Unsupervised Hypernymy Directionality Prediction Using Context Terms

Anonymous EACL submission

Abstract

Hypernymy directionality prediction is an important task in Natural Language Processing due to its significant usages in natural language understanding and generation. Many supervised and unsupervised methods have been proposed for this task, but existing unsupervised methods do not leverage distributional pre-trained vectors from neural language models, as supervised methods typically do. In this paper, we present a simple yet effective unsupervised method for hypernymy directionality prediction that exploits neural pre-trained word vectors in context, based on the distributional informativeness hypothesis. Extensive experiments on seven datasets demonstrate that our method outperforms or achieves comparable performance to existing unsupervised and supervised methods.

1 Introduction

Hypernymy, an Is-A relation, has garnered significant attention in the field of Natural Language Processing (NLP). It constitutes a transitive and asymmetric semantic link between a hypernym (also referred to as a superordination or a superset) and a hyponym (also referred to as a subordination or a subset) (Lyons, 1977). For instance, *mammal* is a hypernym of *elephant*, and *fruit* is a hypernym of *banana*. This hypernymy semantic relation plays a crucial role in various challenging NLP tasks, such as knowledge base construction (Snow et al., 2006; Navigli et al., 2011), natural language inference (Dagan et al., 2015; Williams et al., 2018), textual entailment (Dagan et al., 2015), question answering (Huang et al., 2008), text classification (Jang et al., 2021), and text generation (Biran and McKeown, 2013).

Hypernym detection is generally a two-step process: identifying hypernymy relations and predicting the directionality of those relations. Hypernymy detection distinguishes hypernymy from other semantic relations, such as synonymy and antonymy. Directionality prediction, on the other hand, identifies which word in a given hypernymy pair is the hypernym and which word is the hyponym. For example, given the pair “animal” and “cat”, directionality prediction would determine whether “animal” is a hypernym of “cat” or vice versa. In this paper, our focus is on the problem of directionality prediction—determine whether A is a hypernym of B or B is a hypernym of A.

For hypernymy directionality prediction, there exist a wealth of unsupervised methods (Weeds and Weir, 2003; Clarke, 2009; Kotlerman et al., 2010; Lenci and Benotto, 2012; Santus et al., 2014). Many of these metrics are based on the distributional inclusion hypothesis (Weeds et al., 2004; Kotlerman et al., 2010) and the distributional informativeness hypothesis (Santus et al., 2014). However, these existing methods, which were developed some time ago, do not take advantage of the recent pre-trained distributional vectors from neural language models, such as BERT (Devlin et al., 2018) and fastText (Bojanowski et al., 2017). Additionally, most methods typically require a validation set to tune the threshold for their metrics in order to accurately identify the directionality.

In this paper, we propose a simple yet effective unsupervised metric, DECIDE[1], for hypernymy directionality prediction using pre-trained neural word embedding. In our experiments involving 7 datasets, DECIDE shows superior or comparable performance to existing unsupervised metrics. We also compare our metric with state-of-the-art supervised methods, showing superiority in handling previously unseen data samples. We show that existing supervised methods report optimistic performance due to information overlap between the train and test partitions of a datasets.

[1] DECIDE is an anagram of the bold letters from **C**entroid **Di**stance in **D**istributional Cont**E**xt.

2 Related Works

Several unsupervised directional measures have been proposed to tackle hypernymy prediction, especially in the early stages of research. Weeds et al. (2004) introduced the notion of distributional generality, highlighting that more general words tend to manifest across a broader spectrum of contexts compared to specific ones. Their research relied on the assumption that the contexts of a hyponym are expected to be included in those of its hypernym, known as the distributional inclusion hypothesis. Building upon this, Clarke (2009) employed a partially ordered vector space to formalize distributional generality, while Lenci and Benotto (2012) extended the notion further by proposing that more general terms should exhibit high recall and low precision. Santus et al. (2014) introduced an entropy-based measure, SLQS, considering that hypernyms' typical linguistic contexts might be less informative than those of hyponyms, known as the distributional informativeness hypothesis. They proposed a measure based on the intersection of mutually dependent contexts of target words.

With the ascent of deep learning models, supervised strategies have emerged to adapt word embeddings through joint optimization models during pretraining or retrofitting models during fine-tuning. The former approaches reshaped the entire embedding space e.g., (Levine et al., 2020), which can be computationally expensive. In contrast, the latter methods (Yu et al., 2015; Luu et al., 2016; Vendrov et al., 2016) fine-tuned word vectors to align with external linguistic constraints. While these methods are applicable to any pre-trained distributional space, they only modify the vectors of words seen in constraints, leaving unseen word vectors unmodified. Glavaš and Vulic (2019) attempted to address this issue by building a model, named GLEN, which learns a function during training that can be used for unseen word pairs. All of these use lexical resources like WordNet to (weakly) supervise the models.

