Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5708–5713
Marseille, 11–16 May 2020

Are Word Embeddings Really a Bad Fit for the Estimation of Thematic Fit?

Emmanuele Chersoni[1], Ludovica Pannitto[2], Enrico Santus[3], Alessandro Lenci[4], Chu-Ren Huang[1]
The Hong Kong Polytechnic University [1], University of Trento[2],
Massachussetts Institute of Technology[3], University of Pisa[4]
Chinese and Bilingual Studies, 11 Yuk Choi Road, Hung Hom, Hong Kong (China)[1]
Centre for Mind/Brain Sciences, Corso Bettini 31, Rovereto (Italy)[2]
MIT CSAIL, 32 Vassar Street, Cambridge, MA (United States)[3]
University of Pisa, Via Santa Maria 36, 56126 Pisa (Italy)[4]
{emmanuele.chersoni,churen.huang}@polyu.edu.hk[1], ludovica.pannitto@unitn.it[2],
esantus@mit.edu[3], alessandro.lenci@unipi.it[4]

Abstract

While neural embeddings represent a popular choice for word representation in a wide variety of NLP tasks, their usage for *thematic fit modeling* has been limited, as they have been reported to lag behind syntax-based count models. In this paper, we propose a complete evaluation of count models and word embeddings on thematic fit estimation, by taking into account a larger number of parameters and verb roles and introducing also dependency-based embeddings in the comparison. Our results show a complex scenario, where a determinant factor for the performance seems to be the availability to the model of reliable syntactic information for building the distributional representations of the roles.

Keywords: Semantics, Cognitive Methods, Statistical and Machine Learning Methods

1. Introduction

In recent years, vectors derived from neural network training have quickly replaced the old, count-based Distributional Semantic Models (DSMs) as a *de facto* standard for word representation in NLP.[1] Tools such as Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) have provided the research community with an efficient and scalable method for training vector representations, generally referred to as *word embeddings*. Moreover, the embeddings have been reported to have an advantage over the old count models also in terms of performance in several NLP tasks (Baroni et al., 2014).[2]

In this scenario, thematic fit estimation represents an exception. Concretely, the task consists in estimating a typicality score for a filler noun given a verb role (e.g., a system has to predict how plausible a *cake* is as a patient of the verb *to eat*). It is generally evaluated by assessing the correlation between collections of human judgements and DSM outputs, and it represents an important benchmark for the capacity of the models of capturing compositional meaning (Lenci, 2018).

In a systematic comparison between count-based models and neural embeddings, (Baroni et al., 2014) showed that the latter outperform the former in almost all the evaluation tasks, while in the thematic fit task they were vastly outperformed by dependency-based count models. The results by Baroni and colleagues were later confirmed by (Sayeed et al., 2016), who reported that the Word2Vec embeddings were largely outperformed by many of the previous thematic fit models based on count vectors. Despite the progress made in the recent literature and the introduction of several new architectures and improvements, the studies following the first evaluation attempts have only focused on count models.

In the present contribution, we propose a more systematic comparison between embeddings and count-based models on thematic fit estimation. Compared to earlier evaluations, which only tested CBOW vectors on agents and patients datasets, we evaluate both the Word2Vec architectures on a wider variety of roles, as well as the dependency-based word embeddings by (Levy and Goldberg, 2014). Additionally, since the best thematic fit models make use of syntactic information to build 'prototypical' representations of the verb roles, we test the importance of such information for the model performance.

2. Related Work

According to a long tradition of psycholinguistic studies, human semantic memory stores a generalized knowledge about events and their participants (McRae et al., 1998; McRae et al., 2005; Hare et al., 2009). The typicality of the combinations of verbs and arguments has important consequences for sentence processing, as typical combinations require less effort from human comprehenders (Bicknell et al., 2010; Matsuki et al., 2011). The thematic fit can be defined as the degree of plausibility of a noun filler for a given verb role (e.g. nouns like *pizza, cake, ice cream* would all have a high thematic fit value for the patient role ot *to eat*).[3] Because of this relationship between thematic fit and sen-

[1]Throughout the entire paper, we will refer to the traditional distributional vectors based on co-occurrence counts as *count models* (Baroni et al., 2014).

[2]See however also the research works by (Levy et al., 2015; Lebret and Collobert, 2015; Gamallo, 2017), where it is shown how count-based models can recover from the deficit by means of hyperparameter optimization.

