

Evaluation of the system for extraction of multi-word expressions and prediction of their translations

Katerina Zdravkova

University Sts Cyril and Methodius, Skopje

katerina.zdravkova@finki.ukim.mk

Aleksandar Petrovski

International Slavic University, Sveti Nikole

a.petrovski.sise@gmail.com

Four complementary phases:

- extraction of all candidate groups of words that appear in each language at least twice,
- syntactical filtering of obtained candidates, using a predefined set of eligible syntactic expressions,
- prediction of potential translation equivalents from corresponding pairs of aligned sentences
- cross-evaluation of candidate translations, interchanging the source and the target language.

Extraction

- 3500 candidate MWEs, including many useless:

тоа би ја / toa bi ja, instead of тоа би ја усреќило / toa bi ja usrekjilo = that will make her happy

печа тој со / reche toj so, instead of пече тој со недоверба / reche toj so nedoverba = he said with mistrust;

Syntactical filtering

- Less than 500 phrases, sometimes inflections of the same phrase:

атомската бомба / atomskata bomb = the atomic bomb, атомски бомби / atomski bombi = atomic bombs

обичен човек / obichen chovek = an ordinary man, обичните луѓе / obichnite lugje = the ordinary men or the ordinary people,

шаховска табла / shahovska tabla = a chess board, шаховската табла / shahovskata tabla = the chess board

Translation with cross evaluation

- Candidate MWEs obtained without syntactical filtering

968 English candidate MWEs and their translations

- Parallel aligned MWEs

initially: English to Macedonian

Extension: English to Macedonian and Slovene

Trilingual aligned MWEs

[illegible]

Aligned candidate
multiword expressions

[illegible]

Consistency and completeness of MWEs

$$HHI = \sum_{i=1}^n S_i^2 \quad RC = \frac{HHI}{[MWE]}$$

$$DG = \frac{\text{length}(\text{generated MWE})}{\text{length}(\text{complete MWE})}$$

$$TC = \frac{1}{m} \sum_{j=1}^m S_j^2 DG_j^2$$

Conclusions and further work

- Incorporation of MWE lexical entries
- Extension with semi-fixed and flexible MWEs
- Study the lexical cohesion, and extend the document-level translation to a larger collection
- Integration of the model into a hierarchical phrase-based SMT system
- Extraction of the common knowledge about MWEs out of a continuous context and its incorporation into a translation system capable to competently deal with them

Crucial problems

- Manual translation was either inconsistent or had

"an artistic freedom"

- Inflectional paradigms.

which are richer in the

Slavic languages can

influence the translation

- The context in which the

same target MWE can

influence its translation

Language	English	Macedonian	Slovene
Multithread expression	smell of her hair	(mirisot) na njzmatata koska	vonj njenih las
Mac 1: ... (prijetatir) miris na njzmatata koska			
Mac 2: ... mirisot na njzmatata koska			
Multithread expression	ideologically neutral	идеолошки неутрален	ideološko nevtralen/a
Slov 1: ... (novebna beseda ... ni bila) ideološko nevtralna			
Slov 2: ... (predmet govora ni bil) ideološko nevtralen			
Multithread expression	against us	proti nas	po robu (proti nam)
Slov 1: ... (nikdar ne) postavil po robu			
Slov 2: ... (in se nam) postavil po robu			

Language	English	Macedonian	Slovene
Multitext expression	for more than half an hour	(ja) potoe od polovna čas	za ve kot pol ure
Eng 1: ... and never for more than half an hour at a time			
Maced 1: ... в наредна potoe od polovna čas			
Eng 2: ... to turn off the television for more than half an hour			
Maced 2: ... за do domet iskljuen televizora potoe od polovna čas			
Multitext expression	definitive edition	definitivnoto izdanie	no translation (definitno izdaja)
Eng 1: ... (the) eleventh edition is the definitive edition...			
Slov 1: ... (enajsta izdaja) je dokonana			
Eng 2: ... (we were producing a) definitive edition...			
Slov 2: ... (pripravljali sme) končno izdajo			
Multitext expression	(in) the remote	vo oddalno za prenosuvalje	no translation (prepisovna ekipa)
Maced 1: ... do zavrane raboto vo oddalno za prenosuvalje			
Eng 1: ... (down to final touching-up) by the remote squad			
Slov 1: ... (pa do konne dotrpanje) prepisovna ekipa			
Maced 2: ... nikadisi ne raboto vo oddalno za prenosuvalje			
Slov 2: ... nikadri niso bili v (prepisovni) ekipi			
Eng 2: A has never (in) the remote squad			

Language	English	Macedonian	Slovene
Multivord expression	the seconds were ticking by	секунde se tikale (prijeto)	sekunde so tikale z mano
Mac 1:	секунде минуваа отприлика ...		
Mac 2:	секунде минуваа брзо, додека ...		
Multivord expression	almost on a level with	ravno na isto (sto)	no translation
Mac 1:	... ravno na isto sto od ...		
Mac 2:	... ravno na isto ravnanje do ...		
Slov 1:	... skoraj na ravni z ...		
Slov 2:	... skoraj na ravni z ...		
Multivord expression	the last thing	zadnja rabota што mora da se odradi	zadnja stvar što je morala ...
Eng 1:	the last thing for you to understand ...		
Eng 2:	the last thing you must realize ...		