PolyU-CBS at the FinSim-2 Task: Combining Distributional, String-Based and Transformers-Based Features for Hypernymy Detection in the Financial Domain

Emmanuele Chersoni
emmanuelechersoni@gmail.com
The Hong Kong Polytechnic University
Hong Kong, China

Chu-Ren Huang
churen.huang@polyu.edu.hk
The Hong Kong Polytechnic University
Hong Kong, China

ABSTRACT

In this contribution, we describe the systems presented by the PolyU CBS Team at the second Shared Task on Learning Semantic Similarities for the Financial Domain (FinSim-2), where participating teams had to identify the right hypernyms for a list of target terms from the financial domain.

For this task, we ran our classification experiments with several distributional, string-based, and Transformer features. Our results show that a simple logistic regression classifier, when trained on a combination of word embeddings, semantic and string similarity metrics and BERT-derived probabilities, achieves a strong performance (above 90%) in financial hypernymy detection.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Lexical semantics**; **Language resources**.

KEYWORDS

Financial NLP, Hypernymy Detection, Distributional Models

ACM Reference Format:
Emmanuele Chersoni and Chu-Ren Huang. 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 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3442442.3451387

1 INTRODUCTION

The *hypernymy* relation, linking together a term and its superordinate concept (e.g. *cat-animal*), has always been of essential importance for the research in computational lexical semantics [37], given its role as the backbone of the structure of ontologies [5–7, 18] and in the recognition of entailment relations [11, 45].

With the increasing number of lexical networks and resources dedicated to specific textual domains, the research community consequently saw an increase in the interest in techniques for domain adaptation [14]. As Natural Language Processing technologies are more and more frequently used in accounting and finance [25], the FinSim competition has been recently introduced to test the capacity of these models to perform tasks of term categorization in the financial domain [13].

In the second edition of the shared task, organized by Fortia Financial Solutions [1] and co-located with the Web Conference 2021, the organizers provided an expanded list of financial terms, to be matched with a set of 10 candidate hypernyms [26]. The hypernyms correspond to 10 of the high-level concepts in the Financial Industry Business Ontology (FIBO). [2]

In the present paper, we introduce the systems developed by the PolyU-CBS team in the context of the FinSim2 shared task on hypernymy detection. Our top system, achieving an accuracy score of 90.6% accuracy on the shared task test set, is based on a combination of word embeddings, semantic and string similarity metrics and BERT-based features.

2 RELATED WORK

2.1 Hypernymy Detection in Computational Semantics

Like for other semantic relations, the earlier approaches for detecting hypernymy were based on patterns. Patterns are generally very precise for identification of hypernyms-hyponyms [16, 17, 20, 46], even when the task requires to discriminate between multiple relations at once [32], but they suffer from a limited recall, since the two target terms and the pattern have to appear in the same context.

Since mid-2000s, Distributional Semantic Models (DSMs) became a standard *de facto* in computational semantics: their main feature is to represent words as vectors, whose values are derived from the co-occurrence patterns of the words in a text corpus. The more similar the contexts in which the words appear, the more similar their meaning, and the similarity is typically assessed by means of vector cosine [21]. DSMs do not suffer from the recall problem of pattern-based methods, and thus they became the first choice for the research on semantic relations. However, they only provide a quantitative assessment of how similar two words are (a *similarity* score), without saying anything about their specific semantic relation [9]. To address this issue, several researchers focused on the similarity metric, proposing alternatives to cosine that can be more efficient in setting apart hypernyms from other semantically-related words [10, 22, 40, 48]. In parallel, research on semantic relations benefitted from the release of the large datasets for evaluating semantic relations, including hypernymy [3, 22, 23, 42, 43, 47].

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3451387

[1] https://www.fortia.fr/.
[2] https://spec.edmcouncil.org/fibo/

In a second phase, thanks to the public availability of easy-to-use packages for training word embeddings [4, 27, 33], the focus shifted on using these vectors as features for supervised classifiers [2, 39, 47], sometimes combining distributional information with pattern-based methods [36, 44], sometimes injecting in the vector space extra lexical knowledge from an external ontology [30, 31] (e.g. WordNet [28]). Shared tasks on the topic of the automatic identification of semantic relations have been regularly organized [49, 52], showing that it is often difficult to discriminate hypernymy from similar semantic relations, such as synonymy and co-hyponymy.

2.2 Hypernymy Detection for Ontology Creation

Hypernyms have always received a lot of attention also in taxonomy and ontology research: since they correspond to higher-level categories for target concepts, they play an important role in the organization of the terminology for a given domain. In the earlier shared task that were organized within the SemEval community [5, 6], hypernymy detection was treated as a binary classification task, meaning that, given a candidate pair of terms, a system had to predict whether a hypernym-hyponym relation was holding between them. In more recent shared tasks, the problem was reformulated as *hypernymy discovery* [8]: given a target term, systems have to find potential hypernyms in a domain-specific search space.

