Automatic Extraction of Verbal Locutions for Estonian: Validating Results with Pre-existing Phrasal Lexicons

Heiki-Jaan Kaalep, Kadri Muischnek

Dept. of Estonian and Finno-Ugric Linguistics, University of Tartu, Tiigi 78 – 206, Tartu, Estonia, 50410 {hkaalep,kmuis}@psych.ut.ee

Gaël Harry Dias

Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Monte da Caparica, Portugal, 2725-114 ddg@di.fct.unl.pt

Abstract

Knowing the common expressions of a language is of uttermost importance for an adequate knowledge, description and computational treatment of the language. In order to create a repository of such expressions, one has to rely heavily on text corpora. When using a statistical tool for extracting multiword units from text, one has to harmonise the language-independent algorithm with the specific features of the multiword units and language under consideration. The paper presents a case study to demonstrate a successful way of combining linguistic and statistical processing: extracting Estonian multiword verbs from a text corpus. We evaluate the results by comparing them to a database of multiword verbs, built manually from existing dictionaries beforehand.

Keywords: collocations, multiword units, statistical language processing, evaluation.

1. Introduction

In order to analyse and synthesise sentences of a language, it is not sufficient if one knows the words and syntax rules of that language. In addition, one has to be aware of the common expressions, which may be idioms as well as simple frequent phrases.

A database of Estonian expressions has been compiled from human-oriented dictionaries. It is available at http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html. However, the usage of the expressions in real-life texts has not been explored.

Fortunately, language-independent computational tools have been developed in order to identify and extract multiword units from electronic text corpora (Dias et al. 2000). So, a simple procedure can be used to test the validity of pre-existing resources against text corpora: run a statistical program, find expressions among multiword unit candidates, compare the results with the existing database, and add new information.

However, the program may find expressions that make little sense for a linguist, and fail to find those that one would identify from the text by hand. In order to get most out of a statistical tool, we believe we must take into account the linguistic properties of the text and the expressions, as well as the requirements of the statistical tool.

2. Statistical tool

For the specific case of extracting Estonian multiword verbs, we tailored a statistical tool SENTA (Software for Extracting N-ary Textual Associations) developed by (Dias et al. 2000). Below we briefly describe its underlying principles.

2.1. The Mutual Expectation measure

By definition, multiword units are groups of words that occur together more

often than expected by chance. From this assumption, we define a mathematical model to describe the degree of cohesiveness between the words of an *n*-gram.

First, we define the normalised expectation (NE) existing between *n* words as the average expectation of the occurrence of one word in a given position knowing the occurrence of the other *n*-1 words also constrained by their positions. For example, the average expectation of the *3*-gram "*vahi alla võtma*" (*take into custody*) must take into account the expectation of occurring "*võtma*" after "*vahi alla*", but also the expectation of "*alla*" linking together "*vahi*" and "*võtma*" and finally the expectation of occurring "*vahi*" before "*alla võtma*". The basic idea of the normalised expectation is to evaluate the cost of the possible loss of one word in an *n*-gram. The less an *n*-gram accepts the loss of one of its components, the higher its normalised expectation will be. We define the normalised expectation as the probability of an *n*-gram, divided by the arithmetic mean of the probabilities of *n-1*-grams it contains:

$$NE = \frac{prob(n-gram)}{\frac{1}{n}\sum prob(n-1-grams)}$$

So, the more *n-1*-grams occur somewhere else besides inside the *n*-gram, the bigger the arithmetic mean will be, and consequently, the smaller the NE will be.

