EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

SFB 833
The Construction of Meaning

PP Attachment: where do we stand?

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

The problem

Prepositional phrase (PP) attachment is a well-known structural ambiguity in natural language parsing. PP attachment is particularly difficult because resolving ambiguities requires semantic cues.

PP PP

Wir	begrüßen	die	Neulinge	unter	uns	.
We	welcome	the	newcomers	under/among	us	.

Research questions

- There are several proposed techniques to improve PP attachment disambiguation in the literature. What is the influence of these different techniques?
- Do association scores learnt from large, automatically annotated corpora still improve PP-attachment disambiguation in a neural network/embedding setup?
- What problems are remaining for a state-of-the-art disambiguator?

PP attachment with multiple attachment sites

PP

Goetsch	bezog	sich	auf	einen	Vorschlag	der	Opposition	von	voriger	Woche	.
Goetsch	referred	(herself)	to	a	proposal	of-the	opposition	of	last	week	.

PP PP

Es	gibt	viel	Beifall	auf	der	Kundgebung	für	Margret	Mönig-Raane	.
There	is	much	approval	at	the	rally	for	Margret	Mönig-Raane	.

- PP attachment is traditionally treated as a binary classification task, where the attachment site is the *preceding noun* or the *main verb*.
- However: in reality multiple attachment sites at potentially large distances (3.15 on average in the German TüBa-D/Z treebank).

Association scores

- Normalized PMI [Bouma, 2009]:

$$\text{N-PMI}(x,y) = \left(\log \frac{p(x,y)}{p(x)p(y)}\right) \Big/ -\log p(x,y)$$

- Specific interaction information [Van de Cruys, 2011]:

$$\text{SI}_1(x,y,z) = \log \frac{p(x,y)p(y,z)p(x,z)}{p(x)p(y)p(z)p(x,y,z)}$$

- Specific correlation [Van de Cruys, 2011]:

$$\text{SI}_2(x,y,z) = \log \frac{p(x,y,z)}{p(x)p(y)p(z)}$$

Normalized PMI for the triples:

⟨CANDIDATE,OBJECT⟩
⟨CANDIDATE,PREPOSITION⟩
⟨CANDIDATE,PREPOSITION+OBJECT⟩

Specific interaction information and total correlation for the triple:

⟨CANDIDATE,PREPOSITION,OBJECT⟩

Topological fields

Topological fields are commonly used to describe the regularities in German word order. The distributions of syntactic relations vary significantly across topological fields, which can inform attachment choices, including PP attachment.

	VF	LK	MF	RK	NF
MC:	Gestern	hat	er häufiger	angerufen	als heute
	Yesterday	has	he more-often	called	than today
MC:	Er	ruft	häufig	an	
	He	calls	frequently	up	
SC:		der	noch häufiger	anruft	als er
		who	more often	calls	than him

Neural network scoring model

We use a feed-forward neural network to score each candidate attachment site, given the preposition (P), (preposition) object (Obj), and attachment site (S):

- **Input layer:** word embeddings, tag embeddings, and topological fields of P, Obj, and S. Normalized PMI, specific interaction information and total correlation scores as described above.
- **Hidden layer:** 100 neurons, with ReLU activations
- **Output layer:** The logistic function, trained to estimate the attachment probability.

The scoring model computes the probability of each attachment site and chooses the site with the highest probability.

Results

The association scores were computed using text from the German newspaper *taz* (1986 to 2009, 28.8 million sentences, 393.7 million tokens) after annotation with a state-of-the-art dependency parser. The PP disambiguator is trained using the dataset of [de Kok et al., 2017].

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.	**86.7%**

Qualitative analysis

We hand-classified a random sample of 100 errors made by the PP attacher.

Type	#
Incorrect	44
Irrelevant	36
Discourse	13
Other	7

This analysis shows that more than half of the remaining errors are irrelevant or the attachments could not be solved at the sentence level.