Data Categorization and Model Weighting Approach for Language Model Adaptation in Statistical Machine Translation

Mohammed AbuHamad[1], Masnizah Mohd[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi Selangor, Malaysia

***Abstract*—Language model encapsulates semantic, syntactic and pragmatic information about specific task. Intelligent systems especially natural language processing systems can show different results in terms of performance and precision when moving among genres and domains. Therefore researchers have explored different language model adaptation strategies in order to overcome effectiveness issue. There are two main categories in language model adaptation techniques. The first category includes the techniques that based on the data selection where task-oriented corpus can be extracted and used to train and generate models for specific translations. While the second category focuses on developing a weighting criterion to assign the test data to specific model corpus. The purpose of this research is to introduce language model adaptation approach that combines both categories (data selection and weighting criterion) of language model adaptation. This approach applies data selection for specific-task translations by dividing the corpus into smaller and topic-related corpora using clustering process. We investigate the effect of different approaches for clustering the bilingual data on the language model adaptation process in terms of translation quality using the Europarl corpus WMT07 that includes bilingual data for English-Spanish, English-German and English-French. A mixture of language models should assign any given data to the right language model to be used in the translation process using a specific weighting criterion. The proposed language model adaptation has achieved better translation quality compare to the baseline model in Statistical Machine Translation (SMT).**

***Keywords*—Language model adaptation; statistical machine translation; clustering**

I. Introduction

Language models are considered as important knowledge sources for different natural language processing applications. Language model encapsulates semantic, syntactic and pragmatic information about specific task. Language model has been widely adopted and investigated in speech recognition domain in the last two decades. Recently, dual learning in language models also have been applied in statistical machine translation (SMT) and neural machine translation (NMT) [1]. Normally, the size and domain of the language models can significantly influence the translation quality. According to [2], each doubling in the training data used to build a language model improves the translation approximately 0.5 BLEU (Bilingual Evaluation Understudy) points. Usually, intelligent systems especially natural language processing systems can show different results in terms of performance and precision when moving among genres and domains. Thus, adaptation process must be considered to such applications.

The state-of-the-art SMT systems nowadays involve many components, such as reordering models, language models, translation models, etc. Language models are considered as an essential component of current SMT systems. They influence the selection and the reordering of text translation candidates by estimating the probability that a given text translation is a proper translation according to the current translation hypothesis. Language models can be influenced by the size and topics covered in the constructive corpus. Since SMT can be considered as the pattern recognition approach for machine translation, the results can be enhanced using different methods. One of the approaches is to apply language model adaptation technique to consider the tested data for better translation.

This research is concentrated to study different methods of language model in SMT system and to introduce language model parameters that can adapt to the input text. In statistical machine translation, there is a number of adaptation techniques have been applied to handle this issue by estimating the model parameters from some data and adapting to translate sentences which might not be covered in the training process. For this reason, language model adaptation techniques need to be explored for SMT applications. Basically, language model adaptation techniques can be referred to two main categories. The first category includes the techniques that based on the data selection where task-oriented corpus can be extracted and used to train and generate models for specific translations. On the other hand, the second category focuses on developing a weighting criterion to assign the test data to specific model corpus. The proposed new approach to language model adaptation combines both strategies of the previous two categories of language model adaptation. At first, this approach applies data selection for specific-task translations by dividing the corpus into smaller and topic-related corpora using clustering process. This step can be performed either in a fully unsupervised manner or by considering supervised labels according to specific bilingual corpora. In this case, each subset covers specific characteristics or topic. Afterwards, several language models can be built based on these topic-related corpora and weighting criterion can be defined to assign any given data to the right language model to be translated. Using N-gram mixture of specialized sub-language models to implement overall language model is the basic idea behind our approach to enhance the quality and precision of SMT system. The idea behind using N-gram to implement the sub-language models is that N-gram has been widely popular due to the reliability and robustness of their estimates as well as simplicity to be measured [3]. This paper is organised as follows: the following section reviews some related work, followed by a section representing the proposed approach. The basic two steps of the proposed approach, bilingual data clustering and weights estimation of language models mixture, are presented in separate sections including results of experiments on each method. The final section summarises the work and presents the conclusion.

