Proceedings of the Federated Conference on Computer Science and Information Systems pp. 517–525
DOI: 10.15439/2016F304
ACSIS, Vol. 8. ISSN 2300-5963

# Exploration for Polish-* bi-lingual translation equivalents from comparable and quasi-comparable corpora

Krzysztof Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008 Warszawa, Poland
Email: kwolk@pja.edu.pl

Krzysztof Marasek
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008 Warszawa, Poland
Email: kmarasek@pja.edu.pl

Agnieszka Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008 Warszawa, Poland
Email: awolk@pja.edu.pl

***Abstract*—In contemporary world, translation becomes a critical need of the time. Parallel dictionaries have now become a most accessible source by humans, but confines are there as they do not offer good quality translation function, because of neologisms and words that are out of vocabulary. To overcome this problem in the usage of statistical translation systems is becoming more and more important in maintaining the eminence and quantity of the training data. But due to the limitations in these systems they have very limited availability for few languages and very limited narrow text areas. The purpose of this research is to bring calculation time up gradation via GPU acceleration, tuning script introduction and the enhancement and improvements in the methodologies of the contemporary comparable corpora mining through re-implementation of analogous algorithms through Needleman-Wunch algorithm. Experiments have been conducted on multiple language data which were extracted on numerous domains from Wikipedia. For the sake of Wikipedia, multiple cross-lingual contrasts and comparison were established. Optimistic impact on the both quantity and quality of mined data was observed due to such changes and adaptation. The solution is language independent and highly practical especially for under-resourced languages.**

## I. Introduction

THE purpose of the research is to organize the language models and parallel and comparable corpora. This process advances the quality of SMT through riddling of parallel corpora and it also works through extraction of supplementary parallel data via the resulting corpora. In order to improve the language spring of the SMT systems, alteration measures and interpolation methods are applied to the obtainable prepared data. For this, various experiments were led by using wide domain (TED presentations on variety of topics).

SMT system's assessment was functioned on random samples of analogous data by utilizing automated algorithms (BLEU metric) in order to assess the possible usage and standard of the SMT systems' output. In addition, human evaluation was conducted in order to measure the impact of newly obtained corpora on translation error reduction. [1]

While experiments are discussed, the utilization of the software Moses Statistical Machine Translation Toolkit [2] is done. Further, the symmetrisation is done using Berkeley Aligner [3] and translation models training is done using multi-threaded application the GIZA++ tool. Only from single language data base, the statistical models are shaped well by utilizing SRILM (SRI Language Modelling toolkit). Furthermore, the data from external domain is adapted. In the situation of parallel modelling, in-domain data collection is found using, Moore-Lewis Filtering [4] while single-language models are linearly interpolated [5].

Finally, methods recommended in the Yalign [6] parallel data mining tool are upgraded and critically evaluated. By using the tool in a multi-threaded way and by employing graphics processing units (GPUs), its speed was also amplified. Furthermore, by utilizing Needleman-Wunch [7] algorithm and by developing a tuning script that is used to regulate mining parameters to fix domain supplies, its quality is improved as well.

In the tests, the resultant SMT systems out-performed the baseline systems in terms of BLEU metric and error reduction.

## II. Corpora Types

A corpus includes a large collection of texts stored up on a computer. These text compilations are known as corpora. Usually in linguistic fields, parallel corpus as a term is used with the reference to texts which are the source of translation of each other. In order to deal with the statistical machine translation, we are significantly considerate about parallel corpora. These are paired with text through another language. For the preparation of parallel texts for the call for statistical machine translation, it may require removing the text from HTML, web crawling and sentence structure [5] and performing document alignment.

Two major kinds of parallel corpora exist in two different languages. One is comparable corpus in which common texts are present and their content is also the same. Polish and English newspaper's articles are best example of comparable corpora. The second one is the translation corpus, in this type of corpus the text of first language (e) is the translation of text in second language (f). It is significant to recognize that the word "comparable corpora" signifies the texts in two different languages, they are common in the content but they are not common translations of one another. [5]

In order to assess a parallel text, pattern arrangement is used which mentions common texting sections (approximately, sentences), that is a significant requirement for examination.

Within first and second language machine translation algorithms for translation are often trained, using common fragments. This includes a first and second language corpus that is an element for element translation of the first language corpus. In such kind of training it may involve huge training sets which can be removed from huge corpora of common sources, like databases of the news articles written in the first and second languages while describing common events [5]. Due to this complicatedness, it is problematic to obtain high quality parallel data, significantly for uncommon languages. Comparable corpora are the key to the solution of problem of absence of data for rare language pairs that are under-resourced languages and other subject domains. It is easier and conceivable to use comparable corpora to achieve straight knowledge for the purposes of translations. This data is considered to be a precious foundation of knowledge for other information dependent and cross-lingual tasks. This data is not as rare as parallel, even for Polish-* languages. On the flip side, single language data is accessible in huge qualities [5].