Similar to early unsupervised measures, we introduce an unsupervised directionality measure, named DECIDE, which is based on the idea of distributional generality, specifically the distributional informativeness hypothesis. However, DECIDE is differentiated from previous work in that it takes advantage of neural word embeddings for context words, and does not require setting a threshold to decide directionality.

3 Our Proposed Method: DECIDE

In this section, we present our measure for identifying the hypernymy directionality between a given hypernymy pair. Our measure operationalizes the distributional informativeness hypothesis (Santus et al., 2014), which states that more general terms tend to occur in more general and diverse contexts than specific terms. For example, the words that occur around "animal" can come from generic animal characteristics, and their habitats, whereas context words of "cat" are more specific to cats.

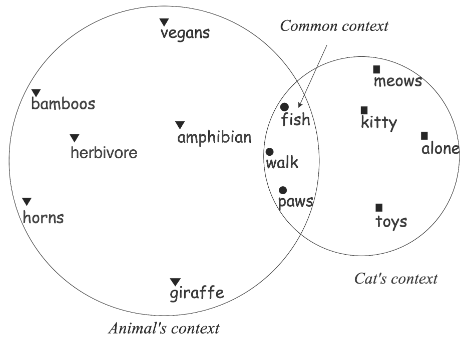

Figure 1: 2D visualization of context word embedding of a Hypernym (Animal) and Hyponym (Cat).

Based on the distributional informativeness hypothesis, we hypothesize that the context words of a hypernym would have a broader distribution compared to its hyponym's context words in terms of their meanings. To obtain the context words of given two terms $term_1$ and $term_2$ in a hypernymy relation, we first collect all sentences that contain each term from a large corpus. Subsequently, we tokenize these sentences using white spaces and punctuation, and remove stop words and tokens solely composed of numbers or symbols, retaining the remaining words as context words. For instance, in Figure 1 the two circles represents the context words of two terms *Animal* and *Cat*. Using these context words, we then identify the common context words (intersecting region of the two circles in Figure 1). Then, we calculate the mean vector of those common context words, m. From the unique context words for $term_1$ (e.g., triangles in Figure 1) and $term_2$ (e.g., rectangles in Figure 1), we determine the minimum number of unique context words, n, and then select the n farthest unique context words for $term_1$ and $term_2$, C_1' and C_2' , respectively. Finally, we compare the average distance between C_1' and C_2' from m. This process is expressed in Figure 2.

DECIDE(C_1, C_2)
Input:
C_1 = context words unique to $term_1$
C_2 = context words unique to $term_2$

$n = min(|C_1|, |C_2|)$
C_1' = n farthest context words from C_1
C_2' = n farthest context words from C_2

m = the average embedding of the common context words

if $\frac{1}{n}\sum_{c \in C_1'}(c - m) > \frac{1}{n}\sum_{c \in C_2'}(c - m)$:

return: $term_1$ is a *hypernym* of $term_2$
else:
return: $term_2$ is a *hypernym* of $term_1$

Figure 2: Synopsis of DECIDE for determining hypernym direction.

Dataset	Original-Pairs	Atmost 1-Entity Unseen	Both-Entities Unseen
Bless	1,277	241	0
Weeds	1,321	175	3
EVAluation	3,035	1,799	18
LenciBenotto	1,724	1,524	224
Medical	3,256	3,185	1,545
Music	5,455	5,115	1,802
Comp.Sci	331	331	247

Table 1: The number of hypernym-hyponym pairs in each data set. Second colum shows the number of original entity-pairs. Third column shows the number of entity-pairs where at least one entity of the pair is not present in the training data. The fourth column shows the number of entity-pairs where both the entities are not present in the training data.

4 Experiments

We evaluate our approach on seven real-life datasets in four domains: general, medicine, music, and computer science. The datasets contain hypernym-hyponym pairs (u, v) with corresponding labels indicating the direction. The dataset statistics is shown in Table 1. To represent u and v, we use fastText (Bojanowski et al., 2017), pre-trained distributed vectors ($d = 300$) trained on Wikipedia.[2]

Hypernymy datasets: The datasets from the general domain are **Bless** (Baroni and Lenci, 2010), **Weeds** (Weeds et al., 2014), **EVAlution** (Santus et al., 2015) and **LenciBennotto** (Benotto, 2015). The **medicine** and **music** datasets are from the SemEval-2018 Task9 Hypernym Discovery (Camacho-Collados et al., 2018). We use the test sets partitioned by Shwartz et al. (2016) for Weeds, EVAluation and LenciBennotto.