The recent effort of the FinSim shared task [13] proposed to identify hypernyms in the financial domain through a multiclass classification task: the target terms had to be assigned one out of eight (mutually exclusive) hypernymy labels, which corresponded to higher-level classes in the FIBO ontology. As an extra challenge, the dataset was extremely small (199 items in total), making it difficult to tackle the task with the modern deep learning approaches.

Among the six participating systems, the top scores were achieved by the IITK team [19], which presented a combination of a rule-based approach and of a Bernoulli Naive Bayes classifier applied to the word embeddings of the target terms. The former approach took advantage of the fact that several dataset terms contained the name of the right hypernym label (e.g. the hypernym of *Closed End Fund* is *Funds*), while the latter was based on Word2Vec embeddings [27] trained on a corpus of financial prospectuses.

3 EXPERIMENTAL SETTINGS

3.1 Datasets

The shared task organizers provided a training set and a test, respectively with 614 and 212 target terms (examples are shown in Table 1). The task consists of classifying the target terms into one out of ten hypernym classes: **Bonds, Credit Index, Equity Index, Forward, Funds, Future, MMIs, Options, Stocks, Swap**. The labels correspond to high-level classes in the FIBO ontology.

The data in the training set come with the gold standard labels, while those are absent in the test set. For running the evaluation of our systems, we used a 80:20 random split of the training data (491 and 123 terms), opting for a 10 runs Monte Carlo cross-validation. The reported results refer to the average values for the evaluation metrics over the 10 runs.

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

Table 1: Examples of term-hypernym pairs.

3.2 Metrics

The organizers provided scripts to evaluate the predictions in terms of *Accuracy* and *Mean Rank*. Notice that 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. Accuracy and Mean Rank are defined as follows:

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

$$MeanRank = \frac{1}{n} * \sum_{i=1}^{n} rank_i \tag{2}$$

Notice that $rank_i$ 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 [7].

3.3 Features

As baseline features (*base*), we used the 300 dimensions of **Word2Vec word embeddings** trained on a corpus of financial prospectuses. Those were provided by the shared task organizers. For terms composed by multiple words, we simply represented them by summing the vectors of the single words [29].

As a second set of features, we used **similarity metrics** (*sim*) between the vectors of the target terms and those of the hypernyms. Notice that for the multiword hypernyms, the vector representations have also been composed via vector sum. As similarity metrics, we chose *cosine similarity*, which is the standard metric in Distributional Semantics, and the *Spearman correlation* between the vectors, as recent studies on word embeddings have shown that cosine may be outperformed by rank-based metrics in several similarity estimation tasks [38, 41, 51]. For each target term, we computed these two similarity scores between the term and each one of the candidate hypernyms, and thus we extracted a total of 20 features for each dataset instance (2 features ∗ 10 classes).

Moreover, keeping in mind the "label inclusion" effect pointed out by Keswani et al. [19], we included some **string-based features** (*str*) to reflect surface-level similarities between terms and hypernyms. More specifically, we computed 1) the *Jaccard similarity*; 2) a boolean feature, equal to 1 if the hypernym was included in the target term and 0 otherwise. Since those features were computed for each possible term-hypernym pairing, we extracted a total of 20 features for each dataset instance (2 features ∗ 10 classes).

Finally, we also used the language modeling capabilities of the recently-introduced **Transformers** models to derive extra features. For each possible term-hypernym pairing, we generated a sentence with the form "The `term` is a `hypernym`", a typical co-occurrence pattern for word pairs in hypernymy-hyponymy relation. We call this sentence the *probe sentence*. We decided to engineer these features since recent studies showed that Transformers have interesting capabilities in associating nouns with their hypernyms, especially when the latter have to be picked from a closed set of possible answers [15, 35]. We tested with the *GPT-2* [34] and with the *BERT* [12] model, using respectively `GPT-2 Base` [3] and `BERT-large-cased`. [4]

For `GPT-2`, we extracted two different features per term-class pairing, for a total of 20 ($2 * 10$ features per instance): the *probability of the whole probe sentence*, and the *probability of the last token of the sentence*. For `BERT`, the procedure was more complex: given a probe sentence with a given term-hypernym pair, we replace each word composing the hypernym label with a *[MASK]* token, and we use the masked language modeling functionality of BERT to assign a probability to the word under the mask. As BERT-based features, we then take the *average of the token probabilities* and the *maximum probability score*. For example, for the probe sentence *The S%P 100 Index is an Equity Index*, we generate two masked sentences:

- *The S%P 100 Index is an [MASK] Index.*
- *The S%P 100 Index is an Equity [MASK].*

We took the average and the maximum of the probability values for each label, for a total of 20 BERT-based features ($2 * 10$).

