Mixing Textual Data Selection Methods for Improved In-Domain Data Adaptation

Krzysztof Wołk(✉)

Polish-Japanese Academy of Information Technology,
Koszykowa 86, Warsaw, Poland
kwolk@pja.edu.pl

Abstract. The efficient use of machine translation (MT) training data is being revolutionized on account of the application of advanced data selection techniques. These techniques involve sentence extraction from broad domains and adaption for MTs of in-domain data. In this research, we attempt to improve in-domain data adaptation methodologies. We focus on three techniques to select sentences for analysis. The first technique is term frequency–inverse document frequency, which originated from information retrieval (IR). The second method, cited in language modeling literature, is a perplexity-based approach. The third method is a unique concept, the Levenshtein distance, which we discuss herein. We propose an effective combination of the three data selection techniques that are applied at the corpus level. The results of this study revealed that the individual techniques are not particularly successful in practical applications. However, multilingual resources and a combination-based IR methodology were found to be an effective approach.

Keywords: Text domain adaptation · In-domain adaptation · Data filtration
Corpora adaptation · Machine learning

1 Introduction

The performance of statistical machine translation (SMT) [1] is heavily dependent on the quantity of training data and the domain specificity of the test data as it relates to the training data. An obstacle to optimal performance is the data-driven system not guaranteeing optimum results if either the training data or testing data are not uniformly distributed. Domain adaptation is a promising approach to increasing the quality of domain-specific translation systems with a mixture of out-of-domain and in-domain data.

The prevalent adaptation method is to choose the data to target the field or domain from a general cluster of documents within the domain. However, the method is applied when the quantity of data is adequately wide to cover some sentences that will exist in the targeted field. Moreover, a domain-adapted machine translation system can be attained through training that uses a chosen subset of the data. Axelrod et al. [2] explained this point as a quasi-in-domain sub-part of the corpus instead of using the complete corpus data.

The present paper focuses entirely on these auxiliary data selection methodologies, which have advanced the development of narrow-domain SMT systems. Translation

Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 746, pp. 367–377, 2018.
https://doi.org/10.1007/978-3-319-77712-2_35

models that are trained in this way will have the benefit of enhanced word arrangements. Furthermore, the system can be modified to avoid redundant pairs of phrases; moreover, proper estimation can support reorganization of the elements of target sentences.

The similarity measurement has an immense influence on the translation quality. In this study, data selection methods that can boost the quality of domain-specific translation were explored. To this end, data selection criteria were carefully analyzed. Two models are thus considered in this paper. One is based on term frequency–inverse document frequency (tf-idf); the other is a perplexity-based approach. These two techniques have roots in information retrieval (IR) and language modeling for SMT. A third approach uses the Levenshtein distance, which is then analyzed.

An evaluation revealed that each of these methods has advantages and disadvantages. First, the tf-idf technique employs the text as a set of words and recovers sentences that are similar. Although this approach helps reduce the number of out-of-vocabulary words, it does not filter bad data. On the other hand, perplexity-oriented measurement tools leverage an n-gram language model (LM), which considers both a grammar's word order and term distribution. It filters irrelevant phrases and out-of-vocabulary words. However, the quality of the filtering depends largely on the in-domain LM and quasi-in-domain sub-parts.

The methodology based on the Levenshtein distance is more stringent in its approach than the other two. It is intended to explore the similarity index; however, in terms of performance, it does not surpass the others on account of its reliance on data generalization. As the number of factors increases, the complexity of the similarity judgment also increases. This scenario can be depicted by a pyramid to show the relevant intensities of multiple approaches. The Levenshtein distance approach is at the top of the pyramid, perplexity is in the middle, and the tf-idf approach is at the bottom. The positive and negative aspects of each method can be addressed by considering all these criteria. If we consider additional factors, then the criterion at each highest point will become stricter.

In this study, the above measurement approaches were combined and compared to each separate method and to the modified Moore–Lewis filtering implementation in the Moses SMT system. A comparative experiment was conducted using a generalized Polish–English language movie-subtitle corpus and in-domain TED lecture corpus. The SMT systems were adapted and trained accordingly. Utilizing the bilingual evaluation understudy (BLEU) metric, the testing results revealed that the designed approach produced a promising performance.

