Towards translation of educational resources using GIZA++

Ivan Obradović, Dalibor Vorkapić, Ranka Stanković, Nikola Vulović, Miladin Kotorčević

The Seventh International Conference on eLearning (eLearning-2016),
29 – 30 September 2016, Belgrade, Serbia

TOWARDS TRANSLATION OF EDUCATIONAL RESOURCES USING GIZA++

IVAN OBRADOVIĆ
University of Belgrade, Faculty of Mining and Geology, ivan.obradovic@rgf.bg.ac.rs

DALIBOR VORKAPIĆ
University of Belgrade, Faculty of Mining and Geology, dalibor.vorkapic@rgf.bg.ac.rs

RANKA STANKOVIĆ
University of Belgrade, Faculty of Mining and Geology, ranka.stankovic@rgf.bg.ac.rs

NIKOLA VULOVIĆ
University of Belgrade, Faculty of Mining and Geology, nikola.vulovic@rgf.bg.ac.rs

MILADIN KOTORČEVIĆ
University of Belgrade, Faculty of Mining and Geology, miladin.kotorcevic@rgf.bg.ac.rs

Abstract: *E-learning courses are becoming progressively popular. Thanks to the Internet and new technologies, education has never been more available to everyone. The main obstacle to studying new subjects is often the language, given the number of different languages in which educational resources are published as well as the corresponding cultural context. That is why the tools for translation of e-learning courses and translation support are nowadays one of the most important topics in this area. E-learning course translation is a very special service that requires specific subject-matter expertise and high technical skills from everyone involved. This paper presents the current state of research in course translation. The translation of electronic courses is an ongoing activity at the University of Belgrade Faculty of Mining and Geology and the first results using the GIZA++ tool for training statistical translation models will be presented. The paper also describes the translation memory in the form of parallel sentences or phrases required by GIZA++ for the learning algorithm.*

Keywords: *E-Learning, GIZA++, translation memory*

1. INTRODUCTION

Massive Open Online Courses (MOOCs) are becoming very popular recently. More than 200 universities around the world are involved in their creation, with the involvement of more than 500 Universities, more than 4200 courses on offer and around 35 million users being actively registered [1]. MOOCs have major contribution to lifelong education. They are a tool to help identify and fill the gap that exists in the digital skills of workers across Europe. The language barrier is the biggest obstacle that stands in the way of the development of online courses as the majority of such courses are offered in English. Thus a growing need for translating MOOC content. The solutions provided so far have been fragmentary, human-based, and implemented off-line by the majority of course providers. [2]

TraMOOC (Translation for Massive Open Online Courses) is a Horizon 2020 collaborative project aiming at providing reliable machine Translation for MOOCs. The main result of the project will be an online translation platform, which will utilize a wide set of linguistic infrastructure tools and resources in order to provide accurate and coherent translation to its end users. [3] TraMOOC constitutes a solution to online course content translation that aims at eleven target languages, is automatic – i.e. it is based on statistical machine translation (SMT) techniques – and is therefore easily extendable to other languages, adaptable to various types of educational content genre, independent of course domain, and designed to produce translations online via its integration in the use-case platforms.

TraMOOC translation includes all types of text genre included in MOOCs: assignments, tests, presentations, lecture subtitles, forum text, from English into eleven languages, i.e. German, Italian, Portuguese, Greek, Dutch, Czech, Bulgarian, Croatian, Polish, Russian, Chinese, which constitute strong use-cases, many of them hard to translate into and with relatively weak machine translation (MT) support. Phrase-based and syntax-based SMT models are developed to address language diversity and support the language independent nature of the methodology. For high-quality MT and to add value to existing infrastructure, extensive advanced bootstrapping of new resources is performed, while at the same time innovative multi-modal automatic and human evaluation schemata are applied. For human evaluation, an innovative, strict-access control, time- and cost-efficient crowdsourcing set-up is used, while translation experts,

domain experts and end users are also involved. Results are combined into a feedback vector and used to refine parallel data and retrain translation models towards a more accurate second-phase translation output. The project results will be showcased and tested on the Iversity [4] MOOC platform and on the VideoLectures.NET digital video lecture library. The translation engine employed in TraMOOC is Moses [5], the most widely used SMT toolkit available in academia as well as in commercial environments, mainly due to its flexibility, modularity, open-source licence, and competitive translation results. [2]

2. RELATED WORK

Coursera, a leading MOOC provider [5], has announced a partnership with ten top organizations from eight countries to translate complete course lectures across multiple disciplines for students around the world, for free. There are three basic approaches to course translation: human translators (traditional), translation using CAT (Computer Aided Translation) tools and machine translation. It is, of course, possible to combine some or all of them.

