

The FinSim 2020 Shared Task: Learning Semantic Representations for the Financial Domain

Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, Dialekti Valsamou-Stanislawski

Fortia Financial Solutions

{ismail.elmaarouf, youness.mansar, virginie.mouilleron, dialekti.valsamou}@fortia.fr

Abstract

The FinSim 2020 shared task, colocated with the FinNLP workshop, offered the challenge to automatically learn effective and precise semantic models for the financial domain.

Going beyond the mere representation of words is a key step to industrial applications that make use of Natural Language Processing (NLP). This is typically addressed using either unsupervised corpus-derived representations like word embeddings, which are typically opaque to human understanding but very useful in NLP applications or manually created resources such as taxonomies and ontologies, which typically have low coverage and contain inconsistencies, but provide a deeper understanding of the target domain.

FinSim is inspired from previous endeavours in the Semeval community, which organized several competitions on semantic/lexical relation extraction between concepts/words. To the best of our knowledge, FINSIM 2020 was the first time such a task was proposed for the Financial domain.

The shared task attracted 6 participants and systems were ranked according to 2 metrics, Accuracy and Mean rank. The best system beat the baselines by over 20 points in accuracy.

1 Introduction

The FinSim 2020 shared task, organized by Fortia Financial Solutions¹, an AI startup with expertise in Financial Natural Language Processing (NLP), was part of the second edition of the 2nd Workshop on Financial Technology and Natural Language Processing (FinNLP²). It focused on automatically learning effective and precise semantic models adapted to the financial domain. More specifically it addressed the task of automatic categorization of financial instrument terms. The range of financial instrument is vast and the category encompasses all sorts of tradable contracts. Financial instruments can be challenging as, while traditional instruments such as

Bonds or *Stocks* are straightforward, other instruments such as *Futures* pose a number of difficulties as they may apply to various underlying instrument types (e.g. bond futures, equities futures). Similarly, while the a feature is key to the definition of some instruments such as future contracts, its presence is not critical to the definition others. Thus, the challenge of automatic classification of financial instruments, is coupled with a challenge of semantic analysis.

Going beyond the mere representation of words is a key step to industrial applications that make use of Natural Language Processing (NLP). The semantic models those applications rely on is critical to their success in addressing traditional semantic tasks such as Named Entity Recognition, Relation Extraction, or Semantic parsing.

NLP applications can rely on annotated datasets, but there are also approaches which leverage manual resources like taxonomies, ontologies, and knowledge bases, for their source of knowledge. Indeed, creating annotated dataset is a costly endeavour and it is challenging to design an annotation dataset that can be exploited for other tasks than the ones it was initially designed for. Thus, on one end of the spectrum, there are approaches which typically rely on grammar or regular expressions and heavily rely on the quality of a manually created resource.

On the other end of the spectrum, there are Machine Learning (ML) approaches which attempt to automatically build semantic resources from raw text, like word embeddings, that are typically opaque to human understanding. In the literature, and e.g. in competitions, such unsupervised approaches have been more successful in building effective NLP applications. In industrial applications, both approaches have met success and it is true that in some contexts, approaches relying on manually crafted are often preferred to pure ML approaches because the former provide more control and are more predictable.

Finally there are approaches which attempt to automatically make use of manual resources but also rely on automatically derived word representations in order to build hybrid models. This is to these approaches that the task is addressed.

The FinSim task provided a raw corpus of financial prospectuses, from which to derive automatic representations, a train set of financial instrument terms classified by types of financial instruments, as well as mappings to an ontology of the financial domain, namely FIBO (The Financial Industry

¹<https://www.fortia.fr/>

²<https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2020/home>

Business Ontology³). There are also resources available on the internet such as the Investopedia dictionary of financial terms⁴, and classifications such as the CFI⁵.

The paper is structured as follows: section 2 will introduce previous work related to the shared task, and section 3 will describe it in detail. Section 4 will introduce participants and section 5 will present the results.

