Extraction of potential multi-word expressions in a parallel corpus

Katerina Zdravkova, University Sts Cyril and Methodius, Skopje
Aleksandar Petrovski, International Slavic University, Sveti Nikole

In order to deal with multi-word expressions in a multilingual environment, we proposed a system for extraction of potential multi-word expressions and prediction of their translations from a multilingual corpus. The system consists of four phases: extraction of potential MWEs existent in the sentence aligned parallel corpora, their syntactical filtering in the source language, prediction of potential translation equivalents and finally, evaluation of the obtained results with the reverse system.

The system is currently under construction. It is based on Multext-East parallel corpora of 16 mutually sentence aligned languages [1]. In this paper, we present the present status of the extraction part, which was performed over Orwell's novel 1984.

The preprocessing phase was rather long, due to the lack of an electronic version of the novel. The novel was first converted into Microsoft Word using ABBYY FineReader, automatically checked for spelling errors and then manually corrected during machine learning of the rules for morphological analysis and synthesis of nouns, adjectives and verbs [2]. The conversion of the text into XML was done using the program UpCast, it was afterwards tokenized using the Perl program *mltokenizer*, and at the end, it was sentence aligned with the English version using the *Vanilla aligner* [3]. The alignment was manually verified and all the inconsistencies were polished [4].

The extraction of potential multi-word expressions is done using the programming language Python (http://www.python.org/) in parallel on Windows and on Mac OS X. The XML version was converted into a new version in which uppercase characters were converted to lowercase, and all the punctuation marks were removed.

The extraction process was divided into two sub-phases: separation of all the repeated blocks of words and removing of the blocks generated from longer blocks.

The first process ended up with 15463 blocks of words, which appeared in total 53246 times. The longest repeated block of words that appeared at least twice in the novel consisted of 39 words "*дури и по големите потреси и навидум неотповикливите промени секогаш се обновувал истиот модел исто како што и жироскопот секогаш се враќа во состојба на рамнотежа без состојба колку силно ќе биде турнат на една или на друга страна*". More than 10 words appeared in the 24-word phrase "*тие понатаму се делеле на разни начини носеле безброј различни имиња а нивната бројност исто како и нивниот меѓусебен однос се менувале од ера*", then the 20-words phrase "*знам дека завршуваше со еве една свеќа да ти го осветли патот еве еден џелат да ти го скине вратот*", 17-word phrase "*можеби уште од крајот на каменото доба во светот постоеле три категории луѓе високи средни и ниски*", and at the end, the 12-word long phrase written in the Newspeak "*тајмс 03.12.1983 дневна заповед гб дуплоплуснедобро одн нелица одново*".

Each longer block of words contains its own blocks. For example, the longest block of 39 words also generates two blocks with 38 words, three blocks with 37 words etc., or in total, 741 blocks with at least two words (presented in the table below).

length of the block	frequency	generated smaller blocks
39 words	2	741
24 words	2	276
20 words	2	190
17 words	2	136
12 words	2	66
	10	1409

Whenever a smaller block is generated by a longer block which already exists in the list of potential MWEs has the same frequency as the longer one, it is removed from the list of potential MWEs. If the frequency of the smaller block exceeds the frequency of its parental block (for example, "*не можеше да*" generates "*можеше да*" and "*не можеше*" / "*можеше да*" exists uniquely 117 times, while "*не можеше*" exists uniquely 10 times), it is added to the list of potential MWEs with a unique appearance, called PotentialUniqueMWE list. The filtering process reduced the number of 15463 blocks of words to only 10170 potential unique MWEs. Unfortunately, most of them have no value for further processing of multi-word expressions. The list of the most frequent blocks proves the claim.

frequency	block	frequency	block	frequency	block
671	да се	137	не беше	76	не можеше да
278	да го	130	и да	76	тоа што
196	дури и	128	тоа беше	63	му беше
193	можеше да	125	и со	63	не го
188	како да	120	не е	49	како и
164	да биде	118	никогаш не	37	сето тоа
158	да ја	117	да ги	37	тој се
155	не се	109	и на	37	му го
154	што се	86	и се	37	се случи
143	му се	86	не можеше	37	можеа да
141	за да	86	да не	37	како да се
141	може да				

The first filtering was done using the linguistic development environment NooJ. By restricting the blocks to these PoS sequences: Adj N (*безнадежна љубов*), Adj Adj Noun (*друго човечко суштество*), Adj N N (*мал број луѓе*), Adj N Adj N (not found in PotentialUniqueMWE), N N (*безумие безумие*), N N N (not found in PotentialUniqueMWE), Adj N Prep N (*будното око на полицијата*), Adj N Adv (*неколку минути подоцна*), Adj N Prep Adj N (*дневната заповед на големиот брат*), Adv N Adv N (*повеќе храна повеќе облека*) 491 nominal phrases have been extracted. By manual inspection, most of them will be beneficial in the multilingual system.

On a contrary, the sequences of 563 smaller verbal blocks extracted the compound tenses created with the auxiliary verbs *биде / сум* (to be): *беше можно*, *беше неопходно*, *е дозволено*, *е невозможно* and *има* (to have): *имал право*. Therefore, verbal phrases will be explored more during the next stage of the system development.

References:

1. Erjavec, T. "MULTEXT-East: morphosyntactic resources for Central and Eastern European languages", Language Resources and Evaluation, Volume 46, Issue 1 (2012): 131-142.
2. Ivanovska, A., Zdravkova, K., Erjavec, T., Džeroski, S. "Learning rules for morphological analysis and synthesis of Macedonian nouns,adjectives and verbs" in Proceedings of 5th Slovenian and 1st international Language Technologies Conference, Jozef Stefan Institute, Ljubljana: (2006) 140-145
3. Vojnovski, V., S. Džeroski, and T. Erjavec, 2005. "Learning PoS tagging from a tagged Macedonian text corpus". In Proceedings of SiKDD 2005 (Conference on Data Mining and Data Warehouses), Ljubljana, Slovenia: (2005) 199-202.