Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 10–16
Marseille, 24 June 2022

Discovering Financial Hypernyms by Prompting Masked Language Models

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, Chu-Ren Huang
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
The Hong Kong Polytechnic University, Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong
{peng-bo.peng,emmanuele.chersoni,yu-yin.hsu,churen.huang}@polyu.edu.hk

Abstract

With the rising popularity of Transformer-based language models, several studies have tried to exploit their masked language modeling capabilities to automatically extract relational linguistic knowledge, although this kind of research has rarely investigated semantic relations in specialized domains. The present study aims at testing a general-domain and a domain-adapted Transformer model on two datasets of financial term-hypernym pairs using the prompt methodology. Our results show that the differences of prompts impact critically on models' performance, and that domain adaptation to financial texts generally improves the capacity of the models to associate the target terms with the right hypernyms, although the more successful models are those which retain a general-domain vocabulary.

Keywords: Transformers, Semantic Relations, Language Modeling, Financial Natural Language Processing

1. Introduction

Since their introduction, Transformer architectures (Vaswani et al., 2017; Devlin et al., 2019) have quickly become the dominant paradigm in modern Natural Language Processing (NLP). On the one hand, their capacity for generating contextualized representations of words in context has led to performance improvements in several supervised tasks. On the other hand, the *masked language modeling* abilities of models like BERT attracted the attention of linguists and NLP scientists to propose experiments with *natural language prompts* to probe the semantic and pragmatic knowledge in the internal representations of the networks (Ettinger, 2020; Ravichander et al., 2020; Pandia et al., 2021; Hanna and Mareček, 2021). Roughly speaking, a prompt is a natural language sentence in which a token has been masked, such that a language model has to predict the hidden token and reconstruct the original sentence. The assumption of the literature on probing language models is that, given a prompt, "filling the gap" requires some specific linguistic knowledge. For example, with a prompt like *A robin is a type of* [MASK], a language model will be able to assign the highest probability to *bird* for that given prompt only if it possesses some knowledge of lexical-semantic relations (and more specifically, the hyponymy-hypernymy relation existing between *robin* and *bird*).

As NLP technologies are frequently used in accounting and finance, detecting hypernymy and other semantic relations can substantially improve results in financial tasks, such as numeral understanding and records management. Hypernyms correspond to higher-level categories for target concepts, and thus they play an important role in the organization of the terminology of specialized domains (Espinosa-Anke et al., 2016).

Despite the popularity of prompt-based methods in NLP (Liu et al., 2021), there are still open questions about their usage in specialized domains: *Can they retrieve lexical-semantic relations in a specialized domain? What is the impact of domain adaptation on relation discovery? And how does the choice of different linguistic prompts affect the models' performance?*

To answer these questions, in our paper, we focus on the specific problem of *hypernymy discovery in the financial domain*. We use the data from two benchmarks in recent FinSim shared tasks (El Maarouf et al., 2021; Mansar et al., 2021). We treat the problem as an unsupervised task and test three different Transformer models (a general domain model and two domain-adapted ones) by using 5 types of prompts, and we report their results in identifying the right hypernym for the financial terms in the datasets. We found that domain adaptation tends to improve the retrieval of the right hypernym. Surprisingly, however, we found that a general-domain vocabulary leads to better retrieval performance than a finance-specific one.

2. Related Work

Lexical-semantic relations such as hypernymy have been investigated in computational linguistics for a long time, especially in the Distributional Semantics community (Weeds and Weir, 2003; Lenci and Benotto, 2012; Weeds et al., 2014; Roller et al., 2014; Santus et al., 2014a; Santus et al., 2014b; Santus et al., 2015; Santus et al., 2016; Chersoni et al., 2016; Roller and Erk, 2016; Shwartz et al., 2017; Liu et al., 2019; Xiang et al., 2020). Hypernymys have received special attention in the literature, since they correspond to higher-level categories of concepts and represent the backbone of ontologies and lexical networks (Chersoni and Huang, 2021). A research trend based on pattern-based methods use external corpora to exploit the co-occurrence of a hyponym and its hypernyms in specific linguistic patterns (e.g., *is a type of*) (Boella and

Di Caro, 2013; Flati et al., 2016; Camacho-Collados and Navigli, 2017). Machine learning models relying on distributional representations as input features have also been trained for prediction and detection of hypernymy relations (Shwartz et al., 2016; Sanchez and Riedel, 2017; Nguyen et al., 2017).

