UWB at SemEval-2018 Task 10: Capturing Discriminative Attributes from Word Distributions

Tomáš Brychcín[1], **Tomáš Hercig**[1,2], **Josef Steinberger**[2], and **Michal Konkol**[1]

[1]NTIS – New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
[2]Department of Computer Science and Engineering,
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{brychcin,tigi,jstein,konkol}@kiv.zcu.cz
http://nlp.kiv.zcu.cz

Abstract

We present our UWB system for the task of capturing discriminative attributes at SemEval 2018. Given two words and an attribute, the system decides, whether this attribute is discriminative between the words or not. Assuming Distributional Hypothesis, i.e., a word meaning is related to the distribution across contexts, we introduce several approaches to compare word contextual information.

We experiment with state-of-the-art semantic spaces and with simple co-occurrence statistics. We show the word distribution in the corpus has potential for detecting discriminative attributes. Our system achieves F1 score 72.1% and is ranked #4 among 26 submitted systems.

1 Introduction

In this paper, we describe our UWB system participating in the pilot shared task on capturing discriminative attributes held at SemEval 2018. Given two words and an attribute, the goal of this task is to decide, whether the attribute is discriminative between them. For example, we can distinguish between the words *car* and *boat* by a discriminative feature (attribute) *wheels*. On the other hand, both *tennis* and *basketball* use a *ball*, so that the *ball* is not discriminative between them. By its nature, capturing discriminative attributes is a binary classification task. In general, there is no assumption on the input words and their attributes (e.g., part of speech, etc.).

While most related works focus on extracting discriminative features from images (Guo et al., 2015; Huang et al., 2016; Lazaridou et al., 2016), this shared task is oriented purely on textual level. The first experiments have been performed by Krebs and Paperno (2016) and have shown the promising potential of this task.

The fundamental assumption of our work is *Distributional Hypothesis*, i.e., two words are expected to be semantically similar if they occur in similar contexts (they are similarly distributed across the text). This hypothesis was formulated by Harris (1954) several decades ago. Today it is the basis of state-of-the-art distributional semantic models (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017). We present several approaches, which rely on Distributional Hypothesis and employ the word contexts for statistical comparison of their meanings.

2 Proposed Approach

Given two words $w_1 \in \boldsymbol{V}$, $w_2 \in \boldsymbol{V}$ and the attribute $a \in \boldsymbol{V}$, where $\boldsymbol{V}$ is a word vocabulary. The task is to predict, whether the attribute a is discriminative between the words w_1 and w_2, which leads to a binary classification task.

We propose several metrics, which estimate the degree to which the attribute a is important for the word w. We denote this importance as $\varphi(w, a) \in \mathbb{R}$. Clearly, if the attribute is important for one word and not for the other, it is likely to be discriminative. In general, we do not place any assumption on the importance metric $\varphi(w, a)$. We transform this score onto the binary vector $\mathbf{b}_{w,a}$ containing exactly one non-zero element (*one-hot vector*). Let $\mathcal{T} : \mathbb{R} \mapsto \{0, 1\}^b$ be the transformation function so that $\mathbf{b}_{w,a} = \mathcal{T}\big(\varphi(w, a)\big)$. In our case, we split the scores $\varphi(w, a)$ for all pairs (w, a) from training data into b bins according to $\frac{100}{b}\%$ quantiles. The bin, where the importance score belongs to, represents the value 1 in the vector $\mathbf{b}_{w,a}$.

Having the one-hot vectors $\mathbf{b}_{w_1,a}$ and $\mathbf{b}_{w_2,a}$ for the pair of words w_1 and w_2, we represent the discriminativeness of the attribute a as a conjunction matrix $\mathbf{C}_{w_1,w_2,a} = \mathbf{b}_{w_1,a}\mathbf{b}_{w_2,a}^\top$ (note $\mathbf{b}_{w_1,a}$ is a column vector). The matrix $\mathbf{C}_{w_1,w_2,a} \in \{0, 1\}^{b \times b}$

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 935–939
New Orleans, Louisiana, June 5–6, 2018.

has exactly one non-zero element at the coordinates given by the bins, onto which the scores $\varphi(w_1, a)$ and $\varphi(w_2, a)$ are mapped. Values in the matrix are used as binary features for the classifier. The main motivation behind this binarization is to allow combining different importance metrics on different scale.

