ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20180702105206
Original scientific paper

Web Sites Internationalization using Domain Translator

Danijel RADOŠEVIĆ, Andrija BERNIK, Nikola MRVAC

Abstract: The paper presents the concept of the Domain Translator (DT) and the example of its implementation for the purpose of web sites internationalization. The majority of Croatian web sites, especially in the domain of academic community, offer only their narrowed versions containing some general information in foreign languages. The users that do not speak Croatian try to cope with this problem by using translators like Google Translate[1], but only with a limited success, because of various problems, including the loss of local heritage like local names, shortcuts, and appropriate textual context. In the scope of this paper, a Domain Translator in a form of web browser extension was built, as well as the database of lexical artifacts covering an example of faculty web site. Given translations were compared with the corresponding results of the Google Translate and Bing Translator[2].

Keywords: machine translation; Domain Translators; web sites internationalization

1 INTRODUCTION AND MOTIVATION

Today it is expected from the web site of an academic institution in the non-English speaking regions to have a foreign language version, mostly in English. However, it is not easy to achieve this goal in the completeness, mostly due to the lack of human resources (the need for translation) and the need to adapt the database and other software resources to support multilingualism. For these reasons, the majority of university web sites in Croatia have their foreign language versions that cover only some general information and other information that is rarely changed. However, there are some categories of users who do not speak Croatian, especially the foreign students, who need some information on a daily basis. That includes all news, schedule of the lectures, Learning Management System (Moodle for most of the Croatian universities), notifications of student office, web services of library, applications for final and graduate papers, and many others.

According to the past experiences, the students mostly cope with this problem by using Google Translate and similar services, to translate the texts into English or their domicile languages. The first problem with these services is that they sometimes do not give translations for all web pages (e.g. that was noticed for some pages inside the foi.unizg.hr domain). After that, they do not know the local heritage, like different names, abbreviations, existing official translations, etc., so some invariant words could be translated, causing confusion, or some inappropriate translation. In addition, from some reasons, some parts of menus and other texts could remain untranslated (Fig. 1).

The things could be even worse when there are some special kinds of text, like file names (often written with "_" instead of blank spaces, as shown in Fig. 1), text combined with images, PDF documents and other where some local adaptation is needed.

All these problems point to a solution in a form of a highly adapted translator that could solve the mentioned problems and offer a translation that preserves the local heritage, along with comparable or even better quality of translation. To achieve this goal, there are some assumptions in favour of the local oriented translators. At first, they should translate only in one direction (e.g. from Croatian to English). After that, languages like Croatian use much more lexical forms in relation to the English language in order to express genders, cases, different conjugations etc.

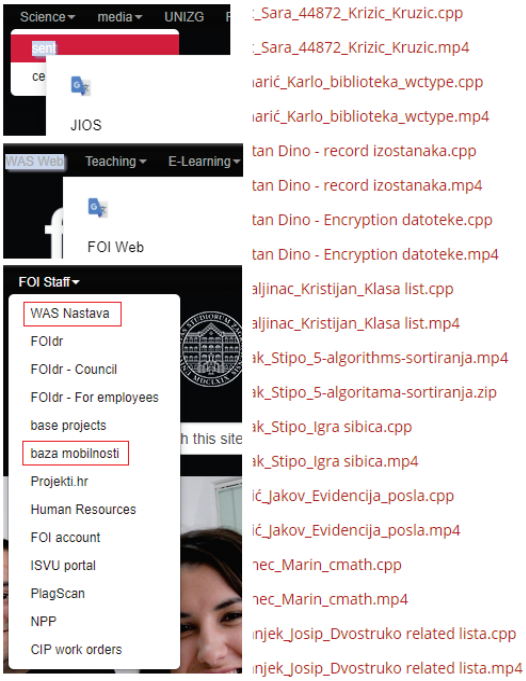

Figure 1 Google's problems with local heritage

For that reason, the correct form in the source language will be in most cases associated with the correct form in the English language, which usually does not work in the opposite direction. In addition, some special situations like usage of "_" instead of spaces, text covered by images etc. could be solved in locally specialized translators, offering the users of translated web service the experience that is not far from the web experience of the domestic users.

In this paper the example of the Domain Translator, implemented in a form of web browser extension will be presented. This form was chosen for the testing purposes, because it does not require any changes in the covered Web pages.

