WG2: **Improving Parsing Medical Discourse using Cascades of Multiword Expression Recognition**

Dimitrios Kokkinakis
Department of Swedish, Språkbanken
University of Gothenburg
Sweden
dimitrios.kokkinakis@svenska.gu.se

Abstract

The motivation of this study is to investigate text mining techniques in relation to a specific discourse, namely biomedicine. More specifically, we are interested on event extraction, where syntactic analysis is an important component for the task, which we believe can be more adequately addressed by identifying multiword expressions (MWEs) and integrating those to the syntactic parser. This paper discusses how a cascaded finite state constituent parser for Swedish behaves, with respect to parsing accuracy, where various types of MWE are successively recognized and introduced into the parser.

1 Introduction

There is a lot of research worldwide to define a typology of Multiword Expressions (MWEs). Thus far there has not been any agreement upon what are the types that should be part of such scheme and therefore different researchers incorporate various views on such content. This implies that a clear-cut consensus has not been yet reached on what to be included in MWE and what not. During the last 10-15 years a series of workshops on MWE have been held try to address that issue while at the same time great emphasis is put on computational approaches for the recognition and integration of MWE into Natural Language Processing (NLP) applications. The motivation of this work is to investigate text mining techniques in relation to a specific discourse, namely biomedicine. Using a cascaded finite state constituent parser (Abney, 1996) various types of MWE are successively introduced, and at each stage we show the benefits of the MWE recognition and annotation. MWE recognition results to a more compact text representation, simpler rules for the constituent recognition and improved precision and recall figures on the recognition of grammatical functions, for sentences that contain certain types of MWEs. The benefits of such approach are many, e.g. few new rules need to be added to the original cascades (i.e. to the general grammar rules) while the existing ones do not have to be substantially modified, since identified MWE can be assigned a single part-of-speech label. For instance general language multiword adverbs, determiners or prepositions have been manually added to the part-of-speech taggers[1] lexicon; *på grund av* 'because of' is already assigned the part-of-speech *preposition*, while lexicalized idioms received the part-of-speech *adverb*. On the other hand, entities and terminology are not automatically receiving a specific part-of-speech but each token in such constructions is assigned a conll-like IOB-*flag*, O (out), B (begin) or I (inside); e.g. *B-obstruktiv I-sömnapné* 'obstructive sleep apnea'. This flag is also augmented with M (medical term) or other, related to the various types of entities. Due to this, we chose to add a new set of rules to the parser (i.e. a new cascade) that can distinguish and appropriately label the MWE annotations obtained during terminology and entity annotation. This results into finer-grained syntactic labels. For instance, instead of the more generic 'np' $_{np}$*[Akademiska sjukhusets urologklinik]* 'the University hospital's urologic clinic', we get 'np-location' $_{np\text{-}location}$*[Akademiska sjukhusets urologklinik]* and instead of $_{np}$*[*obstruktiv sömnapné*]* 'obstructive sleep apnea', we get $_{np\text{-}medical}$*[obstruktiv sömnapné]*.

[1] For the part-of-speech annotation we use TnT (Brants, 2000) trained on Swedish texts. The tagset used is an augmented MULTEXT tagset for Swedish.

2 Experimental Setup

We randomly selected 10 articles (4000 tokens) from the Swedish medical association's journal[2]. These were tokenized and processed through a pipeline in which each module was dedicated to recognize several different distinct types of MWEs. In order, these annotation layers start with more general language, generic MWEs to more specific and domain ones, namely i) MWE function words (approximately 500), auxiliary verbs and (a set of) phrasal verbs ii) lexicalized idioms (none found in the sample) iii) a small subset of complex lexical items (i.e. constructions[3]) iv) named and time entities and v) medical terminology. The mean of all the 212 MWEs found in the sample was 2.36 tokens (*sd* 0.64).