The remainder of this paper is organized as follows. Related literature is reviewed in Sect. 2. Related models are analyzed in Sect. 3. Section 4 describes evaluation methods. The experiment results and conclusions are outlined in Sect. 5.

2 State of the Art

Existing literature discusses data adaptation for SMT from multiple perspectives, such as finding unknown words from comparable corpora [3], corpora weighting [4], mixing multiple models [5–7], and weighted phrase extraction [8]. The predominant criterion

for data selection is tf-idf, which originated in the area of IR. Hildebrand et al. [9] utilized this IR technique to choose the most similar sentence—albeit with a lower quantity—for translation model (TM) and LM adaptation. The results strengthen the significance of the methodology for enhancing translation quality, particularly for LM adaptation.

In a study much closer to the present research, Lü et al. [10] suggested reorganizing the methodology for offline, as well as online, TM optimization. The results are much closer to those of a realistic SMT system. Moreover, their conclusions revealed that repetitive sentences in the data can affect the translation quality. By utilizing approximately 60% of the complete data, they increased the BLEU score by almost one point.

The second technique in the literature is a perplexity approach, which is common in language modeling. This approach was used by Lin et al. [11] and Gao et al. [12]. In that research, perplexity was utilized as a standard in testing parts of the text in accordance with an in-domain LM approach. Other researchers, such as Moore and Lewis [13], derived the unique approach of a cross-entropy difference metric from a simpler version of the Bayes rule. This methodology was further examined by Axelrod et al. [2], particularly for SMT adaptation, and they additionally introduced an exclusive unique bilingual methodology and compared its results with contemporary approaches. Results of their experiments revealed that, if the system was kept simple yet sufficiently fast, it discarded as much as 99% of the general corpus, which resulted in an improvement of almost 1.8 BLEU points.

Early works discuss separately applying the methodology to either a TM [2] or an LM [10]; however, in [10], Lü suggests that a combination of LM and TM adaptation will actually enhance the overall performance. Therefore, in the present study, TM and LM optimization was investigated through a combined data selection method.

3 Combined Corpora Adaptation Method

Four selection criteria are discussed to describe the examined models: tf-idf, perplexity, Levenshtein distance, and the proposed combination approach.

3.1 TF-IDF

In the approach based on tf-idf, each document D_i is represented as a vector $(\mathrm{wi1},\ \mathrm{wi2}, \ldots, \mathrm{win})$, where n is the vocabulary size. Thus, w_{ij} is calculated as:

$$w_{ij} = tf_{ij} \times \log\left(idf_j\right).$$

where, $tf_{ij}i$ is the term frequency (TF) of the j-th word in the vocabulary in the document D_i and idf_j is the inverse document frequency (IDF) of the j-th word. The similarity between the two texts is the cosine of the angle between the two vectors. This formula is applied in accordance with Lü et al. [10] and Hildebrand et al. [9]. The

approach supposes that M is the size of the query set and N is the number of similar sentences from the general corpus for each query. Thus, the size of the tf-idf-based quasi-in-domain sub-corpus is defined as:

$$Size_{Cos-IR} = M \times N.$$

3.2 Perplexity

Perplexity focuses on cross-entropy [14], which is the average of the negative logarithm of word probabilities. Consider:

$$H(p,q) = -\sum_{i=1}^{n} p(w_i) \log q(w_i) = -\frac{1}{N} \sum_{i=1}^{n} \log q(w_i),$$

where p is the empirical distribution of the test sample. If w_i appears n times in the test sample of size N, then $q(w_i)$ is the probability of the w_i event approximated from the training set.

Perplexity (pp) can be simply calculated at the base point presented in the system. It is often applied as a symbolic alternative to perplexity for the data selection as:

$$pp = b^{H(p,q)},$$

where b is the basis of measured cross-entropy, and $H(p,q)$ is the cross-entropy as given in [14] which is often used as a substitute for perplexity in data selection [2, 13].

