphological template (כבישיה, שווקיה, רובעיה...). This becomes a problem when weakly masculine related nouns like עמודיה, מבניה were ranked high above similar but feminine nouns like שכונותיה, שדרותיה.
- Nouns similarities which do not agree on the morphological template tend to be other inflections of the target word base form, not semantic similarities. In the above example, the two words with different possessive form were רחובות and רחובותיו which are inflections of the base noun רחוב, just like רחובותיה. The fact that they are being promoted over real semantic similarities is a problem.
- Construct-states inflections of target nouns are frequently identified as very strong similarities. For example: for the noun סיבה, the word סיבת was ranked first. As other inflections, this is not the kind of similarities we would like to find.
- Adjectives behave pretty much the same as nouns. For example: out of the top-10 similarities of the adjective נמוכים, nine match the target morphologically template (פחותים, קטנים, חלשים…). The only exception is the word נמוך which shares the same base form with the target. However, adjectives do not suffer from the other nouns problems: since they have no defined gender (the gender of an adjective is determined by the gender of the noun it modifies) the

gender problem is irrelevant for adjective. Furthermore, the construct-state similarities problem was not observed for adjectives.

- Top similarities for verb tend to be other inflections of the same base form, not semantic similarities. For example: out of the top-10 similarities of the verb הוקדשו, only three were semantically related, while the rest were inflections (מוקדשים, הוקדשה, הקדישו).
- Verb similarities which are really semantically related to the target word, tend to agree on the morphological template. In the above example, the three non-inflections verbs (יוחדו, נתרמו, עסקו) are all matching the template of הוקדשו.

The effect of lemmatization

We saw that the inflectional structure of Hebrew has a major effect on similarities of verbs, verb and adjective targets. Hence, it may be worth to eliminate this structure by lemmatizing the corpus, i.e., replacing every word by its base form. In order to do that, we first parsed the text using a Hebrew dependency parser[2] and then used the morphological properties of each word to find its base form in a Hebrew inflections dictionary[3]. By running the algorithm on the lemmatized corpus, we managed to solve most of the problems were mentioned before:

- Feminine nouns were not demoted. For example: in the lemmatized version, the nouns שכונה, סמטה, כיכר were ranked much higher than in the inflected version.
- Except for few morphological analysis mistakes, inflections of the words were not determined as similarities (since inflections do not exist in the lemmatized corpus). This resulted in more semantic similarities.

Furthermore, by lemmatizing the corpus we reduce its sparsity. However, lemmatizing can also harm the accuracy in some cases. For example, the most similar word found for נפלה is נכבשה, while for the word נפל, the strongest similarity was נהרג. This difference expresses the different character of their contexts. While נפל frequently occurs next to warrior entities, and in this context is similar to נהרג, the word נפלה tends to co-occur with nations and lands, and in that context is similar to נכבשה. Lemmatizing the corpus mixes together all the inflections contexts and creates some “average” representation. Querying the similarities of the lemma can be much less accurate than querying the similarities of a specific inflection. In the similarities of נפל for the lemmatized version, the word נהרג was ranked only 6, while the word נכבש were not even in the top-10. The average meaning of נפל inflections is more similar to words like נלכד and נלקח.

Future work

We would like to find a way to control of the type of similarities we get, hence to separate the syntactic similarities from the semantic ones. As shown, we found a simple way to get mostly semantic similarities, but we still do not know how to get mostly syntactic ones. Moreover, our solution for semantic similarities should also be reconsidered due to the accuracy problem which we discussed. Therefore, it may be worth to try some deeper adaptations. We plan to begin by exploring various kinds of partial lemmatization: we can try lemmatizing only part of the morphology (gender/person…), only specific parts of speech (verb/nouns...) or only part of a training sample (contexts/word), in order to tailor the results to specific desired behaviours.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[2] Goldberg, Y. (2011). Automatic Syntactic Processing of Modern Hebrew. PhD thesis, Ben Gurion University.

[3] [Itai and Wintner, 2008] Itai, A. and Wintner, S. (2008). Language resources for Hebrew. Language Resources and Evaluation, 42(1):75-98.