While concluding, there are four key corpora kinds which are notable. Parallel corpora that are also very uncommon, can be explained as corpora which have the translations of the common file into two and more than two languages. For that such kind of data it is needed for it to be aligned, at the level of sentence as a minimum. Noisy-parallel corpus that contains bilingual structure sentences that are not excellently arranged and they can also have translations with bad quality. Yet, in most cases, bilingual translations of a precise document are present in it. A comparable corpus is structured from unstructured sentences and with un-translated bilingual documents, but the text or the documents need to be about the same topic. Seemingly quasi-comparable corpus also has very non-parallel bilingual and very mixed documents, they may or may not be structured according to the topic [8].

## III. State of the art

If comparable corpora are concerned, numerous efforts (as for Wikipedia) have been observed in order to evaluate parallel data samples. Two core methods for building comparable corpora can be easily separated or illustrated. The most common method is founded on the notion of recovery of cross-lingual knowledge. Second approach is based on the fact that source texts or documents should be interrelated by using random translation systems. After that the translated documents can be compared with the texts that are written in the most targeted language and the basic purpose is to find out the pairs of common pairs within the documents.

An exciting idea for searching for parallel data inside the Wikipedia was mentioned and explained in [9]. Firstly, the idea is to utilize an online machine translation (MT) system to decode the language, using translating techniques to translate Dutch language into English language on Wiki pages, and after that try to compare original EN pages with translated ones. This idea, though seems computationally impractical, is interesting but perplexed problem. Their second method uses a dictionary generated from the hyperlinks and Wikipedia titles that are shared between documents. Inappropriately, the second method involves the generation of Wikipedia titles by using dictionary and the hyperlinks that are being shared between texts. The second method was improved in [9] by a range of spare confines of the communication between the portions of the concern document and with the help of the introduction of added measure on the bases of similarity. They report that in [9] the accuracy (number of correct translation pairs over total strength of the applicants) is approximately 21% and at this stage in the recommended method [10], the accuracy is around 43%.

Yasuda and Sumita [12] projected a MT bootstrapping structure on the basis of figures that helps to create a sentence-structured corpus. Sentence structure and alignment is accomplished by utilizing a bilingual lexicon which is spontaneously upgraded by the structured sentences automatically. They use corpus that has previously been structured for the early training session. Their recommendations showed that 10% of Japanese Wikipedia sentences have an equivalent on the English Wikipedia.

Tyers and Pienaar in [10] gave the idea to give lead to internal Wiki links. A bilingual dictionary is removed and evaluated on the bases of Wikipedia link structure. In the work of these authors, they actually calculated the normal disparity for numerous languages that are linked between Wikipedia pages. Results showed that 69-92% depends on the specific language according to the precision of the method.

The authors in [13] attempt to raise the best skill in parallel data mining with the help of modelling of document-level arrangement by utilizing the observational technique, so that parallel sentences can greatly and most commonly be found in propinquity. Authors also use explanation that is available on Wikipedia and a mechanically injected lexicon model. For that authors report 80% precision and 90% recall.

In [14] author introduces an instinctive arrangement methodology for parallel textual documentation fragments which utilizes a phrase-based SMT system and textual entailment method. The author mentions that important up gradation in SMT quality were adopted (BLEU increased by 1.73) by utilizing this arranged data between French and German languages.

M. Volk and M. Plamada also explained another method for discovering Wikipedia which was explained in [14]. Previously explained methods differ from their solutions. In these methods parallel data was restricted to the monotonically control of the arrange algorithms that were used in order to match the candidate sentences. Their algorithm pays no heed to the position of a candidate in the text and, instead, ranks candidates by means of modified measurements that combine contradictory similarity criteria [15]. Additionally, the authors limit the process of mining towards a specific domain and examine the semantic equality of took out pairs. Mining accuracy is 39% in the work of M. Plamada and M. Volk, while 26% are for loud parallel sentences, with other remained sentences misaligned. Reportedly they say that an up gradation of 0.5 points happened in the BLEU metric out of domain data, but no prominent improvement took place in-domain data.

In [16] the authors suggested to use titles and few meta-information only, like time for specific document and publication data, neglecting its full-fledged contents, to lessen the cost of development of the comparable corpora. The similarity of cosine of the title terminology as frequency vectors was utilized to match the contents and the contents of the matched pairs. In the research explained in [17], the two authors came up with a document of resemblance of measure which is based on the occasions. In order to count the values of this metric, they present documents as sets of occasions and events. These occasions are time based and are based on the geographical terms that are found in the texts. Documents that are targeted ranked and based on geographical orders. The writers in [18] also recommend a programmed method in order to build a similar corpus from the website by utilizing news web pages like Twitter and Wikipedia. They mine things, URLs of web pages, filtering within time limitations and the specified lengths of the documents as features for the congregation and the categorization of the similar data.