2 Related work

The task proposed at FinSim 2020 is a task of hypernym categorization: given a training set of terms and a fixed set of labels, participants are asked to learn to categorize new terms to their most likely hypernym. Two words are said to be in a hypernymy (or ISA) relation if one of them can be conceived as a more generic term (e.g. “seagull” and “bird”).

2.1 Literature on hypernymy extraction

Semantic relation extraction is a topic largely discussed in the literature and has been addressed from a variety of angles. Seminal work include the creation and use of hyperonym lexical patterns [Hearst, 1992] to extract hyponym-hypernym pairs. Substantive work draws from automatic thesaurus construction (see [Grefenstette, 1994]) which led to work on distributional analysis, which is the basis for a lot of data-driven work including [Lin, 1998] or [Baroni and Lenci, 2009].

More recently, neural networks have been used to learn word representations from text and proved very effective in a variety of NLP tasks ([Mikolov *et al.*, 2013a], [Devlin *et al.*, 2019], [Radford *et al.*, 2019]). Such data-driven approaches capture a lot of similarities between terms in context, however it is not clear how those similarities relate to handcrafted relations such as hypernymy.

Finally, another strand of research to which we cannot do justice here, makes use of knowledge bases to operate relation extraction such as hypernymy relations. This domain tends to focus on names rather than nouns and in general, systems are not relation-specific but tend to cover multiple relation types. Work include [Mintz *et al.*, 2009], or [Zeng *et al.*, 2015]. A number of approaches proposed to create knowledge base embeddings, in which the similarity between terms or names is automatically derived from the structure of the knowledge base (see e.g. [Wang *et al.*, 2014]).

2.2 Hypernymy relation extraction shared tasks

In terms of shared tasks, Taxonomy Extraction Evaluation (TExEval, [Bordea *et al.*, 2015]) was the first shared task on taxonomy induction by focusing on the last step of organizing the taxonomy into hypernym-hyponym relations between (pre-detected) terms in four different domains (chemicals, equipment, foods, science). Because they did not provide a corpus, participants were limited in the data they could use and had to structure a list of terms into a taxonomy (with the possibility of adding intermediate concepts).

The second edition of Tex-Eval ([Bordea *et al.*, 2016]) proposed the same challenge but focusing on multilinguality (English, Dutch, Italian and French) and 3 target domains (Environment, Food and Science). This time, the organizers provided a script to build a corpus from Wikipedia⁶.

[Jurgens and Pilehvar, 2016] addressed the problem of classifying new terms against an existing taxonomy, a task they called taxonomy enrichment. This task relied on Wordnet⁷ and asked participants to attach a given word to, or merge it with, an existing WordNet synset. For each word, participants were provided a definition in natural language. The construction of the dataset (1,000 words) proved difficult, particularly in the identification of the appropriate synsets to associate a given term with, for reasons listed in their paper and mainly related to the structure of Wordnet.

[Camacho-Collados *et al.*, 2018] proposed a multilingual (English, Spanish, and Italian), multi-domain (Medical and Music) task for hypernym discovery. The task put forward the necessity of providing a corpus to limit the search space for hypernyms; as opposed to [Bordea *et al.*, 2016] which used an Encyclopaedic corpus, [Camacho-Collados *et al.*, 2018] provided a web-base corpus (3-billion word UMBC corpus⁸) as well as data from Pubmed⁹. The task provided participants with a list of hyponym-hypernym pairs, and, despite the fact that both terms occurred in the corpus, there was no guarantee that there were hypernymy contexts.

Finally, there are also a large number of datasets and challenges that specifically look at how to automatically extract relations in order to populate knowledge bases such as DBpedia or Wikidata. The Knowledge Base Population track (KBP) at the NIST Text Analysis Conference¹⁰ is a popular series which focus on relations involving Named entities rather than words of the language (see [Shen *et al.*, 2014] for more details).

