

# PP Attachment: Where do We Stand?

Daniël de Kok and Jianqiang Ma and Corina Dima and Erhard Hinrichs

SFB 833 and Seminar für Sprachwissenschaft

University of Tübingen, Germany

{ddekok, jma, cdima, eh}@sfs.uni-tuebingen.de

## Abstract

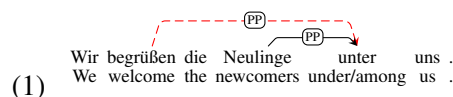
Prepositional phrase (PP) attachment is a well known challenge to parsing. In this paper, we combine the insights of different works, namely: (1) treating PP attachment as a classification task with an arbitrary number of attachment candidates; (2) using auxiliary distributions to augment the data beyond the hand-annotated training set; (3) using topological fields to get information about the distribution of PP attachment throughout clauses and (4) using state-of-the-art techniques such as word embeddings and neural networks. We show that jointly using these techniques leads to substantial improvements. We also conduct a qualitative analysis to gauge where the ceiling of the task is in a realistic setup.

## 1 Introduction

Prepositional phrase (PP) attachment is a well-known structural ambiguity in natural language parsing (Hindle and Rooth, 1993), that even modern parsers have difficulty coping with. For example, Kummerfeld et al. (2012) investigated parsing error types across a large number of parsers for English and found that PP attachment and clause attachment are the most difficult constructions. Mirroshandel et al. (2012) show that in a second-order graph parser for French, 8 of the 13 most common error types relate to PP attachment. We found in our experiments with the parser of de Kok and Hinrichs (2016) that most errors were made in PP attachment (18.42% of all labeled attachment errors).

What makes PP attachment particularly difficult is that the ambiguities can often not be solved using only structural preferences. Example 1 from

German shows the difficulty of the problem in its full glory, where the preposition *unter* “under/among” is attached to *Neulinge* “newcomers”. However, the PP could attach to *begrüßen* “welcome” when the complement of the preposition is a locative noun phrase (e.g. *offenem Himmel* “open skies”).



Spread throughout the literature, there are many important observations about and approaches to the task of PP attachment, but they have never been properly combined. We will first discuss them briefly below, and then summarize the contributions of this paper.

Most work in PP attachment assumes that a preposition attaches to either the immediately preceding noun (phrase) or the main verb (Hindle and Rooth, 1993; Volk, 2002). Some other work does take multiple nouns candidates into consideration, but only nouns that are within a certain window preceding the preposition (Ratnaparkhi, 1998; Belinkov et al., 2014) or all the nouns in the sentence (Foth and Menzel, 2006). Using examples from German, de Kok et al. (2017) show that these crude approaches are problematic. In German, there are typically more than two possible attachment sites. In fact, they show that 30% of the training instances could not even be described in this typical binary classification setup. Moreover, PPs can attach over relatively long distances and the preposition can precede its head (e.g. in PP topicalization). They also show that the task of PP attachment with multiple noun candidates is considerably more difficult than the traditional binary classification task. On the other hand, de Kok et al. (2017) also show that many spurious heads can be eliminated by exploiting relatively shallow

clause structure annotations.

Previous work has shown that bi-lexical preferences are effective in solving PP attachment ambiguities (Brunner et al., 1992; Whittemore et al., 1990). Two words have a strong bi-lexical preference if the words are likely to occur in a head-dependent relation. These preferences are usually stated in terms of information-theoretical measures, such as point-wise mutual information. Since hand-annotated treebanks usually do not have enough material to obtain reliable bi-lexical statistics, these statistics were extracted from raw text (Volk, 2001), automatically tagged (Ratnaparkhi, 1998), chunk parsed (Volk, 2002) or parsed (Hindle and Rooth, 1993; Pantel and Lin, 2000; Mirroshandel et al., 2012) corpora, resulting in *auxiliary distributions*. Since these seminal works in PP attachment, parsers have become faster (Kübler et al., 2009) and more accurate (Chen and Manning, 2014), opening the possibility to obtain better co-occurrence statistics.

