Is "Universal Syntax" Universally Useful for Learning Distributed Word Representations?

Ivan Vulić and **Anna Korhonen**
Language Technology Lab
DTAL, University of Cambridge
`{iv250, alk23}@cam.ac.uk`

Abstract

Recent comparative studies have demonstrated the usefulness of dependency-based contexts (DEPS) for learning distributed word representations for similarity tasks. In English, DEPS tend to perform better than the more common, less informed bag-of-words contexts (BOW). In this paper, we present the first cross-linguistic comparison of different context types for three different languages. DEPS are extracted from "universal parses" without any language-specific optimization. Our results suggest that the universal DEPS (UDEPS) are useful for detecting functional similarity (e.g., verb similarity, solving syntactic analogies) among languages, but their advantage over BOW is not as prominent as previously reported on English. We also show that simple "post-parsing" filtering of useful UDEPS contexts leads to consistent improvements across languages.

1 Introduction

Dense real-valued distributed representations of words known as word embeddings (WEs) have become ubiquitous in NLP, serving as invaluable features in a broad range of NLP tasks, e.g., (Turian et al., 2010; Collobert et al., 2011; Chen and Manning, 2014). The omnipresent `word2vec` skip-gram model with negative sampling (SGNS) (Mikolov et al., 2013b) is still considered the state-of-the-art word representation model, due to its simplicity, fast training, as well as its solid and robust performance across a wide variety of semantic tasks (Baroni et al., 2014; Levy et al., 2015).

The original implementation of SGNS learns word representations from local bag-of-words contexts (BOW). However, the underlying SGNS model is equally applicable to other context types.

Recent comparative studies have demonstrated the usefulness of *dependency-based contexts (DEPS)* (Padó and Lapata, 2007) for the task. In comparison with BOW, syntactic contexts steer the induced semantic spaces towards functional similarity (e.g., *tiger:cat*) rather than towards topical similarity/relatedness (e.g., *tiger:jungle*). DEPS-based embeddings outperform the less informed BOW-based embeddings in a variety of similarity tasks (Bansal et al., 2014; Levy and Goldberg, 2014a; Hill et al., 2015; Melamud et al., 2016). However, these studies have all focused solely on English. A comparison extending to additional languages is required before any cross-lingual generalisations can be drawn.

Following recent initiatives on language-agnostic and cross-linguistically consistent *universal natural language processing* (i.e., universal POS (UPOS) tagging and dependency (UD) parsing) (Nivre et al., 2015), this paper is concerned with two important questions:

(Q1) Can one usefully replace the DEPS extraction pipeline optimised for tools developed for English with a pipeline that relies on language-universal syntactic processing (UDEPS)?

(Q2) Are UDEPS universally better than BOW for learning distributed word representations in other languages?

Regarding Q1, the results show that it is possible to replace original DEPS with UDEPS for English and to obtain benchmarking results with only a slight drop in performance. As for Q2, the framework is not equally effective in other languages, as suggested by the performance in Italian and German, which sheds new light on the usefulness of BOW and dependency-based contexts. Further, the results reveal that even a simple preliminary "post-parsing" selection of useful UDEPS contexts leads to consistent improvements across languages, especially in detecting functional similarity.

This focused contribution is the first cross-linguistic comparison of different context types for learning word representations in three languages, reaching beyond English. It also constitutes a first completely language-universal and widely applicable framework for UDEPS extraction.

2 Methodology

Universal Multilingual Resources The departure point in our experiments is the Universal Dependencies project (McDonald et al., 2013; Nivre et al., 2015) which develops cross-linguistically consistent treebank annotation.[1] The annotation scheme leans on the universal Stanford dependencies (de Marneffe et al., 2014) complemented with the Google universal POS tagset (Petrov et al., 2012) and the Interset interlingua for morphological tagsets (Zeman and Resnik, 2008). It provides a universal and consistent inventory of categories for similar syntactic constructions across languages.

The main aim of the "universal initiative" is to facilitate cross-lingual and multilingual learning (e.g., multilingual parser development, typologies) by capturing structural similarities across languages and by exploiting connections that exist naturally between them (Berg-Kirkpatrick and Klein, 2010; McDonald et al., 2011; Cohen et al., 2011; Naseem et al., 2012). Here, we test the ability of such a universal annotation scheme to encode potentially useful semantic knowledge cross-linguistically; in this case, to yield more informed UDEPS contexts for improved word embeddings.

The extraction of UDEPS as the new variant of dependency-based contexts is completely language-agnostic on purpose: exactly the same procedure is followed for each language in comparison in order to make the representation learning framework completely universal.

