Expanding Machine Translation Training Data with an Out-of-Domain Corpus using Language Modeling based Vocabulary Saturation

Burak Aydın burak.aydin@tubitak.gov.tr
TÜBİTAK-BİLGEM, Gebze 41470, KOCAELİ, TURKEY
Department of Computer Engineering, Boğaziçi University,
Bebek 34342, İSTANBUL, TURKEY

Arzucan Özgür arzucan.ozgur@boun.edu.tr
Department of Computer Engineering, Boğaziçi University,
Bebek 34342, İSTANBUL, TURKEY

Abstract

The training data size is of utmost importance for statistical machine translation (SMT), since it affects the training time, model size, decoding speed, as well as the system's overall success. One of the challenges for developing SMT systems for languages with less resources is the limited sizes of the available training data. In this paper, we propose an approach for expanding the training data by including parallel texts from an out-of-domain corpus. Selecting the best out-of-domain sentences for inclusion in the training set is important for the overall performance of the system. Our method is based on first ranking the out-of-domain sentences using a language modeling approach, and then, including the sentences to the training set by using the vocabulary saturation filter technique. We evaluated our approach for the English-Turkish language pair and obtained promising results. Performance improvements of up to +0.8 BLEU points for the English-Turkish translation system are achieved. We compared our results with the translation model combination approaches as well and reported the improvements. Moreover, we implemented our system with dependency parse tree based language modeling in addition to the n-gram based language modeling and reported comparable results.

1 Introduction

Most of the statistical methods that attempt to solve natural language processing problems achieve better results with increasing training data sizes. In statistical machine translation, the amount of data directly affects a system's overall success. Increasing the size of the training data by effectively utilizing the data available in other domains (e.g. web, news, medical) using domain adaptation and data selection techniques is a promising research direction for improving the performance of an SMT system. This is especially important for low-resource languages and domains, for which there are only limited amounts of training data available. The English-Turkish language pair is an example low-resource language pair for machine translation. Most of the publicly available corpora for this language pair contain only thousands of sentences, whereas the training sets of language pairs with more resources (e.g. English-French) usually contain millions of sentences. The number of parallel sentences in the training set for English-Turkish hardly reaches to millions, even when all available corpora from different domains are combined.

In this paper, we introduce an approach that effectively combines different data selection methods for expanding in-domain training data with the available out-of-domain data in statistical machine translation. The method first scores the sentences in the out-of-domain corpus based on their similarities to the in-domain corpus using a language modeling approach. Then, it adapts the vocabulary saturation filter technique, which has recently been proposed in (Lewis and Eetemadi, 2013) for reducing the training data and model sizes, to the domain adaptation problem. The proposed approach is applied to English-Turkish machine translation by using n-gram based as well as dependency parse tree based language modeling and improvements in terms of BLEU scores are achieved.

The paper is organized as follows. Section 2 presents the related work on data selection and domain adaptation in SMT. Section 3 briefly explains the Vocabulary Saturation Filter algorithm and the proposed approaches. Section 4 describes the data and the experimental setup, and provides the obtained results. Section 5 discusses the results of the study, and Section 6 outlines possible future directions for research.

2 Related Work

In statistical machine translation, several approaches have been proposed for data selection, domain adaptation, and data preprocessing and cleaning. Eck et al. (2005) selected a subset of a monolingual corpus and human-translated it to use for a low-resource language pair. The aim for selecting a suitable subset is not only for decreasing the model size, but also for improving the translation quality as in the work of Okita (2009).

Data selection and preprocessing methods generally aim to be successful in reducing data and model size significantly with a minimum score loss (Lewis and Eetemadi, 2013). Domain adaptation techniques on the other hand, target not only to optimize the data and model size, but also to improve system score. A number of different approaches including language modeling (Bulyko et al., 2007) and source-sentence classification (Banerjee et al., 2010) have been proposed for domain adaptation. Machine translation systems mostly work well only in one domain and domain adaptation techniques usually improve the score of a system in that domain. Wang et al. (2012) attempted to build a system that works well in multiple-domains simultaneously. Their method tries to use models of different domains in a combined system and automatically detects the domain and its parameters at runtime. Axelrod et al. (2011) compared different data selection methods by selecting subsets from a large general domain parallel corpus and proposed a new data selection method, which uses bilingual cross-entropy difference. In addition to the usage of conventional n-gram language models, Duh et al. (2013) used neural language models to select training data from general domain.

