Aligning Estonian and Russian news industry keywords with the help of subtitle translations and an environmental thesaurus

Andraž Repar
International Postgraduate School / Jamova 39, 1000 Ljubljana, Slovenia
andraz.repar@ijs.si
Andrej Shumakov
Ekspress Meedia / Narva mnt 13, 10151 Tallinn, Estonia

Abstract

This paper presents the implementation of a bilingual term alignment approach developed by Repar et al. (2019) to a dataset of unaligned Estonian and Russian keywords which were manually assigned by journalists to describe the article topic. We started by separating the dataset into Estonian and Russian tags based on whether they are written in the Latin or Cyrillic script. Then we selected the available language-specific resources necessary for the alignment system to work. Despite the domains of the language-specific resources (subtitles and environment) not matching the domain of the dataset (news articles), we were able to achieve respectable results with manual evaluation indicating that almost 3/4 of the aligned keyword pairs are at least partial matches.

1 Introduction and related work

The ability to accurately align concepts between languages can provide significant benefits in many practical applications. For example, in terminology, terms can be aligned between languages to provide bilingual terminological resources, while in the news industry, keywords can be aligned to provide better news clustering or search in another language. Accurate bilingual resources can also serve as seed data for various other NLP tasks, such as multilingual vector space alignment.

In this paper, we describe the experiments on an Estonian-Russian dataset of news tags — labels that were manually assigned to news articles by journalists and editors at Ekspress Meedia, one of the largest news publishers in the Baltic region. The dataset contains both Estonian and Russian tags, but they are not aligned between the two languages. We adapted the machine learning term alignment approach described by Repar et al. (2019) to align the Russian and Estonian tags in the dataset.

The alignment approach in Repar et al. (2019) is a reproduction and adaptation of the approach described by Aker et al. (2013a). Repar et al. (2019) managed to reach a precision of over 0.9 and therefore approach the values presented by Aker et al. (2013a) by tweaking several parameters and developing new machine learning features. They also developed a novel cognate-based approach which could be effective in texts with a high proportion of novel terminology that cannot be detected by relying on dictionary-based features. In this work, we perform the implementation of the proposed method on a novel, Estonian-Russian language pair, and in a novel application of tagset alignment.

Section 1 lists the related work, Section 2 contains a description of the tag dataset used, Section 3 describes the system architecture, Section 4 explains the resources used in this paper, Section 5 contains the results of the experiments and Section 6 provides conclusions and future work.

2 Dataset description

The dataset of Estonian and Russian tags was provided by Ekspress Meedia as a simple list of one tag per line. The total number of tags was 65,830. The tagset consists of keywords that journalists assigne to articles to describe an article's topic, and was cut down recently by the editors from more than 210,000 tags.

The number of Russian tags was 6,198 and they were mixed with the Estonian tags in random order. Since Russian and Estonian use different writing scripts (Cyrillic vs Latin), we were able to separate the tags using a simple regular expression to detect Cyrillic characters. The vast majority of the tags are either unigrams or bigrams (see Table 1 for details).

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 71–75

Grams	Estonian	Russian
1	0.49	0.49
2	0.44	0.41
3	0.05	0.06
4	0.01	0.02
> 4	0.01	0.02

Table 1: An analysis of the provided dataset in terms of multi-word units. The values represent the ratio of the total number of tags for the respective language. The total number of Estonian tags was 59,632, and the total number of Russian tags was 6,198. The largest Estonian tag was a 14-gram and the largest Russian tag was an 11-gram, but the vast majority of tags are either unigrams or bigrams.

3 System architecture

The algoritm used in this paper is based on the approach described in Repar et al. (2019) which is itself a replication and an adaptation of Aker et al. (2013b). The original approach designed by (Aker et al., 2013b) was developed to align terminology from comparable (or parallel) corpora using machine-learning techniques. They use terms from the Eurovoc (Steinberger et al., 2002) thesaurus and train an SVM binary classifier (Joachims, 2002) (with a linear kernel and the trade-off between training error and margin parameter c = 10). The task of bilingual alignment is treated as a binary classification - each term from the source language *S* is paired with each term from the target language *T* and the classifier then decides whether the aligned pair is correct or incorrect. (Aker et al., 2013b) use two types of features that express correspondences between the words (composing a term) in the target and source language:

- 7 dictionary-based (using Giza++) features which take advantage of dictionaries created from large parallel corpora of which 6 are direction-dependent (source-to-target or target-to-source) and 1 direction-independent - resulting in altogether 13 features, and
- 5 cognate-based (on the basis of (Gaizauskas et al., 2012)) which utilize string-based word similarity between languages.

To match words with morphological differences, they do not perform direct string matching but utilize Levenshtein Distance. Two words were considered equal if the Levenshtein Distance (Levenshtein, 1966) was equal or higher than 0.95. For closed-compounding languages, they check whether the compound source term has an initial prefix that matches the translation of the first target word, provided that translation is at least 5 characters long.