2.1 Context Types

Prequel: Representation Model For all the context types, we opt for the standard and robust choice in vector space modeling: SGNS (Mikolov et al., 2013b; Levy et al., 2015). In all our experiments we use `word2vecf`, a reimplementation of `word2vec` which is capable of learning from arbitrary $(word, context)$ pairs.[2] Keeping the representation model fixed across experiments and varying only the context type allows us to attribute any differences in results to a sole factor: the context type.

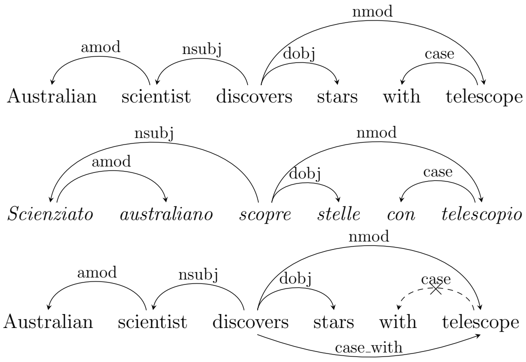


Figure 1: An example of extracting dependency-based contexts from UD parses (UDEPS) in English and Italian. **Top**: the example sentence in English taken from (Levy and Goldberg, 2014a), now UD-parsed. **Middle**: the same sentence in Italian, UD-parsed. Note the very similar structure of the two parses. **Bottom**: the intuition behind UDEPS-ARC. The uninformative short-range *case* arc between *with* and *telescope* is removed, and another "pseudo-arc" now specifying the exact link type (i.e., *case_with*) between *discovers* and *telescope* is added.

BOW The English sentence from Fig. 1 is used as the running example for all context types. Given the target word w and the window size k, the BOW context simply comprises all $2k$ word pairs (w, v), where v is found in the window of k words preceding w or k words following w, e.g., BOW with $k = 2$ extracts the following contexts v for the word *discovers* from Fig. 1: *Australian*, *scientist*, *stars*, *with*. Note that BOW may miss valid longer-range contexts (e.g., *telescope*) while including some accidental (e.g., *Australian*) or uninformative ones (e.g., *with*).

POSIT A more informed variant of BOW is *positional contexts*. It includes extra information on the actual sequential position of each context word (Levy and Goldberg, 2014b). Given the same example, POSIT with $k = 2$ extracts the following contexts for *discovers*: *Australian_-2*, *scientist_-*

---

[1] We use the latest Version 1.2 UD treebanks: http://universaldependencies.org/

[2] https://bitbucket.org/yoavgo/word2vecf
For details concerning the implementation and learning, we refer the interested reader to (Goldberg and Levy, 2014; Levy and Goldberg, 2014a).

1, *stars_+2*, *with_+1*. This context type has not been studied systematically in relation to learning WEs. POSIT suffers from the same issues with locality as BOW, but its shallow positional annotations may capture additional shallow syntactic phenomena in the data. Therefore, POSIT may be considered a link from BOW towards DEPS.[3]

UDEPS-NAIVE Given a corpus of parsed sentences, for each target w with modifiers $m_1, \ldots, m_k$ and head h, w is paired with context elements $m_1\_r_1, \ldots, m_k\_r_k, h\_r_h^{-1}$, where r is the type of the UD relation between the head and the modifier (e.g., *amod*), and r^{-1} denotes an inverse relation. A *naive* version of the UD-based model extracts contexts from the parsed corpus without any post-processing. The UDEPS-NAIVE contexts of *discovers* are now: *scientist_nsubj*, *stars_dobj*, *telescopio_nmod*. They capture longer-range relations (e.g., *telescope*) and filter out "accidental contexts" (e.g., *Australian*). In addition, the typed dependencies reveal more than POSIT and BOW about the nature of the relation in context.

UDEPS-ARC However, UDEPS-NAIVE also produces uninformative context pairs such as *(telescope, with_case)*, and it does not specify the type of e.g. the *nmod* relation between *discovers* and *telescope* which are linked through the preposition *with*. Our intuition is that a simple post-hoc intervention into the UDEPS context extraction may yield even more focused contexts. UDEPS-ARC leans on the idea of *arc collapsing* from prior work (Levy and Goldberg, 2014a; Melamud et al., 2016) that we now adjust to the UD annotation scheme. The difference to UDEPS-NAIVE is as follows: For each pair of words linked through *case* (e.g., *discovers* and *telescope*), we introduce a new "pseudo-arc" which is typed by the actual *case*/preposition. This results in a new context for *discovers*: *telescope_case_with* and also for *telescope*: *discovers_case_with*$^{-1}$ (Fig. 1). In addition, we remove the uninformative *case* arc and its associated contexts: *(with, telescope_case*$^{-1}$*)*, *(telescope, with_case)* from the training pairs.

