CECL: a New Baseline and a Non-Compositional Approach for the Sick Benchmark

Yves Bestgen
Centre for English Corpus Linguistics
Université catholique de Louvain
yves.bestgen@uclouvain.be

Abstract

This paper describes the two procedures for determining the semantic similarities between sentences submitted for the SemEval 2014 Task 1. MeanMaxSim, an unsupervised procedure, is proposed as a new baseline to assess the efficiency gain provided by compositional models. It outperforms a number of other baselines by a wide margin. Compared to the word-overlap baseline, it has the advantage of taking into account the distributional similarity between words that are also involved in compositional models. The second procedure aims at building a predictive model using as predictors MeanMaxSim and (transformed) lexical features describing the differences between each sentence of a pair. It finished sixth out of 17 teams in the textual similarity sub-task and sixth out of 19 in the textual entailment sub-task.

1 Introduction

The SemEval-2014 Task 1 (Marelli et al., 2014a) was designed to allow a rigorous evaluation of compositional distributional semantic models (CDSMs). CDSMs aim to represent the meaning of phrases and sentences by composing the distributional representations of the words they contain (Baroni et al., 2013; Bestgen and Cabiaux, 2002; Erk and Pado, 2008; Grefenstette, 2013; Kintsch, 2001; Mitchell and Lapata, 2010); they are thus an extension of Distributional Semantic Models (DSMs), which approximate the meaning of words with vectors summarizing their patterns of co-occurrence in a corpus (Baroni and Lenci, 2010; Bestgen et al., 2006; Kintsch, 1998; Landauer and Dumais, 1997). The dataset for this task, called SICK (*Sentences Involving Compositional Knowledge*), consists of almost 10,000 English sentence pairs annotated for relatedness in meaning and entailment relation by ten annotators (Marelli et al., 2014b).

The rationale behind this dataset is that "understanding when two sentences have close meanings or entail each other crucially requires a compositional semantics step" (Marelli et al., 2014b), and thus that annotators judge the similarity between the two sentences of a pair by first building a mental representation of the meaning of each sentence and then comparing these two representations. However, another option was available to the annotators. They could have paid attention only to the differences between the sentences, and assessed the significance of these differences. Such an approach could have been favored by the dataset built on the basis of a thousand sentences modified by a limited number of (often) very specific transformations, producing sentence pairs that might seem quite repetitive. An analysis conducted during the training phase of the challenge brought some support for this hypothesis. The analysis focused on pairs of sentences in which the only difference between the two sentences was the replacement of one content word by another, as in *A man is singing to a girl* vs. *A man is singing to a woman*, but also in *A man is sitting in a field* vs. *A man is running in a field*. The material was divided into two parts, 3500 sentence pairs in the training set and the remaining 1500 in the test set. First, the average similarity score for each pair of interchanged words was calculated on the training set (e.g., in this sample, there were 16 sentence pairs in which *woman* and *man* were interchanged, and their mean similarity score was 3.6). Then, these mean scores were used as the similarity scores of the sentence pairs of the test sample

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 160–165,
Dublin, Ireland, August 23-24, 2014.

in which the same words were interchanged. The correlation between the actual scores and the predicted score was 0.83 (N=92), a value that can be considered as very high, given the restrictions on the range in which the predicted similarity scores vary (min=3.5 and max=5.0; Howell, 2008, pp. 272-273). It is important to note that this observation does not prove that the participants have not built a compositional representation, especially as it only deals with a very specific type of transformation. It nevertheless suggests that analyzing only the differences between the sentences of a pair could allow the similarity between them to be effectively estimated.

Following these observations, I opted to try to determine the degree of efficacy that can be achieved by two non-compositional approaches. The first approach, totally unsupervised, is proposed as a new baseline to evaluate the efficacy gains brought by compositional systems. The second, a supervised approach, aims to capitalize on the properties of the SICK benchmark. While these approaches have been developed specifically for the semantic relatedness sub-task, the second has also been applied to the textual entailment subtask. This paper describes the two proposed approaches, their implementation in the context of SemEval 2014 Task 1, and the results obtained.

2 Proposed Approaches

2.1 A New Baseline for CDSM

An evident baseline in the field of CDSM is based on the proportion of common words in two sentences after the removal (or retaining) of stop words (Cheung and Penn, 2012). Its main weakness is that it does not take into account the semantic similarities between the words that are combined in the CDSM models. It follows that a compositional approach may seem significantly better than this baseline, even if it is not compositionality that matters but only the distributional part. At first glance, this problem can be circumvented by using as baseline a simple compositional model like the additive model. The analyses below show that this model is much less effective for the SILK dataset than the distributional baseline proposed here.