After the introduction of Transformers in NLP, several researchers tried to take advantage of their abilities of *masked language modeling* to analyze to what extent they are able to associate nouns with their hypernyms. A simple methodology consists of feeding the masked language model with a sentence of the form "The TERM is a HYPERNYM." then masking the hypernym token and letting the model fill the blank spot. Although the results were not always consistent, previous work showed that the Transformers can perform the hypernymy discovery task well, especially when the right hypernymy has to be picked from a close set of candidates (Ettinger, 2020; Ravichander et al., 2020). Moreover, Chersoni and Huang (2021) recently reported a positive effect of Transformer-based features in supervised hypernymy detection for the financial domain. The target term was masked in a manually constructed probe sentence, and a pre-trained Transformer-based language model was asked to assign probability scores to the candidate hypernyms of the target terms.

In the last few years many domain-adapted versions of BERT and other Transformer architectures have been made available by NLP researchers (Araci, 2019; Yang et al., 2020; Liu et al., 2020). However, to the best of our knowledge, the impact of domain adaptation on the systems' capacity for retrieving lexical-semantic relations has not yet been explored. In theory, a Transformer that has been adapted to a specific domain should have access to a more specific lexical-semantic knowledge for the words of that domain, and therefore one would expect it to perform better in term categorization tasks.

In the present work, we compared a general domain BERT (Devlin et al., 2019) and two domain-adapted FinBERT models (Yang et al., 2020) on two datasets for financial hypernymy detection that have been used for the recent FinSim shared tasks (Keswani et al., 2020; El Maarouf et al., 2021; Mansar et al., 2021). We adopted the masked language modeling approach, feeding the model with 5 types of natural language prompts (Hanna and Mareček, 2021), and we analyzed the capacity of the systems to associate the terms in the datasets with the correct hypernym labels.

3. Experimental Settings

3.1. Datasets

For our study, we used the datasets from the FinSim (El Maarouf et al., 2021) and the FinSim-2 shared task (Mansar et al., 2021). The FinSim dataset is composed of a training set and a test set of, respectively, 100 and 99 financial terms and their corresponding hypernyms, which a system has to identify out of 8 possible alternatives (**Bonds, Forward, Funds, Future, MMIs, Option, Stocks, Swap**). The FinSim-2 has a training set of 614 terms and a test set of 212 terms, and 10 possible hypernym labels (the same as FinSim, with the addition of **Credit Index**, and **Equity Index**). Examples of instances from FinSim-2 are shown in Table 1. In both datasets, the hypernyms correspond to the high-level classes of the Financial Business Ontology (FIBO) [1]. Since the gold labels of the FinSim-2 test set are not publicly accessible, we were able to conduct experiments only on the items of the training set. On the other hand, most of the one-word hypernym pairs are the same in both the FinSim-1 and the FinSim-2 datasets. Thus, we merged the two datasets and to delete the duplicates. After this step, we obtained 202 one-word and 405 two-word term-hypernym pairs (607 pairs in total). The number of unique word-types in the dataset vocabulary is 546, among which 185 and 134 words are not included in the general-domain and in the finance-specific vocabularies of the models, respectively.

Term	**Label**
S&P 100 Index	Equity Index
Green Bond	Bonds
Index Forward	Forward
Preference Share	Stocks

Table 1: Examples of term-hypernym pairs from the FinSim-2 dataset.