2.1 Annotation and Input/Output Format

The MWE annotations obtained by the various layers previous outlined steps are normalized to the IOB format (iv and v), while the rest of the MWEs (i, ii, iii), are marked with underscore, *på grund av* 'because of' becomes *på_grund_av* which is a practical and simple way to capture all these cases. After the MWE recognition step follows part-of-speech tagging [4] and partial parsing. The partial parsing implies more, since the input to the parser is semantically enhanced and contains a mixture of semantic and morphosyntactic labels.

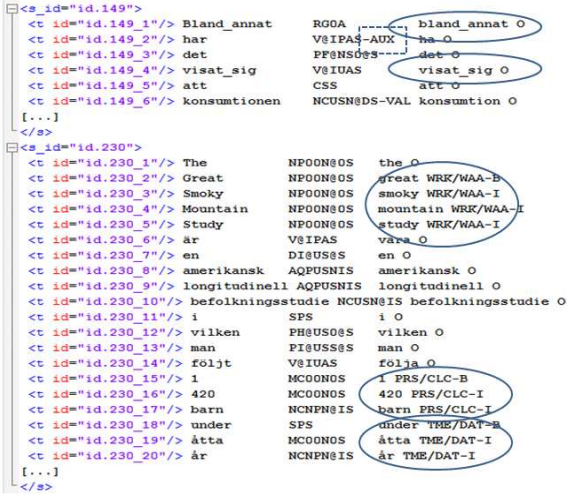

Figure 1: Input to the parser with MWE examples and their annotation.

[2] These 10 articles, both as raw text or annotated with all MWE automatically identified, are available by the author.
[3] See <http://spraakbanken.gu.se/eng/sweccn>.
[4] Note, that auxiliaries are marked as such using a *post* part-of speech filter which assigns the feature AUX to help verbs based on near context (see token with id 149_2 in Figure 1).

The fragments in Figure 1 illustrate some of these annotations. The first one (*Bland annat har det visat sig att konsumtionen [...]*, i.e. 'Among other things, it has been found that the consumption [...]') shows a MWE adverb, a phrasal verb and a help verb; while the second example (*The Great Smoky Mountain Study är en amerikansk longitudinell befolkningsstudie i vilken man följt 1 420 barn under åtta år [...]*, i.e. 'The Great Smoky Mountain Study is a U.S. longitudinal population study in which they followed 1420 children for eight years [...]') shows the recognition of a 'work" named entitiy.

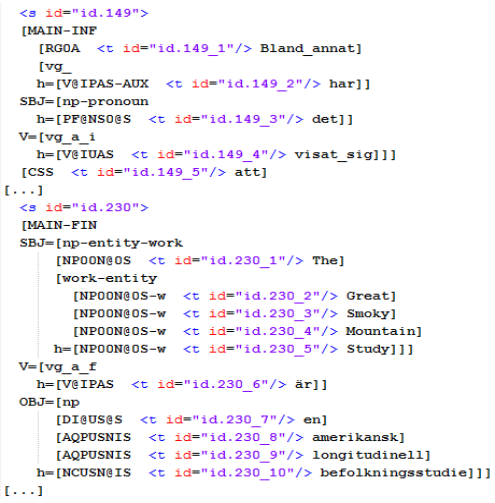

Figure 2: Parser output with MWE and their annotation ('h=' points to the head of a phrase and 'SBJ', 'OBJ' and 'V' are grammatical functions).

3 Results

As might be expected the best results are shown in sentences where longer named entities and medical terms are recognized. Table 1 shows bracketing Precision, Recall and F-scores, with the stepwise recognition of the MWEs, using the evalb scorer <http://nlp.cs.nyu.edu/evalb/>.

	Precision	Recall	F-score
i*	77.92%	76.79%	77.35%
iii*	77.96%	76.82%	77.39%
iv*	80.71%	79.93%	80.32%
v*	86.67%	87.10%	86.89%

Table 1. Evaluation results; '*' see Section 2.

References

Steven Abney. (1996). Partial Parsing via Finite-State Cascades. *Journal of Natural Language Engineering*, 2(4): 337-344.

Thorsten Brants. (2000). TnT – A Statistical Part-of-Speech Tagger. The *6th Conference on Applied NLP*. Pp 224-231. USA.