THE EFFECT OF PARALLEL CORPUS QUALITY VS SIZE IN ENGLISH-TO-TURKISH SMT

Eray Yıldız[1]Ahmed Cüneyd Tantuğ[2]and Banu Diri[3]

[1]Department of Computer Engineering,
Yildiz Technical University, Istanbul, Turkey
yildizeray@hotmail.com.tr
[2]Computer and Informatics Faculty,
Istanbul Technical University, Istanbul, Turkey
tantug@itu.edu.tr
[3]Department of Computer Engineering,
Yildiz Technical University, Istanbul, Turkey
banu@ce.yildiz.edu.tr

ABSTRACT

A parallel corpus plays an important role in statistical machine translation (SMT) systems. In this study, our aim is to figure out the effects of parallel corpus size and quality in the SMT. We develop a machine learning based classifier to classify parallel sentence pairs as high-quality or poor-quality. We applied this classifier to a parallel corpus containing 1 million parallel English-Turkish sentence pairs and obtained 600K high-quality parallel sentence pairs. We train multiple SMT systems with various sizes of entire raw parallel corpus and filtered high-quality corpus and evaluate their performance. As expected, our experiments show that the size of parallel corpus is a major factor in translation performance. However, instead of extending corpus with all available "so-called" parallel data, a better translation performance and reduced time-complexity can be achieved with a smaller high-quality corpus using a quality filter.

KEYWORDS

Machine Translation, Machine Learning, Natural Language Processing, Parallel Corpus, Data Selection

1. INTRODUCTION

A corpus which is comprised of aligned sentences that are translations of each other, namely a parallel corpus is an essential training data for statistical machine translation (SMT) [3]. Also a parallel corpus is useful for other natural language processing applications such as cross-language information retrieval, word disambiguation and annotation projection. Building a corpus that includes vast amount of parallel sentences is one of the most time-consuming and important works for a high-performance SMT system [10]. Training the translation model component in SMT requires large parallel corpora for the parameters to be estimated [24]. Therefore, higher translation accuracy can be achieved when machine translation systems are trained on increasing amounts of training data [12]. The quality of an SMT system output extremely depends on the

Natarajan Meghanathan et al. (Eds) : ICCSEA, SPPR, VLSI, WiMoA, SCAI, CNSA, WeST - 2014
pp. 21–30, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4710

quality of sentence pairs. A small portion of the available parallel text collections are generated naturally within multi-language organizations and governments (like UN, European Parliament, Canada) for a limited set of languages. The manual compilation of a parallel corpus is too expensive, so most of available parallel corpora are generated automatically. Automatic methods for compiling parallel sentence pairs are imprecise and using such a low-quality training corpus that has many non-parallel sentence pairs would cause low quality translations [24]. The noise in an automatically generated corpus might be due to any difference between the contents of source and target documents, non-literal translations or sentence alignment mistakes. It is infeasible to manually eliminate alignment errors in a large parallel corpus; therefore, an automatic evaluation method is desirable to determine if a parallel sentence pair is accurate or not [14]. Such an automatic evaluation method can depend on multiple factors such as overlapping words, sentence lengths, grammar correctness, fluentness, and word usage correctness.

This study focuses on the effect of using a large parallel text versus using a filtered high-quality corpus. The experiments are carried out for an English-to-Turkish SMT task.

Section 2 gives brief information about related previous studies while section 3 is devoted to the details for our training data. Section 4 introduces our method for filtering parallel sentence pairs and experimental results. The final section includes conclusions and future work.

2. RELATED WORK

The process of filtering out non-parallel sentence pairs is considered as a post-processing step of bilingual data mining process. Gale and Church [7] measure the rate of lengths between two bilingual sentences. The length based approaches work remarkably well on language pairs with high correlation in sentence lengths, such as French and English. On the other hand, the performance of a length based sentence aligner significantly decreases for the language pairs with low correlation on length, such as Chinese and English [28]. Chen and Nie [4] build a system in order to search the web for gathering English-Chinese parallel data and clean their gathered data using sentence length features and language detecting methods. Resnik and Smith [21] also build a system for mining parallel text on the web. This system introduces a translation similarity score for data cleaning. The similarity score is symmetric word-to-word model, which controls whether the translation equivalents of the word in the source sentence occur in the target sentence.