Language	Tagging Acc.	LAS [UAS]
English (EN)	0.952	0.852 [0.875]
German (DE)	0.923	0.802 [0.850]
Italian (IT)	0.970	0.884 [0.907]

Table 1: Universal POS tagging accuracy scores and labeled (LAS) vs unlabeled (UAS) attachment scores of universal dependency parsing.

3 Experimental Setup

Evaluation Our cross-linguistic study is made possible not only thanks to the "universal NLP" initiative but also owing to the benchmarking evaluation sets for other languages beyond English (i.e., IT, DE) that have very recently become available, e.g., (Leviant and Reichart, 2015). We evaluate SGNS with different context types from sect. 2.1 across the three languages on two benchmarking tasks and datasets: (1) semantic similarity on SimLex-999 (Hill et al., 2015) translated and re-scored by native speakers in EN, DE, and IT (Leviant and Reichart, 2015), and (2) word analogies on the Google dataset (Mikolov et al., 2013a) made available in IT (Berardi et al., 2015) and DE (Köper et al., 2015) only recently.

WE Induction: Data All the word representations in comparison are induced from the Polyglot Wikipedia data (Al-Rfou et al., 2013).[4]

UPOS Tagging and UD Parsing The Wikipedia corpora were UPOS-tagged using a state-of-the art system TurboTagger (Martins et al., 2013).[5] TurboTagger was trained using suggested settings without any further parameter fine-tuning (SVM MIRA with 20 iterations) on the TRAIN+DEV portion of the UD treebank annotated with UPOS tags. Following that, the Wikipedia data were UD-parsed[6] using the graph-based Mate parser v3.61 (Bohnet, 2010)[7] and the same regime: suggested settings on the TRAIN+DEV UD treebank portion.[8] The performance of the models measured on the TEST portion of the UD treebanks is reported in Tab. 1.

---

[3] Results with another context type relying on substitute vectors (Yatbaz et al., 2012; Melamud et al., 2015) are omitted due to its subpar performance in our experiments as well as across a variety of semantic tasks in a recent English-focused study (Melamud et al., 2016).

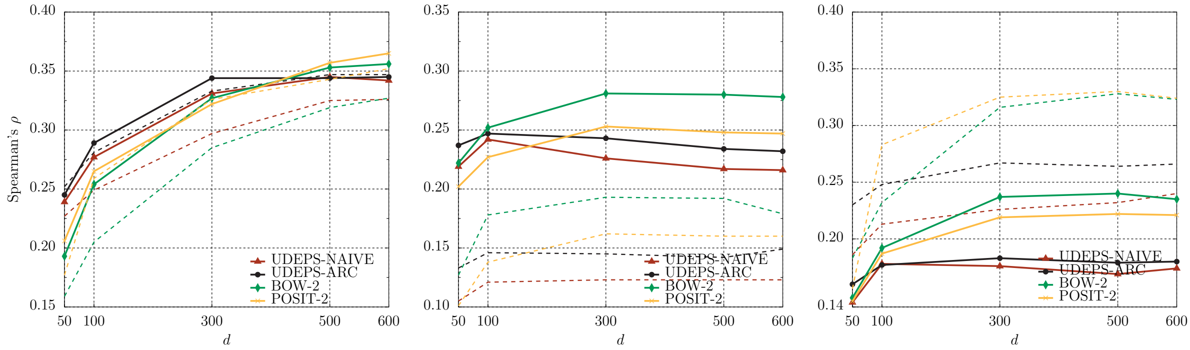
[4] https://sites.google.com/site/rmyeid/projects/polyglot

[5] http://www.cs.cmu.edu/ ark/TurboParser/

[6] Besides EN, DE, and IT, we also UPOS-tagged and UD-parsed Wikipedias in NL, ES, and HR. We believe that the full UPOS-tagged and UD-parsed Wikipedias in six languages are a valuable asset for future research and we plan to make the resource publicly available at: `http://ltl.mml.cam.ac.uk/resources/`

[7] https://code.google.com/archive/p/mate-tools/

[8] We opted for the Mate parser due to its speed, simplicity, and state-of-the-art performance according to very recent parser evaluations (Choi et al., 2015).



(a) English (b) German (c) Italian

Figure 2: Results in the semantic similarity task on SimLex-999 for three languages using different context types in the SGNS model. Solid lines denote the results on all words from SimLex-999, while thinner dashed lines show results on the verb portion of SimLex-999 (222 verb pairs).