Based on NE, we can now define the Mutual Expectation Measure. One effective criterion for multiword unit identification is simple frequency (Daille 1995). From this assumption, we pose that between two *n*-grams with the same normalised expectation, the most frequent *n*-gram is more likely to be a multiword unit:

$$ME = prob(n-gram) \times NE(n-gram)$$

2.2. The GenLocalMaxs Algorithm

Once we have calculated the ME for an *n*-gram, as well as for its *n-1*-grams and

shorter -grams contained in it, we use the GenLocalMaxs algorithm to decide which one among them to choose. It assumes that an *n*-gram is a multiword unit if the degree of cohesiveness between its *n* words is higher or equal than the degree of cohesiveness of any sub-group of (*n-1*) words contained in the *n*-gram and if it is strictly higher than the degree of cohesiveness of any super-group of (*n+1*) words containing all the words of the *n*-gram. In other words, an *n*-gram, let's say *W*, is a multiword unit if its ME value, *ME(W)*, is a local maximum. Let's define the set of all the (*n-1*)-grams contained in the *n*-gram *W*, by Ω_{n-1} and the set of all the (*n+1*)-grams containing the *n*-gram *W*, by Ω_{n+1}.

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$

if n=2 then

if $ME(W) > ME(y)$ then *W* is a multiword unit

else

if $ME(x) \leq ME(W)$ and $ME(W) > ME(y)$ then *W* is a multiword unit

Figure 1: The GenLocalMaxs algorithm

3. Text preparation

Estonian is a flective language with a free word order. Due to its nature, it is likely that a language independent statistical tool will perform poorly on Estonian. Indeed, statistical systems are designed to identify recurrent and probable associations between wordforms and do not take advantage of the specificities of the language. Thus, eliminating inflectional endings, by using a lemmatiser, would give the statistical software better grounds for finding recurring patterns.

However, at the same time, it is known that expressions tend to contain frozen forms, including inflectional endings, and eliminating them might lose information that

is necessary for recognizing the expression. For example, one may not say "*Human Right*" or "*Humans Right*": "*Human Rights*" is the only correct expression.

Phrasal verbs like "*ära maksma*" (*to pay off)* and idiomatic verbal expressions like "*end tükkideks naerma*" (*to laugh oneself to pieces)* represent a situation that is different from both of the above-mentioned extremes: the verb part may inflect freely, but the other word(s) are frozen forms. Consequently, we tried a pragmatic approach to text preparation: lemmatise only some words (the ones that inflect freely in the expressions), and do not lemmatise others.

4. Experiment

We made our experiment on a 500,000-word sub-corpus of the Corpus of Written Estonian of the 20th Century (http://www.cl.ut.ee/cgi-bin/konk_sj_en.cgi).

In order to extract multiword verbs, we performed the following tasks.

1. Perform a morphological analysis and disambiguation of the corpus.
2. For verbs, keep the lemma form; for other words, keep the original wordform.
3. Select all the possible collocations.
4. Eliminate collocations, not relevant for this particular task: collocations not including a verb, as well as collocations containing pronouns (with a few exceptions), punctuation, certain adverbs etc.
5. Calculate Mutual Expectation and GenLocalMaxs; based on these, make the final choice of extracted phrases.

We processed the corpus four times with SENTA, each time setting a different limit (0 to 3) to the number of words that may intervene the words of a phrase. Then we combined the results and compared them against our database that contained 10816

entries and was based on (EKSS 1988 – 2000), (Saareste 1979), (Hasselblatt 1990), (Õim 1991), (Õim 1993) and (Filosoft). We checked manually all the extracted phrases that were not in the database, and decided whether they should be added.

SENTA extracted 13,100 multiword verb candidates. 2,500 of these, 19%, are such that they should be found in a database of Estonian multiword verbs. The rest are collocations a linguist would rather not present in the database. In fact, 1629 of the 2,500 were expressions that the database already contained; and SENTA found an extra 865 phrases that should be included. We can see that out of the 2,500 multiword verbs that were extracted, only 2/3 were present in the database.