MeanMaxSim, the proposed baseline, is an extension of the classic measure based on the proportion of common words, taking advantage of the distributional similarity but not of compositionality. It corresponds to the mean, calculated using all the words of the two sentences, of the maximum semantic similarity between each word in a sentence and all the words of the other sentence. More formaly, given two sentences $a = (a_1, .., a_n)$ and $b = (b_1, ..b_m)$,

$$MMS = \frac{(\sum_i \max_j sim(a_i,b_j) + \sum_j \max_i sim(a_i,b_j))}{n+m}$$

In this study, the cosine between the word distributional representations was used as the measure of semantic similarity, but other measures may be used. The common words of the two sentences have an important impact on MeanMaxSim, since their similarity with themselves is equal to the maximum similarity possible. Their impact would be much lower if the average similarity between a word and all the words in the other sentence were employed instead of the maximum similarity. Several variants of this measure can be used, for example not taking into account every instance where a word is repeated in a sentence or not allowing any single word to be the "most similar" to several other words.

2.2 A Non-Compositional Approach Based on the Differences Between the Sentences

The main limitation of the first approach in the context of this challenge is that it is completely unsupervised and therefore does not take advantage of the training set provided by the task organizers. The second approach addresses this limitation. It aims to build a predictive model, using as predictors MeanMaxSim but also lexical features describing the differences between each sentence of a pair. For the extraction of these features, each pair of sentences of the whole dataset (training and testing sets) is analyzed to identify all the lemmas that are not present with the same frequency in both sentences. Each of these differences is encoded as a feature whose value corresponds to the unsigned frequency difference. This step leads to a two-way contingency table with sentence pairs as rows and lexical features as columns. Correspondence Analysis (Blasius and Greenacre, 1994; Lebart et al., 2000), a statistical procedure available in many off-the-shelf software like R (Nenadic and Greenacre, 2006), is then used to decompose this table into orthogonal dimensions ordered according to the corresponding part of associations between rows and columns they explain. Each row receives a coordinate on these dimensions and these coordinates are used as predictors of the relatedness scores of the sentence

pairs. In this way, not only are the frequencies of lexical features transformed into continuous predictors, but these predictors also take into account the redundancy between the lexical features. Finally, a predictive model is built on the basis of the training set by means of multiple linear regression with stepwise selection of the best predictors. For the textual entailment sub-task, the same procedure was used except that the linear regression was replaced by a linear discriminant analysis.

3 Implementation Details

This section describes the steps and additional resources used to implement the proposed approaches for the SICK challenge.

3.1 Preprocessing of the Dataset

All sentences were tokenized and lemmatized by the Stanford Parser (de Marneffe et al., 2006; Toutanova et al., 2003).

3.2 Distributional Semantics

Latent Semantic Analysis (LSA), a classical DSM (Deerwester et al., 1991; Landauer et al., 1998), was used to gather the semantic similarity between words from corpora. The starting point of the analysis is a lexical table containing the frequencies of every word in each of the text segments included in the corpus. This table is submitted to a singular value decomposition, which extracts the most significant orthogonal dimensions. In this semantic space, the meaning of a word is represented by a vector and the semantic similarity between two words is estimated by the cosine between their corresponding vectors.

Three corpora were used to estimate these similarities. The first one, the TASA corpus, is composed of excerpts, with an approximate average length of 250 words, obtained by a random sampling of texts that American students read (Landauer et al., 1998). The version to which T.K. Landauer (Institute of Cognitive Science, University of Colorado, Boulder) provided access contains approximately 12 million words.

The second corpus, the BNC (British National Corpus; Aston and Burnard, 1998) is composed of approximately 100 million words and covers many different genres. As the documents included in this corpus can be of up to 45,000 words, they were divided into segments of 250 words, the last segment of a text being deleted if it contained fewer than 250 words.

The third corpus (WIKI, approximately 600 million words after preprocessing) is derived from the Wikipedia Foundation database, downloaded in April 2011. It was built using WikiExtractor.py by A. Fuschetto. As for the BNC, the texts were cut into 250-word segments, and any segment of fewer than 250 words was deleted.

All these corpora were lemmatized by means of the TreeTagger (Schmid, 1994). In addition, a series of functional words were removed as well as all the words whose total frequency in the corpus was lower than 10. The resulting (log-entropy weighted) matrices of co-occurrences were submitted to a singular value decomposition (SVDPACKC, Berry et al., 1993) and the first 300 eigenvectors were retained.

