Evaluating Ways of Adapting Word Similarity

Libby Barak  
Rutgers University  
Newark, NJ, USA  
libby.barak@rutger.edu

Adele Goldberg  
Princeton University  
Princeton, NJ, USA  
adele@princeton.edu

Abstract

People judge pairwise similarity by deciding which aspects of the words’ meanings are relevant for the comparison of the given pair. However, computational representations of meaning rely on all dimensions of the vector representation for similarity comparisons, without considering the specific pairing at hand. Prior work has adapted computational similarity judgments by using the softmax function in order to address this limitation by capturing asymmetry in human judgments. We extend this analysis by showing that a simple modification of cosine similarity offers a better correlation with human judgments over a comprehensive dataset. The modification performs best when similarity between two words is calculated with reference to other words that are most similar and dissimilar to the pair.

1 Introduction

Human interpretation relies implicitly on context, and context therefore plays a key role in judgments of similarity (Medin et al., 1993; Tversky, 1977; Schvaneveldt et al., 1976). For example, while people judge *ascend* and *descend* to be highly similar in context of elevators, in the context of morality, they will likely be judged as highly dissimilar (Gerz et al., 2016). As words are interpreted differently in different contexts, human similarity judgments are often not transitive. For example, *add* is highly similar to *pour* in the context of *cook*, and *add* is similar to *multiply* in the context of *compute*. But *pour* and *multiply* are not at all similar, despite their individual similarity to *add*.

Computational models of language processing critically rely on calculations of semantic similarity as a basic component of many natural language processing methods. Representations of meaning have come a long way in recent years due to the availability of large datasets and novel methods of computation (Peters et al., 2018; Mikolov et al., 2013; Pennington et al., 2014). However, like earlier models, the newer methods represent word meanings as vectors and similarity has been mostly calculated as the geometric distance between vectors. Despite using a multi-dimensional space of features, distance metrics of similarity fails to replicate human behavior in certain recognized ways (Griffiths et al., 2007; Nematzadeh et al., 2017; Bakarov, 2018).

Prior work proposed methods of improving on pre-trained vectors using human-annotated thesaurus data, nuanced vectors to address ambiguity, or by mathematically modifying the vectors to enhance which dimensions are assumed to be useful (Kanerva et al., 2017; Remus and Biemann, 2018; Mu et al., 2017). Importantly, these methods all address the challenge by modifying vectors in a global, non-task-specific way. They primarily use the cosine similarity measure to compare the resulting vectors which fails to address the geometrical constraints evident in human similarity judgments (Tversky, 1977).

Recent work has begun to focus on cognitively-plausible similarity measures. (Nematzadeh et al., 2017) suggest using softmax function instead of cosine or Euclidean distance in order to capture asymmetric similarity judgments (see Equation [1]). Their analysis using a dataset of directional judgments on noun pairs demonstrated that the softmax function correlated better with human judgments than simple cosine function did. We recognize that a possible key advantage of Equation [1] is the inclusion of the semantic relation between the word $w_1$ and other words $w_j$, in the denominator, where $w_j$ is every other word in the lexicon. The denominator may be providing contextual information on how similar the pair of words are in comparison to their similarity to other words.
2 Comparing Softmax and Cosine

We present an analysis of how invoking similarity to other words in the denominator can be incorporated into the cosine function as well. We also extend the method by analyzing various ways of representing context by varying the scope of $w_j$. To compare whether using softmax offers an advantage over the more common cosine measure, when the latter is adapted to include an analogous denominator, we introduce the modified cosine measure, relative-cosine (See Equation 2).

$$P(w_1|w_2) = \frac{\exp(w_1 \cdot w_2)}{\sum_{w_j} \exp(w_j \cdot w_1)}$$ (1)

$$P(w_1|w_2) = \frac{\cos(w_1, w_2)}{\sum_{w_j} \cos(w_j \cdot w_1)}$$ (2)

We compare Equations 1 and 2 on the SimVerb3500 dataset, which includes human similarity judgments for 3500 verb pairs, on the basis of responses from 10 annotators per pair and a manual classification of the type of relation between the verbs, e.g., Antonyms, Synonyms, etc. (Gerz et al., 2016). Verbs are used because they play a central role in understanding the meaning of sentences and they are particularly sensitive to context. We use Word2Vec (Mikolov et al., 2013), pre-trained on the British National Corpus with a 10 word window, using Skip-gram and including part-of-speech tagging. We find that restricting the reference to verbs improves results.

The denominators in the equations only include other words’ similarity to the first word included in the numerator, ensuring that the computations are asymmetric. Because here the human data is based on symmetric similarity judgments, we compare performance on a) the average of the scores for both verbs, b) scores when the higher frequency verb is included, and c) scores when the lower frequency verb is included. The best performance is obtained using the lower frequency verb as the reference point for the denominator. Since high frequency verbs tend to be more polysemous, we hypothesize that the lower frequency verbs provide a more concise context for the denominator.

Our results show that using the context information provided by the denominator leads to a statistically significant improvement in the Spearman correlation between human and computational similarity scores. Relative-cosine and softmax versions perform equally well, with the relative-cosine having marginally better performance over softmax for synonyms (0.04 improvement over the cosine similarity).

3 Evaluating Alternative Methods for Context Representation

Although useful, the methods just described do not necessarily capture how people seem to approach similarity judgment tasks. We had suggested that in comparing add and multiply, people implicitly consider verbs that label other mathematical operations, not every verb in the lexicon. Moreover, people are likely to consider those dimensions that are most relevant to mathematics, rather than features related to the many other contexts in which add might be used.

We find we can increase the correlation with human judgments by incorporating two additional steps. First, we identify the $N$ verbs that are the most similar to only one of the two verbs in the numerator: e.g., verbs that are most similar to add but not to multiply and vice versa (the 10 verbs with the highest difference in cosine similarity to the two words). We subtract the sum of these $N$ vectors from the representation of the two verbs as a way to reduce the role of dimensions less relevant to the comparison at hand. We present here the results with $N = 10$, which produced the best result with no significant improvement when using higher $N$ values.

Second, as reference points in the denominator, we select only the 10 most similar verbs to both members of the pair. That is, $w_j$ in Eq. 2 now spans over only the 10 verbs with the highest cosine similarity to both verbs in the pair. Using this method emphasizes relevant dimensions as it involves comparisons only to relevant verbs, and it achieves the best improvement over baseline (0.08). This method is especially effective for synonyms and hyper/hyponyms, where it leads to 0.10 improvement in its correlation with human judgments. Although the improvements are modest compared with recent achievements, our analysis provides evidence that it is beneficial to adjust pre-trained vectors to the task at hand. Moreover, our analysis supports a cognitively-plausible explanation of how people may approach similarity judgments. We aim to extend these preliminary results with analysis of more robust meaning representations. In addition, we use these results to guide the design of an experimental study.
of human judgment of meaning in context (under submission).

References


