Introduction

Word sketch is a corpus-based summary of a word’s grammatical and collocational behavior that enables the extraction of collocations (and corpus examples) using the Sketch Engine tool. A detailed sketch grammar for Slovene, based on regular expressions over POS tags, was developed for the extraction of lexical data from the Gigafida corpus for the purposes of compiling Slovene Lexical Database. Since the adaptation of the MSTParser for Slovene, lexical data based on the same or similar grammatical patterns can also be extracted from parsed corpus data. We compare the differences between the two “sketch grammars” both in terms of general syntactic analysis (1) and MWE extraction and evaluation (2).

1. Comparison of Parsing Precision/Recall

1.1 Method

For the most frequent lemmas in the gold ssj500k corpus, we compared the recall and precision of both sketch grammars for extraction of collocates (i.e., dependents) within the given set of grammatical relations (4 for nouns, 4 for verb and 2 for adjectives), regardless of their MWE status.

1.2 Results

Overall (see →), for currently comparable set of grammatical relations (10) between selected heads (lemmas) and their dependents (collocates), dependency parser gives slightly higher precision and significantly higher recall. However, the results for both methods vary considerably depending on the type of grammatical relations (see ↓). The most significant differences can be observed in the recall for discontinuous syntactic relations (e.g., prepositional phrases), where dependent is often further away from the head.

2. Comparison of MWE Extraction

2.1 What MWEs were we interested in?

Phraseological units in SLD are defined as word combinations whose meaning or communicative function is not deducible from its parts and have phrasal meaning, as opposed to multi-word units, whose meaning remains non-phraseological.

2.2 Where did we extract them from?

The 100 million word Kres corpus is an extensive collection of Slovene texts with a balanced genre structure. It was sampled from the 1BW Gigafida corpus, with random paragraphs as basic sampling units to ensure better representation of the original Gigafida material.

2.3 What was our gold?

Slovene Lexical Database (SLD) consists of lexical data of various degrees of compositionality: 44,626 collocations, 7,151 grammatical patterns, 8,298 syntactic combinations (compositional), as well as 2,053 multi-word units and 1,446 phraseological units (non-compositional).

2.4 Method

For each of the 10 comparable grammatical relations (see ↓), we chose 3° random MWEs from SLD, whose phraseological core (bigram) can be described by such relation. For the head node of every bigram, we then compared the word sketch for the given gramrel, in particular: a) the position of the MWE collocate within the gramrel sketch (rank); b) the attributed collocational strength (logDice score); and c) the number of matched corpus concordances. Note that the latter does not imply recall, as the retrieved examples may or may not be relevant.

2.5 Results

See Table (⬇)

Idenical rank/score/no. of concordances for both
Higher rank/score/no. of concordances by RegEx-WS
Higher rank/score/no. of concordances by DepPars-WS

The two sketches give very similar results, i.e., high precision for extracting MWEs. The dependency-based word sketch attribute the MWEs a slightly greater collocational strength (logDice score), although this does not usually change the rank position of the collocate in question, as both sketches usually display the MWE collocates in the same (top-level) positions.

Future work

➤ determine syntactic patterns for all MWE types in the SLD database
➤ for these patterns, define comparable grammatical relations in both sketches
➤ develop procedures for automated comparison of the two methods in terms of MWE parsing, extraction and evaluation (beyond core bigrams)
➤ on the basis of results, build a hybrid model that combines the best features of both methods
➤ further explore the SLD gold standard of more than 60,000 MWEs for machine learning (extraction and MWE type classification)

References