The role of question formation in MWE detection

STSM report
Veronika Vincze

April 15, 2015

1 Purpose of the STSM

This Short Term Scientific Mission aimed at exploiting – on the one hand – the experience gained at Stanford University by studying the syntactic and semantic aspects of question formation and – on the other hand – the empirical results achieved at the University of Szeged in the field of MWE detection. Hence, the goals of the visit were twofold: (1) to investigate the relationship of question formation and the semantic and syntactic properties of multiword expressions and (2) to carry out machine learning experiments on MWE detection with the help of question formation.

2 Description of the work achieved during the STSM

The research done during the STSM had three consecutive phases. First, we collected data from available corpora. Then, we carried out qualitative and quantitative analysis of the data to discover significant tendencies in the distribution of (multiword) question words. Lastly, we developed a machine learning algorithm to identify questions that contain verb-particle constructions.

Besides the above activities, I also had the chance to have interesting discussions with members of the Stanford Universal Dependencies team, where one of the main topics was the annotation practices of multiword expressions in the Universal Dependencies framework. In addition, I collected Hungarian idiomatic phrases that contain possessive constructions, to contribute to the ongoing research for a study in the PARSEME WG1 book project.

2.1 Data collection

The first phase of the work was data collection. We consulted the available English and Hungarian corpora – i.e. QuestionBank and the English dataset of the Universal Dependency Treebanks for English and The Szeged

Dependency Treebank for Hungarian – and automatically collected questions from them. In this way, we got a dataset with 4972 English questions and 5668 Hungarian questions. Later on, we filtered the English examples for verb-particle constructions (VPCs) and verb-prepositional combinations, thus yielding a dataset that can be used in our machine learning experiments.

2.2 Data analysis

We carried out some statistical analysis of the data and identified the most frequent question words in both languages. We also examined what question words can be regarded as multiword expressions, based on syntactic tests of grammaticality. We also investigated statistical differences in the distribution of question words, their morphological and syntactic roles, with regard to questions that contained VPCs or not.

2.3 Machine learning experiments

Following the above, we constructed a machine learning system for detecting verb-particle constructions (VPCs) and verb-prepositional combinations in our dataset. For this, we made use of results reported earlier in the literature and also our findings of statistically significant differences in data distribution. Our system exploits a lot of morphological, syntactic, semantic and lexical features.

3 Description of main results

During the STSM, all the above activities yielded interesting results, which are described below.

3.1 Analysis of (multiword) question words

The analysis of corpus-based data led to a number of interesting results. First, there are statistically significant differences in the distribution of question words and particles and their morphological and syntactic roles between questions that contain VPCs and those that contain verb-prepositional phrase combinations. Second, we provided a detailed linguistic analysis of multiword expression question word candidates in English and also of a specific Hungarian MWE type, namely, *mi a ** constructions.

3.2 Results of machine learning experiments

The above mentioned statistically significant differences were included in our machine learning system as features. Our experiments demonstrated that VPCs and non-VPCs can be effectively separated from each other in questions by using a rich feature set, achieving an accuracy score of 92.5%.

Our results achieved on a benchmark dataset were also very similar to those reported in the literature, thus the value of relying on additional features based on WH-words in VPC detection was also shown.

3.3 Universal Dependencies and MWEs

As an empirical result of our discussions with the UD team at Stanford, we started to write a draft version of a detailed guideline for annotating MWEs in English corpora, which we hope to elaborate on in the future.

3.4 Hungarian possessive idioms

Based on freely available online dictionaries and collections, I compiled a dictionary of Hungarian idioms that contain possessive constructions. The dictionary contains approximately 90 items, all of which are assigned to their English equivalents, hence it can serve as a base for further interlingual comparisons.

4 Future collaboration with host institution

We intend to continue our work carried out during the STSM. Future collaboration will likely include continued work on questions, on possessive idioms, and on Universal Dependencies. The host also expressed his interest in visiting future PARSEME meetings and he also plans to visit the University of Szeged in the near future.

5 Foreseen publications resulting from the STSM

We plan to submit papers on the topics of the STSM, for instance, we are working on a joint paper together with Francis Bond and Christiane Fellbaum on cross-linguistic perspectives for possessive idioms.

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

Dan Flickinger, Stanford University: Veronika Vincze's visit has been both rewarding and productive, with interesting and important outcomes resulting from many detailed discussions interwoven with corpus-based investigation and cross-linguistic analysis. We have rarely had a visiting researcher in our laboratory who worked as intensively and as productively as Dr. Vincze has. We have established multiple areas of common research interest and methodology during this visit, and I am confident that we will continue our collaboration in the coming months.