In some circumstances, there may be lack of in-domain bilingual data. Wu et al. (2008) used out-of-domain corpora to train a baseline system and then used in-domain translation dictionaries and in-domain monolingual corpora to improve the in-domain performance. Their method unifies old and newly produced resources in a combined framework.

Moore and Lewis (2010) tried to solve the efficient data selection problem in the language model training step, which is an essential feature of statistical machine translation systems. In this work, they compared the cross-entropy according to domain-specific and non-domain-specific language models for each sentence that is used to produce the non-domain-specific model. They calculated the cross-entropy difference to select the data. Using this approach, they produced better language models to use in their machine translation system.

Bertoldi and Federico (2009) attempted to significantly improve the performance of machine translation by exploiting large monolingual in-domain data. They synthesized a bilingual corpus by translating the monolingual adaptation data. Their work is based on adapting an already developed translation system into another domain in which there is no enough parallel data available.

Dependency parsing based language modeling has also been investigated in many studies. Shen et al. (2008) built a framework to employ a target dependency language model (DLM) for machine translation. It predicts the next child based on the previous children of the current head. DLM was used in many tasks such as sentence realisation (Guo et al., 2008), speech recognition (Lambert et al., 2013), and sentence completion (Gubbins and Vlachos, 2013). In our dependency-based language modeling approach we represent sentences with trigrams (i.e., dependent, head, and dependency type) extracted from their dependency parse trees. The out-of-domain sentences are ranked based on the dependency relation language models learned from the in-domain corpus.

Another relevant approach is the phrase table combination method proposed by Bisazza et al. (2011). They tried to expand the in-domain phrase table with an out-of domain phrase table in an efficient manner. In the fill-up method, they used the phrases from the out-of domain table only if they are not available in the in-domain table. We compared our approach with their method and also with the linear interpolation technique. Our systems obtained better BLEU scores in overall. Additionally, the system in (Bisazza et al., 2011) uses all of the available corpora for phrase table building, whereas ours uses a proportion of the out-of-domain data in addition to the in-domain data.

Recently, Lewis and Eetemadi (2013) proposed the Vocabulary Saturation Filter algorithm. The algorithm tries to significantly reduce the training data size. It starts by counting the n-grams from the beginning of the corpus and when all n-grams of a specific sentence reach to a previously defined threshold frequency t, then this sentence is excluded from the resulting subset that is to be used for machine translation system training. They showed that unigrams are enough to choose a subset, since higher order n-grams resulted in selecting the majority of the original corpus.

Our approach is an effective combination of language modeling and vocabulary saturation filtering (VSF) for expanding training data using an out-of-domain corpus. VSF has originally been used for data size reduction (Lewis and Eetemadi, 2013), but in this study we adapt it to use in expanding in-domain data with an out-of-domain corpus for machine translation. Before applying the VSF technique, we pre-rank the out-of-domain parallel corpus based on the sentence perplexities calculated using an in-domain language model. We investigate using dependency tree based language modeling as well as n-gram based language modeling. We apply the approach to the English-Turkish language pair and report promising results.

3 Method

3.1 Vocabulary Saturation Filter (VSF)

The effect of more data on improving BLEU scores is clearly observed through experiments: as more data is added, BLEU scores increase. However, the relationship between the quantity of data and BLEU is not linear, such that addition of new data does not increase much after some point. There is a saturation point of data and the VSF algorithm attempts to find it (Lewis and Eetemadi, 2013). It selects the data within a threshold and uses it for machine translation. The algorithm is very successful in reducing training data size (Lewis and Eetemadi, 2013). In this paper, we use VSF to improve translation results by expanding the training data with out-of-domain data.

3.2 N-gram Language Modeling based VSF on In-domain Data

The VSF algorithm has originally been proposed to reduce training data size with a possible loss in BLEU scores and its power has been shown for French-English translation by applying the algorithm on the sentences in the given order (Lewis and Eetemadi, 2013). In other words,

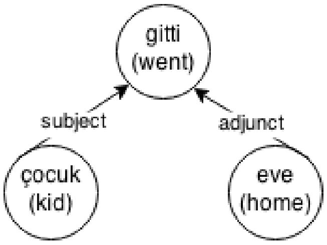

Figure 1: Dependency tree for the sentence "Çocuk eve gitti." (The kid went home.)

no any sentence sorting has been performed before applying the algorithm. The data selection with VSF starts from the beginning of the corpus, hence the sentence order makes difference for the algorithm's choice. Therefore, in this study we propose to pre-sort the sentences using a language modeling based approach before applying the VSF algorithm. The procedure for the described method can be formalized in the following steps:

- Build a language model using the target side of the parallel corpus.
- Score the target side of the parallel corpus based on the previously built language model.
- Rank the parallel corpus with respect to the sentence perplexity scores.
- Apply VSF algorithm on the ordered corpus to select a subset and use it as training data for the SMT system.