Language:	**English**			**German**			**Italian**		
	TOT	SEM	SYN	TOT	SEM	SYN	TOT	SEM	SYN
UDEPS-NAIVE	0.351	0.231	0.446	0.183	0.101	0.276	0.169	0.033	0.282
UDEPS-ARC	0.376	0.247	0.478	0.199	0.091	0.319	0.177	0.033	0.296
DEPS-LEVY	0.390	0.183	0.548	-	-	-	-	-	-
BOW-2	0.581	0.543	0.610	0.334	0.341	0.326	0.225	0.078	0.339
POSIT-2	0.485	0.336	0.607	0.219	0.173	0.271	0.208	0.052	0.330

Table 2: $Acc@1$ scores in the analogy solving task over semantic (SEM), syntactic (SYN) and all analogies (TOT). SGNS with $d = 300$ for all context types. Similar trends are observed with other d-s. DEPS-LEVY refers to pre-trained 300-dimensional EN WEs from (Levy and Goldberg, 2014a).

The results are consistent with prior work on the UD treebanks, e.g., (Tiedemann, 2015).

Training Setup The SGNS preprocessing scheme for English was replicated from (Levy and Goldberg, 2014a) and extended to the other two languages: all tokens were converted to lowercase, and words and contexts that appeared less than 100 times were filtered. Exactly the same vocabularies were used with all context types (approx. 185K distinct EN words, 163K DE words, and 83K IT words). The `word2vecf` SGNS was trained using standard settings: 15 epochs, 15 negative samples, global learning rate 0.025, subsampling rate $1e-4$. All WEs were trained with $d = 50, 100, 300, 500, 600$. BOW-based WEs were trained with $k = 2$ (BOW-2), proven to be the (near-)optimal choice across various semantic tasks in related work (Levy and Goldberg, 2014a; Melamud et al., 2016). The same k was used for POSIT-based WEs (POSIT-2).

4 Results and Discussion

Fig. 2(a)-2(c) show the results on SimLex-999 (Spearman's ρ) for WEs with different d-s, while Tab. 2 displays the $Acc@1$ scores in the analogy solving task. English DEPS with arc collapsing from prior work (Levy and Goldberg, 2014a) (DEPS-LEVY, $d = 300$) obtain ρ of 0.372 on all SimLex pairs, and 0.378 on verb pairs.[9] A comparison with UDEPS-ARC reveals only a slight drop in performance when switching to language-agnostic UDEPS (see Fig. 2(a), Q1).[10]

However, the results are heavily dependent on the actual language: the claims made for English (i.e., DEPS $\geq$ BOW) do not extend to other languages (Q2). A comparison of results from Tab. 1 with the task evaluation also shows that excellent tagging and parsing results do not guarantee a strong task performance.

The results over the verb subset of SimLex also reveal that claims established with English are not necessarily general and true with other languages. For instance, while it has been noted that modeling verb similarity is indeed a difficult problem in English as evidenced by lower correlation scores on SimLex (see Fig. 2(a) and e.g. (Schwartz et al., 2015)), verbs are apparently easier to model in Italian (Fig. 2(c)), and a real challenge in German,

---

[9] Note that the correlation scores for all models on the re-annotated version of SimLex-999 (Leviant and Reichart, 2015) are lower than those on the original SimLex-999.

[10] The comparison is valid since DEPS-LEVY were trained on exactly the same data with the same vocabulary.

Syntactic Relation	**English**	**German**	**Italian**
gram1-adjective-to-adverb	P>B>**A**>N	-	P>B>**A**>N
gram2-opposite	**A**>N>P>B	**A**>B>P>N	**A**>B>P>N
gram3-comparative	P>B>**A**>N	**A**>B>P>N	P>**A**>N>B
gram4-superlative	P>B>**A**>N	**A**>N>B>P	P>B>**A**>N
gram5-present-participle	P>**A**>B>N	**A**>N>P>B	P>B>**A**>N
gram6-nationality-adjective	B>P>**A**>N	B>P>**A**>N	B>P>**A**>N
gram7-past-tense	**A**>P>N>B	**A**>B>N>P	**A**>B>P>N
gram8-plural	B>P>**A**>N	**A**>B>P>N	**A**>P>N>B
gram9-plural-verbs	P>**A**>B>N	**A**>N>B>P	**A**>N>P>B

Table 3: Rankings based on Acc_1 scores over syntactic analogy groups (from the Google dataset). A=UDEPS-ARC, N=UDEPS-NAIVE, B=BOW-2, P=POSIT-2. $d = 300$.

with extremely low correlation scores (Fig. 2(b)).