Leading translation companies, philanthropic organizations, mobile carriers, nonprofits, corporations and universities have joined forces in this partnership. The organizations started with translating selected courses into: Russian, Portuguese, Turkish, Japanese, Ukrainian, Kazakh, and Arabic, as the most popular languages for Coursera students. General approach is that each Coursera Global Translation Partner starts with translation od 3-5 selected courses.

For multilingual support Coursera [7] uses Transifex [8] continuous localization platform, as a cloud-based tool that hosts Coursera's translatable content and allows partner organizations and individuals to easily contribute course translations from anywhere. At this moment, support for user interface in five languages is available, but the long-term goal is to have platform localized to global audiences.

Students can log in to Coursera and check options for type of language support with information about translation offerings in the coming months. In this phase, course lectures are translated via subtitles while all other course material, including quizzes and assignments, are in the source language.

Coursera welcomes at the moment 145 partners across 28 countries offering over 2,000 courses. By joining forces with top organizations globally to produce fully translated course lectures, Coursera with translation partners is producing high-quality education accessible to anyone, anywhere – regardless of what language they speak.

3. TRANSLATION OF EDUCATIONAL RESOURCES - CURRENT APPROACHES

For translation of eLearning resources both language translation, and eLearning skills are necessary. The translation team needs knowledge of various software platforms and custom formats. Data exchange with different platforms can be technically challenging, since there is no common format and schema. [9]

The understanding of adaptation of languages for text and speech is also required, as many eLearning courses have an audio component. Recommendation is that the translation team works closely with the course authors, in order to fine-tune the translation. In order to keep the original style of the course, it is recommended to recruit translators that better capture the essence. Vocabulary, word choices, general style must remain similar in all translated versions. Reference materials from course authors are also helpful, as well as previous translations, glossaries, style guides, translation memory files.

With an iterative and agile approach to translation it is possible to adapt as problems arise or courses are changed during translation. Proper planning is essential. General guidelines and workflow for eLearning course translation should start with an initial representative segment as a pilot, in order to evaluate the quality of the translation and formulate suggestions for the improvement of the rest of the translation. The translation needs several reviews before publishing or preparation for voice recording. [10]

A Computer Aided Translation (CAT) Tool is based on collection of aligned sentence pairs in the form of Translation Memory, which facilitates and speeds up the translator's work. Main key functions of a CAT tool that speed up and improve translation are: [11]

- A CAT tool segments the source text in segments, usually sentences, and uses them to filter and preview the matching segments in a suitable way, usually in specific box, next to or below the source text.
- The source text and translation of each segment are saved together, processed and presented as a translation unit (TU).
- Translation memory (TM) is a database in which CAT tool saves the translation units, so that they can be re-used for later translations. If there are segments that do not match 100 %, the search functions of CAT tools can find them through special "fuzzy search" features.
- CAT tool has support for terminology look-up, display and insertion of the search results into the text being translated.

4. ENVIRONMENT FOR TEXT ALIGNMENT

Preliminary phase for the text alignment (parallelization) consists of XML document (eXtensible Markup Language) preparation according to TEI (Text Encoding Initiative) consortium guidelines. In practice, this step is comprised of marking the divisions, titles, paragraphs and segments using text or XML editing software with support for DTD (Document Type Definition) scheme validation and well-formedness check. This part can be automated using finite-state transducers, but manual intervention is still necessary.

The next key step is aligning the text – parallelization. The aim is to determine for each text segment which segment of the translated text correlates with the segment in the original text. The task is thus to establish the connection between originals and their translations. In this process, segments are paired that sometimes represent whole sentences and sometimes just their parts, depending on the complexity of the sentence or the translation itself.