Additional features are also constructed by:

- Using language pair specific transliteration rules to create additional cognate-based features. The purpose of this task was to try to match the cognate terms while taking into account the differences in writing systems between two languages: e.g. Greek and English. Transliteration rules were created for both directions (source-to-target and target-to-source) separately and cognate-based features were constructed for both directions - resulting in additional 10 cognate-based features with transliteration rules.
- Combining the dictionary and cognate-based features in a set of combined features where the term pair alignment is correct if either the dictionary or the cognate-based method returns a positive result. This process resulted in additional 10 combined features[1].

A subset of the features is described below (For a full list of features, see Repar et al. (2019)):

- *isFirstWordTranslated*: A dictionary feature that checks whether the first word of the source term is a translation of the first word in the target term (based on the Giza++ dictionary).
- *longestTranslatedUnitInPercentage*: A dictionary feature representing the ratio of the longest contiguous sequence of source words which has a translation in the target term (compared to the source term length).
- *Longest Common Subsequence Ratio*: A cognate feature measuring the longest common non-consecutive sequence of characters between two strings
- *isFirstWordCovered*: A combined feature indicating whether the first word in the source

[1]For combined features, a word is considered as covered if it can be found in the corresponding set of Giza++ translations or if one of the cognate-based measures (Longest Common Subsequence, Longest Common Substring, Levensthein Distance, Needleman-Wunsch Distance, Dice) is 0.70 or higher (set experimentally by (Aker et al., 2013b))

term has a translation or transliteration in the target term.

- *isFirstWordCognate*: a binary feature which returns True if the longest common consecutive string (LCST) of the first words in the source and target terms divided by the length of the longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters.

Repar et al. (2019) start by reproducing this approach, but were unable to replicate the results. During the subsequent investigation, they discovered that using the same balance ratio in the training and test sets (i.e. 1:200, which was set by Aker et al. (2013b) to mimic real-world scenarios) have a significant impact on the performance of the algorithm. Furthermore, they filter training set term pairs based on term length and feature values (hence the different training set sizes in Table 2) and develop new cognate-based features.

The system requires several language-specific resources:

- A large parallel corpus to calculate word alignment probability with Giza++. The system in Repar et al. (2019) uses the DGT translation memory (Steinberger et al., 2013).
- A list of aligned terms that serve as training data. The system in Repar et al. (2019) uses the Eurovoc thesaurus (Steinberger et al., 2002). 600 Eurovoc term pairs are used as test data, while the rest is used for training.
- Transliteration rules for the construction of reverse cognate-based features (cognate features are constructed twice: first the target word is transliterated into the source language script, then the source word is transliterated in the target language script).

The constructed features are then used to train the SVM classifier which can be used to predict the alignment of terms between two languages.

4 Resources for the Estonian-Russian experiment

While the DGT translation memory and the Eurovoc thesaurus support all official EU languages, there is no Russian support since Russia is not an EU member state. In order to train the classifier, we therefore had to find alternative resources.

For the parallel corpus, we made experiments with the Estonian Open Parallel corpus[2] and the Estonian-Russian OpenSubtitles corpus from the Opus portal[3]. The OpenSubtitles corpus performed better, most likely due to its much larger size (85,449 parallel Estonian-Russian segments in the Estonian Open Parallel corpus vs. 7.1 million segments in the OpenSubtitles corpus).

While finding parallel Estonian-Russian corpora was trivial due the the list of available corpora on the Opus portal, finding an appropriate bilingual terminological database proved to be more difficult. Ideally, we would want to use a media or news-related Estonian-Russian terminological resource, but to the best of our knowledge, there was none available. Note that the terminological resource needs to have at least several thousand entries: the Eurovoc version used by Repar et al. (2019) contained 7,083 English-Slovene term pairs. We finally settled on the environmental thesaurus Gemet[4], which at the time had 3,721 Estonian-Russian term pairs. For the transliteration rules, we used the Python pip package transliterate [5] to generate the reverse dictionary-based features.

5 Results

Repar et al. (2019) ran a total of 10 parameter configurations. We selected three of those to test on the Estonian-Russian dataset. The first one is the configuration with a positive/negative ratio of 1:200 in the training set, which significantly improved recall compared to the reproduction of Aker et al. (2013b), the second one is the same configuration with additional term filtering, which was overall the best performing configuration in Repar et al. (2019), and the third one is the Cognates approach which should give greater weight to cognate words. As shown in Table 2, the overall results are considerably lower than the results in Repar et al. (2019), in particularly in terms of recall. One reason for this could be that the term filtering heuristics developed in Repar et al. (2019) may not work well for Estonian and Russian as they do for other languages. For example, 1.3 million candidate term pairs were constructed for the English-Slovene lan-

[2] https://doi.org/10.15155/9-00-0000-0000-0000-0002AL

[3] opus.nlpl.eu

[4] https://www.eionet.europa.eu/gemet/en/themes/

[5] https://pypi.org/project/transliterate/

No.	Config ET-RU	Training set size	Pos/Neg ratio	Precision	Recall	F-score
1	Training set 1:200	627,120	1:200	0.3237	0.2050	0.2510
2	Training set filtering 3	30,954	1:200	0.9000	0.0900	0.1636
3	Cognates approach	33,768	1:200	0.7313	0.0817	0.1469

Table 2: Results on the Estonian-Russian language pair. No. 1 presents the results of the configuration with a positive/negative ratio of 1:200 in the training set, no. 2 presents the results of the same configuration with additional term filtering, which was overall the best performing configuration in Repar et al. (2019), and No. 3 presents the results of the Cognates approach which should give greater weight to cognate words.

