CitiusNLP at SemEval-2018 Task 10: The Use of Transparent Distributional Models and Salient Contexts to Discriminate Word Attributes

Pablo Gamallo
Centro Singular de Investigación en
Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela, Galiza
pablo.gamallo@usc.es

Abstract

This article describes the unsupervised strategy submitted by the CitiusNLP team to SemEval 2018 Task 10, a task which consists of predicting whether a word is a discriminative attribute between two other words. The proposed strategy relies on the correspondence between discriminative attributes and relevant contexts of a word. More precisely, the method uses transparent distributional models to extract salient contexts of words which are identified as discriminative attributes. The system performance reaches about 70% accuracy when it is applied on the development dataset, but its accuracy goes down (63%) on the official test dataset.

1 Introduction

The goal of SemEval-2018 Task 10 (Paperno, Lenci and Krebs, To Appear) is to predict whether a word is a discriminative attribute between two other words. The key idea underlying this task is to capture semantic attributes of words in order to discriminate their senses. Distributional semantics is based on the assumption that two words have similar senses if they tend to appear with the same contextual words (Firth, 1957). As contextual words actually refer to the semantic attributes of a given word, I will focus on identifying the most salient word contexts. So, my method to identify discriminative attributes relies on the identification of salient contexts, since they represent the main semantic attributes of a word.

For this purpose, in this paper we will make use of distributional models built with transparent and lexico-syntactic contexts. To capture discriminative attributes, I will rank the most relevant contexts of a word by using lexical association measures between a given word and their contexts. My method is unsupervised and only requires pretrained distributional models.

This paper is organized as follows. The method is described in Section 2. Experiments, results, and a discussion on them are presented in Section 3. Finally, conclusions are addressed in Section 4.

2 The method

As mentioned in the previous section, discriminative attributes might be captured by searching for the most salient contexts of words. For this purpose, the distributional vector space I have adopted is a transparent count-based model with explicit and sparse dimensions. Sparseness reduction is performed by selecting the most salient contexts per word using a filtering strategy (Bordag, 2008; Gamallo and Bordag, 2011; Gamallo, 2017). The filtering strategy to select the most salient contexts consists of selecting, for each word, the S (salient) contexts with highest lexical association scores (e.g. loglikelihood, ppmi, etc). The top S contexts are considered to be the most *relevant* and informative for each word. S is a global, arbitrarily defined constant whose usual values range from 10 to 1000 (Biemann et al., 2013; Padró et al., 2014). In short, I keep at most the S most relevant contexts for each target word. This is an explicit and transparent distributional representation giving rise to a non-zero matrix. By contrast, methods based on dimensionality reduction, such as LSA (Landauer and Dumais, 1997) or neural-based embeddings (Mikolov et al., 2013), make the vector space more compact with dimensions that are not transparent in linguistic terms (Gamallo, 2017).

SemEval-2018 Task 10 to detect discriminative attributes consists of predicting whether a word is a discriminative attribute between two other words. For instance, given a triple <*car, table, wheels*>, the system must determine if the last word of the triple, *wheels*, represents a semantic

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 953–957
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

feature that characterizes the first word, *car*, but not the second one, *table*. The task is a binary classification task. For this particular example, the classifier must return a positive answer since cars have wheels but tables have not. By taking into account the objective of SemEval-2018 Task 10 and my concept of *salient context* introduced above, the classification method I propose is the following very simple rule:

> Given the triplet $< w1, w2, att >$, att is a discriminative attribute of $w1$ and not of $w2$ if att belongs to the most salient contexts of $w1$ and not to those of $w2$.

Concerning the type of context used to represent word distributions, there is a great number of previous studies that evaluate and compare syntactic contexts (usually dependencies) with bag-of-words techniques (Grefenstette, 1993; Seretan and Wehrli, 2006; Padó and Lapata, 2007; Peirsman et al., 2007; Gamallo, 2008, 2009; Levy and Goldberg, 2014; Gamallo, 2017). The cited papers state that syntax-based methods outperform bag-of-words techniques, in particular when the objective is to compute semantic similarity between functional equivalent words, such as detection of co-hyponym/hypernym word relations (i.e. near synonymy).

In my proposal, I use lexico-syntactic contexts to model word distributions. When contexts are defined as lexico-syntactic contexts, I consider that a word is an attribute of $w1$ if that word is the lexical element in at least one of the salient contexts of $w1$. For instance, consider the following three lexico-syntactic contexts:

$[NOUN, with, wheels]$
$[NOUN, nsubj, run]$
$[red, nmod, NOUN]$

If they are salient contexts of the word *car*, then the three lexical words of these three contexts, i.e. *wheels*, *run*, and *red*, will be considered as attributes of *car*.