Parallelization can be performed using ACIDE software [11]. As an end result, three documents are created with

extension *_f_id*, *_s_id* and *_fs*. The first two represent the original documents, whose *seg* labels are tagged with the attribute *id="nx"*, where *x* represents the serial number of the segment. Examples are shown in the image 1.

```
<div>
  <p>
    <seg id="n15">1. An Ocean of Digital Words</seg>
    <seg id="n16">A society of information offers almost
    a limitless amount of information to everyone.</seg>
    <seg id="n17">Without the usage of intelligent,
    efficient applications for information extraction,
    which are based on highly advanced techniques and
    methods, one can benefit only from the smallest part
    of potential offered by new technology (Piskorski
    1999).</seg>
    <seg id="n18">If we define information as a result of
    collecting, processing, manipulating and organizing
    data in order to present new knowledge to the
    recipient, than it can be said that a piece of data
```

```
<div>
  <p>
    <seg id="n15">1. Okean digitalnih reči</seg>
    <seg id="n16">Informatičko društvo stavlja na
    raspolaganje gotovo neograničenu količinu informacija
    svakom pojedincu.</seg>
    <seg id="n17">Bez upotrebe inteligentnih, efikasnih
    aplikacija za ekstrakciju informacija koje se zasnivaju
    na vrlo naprednim tehnikama i metodama, pojedinac nije u
    mogućnosti da iskoristi ni delić nesagledivog
    potencijala koji nude nove tehnologije (Piskorski 1999).
    </seg>
    <seg id="n18">Ako informaciju definišemo kao rezultat
    sakupljanja, obrade, manipulacije i organizovanja
    podataka sa ciljem da se primaocu predstavi novo znanje
```

Image 1: examples of segmented XML texts: English left and Serbian right

Document with the extension *_fs* contains the information about paired segments. The method used in the alignment is based on the number of characters (length of the segment). This approach is very successful (on the average as much as 96% correctly paired segments). Mistakes in pairing, however, must be corrected manually, which is done through the Concordancier software. [13]

The next step is the production of a TMX document [14]. The document consists of <header>, <body>, <p> (paragraph), <tu> (Translation Unit) and <tuv> (Translation unit variant) elements. [15] Metadata code (element <prop>) is attached to each aligned sentence (element <tu>) in order to establish a direct relation to metadata and the original (pdf, edX, docx,…) form of resource document, article, course or other resource. Image 2 presents one part from the TMX document with ID: 1.2010.1.4.

From aligned TMX documents is easy to produce parallel text form for tools like Giza++, or JSON format suitable for web services and Mongo and other NoSQL databases.

```
<tu>
  <prop type="Domain">Gucul-Milojević, 2010, vol. XI:1, ID: 1.2010.1.4</prop>
  <tuv xml:lang="en" creationid="n15 " creationdate="20110513T151548Z">
    <seg>1. An Ocean of Digital Words </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n15 " creationdate="20110513T151548Z">
    <seg>1. Okean digitalnih reči </seg>
  </tuv>
</tu>
<tu>
  <prop type="Domain">Gucul-Milojević, 2010, vol. XI:1, ID: 1.2010.1.4</prop>
  <tuv xml:lang="en" creationid="n16 " creationdate="20110513T151548Z">
    <seg>A society of information offers almost a limitless amount of information to everyone. </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n16 " creationdate="20110513T151548Z">
    <seg>Informatičko društvo stavlja na raspolaganje gotovo neograničenu količinu informacija
    svakom pojedincu. </seg>
  </tuv>
</tu>
<tu>
  <prop type="Domain">Gucul-Milojević, 2010, vol. XI:1, ID: 1.2010.1.4</prop>
  <tuv xml:lang="en" creationid="n17 " creationdate="20110513T151548Z">
    <seg>Without the usage of intelligent, efficient applications for information extraction, which
    are based on highly advanced techniques and methods, one can benefit only from the smallest
    part of potential offered by new technology (Piskorski 1999). </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n17 " creationdate="20110513T151548Z">
    <seg>Bez upotrebe inteligentnih, efikasnih aplikacija za ekstrakciju informacija koje se
    zasnivaju na vrlo naprednim tehnikama i metodama, pojedinac nije u mogućnosti da iskoristi ni
    delić nesagledivog potencijala koji nude nove tehnologije (Piskorski 1999). </seg>
  </tuv>
</tu>
```

