ADAPT at SemEval-2018 Task 9: Skip-Gram Word Embeddings for Unsupervised Hypernym Discovery in Specialised Corpora

Alfredo Maldonado
ADAPT Centre
Trinity College Dublin
Ireland

Filip Klubička
ADAPT Centre
Dublin Institute of Technology
Ireland

firstname.lastname@adaptcentre.ie

Abstract

This paper describes a simple but competitive unsupervised system for hypernym discovery. The system uses skip-gram word embeddings with negative sampling, trained on specialised corpora. Candidate hypernyms for an input word are predicted based on cosine similarity scores. Two sets of word embedding models were trained separately on two specialised corpora: a medical corpus and a music industry corpus. Our system scored highest in the medical domain among the competing unsupervised systems but performed poorly on the music industry domain. Our approach does not depend on any external data other than raw specialised corpora.

1 Introduction

The SemEval-2018 shared task on Hypernymy Discovery sought to study approaches for identifying words that hold a hypernymic relation (Camacho-Collados et al., 2018). Two words have a hypernymic relation if one of the words belongs to a taxonomical class that is more general than that of the other word. For example, the word *vehicle* belongs to a more general taxonomical class than *car* does, as *car* is a type of *vehicle*. Hypernymy can be seen as an *is-a* relationship. Hypernymy has been studied from different angles in the natural language processing literature as it is related to the human cognitive ability of generalisation.

This shared task differs from recent taxonomy evaluation tasks (Bordea et al., 2015, 2016) by concentrating on Hypernym Discovery: the task of predicting (discovering) n hypernym candidates for a given input word, within the vocabulary of a specific domain (Espinosa-Anke et al., 2016). This shared task provided a general language domain vocabulary and two specialised domain vocabularies in English: medical and music industry. For each vocabulary, a reference corpus was also supplied. In addition to these English vocabularies, general language domain vocabularies for Spanish and Italian were also provided. The ADAPT team focused on the two specialised domain English subtasks by developing an unsupervised system that builds word embeddings from the supplied reference corpora for these domains.

Word embeddings trained on large corpora have been shown to capture semantic relations between words (Mikolov et al., 2013a,b), including hypernym-hyponym relations. The word embeddings built and used by the system presented here exploit this property. Although these word embeddings do not distinguish one semantic relation from another, we expect that true hypernyms will constitute a significant proportion of the predicted candidate hypernyms. Indeed, we show that for the medical domain subtask, our system beats the other unsupervised systems, although it still ranks behind the supervised systems.

Even though unsupervised systems tend to rank behind supervised systems in NLP tasks in general, our motivation to focus on an unsupervised approach is derived from the fact that they do not require explicit hand-annotated data, and from the expectation that they are able to generalise more easily to unseen hypernym-hyponym pairs.

The rest of this system description paper is organised as follows: Section 2 briefly surveys the relevant literature and explains the reasons for choosing to use a particular flavour of word embeddings. Section 3 describes the components of the system and its settings. Section 4 summarises the results and offers some insights behind the numbers. Section 5 concludes and proposes avenues for future work.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 924–927
New Orleans, Louisiana, June 5–6, 2018.

2 Related Work

Modern neural methods for natural language processing (NLP) use pre-trained word embeddings as fixed-sized vector representations of lexical units in running text as input data (Goldberg, 2017, ch. 10). However, as mentioned previously, word embedding vectors can be used on their own to measure semantic relations between words in an unsupervised manner by, for example, taking the cosine similarity of two word embedding vectors for which semantic similarity is to be measured.

There are several competing approaches for producing word embedding vectors. One such approach is skip-gram with negative sampling (SGNS), introduced by Mikolov et al. (2013a,b) as part of their Word2Vec software package. The skip-gram approach assumes that a focus word occurring in text depends on its context words (the words the focus word co-occurs with inside a fixed-sized window), but that those context words occur independently of each other. This conditional independence assumption in the context words makes computation more efficient and produces vectors that work well in practice. The negative sampling portion of the algorithm is a way of producing "negative" context words for the focus word by simply drawing random words from the corpus. These random words are assumed to be "bad" context words for the focus word. The positive and negative examples are used by an objective function that seeks to maximise the probability that the positive examples came from the corpus whilst the negative examples did not.

Cosine measures on word embeddings pairs (or even on other distributional lexical semantic representations) give an indication of the overall *semantic relatedness* of the word pairs they represent (Turney and Pantel, 2010), without specifying the type(s) of semantic relation(s) the two words hold. There have been endeavours to train word embeddings that emphasise one semantic relation over another. For example, Nguyen et al. (2016) modified the skip-gram objective function to train word embeddings that distinguished synonymy from antonymy. In a similar vein, Nguyen et al. (2017) developed an algorithm called Hypervec by adapting the skip-gram objective function to emphasise the non-symmetric hypernym-hyponym relations.

Our team indeed implemented a variant of the Hypervec method but failed to obtain better performance scores on the training set than those obtained by using traditional SGNS (see Section 4). Whilst it is possible that a software bug in our implementation could be the cause of this lower performance, we decided to submit the SGNS results to the official shared task due to time constraints.