Let $H_I(p,q)$ and $H_O(p,q)$ be the cross-entropy of the w_i string in accordance with the language model, which is subsequently trained by a general-domain dataset and an in-domain dataset. While examining the target (tgt) and source (src) dimensions of the training data, three perplexity-based variants exist. The first one, known as basic cross-entropy, is defined as:

$$H_{I-src}(p,q).$$

The second is Moore-Lewis cross-entropy difference [13].

$$H_{I-src}(p,q) - H_{G-src}(p,q).$$

which attempts to choose the sentences that are most identical to the ones in I but unlike others in G. Both the standards mentioned above consider only sentences in the source language. Moreover, Axelrod et al. [2] proposed a metric that adds cross-entropy differences to both sides:

$$[H_{I-src}(p,q) - H_{G-src}(p,q)] + \left[H_{I-tgt}(p,q) - H_{G-tgt}(p,q)\right].$$

For instance, candidates with lower scores [3, 15, 16] have a higher relevance to the specific target domain. The sizes of the perplexity-based quasi-in-domain subsets must

be equal. In practice, we work with the SRI Language Modeling (SRILM) toolkit to train 5-gram LMs with interpolated modified Kneser–Ney discounting [17, 18].

3.3 Levenshtein Distance

In information theory and computer science, the Levenshtein distance is regarded as a string metric for the measurement of dissimilarity between two sequences. The Levenshtein distance between points or words is the minimum possible number of unique edits to the data (e.g., insertions or deletions) that are required to replace one word with another.

The Levenshtein distance can additionally be applied to a wider range of subjects as a distance metric. Moreover, it has a close association with pairwise string arrangement.

Mathematically, the Levenshtein distance between two strings a, b (of length |a| and |b|, respectively) is given by $lev_{a,b}(|a|, |b|)$, where:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if \min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise. \end{cases}$$

Here, $1_{(a_i \neq b_j)}$ is the indicator function, which is equal to 0 when $a_i = b_j$; otherwise, it is equal to 1. Furthermore, $lev_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b. The first component has the least correspondence to the deletion (from a to b), the second-closest correspondence to the insertion, and the most correspondence to a match or mismatch.

3.4 Combined Methods[1]

As was first explained by Wang et al. [19], there are three basic processing stages in data selection for domain adaptation. First, we extract sentence pairs from a parallel corpus. A generalized domain corpus is obtained based on significance and corresponding relevance to the targeted domain. Second, the samples are reorganized to maintain the quasi-in-domain sub-corpus. These first two steps are applicable to a general domain monolingual corpus and they are significant for selecting sentences for a language model. Once a large number of sentence pairs are collected, these models are scheduled for data training and will eventually represent the target domain.

In a similar fashion, the similarity index measurement is required to choose the sentences for a quasi-in-domain sub-corpus. For the similarity measurement, three approaches are regarded as the most suitable. First, the tf-idf criterion identifies the similarity by considering the word overlap. This technique is particularly helpful in reducing out-of-vocabulary words. Nevertheless, it is sensitive to irrelevant data in the system. The perplexity-based criterion, on the other hand, is more focused on the

[1] https://Github.Com/Krzwolk/Text-Corpora-Adaptation-Tool.

n-gram word order. Meanwhile, the Levenshtein distance considers the word order, position of the words, and word overlap. Of the three approaches, it is the most stringent.

In this study, a combination of the corpora and language models is used. These three methods are first individually used to identify the quasi-in-domain sub-corpora. They are later combined during the reorganization phase to collectively leverage the benefits of all three metrics. Similarly, these three metrics are joined for domain adaptation during the translation process. Experimental evidence demonstrated the success of this process. In addition, our adaptation tool is freely available for use.

4 Evaluation

To advance machine translation (MT), the quality of the MT results must be evaluated. It has been recognized that using humans to evaluate MT approaches is costly and time consuming [20]. As a result, human evaluation cannot remain abreast of the growing and continual need for MT evaluation. Consequently, the development of automated MT evaluation techniques is critical. Evaluation is particularly crucial for translation between languages of different families, such as Polish and English languages from respective Germanic and Slavic families [20, 21].