5. Evaluating SENTA

How sure can we be that SENTA really found all the multiword verbs that are in the corpus, and that it did not report about verbs that are really not there? For estimating this, we made an experiment with 500 randomly selected multiword verbs that we had in our database before. By checking the corpus manually, we found that 131 out of the 500 could be found in the corpus. In principle, SENTA can find only phrases that occur at least twice in the corpus. The number of such phrases was 71 (out of the 500).

We made 4 experiments with SENTA, where we defined differently the number of words that could possibly occur between the words of a phrase: 0, 1, 2 or 3.

The longer the allowed distance between individual words of a phrase, the more possible phrases SENTA found. It is noteworthy, however, that as the distance grew longer, SENTA stopped finding some phrases that it did with a shorter distance. This is why the combination of the results is better than any individual distance.

In addition to the correct phrases, SENTA extracted some phrases erroneously:

it is possible that all the words of a phrase co-occur in the same sentence, without, however, forming this phrase in that particular sentence. Let us consider "*tagasi tegema*" (*to pay back*) in the context "*tagasi jõudes teeme sotid selgeks*" (*we will pay when we get back*). SENTA extracted the phrase "*tagasi tegema*" (*to pay back*), and this was an error. Just as one might expect, the number of mistakenly extracted phrases grew with the distance we allowed between the words, reaching 15 (7 that never occurred in the corpus plus 8 that occurred once), if the distance was 3.

The 131 phrases we found from the corpus form a random selection from all the phrases that are in the given corpus. In the best possible case, if we combine the results of the experiments with different distances, SENTA will find 57/71=80% of those that occur more than once in the corpus (and close to 99% of those that occur more than 3 times), and 8/60=12% of those that occur once. This evidences a very high recall rate thus balancing the lower precision results.

If a linguist processes a new corpus with SENTA, and picks out only the linguistically good-looking phrases (in our combined experiment 57+7+8=72), (s)he may expect that the final result actually consists of the following: 57/72=79% are phrases that occur in the corpus twice or more, 8/72=11% are phrases that occur in the corpus once, and 7/72=10% are phrases that do not occur in the corpus at all.

6. Conclusion

We have seen that one can accomplish the difficult task of extracting multiword verbs from a corpus, by combining automatic linguistic and statistical processing with manual post-editing. Although the precision rate of the method was low, it was compensated by a high recall rate. This in turn means that SENTA is a useful tool for

lexicography: browsing 13,100 candidates for verb phrases is very different from browsing the 500,000-word corpus for the same verb phrases.

An unexpected result concerns the evaluation of the database: our database is far from perfect, but so are the dictionaries it is based upon! The results obtained by SENTA are immediately usable for syntactic and semantic processing of Estonian.

References

Daille B. (1995) "*Study and Implementation of Combined Techniques for Automatic Extraction of Terminology."* The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Ed. by J. Klavans and P. Resnik. 49-66. Cambridge, MA; London, England: MIT Press.

Dias, G., Guilloré, S., Bassano, J.C., Lopes, J.G.P. (2000) "*Extraction Automatique d'unités Lexicales Complexes: Un Enjeu Fondamental pour la Recherche Documentaire*." Journal Traitement Automatique des Langues, Vol 41:2, Christian Jacquemin (ed.). 447-473. Paris.

EKSS (1988 – 2000) *Eesti kirjakeele seletussõnaraamat.* Tallinn: ETA KKI.

Filosoft – *Tesaurus*. http://ee.www.ee/Tesa/

Hasselblatt, C. (1990) *Das Estnische Partikelverb als Lehnübersetzung aus dem Deutschen*. Wiesbaden.

Õim, A. (1993) *Fraseoloogiasõnaraamat*. Tallinn: ETA KKI.

Õim, A. (1991) *Sünonüümisõnastik*. Tallinn.

Õim, A. (1998) *Väljendiraamat.* Tallinn: Eesti Keele Sihtasutus.

Saareste, A (1979) *Eesti keele mõistelise sõnaraamatu indeks*. Uppsala: Finsk-ugriska institutionen.