In the current research, a methodology that was inspired by the Yalign is being used. The method originally was far from perfect, but after the up gradation and improvements during research, it actually has provided us with excellent mining results.

## IV. Parallel data mining

This study aims at developing new methodologies for obtaining parallel corpora from the sources that are not aligned likewise comparable corpora, quasi-comparable or noisy parallel. We have selected Wikipedia as a base of the data because of the huge number of texts it provides (4,524,017 on EN wiki approximately). Moreover, Wikipedia comprises not only similar documents, but it also includes some text documents which are the translations of one another. This approach can be qualified by the use of measurements in the translations systems of MT.

In data mining, TED corpora, ready for the IWSLT 2015 assessment by FBK, were selected. This domain is very wide and protects numerous subject areas. Data contains nearly 2.5M un-accessible words [19]. The tests were shown on PL-*.

Our idea can be separated and divided into three key steps. Firstly, comparable data is selected, then it is arranged at the article level, and lastly for parallel, the arranged results are mined. The last two steps are important; the reason is that there are huge numbers of differences between documents of Wikipedia. Sentences in the Wiki corpus are mainly not arranged, with translation lines whose assignment does not agree to any textual information in the foreign language. Furthermore, some sentences have no consistent translations within corpus at all. The alignment is very difficult for the correctness. For that sentence alignment should also be practicable with competency that is of practical use in variety of applications. Earlier, a mining tool precedes the data and the text should be ready. Initially, entire data is preserve in a relational database. In the second phase, our tool organizes article pairs and eradicates the articles that seem to be present in only one of the two languages. All these arranged articles at topic-level are checked in order to get rid of XML tags, HTML tags or other noisy data (figures, references, tables, etc.). Lastly, fluent documents are marked with a single ID as an aligned topic, similar corpus. In order to separate the parallel pair's sentence, a decision was taken in an attempt to develop strategy that was designed to program the parallel text mining process after finding the sentences that are near to the translation matches from comparable corpora. This offers chances for finding parallel corpora from bases, as not translated textual documents including the web, which are not limited to any specific language pair.

Though, alignment models for two languages that are of two designated languages particularly the first one is to be created as priority. By using a comparative sentence metric gave an unbalance estimate (a number that is somehow between 0 and 1). This approximation shows that how there is a possibility of being a translation of the two sentences. It also applies pattern alignment, which gave a sequence that increases the quantity of the unique thread (per sentence pair) that is same between the two texts [6]. In order to maintain order alignment, we at first used error friendly and very slow Yalign that used an A* search method [20] to find an ideal alignment between multiple sentences in two particular documents. The algorithm has a polynomial worst time complication. It cannot control alignments that cancel each other or such that form from two

sentences a single sentence [20]. After the sequence alignment, only sentences which have top probability of being translated are placed in the results. The output is checked to deliver top quality corpus. To accomplish this, an action is used: if the sentence has the similar score that is less than the threshold, at that stage pair will not be included. For similarity of sentence metric, the algorithm uses statistical techniques. The classifier must be skilled to find if the sentence pairs are translations between each other. In this research Support Vector Machine (SVM) classifier is used. SVM can give a distant outlook to the parting hyper plane during labelling. This distance can be simply adapted by utilizing a Sigmoid Function to return a value that is similar to similarity between 0 and 1 [21]. Usage of classifier means that the excellence of the rearrangement depends not just on the input but it also depends on the quality of the classifier. For the training of the classifier, high quality parallel data is necessary. For this, we utilized the TED talks [8] corpora. To get a dictionary, we used a phrase table and took 1-grams from it [22].

## V. Development of the Mining Process

Much to our distress, the local Yalign instrument was not practical enough for matters related to calculation in terms of comprehensive and real life parallel data mining. Typical execution required input in simple text or web links and the RAM memory was re-loaded with classifier for every text pair. Moreover, the Yalign software is uniquely stranded. In an effort to speed up the process, this unique solution was developed to supply articles to the tool and load classifier only once per session. The newly developed system also used the multithreading and minimized the mining time by factor of 6.1x, utilizing the four cores and eight thread i7 CPU. The alignment algorithm was replaced to improve accuracy and to make best use of the strengths of the GPUs to fulfil the supplementary requirements of the computations.