2 RELATED WORK

The proposed translation model in this paper is based on statistical approach that uses translation phrases in order to obtain the translation. This approach is in its base similar to the approach of Koehn et al., who have created translation model and decoder to evaluate and compare various translation methods, and their results showed that phrase translation has better performances than word-based methods [1]. Chand performed an empirical survey of the rule based translation systems and statistical based machine translation systems, and it was shown that statistical systems have better performances [2].

Today is one of the mayor challenges in machine translation to make a successful localization of web services, as noted in Jiménez-Crespo research [3]. On the other hand, many enterprises as well as educational institutions from non-English speaking regions have a problem of representing their web pages in foreign languages, mostly in English. As noted by Costales, the field of institutional websites has not been earlier addressed by researchers working in Translation Studies [4]. Therefore, the users that do not speak the local language often use Google Translate to cope with the language problem. As noted by Chen et al., Google provides accurate translation for simple sentences, but has a lot of problems with more complex and, often, domain specific sentences (e.g. related to diabetes) [5].

The problem of domain adaptation is characteristic for statistic machine translation but it can be overcome by domain-specific learning of the translation system [6]. There are different elements of domain adaptation that should be taken into account. One of the most important is to find unseen words [7], and, as found by Choi et al., domain specific abbreviations, technical specifics like PDF-to-text conversion errors, lexical diversity, name expansion, and scattered strings [8].

The usual approach in domain adaptation is to use a large amount of out-of-domain training data combined with a much smaller amount of in-domain training data to optimize translation performance on that particular domain [9, 10]. On the other hand, the approach described in this paper is oriented mostly on usage of in-domain training data, except the corpus of usual words and phrases that are common for different language domains.

For the purpose of machine translation evaluation, probably the most popular method is BLEU. BLEU [11] uses a referent translation which is perceived as "good" to compare it in a relatively simple way with the translation obtained by the translation system. Some other methods, like ORANGE [12], METEOR [13] and LEPOR [14] are developed with the intention of obtaining higher correlation with human judgments. In this research was used the BLEU method, together with the internal metrics of translation system that is based on used translation phrases of different size.

3 TRANSLATION MODEL

The machine translation model used in the Domain Translator arises from the code generation model as described in Radošević and Magdalenić earlier work [15], and the translation model described in [16]. The idea in a base of this approach is to adapt the model that uses the Domain Specific Language (DSL) to produce the code into a translation model that works with a natural language. The main difference is in the complexity of specification language in relation to the target language, where there is a huge misbalance in the case of code generators (DSL vs. target program code) and the translator, where the languages (e.g. Croatian and English) are both very complex.

However, despite the complexity of both used languages, some special cases of translation are used in building of domain-oriented translators:

- only one translation direction,
- relatively narrow lexical domain (for example web contents of some academic institution or an enterprise)
- translation from a language that uses more lexical forms (like cases, genders, different conjugations etc.) into a language that uses less lexical forms (e.g. from Croatian to English) On the other hand, there are some specific requirements for the domain translators:
- local names, expressions, abbreviations, letter capitalization and other parts of the local heritage have to be preserved in the translation,
- specifics of the environment should be taken into account.

These are mostly related to the different forms of web environment, like menus, forms, table contents, PDF documents represented in an html form, usage of images that overlap the text etc. Therefore, for the needs of domain oriented translators, the model described in [16] was adapted, as shown in Fig. 2.

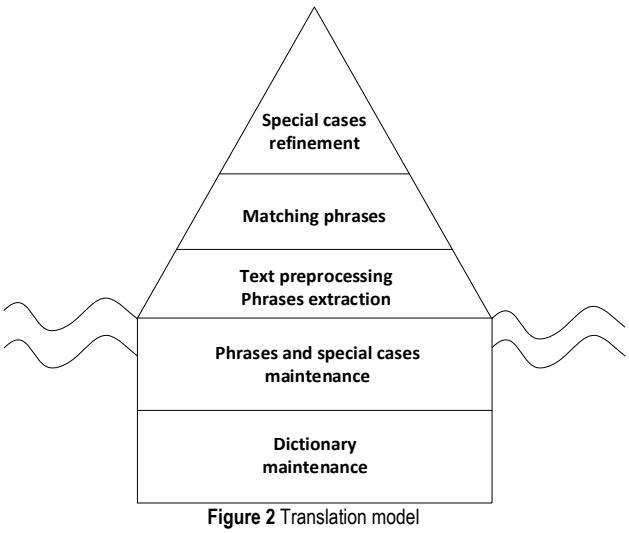

Figure 2 Translation model

The translation model uses a bilingual dictionary, which is maintained by the students using the web interface[3]. The dictionary includes words, lexical tags, pronunciation, different meanings and examples of use with phrases.

