Cross-lingual MWE detection using Distributional Representations and Sense Induction

Participant: Serge Sharoff
Centre for Translation Studies
University of Leeds
Host: Chris Biemann
Department of Informatics
University of Hamburg

Reference: ECOST-STSM-IC1207-130217-080673

Abstract

During my two week visit to the University of Hamburg, I worked with with Chris Biemann and colleagues in his group on detection of multiword expressions. We aimed at describing a specific kind of irregular MWEs, such as *by and large* or *of course*, which do not have an easily interpretable internal syntactic structure. For them, we suggest using a new term: 'closed-class MWEs' in contrast to other much wider classes of such MWEs as light-verb constructions or Named Entities. We propose a method for their supervised detection which relies on clustering the distributional similarity profiles of both single and multiword expression candidates. The model has been outlined during this visit, we will develop and test it with a number of typologically diverse languages. This will result in at least one paper as well as in datasets for closed-class MWEs in at least 20 languages.

1 Purpose of the STSM

The assumption of 1-to-1 correspondence between the tokens and the linguistic units is a source of frequent problems in POS tagging and parsing. This is specifically important in the case of irregular MWEs, which do not conform to many syntactic constraints expected from the well-formed constructions in a language. For example, *by and large* involves a preposition and an adjective connected by a coordinating conjunction, while *of course* violates the assumption that the preposition of joins two noun phrases. It is not reasonable to assign a POS tag or a syntactic dependency link for individual elements in such constructions. On the other hand, many other MWE constructions such as *kick the bucket*, *make a decision* or *New York* have a predictable syntactic structure, while they are still not compositional in their meaning.

Because of their inflexible internal syntactic structure the irregular MWEs lack the possibility of variation and modification. In the discussion below, we will refer to them as 'closed class MWEs' similarly to closed-class words to distinguish them from the more numerous syntactically interpretable MWEs.

Exhaustive lists of closed-class MWEs are not often not available for many languages. Even when they are, such lists are not tuned to the NLP tasks. Also their detection received relatively less attention, they are only mentioned in passing in the most recent MWE workshops.

The goal of this study is to enhance the resources in the Universal Dependencies (UD) treebanks (Nivre et al., 2016) by detecting fixed idiomatic MWEs with high precision.

2 Description of the work carried out during the STSM

For identifying such MWEs we developed a methodology which is is an extension of the method proposed in (Riedl and Biemann, 2015) combined with the use of UD corpora.

1. generating a list of initial MWE candidates via the distributional similarity, frequency and containment filters;
2. using the JoBimText framework (Biemann and Riedl, 2013) to generate topic models on top of Jobim similarity measures to compare the multiword expressions to other single or MW expressions, which are distributionally most similar to them;
3. represent each MWE candidate as a sequence of its POS tags in two versions, a relatively resource-heavy one, which assumes that a reliable POS tagger available, and a lighter one, which is based on unsupervised POS clustering, for example, using Marlin (Müller and Schuetze, 2015)

3 Description of the main results obtained

We experimented with several sources of information for ranking:

1. for each MWE candidate we compare the probability of the lexical sequence against the probability of its POS sequence approximated by the clusters:

Table 1: Examples output for German trigrams ranking by the ratio of POS frequency / lexical frequency

POS trigram	Fp	Word trigram	Fw	Fw/Fp	Comment
APPO KOKOM ADV	568	nach wie vor	559	1.0161	
APPO KOKOM APPR	420	nach wie vor	399	1.0526	
KOUS KON KON	119	Wenn und Aber	110	1.0818	
APPR PWAV APZR	90	von wo aus	83	1.0843	
ADV KON PTKANT	72	ja oder nein	63	1.1428	
KON PRF PTKZU	119	und sich zu	100	1.19	
APPR PWAV APPR	76	von wo aus	63	1.2063	
PIDAT APPR APPR	518	vielen Dank fr	422	1.2274	*vielen Dank*
KOUI PRF PTKZU	234	um sich zu	184	1.2717	
KOUS KON KOKOM	238	ob und wie	184	1.2934	
FM KON CARD	423	I und II	322	1.3136	world wars
FM APPR FM	72	and in the	53	1.3584	engl.
PIS PRF KOKOM	260	man sich als	188	1.3829	
PTKVZ KON PTKA	602	ab und zu	424	1.4198	
KON KON KON	230	und und und	160	1.4375	
PTKVZ KON PTKZU	68	ab und zu	46	1.4782	
PIS PRF PDS	144	man sich das	94	1.5319	
APPR KON PTKA	88	ab und zu	55	1.6	
APPR KOUS KON	144	ohne Wenn und	85	1.6941	→ ”ohne Wenn und Aber”
PTKVZ KOKOM APPR	1204	nach wie vor	706	1.7053	

How much is the lexical sequence more likely than the grammatical sequence; $P(nach\ und\ nach)$ vs $P(PREP\ CONJ\ PREP)$.

2. variability in phrase alignments to a closely related language, as well as to a more distant one.
3. logistic regression for classification on the basis of fixed MWEs marked in the UD treebanks.

4 Future collaboration and foreseen publications

There will be at least one paper, which will be the direct outcome of this work. We aim at submitting it to EMNLP, deadline 14 April. This STSM also led to established cooperation which will lead to two more strands of joint research. One concerns detection of Named Entities, another kind of MWEs, cross-lingually. Another is cross-lingual sense induction for common words, which is beyond MWEs, but still useful in many NLP tasks.

5 Confirmation by the host institution of the successful execution of the STSM

Prof Dr Chris Biemann During the successfully executed visit of Serge Sharoff at the University of Hamburg, we have made great progress on several research projects in natural language processing and multi word units. Specifically, we identified the notion of 'closed class MWEs' and jointly developed a methodology to identify them from monolingual data only. This method will be evaluated on data from the universal dependency treebank and will be described in an upcoming joint publication, most likely submitted to the EMNLP conference. Serge's visit was very inspiring and a great success. I wish he would visit more often.

References

Biemann, C. and Riedl, M. (2013). Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *Journal of Language Modelling*, 1(1):55–95.

Müller, T. and Schuetze, H. (2015). Robust morphological tagging with word representations. In *Proc NAACL*, Denver, Colorado.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proc LREC 2016*.

Riedl, M. and Biemann, C. (2015). A single word is not enough: Ranking multiword expressions using distributional semantics. In *Proc EMNLP*, pages 2430–2440, Lisboa, Portugal.