3.2. Systems and Settings

We used **BERT Base** (Devlin et al., 2019) as a general-domain Transformer model. BERT consists of a series of stacked Transformer encoders, and was trained using a masked language modeling and a next sentence prediction objective on a concatenation of the Books Corpus (Zhu et al., 2015) and of the English Wikipedia. For the domain-specific models, we used two versions of the FinBERT model introduced by Yang et al. (2020), namely **FinBERT BaseVocab** (FV w/ BV) and **FinBERT FinVocab** (FB w/ FV). The main difference is that the former was initialized from the original BERT Base (i.e., it also uses the same general-domain vocabulary) and further pretrained on three financial corpora (the Corporate Reports 10-K & 10-Q from the Securities Exchange Commission [2], the Earnings Call Transcripts from the Seeking Alpha website [3] and the Analyst Reports from the Investext database), while the latter was trained afresh on financial corpora for 1M iterations and uses a domain-specific financial vocabulary. As in Peng et al.(2021), we specifically chose the model by Yang and colleagues because of the availability of two versions obtained with different methods for domain adaptation. This allows us to measure the impact of the vocabulary on task performance.

[1] https://spec.edmcouncil.org/fibo/

[2] https://www.sec.gov/edgar.shtml

[3] https://seekingalpha.com/

Type	Prompt	Example
A	a(n) TERM is a(n) [MASK].	A Share is a [MASK].
B	TERMs are [MASK].	Shares are [MASK].
C	a(n) TERM is a type of [MASK].	A Share is a type of [MASK].
D	a(n) [MASK], such as a(n) TERM.	A [MASK], such as a Share.
E	a(n) TERM is a(n) [MASK], so is a(n) CO-TERM.	A Share is a [MASK], so is a quota.

Table 2: List of the prompt templates.

For each target term in the dataset, we fed a prompt including the term, and asked the masked language models to assign a probability score to each candidate hypernym. The hypernyms were then ranked, for each term, by decreasing probability value. Following Schick and Schütze (2021), we only modeled the probability of the hypernym labels, i.e., the probabilities of the rest of the vocabulary were not taken into account.

We conducted experiments with 5 types of prompts, including using a linking verb to form two basic types of prompts, and using *type-of*, *such-as*, and multiple hyponym. The details of the prompts are shown in Table 2 (Notice that all the prompts have been built with the appropriate determiner *a* or *an* for both the term and the masked hypernym). Type **A**, the classic **is-a** pattern, is the most basic form of hypernym prompting. Specifically, the terms and labels are pluralized in type **B** for checking the consistency of the prediction: if the systems have some actual knowledge about hypernymy-hyponymy relations, we would expect the attribution of a hypernym to a term to be the same, regardless of whether the term is singular or plural. For instance, if the system knows that the hypernym of *apple* is *fruit*, then the system should also be able to recognize the correct hypernym *fruits* for *apples*. However, previous studies showed that, in Transformers' predictions, this is often not the case (Ravichander et al., 2020). For all the other prompts, instead, both the hyponym term and the hypernym label are singular. Type **C** is the **type-of** pattern, a variation of the basic Type A prompt. Type **D** is the **such-as** pattern: although it is a sentence fragment rather than a full sentence, it represents a more natural pattern of co-occurrence of lexemes in a hyponym-hypernym relation (it is quite rare to see the lexemes co-occurring specifically in patterns A-C, except in text like encyclopaediae and wikis). Type D, in particular, has been reported to be one of the most effective ways of prompting the hypernym relation (Hanna and Mareček, 2021).

Type **E** is the **multiple hyponyms** prompt. A co-hyponym, the CO-TERM, is automatically found by using pretrained FastText embeddings (Bojanowski et al., 2017). At first, we looked for the nearest neighbor of each word to find co-hyponym examples. However, after inspecting the results, we chose to use the second nearest neighbor as the hyponym instead of the closest one, because the nearest neighbor always turned out to be the capitalized version of the word itself. As shown by Hanna and Marecek (2021), inserting a co-hyponym in the prompt is likely to query the desired hypernym more precisely. Using off-the-shelf FastText vectors allow us to automatize and speed up the procedure of finding the co-hyponym word. Intuitively, adding a co-hyponym in the sentence makes the prompt more informative, because it gives the language model more semantic information about the general category that needs to be predicted.

