UNBNLP at SemEval-2018 Task 10: Evaluating unsupervised approaches to capturing discriminative attributes

Milton King and **Ali Hakimi Parizi** and **Paul Cook**
Faculty of Computer Science, University of New Brunswick
Fredericton, NB E3B 5A3, Canada
milton.king@unb.ca, ahakimi@unb.ca, paul.cook@unb.ca

Abstract

In this paper we present three unsupervised models for capturing discriminative attributes based on information from word embeddings, WordNet, and sentence-level word co-occurrence frequency. We show that, of these approaches, the simple approach based on word co-occurrence performs best. We further consider supervised and unsupervised approaches to combining information from these models, but these approaches do not improve on the word co-occurrence model.

1 Introduction

In the task of capturing discriminative attributes, a system is presented with three words, and must determine whether the third word — the attribute — characterizes the first word, but not the second. For example, for the triple (*chicken*,*bread*,*legs*), *legs* is a discriminative attribute because chickens typically have legs, but bread typically does not. On the other hand, for the triple (*mother*,*woman*,*female*), *female* is not a discriminative attribute because both mothers and women are typically female. In the case of the triple (*brush*,*chocolate*,*chicken*), *chicken* is not a discriminative attribute because there is no clear relationship between chicken and brushes, or between chicken and chocolate.

In this paper we focus primarily on unsupervised approaches to the task of capturing discriminative attributes. We consider three unsupervised models drawing on information from word embeddings, WordNet (Fellbaum, 1998), and sentence-level word co-occurrence frequency. We then consider three approaches to combining information from these models: one unsupervised majority vote approach, and two supervised approaches. Somewhat surprisingly, we achieve our best F1 score of 0.61 with the remarkably simple approach based on word co-occurrence. None of the approaches to model combination improve over this.

2 Base models

In this section, we discuss three unsupervised models for identifying discriminative attributes that incorporate information from word embeddings, WordNet, and word co-occurrences. We refer to these models as "base models". In Section 3 we describe unsupervised and supervised approaches to combining these base models. Throughout the description of our models we refer to the words in the triples in the dataset as word1, word2, and attribute, respectively.

2.1 Word2vec

If an attribute is a discriminative attribute for word1, then we hypothesize that word1 and the attribute will be more semantically similar than word2 and the attribute. We use similarity of word embeddings as a proxy for semantic similarity.

We train word2vec's skip-gram model (Mikolov et al., 2013) on a snapshot of English Wikipedia from 1 September 2015 containing roughly 2.6 billion tokens, tokenized using the tokenizer available in the Stanford CoreNLP tools (Manning et al., 2014).[1] We use a window size of $\pm$ 8 and 300 dimensions. We remove all words that occur less than 15 times in the corpus. We did not set a maximum vocabulary size. We train our model using negative sampling, and set the number of training epochs to 5.

We then calculate the cosine similarity between the word embeddings for word1 and the attribute ($\cos(\text{word1}, \text{attribute})$), and word2 and the attribute ($\cos(\text{word2}, \text{attribute})$). We label the instance as a discriminative at-

[1] http://nlp.stanford.edu/software/corenlp.shtml

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 1013–1016
New Orleans, Louisiana, June 5–6, 2018.

tribute if $\cos(\text{word1}, \text{attribute})$ is greater than $\cos(\text{word2}, \text{attribute})$.

2.2 WordNet

In this approach we again hypothesize that if an attribute is a discriminative attribute for word1, then word1 and the attribute will be more similar than word2 and the attribute. Here, however, we take an approach loosely inspired by (Lesk, 1986) and (Banerjee and Pedersen, 2002), and measure similarity based on word overlap in definitions, and information available through various lexical relations, in WordNet (Fellbaum, 1998).

For each of word1, word2, and the attribute, we represent that word by a set of words that includes, for each synset for the word, all lemmas in each synset, and all words in the definition and example sentences in each synset.[2] We then optionally also include the same information — i.e., the lemmas, and the words in the definition and example sentences — for hypernyms up to level three, and meronyms. Casefolding was applied to all words in the sets of words representing word1, word2, and the attribute.

An instance is labeled as a discriminative attribute if the size of the intersection of the set of words representing word1 and the set of words representing the attribute is greater than the intersection of the set of words representing word2 and the set of words representing the attribute.

We considered various configurations of this model, differing with respect to the level of hypernyms considered, and whether meronyms were included, for word1, word2, or the attribute. The specific configurations considered, and their average F1 score on the validation data, are shown in Table 1. In subsequent experiments we only use the configuration found to perform best in Table 1.

2.3 Word co-occurrence

We hypothesize that if an attribute is a discriminative attribute for word1, then word1 and the attribute will co-occur more frequently than word2 and the attribute. Various definitions of co-occurrence could be used to operationalize this, for example, co-occurrence within a window of $\pm n$ words, a sentence, or a document. In this preliminary work we consider co-occurrence within a sentence.

[2]We tokenize the definitions and example sentences using a simple regular expression-based tokenizer, and exclude stopwords.

We calculate sentence-level co-occurrences for each pair of (word1,attribute) and (word2,attribute) in the provided shared task datasets using the ukWaC (Ferraresi et al., 2008), a corpus of roughly 1.9 billion tokens automatically constructed from a web crawl of the `.uk` domain. This model then predicts that an attribute is a discriminative attribute if the number of sentences in which word1 and the attribute co-occur is greater than the number of sentences in which word2 and the attribute co-occur. Based on its performance on the validation data (see Section 4), this model was submitted as one of our two official runs.

3 Combined models

In this section, we consider one unsupervised, and two supervised, approaches to combining the individual models discussed in Section 2.

