# Automatic Annotation of Non-English Web Content

Jakub Ševcech and Mária Bieliková
Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava, Ilkovicova3, Bratislava, Slovakia
E-mail: sevcech08@student.fiit.stuba.sk, bielik@fiit.stuba.sk

*Abstract*— **Nowadays we are facing the daily information overload. It is thus difficult to get exactly the information we need. It often happens that while reading, we find a word we do not understand and we would need an explanation or some additional information about this word. For this purpose annotations in the Web environment are created and attached to such words. In this paper we propose a method for an automatic extension of the content available on the Web by adding annotations to selected terms (keywords) in the text. The method is designed to be able to insert annotations into the text written in Slovak with a potential to be language independent. Annotations themselves are obtained through publicly available services providing information retrieval. We adapt created annotations taking into account implicit feedback from users in form of clickthrough data. We evaluate the proposed method in the environment of an educational web-based system.**

***Keywords- Web annotation; keywords; keywords mapping; adaptive annotations***

## I. INTRODUCTION AND RELATED WORKS

While reading a web page, a visitor often encounters a word or a phrase, he does not understand, or he would require some additional information about this term. This situation occurs more frequently if the page contains technical or explanatory text, such as for example in digital libraries [4] and various educational sites [7]. Common scenario that follows when a website visitor encounters an unknown expression is the following: the user opens a new tab in his web browser, displays his favorite search engine, and starts to search for a definition or an explanation of this word. This scenario has several drawbacks for the user and for the author of the website as well. Website visitor has to stop working with the document and has to shift attention to work with other sources in order to search for additional information. The visitor often does not return to the original site anymore, so we lose the reader.

One of the solutions to this situation would be an annotation [5] attached to this word. Such annotation can immediately provide us with an explanation of the unknown word. Content annotation is an active area of research on the Web. Moreover, with advent of the Web 2.0 many applications facilitating annotations were developed.

There are two basic types of tools for adding annotations into documents. Firstly, there are tools that do not focus directly on creating annotations, but on supporting readers of the documents in creating annotations exist. Examples of such tools are Diigo (www.diigo.com/) or AnnotatEd [8]. These tools provide a number of supporting instruments for users, through which they can manually add annotations into a web document and consequently they can share them. However, with the existing amount of information and documents on the Internet, it is impossible for users of these documents to annotate them all.

Secondly, there are annotation tools that aim at creating annotations automatically. The process of creating annotations can be divided into two basic parts: positioning of annotation at the right place and search for information to fill the annotations. To find a location to which it is appropriate to assign the annotation various ATR algorithms [11] or different approaches from the field of Natural Language Processing (NLP) are used. However, satisfactory results are currently achieved only for English texts. To overcome this problem it is possible to use machine translation to translate text into English. We suppose that for this task existing translating mechanisms are sufficient, even though they are still not perfect. We need mainly *important* words in the sentence such as nouns and verbs and these are translated sufficiently in most cases. Since we attach explanations to terms in the original text, we have to find the mappings between extracted keywords and their equivalents in the original text.

There are several ways to search for information to fill annotations. One way is to use pretreated database to retrieve information to fill annotations. Similar method is used for example in the instrument Pannda [3]. This approach can be only used if we create annotations for specific domain. Considering this approach, we need specific database for each particular domain, which is not applicable generally for the Web.

Wikipedia Miner toolkit [2] uses similar approach, but it links keywords in text with corresponding articles in Wikipedia. Thanks to Wikipedia this approach is applicable basically for every domain.

Other way to obtain information to fill the annotations is to use freely available services for information retrieval, such as the tools Gnosis (www.opencalais.com/Gnosis), DictionaryTooltip (www.dictionarytip.com) and many others do it. Using such services, annotations can be created for every domain and they can provide any type of information that is gathered through these services.

In this paper we propose a method for automatic annotations of documents in non-English language and adaptive presentation of the created annotations. We consider the annotation a definition of the word, or a set of links to web pages related to that word found by existing services for information retrieval.