The results of the original approach and language modeling based approach are shown in Table 5 in the results section.

3.3 Dependency Language Modeling based VSF on In-domain Data

In the previous section, the pre-ranking process on the training set is applied based on n-gram language modeling. The corpus is ranked with respect to sentence perplexity scores. We proposed other sorting mechanisms than n-grams based modeling and experimented in end-to-end machine translation systems to investigate their effects. The main idea is to use the dependency relations between the words in a sentence and to identify whether these relations will provide a better ranking for the sentences in the training corpus. Notice that the data selection procedure here is almost the same as the method described in the previous section, the only difference is the sorting mechanism. We tried several combinations of these dependency relation features and reported the results. The representations extracted from the tree in Figure 1 are exemplified in Table 1. The dependency relations for the Turkish sentences are obtained using the parser developed by Eryigit et al. (2008).

After representing the training corpus with the settings exemplified in Table 1, we collected the statistics based on these modified corpus' dependency-based unigrams and sorted the corpus accordingly. For example, for System 1 in Table 1 each word_label_word relation in a sentence is treated as a unigram. Next, VSF is applied through the corpus to select a more efficient subset from the beginning. The selected corpus is used for training statistical machine translation systems and the results are compared with the baseline and the n-gram based ranked VSF systems in Table 6.

We decided to focus on the setting which gives the best BLEU score and applied it for sentence ranking when trying to utilize out-of-domain data.

System	Representation	Example
No dependency	-	çocuk eve gitti
System 1	word_label_word	çocuk_subject_gitti eve_adjunct_gitti
System 2	pos_label_pos	noun_subject_verb noun_adjunct_verb
System 3	word_label_pos	çocuk_subject_verb eve_adjunct_verb
System 4	All	Representations from system 1-3 are combined

Table 1: Dependency-based represantation of the sentence in Figure 1

3.4 Language Modeling based VSF for Out-of-domain Data

As briefly stated before, the method proposed in this paper is a combination of different data selection algorithms, namely Language Modeling (LM) and Vocabulary Saturation Filter (VSF). VSF has originally been proposed to reduce training data and model size with a minimum score loss. However, in this paper, we adapt the VSF approach to increase the training data size using out-of-domain data with the goal of improving the performance of an SMT system.

The VSF algorithm selects the training subset from the corpus by counting the seen n-grams. The algorithm starts to read the corpus from the beginning, so it is more likely that the sentences at the beginning of the corpus will be chosen by the algorithm. This affects the sentence choice. What we propose is to order the out-of-domain data with respect to a language model built from the in-domain data and select a subset from it through VSF in order to add into the in-domain data for training. The approach, whose workflow is shown in Figure 2, may be summarized as in the following steps.

- Build a language model using the target side of the in-domain parallel corpus.
- Score the sentences in the target side of the out-of-domain parallel corpus by the language model produced in the previous step.
- Rank the sentences in the out-of-domain corpus based on sentence perplexity scores.
- Apply the VSF algorithm on the sorted corpus.
- Use the selected out-of-domain corpus sentences together with the in-domain corpus sentences for training a machine translation system.

5-gram language models are built from the in-domain data after tokenization using the SRILM toolkit (Stolcke, 2002). After scoring the out-of-domain sentences with the language model learned from the in-domain data, the sentences are ranked according to their perplexity scores. Since lower perplexity scores correspond to better fitting to the applied language model, sentences with lower perplexity scores appeared at the top of the corpus. Hence, their chances to be selected by the VSF algorithm increased. We applied the VSF technique on the sorted out-of-domain sentences by using the unigrams. In other words, we counted the seen unigrams while selecting the subset from the ranked corpus, since higher order n-grams lead to the selection of almost the entire corpus.

We also ranked the out-of-domain corpus with respect to dependency-based relations investigated in the previous section. Only the best scoring system is applied for the utilization of out-of-domain data and its results are also reported. The results of all systems that utilize out-of-domain data are shown in Table 7.