II. Related Work

Language model adaptation, the matter that has been extensively investigated since the mid-90s in the domain of speech recognition [4],[5] has been witnessing a growing interest in the domain of SMT. One of the earliest methods proposed to handle the adaptation concept in SMT was introduced in [6] which focuses on the translation model, not the language model. This approach has implemented the translation model component as a mixture of translation models, each model is meant to handle a particular topic and to focus its probability on this topic.

Using mixture of models for the purpose of adaptation in SMT has been also explored in [1], in which the mixture of models was implemented for word alignments component. To this end, [7] have proposed to use a mixture of Hidden Markov Model (HMM) alignment models as a replacement of the classic word-alignment model. Since the process of modelling word-alignments mixture is based on soft partitions, each mixture component is meant to handle topic specific alignments. Despite the interesting improvements that this approach has reached considering the alignment error rate, the translation quality seemed to be more restricted due to the large number of processes and heuristics applied to extract phrases after the process of word-alignment.

For the purpose of developing interactive machine translation systems, [8] have proposed other adaptation techniques inspired by the ideas shown in [9] by associating cache language as well as some translation models with the machine translation system. Adopting the same principle as in the cache memories, the main point of adding caches for both translation model and language model is to exploit the fluctuations in words or phrases frequency. These additional caches are merged in the basic translation model and the language model using a log-linear fashion. This study has shown that the language model caches have yielded a remarkable enhancement on the translation quality, although the translation model caches seemed to be incapable yielding further improvements.

Other researchers have used different ways and methods to handle the adaptation problem. For example, [10] have explored different methods to exploit both in-domain and out-of-domain data. In their research, experiments were ranged from using simple series of the entire available data to using more sophisticated combination criterions, such as building several translation models as well as language models to be further merged in a log-linear manner. Using a similar conceptual idea, [11] have also investigated different methods and approaches to make use of all in-domain and out-of-domain data. The basic difference in this work comes from using only the source language data.

It has been obvious that the SMT community has shown a continuously growing interest on language model adaptation aspect. More precisely, researchers have shown recent efforts towards developing more adaptable language models in the SMT system. For instance, [12] have suggested using a query from a set of possible translations for every input sentence. Such query can be used in processing similar sentences while using very large bilingual training corpus, where sentences captured can be used to construct specific language models which are further incorporated in translation time with an actual language model built on the entire available data. Finally, the any given input sentence is re-translated using the final language model. Adopting this approach, the results reported from their work show that the improved language model was able to yield stable and limited enhancements on the single baseline language model.

The similar idea was explored by [13], although in this work term frequency-inverse document frequency (TF-IDF) was used to determine similar data present in the bilingual training corpus, and then use them to build specific language models and translation models. Mixtures of these specific models can be built and activated in translation time based on different weighting criterions. Such work has also been shown in [12], in which the reported results provide a slight but stable enhancement in translation quality.

Similarly, [14] have suggested to perform a categorisation process over the bilingual training corpus in terms of the entropy of each cluster, and then constructing translation models based on these cluster (smaller corpora), which are interpolated using domain prediction during the translation time. The key idea of such work was extended in [15] in which different clustering methods have been examined in terms of their influence on the language model adaptation process, and the resultant translation quality. The bilingual data clusters were used to build different language models which are interpolated using different weightings schemes.

III. Proposed Approach

The basic idea of developing an adaptive language model is to replace the language model part, where the problem of machine translation was defined as follows: given a sentence f from a certain source language, an equivalent sentence eˆ in a given target language that maximizes the posterior probability is to be found. Such a statement can be formalized as:

$$\hat{e} = argmax_e \, Pr(e|f)$$
$$= argmax_e \, Pr(f|e).\, Pr(e) \qquad (1)$$

where $Pr(f|e)$ stands for the translation probability and $Pr(e)$ accounts for penalizing ill formed sentences of the target language. More recently, a direct modelling of the posterior probability $Pr(e|f)$ has been widely adopted, and, to this purpose, different authors [16] proposed the use of the so-called log-linear model, where