Khadivi and Ney [29] develop a rule based algorithm to filter a parallel corpus by using length constraints and translation likelihood measure for the given sentence pairs. In this work, the sentence pairs in a noisy corpus generated from European Union Web Site are reordered so that the noisy sentence pairs are moved to end of the corpus. They report 47.2 BLEU score [19] when the SMT system is trained with the clean corpus (top 97.5% of corpus), whereas the translation score is 46.8 when the system in trained with the entire corpus.

Yasuda et al. [27] develop a training set selection method for an English-Chinese SMT system using a perplexity score. The perplexity score calculated for each sentence by using language model probabilities and the word counts of the sentences. The perplexity of a parallel sentence pair is considered as the geometric means of the perplexities of source and target sentences. They train a SMT system with an initial in-domain corpus and the selected translation pairs whose perplexities are smaller than a threshold. Their method improves BLEU score by 1.76% in comparison with an initial in-domain corpus usage.

Liu and Zhou [14] propose a machine learning based method for evaluating the alignment quality. Linguistic features and constructs extracted from the sentences are unified together into a feature vector to characterize different dimensions of the alignment quality. They use support vector

machines (SVM) to discriminate high-quality and low-quality parallel sentence pairs. The features in this study are the number of misspelled words, language model score for each English sentence, the number of unlinked words provided by a link grammar parser and the translation equivalence measure. The word alignments are obtained by using a comprehensive bilingual dictionary and the word alignment counts are normalized by the sentence lengths to get equivalence measurement. They train their SVM classifier on a 40K training set that is checked by bilingual annotators. Their classifier is reported to have 0.88 precision and recall rates. Taghipour et al. [24] also proposed a classification method for cleaning parallel data. Many features have been tested and used to build the models such as translation probabilities based on IBM translations models [3], the number of the null aligned words, length based features and features based on language model. They chose maximum entropy model for the classification process and they achieved 98.3% accuracy on a 48K Farsi-English parallel corpus expanded with artificial noise.

In the study of Cui et al. [5], an unsupervised method - namely random walk algorithm - is conducted to iteratively compute the importance score of each sentence pair that indicates its quality. Their method utilizes the phrase translation probabilities based on Maximum Likelihood Estimation. They tested their method on various Chinese-English parallel corpus and eventually, their method improves system performance about 0.5~1.0 BLEU.

Munteanu and Marcu [17] train a maximum entropy classifier in order to extract parallel data from large Chinese, Arabic and English non-parallel newspaper corpora. They start from a parallel corpus of 5K sentence pairs to compute word alignments and extract feature values. Their classifier uses the following features: lengths of sentences as well as the length difference and length ratio, percentage of words on each side that have a translation on the other side and the alignment features obtained from word alignment models. They use dictionaries learned from initial parallel corpus whose sizes are 100K, 1M, 10M, 50M and 95M tokens. The precision and recall values of their classifier are 0.97 and 0.45 respectively with the data from Arabic-English in-domain corpus whereas they have 0.94 precision and 0.67 recall with the Arabic-English out-of-domain corpus.

Hoang et al. [10] present a model for extracting parallel sentence pairs from non-parallel corpora resources in different domains. They combine length-based filtering, cognate condition and content similarity measurement. Their content based similarity is the similarity between the target sentence and the translation of the source sentence obtained by a SMT system. Initially they train their SMT system with an initial out-of-domain corpus (50K) and use this system for content based similarity. Afterwards, they filter an in domain comparable corpora (958K). If the candidate sentence pairs pass the condition of similarity measurement, they expand their train data with these new parallel sentence pairs. They retrain the SMT system and repeat the process of expanding parallel corpus as well as improving the SMT system. They expand 50K initial parallel corpus to 95K after 5 iterations and achieve to increase translation BLEU scores from 8.92 to 24.07.

A somehow related approach is Schwenk's [22] lightly supervised training method which is used to translate large amounts of monolingual texts, to filter them and add them to the translation model training data. Another approach that assigns weights to each sentence in the training bitext is proposed in [16]. Foster et al. [8] describe a new approach to SMT adaptation that weights out-of-domain phrase pairs according to their relevance to the target domain. Axelrod et al. [1] propose a method for extracting sentences from a large general-domain parallel corpus by a domain-relevant data selection method.