The hypernym *MMIs*, which is present in both datasets, is not included in BaseVocab, nor in FinVocab (neither in the singular, nor in the plural form). Meanwhile, after pluralization (pattern of type **B**), the hypernym label *Swaps* is not included in BaseVocab, but it is included in FinVocab. During the encoding procedure, the words not included in the vocabulary will be split into subwords, e.g., *Swaps* → *Swap*## and ##*s*. The language model will be unable to guess these hypernyms with one single [MASK] token. Therefore, as the prompt-based learning requires that we convert the hypernym to a corresponding identification number in the vocabulary, the missing hypernym labels are added to the vocabulary of the pretrained language models, so that they can be identified with unique numbers. However, the word representations of these added words are randomly initialized without optimization.

Finally, prompt-based learning requires mapping each hypernym label to a word from the vocabulary of the language model. For two-word hypernyms, e.g., *Equity Index* and *Credit Index*, we first merge them into a single category *Index* and jointly evaluate with other one-word hypernym labels. Then, we perform an extra disambiguation step to discriminate between these two-word hypernyms, by creating an additional prompt. See examples below for the prompts of Type **A** and **C**.

1. *A S&P 100 Index is a/an* [MASK] *Index*.

2. *A S&P 100 Index is a type of* [MASK] *Index*.

In this case, the language model is asked to assign the probabilities of words *Equity* and *Credit* only.

3.3. Metrics

The predictions were evaluated in terms of *Accuracy* and *Mean Rank*. The systems are not expected just to output a prediction for each instance; they have to output a rank of the candidate labels, from the most to the least likely one. The Accuracy and Mean Rank metrics are defined as follows:

Type	BERT Base		FB w/ BV		FB w/ FV		Average	
	ACC	Mean Rank	ACC	Mean Rank	ACC	Mean Rank	ACC	Mean Rank
A	64.42	1.49	69.19	1.41	57.50	1.82	63.70	1.57
B	38.72	2.23	13.84	3.19	43.49	2.39	32.02	2.60
C	72.32	1.65	71.17	1.69	49.75	2.15	64.42	1.83
D	75.12	1.39	**82.21**	**1.28**	39.87	2.27	65.73	1.65
E	74.79	1.38	78.42	1.32	50.74	2.15	**67.98**	**1.62**
Average	**64.78**	**1.63**	62.97	1.78	48.27	2.15		

Table 3: Accuracy(%) and mean rank of hypernym detection on the merged dataset. The best scores are in **bold**.

$$Accuracy = \frac{1}{n} * \sum_{i=1}^{n} I(y_i = y_i^l[0]) \qquad (1)$$

$$MeanRank = \frac{1}{n} * \sum_{i=1}^{n} rank_i \qquad (2)$$

Notice that, in Equation (2), $rank_i$ corresponds to the rank of the correct label if the latter is among the top 3 predictions and 4 otherwise, as in the Semeval 2018 evaluation of the hypernymy discovery task (Camacho-Collados et al., 2018).