3.4 Classification Models

As baseline models, we fed the 300 dimensions of the embedding of the term vector to the following classifiers (all in the standard `scikit-learn` implementation):

- Gaussian Naive Bayes (*NB*)
- Logistic Regression (*LR*)
- Random Forest Classifier (*RF*)

We gradually augmented the models by adding the features by groups and computed the scores for all the combinations over 10 runs with Monte Carlo cross-validation. The only exception were the Transformer-based features, since we do not have access to powerful computational resources and the processing times would have been too long. Therefore, Transformer features (either BERT or GPT-2 ones) have been added only on the top of the best performing model with the other feature groups.

4 RESULTS

Classifier	Baseline	+str	+sim	+all
NB	0.80/1.38	0.84/1.31	0.82/1.33	0.85/1.25
LR	0.86/1.29	0.92/1.12	0.93/1.11	**0.93/1.10**
RF	0.85/1.28	0.85/1.28	0.87/1.23	0.88/1.19

Table 2: Accuracy/mean rank scores for all classifiers with baseline, similarity (*sim*) and string-based features (*str*) and their combination (*all*).

Table 2 shows the results for the models with all the groups of features, except for the Transformers. We can observe that the Logistic Regression (*LR*) classifier, while being close to the others with the baseline features, gradually improved by adding the other groups of features and had performances stably above 0.90 for the accuracy. The *str* and the *sim* both led to improvements and the model using all of them ($LR + all$) turned out to be the best one, although the advantage on the models using only one of the two groups was marginal and only for the Mean Rank metric. Thus, we decided to use this model in combination with the Transformers.

Transformer	Accuracy/Mean Rank
GPT-2	0.90/1.15
BERT	**0.93/1.11**

Table 3: Accuracy/mean rank scores of the best logistic regression classifier augmented with Transformers features.

Transformer	Accuracy/Mean Rank
PolyU-CBS_1	**0.90/1.19**
PolyU-CBS_2	0.89/1.20

Table 4: Accuracy/mean rank scores of the two submitted systems on the shared task test set.

Table 3 shows the performance for the $LR + all$ model augmented with the Transformer features. The Transformers did not seem to improve the general classification accuracy, which was three points lower with GPT-2, while the model with BERT had just a slight increase of the Mean Rank. Given that the model with BERT in Table 3 and the model $LR + all$ in Table 2 were the best performing ones, we submitted their results for the shared task, respectively, as PolyU-CBS_1 and PolyU-CBS_2. On the test set, as it can be seen in Table 4, the BERT features finally gave the model a slight edge over the competitor. This model was also the one that achieved the top score in the FinSim-2 shared task [26].

5 CONCLUSIONS

In this paper, we presented the PolyU-CBS systems that participated into the FinSim-2 shared task on hypernymy detection for the financial domain. Among the features that we tested, both semantic and string similarity features led to important gains over the baseline. On the one hand, the importance of surface string similarities for this task had already been suggested by the winners of the previous edition [19], and on the other hand semantic metrics have a long history as valuable predictors of the hypernymy relation [37].

Concerning the Transformers, we engineered some features inspired by the recent NLP literature on probing tasks. We found only marginal improvements over the systems not using them, but in the end BERT probabilities still gave the model a slight edge over competitors on the test set, confirming that these neural architectures can contribute useful information for hypernymy detection in settings with a close set of labels [15, 35].

We think that these results are extremely promising, especially considering that we did not make use of any domain-specific corpus or resource, with the only exception of the baseline embeddings. In future work, using more specialized models might lead to further improvements. For example, the capacity of the recently-introduced BERT models for finance [1, 24, 50] to model taxonomical relations is, to the best of our knowledge, yet to be explored.

[3] We used the implementation of the `lm-scorer` library.
[4] We used the implementation of the `happy-transformers` library.

REFERENCES

[1] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063* (2019).

[2] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment Above the Word Level in Distributional Semantics. In *Proceedings of EACL*.

[3] Marco Baroni and Alessandro Lenci. 2011. How We BLESSed Distributional Semantic Evaluation. In *Proceedings of the ACL Workshop on GEometrical Models of Natural Language Semantics*.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[5] Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 Task 17: Taxonomy Extraction Evaluation (texeval). In *Proceedings of SemEval*.

[6] Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 Task 13: Taxonomy Extraction Evaluation (Texeval-2). In *Proceedings of SemEval*.

[7] 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 SemEval*.