We also construct a dataset in the Computer Science domain, with the hypernymy broadly defined covering concept-subconcept or topic-subtopic relations. For instance, a hypernym-hyponym pair in this dataset can be ("data structure", "binary search tree"). We use GPT-3 (Brown et al., 2020) to build this dataset using the OpenAI API's create completion functionality. We tailored the prompt to generate a list of 20 subtopic names for a given topic, beginning with "Computer Science" as the initial topic and then using the resulting 20 subtopics as subsequent prompts. The numbers of hypernymy pairs in all the datasets are shown in Table 1.

Context corpus: To obtain context words in the general domain, we use the wiki dump corpus (Goldhahn et al., 2012). For the medicine domain, we use a corpus provided by Camacho-Collados et al. (2018), a 130M-word subset extracted from the PubMed corpus of biomedical literature from MEDLINE. For the music domain, a 100M-word corpus is provided with the original dataset, which includes Amazon reviews, music biographies and Wikipedia pages about music theory and genres (Oramas et al., 2016). Furthermore, for the computer science education domain, we create a corpus by extracting the Wikipedia pages of all the topics and subtopics in our dataset.

4.1 Comparison with Unsupervised Methods

We first compare out methods with existing unsupervised methods: **SLQS** (Santus et al., 2014), **invCL** (Lenci and Benotto, 2012), **ClarkDE** (Clarke, 2009), **cosWeeds** (Lenci and Benotto, 2012), and **weedsPrec** (Weeds et al., 2004). Note that, cosWeeds, ClarkDE, and invCL has a value between 0 and 1; the higher the value, the more likely the directionality holds for the given order. Thus, these metrics need a threshold to decide on the hypernym direction. We choose a threshold of 0.5 for all these 3 methods. SLQS and WeedsPrec do not need a threshold value.

[2] In our preliminary experiments, we also explored the use of Glove (Pennington et al., 2014) and BERT (Devlin et al., 2018) embeddings, and observed that they yielded similar results.

Data	Unsupervised SLQS	invCL	ClarkDE	cosWeeds	WeedsPrec	Supervised GLEN-before	GLEN-after	DECIDE
Bless	0.54	0.51	**0.59**	0.51	0.51	0.89	N/A	0.50
Weeds	0.62	0.53	0.59	0.55	0.43	0.67	**0.66**	0.65
EVALuation	0.63	0.50	0.60	0.50	0.44	0.72	**0.66**	0.63
LenciBenotto	0.62	0.53	0.65	0.56	0.31	0.67	0.60	**0.70**
Medical	0.73	0.60	0.72	0.60	0.26	0.77	0.70	**0.77**
Music	0.64	0.54	**0.66**	0.56	0.34	0.67	0.58	0.65
Comp.Sci	0.82	0.56	0.62	0.60	0.20	0.50	0.53	**0.85**

Table 2: Performance of our measure, DECIDE on hypernymy directionality classification compared to existing unsupervised measures (Accuracy). Note that GLEN-before is included in the table for comparison with GLEN-after to illustrate the memorization problem.

The results in Table 2 show that our measure, DECIDE, outperforms most measures. Over the seven datasets, DECIDE ranks first in five and second in one. DECIDE performs particularly well on domain datasets such as Medical and Com. Sci with 0.77 and 0.85 accuracy, respectively. This is likely because high-quality context words can be obtained for domain datasets. On the general dataset, such as Bless, DECIDE's performance is not as good (0.50 accuracy), but this is also true for the competing methods, as all of them perform relatively poorly on this dataset (accuracy values between 0.51 and 0.59). The second best unsupervised method in our experiment is ClarkDE, which has the best performance on two datasets, Bless (0.59 accuracy) and Music (0.66 accuracy).

4.2 Comparison with Supervised Methods

To compare our unsupervised method with supervised models, we consider GLEN (Glavaš and Vulic, 2019), as this model is conceptually guaranteed to work on unseen pairs. GLEN's inpue is the fastText embedding of the hypernym and the fastText embedding of hyponym. We discard many supervised methods, such as order embedding (Vendrov et al., 2015) and LEAR (Rei et al., 2018), which produce tuned embedding vectors of seen hypernym pairs only and are therefore unable to produce prediction on unseen pairs. We train the GLEN model using the same training setup reported in the original paper and test it on two versions of each of the seven datasets: The first version uses the test data where at most one term of the entity pair may be present in the training data (shown in the third column of Table 1). The second version uses the test data where no terms of the entity pair are present in the training data (shown in the fourth column in Table 1). The results are shown under "GLEN-before" and "GLEN-after" columns in Table 2, respectively.