For the purpose of translator, the specific Web interface was developed[4], Fig. 3.

The structure of phrases database is simple, containing only phrases in both languages, together with the priority value. If the phrase is inherited from the dictionary, the priority depends on the order of meanings, but if the phrase is entered by translator editor's interface, the priority value

is equal to zero (highest priority). In the case when two or more phrases have the zero priority, there is a possibility of cleaning the database, where the priority of older records is being lowered.

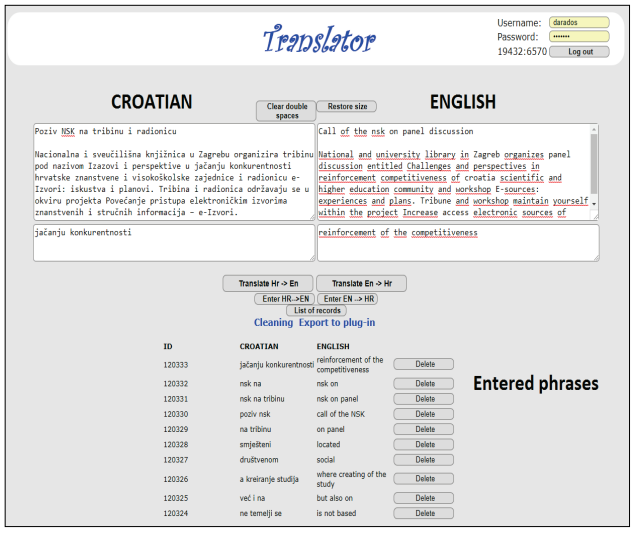

Figure 3 Translator interface

All the translation phrases are being exported to a file for the needs of browser extension. Only the highest priority phrases are included, so the lists of candidates for translation were avoided. That is the difference in relation to the model presented in [16] that simplifies the translation process with no significant loss of accuracy, which is possible because of the translation direction, from the language that uses more conjugational forms to a language that uses less conjugational forms.

The translation process is implemented both in a form of web translation interface and the form of browser extension. In both cases, the translation process starts by text preprocessing. In this phase, the text for translation is being transformed into a list of words and separators. After that, the phrases containing 1-5 words are being formed. The phase of matching phrases tries to find the longest phrase with the translation in the translation phrase database. Such phrase is observed as the candidate for translation. Finally, the candidate for translation should be refined according to the special cases (mostly parts of the local heritage), e.g.:

- local abbreviations (usually in capital letters)
- single words translations (meaning is sometimes different than usage as a part of the sentence)
- professional titles (usually written in a specific way)
- abbreviation used for the days of the week, months of the year etc.

Special challenge in building of Domain Translator in a form of browser extension was to make the translation process fast enough to be used for the purpose of web pages translation. For that reason, each step of the translation process has to be optimized, while the fast matching phrases of different sizes seems to be the main bottleneck.

4 TRANSLATION INTERFACE IN A FORM OF BROWSER EXTENSION

The translation interface in a form of browser extension[5] (Fig. 4) is developed for the purpose of web site internationalization.

The database, exported from the web interface, is being obtained once daily or by user request. The working modes include the list of unrecognized words (which can be copied to the web interface, for learning), statistics about used phrases (for self-evaluation of the translator) and just translation.

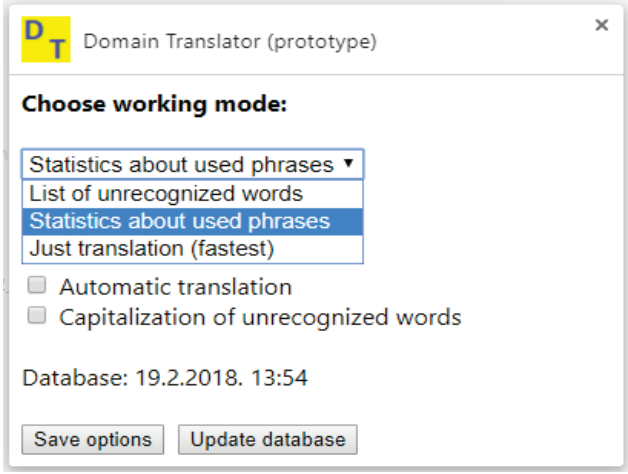

Figure 4 Translator interface in a form of browser extension (options)

It is also possible to automatically translate the pages (without click on the extension button) and to capitalize the unrecognized words.