Image 2: An example excerpt from a TMX document

5. TOWARDS MACHINE TRANSLATION FOR SERBIAN

Moses is a statistical machine translation system written in C++ with library that enables usage of Moses in the JavaScript language. Loading multiple translation systems into the same node process is provided. This means the same process can hold multiple distinctly different translation models (e.g., Chinese to English in IT and English to French in medicine) at the same time, and be able to use those models to translate user-given sentences or paragraphs on-demand. [5]

GIZA++ is an extension of the program GIZA which was developed by the Statistical Machine Translation team during the summer workshop in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU). GIZA++ includes a lot of

additional features. The extensions of GIZA++ were designed and written by Franz Josef Och.

GIZA ++ is installed on the Faculty of Mining and Geology as part of Moses, which is hosted as a virtual machine. It uses the Linux operating system. GIZA is quite a demanding tool, and it therefore requires extra resources. Its execution process requires a larger amount of RAM, which in our case was 16GB.

Corpus Preparation

For our research we used five text collections, three of them being scientific journals and two resources produced within international projects. Total number of documents is 299 in English and the same number in Serbian, while the total of aligned sentences is 67,206.

Haddow et al. [16] give a general MT system overview with details on the training pipeline and decoder configuration using Moses toolkit. [5] In this research we followed their approach, albeit with available resources for Serbian.

To prepare the data for the translation system, we had to perform the following steps:

- **tokenisation:** This means that spaces have to be inserted between (e.g.) words and punctuation.
- **truecasing:** The initial words in each sentence are converted to their most probable casing. This helps reduce data sparsity.
- **cleaning:** Long sentences and empty sentences are removed as they can cause problems with the training pipeline, and obviously mis-aligned sentences are also removed.

Tokenisation launch was initialized by the following sequence:

```
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en \
   < ~/corpus/training/edX.en    \
   > ~/corpus/edX.tok.en

 ~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l sr \
   < ~/corpus/training/edX.sr    \
   > ~/corpus/edX.tok.sr
```

The truecaser first requires training, in order to extract some statistics about the text:

```
~/mosesdecoder/scripts/recaser/train-truecaser.perl \
   --model ~/corpus/truecase-model.en --corpus    \
   ~/corpus/edX.tok.en

 ~/mosesdecoder/scripts/recaser/train-truecaser.perl \
   --model ~/corpus/truecase-model.sr --corpus    \
   ~/corpus/edX.tok.sr
```

Finally cleaning and limiting the length to 80 was performed:

```
~/mosesdecoder/scripts/recaser/truecase.perl \
  --model ~/corpus/truecase-model.en        \
  < ~/corpus/edX.tok.en \
  > ~/corpus/edX.true.en

 ~/mosesdecoder/scripts/recaser/truecase.perl \
  --model ~/corpus/truecase-model.sr        \
  < ~/corpus/edX.tok.sr \
  > ~/corpus/edX.true.sr

~/mosesdecoder/scripts/training/clean-corpus-n.perl \
   ~/corpus/edX.true sr en \
   ~/corpus/edX.clean 1 80
```

Language Model Training

A language model (LM) is used to ensure fluent output, built with the target language, in our case English. Following script creates *lm* folder, positions in it and finally execute command that will build an 3-gram language model.

```
mkdir ~/lm
 cd ~/lm
~/mosesdecoder/bin/lmplz -o 3 <~/corpus/edX.true.en >
edX.arpa.en
~/mosesdecoder/bin/build_binary \
  edX.arpa.en \
  edX.blm.en
```

Finally, we came to the main event - training the translation model. To do this, we ran word-alignment (using GIZA++), phrase extraction and scoring, created lexicalised reordering tables and Moses configuration file, all with a single command. Before starting the command, we created a working folder in which results were stored.

```
nohup nice ~/mosesdecoder/scripts/training/train-
model.perl -root-dir trainedX \
-corpus ~/corpus/edX.clean                    \
-f sr -e en -alignment grow-diag-final-and -reordering
msd-bidirectional-fe \
-lm 0:3:$HOME/lm/edX.blm.en:8                 \
-external-bin-dir ~/mosesdecoder/tools >& edX.out &
```