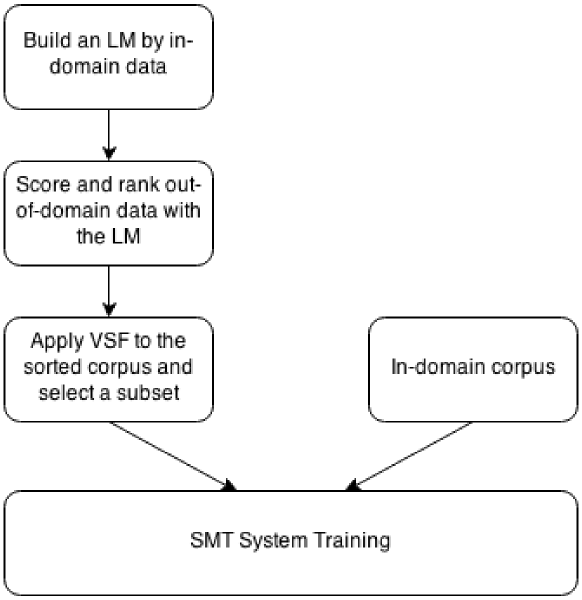
Build an LM by in-domain data
Score and rank out-of-domain data with the LM
Apply VSF to the sorted corpus and select a subset
In-domain corpus
SMT System Training

Figure 2: Workflow of the proposed approach

The proposed approach aims at increasing the score through out-of-domain data selection, rather than solely reducing training data size as in the original VSF method. Using VSF with perplexity-based pre-ranking of out-of-domain sentences for expanding the in-domain training data for statistical machine translation is the main contribution of this paper.

4 Experiments

4.1 The Machine Translation Systems

We used a string-to-string machine translation system for English-Turkish in order to investigate the effects of the proposed method. We implemented the SMT systems with a phrase-based approach (Koehn et al., 2003). We generated word alignments using MGIZA (Gao and Vogel, 2008) and used Moses Open Source toolkit (Koehn et al., 2007) for decoding. The parameters of the system are tuned and optimized with the minimum error rate training (MERT) algorithm (Och, 2003). We tuned the system with 3 different seeds and reported the best result obtained for each setting. We trained conventional 5-gram language models (LMs) from the available parallel corpora. All language models were trained with the SRILM toolkit using the modified Kneser-Ney smoothing technique (Kneser and Ney, 1995) and then, binarized using KenLM (Heafield, 2011). We used the sentence perplexity scores produced by SRILM in order to pre-rank the parallel corpora. Moreover, in our implementation of the VSF algorithm, the threshold t values were chosen in the range $[1, 10]$ for the language pair. The details of the corpora used as in-domain and out-of-domain data are provided in the data section. For n-gram choice in VSF, unigrams are used as in the original study, since higher order n-grams select more than half of the original data. Choosing the major proportion of the data diminishes the system score as shown with the experiments, where all out-of-domain data are added to the training phase.

Data Set	Sentences	Unique Words	Total Words
Turkish	131K	158K	1.8M
English	131K	45K	2.5M

Table 2: WIT training data statistics

Data Set	Sentences	Unique Words	Total Words
Turkish	165K	143K	3.9M
English	165K	60K	4.6M

Table 3: SETIMES training data statistics

4.2 Data

For the experiments, we used WIT (Cettolo et al., 2012) data as in-domain and SETIMES (Tyers and Alperen, 2010) data as out-of-domain data. The WIT[1] corpus contains a collection of transcribed and translated talks and the core is the TED talks. On the other hand, SETIMES[2] corpus is in the news domain, collected from a website covering events in the Balkans. The statistics for the English-Turkish parallel corpora are given in Table 2 and Table 3.

Data Set	Sentences
dev2010	887
test2010	1568
test2011	1433
test2012	1698
test2013	1022

Table 4: English-Turkish test data statistics

As test sets, we used the test2010, test2011, test2012 and test2013 sets. The system is tuned with the dev2010 data set. These test sets were used in the IWSLT[3] competitions in the respective years. These test and development sets also contain collections of talks retrieved from TED talks. The sentence counts for the test and development sets are given in Table 4.