3. TRAINING DATA

We utilized an English-Turkish parallel corpus of one million sentences compiled from various sources with varying quality levels. This corpus contains news text [26], literature text [23], subtitles text [25] and web crawled text [28]. Although the sentences in these sources are supposed to be parallel, a quick inspection of these data sets exposes the existence of serious sentence alignment issues. Parallel sentences pairs from each source are pre-processed in order to overcome the alignment problems and to maintain a basic level of alignment quality. A lexicon based sentence aligner; Champollion [15] is used for a crude elimination of the sentence pairs that seem to be misaligned. Table 1 shows the number of sentences in each corpus after this filtering.

Table 1 Training dataset after basic filtering

Dataset	**Sentences**
News	200K
Subtitles	1,200K
Web	150K
Literature	680K
Total	**2,230K**

4. EXPERIMENTS

Having parallel data described in the previous section, our first investigation is about the quality of each source. In order to alleviate the impact of the dataset size on the translation quality, a fair comparison setup is ensured by generating equal sized versions of the datasets. These clipped versions of datasets include only top 150K parallel sentences from each dataset. Likewise, 1000 parallel sentences are randomly drawn from the remaining parts of these datasets to generate a balanced test set of 4K sentences.

Each 150K bilingual corpus is used to train a separate SMT system using MOSES toolkit [30]. In Table 2, we present the translation performance of each system.

Table 2 Evaluation of SMT systems trained with 150K datasets for each corpus

Training Data	**Training Data Size**	**BLEU Score**
$News_{150K}$	150K	16.59
$Subtitles_{150K}$	150K	4.72
Web_{150K}	150K	19.56
$Literature_{150K}$	150K	9.49

As seen in the table above, there are significant discrepancies in the translation quality. One of the reasons for this result is the varying alignment quality among the datasets. A detailed observation inside the datasets reveals the fact that, even after the first coarse filtering, these "so-called" parallel datasets still contain misaligned pairs or pairs suffering from poor translation quality. Though the performance of a SMT system can be improved by incorporating larger monolingual data for the language model (LM) component, it can be also a good idea to filter out low quality sentence pairs to improve the translation model (TM). Moreover, it is obvious that an effective filtering substantially reduces the time-consuming training phase of SMT even if it does not contribute in translation accuracy. So, a classifier that can discriminate the high quality sentence pairs from low quality ones in the parallel corpus is considered as a beneficial pre-processing step.

4.1 Building The Quality Classifier

For each candidate sentence pair in the corpus, we need a reliable way of deciding whether the two sentences in the pair are the proper translations of each other. We extract some features that are considered as quality indicators from the sentence pairs so that an automatic classifier can be trained. A pair quality depends on not only the correctness of alignment, but also the grammar correctness, fluentness and word usage correctness [14].

An ungrammatical sentence may cause degradations in the translation model, so spelling attributes of a sentence play an important role for evaluating a sentence's quality. Therefore, we opt for using a spell checker [11] to calculate the number of misspellings in English side only, as a representation feature.

Another feature which represents the grammatical correctness is based on a language model. The probability is calculated from the language model for each English sentence and it is used as a fluentness indicator feature. Since the probability of a sentence decreases as its length increase, we introduce the sentence length as a feature also. The BerkeleyLM toolkit [20] is used for constructing an English language model created from the Web 1T Corpus and assigning probabilities to sentences.

The relation between the lengths of the sentences in the pair is the most common feature for many parallel text applications such as sentence alignment. Hence, we selected the lengths of the sentences, as well as the length differences and length ratio as features.

The last feature is the percentage of words on each side that have a valid translation on the other side according to a dictionary. An electronic English-Turkish bilingual dictionary which contains Turkish equivalents of 88,824 English words is used in conjunction with a naïve stemmer for Turkish. This stemmer assumes the first 5 letters of a word form as the lemma, which is linguistically incorrect but sufficiently effective for lookup purposes [28].

In order to train and evaluate a classifier, a train and a test set is required respectively. The train set is generated by sentence pairs randomly selected from the whole corpora and then manually labelled as high-quality or poor-quality. 983 instances are manually labelled as high-quality while 160 instances are manually labelled as poor-quality. These 160 instances labelled as poo-quality are not completely useless instances; their low quality might be due to non-literal translations, grammatical or spelling errors. Also, additional 674 misaligned sentence pairs are generated by randomly matching non-parallel sentences from both sides. These artificially generated noise instances are auto-labelled as poor-quality and joined with the other instances in the train set. Table 3 depicts the detailed information about our training data.