In Reeder [21], Reeder compiled an initial list of SMT evaluation metrics. Further research led to the development of newer metrics. Prominent metrics include Bilingual Evaluation Understudy (BLEU), the National Institute of Standards and Technology (NIST), Translation Error Rate (TER), and the Metric for Evaluation of Translation with Explicit Ordering (METEOR). These metrics were used in the present research for evaluation.

In this study, we employed the most renowned metric, BLEU, which was developed based on a premise similar to that used for speech recognition. It is described in Papineni et al. [16] as follows: "The closer a machine translation is to a professional human translation, the better it is." Accordingly, the BLEU metric is designed to measure how close SMT output is to those of human reference translations. It is important to note that translations, whether SMT or human, may significantly differ in word usage, word order, and phrase length [16].

4.1 Statistical Significance Tests

In cases in which the differences in the above metrics are not significant, a statistical significance test can be performed. The Wilcoxon test [22] (also known as the signed-rank or matched-pairs test) is one of the most renowned alternatives to the Student's t-test for dependent samples. It belongs to the group of non-parametric tests and is used to compare two (and only two) dependent groups that involve two measurement variables.

The Wilcoxon test is employed when the assumptions for the Student's t-test for dependent samples are not valid. For this reason, it is considered an alternative to the latter test. The Wilcoxon test is additionally used when variables are measured on an ordinal scale (in the Student's t-test, the variables must be measured on a quantitative

scale). The requirement for Wilcoxon test application is the potential to rank differences between the first and second variable (the measurement). On an ordinal scale, it is possible to calculate the difference in levels between two variables; therefore, the test can be used for variables calculated on such a scale. In the case of quantitative scales, this test is used if the distributions of these variables are not close to the normal distribution.

Hypotheses for the Wilcoxon test are formulated as:

$$H_0 : F_1 = F_2,$$

$$H_1 : F_1 \neq F_2.$$

In this test, as in the case of the Student's t-test, a third variable is used. The third variable specifies the absolute value of the difference between the values of the paired observations. It involves ranking measurement differences for subsequent observations. First, the differences between measurements 1 and 2 are calculated. Then, the differences are ranked (the results are arranged from lowest to highest), and subsequent ranks are assigned to them. The sum of the ranks is then calculated for differences that are negative and those that are positive (results showing no differences are not significant here). Subsequently, the larger sum (of negative or positive differences) is chosen. This result constitutes that of the Wilcoxon test statistic if the number of observations does not exceed 25.

For larger samples, it is possible to use the asymptotic convergence of the test statistic (assuming that H0 is true) for the normal distribution N(m, s), where

$$m = \frac{n(n+1)}{4},$$

$$\mathrm{s} = \sqrt{\frac{\mathrm{n(n+1)(2n+1)}}{24}}.$$

The Wilcoxon test is also known as the signed-rank test because it requires calculation of ranks assigned to different signs (negative and positive differences). As with the Student's t-test for dependent samples, data missing from one measurement eliminates the entire observation from the analysis. Only the observations measured for the first and second time are considered for the analysis. This is clearly because it is necessary to subtract one result from the other.

5 Results and Conclusions

TED data comprise a unique lecture domain; however, this domain is not as wide as that of the movie subtitles corpus OpenSubtitles (OPEN). An SMT system most effectively operates in a uniquely defined domain, which presents another challenge for the system. If the challenge is not adequately addressed, it can decrease the translation accuracy. The domain adaptation quality largely depends on the training data used to

optimize the language and translation models in the SMT system. This challenge can be addressed by selecting and extracting domain-centric training data from a general corpus and a generalized domain monolingual corpus. The quasi-in-domain sub-corpus is produced through this process.

In this study, experiments were conducted on the Polish–English language pair. The corpora statistics are shown in Table 1. In the Polish (PL) and English (EN) columns, the number of unique words is given for each language; the number of bilingual sentences is given in the "PAIRS" column.

Table 1. Corpora statistics

CORPORA	PL	EN	PAIRS
TED	218,426	104,117	151,228
OPEN	1,236,088	749,300	33,570,553

In Table 2, the corpora statistics are presented for the average sentence lengths for each language and corpus. Both tables expose large disparities between the text domains.

