

Multitask Learning for Cross-Lingual Transfer of Broad-coverage Semantic Dependencies

Maryam Aminian¹, Mohammad Sadegh Rasooli², Mona Diab^{1,3}

¹Department of Computer Science, The George Washington University, Washington D.C.

²Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA

³Facebook AI, Seattle, WA, USA

¹{aminian, mtdiab}@gwu.edu, ²rasooli@seas.upenn.edu

Abstract

We describe a method for developing broad-coverage semantic dependency parsers for languages for which no semantically annotated resource is available. We leverage a multitask learning framework coupled with annotation projection. We use syntactic parsing as the auxiliary task in our multitask setup. Our annotation projection experiments from English to Czech show that our multitask setup yields 3.1% (4.2%) improvement in labeled F1-score on in-domain (out-of-domain) test set compared to a single-task baseline.

1 Introduction

Broad-coverage semantic dependency parsing (SDP)¹ was first introduced in the SemEval shared task (Oepen et al., 2014) and aims to provide semantic analysis of sentences by capturing semantic relations between all content-bearing words in a sentence. The rich graph structure introduced by SDP allows the model to cover a wide range of semantic phenomena such as negation, comparatives, possessives and various types of modifications that have not been previously analyzed in other models such as semantic role labeling (Baker et al., 1998).

Despite all advantages provided by SDP, resources with annotated semantic dependencies are limited to the three languages released in the SemEval shared tasks (Oepen et al., 2014, 2015; Che et al., 2016) namely English, Czech and Chinese. This data scarcity motivates us to use well-known and traditionally used transfer methods such as *annotation projection* for building SDP models for languages without semantically annotated data. In annotation projection, we assume that we have access to sentence-aligned corpora that can be used for transferring semantic annotations from a rich-resource source language to the target language

¹We use *broad-coverage semantic dependencies* and *semantic dependencies* interchangeably throughout this paper.

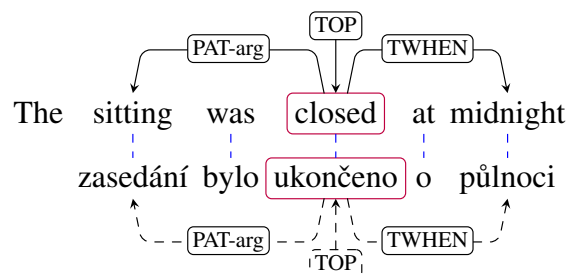


Figure 1: Projecting SDP annotations from an English to a Czech sentence. Semantic dependencies of the English sentence (top) are projected using alignments (dashed lines in the middle) to obtain projected semantic dependencies (bottom) for the target sentence.

through word alignment links. Figure 1 shows an example of annotation projection for semantic dependencies.

Motivated by the large amount of similarities between syntactic and semantic dependencies, we further propose a simple but effective multitask learning framework to leverage supervised syntactic parse information and improve the representation learning capability in the intermediate layers of our semantic parser. Our multitask learning approach, despite its simplicity, yields significant improvements in the performance of the vanilla semantic dependency parser built using annotation projection. We conducted annotation projection experiments from English to Czech. Our experiments show that our multitask setup yields 3.1% and 4.2% improvement in the labeled F1 results on in-domain and out-of-domain evaluation sets respectively. Furthermore, we explore the efficacy of contextualized word representations, BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018) as features in our annotation projection model and find a marginal gain by using those contextual features. To the best of our knowledge, this work is the first study to develop an enhanced semantic depen-

dependency parser through multitasking in the absence of annotated data.