We have employed several experiments with a number of different machine learning algorithms to build our classifier. The WEKA tool [9] is used to train Radial Basis Function (RBF) Network, Random Forest (RF), Multilayer Perceptron and Support Vector Machine (SVM) based classifiers. Table 4 shows the classification results in terms of micro averaged precision, recall and F1 values. All tests are run in a 10-fold cross validation evaluating process.

Table 3: Training Data for Classifiers

Sentence Pair Quality	**Count**
High-Quality	983
Poor-Quality	
Auto-Labeled Artificial Noise	674
Manually Labeled From Corpus	160
Total	**1817**

Table 4: Cross Validation Results of Classification Algorithms

Classifier	**Precision**	**Recall**	**F1**
RBF Network	0.903	**0.964**	0.933
Random Forest	**0.969**	0.960	**0.965**
Multilayer Perceptron	0.938	0.953	0.946
SVM	0.932	0.963	0.947

For our filtering purposes, the classification algorithm that yields the best precision score sounds reasonable because our aim is to guarantee that the filtered corpus only contains high-quality pairs as much as possible, with the cost of leaving out some good pairs. The results show that RF algorithm is the most efficient model for our task. RF is a tree-based ensemble classifier that consists of a number of decision tree classifiers on various sub-samples of the dataset. [2]. Consequently, we have preferred RF as the classification method to produce a higher quality corpus.

It could be said that the classification of the poor quality instances that are manually labelled is harder than the artificial poor quality instances. In order to see the effectiveness of the RF classifier on these 160 instances, 80 manually labelled instances are excluded from training data and used as test instances. The RF classifier labelled 76 instances correctly which means 0.95 accuracy.

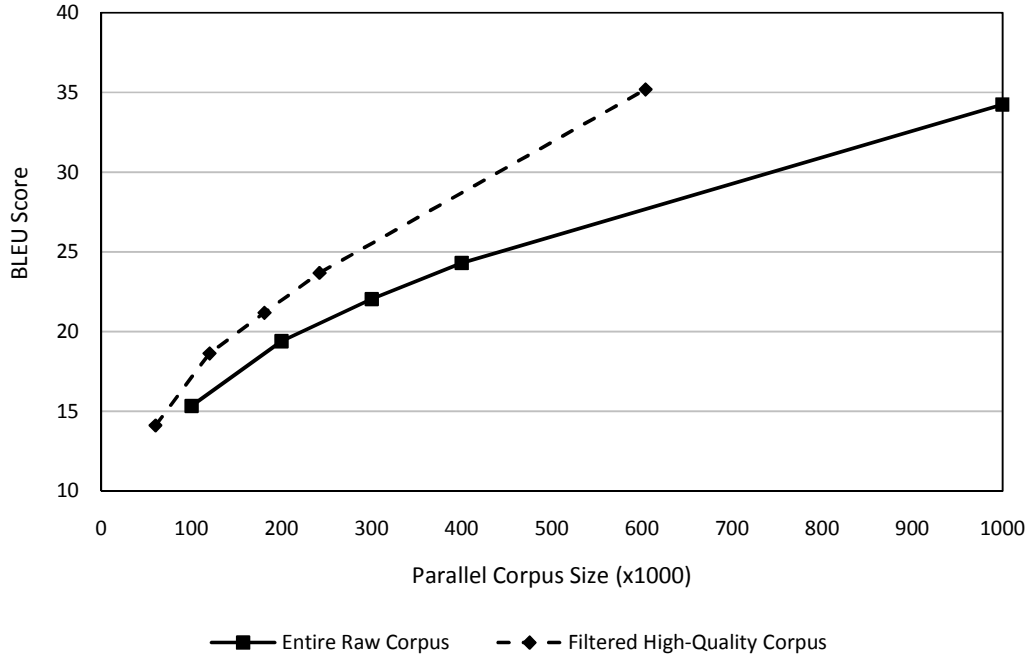


Figure 1 Performance of the SMT system with respect to parallel corpus size and parallel corpus quality

4.2 Experimental Results

Several experiments are conducted with our entire raw parallel corpora and filtered high-quality parallel corpus. To see the impact of the parallel corpus size, we run tests with different corpus sizes up to 1M sentences. The text in the Turkish side of corpus is used for language model training and 10% of each corpus is used as test set. The translation BLEU scores of the SMT outputs are plotted in Figure 1.