To the best of our knowledge, FINSIM 2020 was the first time a task of Hypernymy categorization was proposed for the Financial domain.

3 Task description

3.1 Overview

At FinSim, participants were given a list of carefully selected terms from the financial domain such as “European depositary receipt”, “Interest rate swap” and were asked to design a system which can automatically classify them to the most relevant hypernym (or top-level) concept in an external ontology. For example, given the set of concepts “Bonds”, “Unclassified”, “Share”, “Loan”, the most relevant hypernym of “European depositary receipt” is “Share”. FinSim focused on the category of Financial instruments. A financial instrument is a general category for any contract that can be traded by investors.

⁶<http://wikipedia.org>

⁷<https://wordnet.princeton.edu/>

⁸<http://ebiquity.umbc.edu/blogger/2013/05/01/umbc-webbase-corpus-of-3b-english-words/>

⁹https://www.nlm.nih.gov/databases/download/pubmed_medline.html

¹⁰<https://tac.nist.gov/tracks/index.html>

³<https://spec.edmcouncil.org/fibo/>

⁴<https://www.investopedia.com/financial-term-dictionary-4769738>

⁵https://www.quotemedia.com/apifeeds/cfi_code

3.2 Description of the dataset and labels

We provided participants with (i) a raw corpus from which the participants could extract financial word representations, (ii) an ontology that structures and associate the financial terms with their labels from a carefully designed tagset and (iii) a list of term category pairs that instantiate the ontology concepts.

3.3 The Prospectus corpus

We provided a set of financial prospectuses in English to be used for training embeddings for this task¹¹. Financial prospectuses provide key information for investors and detail investment rules linked to particular financial instruments. The files had been downloaded from various websites and it was forbidden to re-distribute them. The corpus size is estimated to about 10 million tokens. More precisely, the corpus is made of 156 prospectuses in PDF format. Their individual size varies between a dozen pages to several hundreds.

3.4 The FIBO ontology

FIBO is an interesting and pioneering effort (still in progress) to formalize the semantics of the financial domain using a large number of ontologies. More detail can be found on their website¹². Participants were encouraged to use this resource (as well as others) in designing their system and this is why we provided a number of scripts to facilitate its processing.

We also provided a mapping from each of the categories used in FinSim to a concept in the FIBO ontology (in the file `data/outputs/fibo_mapping.json`). In creating this mapping, we chose to map FinSim labels to the most relevant concepts rather than to “instruments” concepts from the instruments ontology. Indeed some instruments, like *Swaps*, have an ontology of their own. Finally, it is worth noting that there is a development version of FIBO which may contain useful content yet not finally released or validated for production.

3.5 The FinSim dataset

Tagset

The first stage for building the dataset involved building the tagset. FinSim focuses on 8 categories of financial instruments (Bonds, Forwards, Funds, Future, MMIs, Option, Stocks, Swap).

These labels refer to the most important and most frequently used types of financial instruments (except for cash deposits). As previously noted, there are multiple classifications available for financial instruments, such as CFI codes. Many organizations design their own classifications or adjust existing ones according to their needs. Categorisation of financial instruments can be approached from two angles:

- a featured-based approach will classify instruments according to their properties (such as whether it contains a maturity condition)
- a kind-based approach will classify them according to their prototypical kind in a list of available kinds (even if they share properties with other kinds of instruments).