$$\Pr(e|f) = \frac{exp \sum_{k=1}^{K} \lambda_k h_k(f,e)}{\sum_{e'} exp \sum_{k=1}^{K} \lambda_k h_k(f,e')} \quad (2)$$

where

$$e' = e_{i\pm1} \quad (3)$$

And the decision rule is given by the expression

$$\hat{e} = \underset{e}{\text{argmax}} \sum_{k=1}^{K} \lambda_k \, h_k(\mathbf{f}, \mathbf{e}) \quad (4)$$

where $h_k(\mathbf{f}, \mathbf{e})$ is a score function representing an important feature for the translation of **f** into **e**, for example the target language model $p(\mathbf{e})$, K is the number of models (or features) and λ_k are the weights of the log-linear combination. Typically, the weights λ_k are optimized during the tuning stage with the use of a development set.

In this research, this function is extended to include multi-language-models by obtaining the probability estimated by a linear mixture of n-grams (smaller word-based language models). This probability can be estimated by the equation (5)

$$p(y) = \sum_{i=1}^{M} w_i p_i(y) \quad (5)$$

where $p_i(y)$ presents the language model trained on a sentence i in the target language. Adopting this formula views the final probability $p(y)$ as a mixture of several language models trained on different parts of the available training data.

The general process of language model adaptation is shown in Fig. 1. Since this process considers adapting the language models built and trained using the target data of parallel corpus, the language models trained using monolingual corpora do not fit this process. Assuming the parallel bilingual corpus is divided into M number of bilingual clusters using specific criterion. For each cluster, several language models are built and then assigned in two language-specific mixtures of models. This procedure is performed off-line over the available training data. Considering an input source text to be translated, this text can be used to define the optimum weights in the source mixture part using Expectation-Maximization (EM). These weights are supposed to contain essential information about the source mixture of the language models featuring the distribution of these models. Such information can be valuable by passing them to the target mixture using a certain process of mapping weights between the source and the target sides. After terminating the mapping process, the mixture of target language models can be employed as language model part in the SMT system. In this research, the mapping process is applied by directly assigning the target weights the same value as the source weights. More sophisticated methods can be easily applied to perform this mapping and they could be more appreciated. However, this research adopts the straightforward method.

The basic idea of language model adaptation in this research includes the process of clustering bilingual training data as discussed in next section (Section 4).

Four language models are trained for each side of the bilingual data (4 language models for the source part plus 4 language models for the target part). For smoothing purpose, a general language model is built based on the entire training data.

The experiments carried out to assess different mechanisms for language model adaptation are implemented based on the Europarl corpus (WMT07 partition). Thus, the bilingual data considered in the experiments includes English-Spanish, English-German and English-French (see Table 1 for some details).