4. Results and Discussion

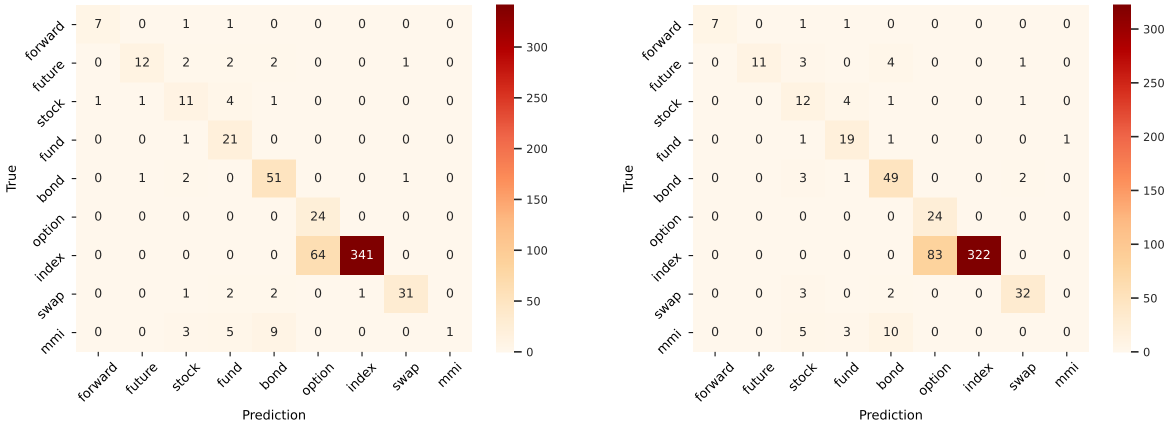
Table 3 illustrates the accuracy and mean rank scores of the three language models and prompt templates on the merged dataset. We observe that the prompt can heavily affect the detection results. For example, without any fine-tuning, the accuracy score of the FinBERT w/ BV model can be changed from 13.84 to 82.21 by changing the prompt from the basic plural type **B** to type **D** (**such-as**). If we exclude the prompt of type **B**, which was included to check the prediction consistency, the average scores tend to improve by using more complex prompts, with the models generally doing better with prompt-types **D** and **E**. The result is in line with the findings of Hanna and Mareček (2021).

The accuracy and mean rank scores for the basic types are lower than the others in both BERT Base and FinBERT w/ BV. It is striking that, after changing the prompt sentence from singular to plural, all the models have a sharp performance drop. FinBERT w/ BV models is, apart from Type **B**, the model achieving the best scores (82.21 in accuracy and 1.28 in mean rank), but on the other hand, it is also the model having the largest drop after the pluralization of the prompt. This finding is consistent with previous studies on hypernymy detection with masked language models (Ravichander et al., 2020; Hanna and Mareček, 2021), and it suggests that the models might only be exploiting some surface lexical cues to predict the hypernyms, rather than learning actual semantic relations between word representations (Rambelli et al., 2020; Pedinotti et al., 2021).

From the language models' perspective, BERT Base and FinBERT w/ BV outperform the FinBERT w/ FV model, confirming previous findings in financial text sentiment analysis and numeral understanding tasks (Peng et al., 2021). Curiously, the only model with finance-specific vocabulary is the only one that achieves its best score with the basic prompt **A** among the 5 prompt types, and it is outperformed by the competitor models with prompts **B**, **C**, and **E**, suggesting that domain-specific vocabulary does not necessarily represent an advantage for this kind of tasks. This contrasts with findings in other domains such as the biomedical domain, where models with a domain-specific vocabulary have been shown to be more efficient (Gu et al., 2021; Portelli et al., 2021). Between the two models with general domain vocabulary, FinBERT w/ BV is generally better at guessing the right hypernyms, with the exceptions of prompt-types **B** and **C**. Excluding the value of the plural prompt, which gives particularly low scores for FinBERT w/ BV, the accuracy and mean rank scores of this model would be 75.16 and 1.43, against 71.66 and 1.48 of BERT Base. Since FinBERT w/ FV was only trained on financial corpora and was missing the training on general-domain text (Yang et al., 2020), the model may not have acquired a good knowledge of the semantic relations. It is also possible that hypernymy patterns such as "is a" or "is a type of" are not very frequent in financial texts, because these patterns are likely to appear in definition-like statements, while financial texts are generally read by specialists that do not need definitions for the meaning of the domain-specific terms.