3.3 Unsupervised Approach Details

Before estimating the semantic similarity between a pair of sentences using MeanMaxSim, words (in their lemmatized forms) considered as stop words were filtered out. This stop word list (n=82), was built specifically for the occasion on the basis of the list of the most frequent words in the training dataset.

3.4 Supervised Approach Details

To identify words not present with the same frequency in both sentences, all the lemmas (including those belonging to the stop word list) were taken into account. The optimization of the parameters of the predictive model was performed using a three-fold cross-validation procedure, with two thirds of the 5000 sentence pairs for training and the remaining third for testing. The values tested by means of an exhaustive search were:

- Minimum threshold frequency of the lexical features in the complete dataset: from 10 to 70 by step of 10.
- Number of dimensions retained from the CA: from 10 to the total number of dimensions available by step of 10.
- P-value threshold to enter or remove predictors from the model: 0.01 and from 0.05 to 0.45 by step of 0.05.

This cross-validation procedure was repeated five times, each time changing the random distribution of sentence pairs in the samples. The final values of the three parameters were selected

on the basis of the average correlation calculated over all replications. For the relatedness sub-task, the selected values were a minimum threshold frequency of 40, 140 dimensions and a p-value of 0.20. For the entailment sub-task, they were a minimum threshold frequency of 60, 100 dimensions and a p-value of 0.25.

4 Results

4.1 Semantic Relatedness Sub-Task

The main measure of performance selected by the task organizers was the Pearson correlation, calculated on the test set (4927 sentence pairs), between the mean values of similarity according to the annotators and the values predicted by the automatic procedures.

Unsupervised Approach: MeanMaxSim. Table 1 shows the results obtained by MeanMaxSim, based on the three corpora, and by three other baselines:

- WO: The word-overlap baseline proposed by the organizers of the task, computed as the number of distinct tokens in both sentences divided by the number of distinct tokens in the longer sentence, optimizing the number of the most frequent words stripped off the sentences on the test set.
- SWL: The word-overlap baseline computed as in WO but using lemmas instead of words and the stop words list.
- ADD: The simple additive compositional model, in which each sentence is represented by the sum of the vectors of the lemmas that compose it (stripping off stop words and using the best performing corpus) and the similarity is the cosine between these two vectors (Bestgen et al., 2010; Guevara, 2011) .

MeanMaxSim	r	Baseline	r
TASA	0.696	WO	0.627
BNC	0.698	SWL	0.613
WIKI	0.696	ADD	0.500

Table 1: Pearson's correlation for MeanMaxSim and several other baselines on the test set.

MeanMaxSim produces almost identical results regardless of the corpus used. The lack of difference between the three corpora was unexpected. It could be related to the type of vocabulary used in the SICK materials, seemingly mostly frequent and concrete words whose use could be relatively similar in the three corpora. MeanMaxSim performance is clearly superior to all other baselines; among these, the additive model is the worst. This result is important because it shows that this compositional model is not, for the SICK benchmark, the most interesting baseline to assess compositional approaches. In the context of the best performance of the other teams, MeanMaxSim is (hopefully) well below the most effective procedures, which reached correlations above 0.80.

Supervised Approach. The supervised approach resulted in a correlation of 0.78044, a value well above all baselines reported above. This correlation ranked the procedure sixth out of 17, tied with another team (0.78019). The three best teams scored significantly higher, with correlations between 0.826 and 0.828.

4.2 Textual Entailment Sub-Task

Only the supervised approach was used for this sub-task. The proposed procedure achieved an accuracy of 79.998%, which ranks it sixth again, but out of 19 teams, still at a respectable distance from the best performance (84.575%).

5 Conclusion

The main contribution of this research seems to be the proposal of MeanMaxSim as baseline for evaluating CDSM. It outperforms a number of other baselines by a wide margin and is very easy to calculate. Compared to the word-overlap baseline, it has the advantage of taking into account the distributional similarity between words that are also involved in compositional models. The supervised approach proposed achieved an acceptable result (sixth out of 17) and it could easily be improved, for example by replacing standard linear regression by a procedure less sensitive to the risk of overfit due to the large number of predictors such as Partial Least Squares regression (Guevara, 2011). However, since this approach is not compositional and its efficacy (compared to others) is limited, it is not obvious that trying to improve it would be very useful.

