Automatic Detection of Multiword Expressions with Dependency Parsers on Different Languages

István Nagy T.

Here, 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 [1] and different version of the Turkish Treebank are also annotated for MWEs [2]. Here, we make use of the Penn Treebank [3], which contains annotation for VPCs, the TIGER corpus [4] and the Szeged Dependency Treebank [5], 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, namely the Stanford parser [6] and the Bohnet parser [7] and examined how they can perform on the Wiki50 corpus [8]. 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 [9]. 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 [5].

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 [10]. 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.

Acknowledgements

István Nagy T. was partially funded by the State of a Hungary, co-financed by the European Social Fund in the framework of TAMOP-4.2.4.A/ 2-11/1-2012-0001 "National Excellence Program".

References

[1] Abeillé, A., Clément, L., Toussenel, F.: Building a Treebank for French. In: Treebanks : Building and Using Parsed Corpora. Springer (2003) 165–188

[2] Eryiğit, G., Ilbay, T., Can, O.A.: Multiword expressions in statistical dependency parsing. In: Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, Dublin, Ireland, Association for Computational Linguistics (October 2011) 45–55

[3] Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics **19**(2) (1993) 313–331

[4] Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: Linguistic interpretation of a German corpus. Research on Language and Computation **2**(4) (2004) 597–620

[5] Vincze, V., Zsibrita, J., Nagy T., I.: Dependency Parsing for Identifying Hungarian Light Verb Constructions. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (October 2013) 207–215

[6] Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Annual Meeting of the ACL. Volume 41. (2003) 423–430

[7] Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of Coling 2010. (2010) 89–97

[8] Vincze, V., Nagy T., I., Berend, G.: Multiword Expressions and Named Entities in the Wiki50 Corpus. In: Proceedings of RANLP 2011, Hissar, Bulgaria, RANLP 2011 Organising Committee (September 2011) 289–295

[9] Vincze, V., Csirik, J.: Hungarian corpus of light verb constructions. In: Proceedings of Coling 2010, Beijing, China, Coling 2010 Organizing Committee (August 2010) 1110–1118

[10] Rácz, A., Nagy T., I., Vincze, V.: 4FX: Light Verb Constructions in a Multilingual Parallel Corpus. In: Proceedings of LREC. (2014)