After starting the command it takes some time to get to the results. In our case, it took about 90 minutes. The result is a file that contains paired Serbian and English words with a factor of accuracy for translation from Serbian to English and from English into Serbian. In the background of Image 3 GIZA++ program the output result of machine translation is shown as a "phrase table", which is analysed in a custom made C# application, filtered, sorted and exported as excel file (Image 3, front). A "phrase table"[1] is a statistical description of a parallel corpus of source-target language sentence pairs, created during the training process.

The frequencies of n-grams in a source language text that co-occur with n-grams in a parallel target language text represent the probability that those source-target paired n-grams will occur again in other texts similar to the parallel corpus. This can be perceived as a kind of dictionary between the source and target languages. Phrase tables and reordering tables are translation model components. Depending on parameters chosen for training process, different phrase translation scores[2] are computed, but the main are:

- inverse phrase translation probability φ(sr|e)
- inverse lexical weighting lex(sr|e)

[1] http://www.statmt.org/moses/glossary/SMT_glossary.html

[2] http://www.statmt.org/moses/?n=FactoredTraining.ScorePhrases

- direct phrase translation probability φ(en|sr
- direct lexical weighting lex(en|sr)
- phrase penalty (always exp(1) = 2.718)

To estimate the phrase translation probability φ(en|sr) we fist sort the extract file is sorted to ensure that all English phrase translations for a Serbian phrase are next to each other in the file. In next step, one Serbian phrase at a time, collect counts and compute φ(en|sr) for that Serbian phrase sr. To estimate φ(sr|en), the inverted file is sorted, and then φ(sr|en) is estimated for an English phrase at a time.

Additional phrase translation scoring parameters can be produced in output: lexical weighting (direct and indirect), word penalty, phrase penalty, Lexical weighting features estimate the probability of a phrase pair or translation rule word-by-word. The word penalty ensures that the translations do not get too long or too short. The phrase penalty feature is a global feature that counts the number of used phrases for all phrase tables cumulatively.

Apart from machine translation, aligned words and multiword expressions can be used for searching and exploring translation variants in large parallel corpora [17]. Volk et al. argue that automatic word alignment allows for major innovations in searching parallel corpora. Some online query systems already employ word alignment for sorting translation variants, but they describe the system for efficiently searching large parallel corpora with a powerful query language [18].

In [19] another approach for extraction of semantically related word pairs, ideally translational equivalents, is presented, from aligned texts in SELFEH, a Serbian-English corpus of texts related to education, finance, health and law, aligned at the sentence level within Intera project. The corpus was lemmatized and the method applied on lemmas of word forms from the corpus, by extracting candidate translational equivalents through a ranking based on lemma frequencies.

Similar experiments with the alignment on the word level were performed also on the Intera English/Serbian corpus [19, 20] with and without lemmatisation and PoS tagging. Authors report the most suitable measure:

$$ranky(x) = (C(x|y) / \Sigma i \epsilon V\ C(i|y)) * (C(x|y) / C(x))$$

where V is the set of word forms i of a target language for which C(i|y) > 0, C(x) is the frequency of occurrences of a word x in the target language, while C(x|y) represents the frequency of a word x from the target language occurring in the same segment with the chosen word y from the source language. Summing is done for all words of the source language. This formula represents a variant of the geometric average.

The SELFEH corpus is part of Biblisha digital library and is used in this research, and a comparison of results is in progress.

