Corpus based analysis for multilingual terminology entry compounding

Andrejs Vasiljevs, Kaspars Balodis
Tilde, University of Latvia
Vienibas gatve 75a, Riga, LV-1004, Latvia
E-mail: andrejs@tilde.lv, kbalodis@gmail.com

Abstract

This paper proposes statistical analysis methods for improvement of terminology entry compounding. Terminology entry compounding is a mechanism that identifies matching entries across multiple multilingual terminology collections. Bilingual or trilingual term entries are unified in compounded multilingual entry. We suggest that corpus analysis can improve entry compounding results by analysing contextual terms of given term in the corpus data.

1. Introduction

This paper addresses some of the problems in terminology consolidation that are discovered in EuroTermBank project. This section briefly describes terminology consolidation challenge and EuroTermBank project. The next section introduces concept of *terminology entry compounding*. The following sections discus possibilities for application of statistical methods for improving entry compounding results as well as first experiments in this field.

Globalization from the one side and growing language awareness from the other side dictates the need to consolidate different national terminology resources, to harmonize international terminology, to provide online access to reliable multilingual terminology. There are number of terminology resources and databases provided by different institutions or national terminology bodies. However these databases are mostly limited in language coverage or are subject specific.

EuroTermBank project (Auksoriute et al., 2006) has a goal to collect, harmonize and disseminate dispersed terminology resources through online terminology data bank. The EuroTermBank was developed by 8 partners from 7 European Union countries – Germany, Denmark, Latvia, Lithuania, Estonia, Poland and Hungary. Web-based terminology data bank www.eurotermbank.com has been developed to provide easy access to centralized terminology resources as well as methodology for harmonization of terminology processes.

Large number of terminology resources was acquired and processed gathering in total 1.5 million terms from about 600 000 term entries, involving almost 30 languages.

To consolidate representation of multilingual terms an automated terminology entry compounding mechanism has been proposed (Vasiljevs, Rirdance, 2007) that identifies matching entries across multiple terminology collections.

2. Terminology entry compounding

Entry compounding solves the problem of unified representation of multiple potentially overlapping term entries that are present in a consolidation of a huge number of multilingual terminology sources. Majority of terminology resources that are available in Eastern European countries are bilingual with a source language mostly being English. Much smaller number of resources is monolingual or has terms in three or more languages.

Since multiple terms in multiple languages can refer to the same concept, the concept is the shared element that must be used to link the terms together in a multidimensional database (Wright, 2005).

EuroTermBank data structure is modeled according to concept-oriented approach to terminology. Terminology entry denotes an abstract concept that has designations or terms as well as definitions in one or more languages. If terminology bank contains entries coming from different collections and designating the same concept we have an obvious interest to merge them into one unified multilingual entry.

For example, if we have term pair *EN computer – LV dators* coming from Latvian IT terminology resource and another term pair *EN computer – LT kompiuteris* from Lithuanian IT terminology resource we may want to join these two into unified entry *EN computer – LV dators – LT kompiuteris.* Such multilingual entry allows to get correspondence between language terms that are not directly available in any terminology resource (in our example new term pair *LV dators – LT kompiuteris).*

But merging entries just on the bases of matching term in one language that is common for these entries will lead to many erroneous term correspondences. For example, if we have LV-EN entry *stumbrs-stick* and ET-EN entry *kang-stick*, we may want to merge these entries into compound entry LV-ET-EN *stumbrs-kang-stick*. But if we would add to this alignment LV-EN entry *rokturis-stick* it would lead to wrong LV-ET translation *rokturis-kang*.

Such problems are obvious due to the frequent ambiguity of terms among subject fields or rarer cases of ambiguity in the context within one subject field. We can conclude that the only error-free method for merging entries is evaluating whether these entries denote the same concept. Unfortunately in practice it is often impossible or very expensive to make comparisons of cross-lingual terminology concepts. There is a lack of experts with sufficient knowledge of respective languages and subject

fields. The task is considerably hindered by the fact that majority of EuroTermBank terminology collections does not have term definitions included.

