Mining Phrases from Syntactic Analysis

Miloš Jakubíček, Aleš Horák, and Vojtěch Kovář

NLP Centre, Faculty of Informatics, Masaryk University,
Botanická 68a, 602 00 Brno, Czech Republic

Abstract. In this paper we describe the exploitation of the syntactic parser synt to obtain information about syntactic structures (such as noun or verb phrases) of common sentences in Czech. These phrases/structures are from the analysis point of view usually identical to nonterminals in the grammar used by the parser to find possible valid derivations of the given sentence. The parser has been extended in such a way that enables its highly ambiguous output to be used for mining those phrases unambiguously and offers several ways how to identify them. To achieve this, some previously unused results of syntactic analysis have been evolved leading to more precise morphological analysis and hence also to deeper distinction among various syntactic (sub)structures. Finally, an application for shallow valency extraction and punctuation correction is presented.

1 Introduction

Usually a derivation tree is presented as the main output of syntactic analysis of natural languages, but currently most of the syntactic analysers for Czech lack precision, i.e. there is large amount (actually, in some cases up to billions) of trees given on the output. However, there are many situations in which it is not necessary and sometimes even not desirable to have such derivation trees, may it be basic information extraction and retrieval, sentence segmentation, transformation of sentences into a predicate-arguments structure, shallow valency extraction or even punctuation correction. In such cases we rather need to mine whole phrases in the given sentence, especially noun, prepositional and verb phrases, numerals or clauses, each of them in a specific way. Moreover, so as not to end up with the same problems as with the standard parser output, we need to identify the structures unambiguously so that they are directly usable for further post-processing.

Therefore we modified the Czech parser synt so that it would be possible to obtain syntactic structures corresponding to the given nonterminal in a number of ways according to the user’s choice. To improve the structure detection we also employed the results of contextual actions used in synt as described in Section 4 which increased the precision of morphological analysis by almost 30%. We also present results of the mining from sample sentences as well as two example applications, first, the shallow extraction of verb valencies from annotated corpora and, second, the punctuation correction based on the described mining.

2 Syntactic Parser synt

Syntactic parser synt has been developed for several years in the in the Natural Language Processing Centre at Faculty of Informatics, Masaryk University. It performs
a chart-type syntactic analysis based on the provided context-free head-driven grammar for Czech. For easy maintenance this grammar is edited in form of a metagrammar (having about 200 rules) from which the full grammar can be automatically derived (having almost 4,000 rules). Contextual phenomena (such as case-number-gender agreement) are covered using the per-rule defined contextual actions.

In recent evaluation [2, p. 77] it has been shown that synt accomplishes a very good recall (above 90 %) but the analysis is highly ambiguous: for some sentences even billions of output syntactic trees can occur. There are two main strategies developed to fight such ambiguity: first, the grammar rules are divided into different priority levels which are used to prune the resulting set of output trees. Second, every grammar rule has a ranking value assigned from which the ranking for the whole tree can be efficiently computed to sort the trees on the output accordingly.

For the purpose of the mining process the internal parsing structure of synt is used, the so called chart, a multigraph which is built up during the analysis holding all the resulting trees. What is important about chart is its polynomial size [3, p. 133] implying that it is a structure suitable for further effective processing – as the number of output trees can be up to exponential to the length of the input sentence, processing of each tree separately would be otherwise computationally infeasible. By processing of the chart we refer to the result of the syntactic analysis, i.e. to the state of the chart after the analysis.

3 Mining Phrases

Several ways how to identify the phrases have been developed respecting the reality that the syntactic structures behind the phrases differ a lot and thus no universal procedure can be used for all of them. Since we want the output of the mining process to be unambiguous, we cover all possible structures found in the resulting chart data structure used during the syntactic analysis.

There are two straightforward approaches for structure detection which consist in extracting the biggest or smallest found structure, however, to achieve high quality results, more sophisticated methods have to be employed for each structure/nonterminal specifically. By speaking about biggest or smallest we mean that with regard to the fact that many of the rules in the grammar used by synt are recursive and thus we have to decide what recursion level should be extracted – e.g. in a simple sentence: Vidím starý velký dům we may extract both velký dům and starý velký dům as noun phrases. This decision is, however, not only nonterminal-specific, but depends also heavily on the particular application of the mining. Sample mining results for various nonterminals are listed in Examples 1–4.

- Example 1. – clause (nested)
  
  **Input:**
  Muž, který stojí u cesty, vede kolo.
  *(A man who stands at the road leads a bike.)*

  **Output:**
  [0–9]: Muž , , vede kolo *(a man leads a bike)*
  [2–6]: který stojí u cesty *(who stands at the road)*

1 In English: I see an old big house.