Topological fields are commonly used to capture the regularities in German word order (Drach, 1937; Höhle, 1986). The distributions of syntactic relations vary significantly across topological fields, which can benefit dependency parsing of German (de Kok and Hinrichs, 2016). We expect topological fields to provide information about the distribution of PP attachment throughout clauses and thus benefit PP attachment disambiguation for German in a similar way as in dependency parsing.

Many tasks in natural language processing have seen substantial improvements in recent years through the use of word embeddings in combination with neural networks. Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) improve the lexical coverage of systems beyond supervised training sets by giving words that occur in similar contexts similar vector representations. Embeddings work especially well with neural networks, as neural networks are able to capture non-linear interactions between features.

Considering these ideas and techniques that can have an impact on modeling PP attachment, the question we want to address is *where do we stand in PP attachment?* Our contributions are threefold: (1) we evaluate PP attachment on a realistic multiple-candidate PP attachment data set for German; (2) we integrate the aforementioned advances in parsing and machine learning and confirm their usefulness for the task; and (3) we per-

form an error analysis to gauge how many of the remaining errors can be attributed to the system.

## 2 PP attachment disambiguation model

Following the discussion in the Introduction, this paper considers a realistic setup for PP attachment disambiguation, where each disambiguation instance involves choosing the correct attachment site from an arbitrary number of candidates. As the number of classes/candidates varies across disambiguation instances, it can not be modeled as a typical multiclass classification. To tackle this setup, we build a *neural candidate scoring model* (Section 2.1) to estimate the probability that the attachment candidate under consideration is the correct attachment site. Then, among all the candidates for the same PP, the candidate with the highest probability is considered to be the correct attachment site.

### 2.1 Neural candidate scoring model

Our neural candidate scoring model uses a feed-forward neural network with three layers. The input layer consists of featurized representations of a <preposition, object of the preposition, candidate> triple. These input features are discussed in more detail in Section 2.2. The network uses a hidden layer with the ReLU activation function (Hahnloser et al., 2000) as its non-linearity. Finally, the output layer uses the logistic function as an activation function to model probabilities. For regularization, dropout (Srivastava et al., 2014) is applied to the input and hidden layers. Following the best practice, we apply batch normalization (Ioffe and Szegedy, 2015) of parameters.

The model parameters are trained using (candidate, probability) pairs that are constructed from the training data. Correct and incorrect attachments are assigned probabilities 1 and 0 respectively. To learn the model parameters, we minimize the cross-entropy loss using mini-batch gradient descent. During learning, the global learning rate follows an exponential decay and the per-parameter learning rate is adjusted using Adagrad (Duchi et al., 2011).

### 2.2 Feature set

**Basic features.** Following Kübler et al. (2007), we use the word form and part-of-speech as features for the preposition, object and candidate. We

augment the absolute distance feature of Kübler et al. (2007) that counts the number of words between the preposition and the candidate, with the logarithm of this distance and the *relative distance*. The relative distance is the number of competing candidates between the candidate and the preposition.

**Word and tag embeddings** Traditional methods for PP attachment represent the word and tag features as one-hot vectors. For the embedding representations of these two types of features, we use the embeddings of de Kok (2015), which were trained on corpora of 800 millions tokens, using WANG2VEC (Ling et al., 2015), a variation of WORD2VEC that is tailored to syntactic tasks.

**Topological fields** As mentioned in the Introduction, topological fields are informative for the distributions of syntactic relations in general. Our analysis of the TüBa-D/Z dependency treebank (Telljohann et al., 2006) for German shows that this observation also holds for the PP attachment relation. For example, when the preposition is in the *initial field*, the preposition is highly likely to attach to the candidate in either the initial field or the *left bracket*. We use the method of de Kok and Hinrichs (2016) to predict the topological fields for all three types of tokens: the preposition, object and candidate. Each of these token will have a corresponding one-hot vector that represents its predicted topological field.

