Identifying Multi-Word Expressions from Parallel Corpora with Kernel Methods and Crowdsourcing

Federico Sangati
FBK, Trento
sangati@fbk.eu

Andreas van Cranenburgh
Huygens ING & University of Amsterdam
andreas.van.cranenburgh@huygens.knaw.nl

Johanna Monti
University of Sassari
jmonti@unisa.it

Introduction

We propose a new methodology for the identification of MWEs from parallel multilingual corpora. Our approach is inspired by one of the most significant properties that characterizes the majority of MWEs, which goes under the name of *non-translatability*: an MWE cannot be translated from one language to another on a word by word basis (Sag et al., 2002; Monti, 2012). The methodology envisions a three-stage process. The first phase makes use of automatic kernel methods for the identification of possible candidate pairs of expressions which has high recall and low precision. In the second phase a "word by word" automatic translation system will filter out those candidate pairs which are literal translations (and therefore not MWEs). In the third phrase, a crowdsourcing system is employed to further validate the list of final candidates.

MWE and non-translatability

There are several types of MWEs that do not admit a literal translation from one language into another, such as:

Fixed expressions e.g., EN. by and large → IT. *da e largo.

Idioms e.g., EN. Call it a day → IT. *Chiamarlo un giorno.

Proverbs e.g., EN. There's no such thing as a free lunch → IT. *Non esiste una cosa come un pranzo gratuito.

Phrasal verbs e.g., EN. Bring somebody down → IT. *Portare qualcuno giù.

A number of MWEs can be translated literally to all other languages, such as proper names and universal proverbs. These are therefore excluded from the scope of the current work. We now briefly describe the three phases we propose.

Phase 1. Kernel methods

The initial phase of our methodology assumes the availability of a large bilingual corpus structured as an aligned list of sentences. These are by now largely available to the scientific community (e.g., Koehn, 2005). The kernel methodology will aim at identifying parallel pairs of sentences which potentially contain MWEs (in one language, in the other, or in both). More precisely, at every iteration the algorithm will detect a pair of sentences in one language which share a certain expression (e.g., "call it a day" in a pair of English sentences) for which the parallel pair of sentences in a different language also shares an expression (e.g., "passare oltre" in a pair of Italian sentences). See table 1 for an illustration.

English	Italian
I feel we will have to **call it a day** at this point.	Credo che a questo punto dobbiamo **passare oltre**.
He would like us to adjourn the vote to the next part-session and **call it a day** for now.	Il relatore chiede di rinviare la votazione alla prossima seduta e, per ora, di **passare oltre**.

Table 1: Example of a parallel text in which the English MWE "call it a day" is translated into the Italian MWE "passare oltre".

The adopted methodology is based on the intuition that a translation of an expression A into B can be considered reliable if we can find two separate sentences in the source language containing A which are paired to two sentences of the target language containing B.

The kernel methodology can be applied efficiently via string kernel methods (Lodhi et al., 2002; Rousu and Shawe-Taylor, 2005) if the parallel data is made of aligned text, while tree kernels can be employed (Collins and Duffy, 2001; Moschitti, 2006; Sangati et al., 2010; van Cranenburgh, 2014) if a parallel treebank is available.

Phase 2. MT filtering

The first phase just described is prone to find many pairs of candidate expressions which do not include MWEs. More precisely table 2 lists all possible outcomes. The last one (4.), includes all non-relevant matches which are literal translations and therefore not MWEs.

	English	Italian
1.	MWE *bring up to date*	× *modernizzare*
2.	× *he died*	MWE *ha tirato le cuoia*
3.	MWE *call it a day*	MWE *passare oltre*
4.	× *aims at adapting*	× *mira ad adattare*

Table 2: All possible outcomes in detecting MWEs from a pair of candidate expressions.

In order to remove the candidate pairs not including MWEs, we envision to adopt a traditional "word by word" translation system, that would be able to detect which candidate pairs are likely to be literal translations of each other.

Phase 3. Crowdsourcing

As a final phase of the methodology we conceive of a crowdsourcing platform which would ask a set of users to manually validate the final list of candidates. This can be done by means of established platforms such as Amazon Mechanical Turk or CrowdFlower or alternatively by making use of educational tools to be proposed to second language learners.

Working groups concerned

WG1: Lexicon-Grammar Interface

WG3: Statistical, Hybrid and Multilingual Processing of MWEs

References

Collins, Michael and Nigel Duffy. Convolution Kernels for Natural Language. In Dietterich, Thomas G., Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*, pages 625–632. MIT Press, 2001.

Koehn, Philipp. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2: 419—444, 2002.

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

Moschitti, Alessandro. Making Tree Kernels Practical for Natural Language Learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Rousu, Juho and John Shawe-Taylor. Efficient computation of gapped substring kernels on large alphabets. *J. Mach. Learn. Res.*, 6:1323–1344, Dec. 2005.

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

Sangati, Federico, Willem Zuidema, and Rens Bod. Efficiently Extract Recurring Tree Fragments from Large Treebanks. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association.

van Cranenburgh, Andreas. Linear average time extraction of phrase-structure fragments. *Computational Linguistics in the Netherlands Journal*, x (accepted for publication):x–y, 2014.