Machine translation research using Giza++ and is usage for eLearning material is in its initial phase, but it is clear that the most effective way of translating is obtained using all three methods of translation (Computer Aided Translation, machine translation and human translation).

```
autonomna pokrajina , odnosno jedinica ||| , autonomous province , or unit ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina , odnosno ||| , autonomous province , or ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina , ||| , an autonomous province , ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina , ||| , autonomous province , ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina ili jedinica lokalne samouprave ||| , Autonomous Province or Unit of Municipal ||| 0.6 0.2 0.2 0.2 0.2 0.6
autonomna pokrajina ili jedinica lokalne ||| , Autonomous Province or Unit of ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina ili jedinica ||| , autonomous province , or unit ||| 0.714286 0.142857 0.142857 0.714286 0.142857 0.142857
autonomna pokrajina ili jedinica ||| , autonomous province or a unit ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina ili jedinica ||| , autonomous province or unit ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina ili ||| , Autonomous Province or ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina ili ||| , autonomous province , or ||| 0.714286 0.142857 0.142857 0.714286 0.142857 0.142857
autonomna pokrajina ili ||| , autonomous province or ||| 0.714286 0.142857 0.142857 0.714286 0.142857 0.142857
autonomna pokrajina ||| , Autonomous Province ||| 0.6 0.2 0.2 0.6 0.2 0.2
autonomna pokrajina ||| , an autonomous pro
autonomna pokrajina ||| , autonomous provi
autonomna pokrajina ||| , autonomous provi
autonomna ||| , Autonomous ||| 0.6 0.2 0.2
autonomna ||| , an autonomous ||| 0.6 0.2
autonomna ||| , autonomous ||| 0.846154 0.
bez obzira na pol , rasu ||| , regardless
bez obzira na pol , ||| , regardless of th
bez obzira na pol ||| , regardless of thei
bez prava odlučivanja . ||| , without deci
bez prava odlučivanja ; ||| , without deci
bez prava odlučivanja ||| , without decisi
da počne sa radom i da ||| , commence oper
da počne sa radom i ||| , commence operati
da počne sa radom ||| , commence operation
dalje obrazovanje i samostalno učenje , ||
dalje obrazovanje i samostalno učenje ||| 
dalje obrazovanje i ||| , further educatio
dalje obrazovanje ||| , further education
dalje ||| , further ||| 0.6 0.2 0.2 0.6 0.2
davanje podrške razvojnom planiranju i |||
davanje podrške razvojnom planiranju ||| ,
davanje podrške razvojnom ||| , providing
davanje podrške ||| , providing support ||
davanje ||| , providing ||| 0.6 0.2 0.2 0.
direktor : ||| , the principal shall : |||
direktor ||| , the principal shall ||| 0.2
direktor ||| , the principal ||| 0.2 0.2 0
```

Word_sr	Word_en	P_sr_en	P_en_sr	P_su
skladu	accordance	0,977011	0,977011	1,9540
, u skladu	, in accordance	0,904762	0,968254	1,8730
1. ovog	1 of this	0,935484	0,935484	1,8709
centar	the Centre	0,935484	0,935484	1,8709
daljem tekstu	further text	0,935484	0,935484	1,8709
daljem	further	0,935484	0,935484	1,8709
tekstu	text	0,935484	0,935484	1,8709
u daljem tekstu	in further text	0,935484	0,935484	1,8709
u daljem	in further	0,935484	0,935484	1,8709
, u	, in	0,913043	0,942029	1,8550
zakonom .	law .	0,925926	0,925926	1,8518
iz stava	from paragraph	0,886792	0,962264	1,8490
Izuzetno od stava	Exceptionally from paragraph	0,818182	0,818182	1,6363
Izuzetno od	Exceptionally from	0,818182	0,818182	1,6363
Republika , autonomna	the Republic , autonomous	0,818182	0,818182	1,6363
aktom o osnivanju	the establishment act	0,818182	0,818182	1,6363
skladu sa ovim i posebnim zakonom	accordance with this law and separate laws	0,818182	0,818182	1,6363
predmeta	subjects	0,866667	0,733333	1,6
za naknadu	employee about the	0,2	0,6	0,8
za naknadu	the employee about the	0,2	0,6	0,8
za naknadu	the	0,2	0,6	0,8

Image 3: The result of machine translation using GIZA++ tool

6. CONCLUSION

Massive Open Online Courses (MOOCs) are becoming very popular. Since they are mostly in English, there is a need to translate them into other languages. GIZZA++ is the right tool for that, but it needs a parallel corpus of significant size, that depends from language and domain. First a DTD scheme neds to be used to validate and check well-formedness. Then it is necessary to pair the text – parallelization. The aim is to determine which element of the text correlates with the translation of the element in the corresponding textNext, the following 3 steps are taken: tokenisation, truecasing and cleaning. At the end, the language model (LM) is used to ensure fluent output, and is thus built with the target language.