¹¹ Available under `data/English_prospectuses` in our data

¹² <https://spec.edmcouncil.org/fibo/>

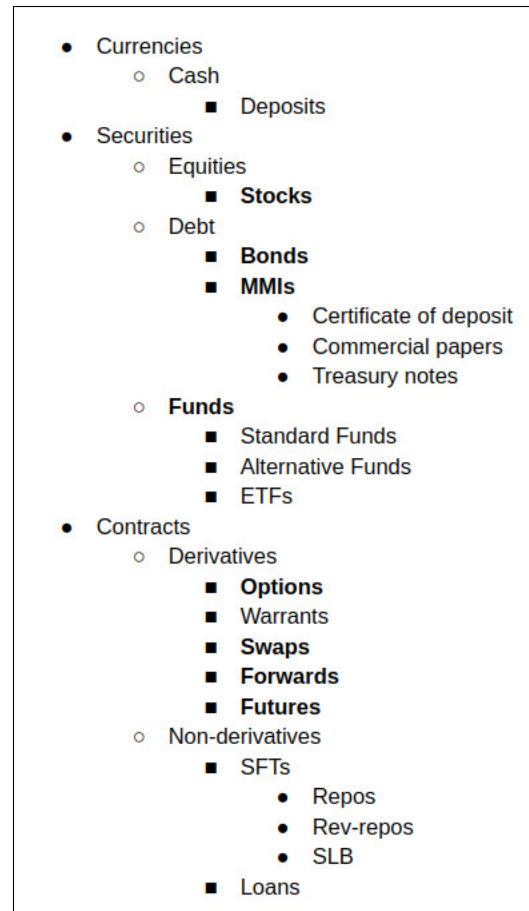


Figure 1: FinSim Asset tree

Termset

The next stage involved selecting appropriate terms of financial instruments and categorizing them. We iteratively and manually built up the lexicon by looking up keywords on the internet and in the prospectus corpus.

Because we wanted to test models’ capacity to generalize from unseen data, we included a set of terms not present in the Prospectus corpus, however the majority of terms had at least one mention. We also selected different types of linguistic expressions. For example, funds are often designated:

- elliptically by naming them by their type, e.g. *SICAV*, *Unit trust*,
- via an acronym, which are known to be very ambiguous forms ,e.g. *AIF*,
- by their role, e.g. *feeder* or *master*
- by selecting larger noun groups including the hypernym, e.g. *hedge fund*, *closed end fund*
- the term itself *fund* or a compound variant, e.g. *subfund*

The dataset was built by two annotators and all were reviewed by a second annotator, expert in the finance domain, who built the asset tree depicted in Figure 1 .

As in [Camacho-Collados *et al.*, 2018] the train and the test sets were of equal size (see Table 1). Careful attention was

Corpus	train	test
Number of terms	100	99

Table 1: Dataset of terms for FinSim 2020

taken to use the same class distribution between train and test datasets. The format of the data was a JSON file containing the terms and their associated hypernym, as {"label": "Option", "term": "Over-the-counter option"}.

3.6 Metrics

As metrics we used Average Accuracy and Mean Rank. For each term x_i with a label y_i from the n samples in the test set, the expected prediction is a top 3 list of labels ranked from most to least likely to be equal to the ground truth by the predictive system \hat{y}_i^l . We note by $rank_i$ the rank of the correct label in the top-3 prediction list, if the ground truth does not appear in the top-3 then $rank_i$ is equal to 4.

Given those notation the accuracy can be expressed as:

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

And the Mean Rank as:

$$Mean_Rank = \frac{1}{n} * \sum_{i=1}^n rank_i$$

A lower value of the Mean Rank is better. This metric is useful because it does not treat all the errors the same way, if the correct label is ranked fourth in the prediction list then its evaluation is penalized more heavily than if it is ranked second. Mean Rank was used by [Camacho-Collados *et al.*, 2018] in their shared task.

3.7 Baselines

Two baselines were provided to help participants design their systems.

The first baseline used pretrained embeddings to compute a representation for the labels and computed the distance between this vector and the vector of each candidate term.

The second baseline split the term into words and using their pretrained embeddings, learnt a Logistic Regression model in a supervised manner from the trainset.

4 Participants and systems

Team	Affiliation
IITK	IITK, India
Anuj	Publicis Sapient, USA
ProsperaMnet	University of Szeged - Hungary
FINSIM20	IIIT & VIT, India
Ferryman	Complex lab, UESTC
AIAI	Rakuten

Table 2: List of the 6 participating teams in the FinSim Shared Task.