In the following subsections, we introduce several approaches to estimate the importance score $\varphi(w, a)$.

2.1 Semantic Spaces

Let $\mathcal{S} : \boldsymbol{V} \mapsto \mathbb{R}^n$ be a semantic space, i.e., a function which projects word w into Euclidean space with dimension n. The meaning of the word w is represented as a real-valued vector $\mathcal{S}(w)$.

We assume, the more similar is the attribute a to the word w in meaning, the more likely a represents some feature of w. We estimate this similarity as a cosine of the angle between the corresponding vectors

$$\varphi(w, a)^{[SS]} = cos\big(\mathcal{S}(w), \mathcal{S}(a)\big). \quad (1)$$

2.2 Word Co-occurrences

We follow the intuition behind the Global Vectors (GloVe) model (Pennington et al., 2014), i.e., that the co-occurrence probabilities have the ability to encode the meaning of words.

We are given the corpus $\boldsymbol{c} = \{c_i\}_{i=1}^{k}$, i.e., a sequence of words $c_i \in \boldsymbol{V}$, where subscript i denotes the position in the corpus. Let $N(w, a)$ denote the weighted frequency of the word w in the context of the word a

$$N(w, a) = \sum_{c_i=w, c_j=a, 1 \le |i-j| \le d} \lambda(|i - j|), \quad (2)$$

where λ is a weighting function. We experiment with two types of weighting: a) *uniform* weighting, where $\lambda(m) = 1$ independently of the distance between words and b) *hyperbolic* weighting, where $\lambda(m) = \frac{1}{m}$. For uniform weighting the equation expresses the number of times the word w occurs in the context of word a. Hyperbolic weighting incorporates the assumption that closer words are more important for each other (the weight decreases with increasing distance).

Let $N(w) = \sum_{a \in \boldsymbol{V}} N(w, a)$ be the number of times any word occurs in the context of w. We estimate the conditional probability of an attribute a given the word w and use it as an importance metric

$$\varphi(w, a)^{[WC-a|w]} = P(a|w) = \frac{N(w, a)}{N(w)}. \quad (3)$$

The core idea is that if a often occurs in the context of w_1 and not in the context of w_2, then a is likely to be discriminative attribute between w_1 and w_2. The similar idea can also be expressed in an opposite way, i.e., to use probability of the word w given the attribute a

$$\varphi(w, a)^{[WC-w|a]} = P(w|a) = \frac{N(w, a)}{N(a)}. \quad (4)$$

2.3 ConceptNet

ConceptNet (Speer and Havasi, 2012) is a large semantic graph, which connects words and phrases with labeled edges. It is based on knowledge collected from many sources, including Wiktionary, WordNet, DBPedia, etc. When ConceptNet is combined with state-of-the-art semantic spaces (e.g., GloVe (Pennington et al., 2014) or SkipGram (Mikolov et al., 2013)) it provides exceptional performance in intrinsic tasks (Speer and Lowry-Duda, 2017).

In this paper, we use ConceptNet API, which enables to measure the relatedness between words[1]. It is built using an ensemble that combines data from ConceptNet, SkipGram, GloVe, and OpenSubtitles 2016, using a variation on retrofitting (Speer et al., 2016). We use the relatedness weight as an importance metric $\varphi(w, a)^{[CN]}$.

3 Experiments

In all our experiments we employ *Maximum Entropy* classifier (Berger et al., 1996) implemented in the Brainy machine learning library (Konkol, 2014). For every importance metric we use mapping onto $b = 5$ bins. This leads to $5 \times 5 = 25$ binary features describing the discriminativeness of an attribute for single importance metric.

We train the classifier on the *validation* dataset[2] proposed by the organizers of this task, containing 2722 manually annotated examples (1364 positive and 1358 negative) with total 576 distinct attributes. We do not use automatically generated data *train.txt*. For the selection of optimal feature

[1] An example of the relatedness between the words *bird* and *bat*: http://api.conceptnet.io/related/c/en/bird?filter=/c/en/bat.

[2] Available at https://github.com/dpaperno/DiscriminAtt.

set we perform 10-fold cross-validation. The official test data consists of 2340 examples (1047 positive and 1293 negative). F1 score is the official evaluation measure of this task. Note the majority class system achieves F1 score 50.1% and 55.3% on the validation and test data sets, respectively.