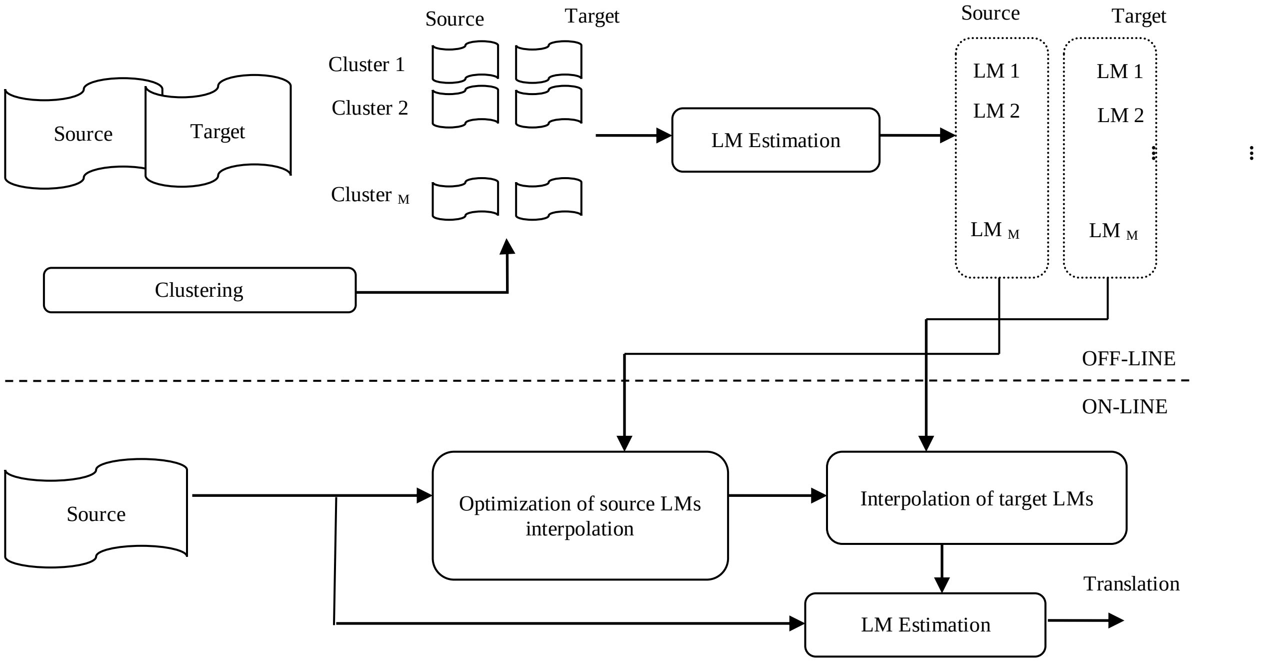

Fig. 1. The Proposed Language Model Adaptation Process.

TABLE I. THE WMT 2007 PORTION OF EUROPARL CORPUS

		Spanish	*English*	*German*	*English*	*French*	*English*
Training Set	**Sentences**	731k		751k		688k	
	Running words	15.7M	15.2M	15.3M	16.1M	15.6M	13.8M
	Average length	21.5	20.8	20.3	21.4	22.7	20.1
	Vocabulary size	103k	64k	195k	66k	80k	62k
Devel	**Sentences**	2000		2000		2000	
	Running words	61K	59k	55k	59k	67K	59k
	Average length	30.3	29.3	27.6	29.3	33.6	29.3
	Out of Vocabulary to WMT 2007	208	127	432	125	144	138
Devtest	**Sentences**	2000		2000		2000	
	Running words	60K	58k	54k	58k	66K	58k
	Average length	30.2	29.0	27.1	29.0	33.1	29.0
	Out of Vocabulary to WMT 2007	207	125	377	127	139	133
Testing set	**Sentences**	3064		3064		3064	
	Running words	92K	85k	82k	85k	101K	85k
	Average length	29.9	27.8	26.9	27.8	32.9	27.8
	Out of Vocabulary to WMT 2007	470	502	1020	488	536	519

The baseline and further SMT systems are implemented using Moses SMT toolkit. The log-linear model weights λ are adjusted and optimised using the well-known method called minimum error rate training (MERT) over the baseline system for the development training set (Devel.) and then adopted in all other systems. Even though it is possible to apply MERT in every language model to obtain the best weights for each individual one, adopting standard weights would better avoid the effects of using many language models in the SMT system. For all experiments, the baseline language model is implemented as 5-gram word-base language model using SRILM toolkit [17]. The language models are constructed based on the target side of the bilingual training data. The language models are smoothed using the extended Kneser-Ney technique presented in [18]. For the final results, the Devtest set is adopted to measure the final quality of the translation in terms of BLEU and TER measures, the BLEU as shown in [16], and Translation Error Rate (TER) as presented in [19],[20],[21].

IV. BILINGUAL DATA CLUSTERING

The basic idea of language model adaptation in this research includes the process of clustering bilingual training data. Since the bilingual data is basically formed of sentences from both source and target data, the clusters are supposed to hold the same features besides the high degree of similarity among their elements. The key point of this process is to categorise the sentences that share similar lexical features into clusters in order to further implement language models for these clusters. This process highlights big benefits to intelligent categorisation of bilingual data since most of bilingual corpora lack the supervised labels of their contents. Thus, this research explores this process in different criterions and their impact on the adaptation process on language model part of SMT system. According to previous studies, some parameters and settings are specified as follows:

a) Processing the training data by considering bilingual sentences as bags of words from source and target languages.

b) Defining the number of clusters as 4 clusters, since previous studies suggested that this number can produce specialised clusters with high similarity among their elements and not too spare.

c) The cosine similarity is adopted to calculate the similarity between sentences and assign sentences to different clusters [22].