Browser extension approaches each node of the web page to find the text to be translated. The translation process uses the model described in section 3, with some implementation differences between the Web interface and the browser extension. At first, it uses a local copy of phrases database, which is searched by using of simple binary search, unlike the web interface version, which uses standard SQL queries.

It is necessary to check the exact match of phrases consisting of 1-5 words. If the phrases are not found, another search tries to find the most similar word that satisfies a threshold of 80% of similarity (according to the Levenstein distance of the strings).

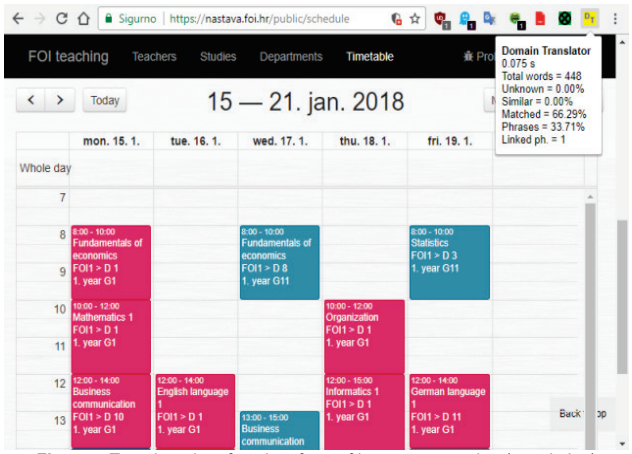

Figure 5 Translator interface in a form of browser extension (translation)

Unrecognized words could be names and remain unchanged in the translation. Fig. 5 shows the example of web translation by a browser extension in the statistical mode, giving the short statistics about used phrases, unknown words, time of translation and similar.

5 MODEL EVALUATION

The developed Domain Translator (DT) has used the database of about 39000 translation phrases, before the domain adaptation. Of that, about 6000 phrases are inherited from the Croatian-English and English-Croatian dictionary, and the rest of 33000 were entered by the Web translation interface. According to the size of phrases, before entering some new phrases within the domain adaptation, the state was as follows (Fn = phrase of *n* words):

F1: 28333 (72,65%)
F2: 7418 (19,02%)
F3: 2272 (5,83%)
F4: 778 (1,99%)
F5: 199 (0,51%)

The database was mostly trained on the pages inside the web service of the Faculty of Organization and Informatics (www.foi.unizg.hr).

For the test purposes, the DT in a form of browser extension was compared to the results of the Google Translate and Bing Translator. The test was conducted on the sample of abstracts of the Printing & Design 2018 Conference[6] that is dedicated to printing. Therefore, the abstract topics were outside the domain in which DT was previously trained. Each abstract was in Croatian and English, where the English version was used as a referential translation for the widely accepted BLEU[7] evaluation method.

The testing sample includes ten documents that were translated with Google Translate, Bing Translator and three times with the DT (before and after each of two phases of the domain adaptation where the specific phrases from test documents were entered). The results for Google and Bing were given using the BLEU metrics, while DT was also tested by using its internal metrics, the coverage of the original text by translation phrases of different size and the number of unknown words.

In the first phase of testing, DT was tested before the domain adaptation and compared to Google and Bing. The results are shown in Tab. 1, columns C2-C5 (for DT) and C9-C10 (Google and Bing). DT had some percentage of unknown words, which varies from 6% to 29%. The highest values are for the texts that use domain-specific special names (e.g. document 8) and specific language variants (e.g. dialect; documents 9 and 10). The coverage of text by translation phrases was relatively weak (14,38% in the average) due to the untrained translation engine. So, it is obvious from the columns C5, C9 and C10 that Google and Bing had better results in this initial phase of testing.

