Swedish multiword expressions and sublanguage parsing – (WG 2 [1,4,3])

Dimitrios Kokkinakis <dimitrios.kokkinakis@svenska.gu.se>
Språkbanken and the Centre for Language Technology
University of Gothenburg, Sweden

Abstract: Multilayer, linguistic annotation of textual corpora (in specialized fields) is important for empirically-based, data-driven language processing and human language technologies. One such annotation is at the syntactic and functional level, i.e. parsing, which apart from the challenge it possesses (usual, deviant and idiosyncratic uses of vocabulary and syntax) is often required as a supporting technology in e.g., information extraction/retrieval, terminology management and knowledge acquisition. Moreover, in several sublanguages, where specialized processing tools have to be adapted which is considered too expensive for many applications and/or languages, means for reducing parsing complexity needs to be explored [5]. We have experimented with means of reducing parsing complexity in a Swedish scientific medical discourse setting. Parsing performance (for a specific task, see below) could be improved by applying a number of pre-processing steps, based on various types of multiword expressions (MWE), guided by the annotations provided by domain-specific lexica, named entity recognition (NER), and the identification of compound function words. The experiment we did used finite-state cascades [1], constituent based shallow parsing, applied on a sample of a Swedish medical text sample. Evaluation was calculated on the extraction of the syntactic relations *Subject* and *Object*. Prior to the evaluation, the parser was made aware of both the shallow semantic annotations from various domain-specific medical resources, the multiword expressions, both terminological ones and the ones obtained from the hybrid NER component as well as the compound function words.

The basic terminological resource used was the Swedish Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), which contains nearly 300,000 terms out of which roughly 14% consist of a single token, while the rest, 86%, are multiword terms, such as *akut lymfatisk leukemi* (acute lymphoblastic leukemia). NER uses, among its resources a set of 30,000 multiword named entities of various classes (person, location, organization, artifact, time, etc.) and length (2-5 tokens), as well as semantic patterns based on a combination of key words and regular expressions. The multiword named entities, such as *Sällskapet mot grymhet mot djur* (Society for the Prevention of Cruelty to Animals) and *Aco omega 3 forte*, have been semi-automatically collected from various Internet (re)sources. The pattern schemata (used also for capturing numerical expressions) have been manually created. For instance: `Journal" "(of" "the|of)(" "{U}[^ ]+)+(" of "{U}[^ ]+|" of the "{U}[^ ]+)?` can be read as strings that start by the keyword *Journal*, followed by *of the* or *of* and followed by a sequence of tokens that start by an upper case letter, and an optional part similarly to its predecessor. This particular pattern can capture e.g. both *Journal of the History of the Neurosciences* and also *Journal of Acquired Immune Deficiency Syndromes*. Finally, the multiword function words consist of 600 tokens, manually categorised as adverbs (*i vilket fall som helst*, in any case), determiners (*en och samma*, same), prepositions (*under loppet av*, during), pronouns (*var och en*, each) and conjunctions (*även om*, although) and added to the part-of-speech tagger's lexicon. In all cases, each MWE is represented as a single token, and

each of its constituents is joined with underscore. The part of speech tagger's lexicon (the parser uses the TnT: Statistical Trigram Tagging [2]) has been extended with all compound function words, manually added in its backup lexicon. Finally, the results from the NER, terminology and compound function word recognition are merged into a single representation format and fed into a syntactic analysis module that has been modified in such a way that can utilize the features provided by the pre-processors, which also results into the effect of slightly increasing the number of rules, but decreasing the complexity of the grammar rules. These pre-processing steps reduce ambiguity at the various levels of the linguistic processing. This can be attributed to: decrease of part-of-speech errors (since part-of-speech tags play a secondary role for the terms and entities, their semantic labels are weighted more), coordination, structural ambiguity reduction (for terms and entities) and simpler, more accurate caption of the functional labels (for instance, a single part-of-speech for a complex function word such as the adverb *i grund och botten* (basically) is easier to handle).

We manually evaluated and compared the results with and without the semantic annotations on a random sample of 200 lengthy sentences, using the syntactic relations as a benchmark in Subject-Verb-Object-triplets. The results showed that the semantic pre-processing resulted in a 16% improvement of accuracy for these relations (with a total precision/recall for Subject 94.3%/96.9% and for Object 88.4%/92.7%). That the recognition of multiword units has a positive effect in the improvement of the parsing results for Swedish and other languages, has been showed in other studies as well [3,4,7]. We believe that MWE have positive impact on syntactic analysis and *should* be identified prior to parsing. However, there are several types of MWE and several resources that can be used in order to achieve better parsing performance, independent of the application in mind, or textual characteristics. In the near future we foresee an increase of MWE, also with the addition of "new" types of suitable resources (at least for Swedish), particularly a *constructicon* [6]. Constructicon (<http://spraakbanken.gu.se/eng/swecxn>) is a database of Swedish constructions under development, a large-scale electronic resource for linguistic, lexicographic, and educational purposes, as well as for language technology applications. Such structures tend to be highly problematic for language technology, and also seem to be quite common. Yet, they are often neglected in both grammars and dictionaries, which focus on general rules and individual words, respectively. Hence, a comprehensive account of construction-specific patterns is so far lacking. Therefore, we started developing such a resource, aiming to make it descriptively adequate, simple enough for large-scale coverage, and formalized to enable computational uses.

References

[1] Abney S. 1997. Part-of-Speech Tagging and Partial Parsing. *Corpus-Based Methods in Language and Speech Processing*. Young S. & Bloothooft G. (eds). Chap. 4:118-136. Kluwer AP.

[2] Brants T. 2000. TnT – A Statistical Part-of-Speech Tagger. *Proc. of the 6th Applied Natural Language Processing (ANLP).* Seattle, WA.

[3] Constant M., Sigogne A. and Watrin P. (2012). Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing. *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics.* Pages 204–212, Jeju, Republic of Korea.

[4] Korkontzelos I. and Manandhar S. 2010. Can Recognising Multiword Expressions Improve Shallow Parsing? *Proc. Of the Annual Conference of the North American Chapter of the ACL*. Pages 636–644, Los Angeles, CA.

[5] Lease M. and Charniak E. 2005. Parsing Biomedical Literature. *IJNLP*. 58-69. LNAI 3651.

[6] Lyngfelt B., Borin L., Forsberg M., Prentice J., Rydstedt R., Sköldberg E. and Tingsell S. 2012. Adding a Constructicon to the Swedish resource network of Språkbanken. *Proc. of KONVENS (LexSem–workshop).* Pp. 452-461. Vienna.

[7] Nivre J. and Nilsson J. 2004. Multiword Units in Syntactic Parsing. *MEMURA 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications. Workshop at LREC.* Lisbon, Portugal.