**Auxiliary distributions** of bi-lexical preferences have been shown to be useful for resolving syntactical ambiguities in general (Johnson and Riezler, 2000; van Noord, 2007), besides their particular benefits for PP attachment as discussed in Section 1. Such bi-lexical preferences can be captured, for example, by point-wise mutual information (PMI) that is estimated from large machine-annotated corpora. Our approach makes use of a state-of-the-art dependency parser (de Kok and Hinrichs, 2016) to parse a large corpus, namely articles from the German newspaper *taz* (*die tageszeitung*) from 1986 to 2009 (28.8 million sentences, 393.7 million tokens). The parser-predicted PP attachments are represented as <preposition, object of the preposition, candidate> triples, which we collect from both ambiguous and unambiguous PP attachment results. Here, unambiguous attachments refer to prepositions that only have one possible attachment site (Ratnaparkhi, 1998).

For bi-lexical association scores, we compute the normalized point-wise mutual information (NPMI) (Bouma, 2009), a normalized version of PMI, for three types of token pairs: (candidate, object), (candidate, preposition) and (candidate, preposition+object). For the last case, each preposition-object combination is considered as one token. NPMI is obtained by normalizing raw PMI into the range  $[-1, 1]$ , which is more favorable for learning. We also extend bi-lexical association scores to tri-lexical association scores by using specific interaction information and total correlation (Van de Cruys, 2011), both of which can simultaneously take into account three variables, which are the preposition, object and candidate in our case. Overall, our auxiliary distributions consist of 5 types of association scores that are estimated from automatically parsed corpora.

### 3 Experiments

For evaluation, we use the recently created PP attachment data set for German (de Kok et al., 2017). In this data set each preposition has multiple head candidates. The average number of candidates per preposition is 3.15. The data set is extracted from TüBa-D/Z, using a set of rules derived from the distributions of prepositions and their heads across topological fields. From this data set, we remove the instances that originate from sentences that were used to train the parser which was used in creating the auxiliary distributions. We split the remaining 43,906 instances with a 4:1 ratio for respectively training and evaluation. Initially, a subset of the training data is used to tune hyper-parameters. Then we train the model on the full training set using the chosen hyper-parameters.<sup>1</sup> Finally, the model performance is evaluated on the test set, using standard per-preposition *accuracy*, i.e the percentage of prepositions that are correctly attached.

#### 3.1 Comparison with baselines

Ideally, we would like to compare the model proposed in Section 2 to earlier approaches for German PP attachment disambiguation, using the new data set with multiple attachment candidates (see Section 3). Previous approaches typically used memory-based learning (Kübler et al., 2007) or

<sup>1</sup>The relevant hyper-parameters are: *number of hidden units: 100; dropout probability input/hidden layers: 0.2/0.05; and word/part-of-speech embedding sizes: 50.*

linear SVMs (Volk, 2001). Since the running time of the memory-based learning implementation on the data set is extremely long and linear SVMs often yield results that are similar to logistic regression on NLP tasks, we build a logistic regression model (LR) as the baseline. Logistic regression is a representative linear model with high computational efficiency. The input representations, regularization and optimization algorithm remain the same for both our model and the LR baseline.

### 3.2 Impact of embeddings and feed-forward neural networks

In the upper half of Table 3.2, we compare the LR baseline with two variations of the proposed neural network model. The baseline and the first variation (NN1) use the same one-hot feature vectors as input, as previous approaches utilize such feature representations. Our NN1 model outperforms the logistic regression baseline (LR) by 11.3% in terms of absolute accuracy improvement. Note that our experiment only uses core features without hand-crafting combinatory features, which would have improved the performance of the LR model. Thanks to the non-linearity, neural networks can implicitly capture useful feature combinations, thus leading to dramatic performance improvement from LR to NN1. Another substantial improvement (13.8%) is obtained by representing the word forms and POS tags with embeddings instead of one-hot vectors (comparing NN2 with NN1). Our lexical coverage analysis shows that the training set only covers 71.7% of the word types that occur in the test set, while the embeddings have the lexical coverage of 89.5%, which can probably account for much of the improved accuracy of NN2. Note that, in both cases, the word forms are used without lemmatization or morphological analyses. The high lexical coverage makes embeddings more robust when linguistic pre-processing is absent or inaccurate.