Table 2. Average sentence lengths

CORPORA	PL	EN
TED	13	17
OPEN	6	7

Multiple versions of the SMT system were evaluated through the experiments. Using the Moses SMT system, we trained a baseline system with no additional data (BASE), a system that employs additional subtitle corpora with no adaptation (NONE), a system adapted using Moore–Lewis filtering (MML) [2] built into Moses, a system using tf-idf adaptation (TF-IDF), a system using perplexity-based adaptation (PP), a system using data selected by the Levenshtein distance (LEV), and, lastly, a system combining the three methods as described in Sect. 3.4 (COMB). In Table 3, we present the amount of data from the OPEN corpus that remained after each filtration method.

Table 3. Number of remaining bi-sentences after filtration

Filtration method	Number of bi-sentences
NONE	33,570,553
MML	1,320,385
TF-IDF	1, 718,231
PP	2,473,735
LEV	1,612,946
COMB	983,271

Additional data were used for training both the bilingual translation phrase tables and language models. The Moses SMT system was used for tokenization, cleaning, factorization, conversion to lower case, splitting, and final cleaning of corpora after splitting. Training of a 6-gram language model was accomplished using the KenLM Modeling Toolkit [17]. Word and phrase alignment was performed using the SyM-GIZA++ tool [23]. Out-of-Vocabulary (OOV) words were addressed using an unsupervised transliteration model [24]. For evaluation purposes, we used an automatically calculated BLEU metric [25] and official International Workshop on Spoken Language Translation (IWSLT) 2012 test datasets[2]. The results are shown in Table 4. Statistically significant results in accordance with the Wilcoxon test are marked with an asterisk '*'; those that are very significant are denoted with '**.'

Table 4. Corpora adaptation results

SYSTEM	BLEU	
	PL → EN	EN → PL
BASE	17.43	10.70
NONE	17.89*	10.63*
MML	18.21**	11.13*
TF-IDF	17.92*	10.71
PP	18.13**	10.88*
LEV	17.66*	10.63*
COMB	18.97**	11.84**

As shown by Table 4, ignoring the adaptation step only slightly improves PL EN translation and degrades EN ← PL translation. As anticipated, other adaptation methods have a rather positive impact on translation quality; however, in some cases, the enhancement is only minor.

The most significant improvement in translation quality was obtained using the proposed method combining all three metrics. It should be noted, however, that the proposed method was not computationally feasible in some cases, even though it produced satisfactory results. In the best-case scenario, fast comparison metrics, such as perplexity, will filter most irrelevant data; however, in the worst-case scenario, most data would be processed by slow metrics.

Summing up, we successfully introduced a new combined approach for the in-domain data adaptation task. In the general case, it provides better adaptation results than those of state of the art methods separately in a reasonable amount of time.

