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

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

In 1984, Makoto Nagao published his machine translation framework between Japanese and English based on the analogy of learning a foreign language [1]. He claimed that while translating one language to another: “Man does not translate a simple sentence by doing deep linguistic analysis”. His idea that “Man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases (very often, into case frame units), then, by translating these fragmental phrases into other language phrases, and finally by composing these fragmental translations into one sentence.**“** started the era of example-based machine translation. In the 21st century, his, at that time, radical and rather unfeasible approach was powered by so called statistical machine translation, which looks for patterns in parallel corpora and matches the most appropriate target pattern to source one [2].

Inspired by both approaches, we propose a system capable of extracting potential multi-word expressions existent in the sentence aligned parallel corpora, their syntactical filtering in the source language, prediction of potential translation equivalents and finally, evaluation of the obtained results with the reverse system.

The prerequisite for the creation of the system are multilingual corpora of sentence aligned texts, preferably given in an xml text format. If the texts are in another format, they can be easily converted into xml using many free formatters, such as FreeFormater or OpenOffice.org XML File Format. If the parallel corpus is not sentence aligned, bilingual sentence aligners, for example, Hunalign or Gargantua can produce satisfactory results. Manual polishing is highly recommended, because all sentence aligners produce mistakes, predominantly connected with the delimiters [3].

All the languages in the multilingual parallel corpora should first pass through a preprocessing phase, where all the punctuation marks and capitalization are removed. After this phase, each sentence becomes a continuous string of lowercase characters separated by spaces. The system goes through these phases:

```
1. Extraction of potential multi-word expressions in the source language from the source
   sentences
     i. Determine MaximumLength, the length of the longest string of characters existing
        in the source language
    ii. Repeat until MaximumLength = 1
          a. Assign the string with MaximumLength to LongestString
          b. Determine the frequency of the LongestString
          c. If Frequency(LongestString) >1
               1. Store the LongestString in the list of unique multi-word
                  expressions
               2. Remove all the exact matches from the text to get a unique
                  appearance of potential multi-word expressions
          d. Decrease MaximumLength
   iii. Store the list of potential unique multi-word expressions for further phases in the
        PotentialUniqueMWE
2. Syntactical filtering of obtained potential multi-word expressions for the languages that
   have an annotated dictionary
     i. Determine a set of syntactic rules which create plausible multi-word expressions
        in the source language and create syntactic grammars
    ii. Apply the grammars to the list of potential multi-word expressions.
```

iii. Manually polish the list of filtered plausible multi-word expressions

iv. Store the list of filtered multi-word expressions in **FilteredUniqueMWE**

3. Prediction of potential translation equivalents of filtered multi-word expressions

    i. Sort FilteredUniqueMWE in an ascending order according to their frequency

    ii. For each filtered source multi-word expression from FilteredUniqueMWE

    - a. Assign the list of target sentences corresponding to the first source sentence to ListOfTargetSentences
    - b. Assign the list of the target sentences corresponding to next source sentence to ParallelListOfTargetSentences
    - c. Find the intersection between each element of ListOfTargetSentences and each element of ParallelListOfTargetSentences and assign the non-empty intersection to IntersectionOfTargetSentences
    - d. Intersect the IntersectionOfTargetSentences with the list of all remaining lists of target sentences

    iii. If IntersectionOfTargetSentences is non-empty, store the potential target multi-word expression in **GeneratedTargetMWE**

4. Evaluate target multi-word expressions with the list of potential multi-word expressions from the target language

    i. Extract the potential multi-word expressions in the target language using the steps from phase 1.

    ii. Store the list in **PotentialTargetMWE**

    iii. Compare each potential multi-word expression from PotentialTargetMWE with all multi-word expression GeneratedTargetMWE

    iv. Whenever the intersection of PotentialTargetMWE with GeneratedTargetMWE is non-empty pairs, store it to **TargetMWE**

    v. Store the pair (FilteredUniqueMWE,TargetMWE)

The proposed system is language independent. It can be implemented for any pair of bilingual parallel texts, which have been previously sentence aligned. Its prototype is currently under construction. It comprises the resources of 16 languages belonging to the fourth version of Multext-East [4]. In this corpus, each language is presented in the xml format, which was the main prerequisite of the system. Furthermore, all the languages are mutually sentence aligned, enabling a consistent evaluation of the correctness and the efficiency of the system. The first two phases have already been completely finished, and the results of the approach are quite promising.

References:

1. Nagao, Makoto. "A framework of a mechanical translation between Japanese and English by analogy principle." (1984): 173-180.

2. Brown, Peter F., et al. "A statistical approach to machine translation. "*Computational linguistics* 16.2 (1990): 79-85.

3. Zlatkovska, M., Zdravkova, K. “Recommendations for the Improvement of Sentence Aligners”, Proceedings of the multidisciplinary conference ‘Language and Culture Interactions via Translation and Interpreting’, 26 - 28.9. 2012, Skopje (in print)

4. Erjavec, T. “MULTEXT-East: morphosyntactic resources for Central and Eastern European languages”, Language Resources and Evaluation, Volume 46, Issue 1 (2012): 131-142.