3 System Description

Our system consists of two components: a **trainer** that learns word vectors using an implementation of the Skip-Gram with Negative Sampling algorithm, and a **predictor** that outputs (predicts) the top 10 hypernyms of an input word based on the trained vectors. These two components and their settings are described here.

Trainer The trainer is a modification of PyTorch SGNS[1], a freely available implementation of the Skip-Gram with Negative Sampling algorithm. One set of vectors per specialised corpus (medicine and music industry) were trained on a vocabulary that consists of the 100,000 most frequent words in each corpus, using a word window of 5 words to the left and 5 words to the right of a sliding focus word. The windows do not cross sentence boundaries. For negative sampling, 20 words were randomly selected from the vocabulary based on their frequency[2]. All vectors had a dimensionality of 300.

Predictor For each input word in the test file, the predictor attempts to produce 10 candidate hypernyms based on the vectors it learned during training. If there is no vector for an input word, no output for that word is given. If the input word is a multiword expression, then the learned vectors for the individual component words are retrieved and averaged together. This averaged vector is interpreted to represent the input multiword expression. After a vector is retrieved (or computed, in the case of averaged multiword expressions), pairwise cosine similarities are taken between this vector and all other vectors (i.e. the vectors corresponding to the other 99,999 most frequent words). The words represented by the 10 highest ranking cosine similarities are output as the 10 candidate hypernyms for the input word or multiword expression.

[1] https://github.com/theeluwin/pytorch-sgns

[2] The frequencies were smoothed by raising them to the power of 0.75 before dividing by the total.

Domain	Approach	MAP	MRR	P@1	P@3	P@5	P@15
medical	SGNS	8.13	20.56	13.20	10.80	8.32	6.33
medical	HV	4.40	13.05	10.60	5.60	4.27	3.10
music	SGNS	1.88	5.34	4.00	2.40	1.89	1.35
music	HV	1.79	5.39	5.00	2.07	1.62	1.28

Table 1: Automatic evaluation results for the submitted system (SGNS) and a Hypervec variant (HV).

As can be seen, our system is completely unsupervised as it does not require corpora with tagged examples of words holding hypernym-hyponym relations or any external linguistic or taxonomical resources.

4 Results

Table 1 shows the results for our SGNS-based approach, which was submitted to the official shared task (SGNS), and for our Hypervec variant (HV), which was not submitted.

Our official submission ranked at eleven out of eighteen on the medical domain subtask with a Mean Average Precision (MAP) of 8.13. However, it ranked first place among all the unsupervised systems on this subtask. On the music industry domain subtask, our system ranked 13th out of 16 places with a MAP of 1.88, ranking 4th among the unsupervised systems. We believe that one reason why the music industry scores are so much lower than the medical results is due to our system not producing an output for 233 of the music industry input words (45% of the total), compared to the 128 medical input words (26%) it failed to predict.

Another aspect that seems to work against our system is its simplistic way of handling multiword expressions, namely by averaging together the individual word's vectors. The total number of multiword expressions in the medical test set is 264, slightly higher than in the music test set, which contains 220 multiword expressions. Similarly, our system does not have a way of predicting multiword expressions as hypernym candidates, as it can only output the unigrams for which it has vector representations. 82% of the medical domain input words have at least one hypernym that is a multi-word expression, whilst 92% of the music industry domain input words have multi-word expression hypernyms.

5 Conclusions and Future Work

We presented a simple but competitive unsupervised system to predict hypernym candidates for input words, based on cosine similarity scores of word embedding vectors trained on specialised corpora.

Unsupervised systems in general tend to have lower performance than supervised systems as they lack explicit information to train on. So we are encouraged that our system beat other unsupervised systems on one corpus, as this gives us more avenues to explore.

One such avenue is to revisit our Hypervec implementation. We suspect that it might require more training epochs than the traditional SGNS method in order to achieve reasonable results. We also seek to experiment with refining pre-trained SGNS word embeddings with Hypervec, rather than training word embeddings from scratch using Hypervec directly.

Another avenue to explore involves incorporating taxonomical information into our word embeddings. One way to achieve this is by retrofitting pre-trained SGNS word embeddings with information derived from existing taxonomies like WordNet (Faruqui et al., 2015). Another way of incorporating taxonomical information is by generating a pseudo-corpus via a random walk over such a taxonomy and then learn SGNS word embeddings in the usual way (Goikoetxea et al., 2015).

These approaches (Hypervec, retrofitting and taxonomy random-walk) however, would relax the unsupervised constraint we followed in our implementation. So yet another avenue to explore is to instead apply different similarity functions that might be more sensitive to the one-way, general-specific nature of hypernymic relationships between words.

Acknowledgements

We thank our anonymous reviewers for their input. The ADAPT Centre for Digital Content Tech-

nology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA. Association for Computational Linguistics.

Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. Supervised Distributional Hypernym Discovery via Domain Adaptation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 424–435, Austin, TX.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1606–1615.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random Walks and Neural Network Language Models on Knowledge Bases. In *Human Language Technologies: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1434–1439, Denver, CO.

Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, Scottsdale, AZ.

Tomas Mikolov, Ilya Stutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS) In Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe, NV.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 454–459, Berlin.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.