## II. METHOD FOR WEB CONTENT ANNOTATION OVERVIEW

Our method consists of four steps:

1. Elimination of redundant parts of web page and selection of a text to be annotated
2. Search for candidate words for annotations
3. Search for information to fill the annotations
4. Adaptation and visualization of annotations

Before searching for the annotations, it is necessary to analyze the document and find the words to which it is appropriate to assign the annotations (steps 1 and 2). As the first step it is necessary to remove redundant parts of the web page such as various navigation elements, advertisements, etc. It is necessary to select only the text that speaks about the main content of the page from the web page body and only this text is postponed for further processing. In our current implementation we use the Readability service (www.readability.com ) for this task.

The second step is a search for candidate words for the assignment of annotations. Currently, machine analysis of a text achieves satisfactory results only for English texts. We believe that this is sufficient for several languages including Slovak language and therefore we translate analyzed text into the English language. Without the use of these results, the quality of found keywords and hence the quality of annotations would decrease significantly. When searching for keywords it is possible to use many different algorithms and services [6, 11]. To connect created annotations to the correct words in the original text, we proposed a method for finding mappings between extracted key words and their equivalents in the original text. This method is key element for annotation acquisition for various languages.

For annotation creation we use publicly available services for information search (see evaluation section for more details). These services provide different types of information and also of different quality according to the required purpose. The annotations may take form of words definitions, links to related sites or multimedia (like video or an image). We use services providing definitions of keywords and services providing links to web pages related to the keyword.

Finally, created annotations are visualized to the user. Before actual visualization we reorder the content of annotation according to the implicit feedback gathered from annotation usage.

We evaluated our method within an educational framework ALEF [7], where created annotations are presented to students. Annotations are in form of tooltip attached to keywords in text. Tooltip contains a list of links to related web pages and definitions of keywords occurred in the learning objects presented by the ALEF. Created annotations are presented along with other types of annotations provided by ALEF. Students can tag or comment learning objects, they can highlight text for better remembering, attach annotations in form of links to external sources or annotate text by simple questions.

Our method is designed to be applicable for any kind of content. Relevance of provided annotations heavily depends on the quality of used services. Dependency on other services can be seen as a disadvantage. However, it enables wider applicability and concentration on other important issues such as language independence or adaptive presentation of acquired annotations.

## III. MAPPING CANDIDATE WORDS

We connect equivalent words in two texts (original and its translation) using bilingual dictionary. We primarily consider Slovak language, which is flexive language with many shapes of words and can represent (considering its syntax) rather large group of languages. Our concern was effectiveness. Effective processing of huge amounts of texts is more important as having exact stemming method. Our hypothesis is that we can work on morphological level on the level of strings with sufficient accuracy considering task of mapping words for the annotation process.

We use a dictionary, in which every word is located as a single shape and for connection of different shapes of words we used a method similar to the method used in Slovak morphology analyzer [9]. We used comparison of words based on Levenshtein distance. Levenshtein distance of two words is a minimal number of Levenshtein edit operations necessary to convert a string of characters to another. During this conversion three operations are allowed: insertion of character, removal of character and replacement of character by another one.

We also adjusted the cost of individual Levenshtein operations depending on the position in word where the operation took place. We take into account the fact that if a letter is changed in the root of a word, the meaning of the word changes significantly. Thus we double the cost of the operation at the beginning of the word. We also take into account the fact that the difference in the shapes of words are just differences in the words suffix, so we let the cost of operations linearly decrease in the last characters of the word.

The first step of mapping words between the text and its translation is the removal of stop-words in both texts. In the process of mapping words we assume that equivalent sentences appear in the same order in both original and the translated text. With this assumption we browse through each pair of both, the sentence and its translation. We move through the words in the translated sentence and we seek for translations of each word in a bilingual dictionary, just like if we tried to translate the translated sentence word by word back into Slovak. We then compare each translation from dictionary using Levenshtein distance with every word in the original sentence. If calculated distance of two words is less than established threshold, we declare them to be mapped.

## IV. ANNOTATIONS ARRANGEMENT ADAPTATION

Adaptation of annotations arrangement is based on an implicit feedback derived from users' behavior. When the user interacts with the content of the annotation, we gather implicit feedback in the form of the fact that the user clicked on presented element of the annotation and that he did not click on the other. Elements of the annotation content are presented in a list, while the user is affected by its arrangement. We therefore do not assign the same weights to

clicks on the elements placed in various positions in the presented list.

We interpret a click as a statement that the clicked element is better or more relevant than other element offered at the same time. Similar approach for interpreting implicit feedback is used in [10], where five strategies for interpreting these statements are proposed.