### 3.3 Impact of topological fields and auxiliary distributions

To test the benefits of using topological fields and auxiliary distributions for the task, we conduct further experiments to test three variations of our model. The NN3 model extends the NN2 model by adding the topological field features. The NN4 model further extends the NN3 model by adding auxiliary distributions that are estimated from all the PP attachments. Finally, the NN5 model ex-

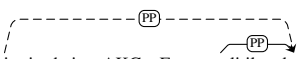
Name	Model	Accuracy
LR	LR with one-hot vectors	56.9%
NN1	NN with one-hot vectors	68.2%
NN2	NN with embeddings	82.0%
NN3	NN2 + topological fields	83.8%
NN4	NN3 + auxiliary all	86.5%
NN5	NN4 + auxiliary unamb.	<b>86.7%</b>

Table 1: Results on PP attachment disambiguation on the logistic regression baseline (LR) and our neural network models (NN\*).

tends the NN4 model by adding auxiliary distributions using only the unambiguous PP attachments. Although the unambiguous attachments are a subset of the *auxiliary all* set, the lexical association distributions of the two sets are different, thus providing extra information to the model. These results are shown in the lower half of Table 3.2. By exploiting topological fields as extra features, model NN3 obtains 1.8% absolute improvements in accuracy over model NN2. Adding *auxiliary all* features on top of NN3 leads to another 2.7% improvement in accuracy. The final 0.2% improvement in accuracy is achieved by adding auxiliary distributions using only the unambiguous PP attachments. These results confirm the usefulness of topological fields and auxiliary distributions.

## 4 Error analysis

To answer the final part of our question “where we stand in PP attachment”, we take a random sample of 100 instances that were incorrectly attached by our most accurate model. We then analyzed each instance by hand and assigned it to one of four types of errors: (1) *incorrect*: the model made a clear attachment error; (2) *discourse*: the attachment can only be resolved with discourse-level information; (3) *irrelevant*: there are two attachment choices that give rise to the same interpretation, where the gold-standard marked one while the model marked the other (see Example 2). (4) *other*: such as possible errors in the gold standard. The results are shown in Table 2.

- (2)  Sie ist Mitarbeiterin beim AKG Frauenpolitik bei den Grünen  
She is employee at-the AKG Women-politics with the Greens

Based on this data analysis, we can conclude that the ceiling for the task is lower than 100%. The 36 *irrelevant* cases and 7 *other* cases could

be seen as shortcomings of the data set, which should mark multiple attachment sites when there is no substantial shift in meaning. The 13 errors that require discourse analysis cannot be resolved as long as PP attachment and consequently parsing are treated as sentence-level tasks. This leaves 44/100 errors that should be solvable by future advancements in PP attachment models, i.e. the accuracy ceiling of the task on the dataset is expected to be around 92.6%.

Type	#
Incorrect	44
Irrelevant	36
Discourse	13
Other	7

Table 2: Error analysis of a random sample of 100 PPs that are incorrectly attached by the best model.

## 5 Conclusion

This paper evaluated a state-of-the-art PP attachment model that combines various insights about the task from the literature on a realistic data set with multiple attachment sites per preposition. We showed that by jointly using these insights, we obtain a very substantial improvement over previous approaches to the task. To answer the question where we stand in PP attachment, we conducted a manual analysis of attachment errors. This analysis showed that for this data set, the margin between the best models and the ceiling (approximately 92.6%) is quickly narrowing. Moreover, any improvements beyond that ceiling requires changes to gold standards to mark multiple correct structures and that certain ambiguities in PP attachment and parsing are resolved with discourse-level information.

The system discussed in this paper is largely language-independent, because it relies on word embeddings and bi-lexical preferences as the primary features. The only exception to this are the topological field features. However, we should point out that the topological field model is also used to describe clause structure in other Germanic languages (e.g. Haeseryn et al. (1997) and Zwart (2014)). Moreover, similar linear precedence constraints have been found for other language families, such as Slavic (Penn, 1999).

In the future, we would like to integrate and evaluate the PP attachment model that was dis-

cussed in this work in a dependency parser. Our aim is to use the representations formed by the feed-forward neural network as additional inputs to the transition classifier. This would combine the power of phrasal representations similar to those proposed by Belinkov et al. (2014) with bi-lexical preferences trained on large corpora.

## Acknowledgments

Financial support for the research reported in this paper was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center “The Construction of Meaning” (SFB 833), project A3.