The number of salient contexts considered per word will be determined experimentally.

3 Experiments

3.1 Resources

The count-based, explicit and transparent distributional model used in the exeperiments was generated from the English Wikipedia (August 2013 dump) containing almost 2 billion tokens. The description of this model is reported in Gamallo (2017), and a version with the 500 most salient contexts per word is freely available.[1] To process the corpus and create the transparent matrices, I used the multilingual PoS tagger of LinguaKit[2] (Garcia and Gamallo, 2015) and DepPattern, a rule-based and multilingual dependency parser (Gamallo, 2015) also taking part of LinguaKit. I also generated other models with different thresholds: from 10 to 2000 salient contexts per word.

As will be described in the next subsection, I will compare the transparent matrix with dense word embeddings, in particular with those reported in Levy and Goldberg (2014), which are publicly available.[3] These embeddings were generated from the same Wikipedia dump as the transparent model. Given that embeddings are opaque and, thereby, their dimensions are not easily associated to specific words, I use Cosine similarity to find discrimative attributes. A word is a discriminative attribute of $w1$ and not of $w2$, if the similarity score between the attribute and $w1$ is higher than a given threshold whereas it is lower in the case of $w2$.

3.2 Preliminary Experiments

To find the best configuration of the proposed system, I carried out several experiments on the train and validation datasets (20,510 examples). As the system is unsupervised, I am not required to separate training from validation. First, I searched for the best lexical association by comparing loglikelihood (Dunning, 1993) and positive pointwise mutual information (ppmi) (Niwa and Nitta, 1994), by using models with 400 and 500 salient contexts. As loglikelihood performed slightly better than ppmi, I chose the former measure to carry out the next experiments. Second, I searched for the best number of salient contexts. For this purpose, several evaluations were made with models from 10 to 2000 salient contexts. Figure 1 shows that the peak is quickly reached with 500 contexts (more than 0.67 accuracy), while performance is getting down slowly as more contexts are added.

[1] http://fegalaz.usc.es/~gamallo/resources/count-models.tar.gz

[2] https://github.com/citiususc/Linguakit

[3] https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

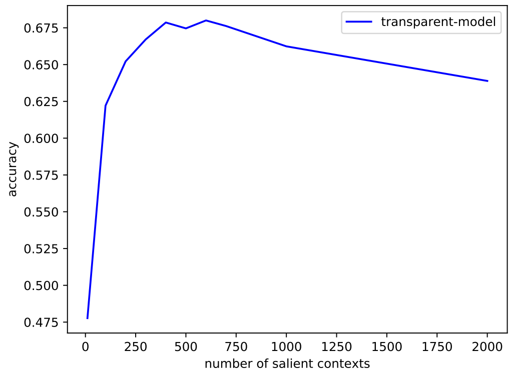

Figure 1: Accuracy of the system at different settings: from 10 to 2000 salient contexts. The experiments were carried out with the development corpus: training + validation.

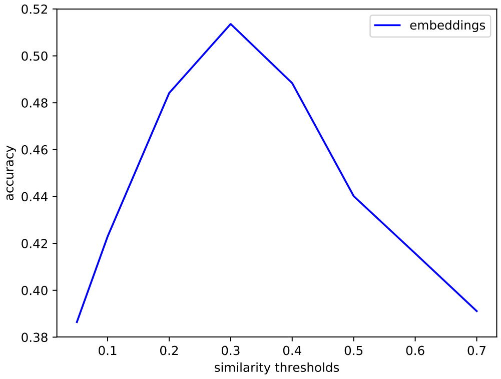

Figure 2: Accuracy of the embedding-based system at different settings: from 0.05 to 0.7 similarity values. The experiments were carried out with the development corpus: training + validation.

models	**accuracy**	**|word-cntx pairs|**
wiki	0.674	36 million
wiki+bnc	0.690	38 million
wiki+bnc+reddit	0.701	45 million

Table 1: Accuracy obtained with three corpus-based models: just Wikipedia, Wikipedia and BNC, Wikipedia, BNC and Reddit. The experiments were carried out with the development dataset: training and validation. All models were built by filtering 500 contexts per word. The last column shows the size of the filtered word-context pairs.

I therefore decided to use models with 500 salient contexts per word for the next experiments.

Next, I merged the Wikipedia-based model with other models generated from two different corpora: British National Corpus (BNC),[4] and a sample with 500 million words from Reddit corpus.[5] Results are shown in Table 1. As expected, accuracy is improved as the model grows.