A total of 6 teams who participated from which 4 submitted a paper to describe their method. The shared task brought together private and public research institutions including Publicis Sapient, IITK, IIIT, VIT and University of Szeged (see Table 2 for more details).

Participating teams explored and implemented a wide variety of techniques and features. In this section, we give a short summary of the methods proposed by each participating team (for further details, all papers are published in the proceedings of the FinNLP 2020 Workshop).

IITK This team’s system is based on a comparison between context-independent word embeddings in the form of Word2vec word vectors [Mikolov *et al.*, 2013b] that were trained on financial prospectuses and context-dependent word vectors using BERT [Devlin *et al.*, 2019]. Their system is a combination of two prediction strategies. The first strategy is a rule-based approach that is applied to test samples that have exactly one label mentioned in the entity that needs to be classified, in this case the top prediction simply the label that was mentioned. The second strategy is based on a Naive Bayes classifier applied to word embeddings. Their system achieved the overall rank 1 in the shared task when based on 100 dimension Word2vec vectors, over-performing larger dimension Word2vec vectors and BERT embeddings.

Anuj This team took advantage of an external data source (*Investopedia*) in order to supplement the terms with their in-domain definition. Their ML system is based on hand-crafted features and bi-gram TF-IDF features followed by a linear SVM model. This system scored 1st on the accuracy metric and second on the overall ranking of the shared task.

ProsperaMnet This team builds their system on sparse word embeddings and an algorithm proposed in [Balogh *et al.*, 2020] that tries to quantify the extent to which a specific dimension of the sparse word vector relates to certain common sense properties of concepts. They compare their approach to a dense-vector baseline and show that their approach works better than the baseline, especially when used with the best regularization hyper-parameter. This System scored second on the Average Rank metric and 3rd in the overall ranking.

FINSIM20 This team compared different types of algorithms under multiple configuration in order to solve the task. They first used either generic Glove word embeddings [Pennington *et al.*, 2014] or fine-tuned on financial prospectuses along with a cosine similarity metric in order to rank the labels that best fit an entity. They also run experiments using a KNN approach either based on the original training set or an extended version of the data-set that they generated using Hearst Patterns [Hearst, 1992]. They also explored graph-based methods where they built a graph in which each entity is a node and then leveraged the relations between nodes to detect hypernymy-hyponymy. Their best approach is based on Universal Sentence Encoder [Cer *et al.*, 2018] and cosine similarity, this approach scored third place overall on the shared task.

5 Results and discussion

5.1 Results

We ranked submissions using the metrics defined in 3.6 and we provided an overall ranking by combining them. IITK came first as it obtained the best performance according to both metrics. ProsperaMnet and Anuj were second depending on the metric. This variation is explained by the fact that the Anuj system was a single class model and only provided a single category as answer (as opposed to a ranked list of labels).

Rank	Team
1	IITK
2	Anuj
3	ProsperaMnet
3	FINSIM20
4	Ferryman
5	AIAI

Table 3: Overall results

Team	Mean rank
IITK	1.21
ProsperaMnet	1.34
Anuj	1.42
FINSIM20	1.43
Ferryman	1.59
AIAI	1.94

Table 4: Mean-Rank Ranking

Team	Accuracy
IITK	0.858
Anuj	0.858
FINSIM20	0.787
ProsperaMnet	0.777
Ferryman	0.757
AIAI	0.545

Table 5: Accuracy Ranking

5.2 Discussion

Most teams used some type of unsupervised word embeddings, either being context-independent like Word2vec [Mikolov *et al.*, 2013a] or Glove [Pennington *et al.*, 2014] or context-dependent like BERT [Devlin *et al.*, 2019] or Universal Sentence Encoder [Cer *et al.*, 2018] while one team built their system on TF-IDF of bi-gram words. The word embeddings are generally averaged and used as is for the subsequent steps in the predictive system.