As expected, the argument that states the more data, the higher SMT accuracy is proven to be valid. When the number of parallel sentences is increased from 100K to 1000K, the BLEU score increases from 15.33 to 34.24 for the entire raw corpus. The tendency in the plot shows that higher BLEU scores can be achieved with the introduction of more parallel data than 1000K.

On the other hand, Figure 1 reveals another important fact that better translation performance can be acquired by less but high-quality training data. By using only 60% of the raw training data, a BLEU score of 35.19 (which means 2.77% relative improvement) can be achieved with a quality classifier based filtering. The reduction in the training data size also leads to reduction of time complexities of training phase of an SMT system which necessitates a complex and time-consuming process of days or weeks long.

5. CONCLUSIONS

This paper discusses the issue of the quality of the bilingual corpus in SMT. We propose the use of a Random Forest classifier to classify sentence pairs as ‘high-quality and ‘poor-quality’ with proposed features based on translation equivalence and grammatical correctness. The experiments show that our filtering method can be used for extracting unsuitable pairs from a noisy parallel

corpus and the remaining pairs can still be effective in achieving results as high as the results of the entire raw corpus. Our results also indicate that there is still need for more parallel data.

The availability of parallel resources in English-Turkish pair is much more limited than the availability of the language pairs that have reliable and big resources such as EuroParl Corpus [13]. Therefore, adding more parallel data is desired, however, the unreliable resources such as Wikipedia and movie subtitles should not be added directly as the training data for a SMT system.

Our filtering method is useful in effective incorporation of these resources in SMT process.
Although we presented our results on English-to-Turkish SMT task specifically, the notions in this study can be easily extended for any language pair.

We plan to use this filtering method in our efforts to build up a large-coverage and high-quality English-Turkish parallel corpus.

ACKNOWLEDGEMENTS

We would like to thank to all members of Istanbul Technical University, Natural Language Processing Research Group and our all labmates especially Ezgi, İsmail, Sami. We are also grateful to Natural Language ProcessingWorkgroup in Yildiz Technical University.

REFERENCES

[1] Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. (2011) "Domain adaptation via pseudo in-domain data selection."Proceedings of the Conference on Empirical Methods in Natural Language Processing.Association for Computational Linguistics.
[2] Breiman, Leo. (2001) "Random forests." Machine learning 45.1 : 5-32.
[3] Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin. (1990) "A statistical approach to machine translation", Computational Linguistics, 16 (2):79–85.
[4] Chen, Jiang and Jian-Yun Nie. (2000) "Parallel Web text mining for cross-language information retrievalal", In Recherché d'Informations Assist ´ee par Ordinateur (RIAO), pages 62–77, Paris.
[5] Cui, Lei, Dongdong Zhang, Shujie Liu, Mu Li and Ming Zhou. (2013) Bilingual data cleaning for smt using graph-based random walk.In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 340-345).
[6] Esplá-Gomis, Miquel. (2009) "Bitextor, a free/open-source software to harvest translation memories from multilingual websites." Proceedings of MT Summit XII, Ottawa, Canada.Association for Machine Translation in the Americas.
[7] Gale, William A. and Kenneth W. Church.(1993) "A program for aligning sentences in bilingual corpora", Comput. Linguist., 19:75–102
[8] Foster, George, Cyril Goutte, and Roland Kuhn. (2010) "Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation", In: Proceedings of the Conference on Empirical Methods in Natural Lan-guage Processing (EMNLP), Cambridge, US-MA, pp 451--459
[9] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten.(2009) "The weka data mining software: An update", ACM SIGKDD explorations newsletter 11.1 : 10-18.
[10] Hoang, Cuong, Nguyen Phuong Thai, and Ho Tu Bao. (2012) "Exploiting non-parallel corpora for statistical machine translation", In Proceedings of The 9th IEEE-RIVF International Conference on Computing and Communication Technologies, pages 97 – 102. IEEE Computer Society.
[11] Idzelis Mindaugas. (2005) "Jazzy: The java open source spell checker", http://jazzy.sourceforge.net/
[12] Koehn, Philipp. (2002) "Europarl: A multilingual corpus for evaluation of machine translation", Information Sciences Institute, University of Southern California.
[13] Koehn, Philipp. (2005) "EuroParl: A Parallel Corpus for Statistical Machine Translation", Machine Translation Summit 2005. Phuket, Thailand.