The results on the analogy task from Tab. 2 suggest the evident advantage of more abundant (but less informed) BOW contexts across all languages. This finding is completely in line with the analyses from prior work on English, e.g., Levy and Goldberg (2014a) report that "DEPS perform dramatically worse than BOW contexts on analogy tasks", but without providing any exact numbers.

Nonetheless, the relative ranking of context types over syntactic analogy sets as highlighted in Tab. 3 marks the evident advantage of the more-informed POSIT and UDEPS-ARC on analogies referring to functional similarity. UDEPS-ARC in German outperforms all other context types on all syntactic analogies, except for the *nationality-adjective* relation. The strongest performance of UDEPS is detected with syntactic analogies where two words in the analogy pair are perfectly replaceable in the given context (e.g., past-tense: *dancing-danced*, *sleeping-slept* or opposite: *sure-unsure*, *honest-dishonest*).

We can also see that POSIT displays a strong performance in detecting functional similarity across all three languages in both tasks (e.g., see the results in Tab. 3 where they outperform BOW). This finding reveals that POSIT should be included as a strong baseline in any follow-up work.

We also analysed the influence of the training data size by learning EN WEs from the EN Wikipedia comprising roughly 13M sentences (same size as the IT Wikipedia). As Tab. 4 shows, the absolute scores are naturally lower with less training data, and we observe a decrease in the performance of UDEPS. However, the decrease is small: these results demonstrate that the reduced performance of UDEPS in IT and DE cannot be attributed solely to smaller training datasets and sparsity of $(word, context)$ pairs.

Set/Model	BOW-2	POSIT-2	NAIVE	ARC
SimLex-all	0.286	0.289	0.271	0.279
SimLex-verbs	0.259	0.286	0.260	0.288

Table 4: Results on SimLex in English with SGNS trained on a reduced EN training set containing the same number of sentences as the entire IT training set ($\approx$ 13M sentences). $d = 300$.

Finally, the consistent improvements of UDEPS-ARC over UDEPS-NAIVE for all three languages on both tasks show the importance of a careful post-hoc selection of informative contexts. Future work will delve deeper into the informative context selection for the WE learning.

5 Conclusion and Future Work

We have presented the first comparison of different context types for learning word embeddings for multiple languages. Dependency-based contexts in different languages are for the first time extracted from "universal" parses made possible by the Universal Dependencies initiative, without any language-specific optimisation.

In sum, our comparison provides no clear answer to the question posed by the title of this paper. However, it shows conclusively that different context types yield semantic spaces with different properties, and that the optimal context type depends on the actual application and language. The usefulness of universal dependency-based contexts is evident with a simple post-parsing context extraction scheme in tasks oriented towards syntactic/functional similarity.

This first cross-linguistic analysis covering only a small set of languages from the same (Indo-European) phylum also reveals that training word embeddings in languages other than English is not trivial, suggesting Anglo-centric assumptions that do not extend to other languages (Bender, 2011). It is therefore essential not to generalise results on English to other languages without clear empirical evidence. Yet, a broader cross-linguistic study involving more languages from other families (with UD treebanks available) and additional experimentation is warranted in order to better guide research on "universal NLP" and language-independent word representation learning.

Acknowledgments

This work is supported by ERC Consolidator Grant LEXICAL (648909). The authors are grateful to Roi Reichart and the anonymous reviewers for their helpful comments and suggestions.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *CoNLL*, pages 183–192.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *ACL*, pages 809–815.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–28.

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to Italy: A comparison of models and training datasets. In *Italian Information Retrieval Workshop*.

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *ACL*, pages 1288–1297.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *COLING*, pages 89–97.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

Jinho D. Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a Web-based evaluation tool. In *ACL*, pages 387–396.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *EMNLP*, pages 50–61.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, pages 4585–4592.

Yoav Goldberg and Omer Levy. 2014. Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and "semantic" structure of continuous word spaces. In *IWCS*, pages 40–45.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR*, abs/1508.00106.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *ACL*, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3:211–225.

André F. T. Martins, Miguel B. Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL*, pages 617–622.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.

Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *ACL*, pages 92–97.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015. Modeling word meaning in context with substitute vectors. In *NAACL-HLT*, pages 472–482.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *NAACL-HLT*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *ACL*, pages 629–637.

Joakim Nivre et al. 2015. Universal Dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. A universal part-of-speech tagset. In *LREC*, pages 2089–2096.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *CoNLL*, pages 258–267.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. In *DepLing*, pages 340–349.

Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384–394.

Mehmet Ali Yatbaz, Enis Sert, and Deniz Yuret. 2012. Learning syntactic categories using paradigmatic representations of word context. In *EMNLP*, pages 940–951.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP*, pages 35–42.