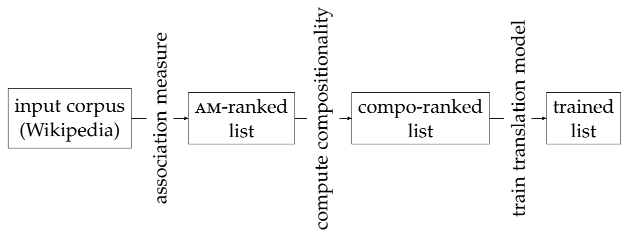
Integration of automatically-acquired multiword expressions in a hybrid machine translation system

Will Roberts and Markus Egg

Department of English and American Studies, Humboldt-Universität zu Berlin

Multiword expression (MWE) acquisition



- Unrestricted identification of MWEs by collecting lexical co-occurrence statistics on all words in Wikipedia.
- Limited pre-processing of the text prior to MWE identification:
  - Extract plain text from the Wikipedia dumps; Segment text into sentences; tokenize and strip out URLs using regular expressions; Remove all punctuation.
  - No further processing (POS-tagging, lemmatisation, case normalisation, removal of numbers or symbols).
  - Unlemmatised text may be useful for capturing the morphological and syntactic fixedness of some idiomatic MWEs (e.g., *spill the beans* but not *spill the bean*).
- Rank MWE candidates using the log-likelihood association measure.
  - Collect word frequency information using the SRILM language modelling toolkit.
  - Count *n*-grams with *n* up to 3 (i.e., we treat MWEs as bigrams and trigrams).

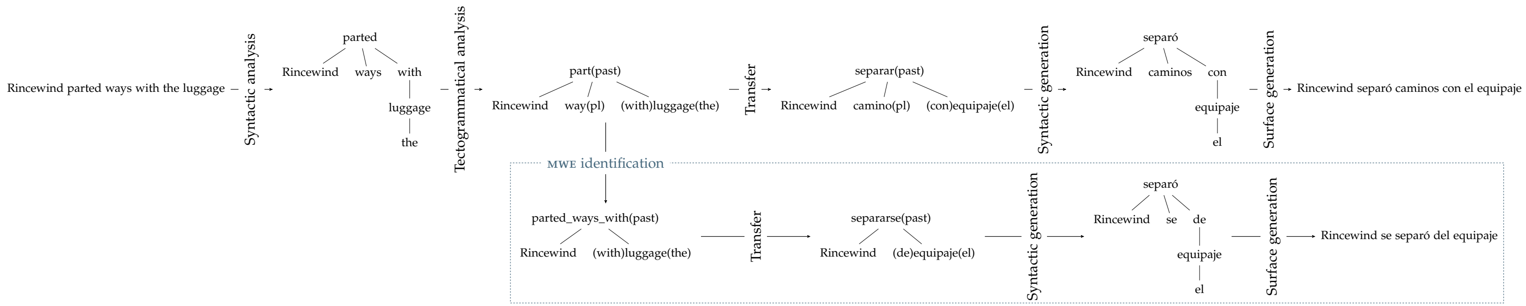
Compositionality ranking

- Take the top 10% from each association-measure-ranked list of MWEs and re-rank these candidates in order of increasing compositionality.
- Based on Salehi et al. (2015), this makes use of word embeddings constructed using `word2vec`:
  - Build a vector representation for every word in the vocabulary, as well as for every MWE, using the extracted Wikipedia text.
  - Greedy string search-and-replace of all occurrences of MWEs.
  - Replace each of these with a single words-with-spaces token.
- Problem: greedy rewriting cannot handle MWEs which overlap.
- Solution: split MWEs into batches with no overlaps.
  - Each batch produces a word embedding space.
  - Compute compositionality scores, and merge batches back together.
- Compositionality score: cosine similarity of MWE vector with its constituent words (arithmetic mean).
  - Do not compute similarity with "stop words" (the 50 most frequent words in the vocabulary).

0.005	*a front for*	–	0.005	–
0.012	*red tape*	-0.056	0.081	
0.191	*stops short of*	0.285	0.097	–

Integration into the TectoMT machine translation system (English-Spanish)

- TectoMT (Žabokrtský et al., 2008) is a hybrid machine translation system built on a pipeline model; statistical analysis phases (e.g., parsing, transfer) are interleaved with rule-based components.
- The system analyses source text up to a high (tectogrammatical) level of abstraction: a dependency graph containing only autosemantic words.



- MWE identification is performed by string matching; successfully identified MWEs are collapsed into a single tectogrammatical node.

Results

- QTLeap test corpus contains 1K sentences, ca. 21K words of text from the IT domain.
- BLEU scores for translation models trained on Europarl and in-domain (1.2M sentences, 24M words) text:

	Europarl	In-domain
Baseline	20.24	26.00
$\theta = 0.1$	20.25	26.46 ***
$\theta = 0.2$	20.19	26.43 **
$\theta = 0.3$	—	26.08
$\theta = 0.4$	—	25.48
$\theta = 0.5$	19.39	24.55

Statistical significance with respect to the baseline: ** $p < 0.01$, *** $p < 0.001$.

Experiment	Training		Test	
	Types	Tokens	Types	Tokens
Europarl				
$\theta = 0.1$	1,093	32,956	1	1
$\theta = 0.2$	5,020	174,015	7	8
$\theta = 0.5$	90,133	2,808,015	220	331
In-domain				
$\theta = 0.1$	837	4,593	0	0
$\theta = 0.2$	3,576	19,586	11	14
$\theta = 0.3$	12,333	67,709	52	95
$\theta = 0.4$	32,126	160,828	138	234
$\theta = 0.5$	61,657	303,724	293	480

Discussion

- Source-only analysis of automatically acquired MWEs improves translation quality for this language pair (+0.46 BLEU points).
- The improvement is only seen for the models built with the in-domain text.
  - An indication that our approach is sensitive to the domain of the training data.
- Evaluation paradigm sensitive to the compositionality of the MWEs.
  - The greatest improvements over the baseline are seen with small values of θ.
  - Including more compositional MWEs ($\theta > 0.3$) eventually reduces BLEU scores below the baseline.
  - Composite *t*-nodes representing compositional MWEs likely cannot be adequately translated by single lexemes.
- Methodology introduced here is:
  - Automatic and wide-coverage, allowing construction of linguistic resources with a minimum of human effort; requires no external lexical resources or language-specific tools.
  - Language-independent.
  - Domain-independent.

References

Bahar Salehi, Paul Cook, and Timothy Baldwin. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 977–983, 2015.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170. Association for Computational Linguistics, 2008.

Working Group 3, PARSEME 6th general meeting, 7–8 April 2016, Struga, Macedonia