After these initial results, the following step was the **domain adaptation**. For that purpose, the translator interface has the "List of unrecognized words" mode (Fig. 4), so the unknown words were listed and entered into translation database by using of the translator Web interface (Fig. 3). By entering all of the unknown words (210 in the example), the results for the DT was somewhat improved, as shown in Tab. 1, column C6.

Entering of the domain-specific translation phrases is somehow a more complex task. To do this, it is necessary to identify them in the text to be translated, or, in the real case, in the lexical domain to be covered by the DT. In this example, the test documents (abstracts of the scientific papers) were merged into one document, together with the full papers in Croatian language. The scripts in Python were used to extract all phrases containing 2-5 words, together with the number of occurrences. After that the phrases already existing in the translator database were excluded. Also, another mechanism of filtering phrases was used. Some phrases are "trivial" i.e. could be obtained by word-to-word translation, so there is no need for them in the translation system. To recognize them, a list of previously acquired English phrases (about 31000 phrases; obtained by the same mechanism as the Croatian phrases) was used. The process was as follows:

- each Croatian phrase was translated word-to-word
- if translated phrase exists in the list of English phrases, then it is probably trivial.

This process has resulted by exclusion of about 25% of phrases, while the others (184 in the example) were entered manually by using the translator Web interface. Entering new translation phrases of 2-5 words had a more significant impact on a BLUE score than entering just the unknown words, as shown in Tab. 1, column C8.

Table 1 Testing results before and after the domain adaptation

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Document	Words	Phrases* %	Unknown %	**BLEU test0**	BLEU test1	Phrases* %	**BLEU test2**	BLEU Google	BLEU Bing
1	189	6,50	8,50	**1,03**	1,55	16,40	**6,22**	4,28	7,76
2	187	8,02	7,49	**7,36**	9,20	21,93	**21,24**	32,49	35,39
3	166	22,89	6,02	**14,61**	17,15	37,95	**34,59**	40,73	46,21
4	176	14,77	8,52	**14,08**	25,43	31,82	**32,02**	28,01	37,09
5	187	18,72	8,56	**7,56**	8,33	47,59	**34,83**	23,49	22,50
6	213	11,88	17,82	**5,37**	8,11	30,99	**19,67**	18,40	19,49
7	104	14,04	12,28	**3,62**	10,39	53,85	**52,81**	30,25	31,25
8	148	9,62	28,85	**21,87**	32,51	39,86	**60,56**	43,40	40,80
9	104	6,73	26,92	**6,64**	25,00	48,08	**64,10**	20,72	23,06
10	188	13,44	18,82	**16,71**	30,29	57,98	**68,39**	49,15	53,84
Average:	166,20	12,66	14,38	**9,89**	16,80	38,65	**39,44**	29,09	31,74

* Phrases were 2-5 words long

Therefore, the results show that DT after appropriate domain adaptation can give comparable or even better translation results than Google and Bing. In addition, some interesting correlations were found between internal metrics of the DT and the BLEU scores, as well as the correlations among the translation results of different translators. The correlation between used translation phrases and BLEU score for untrained DT was relatively weak (0.29), but strong after the domain adaptation (0,87).

Correlation between Google and Bing was very strong (0,97), while the correlation between DT and these two translators was higher for the untrained DT (0,83 for Google and 0,81 for Bing) than for the trained DT (0,65 for Google and 0,58 for Bing).

6 CONCLUSION

This paper presents the concept and one possible implementation of the Domain Translator (DT). The intention is to improve the presentation of Croatian web sites in foreign languages, mostly in English, by preserving the local heritage like local names, abbreviations, capitalization of words etc. Namely, there are many problems noticed in the usage of Google Translate in a form of browser extension. At first, some of the web contents (particularly inside the foi.unizg.hr domain) were not translated at all. In addition, Google Translate has problems with some special kinds of text, like filenames, text covered by a picture, access to some web domains, special names, abbreviations etc.

The main assumptions to make it possible to build a useful DT are that the language domain of some web site is much narrower in relation to the general language domain, that the translation is in only one direction, and awareness of the local context (e.g. images that should be removed to show the text, special ways of writing like using "_" between the words etc.).

Developed prototype was tested by using BLEU machine translation evaluation method, as well as by using internal metrics of the translation system, like usage of different length phrases. The results show that Google Translate and Bing Translator give better results for the clear text expressed in a BLUE metrics, but only for the lexical domain for which DT was not trained before.