### *A. Needleman-Wunsch algorithm (NW)*

Major purpose of this algorithm is to line up two sequences together. At first, it is necessary to define the correspondence among the two elements. It can be explained by using the similarity matrix S in which N represents the number of elements in the first sequence and M represents the number of elements in the second sequence. The algorithm was designed to analyse the matters related to bio-informatics for the assessment of RNA and DNA. However, it can be modified to deal with assessment of textual data. In other words, the given algorithm combines real number and a pair of each element together in the matrix. As the similarity index rises, so is the similarity of the elements concern. For instance, if we have a similarity matrix S which is equal to the numbers between 0 and 1 than by 0 for the two expressions mean that their similarity index is zero whereas 1 means that the two given

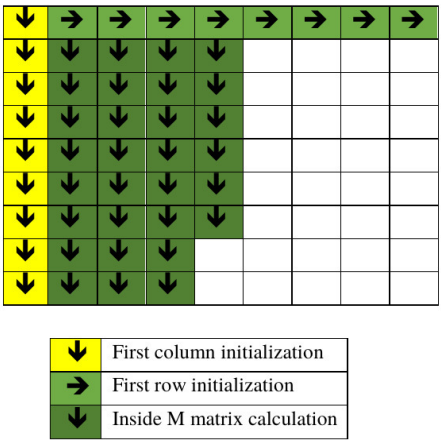


Fig. 1 Needleman-Wunsch S-matrix calculation

words are perfect translations of each other. The significance of similarity matrix index for the consequences of the algorithm is undeniable [20]. After that we will identify the consequences of gap penalty. It is essential particularly when one of the elements of the sequence is connected with the gap in another sequence.

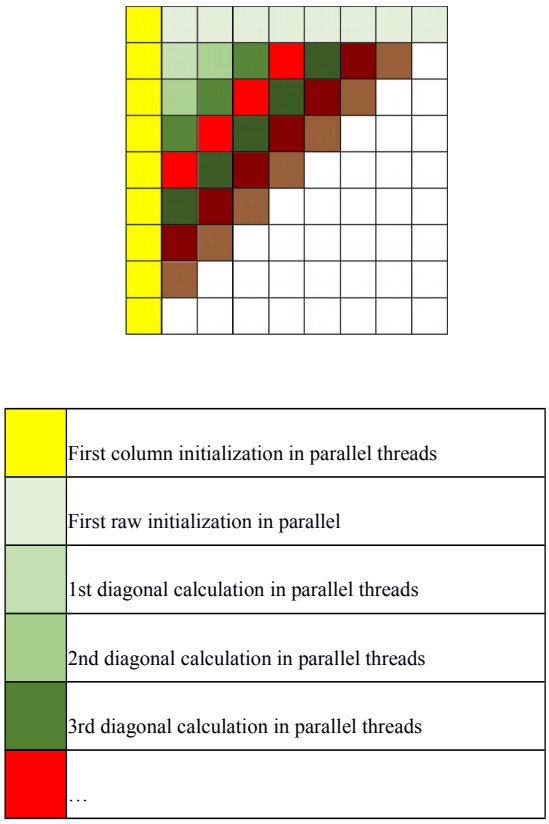


Fig. 2 Needleman-Wunsch S-matrix calculation with parallel threads

Though, such a stride will lead to a penalty (p). The calculation of the S matrix will be executed right from the start of S (0, 0) element which by definition is equal to 0. Once the process of initializing first and second row is started, the algorithm moves to the other elements of the S matrix, taking the cue from the upper left side, leading up to the bottom right side. All of these steps are illustrated in the Fig. 1.

The two (CPU and GPU) NW algorithms are theoretically same but the GPU has a leverage of better performance due to its hardware handicap up to max(n,m) times.

Though, this process is different when it comes to the calculation of the elements of S matrix. At this step forward, we will apply the multiple strands to an optimized level. These processes are so small so that that they can be processed easily by a huge number of Graphics Processing Units (ex. CUDA cores). The major purpose behind this is that we will calculate all the elements in predefined diagonals in parallel way which always starts from upper left and ends at the bottom right, as presented in the Fig. 2 [23].

In an attempt to explain the value of a cell of S(m,n), for all pairs of m and n, the values to its top S(m-1,n) , left S(m,n-1) and top left S(m-1,n-1) must be known in advance and filled with tokenized documents that are to be compared. Where, S(m,n) can be calculated with the help of following equation: [24]

$$S(m,n) = max\{S(m-1) \pm 1, \\ S(m-1,n)-2, S(m,n-1)-2\}$$

Regardless of the results of the A* algorithm, if the coherence between calculation and the gap penalty are defined on the similar patterns as they are defined in the NW algorithm, they will have similar results just if there are supplementary restraints on the way, these ways cannot be led to uphill or left-side in the matrix. Yalign does not compel such terms and conditions, therefore in some cases, expressions can be repeated more than once or misaligned. Most of the cases the algorithm keeps on moving backward and forward to the first two sequences in line. S matrix has been exemplified without any barriers in the Fig. 3.

