Mapping Words Between Slovak Text and its Translation to English

Jakub Ševcech[*]
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
sevcech08@student.fiit.stuba.sk

Abstract

Word alignment in texts translated to different languages is used in various applications such as cross-language information retrieval. To search for equivalent words in text translations various statistical methods, methods based on position of words in phrases and methods based on bilingual dictionaries are used. However it is very difficult to use these methods in languages with big morphological complexity such as Slovak language. We propose method for word alignment for Slovak language with potential to be language independent. Proposed method uses bilingual dictionary for search for equivalent words and Levenshtein distance is used to compare different shapes of inflected words.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; I.7.5 [**Document and Text Processing**]: Document analysis

Keywords

Word-mapping, keywords, annotation

1. Introduction and Related Work

Word alignment is necessary in various applications where we are working with texts in different languages and the vector representation of texts is not satisfactory. Such applications are for example cross-language information retrieval, text summarization or automatic text annotation. In these applications we need to know the exact mapping between equivalent words in text and its translation.

Most of methods used for search for equivalent words in text translations use various statistical approaches [4], but also methods based on positions of words in the text [2] and methods based on dictionaries are used as well. However these methods have problems when text in language with big morphological complexity is processed. One of the main problems in processing of these texts is the presence of inflected words, where the word with the same meaning can be present in various shapes [7]. To transform these words to their base form, various methods for word stemming such as Porter algorithm [9] are used. Stemming reduces words to their stem (base form), that is the same for all morphologically related words. For languages with little morphological complexity, these algorithms are rather simple, but their usage do not improve significantly results of methods for information retrieval. Stemming of words in morphologically complex languages improves significantly the quality of results of methods for information retrieval [8], but the complexity of stemming algorithms for these languages is also greatly increased. In languages such as Slovak, stemming can be performed only manually or automatically with low quality.

We propose a method for mapping equivalent words between text and its translation using bilingual dictionary and Levenshtein distance to compare different shapes of words. Such method can be used not only to identify equivalent words in text translations, but may be also used to automatically associate every word in dictionary with all its shapes. Such list of words could be used as an alternative of methods for word-stemming in languages where it is difficult to create algorithm for word-stemming.

2. Method for Word-Mapping

We propose a method for connecting equivalent words in Slovak text and its translation to English, with potential to be language independent. In the design of the method, we focused on the fact that Slovak is flexive language, thus the same words can have a lot of different shapes. To cover all possible shapes of words in Slovak, we would need a dictionary of all possible shapes of words, that occur in Slovak language. Such dictionary would be huge, would be difficult to work with and even harder to create. We used therefore much smaller dictionary, in which every word is contained as a single shape and for connection of different shapes of words we used method similar to the method used in Slovak morphology analyzer in [5]. We

[*] Bachelor degree study programme in field Informatics. Supervisor: Professor Mária Bieliková, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava.
Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011 and recognized as best bachelor degree paper.

Ševcech, J. Mapping Words Between Slovak Text and its Translation to English. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 113-115

used Levenshtein distance [6] to compare different shapes of words. Our hypothesis is that we can work with words on level of strings of character with sufficient accuracy.

Levenshtein distance is a minimal number of Levenshtein edit operations necessary to convert one string of characters to another. In this conversion three operations are allowed: insertion of character, removal of character and replacement of character by another one. We adjusted the cost of individual Levenshtein operations depending on the position of character in the word, where operation took place. We take into account the fact that if the letter is changed in the root of word, thus the beginning of the word, the meaning of the word changes significantly. Therefore we increased the cost of operation at the beginning of word against the rest of the word. We also take into account the fact that difference in the shape of word is just difference in the words suffix. Therefore we let the cost of Levenshtein operations linearly decrease for last characters of the word.

The first step in process of mapping words between the text and its translation is the removal of stop-words in both texts. Stop-words, are all words that do not assume significance. Such words are for example conjunctions or articles.

The mapping of equivalent words is done according to Algorithm 1. In the process of word-mapping we assume that equivalent sentences are in the same order in both original and the translated text. With this assumption we browse through sentences in the text and we associate the word with its equivalent in the second sentence. For every sentence in the text we pass through all of its words. For each word from translated text, we find its translations using bilingual dictionary. For all translations of that word we calculate Levenshtein distance to words in the sentence in original text. If we find such translation of the word, that have Levenshtein distance to the word in the original sentence smaller than defined threshold, we declare these words equivalent translations, thus mapped words.

The result of processing of the text is a list of words and for each of these words, equivalent words are associated. For every word in the list, multiple mappings can be found. This is due the fact that in the text the same word can be present in multiple shapes or multiple translations of this words can be contained in the dictionary.

Proposed algorithm passes sequentially by phrases in the text an by words in sentences. It is thus easily parallelizable due to its data independence.

3. Evaluation

We experimented with the aim to verify the success rate of proposed method for mapping equivalent words between Slovak text and its translation into English. As a test sample we used part of the textbook with the subject of software engineering. After the step of stop-words removal, the test sample consisted of 1 928 words. We implemented proposed method along with two enhancements. These enhancements were implemented to increase the total number of mapped words. In the first improvement we assume that in most cases, if two words are adjacent in one sentence, they will be adjacent in the translated sentence as well. Therefore the enhancement passes the mapped pairs of words and if both mappings have unmapped neighboring words, they are linked and declared as mappings. Second improvement resolves the different shapes of words (even in non-flexive languages such as English). We preprocessed every entry in the dictionary used in mapping process, so that all English words in this dictionary are stemmed using Porter algorithm. Then in process of word-mapping, we use preprocessed dictionary to search for translations of stemmed words. Using these advancements we have created four functions:

Algorithm 1 Find mapping for words in parallel texts

```
origSentences = splitToSentences(originalText)
translSentences = splitToSentences(translatedText)
for i = 0 to |origSentences| do
    for all origWord in origSentences[i] do
        for all translWord in translSentences[i] do
            translations = findTranslations(translWord)
            for all transl in translations do
                if (levenshteinDistance(transl,origWord) <
                  threshold) then
                    declareMappings(origWord,translWord)
                end if
            end for
        end for
    end for
end for
```

1. basic function,
2. function that takes into account positions of words in the text,
3. function with stemmed dictionary and
4. function that applies both enhancements.