Figures 1a and 1b show the confusion matrices of the two best prompts, **D** and **E**, in the FinBERT w/ BV model, respectively. The two prompts are both good at associating terms to hypernyms. For example, the detection accuracies of *Bond* and *Option* are almost 100%. Moreover, despite the unbalanced distribution of the labels (e.g., Forward), we did not observe accuracy drops for rare labels. Both prompts failed to recognize *MMI*, as expected, as it is not included in the vocabulary of the language model. The merged label *Index* is the primary source of errors for both models, with many instances of *Option* being erroneously associated with *Index*, particularly with type **E** prompts.

We show some term-hypernym pairs that are misclassified by FinBERT models with type **D** and **E** prompts in Table 4. Only the terms CDS and To Be Announced are included in both base and financial vocabularies, while the others are split into subwords. This might have misled the language models, making them unable to guess the right hypernym. On the other hand, some



(a) Type **D**

(b) Type **E**

Figure 1: Confusion matrices of FinBERT w/ BV model with type **D** and **E** prompts, respectively.

		FinBERT w/ BV				FinBERT w/ FV			
Term	Label	Rank		Prediction		Rank		Prediction	
		Type **D**	Type **E**	Type **D**	Type **E**	Type **D**	Type **E**	Type **D**	Type **E**
Sukuk	Bond	2	3	Stock	Stock	2	2	Future	Stock
To Be Announced	Bond	5	7	Future	Future	8	4	Future	Future
CDS	Swap	4	4	Bond	Bond	4	8	Index	Index
Wisdomtree Europe Hedged	Index (Equity Index)	2	2	Option	Option	3	9	Future	Future
CDX Swaption	Index (Credit Index)	2	2	Option	Option	4	2	Future	Option

Table 4: Misclassified terms of FinBERT models with type **D** and **E** prompts, respectively. The rank of probability of the correct label and the prediction result are reported as well.

in-vocabulary terms may have special meanings in the financial domain, e.g. CDS (Credit Default Swap), and they are also misclassified by FinBERT models. This may due to a failure of the language models in extracting the domain-specific meanings of the terms (e.g., the models may interpret CDS as Compact Discs). FinBERT w/ FV model generally got a worse probability rank for the hypernyms, which once again suggests that domain-specific vocabulary does not necessarily represent an advantage for this kind of task.

Finally, for the original two-word hypernyms (*Equity Index* and *Credit Index*) we further analyzed the detection accuracy by creating an additional disambiguation prompt, using the word *Index/Indices*. The language models are asked to fill the [MASK] with only *Equity* or *Credit* as illustrated in Section 3.2. Table 5 shows the accuracy score of two-word hypernyms detection. We still observe large drops for the plural prompts, while the basic type **A**, and the types **D** and **E** are the most effective patterns. Considering all models, type **A** achieves the more stable performance. Patterns D and E also obtained perfect scores, but the average is pulled down by the low performance of FinBERT w/ FV. Among the language models, FinBERT w/ BV is the top-scoring model as it manages to guess all the hypernyms correctly in three cases out of five.

Overall, the results prove that training Transformer language models on specialized corpora can improve hypernymy detection. Apart from the consistency issue with plural prompts, FinBERT w/ BV is the model achieving most often the highest scores, and it tends to perform better than BERT Base for the more informative prompts (types **D** and **E**). BERT Base is still competitive with the domain-adapted model, and shows more consistency with the basic prompts. Finally, FinBERT w/ FV performs the worst of the three, suggesting that knowing financial-specific vocabulary *per se* does not help hypernym detection. Almost all the hypernymy labels and the majority of the terms to be classified were included in both vocabularies, with a slightly better coverage using FV (only 134 missing terms against 185 for BV). The fact that this small advantage did not help the FinBERT w/ FV model suggests that the internal representations of the Transformers are able to efficiently exploit lexical cues from the