In this research, three approaches for bilingual data clustering are investigated. The first approach is a straightforward method in which the training data is applied directly to the clustering process and the resultant clusters will be used to build different language models. The second approach considers the development set in the clustering process in order to overcome the issue of mismatching patterns between training and development sets. The basic idea to be applied for this purpose is to perform a clustering process for the development set, and afterwards categorise the training data according to the patterns obtained by clustering the development set (Fig. 2). After performing a clustering process for the development set, language models can be constructed for each cluster which can be used to categorise each bilingual sentence n from the training data to a specific cluster $\hat{m}$ according to the formula.

$$\hat{m} = argmax_m \cos(t_n^x, d_m^x) + \cos(t_n^y, d_m^y) \quad (6)$$

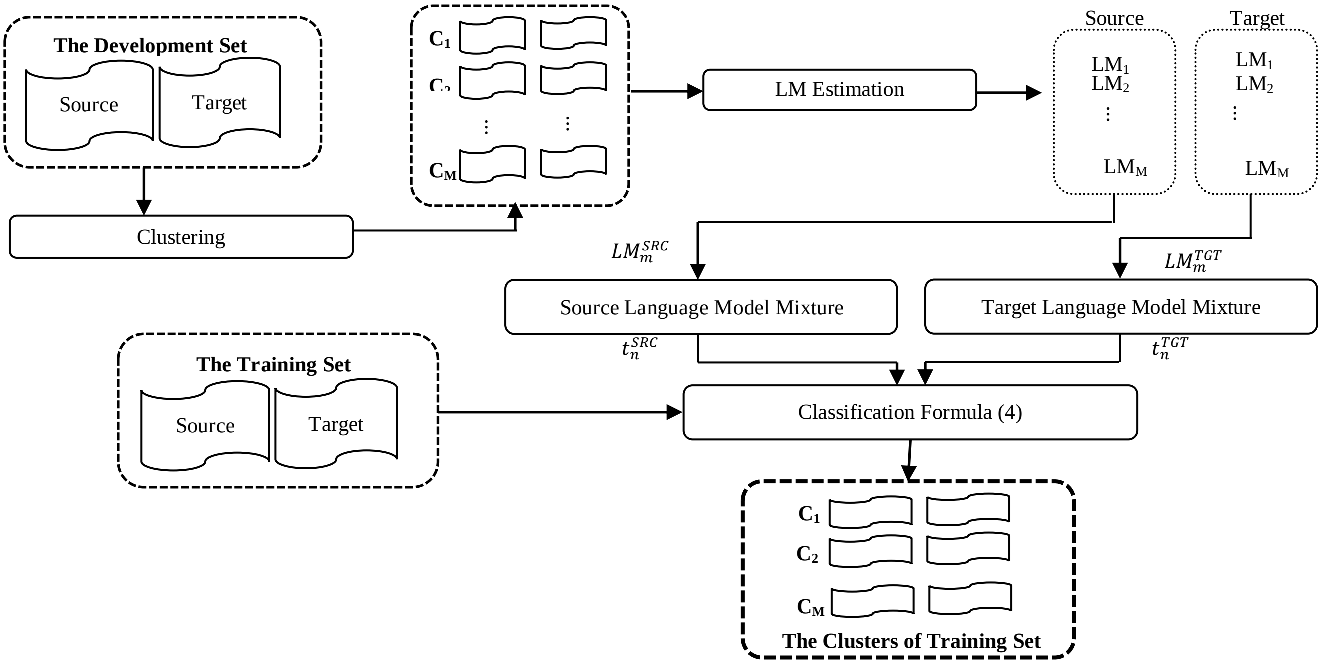

Fig. 2. Clustering Bilingual Training Corpus Considering the Development Set.