[3]In this paper, the usage of the expression *thematic fit* has always to be interpreted as related to the integration of event-

tence processing, several researchers in computational semantics have tried to model this phenomenon, mostly using syntax-based DSMs (Erk et al., 2010; Baroni and Lenci, 2010; Sayeed et al., 2015; Greenberg et al., 2015; Santus et al., 2017). Notice that this research trend developed in parallel with the one aiming at automatically acquiring selectional preferences (Resnik, 1997; Zhang et al., 2019; Zhang et al., 2020), which has mostly been seen as an auxiliary task for improving the performance of systems with different goals, such as semantic role classification (Collobert et al., 2011; Zapirain et al., 2013; Roth and Lapata, 2015) or coreference resolution (Heinzerling et al., 2017). Moreover, although the notions of selectional preference and thematic fit are closely related, the nature of the involved elements is different: discrete semantic types in the former case, gradient compatibility between arguments and thematic roles in the latter one (Lebani and Lenci, 2018).

In the literature using DSMs for modeling thematic fit, the method by (Baroni and Lenci, 2010) turned out to be particularly influential. Given a verb role, this study made use of a corresponding syntactic relation (e.g., the subject for the agent) to extract its typical fillers. The vectors of the typical fillers were then summed to create distributional representations of the prototypical fillers, and the thematic fit of a noun for a role was finally assessed as the cosine similarity between its filler vector and the role prototype.

While word embeddings were taking distributional semantics by storm (Mikolov et al., 2013a; Mikolov et al., 2013b; Pennington et al., 2014; Bojanowski et al., 2017), it is surprising that, after the early studies, such vector representations have not been tested anymore on thematic fit. Moreover, the few works were carried out only with the original Word2Vec embeddings, trained on window-based contexts, and were limited to the Continuous-Bag-of-Words (CBOW) tested on two datasets, in which only the agent and the patient roles are represented. To the best of our knowledge, Skip-Gram vectors have never been tested on thematic fit estimation. Finally, the embeddings trained on syntactic dependencies by (Levy and Goldberg, 2014) proved to be efficient in modeling the *functional similarity* between words that tend to have the same function or structural role in a sentence, and thus they seem good candidates to perform well in the task. [4]

Here, we present a complete evaluation of the above-mentioned models on datasets including *agents, patients, instruments and locations*. Moreover, given the recent claims that incorporating syntax in DSMs does not lead to significant improvements (Lapesa and Evert, 2017), we pay specific attention to a question not addressed yet in the literature: *how essential is syntactic information for building good-quality semantic role prototypes?*

3. Experimental Settings

Datasets. We tested our models on three standard datasets derived from (McRae et al., 1998), (Ferretti et al., 2001) and (Padó, 2007), containing plausibility judgments for verb role-filler pairs. McRae and Padó include, respectively, 1,444 and 414 scores for agents and patients (e.g., *doctor-advise* and *hit-ball*), whereas Ferretti includes judgements for 274 instruments and 248 locations (e.g., *cut-mower* and *teach-classroom*). The scores range from 1 (atypical) to 7 (very typical) and their distribution per role can be observed in Figures 1, 2 and 3.

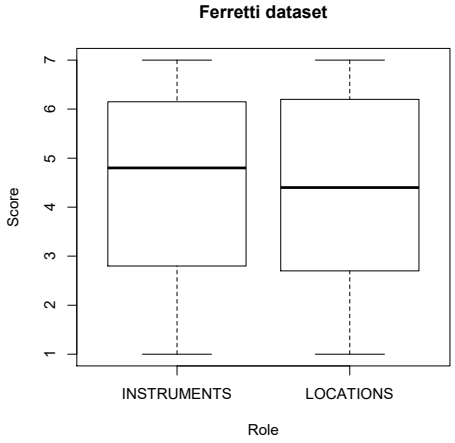

Figure 1: Scores distribution per role in the Ferretti dataset.

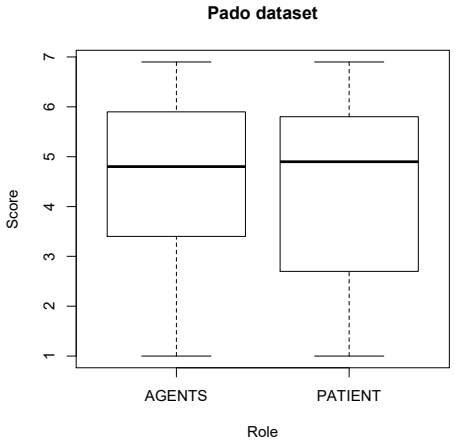

Figure 2: Scores distribution per role in the Padó dataset.