2 Related Work

After the SemEval shared tasks on broad-coverage semantic dependency parsing (Oepen et al., 2014, 2015; Che et al., 2016), there have been many studies to build supervised SDP models (Du et al., 2015; Chen et al., 2018; Almeida and Martins, 2015; Wang et al., 2018; Dozat and Manning, 2018; Stanovsky and Dagan, 2018; Kurita and Søgaard, 2019), however, all efforts were restricted to the three languages released through SemEval shared tasks. There have been extensive number of studies that use annotation projection to cure data scarcity in different tasks such as part-of-speech tagging (Täckström et al., 2013), syntactic parsing (McDonald et al., 2011), semantic role labeling (Padó and Lapata, 2005) and semantic parsing (Hershcovich et al., 2019). Nevertheless, none of the previous works, to the best of our knowledge, looked into using annotation projection for building SDP models for languages without semantically annotated data.

Motivated by the fact that different semantic representations or formalisms cover different aspects of sentence-level semantics, there has been a line of studies to apply multitask learning over different semantic annotations (Peng et al., 2017, 2018; Hershcovich et al., 2018; Kurita and Søgaard, 2019) or target cross-framework meaning representation (Oepen et al., 2019). These studies use the shared semantic information across different representations to enhance the SDP model for a given language, however, none of them addressed the case that no semantically annotated data is available for a language. This paper is the first work that aims to build an SDP model based on cross-lingual transfer without any annotation in the target language of interest.

3 The Parsing Model

For an input sentence $x = x_1, \dots, x_n$ with n words, the goal of a semantic dependency parsing model is to learn binary dependency decisions $y_{i,j} \in \{0, 1\}$ for every head index $0 \leq i \leq n$ and dependent index $1 \leq j \leq n$, where x_0 is the dummy *root* token. For every head-dependent pair (i, j) , such that $y_{i,j} = 1$, the parser finds a label $l_{i,j}$ from a set of predefined semantic dependency labels \mathcal{L} . In most cases, the parsing decision is

decomposed in two steps: unlabeled dependency parsing, and labeling each dependency edge. The only constraint here is that the final semantic graph should be acyclic.

We use the standard model of Dozat and Manning (2018) for which the parsing model is based on a simple head selection algorithm. This model learns dependency edge scores $s^{\text{edge}}(i, j)$ for all possible head-dependent pairs (i, j) . The final parsing decision is a sign function:

$$y_{i,j} = \{s^{\text{edge}}(i, j) \geq 0\}$$

Similarly, the parser learns a labeling function $s_1^{\text{label}}(i, j)$ for every pair that $y_{i,j} = 1$:

$$l_{i,j} = \arg \max_{l \in \mathcal{L}} s_l^{\text{label}}(i, j)$$

Our parsing model uses a deep neural model in which the first layer is the embedding layer that consists of word, part-of-speech tag, and character representations. The second layer consists of deep bidirectional LSTMs (Hochreiter and Schmidhuber, 1997) that construct recurrent representations r_i for every word. The third layer uses four single-layer feed-forward neural networks (FNN) as attention mechanisms for head and dependent binary decisions and label assignments. The final layer uses a bilinear function to score the FNN outputs. For training the model, the sigmoid cross-entropy function is used for the edges, and the softmax cross-entropy function is used for the labels. The two losses are interpolated to calculate the final loss value with a coefficient $0 < \lambda < 1$.

4 Projecting Semantic Dependencies

For a source sentence $x' = x'_1, \dots, x'_m$ with m words, and a target sentence $x = x_1, \dots, x_n$ with n words, we obtain one-to-one alignments by running an unsupervised word alignment algorithm on both directions. We use the intersected alignments $a = a_1, \dots, a_m$ such that $0 \leq a_i \leq n$ where $a_i = 0$ indicates a null or empty alignment. For every source dependency relation $y'_{i,j} \in \{0, 1\}$ where $a_i, a_j \neq 0$, we project the dependency edge and label to the target sentence $y_{a_i, a_j} = y'_{i,j}$ and $l_{a_i, a_j} = l'_{i,j}$ (if $i = 0$ then $a_i = 0$). We then train a supervised parsing model on the projected dependencies. These projected dependencies are usually partial and contain some noise that are caused by different reasons such as translation shifts and alignment errors.

