A “sentence pattern” in modern Natural Language Processing and Computational Linguistics is often considered as a subsequent string of words (n-grams). However, in many branches of linguistics, like Pragmatics or Corpus Linguistics, it has been noticed that simple n-gram patterns are not sufficient to reveal the whole sophistication of grammar patterns. We present a language independent architecture for extracting from sentences more sophisticated patterns than n-grams. In this architecture a “sentence pattern” is considered as n-element ordered combination of sentence elements (words). This paper presents general description of the architecture.

1. This pattern cannot be discovered with only n-gram approach. It could be found with any pattern matching algorithm, however,
2. This pattern cannot be discovered if you don’t know what to look for (e.g. a new corpus of unknown domain).

Pattern Extraction and Frequency Calculation

1. Match all created patterns against the corpus
2. Extract
3. Calculate pattern frequency (PF)
4. Delete quasi patterns (patterns that appeared only once)
5. Calculate weights:

\[ w_p = \frac{PF}{all \quad quasi} \]

Conclusions and Future Work

We presented SPEC, an architecture for extraction of frequent sophisticated patterns from sentences. SPEC can deal with any corpora written in any language. SPEC provides significantly more frequent patterns than only n-gram method.

To do: Compare different combination algorithms. In pattern matching compare grep, regular expressions ad re2[10]. Implement parallel programing (fork, thread).

References