Identifying Multi-Word Expressions from Parallel Corpora with String-Kernels and Alignment Methods

WG1 - WG3

Parseme 5th GM
Iași, 23-24 September 2015

Johanna Monti
jmonti@uniss.it

Federico Sangati
federico.sangati@gmail.com

Mihael Arcan
mihael.arcan@deri.org

Identification of MWE based on **non-literal translatability property**

an MWE cannot be translated from one language to another on a word-for-word basis.

MANUALLY

EVALUATION

AUTOMATICALLY

Annotating the English-Italian TED parallel corpus (WIT^3**)** with MWE-TED corpus: 1500 sentences.

Three phases:

1. **Individual annotation**: 13 annotators, each sentence assigned to at least 2 annotators.
2. **Inter-annotation check**: each annotator confirms or changes the annotations after being confronted with the others' annotations.
3. **Validation**: integration and resolution of possible annotation conflicts.

Two-stage process:

1. Identifying a list of *potential MWEs* pairs via **Parallel String-Kernel** (including discontinuous sequences).
2. Filter out those candidates which are *not MWEs* using
   a. alignment information
   b. co-occurrence statistics