ET	RU	Evaluation
kontsert	концерт	exact match
kosmos	космос	exact match
majandus	экономика	exact match
juhiluba	водительские права	exact match
lõbustuspark	парк развлечений	exact match
unelmate pulm	свадьба	partial match
eesti mees	мужчина	partial match
indiaani horoskoop	гороскоп	partial match
hiina kapsas	капуста	partial match
hulkuvad koerad	собаки	partial match
eesti autospordi liit	эстонский футбольный союз	no match
Kalevi Kull	орел	no match
honda jazz	джаз	no match
tõnis mägi	гора	no match
linkin park	парк	no match

Table 3: Examples of exact, partial and no match tag pairs produced by the system.

guage pair and around one half of those were filtered out during the term filtering phase. On the other hand, only around 33,000 Estonian-Russian candidate pairs out of the total 627,000 survived the term fitering phase in these experiments. Another reason for the lower performance is likely the content of the language resources used to construct the features. Whereas Repar et al. (2019) use resources with similar content (EU legislation), here we have dictionary-based features constructed from a subtitle corpus and term pairs from an environmental thesaurus.

We then used the best performing configuration to try to align the Estonian and Russian tags from the dataset provided by Ekspress Meedia. The size of the dataset (59,632 Estonian tags and 6,198 Russian tags) and the fact that the system must test each possible pairing of source and target tags meant that the system generated around 370 million tag pair candidates which it then tried to classify as positive or negative. This task took more than two weeks to complete, but at the end it resulted in 4,989 positively classified Estonian-Russian tag pairs. A subset of these (500) were manually evaluated by a person with knowledge of both languages provided by Ekspress Meedia according to the following methodology:

- C: if the tag pair is a complete match
- P: if the tag pair is a partial match, i.e. when a multiword tag in one langauge is paired with a single word tag in the other language (e.g. eesti kontsert — концерт, or *Estonian concert — concert*)
- N: if the tag pair is a no match

Of the 500 positively classified tag pairs that were manually evaluated, 49% percent were deemed to be complete matches, a further 25% were evaluated as partial matches, and 26% were considered to be wrongly classified as positive tag pairs. The evaluator observed that "the most difficult thing was to separate people's names from toponyms, such as a famous local singer called "Tõnis Mägi", a district in Talinn called "Tõnismägi"

and a mountain named "Muna Mägi". More examples of exact, partial and no match alignments can be found in Table 3.

6 Conclusions and future work

In this paper, we reused an existing approach to terminology alignment by Repar et al. (2019) to align a set of Estonian and Russian tags provided by the media company Ekspress Meedia. The approach requires several bilingual resources to work and it was difficult to obtain relevant resources for the Estonian-Russian language pair. Given the domain of the tagset, i.e. news and media, the selected resources (subtitle translations and an environmental thesaurus) were less than ideal. Nevertheless, the approach provided respectable results with 74% of the positive tag pairs evaluated to be at least a partial match.

When asssessing the performance of the approach, one has to take into account the fact that the tagset is heavily unbalanced with almost 60,000 Estonian tags compared to a little over 6,000 Russian tags. This means that for many Estonian tags, a true equivalent was simply not available in the tagset.

For future work, we plan to integrate additional features into the algorithm, such as those based on novel neural network embeddings which may uncover additional hidden correlations between expressions in two different languages and may provide an alternative to large parallel corpora which are currently needed for the system for work. In terms of the Estonian and Russian language pair, additional improvements could be provided by taking into account the compound-like structure of many Estonian words. Finally, we will look into techniques that would allow us to pre-filter the initial list of tag pairs to reduce the total processing time.

7 Acknowledgements

The work was supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Ahmet Aker, Monica Paramita, and Rob Gaizauskas. 2013a. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–411, Sofia, Bulgaria. Association for Computational Linguistics.

Ahmet Aker, Monica Paramita, and Rob Gaizauskas. 2013b. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 402–411.

Robert Gaizauskas, Ahmet Aker, and Robert Yang Feng. 2012. Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics*, pages 23–32.

Thorsten Joachims. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Andraž Repar, Matej Martinc, and Senja Pollak. 2019. Reproduction, replication, analysis and adaptation of a term alignment approach. *Language Resources and Evaluation*, pages 1–34.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2013. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*.

Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, pages 101–121.