Detecting Multiword Expressions by Dependency Parsing

István Nagy T. and Veronika Vincze
University of Szeged

WG3: Statistical, Hybrid and Multilingual Processing of MWEs

In this poster, we present how different types of MWEs can be identified by dependency parsers in different languages. In our investigations, we focus on English verb-particle constructions (VPCs), Hungarian light verb constructions (LVCs) and German light verb constructions.

In our experiments, we exploit the fact that some treebanks contain MWE-aware annotations, i.e. there are MWE-specific morphological or syntactic tags in them. For instance, the French Treebank contains explicit annotations for MWEs (Abeillé et al. 2003) and different version of the Turkish Treebank are also annotated for MWEs (Eryiğit et al. 2011). Here, we make use of the Penn Treebank (Marcus et al., 1993), which contains annotation for VPCs, the TIGER corpus (Brants et al. 2004) and the Szeged Dependency Treebank (Vincze et al. 2013), both of which contain annotation for LVCs. In these treebanks, the special relation of the two components of the MWE is distinctively marked by a certain syntactic label. This entails that if a data-driven syntactic parser is trained on a dataset annotated with extra information for MWEs, it will be able to assign such tags as well, in other words, the syntactic parser itself will be able to identify MWEs in texts. In our experiments, we investigate the performance of such dependency parsers for three languages and two different MWE types.

English VPCs

The special relation of the verb and particle within a VPC is distinctively marked in the Penn Treebank, the particle is assigned a specific part of speech tag (RP) and it also has a specific syntactic label (PRT). Thus, parsers trained on the Penn Treebank are able to identify VPCs in texts.

We experimented with two dependency parsers, the Stanford parser (Klein and Manning 2003) and the Bohnet parser (Bohnet 2010) and examined how they can perform on the Wiki50 corpus (Vincze et al. 2011). This corpus contains English Wikipedia articles which are annotated for several types of MWEs – thus for VPCs as well – and named entities. We parsed the texts of Wiki50 with the two parsers, using their default settings and if the parser correctly identified a PRT label, we considered it as a true positive. For evaluation, we employed the metrics precision, recall and F-measure interpreted on VPCs. The two parsers obtained the following results: the Stanford Parser achieved 91.09 (precision), 52.57 (recall) and 66.67 (F-measure) and the Bohnet Parser achieved 89.04 (precision), 58.16 (recall) and 70.36 (F-measure). Thus, precision values are rather high but recall values are lower, which suggests that the sets of VPCs found in the Penn Treebank and Wiki50 may differ significantly.

Hungarian LVCs

The Szeged Dependency Treebank contains manual annotation for light verb constructions (Vincze et al. 2013). Dependency relations were enhanced with LVC-specific relations that can be found between the two members of the constructions. For instance, the relation OBJ-

LVC can be found between the words *döntést* (decision-ACC) and *hoz* "bring", members of the LVC *döntést hoz* "to make a decision".

We used the Bohnet dependency parser to identify LVCs in the legal subdomain of the corpus. We applied 10-fold cross validation here and got the following values: 86.60 (precision), 67.12 (recall), 75.63 (F-measure). According to the results and error analysis, the main advantages of the system are the high precision value on the one hand and the adequate treatment of non-contiguous LVCs on the other hand (Vincze et al. 2013).

German LVCs

In the TIGER corpus, LVCs that consist of a verb and a prepositional phrase are annotated with the relation CVC. The German model of the Bohnet parser trained on the Tiger corpus is able to assign such a label, so we used it in our experiments with its default settings. For evaluation, we selected a subset of the German part of the JRC-Acquis corpus, which has recently been annotated for LVCs (Rácz et al. 2014). If the parser correctly identified a CVC label, we considered it as a true positive. We obtained a result of 84.81 (precision), 60.91 (recall) and 70.90 (F-measure), which indicates that similar to English VPCs, the set of LVCs in the test corpus may just partly overlap with the set of LVCs in the TIGER corpus.

In the poster, we also give a comparative evaluation of the results on the three languages and offer a more detailed error analysis, which we believe to be beneficial for investigating parsing techniques for other types of MWEs and/or other languages as well.

References

Abeillé, Anne; Clément, Lionel; Toussenel, François 2003: Building a treebank for French. In Abeillé, Anne (ed.): Treebanks: building and using parsed corpora. Kluwer, Chapter 10.

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.

Eryiğit, Gülşen; Ilbay, Tugay; Can, Ozan Arkan 2011: Multiword expressions in statistical dependency parsing. In: *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pp. 45–55.

Klein, Dan; Manning, Christopher D. 2003. Accurate unlexicalized parsing. In: *Proceedings of ACL*, pp. 423–430.

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. Submitted to *LREC 2014*.

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.

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