4.3 Results

The models are trained with the corpora described in the previous section. The BLEU score (Papineni et al., 2001) is used as an evaluation metric on the test sets. First, we implemented the baseline systems trained with the original in-domain data. Then, we added all of the out-of-domain data to the baseline system and retrieved the results. Using the whole out-of-domain data did not increase the BLEU score as much as expected. Afterwards, we ranked the out-of-domain data with respect to the sentence-perplexity scores based on a language model built with in-domain data. Then, we selected subsets of the sorted corpus with various VSF frequency threshold settings and used them in the machine translation systems.

[1] https://wit3.fbk.eu/

[2] http://opus.lingfil.uu.se/SETIMES.php

[3] http://www.iwslt2013.org/

System	Sentence Count	test2010	test2011	test2012	test2013
PB baseline	131K	7.94	7.94	8.02	7.09
VSF(t=1)	77K	7.34	7.46	7.48	6.69
VSF(t=2)	92K	7.56	7.51	7.57	7.06
VSF(t=5)	108K	7.60	7.71	7.56	6.83
n-gram sorted data + VSF(t=1)	85K	7.32	7.42	7.59	6.81
n-gram sorted data + VSF(t=2)	100K	7.29	7.60	7.70	6.88
n-gram sorted data + VSF(t=5)	115K	7.54	7.70	7.83	6.94

Table 5: BLEU scores for the system on in-domain data only: t is the frequency threshold for the VSF algorithm

According to the results in Table 5, the data reduction also worked for the English-Turkish system. Using the 58% of the total data, we recovered 93%, 94%, 94%, and 95% of the BLEU scores for test sets test2010, test2011, test2012, and test2013, respectively. On the other hand, compared with the origininal VSF, the language modeling based approach did not improve performance much for this case. This shows that language modeling based sorting for an only in-domain parallel corpus may not be a good metric. If we have only one parallel corpus available and we are to sort it, then metrics such as translation quality of sentence pairs can provide more promising results in the only in-domain data case.

System	Sentence Count	test2010	test2011	test2012	test2013
PB baseline	131K	7.94	7.94	8.02	7.09
ngram based + VSF (t=1)	85K	7.32	7.42	7.59	6.81
System 1 + VSF (t=1)	89K	7.42	7.37	7.59	6.32
System 2 + VSF (t=1)	74K	6.92	6.86	7.34	6.48
System 3 + VSF (t=1)	91K	6.78	7.06	7.32	6.12
System 4 + VSF (t=1)	92K	7.09	6.75	7.14	6.22

Table 6: BLEU scores for the dependency-based ranked systems stated in Table 1 on in-domain data only: t is the frequency threshold for the VSF algorithm

After using the n-gram language modeling based ranking approach, we experimented with various dependency based ranking approaches to select in-domain data. As shown in Table 6 the selection of data with dependency-based rankings did not overperform the n-gram based approach. The representation *System1* was the best scoring representation, hence we experimented with this setting in the machine translation systems that include the utilization of out-of-domain data.

In the English to Turkish translation systems that utilize out-of domain data, the best scoring system for test2012 set uses only 33% of the SETIMES data, which is our out-of-domain data. For test2011, again the same system achieves the best score. The improvement in the BLEU score is around 0.3 points. The improvements over the individual systems for test2011 and test2012 were computed to be statistically significant with a 95% confidence interval ($p<0.05$) (Koehn, 2004). Note that the BLEU scores for this language pair are generally low due to the differences between the Turkish and English languages. Turkish is morphologically more complex and the word orders between this language pair differ as well. Additionally,

System	Sentences	test2011	test2012	test2013
1. Only WIT	131K	7.94	8.02	7.09
2. (1) + SETIMES	+165K	8.00	8.1	7.16
3. (1) + ngram sorted-SETIMES + vsf(t=1)	+53K	**8.24**	**8.38**	7.27
4. (1) + ngram sorted-SETIMES + vsf(t=2)	+66K	8.12	8.2	7.15
5. (1) + ngram sorted-SETIMES + vsf(t=5)	+80K	8.05	8.15	6.88
6. (1) + dep. sorted-SETIMES + vsf(t=1)	+72K	8.27	8.38	7.19
7. (1) + dep. sorted-SETIMES + vsf(t=2)	+93K	7.89	8.11	**7.42**
8. (1) + dep. sorted-SETIMES + vsf(t=5)	+118K	8.18	8.32	7.2
9. linear(WIT + SETIMES)	+165K	7.64	7.81	7.16
10. fillup(WIT + SETIMES)	+165K	7.46	7.84	6.87

Table 7: BLEU scores for the systems utilizing out-of domain data: t is the frequency threshold for the VSF algorithm (Sentence count starting with '+' indicates the additional amount of sentences included to the data of the baseline system shown in the first row)

the size of the data for this pair does not reach to million sentences, which is generally a case for language pairs like French-English. In (Yılmaz et al., 2013), it is discussed that SETIMES data was not helpful to increase the BLEU scores for English-Turkish translation in the IWSLT test sets. However, our results show that this data can be effectively utilized to improve the translation qualities of the corresponding sets.