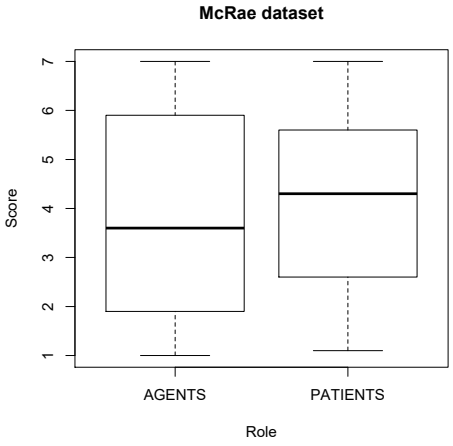

Figure 3: Scores distribution per role in the McRae dataset.

specific world knowledge in sentence comprehension, as in the psycholinguistic literature of reference.

[4] See also (Turney, 2012) for the distinction between *functional similarity* and *domain similarity*.

Verb and Role	Fillers
Agent of *to play*	actor, gamer, violinist
Patient of *to eat*	pizza, sandwich, ice-cream
Instrument of *to cut*	knife, axe, scissors
Location of *to swim*	sea, pool, ocean

Table 1: Verb roles and examples of extracted fillers.

Training Corpora The DSMs have been trained on a concatenation of the BNC (Leech, 1992), the Ukwac corpus (Baroni et al., 2009) and a 2018 dump of the English Wikipedia.

Count-Based Models Our count models are a standard dependency-based DSM (**DEPS**) and Distributional Memory (**DM**). In **DEPS**, the targets and contexts are the 20K most frequent nouns and verbs in the training corpus. Contexts are typed with the dependency link with the target (e.g., the nouns *chef:SUBJ* and *pizza:DOBJ* are contexts for the verb *to cook*). DM (Baroni and Lenci, 2010) is also a dependency-based DSM, but it is enriched with hand-selected lexico-syntactic patterns, and it was shown to achieve extremely high performances in the task (Santus et al., 2017). For each relation, both models include also its inverse (e.g., the verb *cook:DOBJ*$^{-1}$ is a context for the target noun *pizza*).

Word Embeddings As for the *bag-of-words embeddings*, we used both the standard Word2Vec architectures, the Skip-Gram (**SG**) and the Continuous-Bag-of-Words (**CBOW**). Both models have been trained with negative sampling and standard hyperparameters.[5] The *dependency-based embeddings* (**LG-DEPS**) by (Levy and Goldberg, 2014) are based on the Skip-Gram architecture and trained on the syntactic dependencies extracted from the same corpora. This model has also been trained with standard hyperparameters.[6] We tried several settings for the embeddings dimensionality ($d = 100, 200, 300, 400$): we found no significant differences, although higher-dimensional vectors seem to have a slight advantage. In the Results section, all the reported scores are for $d = 300$. The coverage for models and datasets is reported in Table 2. [7]

DSM	Padó	McRae	Instruments	Locations
DEPS	97	95.6	94.6	96.7
DM	100	95	93.9	95.7
SG	99.5	100	100	96.4
CBOW	99.5	100	100	96.4
LG-DEPS	98	96.5	93.6	95.7

Table 2: Model coverage in percentage for each dataset (the total number of rated pairs in brackets).

Role Prototypes As in (Baroni and Lenci, 2010), we extract typical fillers for each verb role and sum them to create the representation of a prototypical role filler. The idea is that the higher the similarity of a filler with a role prototype, the better it will be fitting the role. In order to test the importance of syntactic information for modeling the roles, we compared two different methods of filler selection.

The first method is the classical one, based on syntactic relations. We approximated the agents with the subjects, the patients with the direct object, the instruments with prepositional complements introduced by *with* and the locations with prepositional complements introduced by either *at, on* or *in*. For each word of the datasets, a set of fillers was extracted from DM (**DM-F**) and another set from DEPS (**DEPS-F**).

The second method uses the nouns co-occurring in a window of size w and ranked by association score, independently of the syntactic relation (**BOW-F**). We tested with windows of different size and report the results for $w = 2$, which gave us the best results overall. The performance of BOW-F models is of particular interest: if they can efficiently model thematic fit judgements for verb roles, then the claim by (Lapesa and Evert, 2017) on the non-necessarity of syntactic information will be confirmed also for this task.