The t_n^x and t_n^y present the language model weights to maximise the n sentence probability in the training data on the source and the target sides respectively, based on the linear mixture of source and target language models obtained from clustering the development set. While, d_m^x *and* d_m^y present the language model weights to maximize the n sentence probability in the cluster m of development set on the source and the target sides respectively.

The last approach for clustering the training data is by considering the test set rather than the development set in the previous approach. Since the test set has no target side, the clustering is performed according to the source side of the test set with simple modification on the formula used for the previous approach to become:

$$\hat{m} = argmax_m \cos(t_n^x, d_m^x) \qquad (7)$$

Considering the three different clustering approaches adopted to categorise the bilingual training data, Table 2 shows the translation quality of SMT system developed with different language models. Generally, the TER score achieved by the three approaches has outperformed that achieved by the baseline system while BLEU score has shown different results as shown in Table 2.

TABLE II. TRANSLATION QUALITY USING DIFFERENT LANGUAGE MODELS ADAPTATION APPROACHES

Language Model Adaptation	**English-Spanish**		**English-German**		**English-French**	
Clustering approach	BLEU	TER	BLEU	TER	BLEU	TER
Direct clustering	30.3	54.5	18.0	67.6	32.5	55.0
Clustering/*development set*	30.9	54.6	18.7	67.2	32.9	55.1
Clustering/*test set*	31.0	54.6	18.9	67.1	33.0	55.2
Baseline system	29.7	55.6	18.2	68.4	33.1	56.3

The direct clustering approach has not improved the translation quality in terms of BLEU measure. The observed BLEU score has degraded in all experiments except a slight improvement in the *English-Spanish* corpus. However, the TER score has decreased as an indication of better performance. The second approach, clustering based on the development set, shows a slight improvement of the translation quality in comparison with direct clustering. The BLEU score has not been largely affected by this approach, despite the slight improvement on the BLEU score. In the experiment of the English-French corpus, this approach has not achieved a BLEU score that surpasses the baseline system score. However, it has achieved lower TER score as an indication of better performance not only in the English-French but also in the other corpora. The last approach, clustering based on the test set, has achieved almost the same results from other experiments with slight improvement in the BLEU score. Among the three different approaches, clustering the bilingual training data based on the test set seems to have better effect on the BLEU score.

V. WEIGHTS ESTIMATION CRITERION FOR LANGUAGE MODEL MIXTURE

In the previous experiments, the weighting criterion used to make the interpolation of language models was the most simple and straightforward criterion. That weighting criterion was based on the test set wherein the source side only available. The weights of language model interpolation were obtained using the source side of the entire test set. Despite the fact that this criterion is the most straightforward, it might not be the best option to estimate weights for the language model interpolation due to fact that using the whole test set would favour the weights that model the entire set, regardless probable significant differences on sentence level. This issue can be so important that the desired benefits of building several language models may fade. For this purpose, two other weighting criterions have been investigated. The first one is by considering the sentence level where weights can be estimated for each individual sentence on the source side of the test set. This would enable complete freedom for assigning weights to the language models and getting better results in case the training data is divided into several subsets. However, weights estimated in such criterion could be less reliable due to lack of data that produce estimation (only one sentence).

In the attempt to utilize the capabilities and advantages and minimize the drawbacks of those criterions, another criterion was introduced by combines the previous two weighting criterions (based on the entire test set and sentence level). At first, for each sentence in the test set, weights estimated based on the sentence level are used to classify the test sentences into different categories according to the most weighted language model. Afterwards, for each cluster of sentences, weights are re-estimated to make sure to consider the entire cluster rather than relying only on the sentence level estimation. Fig. 3 shows the simple procedure of this approach. This criterion has the intuitive advantage of reflecting the clustering of the bilingual training data (through the obtained language models) into the test set, without worrying about possible data sparseness that could affect the weights estimated on the sentence level.