3.1 Settings

We estimate word co-occurrence probabilities (Section 2.2) using the English Wikipedia corpus. We experiment with several semantic spaces:

SkipGram is a neural-network based model (Mikolov et al., 2013). Levy and Goldberg (2014) provide pre-trained SkipGram models on English Wikipedia with two sizes of the context window (2 and 5) and their own model with dependency-based context.

GloVe is a log-bilinear model for word representations, which encodes global word co-occurences (Pennington et al., 2014). We use vectors provided by authors of the model, pre-trained on various corpus sizes (6, 42, and 840 billion words)[3].

FastText (Bojanowski et al., 2017) is a character-n-gram-based SkipGram model. We use word vectors pre-trained on English Wikipedia[4].

LexVec is based on factorization of positive point-wise mutual information matrix using proven strategies from GloVe, SkipGram, and methods based on singular value decomposition (Salle et al., 2016). We use pre-trained word vectors provided by the authors of the model[5].

Latent Semantic Analysis (LSA) (Landauer et al., 1998) first creates a word-document co-occurrence matrix and then reduces its dimension by singular value decomposition. We trained the model on English Wikipedia.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) represents the text as a topic distribution. In our case, each value in the word vector corresponds to the probability of this word conditioned by the particular topic.

3.2 Results

In Table 1 we show F1 scores for individual importance metrics including all three approaches, namely, semantic spaces (Section 2.1), word co-occurrences (Section 2.2), and ConceptNet (Section 2.3). The last two columns in the table contain F1 scores for 10-fold cross-validation on the validation dataset and F1 scores on the official test data. All three approaches provide comparable F1 scores on both datasets.

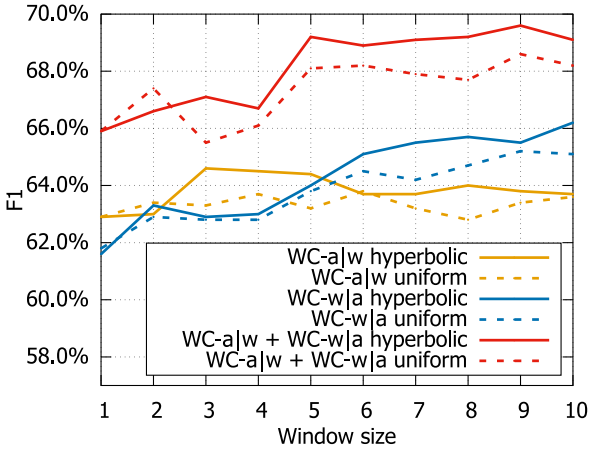

Figure 1: Results of word co-occurrence metric with different weighting and different window size d.

Detailed experiments with different context window sizes $1 \leq d \leq 10$ for estimating co-occurrence probabilities are shown in Figure 1. We show F1 scores achieved by 10-fold cross-validation on the validation dataset. The hyperbolic weighting performs better than uniform for both cases w|a and a|w independently on the size of the context window. Bigger context window seems to be more suitable for capturing discriminativeness. We can see that both metrics enrich each other and their combination leads to significantly better results than using standalone metrics. Based on this graph, we chose context window size $d = 9$ and use it in all further experiments.

Based on the cross-validation F1 scores, we combine different importance metrics to boost the performance (see Table 2). LexVec proved to perform best among semantic spaces. We found out that LDA enrich LexVec and improve the performance by approximately 1%. We believe this is because of the different context type (we used Wikipedia articles as documents for LDA). Significant improvements are achieved when we combine co-occurrence probabilities with semantics spaces or with ConceptNet (both cases give approximately 70% F1 score on both validation and test data). Combining all three approaches to-

[3] Available at https://nlp.stanford.edu/projects/glove.

[4] Available at https://fasttext.cc.

[5] Available at https://github.com/alexandres/lexvec.