The results of experiments on these functions are summarized in the Table 1. We recorded the number of correctly mapped words, the number of incorrectly mapped words and the number of assignments, where the correct words were mapped, but to the same word other, incorrect words were mapped too. The last column of Table 1 shows the ratio of all mapped words to all words in the test sample. We see that the ratio of correctly mapped words to all mapped words in the basic function is more than 90 %. However, the number of all mapped words is not even half of the number of all words in the sample. Similar result was achieved using function enhanced by preprocessed dictionary (second enhancement). The number of correctly mapped words is still more than 90 % of all mapped words, but portion of mapped words to total number of words in the sample increased slightly. Both functions that take into account the position of unmapped words in a sentence (function with first enhancement and with both enhancements) reached the ratio of mapped words to all words more than 80 %, but the number of errors in created mappings is disproportionately increased.

Function that works with both improvements is able to find the largest number of mappings, but it produces many errors. The best ratio between number of found mappings and number of errors is achieved when basic function with stemmed dictionary is used.

Table 1: Results of method for word-mapping and its enhancements.

Function	Correct	Incorrect	More	Mapped/All
Basic	810	51	14	45.38 %
1^{st} enhancement	901	538	195	84.85 %
2^{nd} enhancement	993	60	21	63.59 %
Both enhancements	1040	405	178	96.08 %

4. Word-Mapping Usage Scenario

Currently, methods for keyword extraction do not achieve satisfactory results when texts in languages with big morphological complexity such as Slovak are processed. However keyword extraction in English texts provide much better results. We increased the quality of keywords extracted from Slovak text using method for keyword extraction on text translated via Google Translate into English. Similar approach is used in [3] to compare services for keyword extraction on various types of texts. Slovak and Czech texts automatically translated to English achieved comparable, if not even better results as texts in English. With the help of the method described in this article, we found Slovak equivalents of keywords extracted from text automatically translated into English.

We used this method as a support method for creation of annotations [1] for keywords in Slovak text. Created annotations provide definitions or additional information for these keywords. We evaluated our method within an educational framework ALEF [10], where created annotations are presented to students in form of list of links to related web pages and in form of definitions of keywords occurred in the learning objects presented by ALEF. Annotations associated with keywords were presented along with other types of annotations such as comments, highlights and tags created by students.

Other possible application of proposed method is to associate every word in a dictionary with all its shapes. Such list of words could be used as an alternative to methods for word-stemming in languages where it is difficult to create efficient algorithm to convert words to its base form. To perform this step, function with stemmed dictionary or function with no enhancements are the most suitable. In this task the quality of extracted mappings is the priority. Higher number of found mappings can be achieved by increasing the amount of processed text used as a training sample.

5. Conclusions

We proposed the method for search for mapping of equivalent words in text and its translation into English. We verified this method on the sample in Slovak language, but it is proposed in a way that it can be used for different languages with similar structure. For the proposed method, we developed two improvements that increase the ratio of mapped words to the total number of words in the processed text. The trade-off between the quality and number of found mappings is possible to control by choice from enhancements proposed along with method for word-alignment and by setting the threshold used in comparison of words using Levenshtein distance.

Acknowledgements. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

[1] M. Agosti and N. Ferro. A formal model of annotations of digital content. In: *ACM Trans. Inf. Syst.* , 2007, vol. 26, no. 1, pp. 3+.

[2] A.-M. Barbu. A Positional Linguistics-Based System for Word Alignment. In: *Lecture Notes in Computer Science*, 2004, vol. 3206, pp. 23–30.

[3] M. Barla and M. Bieliková. Ordinary Web Pages as a Source for Metadata Acquisition for Open Corpus User Modeling. In: *Proc. of IADIS WWW/Internet 2010.* IADIS Press, 2010, pp. 227-233.

[4] W. Gale, K. Church. Identifying word correspondences in parallel texts. In: *Fourth DARPA Workshop on Speech and Natural Language*. 1991, pp. 152–157.

[5] E. Garabík. Slovak morphology analyzer based on Levenshtein edit operations. In: *Proc. of Workshop on Intelligent and Knowledge oriented Technologies*. 2006, pp. 2–5.

[6] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In: *Soviet Physics Doklady*. 1966, vol. 10, pp. 707+.

[7] A. Pirkola, T. Hedlund, H. Keskustalo and K. Järvelin. Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. In: *Information Retrieval*. 2001, vol. 4, no 3, pp. 209–230.

[8] M. Popovic and P. Willet. The effectiveness of stemming for natural-language access to slovene textual data. In: *J. Am. Soc. Inf. Sci.*. 1992, vol. 43, no 5, pp. 384–390.

[9] M. F. Porter. An algorithm for suffix stripping In: *Program*, 1980, vol. 14, no. 3, pp. 130–137.

[10] M. Šimko, M. Barla, and M. Bieliková. ALEF: A framework for adaptive Web-Based learning 2.0 In: *Key Competencies in the Knowledge Society, ser. IFIP Advances in Inf. and Communication Technology*, Springer, 2010, vol. 324, ch. 36, pp. 367-378.