Detecting Multiword Expressions by Dependency Parsing

WG3

István Nagy T. and Veronika Vincze
University of Szeged, Hungary
{nistvan, vinczev}@inf.u-szeged.hu

Automatic detection of MWEs by dependency parsers in different languages

ENGLISH verb-particle constructions

- Penn Treebank has VPC annotation
- Bohnet and Stanford parsers were trained
- Evaluated the parsers on Wiki50, manually annotated for VPCs

HUNGARIAN light verb constructions

LVCs were manually annotated in the Szeged corpus
LVC-specific dependency relations
Trained and evaluated the parser on Szeged corpus with 10 fold cross validation

GERMAN light verb constructions

TIGER corpus has LVC annotation
Bohnet parser was trained on TIGER
Evaluated this model on JRC-Acquis manually annotated for LVCs

English VPCs in the Penn Treebank

- **VPC**:
  - Verb + particle: *show off*
  - Compositional or not
- The special relation of the verb and particle within a VPC is distinctively marked in the **Penn Treebank** (Marcus et al., 1993)
- It also has a specific syntactic label (**PRT**)

Turn the light off.
(S (NP-SBJ *) (VP turn (NP the light) (PRT off)))

Automatic detection of English VPCs

- Wiki50: full-coverage VPC annotated corpus where each individual occurrence of a VPC was manually annotated
- Examined how syntactic parsers can perform on Wiki50
- Applied the Stanford (Klein and Manning, 2003) and Bohnet (Bohnet 2010) parsers
- Only 52.57% and 58.16% of annotated VPCs on the Wiki50 had a verb-particle syntactic relation when we used the Stanford and Bohnet parsers
- The parsers achieved high precision scores of about 90%

Edge type	Stanford		Bohnet	
	#	%	#	%
Prt	235	52.57	260	58.16
Prep	23	5.15	107	23.94
Advmod	56	12.52	64	14.32
Sum	**314**	**70.24**	**431**	**96.42**
Other	8	1.79	1	0.22
None	125	27.97	15	3.36
sum	**447**	**100.00**	**447**	**100.00**

Method	Precision	Recall	F-score
Dictionary Lookup	49.77	27.5	35.43
Stanford Parser	**91.09**	52.57	66.67
Bohnet Parser	89.04	58.16	**70.36**

German LVCs in TIGER Corpus

- In the TIGER corpus (Brants et al. 2004), LVCs that consist of a **verb and a prepositional phrase** are annotated
- The **PP** is marked with the relation **CVC** (collocational verb construction)
- **Verb-object pairs** are **excluded** from the annotation

***Abschied nehmen* "to take leave"** – not an LVC here
***zur Diskussion bringen* "to discuss"**
(zur Diskussion)$_{CVC}$ bringen

Automatic detection of German LVCs

- The **Bohnet parser** was trained on the TIGER corpus
- Evaluated the model on the German part of the JRC-Acquis corpus, annotated for LVCs (Rácz et al. 2014)
- 84.81 (precision), 60.91 (recall) and **70.90 (F-score)**
- Same results as the English VPCs

Hungarian LVCs in the Szeged Treebank

LVC: noun + verb: *döntést hoz* "*make a decision*"

- Semi-compositional: the sense of the noun is dominant
- Verb + object, verb + PP, verb + other arguments
- In the **Szeged Dependency Treebank**, dependency relations between the two members of LVCs were enhanced with **LVC-specific relations** (Vincze et al., 2013)
- **OBJ-LVC** relation between the words *döntést* (*decision-ACC*) and *hoz* "*bring*", members of the LVC *döntést hoz "to make a decision"*.

Automatic detection of Hungarian LVCs

- The **Bohnet parser was** trained on the legal subdomain of the corpus.
- 10-fold cross validation was applied:
  - 86.60 (precision), 67.12 (recall), **75.63 (F-score)**
- Classification: two-stage procedure (Nagy et al. 2013)
  - Extract potential LVCs
  - Classify them
- Main advantages:
  - **High precision**
  - Proper treatment of the **non-contiguous LVCs**

Method	Precision	Recall	F-score
Dictionary lookup	78.49	12.29	21.25
Classification	82.84	67.60	74.45
Dependency parser	**86.60**	67.12	**75.63**

Method	Precision	Recall	F-score
Contiguous LVCs			
Classification	87.46	78.54	**82.76**
Dependency parser	**90.08**	73.57	80.99
Non-contiguous LVCs			
Classification	71.03	51.88	60.00
Dependency parser	**79.40**	53.62	**64.01**

References

Bohnet, Bernd 2010: Top accuracy and fast dependency parsing is not a contradiction. In: *Proceedings of Coling 2010, pp. 89–97.*

Brants, Sabine; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; König, Esther; Lezius, Wolfgang; Rohrer, Christian; Smith, George; Uszkoreit, Hans 2004: TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation 2, 597–620.*

Klein, Dan; Manning, Christopher D. 2003. Accurate unlexicalized parsing. In *Annual Meeting of the ACL*, volume 41, pages 423–430.

Nagy T., István, Vincze, Veronika, Farkas, Richárd 2013: Full-coverage Identification of English Light Verb Constructions. In: *Proceedings of IJCNLP 2013*, pp. 329-337.

Marcus, Mitchell P.; Santorini, Beatrice; Marcinkiewicz, Mary Ann 1993: Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics 19(2): 313–330.*

Rácz, Anita, Nagy T., István, Vincze, Veronika 2014: 4FX: Light Verb Constructions in a Multilingual Parallel Corpus. Accepted to *LREC 2014.*

Vincze, Veronika; Zsibrita, János; Nagy T., István 2013: Dependency Parsing for Identifying Hungarian Light Verb Constructions. In: *Proceedings of IJCNLP 2013.*

Vincze, Veronika; Nagy T., István; Berend, Gábor 2011: Multiword expressions and Named Entities in the Wiki50 corpus. In: *Proceedings of RANLP 2011.* Hissar, Bulgaria, pp. 289–295.