In EuroTermBank, a practical solution is proposed by introducing the *terminology entry compounding* approach. Entry compounding is an automated approach for matching terminology entries based on available data. The most reliable indication for matching entries is having unique and unambiguous concept identifiers. The best example is terms from ISO terminology standards. These term entries have an identifier in the form *[Standard_identifier].[term_number]*. Accordingly, all national standards share the same identifier for corresponding entries and can be merged with a very high degree of reliability. Another case of unique internationally applied identification is the usage of Latin names in medicine and biology (with a number of exceptions with different Latin names designating the same concept). If there is no unique identification for concepts in collections, less precise matching criteria are used, namely, the English term and the subject field. English was chosen as the most popular language in term resources.

EuroTermBank uses Eurovoc as a subject field classification. A number of terminology resources use only top classification levels of Eurovoc but there are many resources with detailed classification using Eurovoc sublevels of different depth. For this reason it was decided to take into account only the top classification level for entry compounding. This means that sublevels are equalized to the top classification level.

It is important to understand that entry compounding is a data representation method that does not propose to create new terminology entries. It is a visualization aid that displays matching entries across collections in a consolidated way. Matches are determined by applying a number of criteria and as such cannot be error-free.

As majority of terminology resources integrated in EuroTermBank are bilingual (Table1), we would like to transform data representation from number of separate bilingual entries to unified multilingual record.

Entry languages	*Number of entries*	*Percentage from total*
monolingual	11230	2%
bilingual	398854	68%
3-lingual	45497	8%
4-lingual	69134	12%
5-lingual	48761	8%
>5-lingual	12216	2%

Table 1 Multilinguality of EuroTermBank source records

Entry compounding solves the problem of visual representation of multiple potentially overlapping term entries that are present in a consolidation of a huge number of multilingual terminology sources. At present, the EuroTermBank database contains over 585,711 term entries with more than 1,500,500 terms. When applying entry compounding, over 135,000 or 23% of entries get compounded. Hence entry compounding is a considerable aid for the user in finding the required term, for example, in the translation scenario between language pairs for which term equivalence is not established in existing collections.

Unfortunately abovementioned criteria for entry compounding are insufficient and generate too much incorrect alignments. High recall rate lead also to relatively low precision although we currently do not have exact precision evaluation figures.

If our term entries would include term definitions then we could compare these by human review or by applying automated analysis methods. But because large majority of Eastern European terminology resources do not include definitions we need to look for other sources to depict meaning of terms.

We suggest to use multilingual text corpus as a source were to look for term usage patterns and try to disambiguate its meaning. Of course it is impossible to get term definition from the regular text corpus. But we can intuitively assume that term meaning is related to the context where term usually appears in. This intuition has also some rational basis. For cost and time saving many institutions dealing with terminology creation are not preparing definitions for new terms but instead include in term database several typical examples of usage context.

We can assume that term t in language L1 and s in language L2 are matching (or denoting the same concept) if t and s have similar context patterns in L1 corpus and L2 corpus respectively. By the context pattern we mean characteristic collocates frequently appearing in proximity of term. Because terminology is related to special language (special language uses specific words with specific, preferably unambiguous meaning, in contrast to general language with wide lexicon of usually very ambiguous words) we are interested in those collocate words that are terms from the same subject field. This is also based on common intuition that term in specific subject field should be best described by other terms from this subject field.

3. Proposed method

In the proposed method we try to grasp the intuition that if two terms in different corpora have similar context patterns then they might denote the same concept and more frequent collocations have more impact on term context pattern than less frequent ones.

Let's assume that we have applied simple term compounding for bilingual terminology resources as described previously. For language *L1* term *t* we have several translation candidates $s_1, s_2, ..., s_n$ in language *L2*. Our task is to select the most probable from these candidates by analyzing context patterns of these terms.

Let's denote frequency of term *t* in language *L1* corpus with *count(t)*.