For the three bilingual corpora (English-Spanish, English-German and English-French), the experiments were conducted to investigate the impact of the language model adaptation process using three clustering approaches and three criterions for weight estimation for models mixture. The results are illustrated in Table 3.

For all the bilingual training corpora (English-Spanish, English-German and English-French), the language model adaptation process has increased their performance in terms of translation quality measures (BLEU and TER). Categorising the training data into clusters with high intra-element similarity could lead to better construction of specific language models which can work as a mixture in the translation time with better translation result. Among the clustering approaches adopted in this research this categorisation of bilingual training data, clustering training data based on the development set seems to have the lead ahead of other approaches. Since the language models are built based on the development set in the first place, categorising the training data based on the development set can be more reasonable since both sides of the development set can be reflected on the clusters of the training data. The clustering approach based on the test set may not provide reliable results since the categorisation of training data is categorised based only on the source side of the test set which is the only available data.

Among the three weighting estimation criterion, the hybrid approach has achieved the best results on constructing the language model mixture. As shown in Table 2, categorising the training data based on the development set and adopting the hybrid weight estimation can build the best language model mixtures that can lead to better translation quality in both measures, BLEU and TER. Comparing to the baseline system, the results achieved by language model adaptation have achieved better translation quality. The overall results show that adopting language model adaptation method has provide better translation quality, thus impact the translation performance task in Statistical Machine Translation (SMT) system.

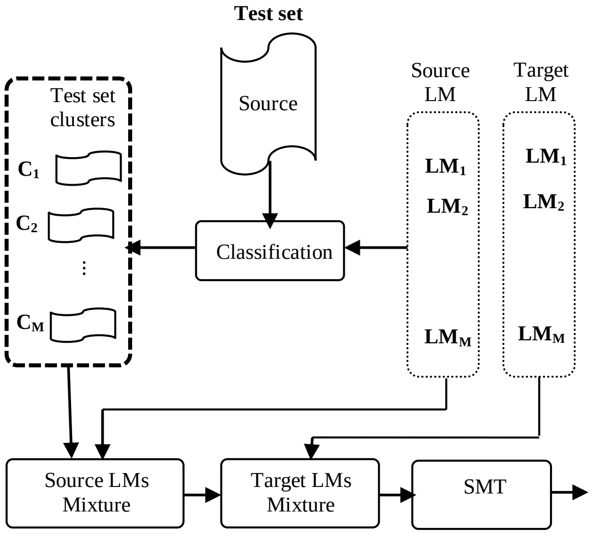

Fig. 3. Hybrid Weighting Criterion.

TABLE III. TRANSLATION QUALITY IN DIFFERENT LANGUAGE MODEL ADAPTATION SETTINGS

		English-Spanish		**English-German**		**English-French**	
Clustering approach	Weighting criterion	BLEU	TER	BLEU	TER	BLEU	TER
Direct clustering	*Based on entire test set*	30.3	54.5	18.0	67.6	32.5	55.0
	Based on sentence level	31.2	53.8	18.6	67.2	32.6	54.8
	Based on hybrid criterion	31.1	54.2	18.4	67.1	32.8	54.2
Clustering/ *Development*	*Based on entire test set*	30.9	54.6	18.7	67.2	32.9	55.1
	Based on sentence level	31.9	54.1	19.8	66.4	33.4	54.6
	Based on hybrid criterion	32.3	53.1	20.2	65.3	34.1	54.1
Clustering/ *testing*	*Based on entire test set*	31.0	54.6	18.9	67.1	33.0	55.2
	Based on sentence level	30.7	54.3	19.1	66.9	32.9	55.4
	Based on hybrid criterion	31.3	54.1	19.0	66.7	33.1	54.6
Baseline system		29.7	55.6	18.2	68.4	33.1	56.3

VI. Conclusion

This work has explored the problem of language model adaptation in SMT using several approaches. Several experiments have been carried out to study language models and their n-gram mixtures which are constructed using different clustering criterion on the bilingual training data. Several clustering approaches were examined and analysed using different means to automatically categorise the bilingual training data with an unsupervised manner. Given the fact that the training data has been well-categorised, independent language models are built based on each cluster. The resultant language models are assigned by weights to form a mixture of language models to construct an adaptive language model component to replace the typical language model component in the SMT system. Several experiments have been carried out to estimate the mixture weights with different degrees of granularity, starting from sentence level and ending with including the entire test set. The results of the conducted experiments show that translation quality can be improved by building different language models rather than using a single language model. These different language models can be trained and weighted based on actual input data to develop a mixture, able to produce better translation quality in terms of both BLEU and TER. In the future, we will try to experiment with a specific sentence weighting method in SMT domain adaptation.