The other proposed sorting metrics related to dependency relations and part-of-speech tags have also shown minor improvements on some of the test sets. For test set test2013, the best scoring system is the one trained with the out-of-domain corpus ranked using dependency relation based language modeling. We compared our approach with the phrase table combination methods between different domains proposed in (Bisazza et al., 2011). Our system outperformed the phrase table combination approach, which did not bring any improvement for the English-Turkish language pair on the data sets used.

Additionally, we compared our system with TUBITAK's best system for English-Turkish translation in the IWSLT 2013 evaluation campaign by implementing their work. In (Yılmaz et al., 2013), it is stated that adding all of the SETIMES data did not improve the performance of their system, it even decreased it. Their best system was trained with hierarchical phrase-based translation (Chiang, 2007) and made use of morphological and lexical features specific to the Turkish language. We adopted their work and also reported that the addition of the entire out-of-domain data to the baseline system decreases the BLEU score. Next, we integrated our proposed data selection methods to the system. The results that we obtained are shown in Table 8. The BLEU score has increased by 0.8 points and this score is higher than the best score in the corresponding IWSLT evaluation campaign for English-Turkish translation. The increase was tested using (Koehn, 2004) and computed to be statistically significant with a 95% confidence interval ($p<0.05$). Although our replication of the system by Yılmaz et al. (2013) did not include some of the features that they have used and shown to improve performance, our system is still able to outperform their reported best system for the test2013 data set. In this system setting, we also experimented applying VSF on non-sorted (original) SETIMES data for expanding in-domain data to investigate whether sorting makes a difference or not. The results in Table 8 show that the n-gram sorting based VSF selection approach performs better than the non-sorting based VSF selection approach, even though it did not help much in the experiments where the models were only trained with in-domain data.

System	Sentences	test2013
1. TUBITAK IWSLT Best	131K	8.41
2. (1) + SETIMES	+165K	8.37
3. (1) + ngram sorted-SETIMES + vsf(t=1)	+53K	9.14
4. (1) + ngram sorted-SETIMES + vsf(t=2)	+66K	**9.20**
5. (1) + ngram sorted-SETIMES + vsf(t=5)	+80K	8.58
6. (1) + dep. sorted-SETIMES + vsf(t=1)	+72K	8.66
7. (1) + dep. sorted-SETIMES + vsf(t=2)	+93K	8.61
8. (1) + dep. sorted-SETIMES + vsf(t=5)	+118K	8.75
9. (1) + no-sort-SETIMES + vsf(t=1)	+74K	7.85
10. (1) + no-sort-SETIMES + vsf(t=2)	+94K	8.65
11. (1) + no-sort-SETIMES + vsf(t=5)	+121K	8.91

Table 8: BLEU scores of the IWSLT 2013's best system and the proposed approaches: t is the frequency threshold for the VSF algorithm (Sentence count starting with '+' indicates the additional amount of sentences included to the data of the baseline system shown in the first row)

The results show that the proposed technique successfully utilizes the available out-of-domain data and leads to improvements in BLEU for the specified domain. The approach is more a data selection technique among domains, rather than domain adaptation in which a pre-built system in a specific domain is being adapted to a different domain. It is useful for building better and more successful systems for a domain, where there is not much data, but there is a lot of data in different domains.

5 Discussion

In this study, we introduced an approach for expanding machine translation training data by utilizing an out-of-domain corpus through vocabulary saturation. We proposed using sentence ranking strategies based on n-gram and dependency relation language modeling. We evaluated the proposed methods for English-Turkish translation. This language pair does not have sufficient amount parallel texts, hence it is important to fully utilize and use all the available texts from different domains. Due to the morphological and word-sequential differences between the English and Turkish languages, most translation systems produce low BLEU scores. Turkish is an agglutinative language and is subject-object-verb oriented, whereas English is more compact and subject-verb-object oriented. Our results show that the proposed technique leads to significant improvement upon the best English-Turkish translation system reported in the IWSLT 2013 evaluation campaign. It also significantly outperforms the phrase table combination approach proposed for utilizing out-of-domain data by Bisazza et al. (2011).