Given the small size of the training data, some teams tried to extend the dataset either by using an external source of term definitions or by automatically extracting hypernym examples using Hearst Patterns [Hearst, 1992].

Baseline	Mean rank	Accuracy
Baseline 2	1,838	0,606
Baseline 1	2,111	0,505

Table 6: Baseline Performance

The most common unsupervised approach for this classification was using the cosine similarity between the term representation and the label representation in the embedding space, the labels are then ranked by decreasing order of similarity. Since the training sample is small, most teams based their approach on a model that learns linear boundaries between the target classes like a linear SVM or a logistic regression.

6 Conclusions and perspectives

This paper introduced FinSim, the first shared task in Hypernymy categorization to focus on the financial domain. This task attracted 6 teams across the world, although 20 teams initially expressed their interest. The challenge posed by FinSim is how to appropriately make use of corpus-derived representations, such as word embeddings, with existing manually designed taxonomies. In order to do that, it drew from previous similar shared tasks and proposed a train set of terms along with their categories, from a tagset of financial instruments.

The task was addressed by a variety of approaches, both supervised and unsupervised, and attempting to make use of external resources such as Investopedia or FIBO, and pretrained embeddings such as Glove or BERT, or using more traditional ngram counts as features for their models.

The best team reached 0.858 accuracy, which largely beats the baselines (0.5 and 0.6), which means that, despite the small size of the corpus, the effort put in modeling paid off.

The FinSim shared task made it easy for participants to access data by providing scripts for data processing and baseline models as guidance.

This task focused on financial instruments. Obviously one way this task could be extended, would be by selecting a larger number of financial instruments. One of the pieces of feedback from participants was that the size of the corpus was small, and powerful methods did not work. Another direction for future work is to look at different semantic categories, as provided in FIBO, e.g. types of business entities, types of rates and indicators. Another perspective would be to change the type of task and turn it into a Named Entity Recognition task, but that would involve a substantial dataset creation. Finally it is also envisaged to extend the task to other languages such as French or German.

Acknowledgments

This work was fully supported by Fortia Financial Solutions.

References

[Balogh *et al.*, 2020] Vanda Balogh, Gábor Berend, Dimitrios I. Diochnos, and György Turán. Understanding the semantic content of sparse word embeddings using a commonsense knowledge base. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*,

- The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7399–7406. AAAI Press, 2020.
- [Baroni and Lenci, 2009] Marco Baroni and Alessandro Lenci. One distributional memory, many semantic spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 1–8, Athens, Greece, March 2009. Association for Computational Linguistics.
- [Bordea et al., 2015] Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of SemEval 2015, co-located with NAACL HLT 2015*, Denver, Col, USA, 2015.
- [Bordea et al., 2016] Georgeta Bordea, Els Lefever, and Paul Buitelaar. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, CA, USA, 2016.
- [Camacho-Collados et al., 2018] Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, , and Horacio Saggion. Semeval-2018 task 9: Hyponym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States, 2018.
- [Cer et al., 2018] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics.
- [Devlin et al., 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [Grefenstette, 1994] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, London, Dordrecht, 1994.
- [Hearst, 1992] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, 1992.
- [Jurgens and Pilehvar, 2016] David Jurgens and Mohammad Taher Pilehvar. Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, CA, USA, 2016.
- [Lin, 1998] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics (Coling)*, Montreal, Quebec, Canada, 1998.
- [Mikolov et al., 2013a] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13*, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.
- [Mikolov et al., 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [Mintz et al., 2009] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.
- [Pennington et al., 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
- [Radford et al., 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- [Shen et al., 2014] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2014.
- [Wang et al., 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI’14*, page 1112–1119. AAAI Press, 2014.
- [Zeng et al., 2015] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762, 2015.