| | a | d | e | g | f |
|---|---|---|---|---|---|
| a | X | | | | |
| d | | X | | | |
| c | X | | | | |
| d | | X | | | |
| e | | | X | X | X |

Fig. 3 S matrix pass-through without constraints

The alignment result in this scenario is:

a, d, a, d, e, g, f

a, d, c, d, e, −, −

In the same problem, the NW would react as presented in Fig. 4:

| | a | d | e | g | f |
|---|---|---|---|---|---|
| a | X | | | | |
| d | | X | | | |
| c | | X | | | |
| d | | X | | | |
| e | | | X | X | X |

Fig. 4 S matrix pass through with NW

The alignment result using NW would be:

a, d, −, −, e, g, f

a, d, c, d, e, −, −

For second example, we will assume that the very first sequence in a row is “Tablets make children very addicted” and second one says that “Tablets make people spoil children”. Fig. 5 given below presents a solution to this problem by the help of A* algorithm and Fig. 6 shows it with NW, without any restrains.

| | Tablets | make | children | very | Addicted |
|---|---|---|---|---|---|
| tablets | X | - | - | - | - |
| make | - | X | - | - | - |
| people | X | - | - | - | - |
| spoil | - | - | - | - | - |
| children | - | - | X | X | X |

Fig. 5 A* alignment without constraints

Due to lack of any restrains, repetitions and bad alignments are most likely to be made by envisaging the blemish A* algorithm, which is applicable in the Yalign program. However, some of the sentences can be easily passed over during the checking of the alignment. That is why NW with GPU optimization is the most preferred algorithm.

| | Tablets | make | children | very | addicted |
|---|---|---|---|---|---|
| tablets | X | - | - | - | - |
| make | - | X | - | - | - |
| people | - | - | - | - | - |
| spoil | - | - | - | - | - |
| children | - | - | X | - | - |

Fig. 6 NW alignment with constraints

### *B. Other improvements to data mining methodology*

The SVM classifier is used to define the quality of the alignment; it creates a trade-off between the recall and the precision. Two configurable variables can be found in the classifier.

- Threshold: the acceptance of an alignment is 'Good' if the confidence threshold if high. For less recall and more precision, the value should be lowered. The probability estimated through the support vector machine is the 'confidence' that classifies it as 'is a translation' or 'is not a translation' [25].
- Penalty: the alignment allows control of the amount of 'slipping ahead' [6] if you are aligning subtitles. There would then be no extra or fewer paragraphs, and the alignment would be one-to-one while the penalty would be high. The penalty would be lower if the translations of the alignments are similar and there are no extra paragraphs.

These parameters are automatically selected during the training; however, they can be manually changed if necessary. A tuning algorithm is also introduced in our implementation of the solution[1] used. In this research, it allows adjustments for better accuracy. Random articles of the corpus must be extracted to perform the tuning; humans can manually align these random articles. Given the information provided, the tuning mechanism finds the classifier value through randomly selected parameters; it tries to find the output that would be as humanly similar as possible.

The improvements that were debated earlier, deal entirely with heuristics utilized in the mining tool and they can be implemented to any fluent textual data. Though, Wikipedia has huge extra sources of cross-lingual information that need to be utilized. Firstly, the theme domain of Wikipedia cannot be enclosed into a particular domain; this page covers approximately any topic in question. Due to the fact that these articles frequently contain complex vocabulary, and that is why approaches of statistical mining can skip many of the parallel sentences. The answer to this query can be extracted dictionary by utilizing the article titles from Wikipedia (Fig. 7) and moreover it can be implemented into web crawler tool.


Fig. 7 Sample of bi-lingual Wikipedia page title

In [11] authors say that the precision of this dictionary can be attained at 92%. For that this dictionary can be utilized not only for the delay of the parallel corpora but in the classifier phase of exercise as well.

Secondly, figures contain best quality parallel sentences and explanations. It is likely to attain pictures and graphics with the help of hyperlinks and analysis the pictures. Similarly, same is for any figures, maps, tables, videos, audio or even compound media on Wikipedia. Tactlessly not whole knowledge can be removed from Wikipedia dumps and it is essential to utilize a web crawler that is right for this assignment (Fig. 8). It also can be predicted that simply only cross-language knowledge which are marked with simple links can be removed.


Fig. 8 Specimen of bi-lingual character caption

On Wikipedia it is very likely for the sentences to be equal linguistically, if they are referenced with the similar publication. Such an analytical approach, combined with other comparative tactics, can move us forward to better accuracy in a parallel text sequence discovery task (Fig. 9).