[14] Liu, Xiaohua, and Ming Zhou. (2010) "Evaluating the quality of web-mined bilingual sentences using multiple linguistic features", Asian Language Processing (IALP), 2010 International Conference on. IEEE, 2010

[15] Ma, Xiaoyi. (2006) "Champollion: A Robust Parallel Text Sentence Aligner", LREC 2006: The Fifth Interna-tional Conference on Language Resources and Evaluation

[16] Matsoukas, Spyros, Antti-Veikko I. Rosti, and Bing Zhang. (2009) "Discriminative Corpus Weight Estimation for Machine Translation", In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore : 708--717

[17] Munteanu, Dragos Stefan, and Daniel Marcu. (2005) "Improving machine translation performance by exploiting non-parallel corpora", Computational Linguistics 31.4 : 477-504.

[18] Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. (1993) "The mathematics of statistical machine translation: Parameter estimation", Computational linguistics, vol. 19, 1993 : 263– 311.

[19] Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. (2002) "BLEU: A Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics :311-318

[20] Pauls, Adam, and Dan Klein. (2011) "Faster and smaller n-gram language models", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics,, Portland, Oregon.

[21] Resnik, Philip, and Noah A. Smith. (2003) "The web as a parallel corpus", Computational Linguistics 29.3 (2003): 349-380.

[22] Schwenk, Holger. (2008) "Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation", In Proc. of the International Workshop on Spoken Language Translation

[23] Taşçı, Şerafettin, A. Mustafa Güngör, and Tunga Güngör. (2006) "Compiling a Turkish-English Bilingual Corpus and Developing an Algorithm for Sentence Alignment", International Scientific Conference Computer Science

[24] Taghipour, Kaveh, Nasim Afhami, Shahram Khadivi and Saeed Shiry. (2010) "A discriminative approach to filter out noisy sentence pairs from bilingual corpora", Telecommunications (IST), 5th International Symposium on 2010 : 537-541.

[25] Tiedemann, Jörg. (2009) "News from opus - a collection of multilingual parallel corpora with tools and inter-faces", In Recent Advances in Natural Language Processing, volume V, Amsterdam/Philadelphia

[26] Tyers, Francis M., and Murat Serdar Alperen. (2010) "SETimes: a parallel corpus of Balkan languages", In: Proceedings of the multiLR workshop at the language resources and evaluation conference, LREC2010, Malta, pp 49–53

[27] Yasuda, Keiji, Ruiqiang Zhang, Hirofumi Yamamoto and Eiichiro Sumita. (2008) "Method of Selecting Training Data to Build a Compact and Efficient Translation Model", In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Hyderabad, India

[28] Yıldız, Eray, Tantuğ, A.Cüneyd. (2012) "Evaluation of Sentence Alignment Methods for English-Turkish Par-allel Texts", LREC 2012: The International Con-ference on Language Resources and Evaluation. Istanbul

[29] Khadivi, Shahram, and Hermann Ney. (2005) "Automatic filtering of bilingual corpora for statistical machine translation."Natural Language Processing and Information Systems. Springer Berlin Heidelberg : 263-274.

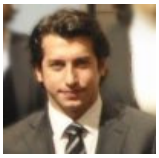
[30] Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, ... and Evan Herbst . (2007) "Moses: Open source toolkit for statistical machine translation", In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions : 177-180.

AUTHORS

Eray YILDIZ
Graduate Student in Yildiz Technical University, Computer Engineering Department, Istanbul, Turkey B.Sc. : Computer Engineering, Kocaeli University, 2011

Ahmed Cüneyd TANTUĞ

Assistant Professor, Istanbul Technical University Computer Engineering Department, Turkey.
Ph.D. : Computer Engineering, Istanbul Technical University, 2007
M.Sc. : Computer Engineering, Istanbul Technical University, 2002
B.Sc. : Control & Computer Engineering, Istanbul Technical University, 2000



Banu DİRİ
Associate Professor, Yildiz Technical University, Computer Engineering Department , Istanbul, Turkey
Ph.D. : Computer Engineering, Yildiz Technical University, 1999
M.Sc. : Computer Sciences Engineering, Yildiz University, 1990
B.Sc. : Computer Sciences Engineering, Yildiz University, 1987