Frequency of s_1, s_2, ..., s_n in *L2* corpus will be denoted with $count(s_1)$, $count(s_2)$, ..., $count(s_n)$.

We denote collocations of term *t* with $coll_1(t)$, $coll_2(t)$, ..., $coll_m(t)$ and respective frequency of these collocations in proximity with *t* with $count(t, coll_1(t))$, $count(t, coll_2(t))$, ..., $count(t, coll_m(t))$.

We will select those collocations of the term *t* in language *L1* whose frequency is higher than certain threshold *p*.

This means that we will select $coll_j(t)$, where

$$\frac{count(t, coll_j(t))}{count(t)} > p.$$

For every such collocation we will find translation candidate $x_1, x_2, ..., x_k$ in language *L2*. For every candidate translation s_i of the term *t*:

if $\frac{count(s_i, x_1) + count(s_i, x_2) + ... + count(s_i, x_k)}{count(s_i)} > p$

then we will add to the score of this candidate the lowest from the numbers

$$\frac{count(s_i, x_1) + count(s_i, x_2) + ... + count(s_i, x_k)}{count(s_i)}$$

and $\frac{count(t, coll_j(t))}{count(t)}$.

Now let's do the same calculation from reverse side – for every translation candidate si in language L2 we will select collocations whose frequency is higher than certain threshold p.

This means that we will select $coll_j(s_i)$, where

$$\frac{count(s_i, coll_j(s_i))}{count(s_i)} > p.$$

For every such collocation we will find translation candidates $x_1, x_2, ..., x_k$ in language *L1*.

If these translations appear in context with t frequently enough passing our threshold p:

$$\frac{count(t, x_1) + count(t, x_2) + ... + count(t, x_k)}{count(t)} > p,$$

then we will add to the score of this candidate the lowest from the numbers

$$\frac{count(t, x_1) + count(t, x_2) + ... + count(t, x_k)}{count(t)}$$ and

$$\frac{count(s_i, coll_j(s_i))}{count(s_i)}.$$

We will assume that translation candidate s_i with the highest resulting score is the most probable equivalent of term *t* in language *L2*.

4. Experimental results

To test the proposed method we carried out experiment on compounding of Latvian and Lithuanian terms. For this experiment we used JRC-Acquis Multilingual corpus v3.0 which is the largest publicly available source of corpus data for Latvian and Lithuanian (Steinberger et al., 2006). Latvian corpus contains 22 906 documents with 27 592 514 words. Lithuanian corpus contains 23 379 documents with 26 937 773 words.

It could be asked why not to use well proven statistical alignment methods to align terms from these corpora as these are highly parallel texts mostly being translations from the same source (English). But as we want to find a method for more general case of lack of parallel in-domain data, we split this corpus in 2 parts. For Latvian corpus we used the first part and for Lithuanian – the second. In such a way we got sufficiently large corpus of un-parallel texts for Latvian and Lithuanian.

For experiment we selected 27 Lithuanian terms and 80 corresponding Latvian term candidates. Only terms with at least 50 occurrences in corpus were selected and only Lithuanian terms for which there were at least one correct and one incorrect Latvian term were selected. Correct translation was depicted by human terminologist.

Every Lithuanian term had from 2 to 8 candidate translations in Latvian from which only 1 to 4 were correct.

We implemented the proposed method and made experiments with different settings of threshold parameter p.

The size of window for collocations was 10 words to the left and right of the term occurrences. As Latvian and Lithuanian are highly inflected languages, morphological normalization was applied.

To measure the usefulness of our method we chose the value of threshold parameter p = 0.002.

We say that our method gives correct result on Latvian term *s* (which is translation candidate of Lithuanian term *t*):

- If *s* is a correct translation of *t* and its score is at least 5% higher than for every other incorrect translation candidate of *t*.
- If *s* is not a correct translation of *t* and its score is at least 5% lower than for any other correct translation of *t*.

- If scores of correct translation and an incorrect translation of *t* differ by less than 5% then we say that there is not enough difference in score.
- Otherwise we say that our method gives wrong result.