Given these preliminary experiments, I submitted the two best configurations to the test evaluation (2,340 examples):

syst.	**meas.**	**saliency**	**corpora**
run1	loglike	500 ctxs	wiki+bnc
run2	loglike	500 ctxs	wiki+bnc+reddit

In my preliminary experiment, I also used the word embeddings described in the previous subsection to capture discriminative attributes. As mentioned above, a word is considered to be an attribute of a target word if their similarity is higher than a specific threshold, otherwise it is not a discriminative attribute. Several similarity scores were set to determine whether a word is an attribute or not. Figure 2 shows that the best similarity threshold is around 0.3 (cosine value). Accuracy drops dramatically with higher threshold values. The best accuracy reached by this strategy is about 20 points below the best models based on salient contexts. Therefore, for this particular task, transparent models consistently outperform word embeddings.

3.3 Official Test

The test dataset consists of 2,340 examples. My *run1* (wiki+bnc) merely reached 0.625 accuracy while *run2* (wiki+bnc+reddit) reached 0.634. These results are very far below those obtained with the development dataset, which is nevertheless 10 times larger.

3.4 Discussion

With regard to the rest of teams at the shared task, my *run2* is just in the middle of the ranking (13 out of 26 runs). However, its performance in the development dataset (0.700 accuracy) is close to the third best system. I am not able to explain the difference between the development and the test dataset. It would require a deep error analysis to understand that significant difference. This disparity is not due to a difference in the corpus frequency of the words included in the test set. I have checked the frequency of all words (test and

[4] https://corpus.byu.edu/bnc/

[5] https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/

development) and there is no important contrast at this regard. The reason could just be that the test dataset might contain more difficult triples.

The best system at the shared task achieves 0.75, leading by 12 points my *run2*. Even though the score of my system is lower, it is worth mentioning that my strategy is fully unsupervised and no tunning or specific configuration has been carried out to adapt the system to the test dataset.

4 Conclusions and Future Work

I presented a very basic unsupervised strategy to predict whether a word is a discriminative attribute between two other words. The current strategy relies on the correspondence between discriminative attribute and context saliency, and it works on transparent distributional models to extract salient contexts of words.

As I observed that accuracy improves as the corpus grows, in future work, I will compile specific text corpora for just the words of the test. This should lead to select more salient contexts (and so more discriminative attributes) per word. In addition, I will make new experiments with relational lexical resources, such as WordNet, to compare them with distributional models in this particular task.

Acknowledgments

This work has received financial support from a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, TelePares (MINECO, ref:FFI2014-51978-C2-1-R), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

Biemann, C., and Riedl M. 2013. Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Stefan Bordag. 2008. A Comparison of Co-occurrence and Similarity Measures as Simulations of Context. In *9th CICLing*, pages 52–63.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

J.R. Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32.

Pablo Gamallo. 2008. Comparing window and syntax based strategies for semantic extraction. In *PROPOR-2008*, pages 41–50. Lecture Notes in Computer Science, Springer-Verlag.

Pablo Gamallo. 2009. Comparing different properties involved in word similarity extraction. In *14th Portuguese Conference on Artificial Intelligence (EPIA'09), LNCS, Vol. 5816*, pages 634–645, Aveiro, Portugal. Springer-Verlag.

Pablo Gamallo. 2015. Dependency parsing with compression rules. In *Proceedings of the 14th International Workshop on Parsing Technology (IWPT 2015)*, pages 107–117, Bilbao, Spain. Association for Computational Linguistics.

Pablo Gamallo. 2017. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 51(3):727–743.

Pablo Gamallo and Stefan Bordag. 2011. Is singular value decomposition useful for word simalirity extraction. *Language Resources and Evaluation*, 45(2):95–119.

Marcos Garcia and Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. In *Languages, Applications and Technologies*, volume 563 of *Communications in Computer and Information Science*, pages 65–75, Switzerland. Springer. Revised Selected Papers of the Symposium on Languages, Applications and Technologies (SLATE 2015).

Gregory Grefenstette. 1993. Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches. In *Workshop on Acquisition of Lexical Knowledge from Text SIGLEX/ACL*, Columbus, OH.

T.K. Landauer and S.T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquision, induction and representation of knowledge. *Psychological Review*, 10(2):211–240.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA*, pages 302–308.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 304–309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Padó and Mirella Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014. Nothing like good old frequency: Studying context filters for distributional thesauri. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 419–424.

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2007. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *CoSMO Workshop*, pages 9–16, Roskilde, Denmark.

Violeta Seretan and Eric Wehrli. 2006. Accurate Collocation Extraction Using a Multilingual Parser. In *21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 953–960.