[2] iwslt.org

References

1. Brown, P., Pietra, V., Pietra, S., Mercer, R.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**, 263–311 (1993)
2. Axelrod, A., He, X., Gao, J.: Domain adaptation via pseudo in-domain data selection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), pp. 355–362. Association for Computational linguistics, Stroudsburg (2011)
3. Daumé III, H., Jagarlamudi, J.: 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 (ACL-HLT 2011), pp. 407–412. Association for Computational Linguistics, Stroudsburg (2011)
4. Koehn, P., Haddow, B.: Towards effective use of training data in statistical machine translation. In: Proceedings of the 7th ACL Workshop on Statistical Machine Translation, pp. 317–321. Association for Computational Linguistics, Stroudsburg (2012)
5. Civera, J., Juan, A.: Domain adaptation in statistical machine translation with mixture modelling. In: Proceedings of the 2nd ACL Workshop on Statistical Machine Translation, pp. 177–180. Association for Computational Linguistics, Stroudsburg (2007)
6. Foster, G., Kuhn, P.: Mixture-model adaptation for SMT. In: Proceedings of the 2nd ACL Workshop on Statistical Machine Translation, pp. 128–136. Association for Computational Linguistics, Stroudsburg (2007)
7. Eidelman, E., Boyd-Graber, J., Resnik, P.: Topic models for dynamic translation model adaptation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers (ACL 2012), vol. 2, pp. 115–119. Association for Computational Linguistics, Stroudsburg (2012)
8. Matsoukas, S., Rosti, A., Zhang, B.: Discriminative corpus weight estimation for machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), vol. 2, pp. 708–717. Association for Computational Linguistics, Stroudsburg (2009)
9. Hildebrand, A.S., Eck, M., Vogel, S., Waibel, A.: Adaptation of the translation model for statistical machine translation based on information retrieval. In: Proceedings of EAMT 10th Annual Conference, Budapest, Hungary, 30–31 May 2005, pp. 133–142. Association for Computational Linguistics, Stroudsburg (2005)
10. Lü, Y., Huang, J., Liu, Q.: Improving statistical machine translation performance by training data selection and optimization. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), pp. 343–350. Association for Computational Linguistics, Stroudsburg (2007)
11. Lin, S., Tsai, C., Chien, L., Chen, K., Lee, L.: Chinese language model adaptation based on document classification and multiple domain-specific language models. In: Kokkinakis, G., Fakotakis, N., Dermatas, E. (eds.) Proceedings of the 5th European Conference on Speech Communication and Technology, pp. 1463–1466. International Speech Communication Association, Grenoble (1997)
12. Gao, J., Goodman, J., Li, M., Lee, K.: Toward a unified approach to statistical language modeling for Chinese. ACM Trans. Asian Lang. Inf. Process. **1**, 3–33 (2002). https://doi.org/10.1145/595576.595578
13. Moore, R., Lewis, W.: Intelligent selection of language model training data. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pp. 220–224. Association for Computational Linguistics, Stroudsburg (2010)

14. Koehn, P.: Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In: Proceedings of the Antenna Measurement Techniques Association (AMTA 2004), pp. 115–124. Springer, Berlin (2004)
15. Mansour, S., Ney, H.: A simple and effective weighted phrase extraction for machine translation adaptation. In: Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT 2012), pp. 193–200. Springer, Heidelberg (2012)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Workshop on Automatic Summarization (ACL 2002), pp. 311–318. Association for Computational Linguistics, Stroudsburg (2002). https://doi.org/10.3115/1073083.1073135
17. Stolcke, A.: SRILM-an extensible language modeling toolkit. Paper presented in the 7th International Conference on Spoken Language Processing, ICSLP 2002 - INTERSPEECH, Denver, Colorado, USA (2002)
18. Chen, S., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL 1996), pp. 310–318. Association for Computational Linguistics, Stroudsburg (1996). https://doi.org/10.3115/981863.981904
19. Wang, L., Wong, D., Chao, L., Lu, Y., Xing, J.: A systematic comparison of data selection criteria for SMT domain adaptation. Sci. World J. **2014**, 745485 (2014). https://doi.org/10.1155/2014/745485
20. Hovy, E.: Toward finely differentiated evaluation metrics for machine translation. Paper presented in the Proceedings of the EAGLES Workshop on Standards and Evaluation Conference, Pisa, Italy (1999)
21. Reeder, F.: Additional Mt-eval references. Technical report, International Standards for Language Engineering, Evaluation Working Group (2001)
22. Oyeka, I.C.A., Ebuh, G.U.: Modified Wilcoxon signed-rank test. Open J. Stat. **2**, 172–176 (2012). https://doi.org/10.4236/ojs.2012.22019
23. Junczys-Dowmunt, M., Szał, A.: SyMGiza++: symmetrized word alignment models for statistical machine translation. In: Proceedings of the 2011 International Conference on Security and Intelligent Information Systems, pp. 379–390. Springer, Heidelberg (2002). https://doi.org/10.1007/978-3-642-25261-7_30
24. Durrani, N., Koehn, P., Hoang, H., Sajjad, H.: Integrating an unsupervised transliteration model into statistical machine translation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 148–153. Association for Computational Linguistics, Stroudsburg (2014). https://doi.org/10.3115/v1/e14-4029
25. Wołk, K., Marasek, K.: Enhanced bilingual evaluation understudy. Lect. Notes Inf. Theory **2**, 191–197 (2014). https://doi.org/10.12720/lnit.2.2.191-197