After the domain adaptation, the results become comparable, somewhere even better for the DT, and encouraging for further development of the proposed concept.

In the future work, it is planned to improve the translation model by including more elements of our SCT generator model [15] and Autogenerator [17] that are used in the field of code generation.

Acknowledgements

We would like to thank prof. Višnja Fara for her work on the online dictionaries and the help in the linguistic part.

7 REFERENCES

[1] Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*(48-54). Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073462

[2] Chand, S. (2016). Empirical survey of machine translation tools. In *Research in Computational Intelligence and Communication Networks (ICRCICN), 2016 Second International Conference on* (181-185). IEEE. https://doi.org/10.1109/ICRCICN.2016.7813653

[3] Jiménez-Crespo, M. A. (2016). What is (not) web localization in translation studies? *The Journal of Internationalization and Localization*, *3*(1), 38-60. https://doi.org/10.1075/jial.3.1.03jim

[4] Costales, A. F. (2012). The internationalization of institutional websites: The case of universities in the European Union. *Translation research projects*, 4, 51-60.

[5] Chen, X., Acosta, S., & Barry, A. E. (2016). Evaluating the accuracy of Google translate for diabetes education material. *JMIR Diabetes*, *1*(1), e3. https://doi.org/10.2196/diabetes.5848

[6] Pecina, P., Toral, A., Way, A., Papavassiliou, V., Prokopidis, P., & Giagkou, M. (2011). Towards using web-crawled data for domain adaptation in statistical machine translation.

[7] Daumé III, H. & Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*(407-412). Association for Computational Linguistics.

[8] Choi, E., Horvat, M., May, J., Knight, K., & Marcu, D. (2016). Extracting Structured Scholarly Information from the Machine Translation Literature. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. European Language Resources Association.

[9] Koehn, P. & Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation* (224-227). Association for Computational Linguistics. https://doi.org/10.3115/1626355.1626388

[10] Bertoldi, N., Cettolo, M., & Federico, M. (2013). Cache-based on line adaptation for machine translation enhanced computer assisted translation. In *MT-Summit* (35-42).

[11] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

[12] Lin, C. Y. & Och, F. J. (2004). Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 501). Association for Computational Linguistics. https://doi.org/10.3115/1220355.1220427

[13] Lavie, A. & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, *23*(2-3), 105-115. https://doi.org/10.1007/s10590-009-9059-4

[14] Han, L. (2017). LEPOR: An Augmented Machine Translation Evaluation Metric. *arXivpreprint arXiv: 1703.08748*.

[15] Radošević, D. & Magdalenić, I. (2011). Source code generator based on dynamic frames. *Journal of Information and Organizational Sciences*, *35*(1), 73-91.

[16] Radošević, D., Magdalenić, I., Andročec, D., Bernik, A., & Kaniški, M. (2017). A machine translation model inspired by code generation. In *28th Central European Conference on Information and Intelligent Systems (CECIIS 2017)*.

[17] Magdalenić, I., Radošević, D., & Orehovački, T. (2013). Autogenerator: Generation and execution of programming

code on demand. *Expert Systems with Applications*, *40*(8), 2845-2857. https://doi.org/10.1016/j.eswa.2012.12.003

Contact information:

Danijel RADOŠEVIĆ, PhD, Full Professor
University of Zagreb, Faculty of Organization and Informatics,
Pavlinska 2, 42000 Varazdin, Croatia
danijel.radosevic@foi.hr

Andrija BERNIK, PhD
University North,
104. brigade 1, 42000 Varazdin, Croatia
andrija.bernik@unin.hr

Nikola MRVAC, PhD, Full professor
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10000 Zagreb, Croatia
nikola.mrvac@grf.hr

[1] Google Translate in a form of browser extension is available at https://chrome.google.com/webstore/category/extensions
[2] https://www.bing.com/translator
[3] Croatian-English and English-Croatian online dictionary interface is available at http://gpml.foi.hr/dictionary/
[4] Croatian-English and English-Croatian translator interface is available at http://gpml.foi.hr/translator/
[5] The prototypes for Chrome and Firefox are available at http://gpml.foi.hr/laboratory/index.php?id=translators
[6] http://tiskarstvo.net/
[7]The tests have been conducted using Tilde Interactive BLEU score evaluator, https://www.letsmt.eu/Bleu.aspx