5 Multitask Learning with Syntax

Modeling auxiliary tasks in a multitask learning framework allows the main task to benefit from structural or statistical similarities found in one or more auxiliary tasks to improve the model learned for a target task (Caruana, 1997). Given the large amount of (labeled and unlabeled) correlations existing among syntactic and semantic dependencies, we consider syntactic dependency parsing as the auxiliary task for semantic dependency parsing.

In order to find out the best parameter sharing structure, we try the following parameter sharing variations: 1) sharing embedding and recurrent layers, 2) sharing embedding and recurrent layers with an additional task-specific recurrent layer, 3) sharing all three layers, but with an additional task-specific recurrent layer, and 4) sharing all intermediate layers. Figure 2 shows the first case for which only the first two layers are shared between the two tasks. The overall loss value for the multitask model is computed by interpolating semantic and syntactic losses using an interpolation coefficient ω which is tuned on the development data. We use projected semantic dependencies and syntactic dependency parses generated using a supervised parser to train the multitask model. Thus the training data for the target language has projected semantic annotations plus fully parsed syntactic trees.

6 Experiments and Results

We consider English as the source language and Czech as the target language. We use the SemEval 2015 (Oepen et al., 2015) in-domain and out-of-domain test sets to evaluate our models. Since the PSD (Prague semantic dependencies) annotation is available for both English and Czech, we use that throughout our experiments. We use Giza++ (Och and Ney, 2003) with its default configuration to obtain intersected word alignments on the Europarl parallel corpus (Koehn, 2005). The training data used in our projection experiments is drawn from Europarl which contains text from the political domain. The in-domain Czech test set provided by the SemEval 2015 contains translated texts from corresponding sections of WSJ in the newswire domain, whereas the out-of-domain evaluation set for Czech (also provided by SemEval 2015) is drawn from Prague Dependency Treebank 3.0 (Hajič et al., 2012) which mainly contains text from journals and scientific articles, thus considered of a fairly differ-

ent domain compared to Europarl (political).

We explore efficacy of multitasking in our annotation projection model by comparing the multitask results with the single-task baseline model that does not use any multitasking. The training corpus of Czech with projected annotations contains 612k sentences but due to computational limitations, we train all models on a sample of 80k sentences² randomly selected from original projections. In order to simulate a fully unsupervised approach, we use 5% of the projected data as the held-out data during training.

Parsing Parameters We use the structural skip-gram model of Ling et al. (2015) for English word embeddings and run word2Vec (Mikolov et al., 2013) on Wikipedia text to acquire the word vectors for Czech. We use UDpipe (Straka and Straková, 2017) pretrained models v1.2.0 (trained on the Universal Dependencies v2.0) to produce automatic part-of-speech tags. We train the biaffine dependency parser of Dozat and Manning (2017) on the Universal Dependencies corpus v2.0 (Nivre et al., 2017) to generate supervised syntactic parses in our multitask learning experiments. All modules are implemented using the Dynet library (Neubig et al., 2017).

We mainly use the hyper-parameters used in Dozat and Manning (2018) except that we use a character BiLSTM without any linear transformation layers. We use word and part-of-speech vectors of size 100, with 3-layer LSTMs of size 600, and feed-forward layers of size 600. We use a dropout of probability of 0.2 for words and part-of-speech tags, and 0.25 for the recurrent and unlabeled feed-forward layers, and 0.33 for the labeled feed-forward layers. The interpolation constants λ and ω are set to 0.025 and 0.975 respectively to prioritize the semantic task as our main task in the multitask framework. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 on minibatches of approximately thousand tokens. We also concatenate the contextual vectors to the input layer as additional features to the parser. We use the pretrained ELMO embeddings (Peters et al., 2018) of size 1024 from (Che et al., 2018; Fares et al., 2017). Their model is trained on the set of 20-million-words data randomly sampled from the raw texts released by the CoNLL 2018 shared task for Czech and uses the same model and

²The sample size is selected during development experiments.