Fillers were assigned a typicality score based on Local Mutual Information (LMI) (Evert, 2004), as in the evaluation by Baroni and colleagues. For the *BOW-F* set, the score is computed with Eq. 1, while for the syntax-based sets is based on Eq. 2.

$$LMI(v, f) = log\left(\frac{O_{v,f}}{E_{v,f}}\right) * O_{v,f} \quad (1)$$

$$LMI(v, r, f) = log\left(\frac{O_{v,r,f}}{E_{v,r,f}}\right) * O_{v,r,f} \quad (2)$$

$O_{v,f}$ and $E_{v,f}$ are respectively the co-occurrence count between a verb v and a filler f and the expected count under independence (the formula is the same for the syntactic method, but adding the third element of the syntactic relation r). For each DSM and filler type, the vectors of the top scoring k fillers for each role were summed to build the prototypes. After testing with $k = 10, 20, 30, 40, 50$, we observed that the number of fillers did not significantly affect the performance, coherently with the findings of (Greenberg et al., 2015). The reported results have been obtained with $k = 20$, as in the works by (Baroni et al., 2014) and (Sayeed et al., 2016). The 5 DSMs have been evaluated with all 3 filler types, making 15 different models. Finally, we measured the cosine similarity between roles prototypes and fillers in the datasets, and we computed the Spearman correlation between scores and human judgements.

4. Results and Discussion

The scores for agents and patients datasets and those for instruments and locations follow quite different patterns. In Table 3, DM turns out to be by far the best model on both datasets (the margin is significant at $p < 0.05$ on Padó's data)[8], and all models based on DM fillers perform better.

[5] Learning rate from 0.025 to 0.0001, 0.001 as downsampling threshold, 5 negative samples and 10 as window size.

[6] 100 as a frequency threshold for words and contexts, 15 negative samples (the other parameters are the same of the BOW counterparts).

[7] The small differences in coverage do not significantly affect the results presented in Section 4.

[8] p-values computed with Fisher's r-to-z transformation.

	Padó			McRae		
	DM-F	BOW-F	DEPS-F	DM-F	BOW-F	DEPS-F
SG	0.372	0.161	0.313	0.287	0.144	0.232
CBOW	0.317	0.134	0,248	0.270	0.107	0.214
LG-DEPS	0.345	ns	0.272	0.318	0.090	0.251
DM	**0.500**	0.251	0.462	**0.342**	0.112	0.293
DEPS	0.383	0.218	0.388	0.278	0.056	0.247

Table 3: Spearman correlations for the Padó and McRae dataset for all models with all filler sets.

	Instruments			Locations		
	DM-F	BOW-F	DEPS-F	DM-F	BOW-F	DEPS-F
SG	**0.425**	0.341	0.433	0.336	**0.379**	0.333
CBOW	0.376	0.301	0.376	0.335	0.365	0.317
LG-DEPS	0.316	0.219	0.313	0.262	0.247	0.248
DM	0.368	0.166	0.341	0.225	0.126	0.203
DEPS	0.301	0.157	0.249	0.200	0.168	0.261

Table 4: Spearman correlations for the Instruments and Locations dataset for all models with all filler sets.

This is not surprising: DM is a carefully crafted syntactic DSM, and the addition of lexical syntactic patterns has been hypothesized to have a positive impact in this task (Sayeed et al., 2015).

However, the less-refined DEPS model perform similarly to the SG embeddings, which in turn perform always better than the CBOW ones. The performance of LG-DEPS is also close to the SG one: syntactic dependencies, as suggested by some recent contributions in the literature (Li et al., 2017; Lapesa and Evert, 2017), do not improve model performance. As for the fillers, syntax seems instead to play an important role: if we compare models with BOW-F fillers with those making use of the "syntactic" sets, we observe large and significant drops for every model on both datasets, to the point that many correlations on McRae become non-significant. DM fillers are clearly better than the DEPS one, suggesting that syntactic information is more useful to select typical contexts.