Examples of results are in Figure 1 and Figure 2. On X axis there are different values of threshold p and on Y axis are the scores for term pair.

Results of experiment showed that our method gave correct answer in 61% of cases (for 49 out of 80 Latvian terms).

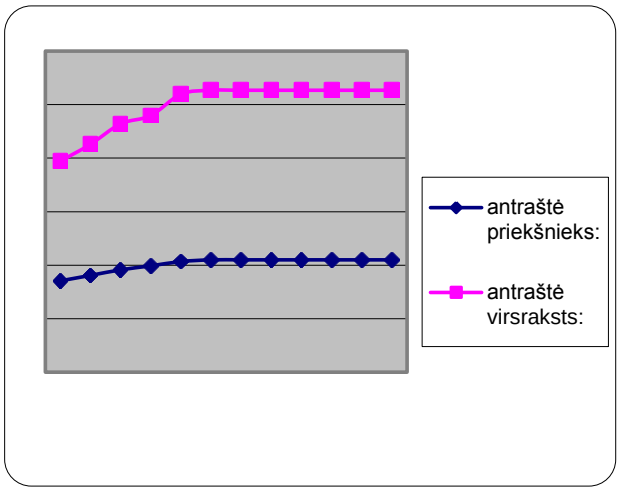

Figure 1 Correct Latvian term *virsraksts* for Lithuanian term *antraštė* achieved significantly higher score than wrong translation *priekšnieks*

For 21% (17 out of 80 Latvian terms) there was not enough difference in score.

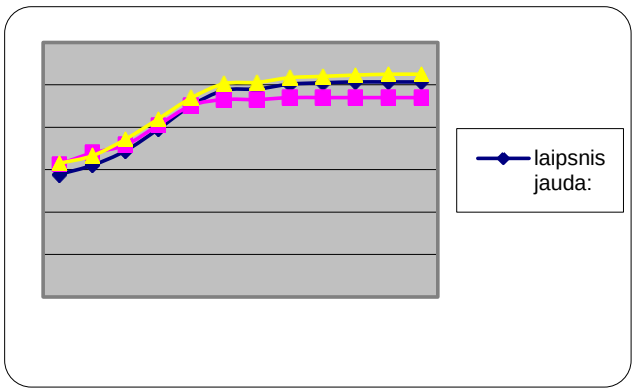

Figure 2 Example of insufficient difference in score for Latvian term *jauda* and Lithuanian term *laipsnis.*

For 18% (14 out of 80 Latvian terms) the method gave wrong result.

5. Conclusions and Future Work

Proposed method of statistical corpus analysis of term context demonstrates promising results to improve automated terminology entry compounding.

These results encourage further research for different language pairs and in different domains.

6. Acknowledgements

This work was supported by European Social Funds.

7. References

Auksoriute A. 2006 Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project., ISBN 9984-9133-4-1, Riga

Henriksen L., Povlsen C., Vasiljevs A. 2005. EuroTermBank – a Terminology Resource based on Best Practice. In *Proceedings of LREC 2006, the 5th International Conference on Language Resources and Evaluation*, Genoa, on CD-ROM , May 2006

Vasiļjevs A., Skadiņš R. 2005. Eurotermbank terminology database and cooperation network. In: *Proceedings of the Second Baltic Conference on Human Language Technologies*, Tallinn, pp. 347-352.

Vasiljevs A., Rirdance S. 2007. Consolidation and unification of dispersed multilingual terminology data, *International Conference RANLP 2007 (Recent Advances in Natural Language Processing*), Borovets, Bulgaria, 2007

Wright, Sue Ellen 2005. A Guide to Terminological Data Categories – Extracting the Essentials from the Maze. In *Proceedings of TKE 2005, the 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen, pp. 63-77.

Steinberger R., Pouliquen B., Widiger A., Camelia Ignat C., Erjavec T., Tufiş D., Varga D. 2006, The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages, In *Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006*, May 2006, Genoa, Italy