UZH at SemEval-2020 Task 3: Combining BERT with WordNet sense embeddings to predict graded word similarity changes

Li Tang
University of Zurich, Rämistr. 71, 8006 Zürich
li.tang@uzh.ch

Abstract

CoSimLex is a dataset that can be used to evaluate the ability of context-dependent word embeddings for modeling subtle, graded changes of meaning, as perceived by humans during reading. At SemEval-2020, task 3, subtask 1 is about "predicting the (graded) effect of context in word similarity", using CoSimLex to quantify such a change of similarity for a pair of words, from one context to another. Here, a meaning shift is composed of two aspects, a) discrete changes observed between different word senses, and b) more subtle changes of meaning representation that are not captured in those discrete changes. Therefore, this SemEval task was designed to allow the evaluation of systems that can deal with a mix of both situations of semantic shift, as they occur in the human perception of meaning. The described system was developed to improve the BERT baseline provided with the task, by reducing distortions in the BERT semantic space, compared to the human semantic space. To this end, complementarity between 768- and 1024-dimensional BERT embeddings, and average word sense vectors were used. With this system, after some fine-tuning, the baseline performance of 0.705 (uncentered Pearson correlation with human semantic shift data from 27 annotators) was enhanced by more than 6%, to 0.7645. We hope that this work can make a contribution to further our understanding of the semantic vector space of human perception, as it can be modeled with context-dependent word embeddings in natural language processing systems.

1 Introduction

Context-dependent word embeddings such as BERT (Devlin et al., 2019) provide semantic vector representations of words which depend on the surrounding context, capturing not only discrete differences in word sense, but also the more graded effects of context on word meaning (Armendariz et al., 2019). While discrete shifts in meaning can often be aligned with word senses for polysemous English words in the WordNet knowledge graph (Fellbaum, 1998), more subtle, graded shifts in meaning cannot be fully represented in this manner. As such embeddings are vectors, for which distances can be calculated, relative shifts in meaning can in principle be modeled by distances in that semantic vector space. However, it is not fully understood how well such distances based on context-dependent word embeddings actually represent graded word similarity changes as perceived by humans while reading. While their ability to predict human cognitive data collected during reading has been investigated, indicating a non-random correspondance (Hollenstein et al., 2019), additional approaches are needed to better assess their ability to model human perceptions of word semantics under natural conditions.

By compiling the CoSimLex dataset, Armendariz et al. (2019) enable an evaluation of such embeddings in terms of their ability to model human perceptions of meaning, when performing similarity judgements in context. In other words, they capture both graded and discrete shifts in meaning, depending on the contexts in which a pair of words appear. This evaluation in SemEval-2020 therefore enables new research questions to be asked, compared to previously compiled datasets for the modeling of word semantics,

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 166–170
Barcelona, Spain (Online), December 12, 2020.

by reflecting the gradedness of human semantic judgements. While CoSimLex provides a multi-lingual dataset, the described system in this paper only predicts graded effects of context on word meaning for the largest subset of CoSimLex, which is composed of 340 English word pairs in two different contexts. An extension of this system to other languages is in principle feasible, but would require considerably more work as a key system component described here, the average word sense vectors based on the work of Loureiro and Jorge (2019), only exist for English words, at present.

To improve upon the relatively strong baseline script provided with the task, an approach was developed that adds a multi-parametric semantic space distortion correction model to the baseline algorithm. In the unmodified baseline, a pre-trained 768-dimensional BERT model generates a context-dependent embedding for each word in the pair. Then the similarity between those vectors is calculated, using cosine distance. Once the same process has been repeated for the same word pair in the second context, both similarity values can be compared using subtraction. This results in a positive or negative value, representing the "semantic shift" for this word pair, depending on the meanings of those two words shifting towards being more or less similar. In many cases, this perception of meaning shift does not reflect the use of a different word sense, but something more gradual, in the human semantic vector space.

In the baseline algorithm, which word sense (according to WordNet) is used in both contexts is not determined. To experiment with a model that would either boost or shrink the result produced by the baseline, modeling a distortion correction that would approximate a more human judgement of semantic shift, an algorithm was developed for performing a number of comparisons between the embeddings produced by the baseline, and average embeddings for different word senses for that word. The latter aspect taps into semantic knowledge captured in a computationally accessible form, as semantic vectors, using a knowledge graph called WordNet (Fellbaum, 1998).