[8] Jose Camacho-Collados and Roberto Navigli. 2017. BabelDomains: Large-scale Domain Labeling of Lexical Resources. In *Proceedings of EACL*.

[9] Emmanuele Chersoni, Giulia Rambelli, and Enrico Santus. 2016. CogALex-V Shared Task: ROOT18. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

[10] Daoud Clarke. 2009. Context-Theoretic Semantics for Natural Language: An Overview. In *Proceedings of the EACL workshop on Geometrical Models of Natural Language Semantics*.

[11] Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies* 6, 4 (2013), 1–220.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

[13] Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. 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*.

[14] Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. Supervised Distributional Hypernym Discovery Via Domain Adaptation. In *Proceedings of EMNLP*.

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

[16] Gregory Grefenstette. 2015. INRIASAC: Simple Hypernym Extraction Methods. *arXiv preprint arXiv:1502.01271* (2015).

[17] Marti A Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of COLING*.

[18] David Jurgens and Mohammad Taher Pilehvar. 2016. SemEval-2016 Task 14: Semantic Taxonomy Enrichment. In *Proceedings of SemEval*.

[19] Vishal Keswani, Sakshi Singh, and Ashutosh Modi. 2021. IITK at the FinSim Task: Hypernym Detection in Financial Domain via Context-Free and Contextualized Word Embeddings. *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

[20] Els Lefever, Marjan Van de Kauter, and Véronique Hoste. 2014. HypoTerm: Detection of Hypernym Relations Between Domain-Specific Terms in Dutch and English. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 20, 2 (2014), 250–278.

[21] Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual review of Linguistics* 4 (2018), 151–171.

[22] Alessandro Lenci and Giulia Benotto. 2012. Identifying Hypernyms in Distributional Semantic Spaces. In *Proceedings of *SEM*.

[23] Hongchao Liu, Emmanuele Chersoni, Natalia Klyueva, Enrico Santus, and Chu-Ren Huang. 2019. Semantic Relata for the Evaluation of Distributional Models in Mandarin Chinese. *IEEE access* 7 (2019), 145705–145713.

[24] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. In *Proceedings of IJCAI*.

[25] Tim Loughran and Bill McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research* 54, 4 (2016), 1187–1230.

[26] Youness Mansar, Kang Juyeon, and Ismail El Maarouf. 2021. FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain. In *Proceedings of The Web Conference*.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781* (2013).

[28] George A Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[29] Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science* 34, 8 (2010), 1388–1429.

[30] Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical Embeddings for Hypernymy Detection and Directionality. *Proceedings of EMNLP* (2017).

[31] 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 ACL*.

[32] Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Demantic Relations. In *Proceedings of COLING-ACL*.

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI blog* 1, 8 (2019), 9.

[35] Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT. In *Proceedings of *SEM*.

[36] Stephen Roller and Katrin Erk. 2016. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In *Proceedings of EMNLP*.

[37] Enrico Santus. 2016. *Making Sense: From Word Distribution to Meaning*. Ph.D. Dissertation. Hong Kong Polytechnic University.

[38] Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. 2016. Testing APSyn Against Vector Cosine on Similarity Estimation. (2016).

[39] Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine Features in a Random Forest to Learn Taxonomical Semantic Relations. In *Proceedings of LREC*.

[40] Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. 2014. Chasing Hypernyms in Vector Spaces with Entropy. In *Proceedings of EACL*.

[41] Enrico Santus, Hongmin Wang, Emmanuele Chersoni, and Yue Zhang. 2018. A Rank-Based Similarity Metric for Word Embeddings. In *Proceedings of ACL*.

[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 ACL Workshop on Linked Data in Linguistics: Resources and Applications*.

[43] Silke Scheible and Sabine Schulte Im Walde. 2014. A Database of Paradigmatic Semantic Relation Pairs for German Nouns, Verbs, and Adjectives. In *Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*.

[44] Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving Hypernymy Detection with an Integrated Path-Based and Distributional Method. In *Proceedings of ACL*.

[45] Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms Under Siege: Linguistically-Motivated Artillery for Hypernymy Detection. In *Proceedings of EACL*.

[46] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning Syntactic Patterns for Automatic Hypernym Discovery. In *Proceedings of NIPS*.

[47] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to Distinguish Hypernyms and Co-hyponyms. In *Proceedings of COLING*.

[48] Julie Weeds and David Weir. 2003. A general Framework for Distributional Similarity. In *Proceedings of EMNLP*.

[49] Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. 2020. The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*.

[50] Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097* (2020).

[51] Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Y Hammerla. 2019. Correlation Coefficients and Semantic Textual Similarity. In *Proceedings of NAACL*.

[52] Michael Zock, Alessandro Lenci, and Stefan Evert. 2016. Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*.