System for extraction of potential multi-word expressions and prediction of their translations from a multilingual corpus

WG 2: Parsing Techniques for MWEs

Katerina Zdravkova, University Sts Cyril and Methodius, Skopje
Aleksandar Petrovski, International Slavic University, Sveti Nikole

**Inspiration**

1. Example based machine translation (Makoto Nagao, 1984)
2. Statistical machine translation (Peter Brown, 1990)

**Goal**

- System capable of:
  - extracting potential multi-word expressions from sentence aligned parallel corpora
  - syntactical filtering in the source language
  - prediction of potential translation equivalents
  - evaluation of the obtained results with the reverse system

**Prerequisites**

- Corpora presented in xml format
  - FreeFormatter
  - OpenOffice.org XML File Format
- Sentence aligned corpus
  - Hunalign
  - Gargantua
- Manual polishing of sentence aligned corpora is inevitable

**Preprocessing phase**

- Preprocessing:
  - Remove all the punctuation marks and capitalization
- Result:
  - Each sentence is a continuous string of lowercase characters separated by spaces

**Phase 1: Extraction of potential MWEs**

- Determine MaximumLength, the length of the longest string of characters existing in the source language
- Repeat until MaximumLength = 2
  - Assign the string with MaximumLength to LongestString
  - Determine the frequency of the LongestString
  - IF Frequency(LongestString) >1
    - Store the LongestString in the list of unique multi-word expressions
  - Remove all the exact matches from the text to get a unique appearance of potential multi-word expressions
  - Decrease MaximumLength
- Store the list of potential unique multi-word expressions for further phases in the PotentialUniqueMWE

**Phase 2: Syntactical filtering of potential MWEs**

- Determine a set of syntactic rules which create plausible multi-word expressions in the source language and create syntactic grammars
- Apply the grammars to the list of potential multi-word expressions
- Manually polish the list of filtered plausible multi-word expressions
- Store the list of filtered multi-word expressions in FilteredUniqueMWE

**Phase 3: Prediction of potential translation equivalents**

- Sort FilteredUniqueMWE in an ascending order according to their frequency
- For each filtered source multi-word expression from FilteredUniqueMWE
  - Assign the list of target sentences corresponding to the first source sentence to ListOfTargetSentences
  - Assign the list of the target sentences corresponding to next source sentence to ParallelListOfTargetSentences
  - Find the intersection between each element of ListOfTargetSentences and each element of ParallelListOfTargetSentences and assign the non-empty intersection to IntersectionOfTargetSentences
  - Intersect the IntersectionOfTargetSentences with the list of all remaining lists of target sentences
  - If IntersectionOfTargetSentences is non-empty, store the potential target multi-word expression in GeneratedTargetMWE

**Phase 4: Evaluation of target MWEs**

- Extract the potential multi-word expressions in the target language using the steps from phase 1.
- Store the list in PotentialTargetMWE
- Compare each potential multi-word expression from PotentialTargetMWE with all multi-word expression GeneratedTargetMWE
- Whenever the intersection of PotentialTargetMWE with GeneratedTargetMWE is non-empty pairs, store it to TargetMWE
- Store the pair (FilteredUniqueMWE, TargetMWE)

**Advantages**

- Language independent
- It can be implemented for any pair of bilingual parallel texts, which have been previously sentence aligned

**Implementation**

- Extraction of potential MWEs: done
- Syntactical filtering of obtained potential MWEs: partially done
- Prediction of potential translation equivalents: under construction
- Evaluation of target MWEs: start, March 2014