Acknowledgements

Yves Bestgen is Research Associate of the Belgian Fund for Scientific Research (F.R.S-FNRS).

References

Aston Guy, and Burnard Lou (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baroni, Marco, and Lenci Alessandro (2010) Distributional memory: A general framework for corpus-based semantics, *Computational Linguistics*, 36, 673-721.

Baroni, Marco, Bernardi, Raffaella, and Zamparelli, Roberto (2013). Frege in space: a program for compositional distributional semantics. In Annie Zaenen, Bonnie Webber and Martha Palmer. *Linguistic Issues in Language Technologies* (LiLT), CSLI Publications.

Berry, Michael, Do, Theresa, O'Brien, Gavin, Krishna, Vijay, and Varadhan, Sowmini (1993). Svdpackc: Version 1.0 user's guide, Technical Report Number CS-93-194, University of Tennessee, Knoxville, TN.

Bestgen, Yves, and Cabiaux, Anne-Franoise (2002). L'analyse sémantique latente et l'identification des métaphores. In *Actes de la 9me Conférence annuelle sur le traitement automatique des langues naturelles* (pp. 331-337). Nancy : INRIA.

Bestgen, Yves, Degand, Liesbeth, and Spooren, Wilbert (2006). Towards automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes*, 41, 175-193.

Bestgen, Yves, Lories, Guy, and Thewissen, Jennifer (2010). Using latent semantic analysis to measure coherence in essays by foreign language learners? In Sergio Bolasco, Isabella Chiari and Luca Giuliano (Eds.), *Proceedings of 10th International Conference on Statistical Analysis of Textual Data*, 385-395. Roma: LED.

Blasius, Jorg, and Greenacre, Michael (1994). Computation of Correspondence Analysis. In Michael Greenacre and Jorg Blasius (eds.), *Correspondence Analysis in the Social Sciences*, pp. 53-75. Academic Press, London.

Cheung, Jackie, and Penn, Gerald (2012). Evaluating distributional models of semantics for syntactically invariant inference. In *Conference of the European Chapter of the Association for Computational Linguistics*, 33-43, Avignon, France.

de Marneffe, Marie-Catherine, MacCartney, Bill, and Manning, Christopher (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th Edition of the Language Resources and Evaluation Conference*. Genoa, Italy.

Deerwester, Scott, Dumais, Susan, Furnas, George, Landauer, Thomas and Harshman, Richard (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Erk, Katrin, and Pado, Sebastian (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 897-906, Honolulu, Hawaii.

Grefenstette, Edward (2013). *Category-theoretic quantitative compositional distributional models of natural language semantics*. PhD Thesis, University of Oxford, UK.

Guevara, Emiliano (2011). Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics*, 135-144, Oxford, UK.

Howell, David (2008). *Méthodes statistiques en sciences humaines*. Bruxelles, Belgique: De Boeck Université.

Kintsch, Walter (1998). *Comprehension: A Paradigme for Cognition*. New York: Cambridge University Press.

Kintsch, Walter (2001). Predication. *Cognitive Science*, 25(2), 173-202.

Landauer, Thomas, and Dumais, Susan, (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2), 211-240.

Landauer, Thomas, Foltz, Peter, and Laham, Darrell (1998). An introduction to latent semantic analysis, *Discourse Processes*, 25, 259-284.

Lebart, Ludovic, Piron Marie, et Morineau Alain (2000). *Statistique exploratoire multidimensionnelle* (3e édition), Paris: Dunod.

Marelli, Marco, Bentivogli, Luisa, Baroni, Marco, Bernardi, Raffaella, Menini, Stefano, and Zamparelli, Roberto (2014a). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of SemEval-2014: Semantic Evaluation Exercises*. Dublin, Ireland.

Marelli, Marco, Menini, Stefano, Baroni, Marco, Bentivogli, Luisa, Bernardi, Raffaella, and Zamparelli, Roberto (2014b). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

Mitchell, Jeff, and Lapata, Mirella (2010). Composition in distributional models of semantics. *Cognitive Science*, 34, 1388-1429.

Nenadic, Oleg, and Greenacre, Michael (2007). Correspondence analysis in R, with two- and three-dimensional graphics: the CA package, *Journal of Statistical Software*, 20(3), 1-13.

Schmid, Helmut (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the 1994 International Conference on New Methods in Language Processing*, 44-49, Manchester, UK.

Toutanova, Kristina, Klein, Dan, Manning, Christopher, and Singer, Yoram (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceddings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistic 2003*, 252-259, Edmonton, Canada.