## References

- Yonatan Belinkov, Tao Lei, Regina Barzilay, and Amir Globerson. 2014. Exploring compositional architectures and word vector representations for prepositional phrase attachment. *Transactions of the Association for Computational Linguistics*, 2:561–572.
- Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2009)*, pages 31–40.
- Hans Brunner, Greg Whittemore, Kathleen Ferrara, and Jiamiene Hsu. 1992. An assessment of written/interactive dialogue for information retrieval applications. *Human-Computer Interaction*, 7(2):197–249.
- Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.
- Daniël de Kok and Erhard Hinrichs. 2016. Transition-based dependency parsing with topological fields. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–7, Berlin, Germany, August. Association for Computational Linguistics.
- Daniël de Kok, Corina Dima, Jianqiang Ma, and Erhard Hinrichs. 2017. Extracting a PP attachment data set from a German dependency treebank using topological fields. In *International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 89–98.
- Daniël de Kok. 2015. Bootstrapping a neural net dependency parser for German using CLARIN resources. In *Proceedings of the CLARIN 2015 conference*.

- Erich Drach. 1937. *Grundgedanken der Deutschen Satzlehre*.
- John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.
- Kilian A. Foth and Wolfgang Menzel. 2006. The benefit of stochastic PP attachment to a rule-based parser. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 223–230, Sydney, Australia, July. Association for Computational Linguistics.
- Walter Haeseryn, Kirsten Romijn, Guido Geerts, Jaap de Rooij, and Maarten C. van den Toorn. 1997. *Algemene nederlandse spraakkunst*, volume 2. Martinus Nijhoff, Groningen, The Netherlands.
- Richard H.R. Hahnloser, Rahul Sarpeshkar, Misha A. Mahowald, Rodney J. Douglas, and H. Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951.
- Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120, March.
- Tilman Höhle. 1986. Der Begriff ‘Mittelfeld’. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne, editor, *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen*, pages 329–340.
- Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 448–456.
- Mark Johnson and Stefan Riezler. 2000. Exploiting auxiliary distributions in stochastic unification-based grammars. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference*, pages 154–161. Association for Computational Linguistics.
- Sandra Kübler, Steliana Ivanova, and Eva Klett. 2007. Combining dependency parsing with PP attachment. In *Fourth Midwest Computational Linguistics Colloquium*.
- Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency parsing*, volume 1. Morgan & Claypool Publishers.
- Jonathan K. Kummerfeld, David Hall, James R. Curran, and Dan Klein. 2012. Parser showdown at the Wall Street corral: An empirical investigation of error types in parser output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1048–1059, Jeju Island, Korea, July. Association for Computational Linguistics.
- 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, May–June. 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 C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.
- Seyed Abolghasem Mirroshandel, Alexis Nasr, and Joseph Le Roux. 2012. Semi-supervised dependency parsing using lexical affinities. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 777–785, Jeju Island, Korea, July. Association for Computational Linguistics.
- Patrick Pantel and Dekang Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 101–108, Hong Kong, October. Association for Computational Linguistics.
- Gerald Penn. 1999. Linearization and WH-extraction in HPSG: Evidence from Serbo-Croatian. In Robert D. Borsley and Adam Przepiórkowski, editors, *Slavic in Head-Driven Phrase Structure Grammar*, pages 149–182.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.
- Adwait Ratnaparkhi. 1998. Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1079–1085, Montreal, Quebec, Canada, August. Association for Computational Linguistics.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.
- Heike Telljohann, Erhard Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2006. *Stylebook for the Tübingen treebank of written German (TüBa-D/Z)*.

- Tim Van de Cruys. 2011. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 16–20, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Gertjan van Noord. 2007. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 1–10, Prague, Czech Republic, June. Association for Computational Linguistics.
- Martin Volk. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of Corpus Linguistics*, volume 200.
- Martin Volk. 2002. Combining unsupervised and supervised methods for pp attachment disambiguation. In *Proceedings of the 19th International Conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.
- Greg Whittemore, Kathleen Ferrara, and Hans Brunner. 1990. Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 23–30, Pittsburgh, Pennsylvania, USA, June. Association for Computational Linguistics.
- Jan-Wouter Zwart. 2014. *The syntax of Dutch*. Cambridge University Press.