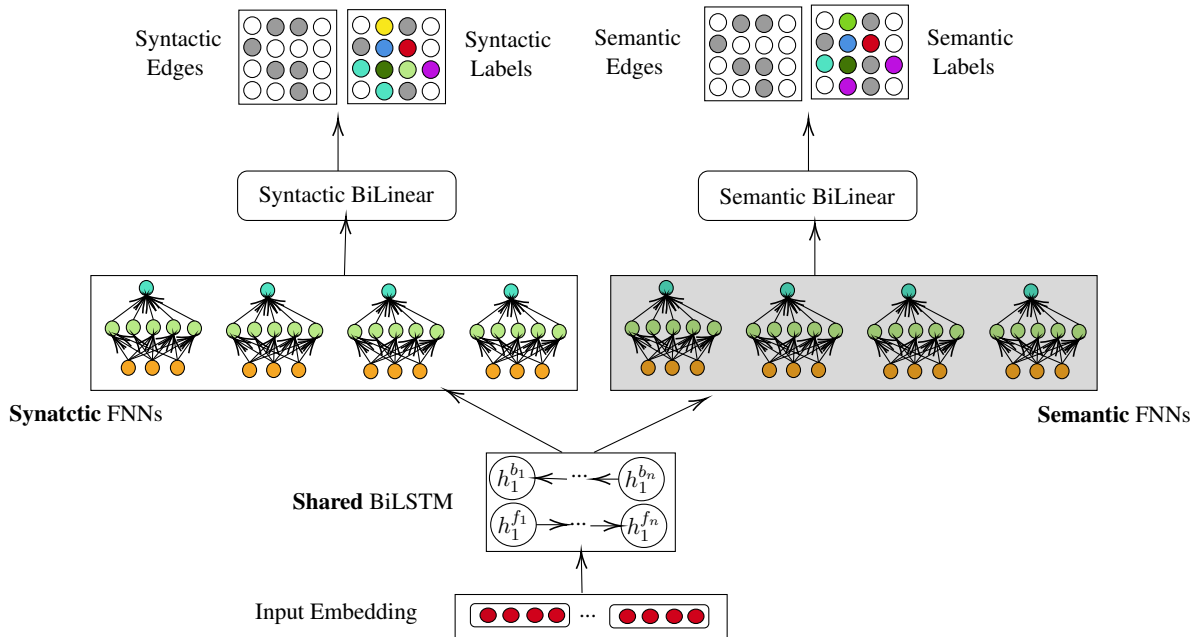


Figure 2: Multitask architecture with shared embedding and recurrent layers across the two tasks.

Model	Shared		Task	In-domain	Transfer		+ ELMO		+ mBERT		Supervised	
	RNN	FNN			data	LF	UF	LF	UF	LF	UF	LF
Single				✓	57.5	74.3	57.4	74.5	56.3	75.0	85.4	91.1
					58.8	75.8	59.0	75.8	57.9	76.3	70.4	86.8
Mutitask	✓	✓		✓	59.3	76.4	59.3	76.9	58.3	77.0	85.1	90.1
					61.2	78.2	61.5	78.2	60.6	78.6	70.8	87.0
	✓	✓	✓	✓	58.3	75.7	59.3	76.7	58.4	77.1	85.1	91.0
					60.5	77.9	61.4	78.5	60.4	78.6	70.8	87.3
	✓			✓	57.7	75.2	59.2	75.7	55.7	75.0	84.3	90.2
				59.9	77.1	61.4	77.5	58.2	76.6	69.7	86.4	
✓		✓	✓	58.7	75.3	58.6	76.0	57.6	76.4	85.4	91.0	
				60.8	77.3	60.6	77.5	59.9	78.1	71.1	87.3	

Table 1: Results on the Czech SemEval test data. LF and UF denote Labeled and Unlabeled F1 respectively. The Transfer column does not use contextualized word embeddings. *Task RNN* refers to an extra task-specific embedding in the multitask setting. The shaded rows show results on out-of-domain test data.

hyper-parameters as Peters et al. (2018). We use the pretrained multilingual BERT models (Devlin et al., 2019) of size 768 from Xiao (2018) with 12 layers and 12 heads. Due to computational limitations, we only use the pretrained BERT models in the input layer without finetuning.