To integrate such knowledge into the logic of the baseline, pre-trained embeddings for all discrete word senses stored in WordNet were obtained from the work of Loureiro and Jorge (2019). In brief, Loureiro and Jorge calculated average vectors from BERT embeddings for each English word sense recorded in WordNet, based on a large corpus in which those distinct word senses were labeled with their WordNet identifiers. For a particular target word in a word pair in our task, those pre-trained embeddings for word senses are then compared to standard 1024-dimensional BERT-based embeddings that do not take word sense knowledge in WordNet into account, as in the baseline. Basically, this correction model was designed to either boost or shrink the similarity shift value generated by the baseline, depending on a few factors derived from parametrized semantic vector space calculations, aiming to improve the system's ability to model human perceptions of meaning shift. In other words, the model aims to reduce distortions in the original BERT-based semantic space, from the perspective of human semantic perception. After experimenting with a few parameters that influence the balance between different distortion corrections in this model, a final combination was chosen that substantially improved the match with the human perception of meaning shift. This fine-tuning of the system was guided by human inspection, based on the author's subjective perception of semantic shift. During such inspection, did the meaning of the two words become more or less similar in the second context, compared to the first? Was it a substantial shift, to a semantically very different word sense, or a more subtle discrete or graded shift? In total, 32 combinations of 4 parameters were evaluated in this way, before no further improvement of the system was observed. This combination also improved the performance in the SemEval-2020 evaluation task, under the conditions of post-evaluation in CodaLab.

Note that the true labels for the evaluation set in this task, the graded similarity changes obtained from the 27 human annotators, for 340 word pairs, were not available during system development and fine-tuning, as reported in Table 1, preventing other methods for hyper-parameter optimization. However, a small practice set of 10 labeled word pairs was provided by the task organizers, for getting started with algorithm design. Once the main logic of the algorithm was developed, fine-tuning was performed as described above, to improve the ability of the system output to model human perception.

2 Background

The English language subset of the CoSimLex dataset, the focus of this work, consists of 340 word pairs, with each word pair occurring in 2 different multi-sentence contexts. See Armendariz et al. (2019) for an example, and a more detailed description of the theory, the human annotation method, as well as the calculation of meaning shift. Note that the UZH system only performs subtask 1, the graded similarity shift between context 1 and 2 for a particular word pair, for the English language subset, for the reasons outlined above.

3 System Overview

The UZH system uses the similarity values generated by the baseline script as input to a semantic space distortion correction model for either boosting (increasing) or shrinking (decreasing) that value, based on a range of calculations in two complementary semantic vector spaces. The basic intuition behind this design is to reduce distortions in the semantic vector space modeled by the baseline algorithm, resulting from differences in the way BERT represents meaning shifts using cosine distances in vector space, compared to perceptions of semantic shift by the 27 human annotators. One of the main ideas was that averaged vectors for word senses from large training data would allow the empirical discovery of such distortion corrections for semantic vector space, using a limited number of trials, guided by visual inspection. Therefore, in each iteration of system fine-tuning, each word pair was printed with both contexts, together with the baseline prediction of semantic shift, as well as the output and behavior of the distortion correction model.

To match the type of BERT model used to generate those pre-trained word sense embeddings (1024 dimensions) with the BERT model used in the baseline script (768 dimensions), the 1024 cased version of the BERT model, trained on the book corpus, was used for this comparison with the averaged word sense vectors.

In a first step, distances between embeddings for all word senses and the context-dependent 1024-dimensional BERT embedding are calculated. Then, the word sense embedding with the highest similarity to the target word embedding is selected for further calculations (vector WSEsel), representing the most relevant discrete word sense. By calculating distances in semantic vector space, compared to this WSEsel reference, the model can apparently better reflect subtle, gradual sense shifts for the same word sense, in different contexts. Distances to WSEsel are also calculated for the second target word in the word pair, for comparison, thereby generating all required parameters for the correction model, as outlined in Equation 1. Fine-tuning of those 4 parameters then helped to determine the best balance between the factors influencing the reduction of distortion.

Equation 1: $s = (a + b) * (w2 + sl * w3 + as * w4)$

s: similarity (prediction), a: baseline prediction, using 768-dimensional BERT embeddings (BBE768) b: 1024-dimensional BERT prediction (difference to a, divided by $w1$ weight; negative if a was negative), sl: average of both distances between the most similar word sense embedding (WSEsel) and the 1024-dimensional BERT embedding (BBE1024) for the target words as: change of distance between the WSEsel and BBE1024 for the target words, between context 1 and context 2.