The results for the other roles (Table 4) instead show a clear advantage of the BOW embeddings over count-based models. SG embeddings are again the best, followed by the CBOW ones, and LG-DEPS vectors, unexpectedly, are again lagging behind their BOW counterparts. Disappointing results for dependency-based embeddings have also been reported by (Gamallo, 2017), in comparison to syntactic count-based vectors. A possible cause, as suggested by (Asr et al., 2016; Sahlgren and Lenci, 2016), could be found in the fact that embedding models are suboptimal when trained on smaller data sizes, and this could be especially true with sparse dependency contexts. It is also interesting to observe that, with instruments and locations, dropping syntactically-selected fillers does not always cause huge correlation drops for the embedding models, and in some cases it even leads to improvements (cf. the SG and CBOW scores for Locations). Probably, using prepositions to select the fillers turns out to be a rough approximation, and the prototypes are so noisy that there are no gains with respect to an extraction based on word windows.

The Location role is the most challenging for most models, and this fact can be related to some of the results in the literature. In a single-word priming experiment, (Ferretti et al., 2001) found that verbs activate knowledge of typical agents, patients and instruments, but not locations. The explanation proposed by the authors was that location information is not salient in an event description, unless the event is described as ongoing (i.e. the verb is in a progressive tense). Indeed, a following study by (Ferretti et al., 2007) manipulated verb aspect and found that locations were primed by verbs only when the aspect was imperfective (e.g. priming was obtained for *was skating - arena* but not for *had skated - arena*). Since the role of the event temporal structure has not been explored so far by DSMs of event knowledge, it might be interesting for future models to try to incorporate this kind of information.

5. Conclusions

In this paper, we proposed a new, thorough evaluation of word embeddings on the thematic fit task. Contrary to previous findings, our results show that the performance of the models depends on two main factors: the verb roles to be modeled and the filler selection method to build the role prototypes. For agents and patients, having clean, syntactically-selected fillers is a pro, while for other roles syntactic information is probably too noisy. [9] SG embeddings achieved solid results on the task, performing similarly to a standard dependency-based model on the Padó and McRae datasets (it only lags behind the carefully-crafted DM), and outperforming all count-based competitors on the Ferretti ones, even without syntactic fillers. Our scores were obtained simply by training the model with standard parameters and with no refined context selection. Thus, we conclude that word embeddings are not always a bad fit for the thematic fit task.

Given the growing interest for the psychological plausibility of word embeddings and for their performance on cognitively-motivated benchmarks (Søgaard, 2016; Mandera et al., 2017; Bakarov, 2018; Schwartz and Mitchell,

[9] It should be noticed that state-of-the-art neural systems for this task are trained on semantic role labels (Tilk et al., 2016; Hong et al., 2018), and thus they avoid -at least in theory- the problem of dealing with the ambiguity of the prepositions.

2019; Hollenstein et al., 2019), future experiments might add thematic fit estimation to the set of tasks in which they could be tested, by carefully taking into account the impact of factors such as the size of training data and the linguistic information available to the models. Another possible direction of work could aim at adapting the recently-introduced contextualized embeddings (Radford et al., 2018; Peters et al., 2018; Devlin et al., 2019) to the task.

6. Acknowledgements

We would like to thank the three anonymous reviewers for the positive feedback and for their useful comments.

7. Bibliographical References

Asr, F. T., Willits, J., and Jones, M. N. (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. In *Proceedings of CogSci*.

Bakarov, A. (2018). Can Eye Movement Data Be Used As Ground Truth For Word Embeddings Evaluation? In *Proceedings of the LREC Workshop on Linguistic and Neurocognitive Resources*.

Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In *Proceedings of ACL*.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of Event Knowledge in Processing Verbal Arguments. *Journal of Memory and Language*, 63(4):489–505.

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.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

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*.

Erk, K., Padó, S., and Padó, U. (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating Verbs, Situation Schemas, and Thematic Role Concepts . *Journal of Memory and Language*, 44(4):516 – 547.

Ferretti, T. R., Kutas, M., and McRae, K. (2007). Verb Aspect and the Activation of Event Knowledge. *Journal of Experimental Psychology: Learning, memory, and cognition*, 33(1):182.

Gamallo, P. (2017). Comparing Explicit and Predictive Distributional Semantic Models Endowed with Syntactic Contexts. *Language Resources and Evaluation*, 51(3):727–743.

Greenberg, C., Sayeed, A. B., and Demberg, V. (2015). Improving Unsupervised Vector-space Thematic Fit Evaluation via Role-filler Prototype Clustering. In *Proceedings of HLT-NAACL*.

Hare, M., Jones, M., Thomson, C., Kelly, S., and McRae, K. (2009). Activating Event Knowledge. *Cognition*, 111 2:151–67.