6.1 Results

Table 1 shows the results on in-domain and out-of-domain data with and without contextual word embeddings. The Single row shows the baseline where we use Czech projection data to train the model. The Multitask rows show the results when we utilize syntactic parses through multitasking.

The last column shows results of the supervised model trained on the gold data provided as part of the SemEval 2015 shared task, whereas other columns are trained on the transferred/projected annotations. The + ELMO and + mBERT columns in Table 1 show results when we add ELMO and BERT pretrained embeddings as additional features in the input layer. It is worth emphasizing that the ELMO and BERT embeddings are not cumulative and their results are reported from separate models.

Comparing the labeled F1 scores for different multitask models, we observe that all multitask models outperform the Single baseline, regardless of the architecture used to train the model.

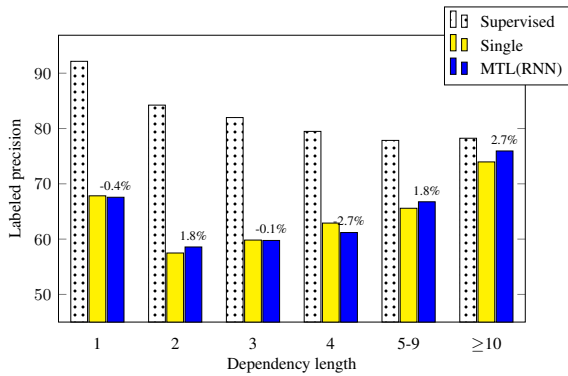


Figure 3: Labeled precision of the best-performing multitask model compared to the single-task and supervised model for different dependency lengths.

We also observe that multitask models yield a larger increase on the out-of-domain test set compared to the in-domain test set which illustrates the particular power of multitask model to improve SDP model in truly low-resource settings where in-domain training data might not be available. As we see in the results, the multitask model with a shared recurrent layer slightly outperforms other models. We also see marginal gains from using the ELMO embeddings, and some gain in unlabeled score in using BERT without seeing improvement in labeled accuracy.

Comparing our results with the supervised model, we observe that multitasking helps the target SDP model obtain closer performance to the supervised model on out-of-domain data which further highlights the power of multitasking for low-resource settings.

Analyzing Different Dependency Lengths We analyze the performance of our best performing multitask model on different semantic dependencies. Figure 3 illustrates labeled precision of the best performing multitask model compared to the single-task and supervised models for different semantic dependency lengths. Length of a dependency is defined as number of tokens located between the semantic head and its dependent. Numbers shown above each plot denotes the improvement obtained from the multitask model compared to the single-task model. Interestingly, the multitask model yields larger improvement on longer semantic dependencies compared to the shorter ones, such that its precision for semantic dependencies with length ≥ 10 is noticeably close to the supervised results. This finding further highlights the power of syntactic representations in capturing long

distance relations which is injected to our model through the shared RNN layer between syntax and semantics.

7 Conclusion

We have described a semantic dependency parsing model based on annotation projection that do not use any annotated semantic data in the target language. We enhance the target semantic model by incorporating syntax in a multitask learning framework. We demonstrate that our multitask model outperforms the single-task model on both in-domain and out-of-domain test sets on the Czech language.

Acknowledgments

We would like to acknowledge the useful comments by three anonymous reviewers who helped in making this publication more concise and better presented.

