Translation of Multi-word Expressions into Under-resourced Language: Case of English-Latvian Statistical Machine Translation

Inguna Skadiņa (WG3)
Institute of Mathematics and Computer Science, University of Latvia
e-mail: inguna.skadina@lumii.lv

Abstract

This poster presents two sets of experiments aiming to find best way for MWE treatment in English – Latvian statistical machine translation system. Two typical approaches – pattern-based and statistical – are applied for MWE identification. For pattern-based approach three ways how to integrate MWEs were investigated – (1) bilingual pairs of automatically extracted MWE candidates were added to the parallel corpus and the SMT system was retrained, (2) automatically extracted MWE candidates were added as a second translation table and, (3) MWEs were marked in the translation table. Although automatic evaluation results did not show significant improvement, manual inspection of translations has demonstrated some improvement in fluency and adequacy of MWE translations.

Experiments with linguistically motivated MWEs

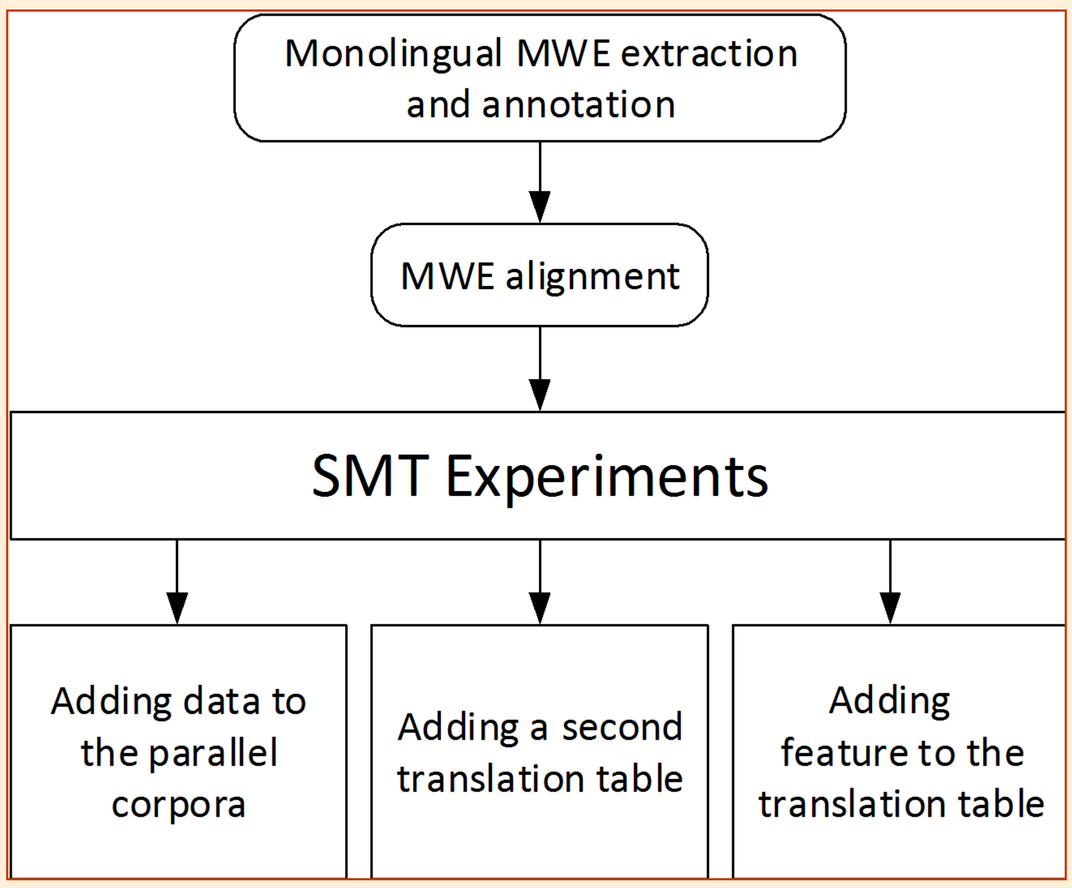
The JRC-Acquis corpus

Training data

- 1.47 million unique parallel sentence pairs
- 41,689 extracted multiword expressions

Tunning data: 1134 random sentences

Test data: 1599 random sentences



Method	BLEU
Baseline	62.40
Baseline + MWE training data	62.44
Two translation tables	62.55
Additional feature	62.27

English: the hazard represented by biological agents
Reference: draudiem , ko rada bioloģiskie aģenti
Baseline: draudiem , ar bioloģiskiem aģentiem
Improved: draudiem , ko rada bioloģiskie aģenti

Experiments with MWE candidates extracted using association measures

DGT-TM corpus

- **Training data:** 1, 63 million parallel sentences
- **Tunning data:** 2000 random sentences
- **Test data:** 1000 random sentences

System	English	Latvian	BLEU
Baseline			46.35
Minimal freq. >2	1,087,932	795,063	**44.86**
Freq. and cost >9	1,074,112	556,695	44.57
Freq. for Latvian >4, freq. for English >9	98,843	88,943	45.13

English: the commission concludes that the scheme in review is incompatible with the common market.
Human: tādēļ komisija secina, ka **apskatāmā** nodokļu shēma nav saderīga ar kopējo tirgu .
Baseline: komisija secina, ka šī **atbalsta** shēma ir uzskatāma par nesaderīgu ar kopējo tirgu.
Improved SMT: komisija secina, ka **apskatāmā** shēma nav saderīga ar kopējo tirgu .

Acknowledgements

The research was supported by Grant 271/2012 from the Latvian Council of Science.