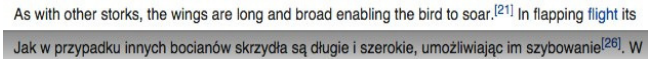


Fig. 9 Sample of bilingually referenced sentence

[1]https://github.com/krzwolk/yalign

## VI. Assessment of obtained parallel corpora and conclusions

By the help of techniques explained earlier we were capable of creating comparable corpora for many PL-* language pairs and later, probe them for parallel phrases. We paired Polish (PL) with Arabic (AR), Czech (CS), German (DE), Greek (EL), English (EN), Spanish (ES), Persian (FA), French (FR), Hebrew (HE), Hungarian (HU), Italian (IT), Dutch (NL), Portuguese (PT), Romanian (RO), Russian (RU), Slovene (SL), Turkish (TR), Vietnamese (VI). Statistics of the resulting corpora are presented in Table I.

In order to assess the corpora quality and usefulness, we trained the baseline SMT systems by utilizing the WIT[2] data (BASE). We also augmented them with resulting mined corpora both as parallel data as well as the language models (EXT). The additional corpora were domain adapted through the linear interpolation and Modified Moore-Lewis filtering [26]. Tuning of the system was not executed during experiments due to the volatility of the MERT [27]. However, usage of the MERT would have an overall positive impact on MT system in general [27]. The results are showed in Table II.

The assessment was based on sets of official test sets from IWSLT 2013[3] conference. Bilingual Evaluation Understudy (BLEU) measurement was used to score the progress. As it was expected earlier, sets of supplementary data enhance the general quality of translation for each and every language.

In order to verify the importance of our results we conducted the significance tests for 4 divers languages. The decision was made to use the Wilcoxon test. The Wilcoxon test (also known as the signed-rank test or the matched-pairs test) is one of the most popular alternatives for the Student's t-test for dependent samples. It belongs to the group of non-parametric tests. It is used to compare two (and only two) dependent groups that is two measurement variables. The significance tests were conducted to evaluate how the improvements differ from each other. Changes with low significance were marked with *, significant changes were marked with ** and very significant with *** in presented Tables III and IV.

Bi-lingual sentence extraction has particular importance in dealing with unsubstantiated learning processes for multiple tasks involved. With the help of this methodology we can easily resolve the dissimilarities between Polish and other languages. It is a method that is independent from language matters, it is adaptable to new environments for any language pair. Our experiments validated the performance of the method. The corpora received as consequence of the experiments, can maximize the quality of MT in an under-resourced text domain. However, in few scenarios, only small differences have been observed in BLEU scores. Keeping that aside, it can be said that even such small differences, can influence the real life situations positively especially for infrequent translation cases. Moreover, the results of our work are freely accessible for research community (corpora is hosted at OPUS[4] and tools at GitHub[5]). In order to see things from sensible outlook, we can say that such methodology does not require large scale training or special language specific

Table I.

Results of mining after progress

| Language Pair | Number of bi-sentences | Number of unique PL tokens | Number of unique foreign tokens |
|---|---|---|---|
| PL-AR | 823,715 | 1,345,702 | 1,541,766 |
| PL-CS | 62,507 | 197,499 | 206,265 |
| PL-DE | 169,739 | 345,266 | 341,284 |
| PL-EL | 12,222 | 51,992 | 51,384 |
| PL-EN | 172,663 | 487,999 | 412,759 |
| PL-ES | 151,173 | 411,800 | 377,557 |
| PL-FA | 6,092 | 31,118 | 29,218 |
| PL-FR | 51,725 | 215,116 | 206,621 |
| PL-HE | 10,006 | 42,221 | 47,645 |
| PL-HU | 41,116 | 130,516 | 136,869 |
| PL-IT | 210,435 | 553,817 | 536,459 |
| PL-NL | 167,081 | 446,748 | 425,487 |
| PL-PT | 208,756 | 513,162 | 491,855 |
| PL-RO | 6,742 | 38,174 | 37,804 |
| PL-RU | 170,227 | 365,062 | 440,520 |
| PL-SL | 17,228 | 71,572 | 71,469 |
| PL-TR | 15,993 | 93,695 | 92,439 |
| PL-VI | 90,428 | 240,630 | 204,464 |

[2]https://wit3.fbk.eu/mt.php?release=2013-01

[3]Iwslt.org

[4] http://opus.lingfil.uu.se/Wikipedia.php

[5]https://github.com/krzwolk/yalign

TABLE II.