References

- Mariana S. C. Almeida and André F. T. Martins. 2015. [Lisbon: Evaluating TurboSemanticParser on multiple languages and out-of-domain data](#). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 970–973, Denver, Colorado. Association for Computational Linguistics.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. [The Berkeley FrameNet project](#). In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.
- Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.
- Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. [Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.
- Wanxiang Che, Yanqiu Shao, Ting Liu, and Yu Ding. 2016. [SemEval-2016 task 9: Chinese semantic dependency parsing](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1074–1080, San Diego, California. Association for Computational Linguistics.

- Yufei Chen, Sheng Huang, Fang Wang, Junjie Cao, Weiwei Sun, and Xiaojun Wan. 2018. [Neural maximum subgraph parsing for cross-domain semantic dependency analysis](#). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 562–572, Brussels, Belgium. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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. Association for Computational Linguistics.
- Timothy Dozat and Christopher D Manning. 2017. [Deep biaffine attention for neural dependency parsing](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
- Timothy Dozat and Christopher D. Manning. 2018. [Simpler but more accurate semantic dependency parsing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.
- Yantao Du, Fan Zhang, Xun Zhang, Weiwei Sun, and Xiaojun Wan. 2015. [Peking: Building semantic dependency graphs with a hybrid parser](#). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 927–931, Denver, Colorado. Association for Computational Linguistics.
- Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. [Word vectors, reuse, and replicability: Towards a community repository of large-text resources](#). In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.
- Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Sebecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. [Announcing Prague Czech-English dependency treebank 2.0](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association (ELRA).
- Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2018. [Multitask parsing across semantic representations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 373–385, Melbourne, Australia. Association for Computational Linguistics.
- Daniel Hershcovich, Zohar Aizenbud, Leshem Choshen, Elior Sulem, Ari Rappoport, and Omri Abend. 2019. [SemEval-2019 task 1: Cross-lingual semantic parsing with UCCA](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1–10, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.
- Shuhei Kurita and Anders Søgaard. 2019. Multi-task semantic dependency parsing with policy gradient for learning easy-first strategies. *arXiv preprint arXiv:1906.01239*.
- Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. [Two/too simple adaptations of word2vec for syntax problems](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado. Association for Computational Linguistics.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. [Multi-source transfer of delexicalized dependency parsers](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.
- Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.
- Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, et al. 2017. [Universal Dependencies 2](#). LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

- Stephan Oepen, Omri Abend, Jan Hajič, Daniel Herscovich, Marco Kuhlmann, Tim O’Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdenka Uresova. 2019. **MRP 2019: Cross-framework meaning representation parsing**. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 1–27, Hong Kong. Association for Computational Linguistics.
- Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. 2015. **SemEval 2015 task 18: Broad-coverage semantic dependency parsing**. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado. Association for Computational Linguistics.
- Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. **SemEval 2014 task 8: Broad-coverage semantic dependency parsing**. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland. Association for Computational Linguistics.
- Sebastian Padó and Mirella Lapata. 2005. **Cross-linguistic projection of role-semantic information**. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.
- Hao Peng, Sam Thomson, and Noah A. Smith. 2017. **Deep multitask learning for semantic dependency parsing**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2048, Vancouver, Canada. Association for Computational Linguistics.
- Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith. 2018. **Learning joint semantic parsers from disjoint data**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1492–1502, New Orleans, Louisiana. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. **Deep contextualized word representations**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Gabriel Stanovsky and Ido Dagan. 2018. **Semantics as a foreign language**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2412–2421, Brussels, Belgium. Association for Computational Linguistics.
- Milan Straka and Jana Straková. 2017. **Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes**. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.
- Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. **Token and type constraints for cross-lingual part-of-speech tagging**. *Transactions of the Association for Computational Linguistics*, 1:1–12.
- Yuxuan Wang, Wanxiang Che, Jiang Guo, and Ting Liu. 2018. **A neural transition-based approach for semantic dependency graph parsing**. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Han Xiao. 2018. bert-as-service. <https://github.com/hanxiao/bert-as-service>.