3.1 Majority vote

In this unsupervised approach we use a majority vote of the output of the word2vec, WordNet, and word co-occurrence models. We label an attribute as a discriminative attribute if at least two of the three models predict that it is. This approach was submitted as our second official run, again based on its performance over the validation data (see Section 4), and because we are particularly interested in unsupervised approaches to this task.

3.2 Supervised: output

In this supervised approach, we represent each instance as a vector of three binary features, corresponding to the output of the word2vec, WordNet, and word co-occurrence models. We then train a logistic regression classifier on these representations of the instances. Specifically, we use the logistic regression implementation available in scikit-learn (Pedregosa et al., 2011), with l2 normalization using the liblinear solver for a maximum of 100 iterations and a stopping criteria of 0.0001.

3.3 Supervised: features

In this supervised approach we use a total of 8 features that are based on the information used by the word2vec, WordNet, and word co-occurrence models. The following features are used:

1. the cosine similarity between the word embeddings for word1 and the attribute, based

Synsets	Hypernymy level 1	Hypernymy level 2	Hypernymy level 3	Meronymy	Validation average F1
w1,w2,att	w1,w2,att	w1,w2,att	w1,w2,att	w1,w2,att	0.544
w1,w2,att	w1,w2	w1,w2	w1,w2	w1,w2	0.566
w1,w2,att	w1,w2	w1,w2	w1,w2	w1	**0.567**
w1,w2,att	w1,w2			w1,w2	0.565
w1,w2,att	w1				0.553
w1,w2,att					0.553

Table 1: F1 score on the validation data for the WordNet method. Each row corresponds to a different configuration for this model, with information for word1 (w1), word2 (w2), and the attribute (att) taken from the indicated relations in WordNet. The best F1 is indicated in boldface.

on the word2vec approach (Section 2.1);

2. the cosine similarity between the word embeddings for word2 and the attribute;
3. the size of the intersection between the set of words representing word1 and the set of words representing the attribute, as formed for the WordNet approach (Section 2.2);
4. the size of the intersection between the set of words representing word2 and the set of words representing the attribute;
5. 3−4, i.e., the difference between the previous two features;
6. the number of times word1 and the attribute co-occur, using the sentence-level approach to co-occurrence (Section 2.3);
7. the number of times word2 and the attribute co-occur;
8. 6−7, i.e., the difference between the previous two features.

Similarly to the supervised: output approach (Section 3.2), we train a logistic regression classifier (using the same settings as for that model) on these representations of the instances.

4 Results

Table 2, shows the average F1 score for each of our models on the validation and test sets. For the test set, the supervised models (supervised: output and supervised: features) were trained on the validation data, and tested on the test set; for the validation data, results for the supervised models are for 10-fold cross-validation.[3]

[3]We did not use the training data, which was not constructed in the same way as the test data, for training our supervised models. In preliminary experiments we considered models trained on the training data, and tested on the validation data, but found the performance to be relatively poor.

	Model	Average F1	
		Validation	Test
	Word2vec	0.57	0.58
	WordNet	0.57	0.56
✓	Word co-occurrence	**0.61**	**0.61**
✓	Majority vote	0.60	**0.61**
	Supervised: output	0.59	**0.61**
	Supervised: features	0.60	0.59

Table 2: Average F1 score for each of our models on the validation and test sets. Officially submitted runs are indicated with checkmarks. The highest F1 for each dataset is shown in boldface.

On the validation data, the word co-occurrence model achieved the highest F1 of the base models of 0.61, and indeed the highest F1 overall; none of the approaches to combining information from the base models (i.e., majority vote, supervised: output, or supervised: features) improved over the word co-occurrence model. The word co-occurrence model was therefore submitted as an official run. The majority vote and supervised: features models achieved the next best F1 of 0.60. Keeping with our primary interest of exploring unsupervised approaches to this task, the majority vote model was selected as our second official run.

Turning to results on the test set, the word co-occurrence, majority vote, and supervised: output models all achieved the highest F1 of 0.61. That the word-cooccurrence model outperforms the other two base models — word2vec and WordNet — shows that sentence-level word co-occurrence is more informative about discriminative attributes than the information carried by

word embeddings and the information available in WordNet, at least as it has been incorporated in these models. That none of the combined models is able to improve on the best base model suggests that, although these models are based on very different sources of information, they are not complementary.

5 Conclusions

In this paper we evaluated three unsupervised models for capturing discriminative attributes based on information from word embeddings, WordNet, and sentence-level word co-occurrence frequency. Surprisingly we found that the simple approach based on word co-occurrence performed best. We further considered supervised and unsupervised approaches to combining information from these models, but were unable to improve on the word co-occurrence model.

In future work, because of its relatively good performance, we intend to further explore the word co-occurrence model. In this work we only considered sentence-level co-occurrence. In future work we intend to consider other definitions of co-occurrence, such as co-occurrence within a window of $\pm n$ words, and document-level co-occurrence. We also only considered raw frequency in the word co-occurrence model. As an alternative to this, we also intend to consider using various lexical association measures, such as pointwise mutual information (Church and Hanks, 1990) and log-likelihood ratio (Dunning, 1993). In a similar vein, we also intend to explore the impact of the window size and number of dimensions on the word2vec model. Finally, we intend to consider other WordNet-based measures of similarity (e.g., Resnik, 1995; Jiang and Conrath, 1997).

References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 136–145, Mexico City, Mexico.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

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

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4): Can we beat Google?*, pages 47–54, Marrakech, Morocco.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, pages 19–33, Taipei, Taiwan.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, Toronto, Canada.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations, 2013*, Scottsdale, USA.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.