Multi-Word Expressions in a parallel bilingual spoken corpus: data annotation and initial identification results

PARSEME-5 – WG1, WG3

Johanna Monti
University of Sassari
Sassari, Italy
jmonti@uniss.it

Federico Sangati
Fondazione Bruno Kessler
Trento, Italy
sangati@fbk.eu

Mihael Arcan
National University of Ireland
Galway, Ireland
mihael.arcan@deri.org

Multiword expressions (MWEs) represent one of the major challenges for all Natural Language Processing (NLP) applications and in particular for Machine Translation (MT) (Sag et al., 2002). Their morpho-syntactic, semantic and pragmatic idiomaticity (Baldwin and Kim, 2010) together with translational asymmetries, i.e. the differences between an MWE in the source language and its translation, prevent technologies from using systematic criteria to properly process and translate MWEs.

The current work presents the first results of a new methodology for the identification of Multiword Expressions (MWEs) which meet the needs of the MT community to have parallel corpora annotated with MWEs, useful both for SMT training purposes and MWE translation quality evaluation. Our approach is inspired by one of the properties that characterises the majority of MWEs, which goes under the name of non-literal translatability, i.e. an MWE cannot be translated from one language to another on a word by word basis (Sag et al., 2002; Monti, 2012). This property is mainly shared by idioms (e.g., it's raining cats and dogs → it. *sta piovendo cani e gatti), but also by many collocations (e.g., heavy rain → it. *pioggia pesante), fixed expressions (e.g., by and large → it. *da e largo), proverbs (e.g., there's no such thing as a free lunch → it. *non esiste una cosa come un pranzo gratuito), phrasal verbs (e.g., Bring somebody down → it. *Portare qualcuno giù) among others.

The methodology is based on a two-stage process: in the first stage we use parallel String-Kernel methods for the identification of candidate MWEs and their corresponding translations, whereas the second stage is aimed at filtering incorrect candidate pairs. We have refined this methodology while developing the English-Italian MWE-TED corpus, which contains 1.5K sentences and 31K EN tokens. We have used the WIT^3 web inventory (Cettolo et al., 2012) which offers access to a collection of transcribed and translated talks. The core of WIT^3 is the TED Talks corpus, that basically redistributes the original content published by the TED Conference website.

The focus of our poster is to provide a description of the first results of our research, and in particular the general approach, guidelines and procedures adopted for annotating the parallel corpus with all MWEs (with no restrictions to a specific type) together with the MWE annotation statistics.

Acknowledgement We greatly acknowledge the PARSEME IC1207 COST Action for supporting this work. We also thank the three anonymous reviewers for the very useful comments.

References

Timothy Baldwin and Su Nam Kim. 2010. *Handbook of Natural Language Processing*, chapter Multiword Expressions, pages 267–292. 1. CRC Press, Boca Raton, USA, second edition edition.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268. Trento, Italy.

Johanna Monti. 2012. *Multi-word unit processing in Machine Translation - Developing and using language resources for Multi-word unit processing in Machine Translation*. Ph.D. thesis, University of Salerno.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.