In summary, the conceptual stages of the algorithm that apply the correction model to the baseline script outputs are:

1. **Reference similarity prediction** generated by the baseline script, using BBE768, resulting in output a

2. **Correction 1** generated by the 1024 cased version of the BERT model trained on the book corpus (BBE1024), resulting in output b

3. **Correction 2** generated by the comparison of BBE1024 (both words in the word pair) with WSEsel

The pre-trained word sense embeddings can be considered a summarization of a population of embeddings in semantic vector space, for that particular word sense, capturing vector representation of that word sense. Due to the averaging, this vector is independent of variations in the context that do not affect the annotation of word sense in its discrete form. For example, the word 'play' can refer to child's play, playing a role in a movie, and playing an instrument while making music (three word senses). Therefore, it can capture each distinct word sense as a different semantic vector, based on how word senses are stored in the WordNet knowledge graph. Hence, the distortion correction model described here uses the distance between the context-dependent embedding for a target word, which can reflect graded shifts in meaning, and those summarized embeddings for that word sense, for the larger discrete shifts in meaning. While the 768-dimensional and 1024-dimensional BERT models may have distinct advantages and disadvantages in modeling human perceptions of meaning, the use of WSEsel provides an orthogonal way to introduce additional information in assessing such distortions.

The hyper-parameters in this distortion correction model were adjusted based on multiple submissions, as outlined above. See Table 1. During visual inspection, model predictions were monitored to detect cases where the correction model showed an exaggerated or too weak boosting of graded meaning shifts, compared to human perception of meaning shifts by the author (which may not correspond perfectly to the average of the 27 human annotators in CoSimLex.

As it was impossible, under post-evaluation conditions, to use standard methods for supervised learning on this task, as outlined above, it is plausible that the performance of the correction model may be further enhanced, once the labels are used for such parameter optimization. See below for a few experiments on the labels shared by the conference organizers after the submission of this paper. In other words, the effort for fine-tuning described here may well have found a local rather than a global optimum. Further investigation into the most salient aspects of the algorithm could therefore lead to useful insights into the modeling of human perceptions of meaning shifts using a complementary set of BERT and other embeddings, which seem to have unique advantages that can be combined in a synergistic manner. The design of the presented algorithm may be a starting point for such investigations.

For details on the algorithm, see equation 1, and the code (written in Python 3.7.6) at `https://github.com/lilytang2017/semeval2020`

4 Experimental Setup

As no supervised learning could be applied during fine-tuning of model parameters, no data splits were performed. Python libraries used: bert_embedding, numpy, pandas, and scipy. Evaluation measures were predefined by the task organizers, as described by Armendariz et al. (2019), basically measuring correlation.

After paper submission, the conference organizers shared the gold labels, reflecting the human perception of meaning shifts in CoSimLex, as perceived by 27 human annotators. Based on those labels, additional combinations of hyperparameters were investigated, beyond those shown in Table 1.

5 Results

Table 1 shows the result scores (correlations) for 4 adjustable parameters (weights) in the distortion correction model, see Equation 1 above. They reflect the balance between contributions from different embeddings and vector calculations to the final output, which aims to reduce distortions in the semantic vector space modeled by context-dependent embeddings.

The best score obtained with this correction model was 0.7645, reaching the top rank at the time of submission, in the post-evaluation phase of SemEval-2020. As this is a considerable improvement over the baseline prediction, which reached a score of 0.705, the described correction model is able to make a contribution by combining the strengths of different semantic vectors and vector calculations.

After obtaining the gold labels, 70 additional combinations of hyper-parameters were evaluated. Under those conditions, a slight improvement to a score of 0.7649 was observed, with the parameters: w1 = 23, w2 = 3, w3 = 6, w4 = 2.5. Parameter w3 was observed to be optimal near value 6, when testing values up to 20. Each run took about 30 min for each hyper-parameter combination.

w1	w2	w3	w4	score
3	5	5	0	0.613
3	3.8	2.2	0.2	0.736
3	3.8	2.2	0.9	0.7525
5	2.6	2.9	1.3	0.7602
5	2	4.5	1.5	0.7605
5	4	5	1.8	0.7633
5	3	6	2.5	**0.7645**

Table 1: *Scores obtained with different parameters in the correction model, guided by visual inspection (human perception of meaning shifts, by the author). Best score in bold.*

6 Conclusion

As the application of the described multi-parameter correction model to the predictions of the baseline script were able to improve performance by over 6%, we can conclude that the further investigation of this approach for reducing distortions in the BERT semantic space could help us better model the human semantic vector space at word level, as far as meaning shifts captured in CoSimLex are concerned.

References

Carlos S. Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Marko Robnik-Šikonja, Ivan Vulić, Mohammad Taher Pilehvar. 2019. *SemEval-2020 Task 3: Graded Word Similarity in Context (GWSC)* Proceedings of the 14th International Workshop on Semantic Evaluation, 2020.

Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), p4171-4186.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database* MIT Press, 1998.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, Ce Zhang. 2019. *CogniVal: A Framework for Cognitive Word Embedding Evaluation* Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). 2019,

Daniel Loureiro, Alıpio Mario Jorge. 2019 *Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation* arXiv:1906.10007v1.