Table 2 shows the results. As can be seen, DECIDE outperforms GLEN-after on five datasets, while GLEN-after outperforms DECIDE on two datasets by a narrow margin (0.66 vs 0.65, and 0.66 vs 0.63). Note that there are no results for GLEN-after on the Bless dataset, as the number of instances of this dataset is zero after overlap removal. When we compare DECIDE with GLEN-before, for which either the hypernym or hyponym entities (but not both) from the test data may present in the training data, GLEN's performance improves substantially. In fact, in this case GLEN outperforms DECIDE on four out of seven datasets. This validates that supervised methods, like GLEN, can boost their performance through information overlap between the training and test data, even if only one element of the hypernym pair is in the training data. This phenomenon was also reported by (Levy et al., 2015), who showed that supervised methods for this task suffer from the memorization problem, in which the model memorizes prototypical hypernyms ("general words"), thereby failing to generalize for word pairs where those prototypical hypernyms are not part of the training data.

5 Conclusion

Our contributions are three folds: First, we introduced a new measure, DECIDE, for hypernymy directionality prediction that does not require setting a threshold. DECIDE can be worked with any neural pre-trained distributional space. Second, our extensive experiments showed that DECIDE outperforms or is on par with existing unsupervised and supervised methods on previously unseen samples, demonstrating its effectiveness. Lastly, we also showed that existing supervised methods do not generalize well on unseen samples, corroborating the previously reported claim of the memorization problem by Levy et al. (2015). Our code and dataset will be available at GitHub: http://anonymous.

6 Limitations

The proposed measure, DECIDE, may exhibit sensitivity to the choice of corpus used to retrieve context words, similar to other context-based measures, e.g., (Clarke, 2009; Lenci and Benotto, 2012; Santus et al., 2014). For example, a corpus of Wikipedia articles may yield different results from a corpus of scientific papers. Further investigations into the nature of context and how it affects hypernymy directionality would be beneficial, as well as studies on how to obtain the typical context of a term.

In addition, our method does not incorporate the frequency of context words while remarkably, it outperforms other measures even without considering frequencies. However, frequency could also play an important role in hypernymy directionality, as shown in previous work, e.g., (Clarke, 2009; Lenci and Benotto, 2012; Santus et al., 2014). Therefore, combining our current distributional space distances with frequency information could lead to further improvements. We leave this exploration for future work.

7 Ethical Consideration

As with any measures, inaccuracies in the predictions made by our proposed measure could potentially result in unintended and erroneous outcomes in applications. For example, if the measure is used to predict the hypernymy directionality between two terms in a medical context, a wrong prediction could lead to a misdiagnosis or incorrect treatment. It is important to use our measure responsibly and to be aware of its limitations. It is also important to validate the predictions of the measure against other sources of information before using them in any critical applications.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Giulia Benotto. 2015. Distributional models for semantic relations: A sudy on hyponymy and antonymy.

Or Biran and Kathleen McKeown. 2013. Classifying taxonomic relations between pairs of wikipedia articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 788–794.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*, pages 712–724.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 112–119.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2015. Recognizing textual entailment: models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Goran Glavaš and Ivan Vulic. 2019. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the 2008 Conference on empirical methods in natural language processing*, pages 927–936.

Hyeju Jang, Seojin Bang, Wen Xiao, Giuseppe Carenini, Raymond Ng, and Young ji Lee. 2021. Kw-attn: knowledge infused attention for accurate and interpretable text classification. In *Proceedings of deep learning inside out (DeeLIO): the 2nd workshop on knowledge extraction and integration for deep learning architectures*, pages 96–107.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 75–79.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.

Anh Tuan Luu, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 403–413.

John Lyons. 1977. *Semantics: Volume 2*, volume 2. Cambridge university press.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, volume 11, pages 1872–1877.

Sergio Oramas, Luis Espinosa Anke, Mohamed Sordo, Horacio Saggion, and Xavier Serra. 2016. Elmd: An automatically generated entity linking gold standard dataset in the music domain. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3312–3317.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring lexical entailment with a supervised directional similarity network. *arXiv preprint arXiv:1805.09355*.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 38–42.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2016. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. *arXiv preprint arXiv:1612.04460*.

Rion Snow, Dan Jurafsky, and Andrew Y Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 801–808.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. *International Conference on Learning Representation*.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *COLING 2004: Proceedings of the 20th international conference on Computational Linguistics*, pages 1015–1021.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.