RESULTS OF MT EXPERIMENTS

| LANGUAGE | SYSTEM | DIRECTION | BLEU | LANGUAGE | SYSTEM | DIRECTION | BLEU | LANGUAGE | SYSTEM | DIRECTION | BLEU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PL-AR | BASE | →PL | 19.67 | PL-FA | BASE | →PL | 14.21 | PL-PT | BASE | →PL | 27.07 |
| | EXT | →PL | 21.78 | | EXT | →PL | 14.32 | | EXT | →PL | 29.14 |
| | BASE | ←PL | 20.98 | | BASE | ←PL | 16.87 | | BASE | ←PL | 30.11 |
| | EXT | ←PL | 23.12 | | EXT | ←PL | 17.03 | | EXT | ←PL | 31.33 |
| PL-CS | BASE | →PL | 12.21 | PL-FR | BASE | →PL | 19.07 | PL-RO | BASE | →PL | 22.16 |
| | EXT | →PL | 12.98 | | EXT | →PL | 20.01 | | EXT | →PL | 22.26 |
| | BASE | ←PL | 13.44 | | BASE | ←PL | 21.13 | | BASE | ←PL | 25.01 |
| | EXT | ←PL | 14.21 | | EXT | ←PL | 21.56 | | EXT | ←PL | 25.67 |
| PL-DE | BASE | →PL | 23.68 | PL-HE | BASE | →PL | 17.03 | PL-RU | BASE | →PL | 12.36 |
| | EXT | →PL | 24.91 | | EXT | →PL | 17.65 | | EXT | →PL | 13.51 |
| | BASE | ←PL | 26.61 | | BASE | ←PL | 18.18 | | BASE | ←PL | 13.58 |
| | EXT | ←PL | 26.87 | | EXT | ←PL | 18.54 | | EXT | ←PL | 14.32 |
| PL-EL | BASE | →PL | 14.27 | PL-HU | BASE | →PL | 14.62 | PL-SL | BASE | →PL | 12.11 |
| | EXT | →PL | 14.67 | | EXT | →PL | 15.23 | | EXT | →PL | 12.57 |
| | BASE | ←PL | 17.22 | | BASE | ←PL | 17.18 | | BASE | ←PL | 14.26 |
| | EXT | ←PL | 17.28 | | EXT | ←PL | 17.81 | | EXT | ←PL | 14.61 |
| PL-EN | BASE | →PL | 15.91 | PL-IT | BASE | →PL | 18.83 | PL-TR | BASE | →PL | 11.59 |
| | EXT | →PL | 17.01 | | EXT | →PL | 19.87 | | EXT | →PL | 12.68 |
| | BASE | ←PL | 17.09 | | BASE | ←PL | 21.19 | | BASE | ←PL | 13.07 |
| | EXT | ←PL | 18.43 | | EXT | ←PL | 21.34 | | EXT | ←PL | 13.44 |
| PL-ES | BASE | →PL | 16.35 | PL-NL | BASE | →PL | 18.29 | PL-VI | BASE | →PL | 12.66 |
| | EXT | →PL | 17.92 | | EXT | →PL | 20.13 | | EXT | →PL | 14.12 |
| | BASE | ←PL | 18.34 | | BASE | ←PL | 20.79 | | BASE | ←PL | 14.11 |
| | EXT | ←PL | 18.65 | | EXT | ←PL | 21.45 | | EXT | ←PL | 15.17 |

rammer resources, and despite of that they produce gratifying results.

Because statistically classified data contains some amounts of noisy data, in future we plan to develop precise filtering strategies for bi-lingual corpora. The results of current solution are highly related to SVM classifier. In other words, we plan to train more classifiers for different text domains in order to discover more bi-lingual sentences.

## VII. ACKNOWLEDGEMENTS

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and was backed by the PJATK legal resources.

## VIII. REFERENCES

[1] K. Wołk, K. Marasek. „Real-Time Statistical Speech Translation." *In: New Perspectives in Information Systems and Technologies*, Volume 1. Springer International Publishing, 2014, p. 107-113. http://dx.doi.org/10.1007/978-3-319-05951-8_11

[2] K. Wołk, K. Marasek. „Polish–English Speech Statistical Machine Translation Systems for the IWSLT 2013". *In: Proceedings of the 10th International Workshop on Spoken Language Translation*, Heidelberg, Germany. 2013, p. 113-119. http://dx.doi.org/10.13140/RG.2.1.1128.9204

[3] A. Haghighi et al. "Better word alignments with supervised ITG models." *In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. Association for Computational Linguistics, 2009, p. 923931.

[4] P. Koehn. „Statistical machine translation." *Cambridge University Press*, 2009. http://dl.acm.org/citation.cfm?doid=1380584.1380586

[5] G. Berrotarán, R. Carrascosa, A. Vine „Yalign documentation", https://yalign.readthedocs.org - accessed 01/2015

[6] R. Dieny, J. Thevenon, J. Martinez-Delrincon, J. C. Nebel. „Bioinformatics inspired algorithm for stereo correspondence." *International Conference on Computer Vision Theory and Applications*, March 5–7, Vilamoura - Algarve, Portugal, 2011.