We use the following for reordering list of annotations:

1. **Click > Skip Above**, where the element user clicked is better than all the elements listed on higher positions and which user did not click.
2. **Click > No-Click Next**, where the element user clicked is better than the immediately following element which user did not click.

Based on these strategies and users' feedback, we get a set of statements about the quality of the provided information. These statements are then used to rearrange the content of the annotations. We consider these statements as oriented edges of graph, where the elements of content of the annotation are the nodes of the graph. The statements gathered from users' feedback can repeat for the same elements and moreover they can be contradictory. We therefore combine repetitive edges to one, where the number of combined edges is stored in the weight of the result edge.

Subsequently we use adapted PageRank algorithm that takes into account the weights of the edges and we calculate the rating of nodes. Using PageRank we ensure that contradictory statements are taken into account. When we arrange nodes by their decreasing rating, we obtain the new order for the elements of content of the annotation.

## V. EVALUATION

We experimented with the aim to evaluate the success rate of the proposed method for mapping equivalent words between the Slovak text and its translation into English. As a test sample we used part of the textbook for the course Principles of software engineering. After removing stop-words, the test sample consisted of 1 928 words.

We implemented the proposed method along with two enhancements. The first improvement maps the words that failed to connect using the basic method. In this improvement we assume that in most cases, if two words are adjacent in one sentence, they will be adjacent in the translated sentence too. We therefore proposed a method that passes the mapped words in sentence and if both mappings have unmapped neighbouring 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 preprocess every entry in the dictionary used in mapping process, so that all English words in this dictionary are stemmed using Porter algorithm. Stemming reduces words to their stem that is the same for all morphologically related words. In the process of words mapping we use preprocessed dictionary to find translations of stemmed words.

Using these advancements we created four functions:

- basic function,
- function taking into account positions of unmapped words,
- function using stemmed dictionary and
- function applying both enhancements.

The value of threshold used in our experiments during comparison of words using Levenshtein distance was equal to the sum of the costs of three Levenshtein operations.

The results of experiments on these functions are summarized in Table 1 and Table 2.

TABLE I. SUCCESS RATE OF MAPPING FUNCTIONS FOR EQUIVALENT WORDS.

| Function | Correct | Incorrect | More |
|---|---|---|---|
| Basic | 92.75% | 5.82% | 1.60% |
| 1st enhancement | 55.14% | 32.92% | 11.93% |
| 2nd enhancement | 92.45% | 5.58% | 1.95% |
| Both enhancements | 64.07% | 24.95% | 10.96% |

TABLE II. NUMBER OF MAPPED WORDS COMPARED TO TOTAL NUMBER OF WORDS.

| Function | Mapped words number / Total words number |
|---|---|
| Basic | 45.38% |
| 1st enhancement | 84.85% |
| 2nd enhancement | 63.59% |
| Both enhancements | 96.08% |

In evaluating the success rate of the function for mapping of equivalent words (Table 1), we recorded the number of correctly mapped words, incorrectly mapped words and the number of assignments, where the correct words were mapped, but along with these words, other incorrect words were attached as well. Table 2 shows the ratio of all mapped words to all words in the test sample.

We see that the ratio of correctly mapped to all mapped words when basic function was used is more than 90 %. However, the number of all mapped words is only a little more than 45 % of all words in the sample. Similar results were achieved in function using a preprocessed dictionary. The number of correctly mapped words stays above the limit of 90 % of all mapped words, but the portion of mapped words to all words in the sample increased slightly. Both functions taking 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 mapping is disproportionately increased.

The function working with both improvements is able to find the largest number of mappings, but it produces many errors. The best ratio between the number of found mappings and the number of errors is achieved when the basic function with stemmed dictionary is used. This ratio can be even better, if better dictionary is used.

As the next step, we focused on evaluation of the quality of information we gathered through publicly available services. We selected 16 texts on software engineering. In these texts, we extracted keywords using the AlchemyAPI service. Then we gathered additional information using Google Search, SlideShare, Dbpedia and DictService.

Google Search and SlideShare took the keyword as an argument and returned a list of hyperlinks to related resources. DictService returned definitions from different dictionaries for the query in form of the keyword. When searching for keywords using AlchemyAPI, we were able to search for concepts of the processed document. With these concepts we received a link to corresponding resource in Dbpedia. We used SPARQL to seek for websites related to that resource.