Heinzerling, B., Moosavi, N. S., and Strube, M. (2017). Revisiting Selectional Preferences for Coreference Resolution. In *Proceedings of EMNLP*.

Hollenstein, N., de la Torre, A., Langer, N., and Zhang, C. (2019). CogniVal: A Framework for Cognitive Word Embedding Evaluation. In *Proceedings of CONLL*.

Hong, X., Sayeed, A., and Demberg, V. (2018). Learning Distributed Event Representations with a Multi-Task Approach. In *Proceedings of *SEM*.

Lapesa, G. and Evert, S. (2017). Large-scale Evaluation of Dependency-Based DSMs: Are They Worth the Effort? In *Proceedings of EACL*.

Lebani, G. E. and Lenci, A. (2018). A Distributional Model of Verb-Specific Semantic Roles Inferences. *Language, Cognition, and Computational Models*, pages 118–158.

Lebret, R. and Collobert, R. (2015). Rehabilitation of Count-Based Models for Word Vector Representations. In *Proceedings of CICLING*. Springer.

Leech, G. (1992). 100 Million Words of English: The British National Corpus (BNC). *Language Research*, 28(1):1–13.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual review of Linguistics*, 4:151–171.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of EMNLP*.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Li, B., Liu, T., Zhao, Z., Tang, B., Drozd, A., Rogers, A., and Du, X. (2017). Investigating Different Syntactic Context Types and Context Representations for Learning Word Embeddings. In *Proceedings of EMNLP*.

Mandera, P., Keuleers, E., and Brysbaert, M. (2017). Explaining Human Performance in Psycholinguistic Tasks with Models of Semantic Similarity Based on Prediction and Counting: A Review and Empirical Validation. *Journal of Memory and Language*, 92:57–78.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., and McRae, K. (2011). Event-based Plausibility Immediately Influences On-line Language Comprehension. *Journal of experimental psychology. Learning, memory, and cognition*, 37 4:913–34.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, 38:283–312.

McRae, K., Hare, M., Elman, J. L., and Ferretti, T. (2005). A Basis for Generating Expectancies for Verbs from Nouns. *Memory & Cognition*, 33(7):1174–1184.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119.

Padó, U. (2007). *The Integration of Syntax and Semantic Plausibility in a Wide-coverage Model of Human Sentence Processing*. Ph.D. thesis.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of NAACL*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Resnik, P. (1997). Selectional Preference and Sense Disambiguation. In *Tagging Text with Lexical Semantics: Why, What, and How?*

Roth, M. and Lapata, M. (2015). Context-Aware Frame-Semantic Role Labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Sahlgren, M. and Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of EMNLP*.

Santus, E., Chersoni, E., Lenci, A., and Blache, P. (2017). Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.

Sayeed, A., Demberg, V., and Shkadzko, P. (2015). An Exploration of Semantic Features in an Unsupervised Thematic Fit Evaluation Framework. In *Italian Journal of Linguistics*.

Sayeed, A., Greenberg, C., and Demberg, V. (2016). Thematic Fit Evaluation: An Aspect of Selectional Preferences. In *Proceedings of the ACL Workshop for Evaluating Vector Space Representations for NLP*.

Schwartz, D. and Mitchell, T. (2019). Understanding Language-Elicited EEG Data by Predicting It from a Fine-Tuned Language Model. In *Proceedings of NAACL*.

Søgaard, A. (2016). Evaluating Word Embeddings with fMRI and Eye-Tracking. In *Proceedings of the ACL Workshop on Evaluating Vector-Space Representations for NLP*.

Tilk, O., Demberg, V., Sayeed, A. B., Klakow, D., and Thater, S. (2016). Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.

Turney, P. D. (2012). Domain and Function: A Dual-Space Model of Semantic Relations and ompositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Zapirain, B., Agirre, E., Marquez, L., and Surdeanu, M. (2013). Selectional Preferences for Semantic Role Classification. *Computational Linguistics*, 39(3):631–663.

Zhang, H., Ding, H., and Song, Y. (2019). SP-10K: A Large-Scale Evaluation Set for Selectional Preference Acquisition. In *Proceedings of ACL*.

Zhang, H., Bai, J., Song, Y., Xu, K., Yu, C., Song, Y., Ng, W., and Yu, D. (2020). Multiplex Word Embeddings for Selectional Preference Acquisition. *arXiv preprint arXiv:2001.02836*.