Type	Bert Base	FB w/ BV	FB w/ FV	Average
A	**100.00**	**100.00**	94.57	**98.19**
B	69.14	78.27	71.85	73.09
C	72.10	74.07	67.90	71.36
D	**100.00**	**100.00**	76.54	92.18
E	99.51	**100.00**	89.63	96.38
Average	88.15	**90.47**	80.10	

Table 5: The accuracy(%) scores of two-word hypernym label detection. The best scores are in **bold**.

context to make their predictions, even when the target words are not included in their vocabulary. However, it should also be noticed that FinBERT w/ FV achieved a higher accuracy than the competitors with the plural prompt of type **B** (see Table 3). This might be due to the fact that this model is the only one that includes the pluralized forms of all the hypernym labels (except for MMI) in its vocabulary.

5. Conclusion

In this paper, we proposed a comparison between general and domain-adapted pretrained language models' performance of the task of financial hypernym detection via masked language modeling. We also tested different types of prompts used to search for hypernym. The results indicate that the domain adaptation can improve the language model's capacity to retrieve the right hypernym, although the models are more efficient when they also retain a general-domain vocabulary. In addition, we observed that different prompts have an important impact on hypernym detection; that is, more natural and informative prompts generally lead to better scores. Future work will experiment with new methods to refine the Transformers' internal representations to identify hypernyms and other lexical-semantic relations. For example, one could explore fine-tuning the model on triples from knowledge graphs (Bosselut et al., 2019), or extracting relation embeddings from the language model output (Ushio et al., 2021).

Acknowledgements

The authors acknowledge the support from the projects "Analyzing the semantics of Transformers representations for financial natural language processing" (ZVYU) and "Conversational Brains: A Multidisciplinary Approach" (ZG9X), funded by the Hong Kong Polytechnic University.

6. Bibliographical References

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*.

Boella, G. and Di Caro, L. (2013). Supervised Learning of Syntactic Contexts for Uncovering Definitions and Extracting Hypernym Relations in Text Databases. In *Machine Learning and Knowledge Discovery in Databases*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of ACL*.

Camacho-Collados, J. and Navigli, R. (2017). Babel Domains: Large-Scale Domain Labeling of Lexical Resources. In *Proceedings of EACL*.

Camacho-Collados, J., Delli Bovi, C., Espinosa-Anke, L., Oramas, S., Pasini, T., Santus, E., Shwartz, V., Navigli, R., and Saggion, H. (2018). SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*.

Chersoni, E. and Huang, C.-R. (2021). PolyU-CBS at the FinSim-2 Task: Combining Distributional, String-Based and Transformers-Based Features for Hypernymy Detection in the Financial Domain. In *Companion Proceedings of the Web Conference*.

Chersoni, E., Rambelli, G., and Santus, E. (2016). CogALex-V Shared Task: ROOT18. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

El Maarouf, I., Mansar, Y., Mouilleron, V., and Valsamou-Stanislawski, D. (2021). The FinSim 2020 Shared Task: Learning Semantic Representations for the Financial Domain. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Espinosa-Anke, L., Camacho-Collados, J., Delli Bovi, C., and Saggion, H. (2016). Supervised Distributional Hypernym Discovery via Domain Adaptation. In *Proceedings of EMNLP*.

Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Flati, T., Vannella, D., Pasini, T., and Navigli, R. (2016). MultiWiBi: The Multilingual Wikipedia Bitaxonomy Project. *Artificial Intelligence*, 241:66–102.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Hanna, M. and Mareček, D. (2021). Analyzing BERT's Knowledge of Hypernymy via Prompting. In *Proceedings of the EMNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackBoxNLP)*.

Keswani, V., Singh, S., and Modi, A. (2020). IITK at the FinSim Task: Hypernym Detection in Financial Domain via Context-free and Contextualized Word Embeddings. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Lenci, A. and Benotto, G. (2012). Identifying Hypernyms in Distributional Semantic Spaces. In *Proceedings of * SEM*.