[7] G. Musso. „Sequence alignment (Needleman-Wunsch, Smith-Waterman)", http://www.cs.utoronto.ca/~brudno/bcb410/lec2notes.pdf.

[8] M. Cettolo, C. Girardi, M. Federico. "Wit3: Web inventory of transcribed and translated talks." *In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. 2012, p. 261-268.

[9] M. Mohammadi; N. Ghasemaghaee. „Building bilingual parallel corpora based on Wikipedia." *In: Computer Engineering and Applications (ICCEA)*, 2010 Second International Conference on. IEEE, 2010, p. 264-268. http://dx.doi.org/10.1109/ICCEA.2010.203

[10] F. M. Tyers, J. A. Pienaar. „Extracting bilingual word pairs from Wikipedia", *Collaboration: interoperability between people in the creation of language resources for less-resourced languages 19*, 2008, p. 19-22.

[11] J. R. Smith, C. Quirk, K. Toutanova. „Extracting parallel sentences from comparable corpora using document level alignment." *In:*

TABLE III.

SIGNIFICANCE TESTS PL ->*

| Language Pair | Number of bi-sentences |
|---|---|
| PL-EN | 0.0103** |
| PL-CS | 0.0217** |
| PL-AR | 0.0011*** |
| PL-VI | 0.0023*** |

TABLE IV.

SIGNIFICANCE TESTS * ->PL

| Language Pair | Number of bi-sentences |
|---|---|
| PL-EN | 0.0193** |
| PL-CS | 0.0153** |
| PL-AR | 0.0016*** |
| PL-VI | 0.0027*** |

*Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, p. 403-411.

[12] K. Yasuda, E. Sumita. „Method for building sentence-aligned corpus from wikipedia". *In: 2008 AAAI Workshop on Wikipedia and Artificial Intelligence (WikiAI08),* 2008, p.263-268.

[13] S. Pal, P. Pakray, S. K. Naskar. "Automatic Building and Using Parallel Resources for SMT from Comparable Corpora." *In: Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, 2014. p. 48-57.

[14] M. Plamada, M. Volk. "Mining for Domain-specific Parallel Text from Wikipedia." *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, ACL 2013, 2013, p.112-120. http://dx.doi.org/10.5167/uzh-80043

[15] A. Aker, E. Kanoulas, R.J. Gaizauskas. "A light way to collect comparable corpora from the Web". *In: LREC*, 2012, p. 15-20.

[16] J. Strötgen, M. Gertz, C. Junghans. "An event-centric model for multilingual document similarity." *In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM*, 2011, p. 953-962. http://dx.doi.org/10.1145/2009916.2010043

[17] M.L. Paramita et al. "Methods for collection and evaluation of comparable documents." *In: Building and Using Comparable Corpora*. Springer Berlin Heidelberg, 2013, p. 93-112. http://dx.doi.org/10.1007/978-3-642-20128-8_5

[18] D. Wu, P. Fung. "Inversion transduction grammar constraints for mining parallel sentences from quasicomparable corpora." *In: Natural Language Processing– IJCNLP 2005*. Springer Berlin Heidelberg, 2005, p. 257-268. http://dx.doi.org/10.1007/11562214_23

[19] J.H. Clark et al. "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability." *In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, p. 176181.

[20] S. Adafre; M. De Rijke. „Finding similar sentences across multiple languages in Wikipedia." *In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006, p. 6269.

[21] K. Wołk, K. Marasek. "A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation." *In: New Perspectives in Information Systems and Technologies*, Volume 1. Springer International Publishing, 2014, p. 229-237. http://dx.doi.org/10.1016/j.protcy.2014.11.024

[22] A. Axelrod, X. HE, J. Gao. "Domain adaptation via pseudo in-domain data selection." *In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2011, p. 355-362.

[23] K. Wołk, K. Marasek. "Tuned and GPU-accelerated parallel data mining from comparable corpora." *In: Text, Speech, and Dialogue. Springer International Publishing*, 2015, p. 32-40. http://dx.doi.org/10.1007/978-3-319-24033-6_4

[24] C. S. Khaladkar. "An Efficient Implementation of Needleman Wunsch Algorithm on Graphical Processing Units", PHD Thesis, School of Computer Science and Software Engineering, The University of Western Australia, 2009.

[25] https://github.com/machinalis/yalign/issues/3 accessed 10.11.2015

[26] R. Roessler. "A GPU implementation of NeedlemanWunsch, specifically for use in the program pyronoise 2." *Computer Science & Engineering*, 2010.

[27] T. Joachims. "Text categorization with support vector machines: Learning with many relevant features." *Lecture Notes in Computer Science vol 1398*, 2005, p. 137-142. http://dx.doi.org/10.1007/BFb0026683