Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures *

Eduard Bejček, Pavel Straňák, Pavel Pecina
Charles University in Prague, Institute of Formal and Applied Linguistics
{bejcek,stranak,pecina}@ufal.mff.cuni.cz
WG4: Annotating MWEs in Treebanks

Abstract

We deal with syntactic identification of occurrences of multiword expression (MWE) from an existing dictionary in a text corpus. The MWEs we identify can be of arbitrary length and can be interrupted in the surface sentence. We analyse and compare three approaches based on linguistic analysis at a varying level, ranging from surface word order to deep syntax. We use the dictionary of multiword expressions SemLex, that includes deep syntactic dependency trees of its MWEs.

1 Introduction

The Prague Dependency Treebank (PDT) of Czech and the associated lexicon of MWEs SemLex offer a unique opportunity for experimentation with MWEs. In this proposal, we focus on identification of their syntactic structures in the treebank using various levels of linguistic analysis and matching algorithms. We compare approaches operating on manually and automatically annotated data with various depth of annotation.

The task of **identification** of MWE occurrences (as opposed to **acquisition**) expects a list of MWEs (types) as the input and identifies their occurrences (instances) in a corpus. This may seem to be a trivial problem. However, the complex nature of this phenomenon gives rise to problems on all linguistic levels of analysis: morphology, syntax, and semantics.

The definition of MWEs is rather unimportant for the purpose of this paper. We already have the decision work done: we have a lexicon of MWEs and we simply try to find all instances of these expressions (subtrees) in a text, whatever form the expression may take in a sentence.

*Presented at The 9th Workshop on MWEs

2 Data

In this work we use two datasets: Czech National Corpus (**CNC**), version SYN2006-PUB (600 mil. words, www.corpus.cz) and the Prague Dependency Treebank (**PDT**[1]), version 2.5 (approx. 0.8 mil. words that have three layers of annotation: morphological, analytical and tectogrammatical). We run and compare results of our experiments on both manual annotation of PDT, and automatic analysis of both PDT and CNC.

SemLex is the lexicon[2] of all the MWEs annotators identified during the annotation of PDT 2.5 t-layer (Straňák, 2010).

There are three attributes of SemLex entries crucial for our task:

LEMMATIZED – "Lemmatised basic form", i.e. take the MWE and substitute each word form with its morphological lemma. This attribute is used for the identification of MWEs on the morphological layer.[3]

TREE_STRUCT (TS) – A simplified tectogrammatical dependency tree structure of an entry. Each node in this tree structure has only two attributes: its tectogrammatical lemma, and a reference to its effective parent.

SURFACE_TREE_STRUCT – We extended the lexicon by adding the tree structure from surface syntactic layer. Given the annotated occurrences of MWEs in the t-layer and links from t-layer to a-layer, the extraction is straightforward. The most frequent structure is inserted if different surface syntactic trees are found.[4]

[1] Available at http://ufal.mff.cuni.cz/pdt2.5

[2] Available at http://ufal.mff.cuni.cz/lexemann/mwe

[3] Since it is lemmatized word by word, it may differ from basic form in human readable lexicons ("barking dogs never bite" → "to-bark dog ever to-bite").

[4] In reality the difference between t-layer and a-layer is unfortunately not as big as one could expect. Our

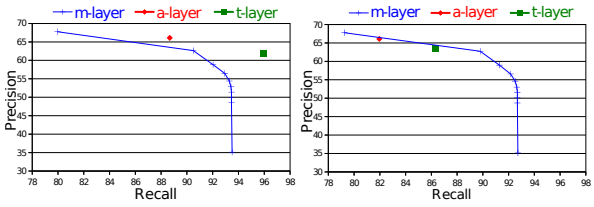

Figure 1: Precision–Recall scores of identification of MWE structures on manually/automatically annotated PDT.

3 Methodology of Experiments

SemLex – with its almost 9,000 types of MWEs and their 22,000 instances identified in PDT – allows us to measure accuracy of MWE identification on various interlinked layers of PDT 2.5. In this section, we present the method for identification of MWEs on t-layer in comparison with identification on a-layer and m-layer. The idea of using tectogrammatical TS for identification is that with a proper tectogrammatical layer[5] this approach should have the highest Precision (together with useable Recall).

Our approach to identification of MWEs in this work is purely syntactic. We simply try to find MWEs from a lexicon in any form they may take (including partial ellipses in coordination, etc.). We do not try to exploit semantics, instead we want to put a solid baseline for future work.

We assume that each occurrence of a given MWE has the same t-lemmas and the same **t-layer** structure anywhere in the text. This tectogrammatical *"tree structure"* (TS) is included in the lexicon. These TSs are taken one by one and we try to find them in the tectogrammatical structures of the input sentences. The criteria for matching are so far only t-lemmas and topology of the subtree.

The **a-layer** is processed in the same manner as t-layer: analytical TS is taken from the SemLex and the algorithm tries to match it to all a-trees.

MWE identification on the **m-layer** is based on matching lemmas (which is the only morphological information we use). The process is parametrised by a width of a window which restricts the maximum distance (in a sentence) of MWE components (irrespective of their order). This method naturally over-generates.

methods would benefit from more unified t-lemmas.

[5]as it is proposed in FGD (Sgall et al., 1986), i.e. with correct lemmatisation, added nodes in place of ellipses, etc.

4 Results

Effectiveness of our method evaluated against gold data in PDT 2.5 is visualised in Figure 1. The approach on the t-layer has the biggest Recall only when used on manually annotated data. Precision is always the best for m-layer approach with the smallest window of two words.

The third experiment with CNC data was manually evaluated on 546 sentences. The m-layer approach with smallest window has the maximal Precision (52) as well as F_1 measure (54); with unlimited window it achieves the maximal Recall (62).

Several reasons, why the t-layer results are not clearly better: 1. our representation of tree structures proved a bit too simple, 2. there are some deficiencies in the current t-layer parser, and 3. t-layer in PDT has some limitations relative to the ideal tectogrammatical layer.

5 Conclusions

The theoretically ideal approach based on t-layer turned out not to perform better, mainly due to the imperfectness of the t-layer implemented in PDT and also due to the low accuracy of automatic parser. It still shows very high Recall, as expected, however Precision is not ideal. Morphology-based MWE identification guarantees high Recall (especially when no limits are put on the MWE component distance) but Precision of this approach is rather low. Using analytical layer might be a good approach for many applications, too. It provides high Precision as well as reasonable Recall.

References

Sgall, P., Hajičová, E., and Panevová, J. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/Reidel Publ. Comp., Praha/Dordrecht.

Straňák, P. 2010. *Annotation of Multiword Expressions in The Prague Dependency Treebank*. Ph.D. thesis, Charles University in Prague.