The proposed approach may easily be integrated to state-of-the-art machine translation systems and applied to other language pairs. Since additional and valuable out-of-domain data is selected through this method, we believe it will lead to improvement of machine translation systems' overall successes for other languages as well. The VSF and n-gram language modeling perplexity-based rankings are algorithms that have already been proposed for machine translation. However, the combination of these approaches for expanding in-domain training data with out-of-domain data is a new approach for statistical machine translation.

6 Future Work

In the proposed methodology, there are several future directions to investigate. Ranking the corpus with an external language model is a possible direction to follow. Other potential avenues for research are using different sentence sorting metrics and features. Instead of the sentence perplexity based scoring method, other features such as sentence length or the feature functions introduced in (Taghipour et al., 2011) can be integrated into the system.

As mentioned up to now, we had data from two different domains in our experiments. We are planning to investigate the effect of adding more out-of domain data from more different domains and to see if the system score will continue to increase or not.

Moreover, we plan to examine the effects of the method specifically on the translation model. That is, we will create phrase-tables from different domains and sort the phrases in the out-of-domain phrase table with respect to an in-domain language model. Then, we will apply VSF on the sorted out-of-domain phrase table to select a subset of phrases to concatenate with the in-domain phrase table in the machine translation system.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Banerjee, P., Du, J., Li, B., Naskar, S., Way, A., and van Genabith, J. (2010). Combining multi-domain statistical machine translation models using automatic classifiers. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. In Federico, M., Hwang, M.-Y., Rödder, M., and Stüker, S., editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 136–143.

Bulyko, I., Matsoukas, S., Schwartz, R. M., Nguyen, L., and Makhoul, J. (2007). Language model adaptation in machine translation from speech. In *ICASSP (4)*, pages 117–120.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. In Cettolo, M., Federico, M., Specia, L., and Way, A., editors, *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Duh, K., Neubig, G., Sudoh, K., and Tsukada, H. (2013). Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

Eck, M., Vogel, S., and Waibel, A. (2005). Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. In *Proc. of the International Workshop on Spoken Language Translation*.

Eryigit, G., Nivre, J., and Oflazer, K. (2008). Dependency parsing of turkish. *Computational Linguistics*, 34(3):357–389.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Gubbins, J. and Vlachos, A. (2013). Dependency language models for sentence completion. In *EMNLP*, pages 1405–1410.

Guo, Y., van Genabith, J., and Wang, H. (2008). Dependency-based n-gram models for general purpose sentence realisation. In *COLING*, pages 297–304.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Accoustics, Speech and Signal Processing*, volume 1.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL).*

Lambert, B., Raj, B., and Singh, R. (2013). Discriminatively trained dependency language modeling for conversational speech recognition. In *INTERSPEECH*, pages 3414–3418.

Lewis, W. and Eetemadi, S. (2013). Dramatically reducing training data size through vocabulary saturation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 281–291, Sofia, Bulgaria. Association for Computational Linguistics.

Moore, R. C. and Lewis, W. D. (2010). Intelligent selection of language model training data. In *ACL (Short Papers)*, pages 220–224.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In Hinrichs, E. and Roth, D., editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Okita, T. (2009). Data cleaning for word alignment. In *ACL/IJCNLP (Student Research Workshop)*, pages 72–80. The Association for Computer Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report.

Shen, L., Xu, J., and Weischedel, R. (2008). A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT*, pages 577–585, Columbus, Ohio. Association for Computational Linguistics.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.

Taghipour, K., Khadivi, S., and Xu, J. (2011). Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 414–421. International Association for Machine Translation.

Tyers, F. M. and Alperen, M. S. (2010). South-east european times: A parallel corpus of the balkan languages.

Wang, W., Macherey, K., Macherey, W., Och, F., and Xu, P. (2012). Improved domain adaptation for statistical machine translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Wu, H., Wang, H., and Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 993–1000, Manchester, UK. Coling 2008 Organizing Committee.

Yılmaz, E., İlknur Durgar ElKahlout, Aydın, B., Özil, Z. S., and Mermer, C. (2013). TÜBİTAK Turkish-English submissions for IWSLT 2013. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.