The presented method yielded promising results, but bigger corpus is needed for better results. Therefore, great efforts are being made for additional text alignment and augmentation of Biblisha library. The detailed evaluation will be performed when we reach at least 100000 sentence pairs. Ur aim is to publish SMT based web service (API) and integrate it with eLearning systems that we use: Moodle and edX,

REFERENCES

[1] Class Central • Discover Free Online Courses & MOOCs https://www.class-central.com/, accessed June, 10th 2016

[2] V. Kordoni, A. Van Den Bosch, K. Kermanidis, V. Sosoni, K. Cholakov, I. Hendrickx and M. Huck, "Enhancing Access to Online Education: Quality Machine Translation of MOOC Content", Proceedings of the 10th edition of the Language Resources and Evaluation Conference, , Portorož, Slovenia, May 2016

[3] TraMOOC H2020 Project, http://tramooc.eu/, accessed August, 1st 2016.

[4] Open Online Courses - Study Anywhere https://iversity.org/, accessed June, 10th 2016

[5] Moses - a statistical machine translation system http://www.statmt.org/moses/, accessed June, 10th 2016

[6] Coursera Blog, Coursera Partnering with Top Global Organizations Supporting Translation Around the World, https://blog.coursera.org/post/50452652317/coursera-partnering, accessed June, 10th 2016

[7] Free Online Courses from Top Universities, https://www.coursera.org/ , accessed June, 10th 2016

[8] Localization Platform for Translating Digital Content https://www.transifex.com/, accessed June, 10th 2016

[9]. eLearning Industry, 10 Easy Steps For Successful eLearning Course Translation https://elearningindustry.com/10-easy-steps-for-successful-elearning-course-translation, accessed June, 10th 2016

[10] eLearning Industry, eLearning Course Translation Workflow https://elearningindustry.com/elearning-course-translation-workflow, accessed June, 10th 2016

[11] Meta Texis, http://www.metatexis.com/cat.htm, accessed June, 10th 2016

[12] I.Obradović, R.Stanković, and M. Utvić, An Integrated Environment for Development of Parallel Corpora (in Serbian). In: Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen (pp. 563-578), B. Tošović (Ed.). Berlin: LitVerlag 2008

[13] Digital library for parallel text Biblisha Online user manual, http://jerteh.rs/biblisha/Documentation.aspx, accessed June, 10th 2016

[14]. TMX 1.4b Specification. (2005). http://www.gala-global.org/oscarStandards/tmx/tmx14b.html, accessed June, 10th 2016

[15]. R. Stanković, C. Krstev, I. Obradović, A. Trtovac and M. Utvić, "A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals", Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, , eds. Nicoletta Calzolari et al., ISBN 978-2-9517408-7-7 23--25 May 2012

[16] B. Haddow, M. Huck, A. Birch, N. Bogoychev and P. Koehn. "The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal, September 2015

[17] G. Johannes, S. Clematide, and M. Volk. "Efficient Exploration of Translation Variants in Large Multiparallel Corpora Using a Relational Database." 4 th Workshop on Challenges in the Management of Large Corpora Workshop Programme. 2016

[18] M. Volk, J. Graën, and E. Callegaro. Innovations in Parallel Corpus Search Tools. In LREC (pp. 3172-3178), 2014

[19] I. Obradović, "A Method for Extracting Translational Equivalents from Aligned Text", "Methods and applications of quantitative linguistics": selected papers of the 8th International Conference on quantitative linguistics (QUALICO) in Belgrade, Serbia, April 26-29, 2012. University of Belgrade, 2013.

[20] D. Vitas and C. Krstev. "Construction and Exploitation of X-Serbian Bitexts". In Cristina Vertan and Walther v. Hahn (eds.) Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation, pp. 207-227, Cambridge Scholars Publishing,. ISBN (13) 978-1-4438-3878-8, 2012.

[21] A. Obuljen, Kvantitativna metoda za poravnanje dvojezičnog korpusa. Internal report, Faculty of Mathematics, University of Belgrade, Serbia, 2009.