Volunteers were then asked to say whether they will find the gathered information useful and whether they are relevant to the analyzed text and to the keywords used to search for this information. The results we obtained are summarized in Table 3 where the ratio of relevant and irrelevant information gathered by different services for information retrieval is shown.

TABLE III. RELEVANCY OF GATHERED INFORMATION.

| Service | Relevant | Irrelevant |
|---|---|---|
| Google Search | 70.01% | 26.98% |
| DBpedia | 63.29% | 31.90% |
| DictService | 59.64% | 40.36% |
| SlideShare | 26.32% | 72.79% |

A small amount of links returned by evaluated services, was corrupted, thus it was not possible to evaluate its relevance (we omit them here). We observed big differences in relevancy of returned information between the compared services. We believe this difference in relevance of information gathered through SlideShare service is caused by the narrow focus of the documents provided by this service. The quality of the annotations thus heavily depends on the services used to gather information.

It is necessary to properly choose services to search for the annotation content. None of the evaluated services reached success rate approaching 100%, so there remains a place for improvement of their utilization, as well as space to emphasize the relevant information. We approach this problem by reordering information according to implicit feedback. Drawback of the approach is that we first need users to click and in such way evaluate the annotations. However, this can be overcome rather quickly considering the power of collaborating users also in connection to specific settings of educational systems, where the users obviously have motivation to go through the resources.

## VI. CONCLUSIONS

We proposed a method for automatic creation of annotations for web pages written in the Slovak language. We proposed the method for mapping equivalent words between the text and its translation. Our method is not constrained by Slovak language. It is open to other languages with a similar structure. It enables effective use of results achieved for keyword extraction in English, which is well elaborated and still evolving.

We evaluated our approach and confirmed that this current quality of language translation is sufficient for such task. This brings new possibilities for analysis and enhancement of non-English web sites.

We evaluated the quality of the information gathered from publicly available services for information retrieval. We showed that the quality of created annotations heavily depends on the quality of used services. Both experiments were performed independently from the educational system in which the proposed method is implemented.

As number of annotation gathered by available services is obviously rather high, we proposed a method for adaptation of annotations, based on user implicit feedback in form of clickthrough data. We used adapted PageRank algorithm to find ratings of annotation content elements, where edges of analyzed graph were created using implicit feedback from users. In our future work we plan to concentrate on this aspect and use term-based user model [1] for effective personalization of annotations.

## ACKNOWLEDGMENT

This work was partially supported by the grants VEGA 1/0675/11/2011-2014, KEGA 028-025STU-4/2010, APVV-0208-10 and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

## REFERENCES

[1] M. Barla, “Towards Social-based User Modeling and Personalization”, Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 3, No. 1, 2011, 52-60.

[2] D. Milne and I. H. Witten, “Learning to link with wikipedia,” in Proc. of the 17th ACM Conf. on Information and Knowledge Management. New York, NY, USA: ACM, 2008, pp. 509-518.

[3] M. Adam, “An approach to automated on-line annotation,” in *Proc. of research project workshop Tools for Acquisition, Organization and Presenting of Information and Knowledge*, 2007, pp. 20–25.

[4] M. Agosti and N. Ferro, “Annotations: Enriching a digital library,” in *Research and AdvancedTechnology for Digital Libraries*, ser. LNCS, Springer, 2003, vol. 2769, pp. 88-100.

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

[6] 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.

[7] 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.

[8] R. Farzan and P. Brusilovsky, “AnnotatEd: A social navigation and annotation service for web-based educational resources,” In *New Rev. Hypermedia Multimedia*, 2008, vol. 14, no. 1, pp. 3-32.

[9] R. Garabík, “Slovak morphology analyzer based on Levenshtein edit operations,” in *Proc. of Workshop on Intelligent and Knowledge oriented Technologie*s, 2006, pp. 2-5.

[10] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, “Accurately interpreting clickthrough data as implicit feedback,” in *Proc. of the 28th annual int. ACM SIGIR conf.* New York, NY, USA: ACM, 2005, pp. 154-161.

[11] Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna, “A comparative evaluation of term recognition algorithms,” in *Proc. of 6th Int. Conf. on Language Resources and Evaluation,* 2008, pp. 2108-2113.