Liu, H., Chersoni, E., Klyueva, N., Santus, E., and Huang, C.-R. (2019). Semantic Relata for the Evaluation of Distributional Models in Mandarin Chinese. *IEEE Access*, 7:145705–145713.

Liu, Z., Huang, D., Huang, K., Li, Z., and Zhao, J. (2020). FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. In *Proceedings of IJCAI*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv preprint arXiv:2107.13586*.

Mansar, Y., Kang, J., and Maarouf, I. E. (2021). The FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain. In *Companion Proceedings of the Web Conference*, pages 288–292.

Nguyen, K. A., Köper, M., Schulte im Walde, S., and Vu, N. T. (2017). Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Proceedings of EMNLP*.

Pandia, L., Cong, Y., and Ettinger, A. (2021). Pragmatic Competence of Pre-trained Language Models through the Lens of Discourse Connectives. In *Proceedings of CONLL*.

Pedinotti, P., Rambelli, G., Chersoni, E., Santus, E., Lenci, A., and Blache, P. (2021). Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge. In *Proceedings of *SEM*.

Peng, B., Chersoni, E., Hsu, Y.-Y., and Huang, C.-R. (2021). Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks. In *Proceedings of the EMNLP Workshop on Economics and Natural Language Processing*.

Portelli, B., Lenzi, E., Chersoni, E., Serra, G., and Santus, E. (2021). BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection. In *Proceedings of EACL*.

Rambelli, G., Chersoni, E., Lenci, A., Blache, P., and Huang, C.-R. (2020). Comparing Probabilistic, Distributional and Transformer-Based Models on Logical Metonymy Interpretation. In *Proceedings of AACL-IJCNLP*.

Ravichander, A., Hovy, E., Suleman, K., Trischler, A., and Cheung, J. C. K. (2020). On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT. In *Proceedings of *SEM*.

Roller, S. and Erk, K. (2016). Relations such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment. In *Proceedings of EMNLP*.

Roller, S., Erk, K., and Boleda, G. (2014). Inclusive Yet Selective: Supervised Distributional Hypernymy Detection. In *Proceedings of COLING*.

Sanchez, I. and Riedel, S. (2017). How Well Can We Predict Hypernyms from Word Embeddings? A Dataset-centric Analysis. In *Proceedings of EACL*.

Santus, E., Lenci, A., Lu, Q., and Im Walde, S. S. (2014a). Chasing Hypernyms in Vector Spaces with Entropy. In *Proceedings of EACL*.

Santus, E., Lu, Q., Lenci, A., and Huang, C.-R. (2014b). Taking Antonymy Mask Off in Vector Space. In *Proceedings of PACLIC*.

Santus, E., Yung, F., Lenci, A., and Huang, C.-R. (2015). Evalution 1.0: an Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the ACL Workshop on Linked Data in Linguistics: Resources and Applications*.

Santus, E., Lenci, A., Chiu, T.-S., Lu, Q., and Huang, C.-R. (2016). Nine Features in a Random Forest to Learn Taxonomical Semantic Relations. In *Proceedings of LREC*.

Schick, T. and Schütze, H. (2021). Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of EACL*, pages 255–269.

Shwartz, V., Goldberg, Y., and Dagan, I. (2016). Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In *Proceedings of ACL*.

Shwartz, V., Santus, E., and Schlechtweg, D. (2017). Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection. In *Proceedings of EACL*.

Ushio, A., Camacho-Collados, J., and Schockaert, S. (2021). Distilling Relation Embeddings from Pre-trained Language Models. In *Proceedings of EMNLP*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Weeds, J. and Weir, D. (2003). A General Framework for Distributional Similarity. In *Proceedings of EMNLP*.

Weeds, J., Clarke, D., Reffin, J., Weir, D., and Keller, B. (2014). Learning to Distinguish Hypernyms and Co-hyponyms. In *Proceedings of COLING*.

Xiang, R., Chersoni, E., Iacoponi, L., and Santus, E. (2020). The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Yang, Y., Uy, M. C. S., and Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.