	Imp. metrics	Training data	Settings	Cross-validation	Test
semantic spaces	SS-GloVe	6B, Wikipedia + Gigaword 5	$n = 300$	62.0%	62.5%
	SS-GloVe	42B, Common Crawl	$n = 300$	62.6%	62.7%
	SS-GloVe	840B, Common Crawl	$n = 300$	62.1%	62.6%
	SS-fastText	1-5B, Wikipedia	$n = 300$	60.4%	63.3%
	SS-LSA	1-5B, Wikipedia	$n = 300$	55.7%	58.3%
	SS-LDA	1-5B, Wikipedia	$n = 200$	60.5%	63.1%
	SS-LexVec	58B, Common Crawl	$n = 300$	**64.1%**	**64.9%**
	SS-LexVec	7B, Wikipedia + News Crawl	$n = 300$	59.3%	64.3%
	SS-SkipGram	1-5B, Wikipedia	$n = 300$, BoW 2	59.0%	60.6%
	SS-SkipGram	1-5B, Wikipedia	$n = 300$, BoW 5	58.0%	62.0%
	SS-SkipGram	1-5B, Wikipedia	$n = 300$, Dependencies	54.9%	56.0%
word co-oc.	WC-a|w	1-5B, Wikipedia	hyperbolic weighting, $d = 9$	63.8%	60.7%
	WC-w|a	1-5B, Wikipedia	hyperbolic weighting, $d = 9$	**65.5%**	65.5%
	WC-a|w	1-5B, Wikipedia	uniform weighting, $d = 9$	63.4%	59.9%
	WC-w|a	1-5B, Wikipedia	uniform weighting, $d = 9$	65.2%	**66.8%**
	CN			**65.1%**	**66.8%**

Table 1: Results for individual importance metrics based on semantic spaces, word co-occurrences, and ConceptNet.

Importance metric combinations	Settings	Cross-validation	Test
SS-LexVec + SS-LDA		65.6%	66.0%
WC-a|w + WC-w|a	hyperbolic weighting	69.6%	68.2%
WC-a|w + WC-w|a	uniform weighting	68.6%	67.1%
WC-a|w + WC-w|a + CN	hyperbolic weighting	70.4%	70.0%
WC-a|w + WC-w|a + SS-LexVec + SS-LDA	hyperbolic weighting	70.6%	69.8%
WC-a|w + WC-w|a + CN + SS-LexVec + SS-LDA	hyperbolic weighting	72.0%	71.3%
WC-a|w + WC-w|a + CN + SS-LexVec + SS-LDA	hyperbolic weighting, conjunction	**73.9%**	**72.1%**
Winner of SemEval 2018			75%

Table 2: Combinations of proposed importance metrics.

gether yields additional improvements (72.0% and 71.3% on validation and test data, respectively).

Our final UWB system combines all three approaches with one extra trick. We create additional binary features represented as a product of each pair of features ($x_a \times x_b$ for $a \neq b$) and add them into the classifier. We do this to better model the dependencies between single features. In the table, we denote this trick as a *conjunction*. Despite the fact that this setting leads to increasing sparseness of the feature set, it boosts F1 score on validation data by 1.9% and on test data by 0.8%.

4 Conclusion

In this paper we described our UWB system participating in SemEval 2018 shared task for capturing discriminative attributes. We explored three approaches based on word distribution in the corpus, including various semantic spaces, co-occurrence probabilities, and ConceptNet. Our best results have been achieved by Maximum Entropy classifier combining all three approaches with careful feature engineering. Our system is ranked #4 among 26 participating systems.

Acknowledgments.

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports under the program NPU I, by university specific research project SGS-2016-018 Data and Software Engineering for Advanced Applications, and by the project MediaGist, EUs FP7 People Programme (Marie Curie Actions), no. 630786.

References

Adam L. Berger, Vincent J. D. Pietra, and Stephen A. D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang. 2015. Learning predictable and discriminative attributes for visual recognition. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 3783–3789. AAAI Press.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Chen Huang, Chen C. Loy, and Xiaoou Tang. 2016. Unsupervised learning of discriminative attributes and visual representations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5175–5184.

Michal Konkol. 2014. Brainy: A machine learning library. In Leszek Rutkowski, Marcin Korytkowski, Rafa Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing*, volume 8468 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Alicia Krebs and Denis Paperno. 2016. Capturing discriminative attributes in a distributional space: Task proposal. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2016. The red one!: On learning to refer to things based on discriminative properties. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 213–218, Berlin, Germany. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alexandre Salle, Aline Villavicencio, and Marco Idiart. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–424, Berlin, Germany. Association for Computational Linguistics.

Robert Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Robert Speer and Joanna Lowry-Duda. 2017. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada. Association for Computational Linguistics.