Acknowledgement

This study was supported by the Universiti Kebangsaan Malaysia grant: GGP-2017-022

References

[1] Di, H., Yingce, X., Tao, Q., Wang, L., Nenghai, Y., Tie-Yan, L., and Wei-Ying, M. 2016. Dual learning for machine translation. In NIPS, pp. 820–828.

[2] Brants, T., Popat, A. C., Xu, P., Och, F. J., Dean, J., 2017. Large language models in machine translation. In: Proceedings of the 2017 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, pp. 858-867.

[3] Xiong, T., Popat, A. C., Xu, P., Och, F. J., Dean, J., 2017. Large language models in machine translation. In: Proceedings of the 2017 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, pp. 858-867.

[4] DeMori, R., Federico, M., 1999. Language model adaptation. Computational Models of Speech Pattern Processing, Springer, pp. 280-303.

[5] Bellegarda, J. R. 2001. An overview of statistical language model adaptation. In: ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition, pp. 165-174.

[6] Lagarda, A., Juan, A., 2003. Topic detection and classification techniques. In: WP4 deliverable, TransType2.

[7] Civera, J., Juan, A., 2017. Domain adaptation in statistical machine translation with mixture modelling. In: Proceedings of the Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, pp. 177-180.

[8] Nepveu, L., Lapalme, G., Langlais,P., Foster, G., 2014. Adaptive language and translation models for interactive machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp. 190-197.

[9] Kuhn, R., De Mori, R., 1990. A cache-based natural language model for speech recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions 12(6), 570-583.

[10] Koehn, P., Schroeder, J., 2017. Experiments in domain adaptation for statistical machine translation. In: Proceedings of the Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, pp. 224-227.

[11] Bertoldi, N., Federico, M., 2009. Domain adaptation for statistical machine translation with monolingual resources. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Athens, Greece, pp. 182-189.

[12] Zhao, B., Eck, M., Vogel, S., 2014. Language model adaptation for statistical machine translation with structured query models. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, Geneva, Switzerland, article 411.

[13] Lü, Y., Huang, J., Liu, Q., 2017. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In: Proceedings of the 2017 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, pp. 343-350.

[14] Yamamoto, H., Sumita, E., 2008. Bilingual cluster based models for statistical machine translation. IEICE - Transactions on Information and Systems 91(3), 588-597.

[15] Sanchis-Trilles, G., Cettolo, M., 2010. Online language model adaptation via n-gram mixtures for statistical machine translation. In: Proceedings of the Conference of the European Association for Machine Translation, Saint Raphaël, France.

[16] Papineni, K., Roukos, S., Ward, T., Wei-Jing, Z., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 311-318.

[17] Stolcke, A., 2002. SRILM-an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing, pp. 257-286.

[18] Chen, S. F., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Computer Speech & Language 13(4), 359-393.

[19] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2016. A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223-231.

[20] Andreas, T., Prasanta, G., Panayiotis, G., Shrikanth, N., 2013. High-quality bilingual subtitle document alignments with application to spontaneous speech translation. Computer Speech & Language 27 (2), 572-591.

[21] Albadr, M. A. A., Tiun, S., & Al-Dhief, F. T. 2018. Evaluation of machine translation systems and related procedures. *ARPN Journal of Engineering and Applied Sciences*, *13*(12), 3961-3972.

[22] Mohd, M, Bsoul, QW, M.Ali, N, M.N, S.Azman, Saad, S, Omar,N, and A.Aziz., M.J. in 2012. Optimal Initial Centroid in K-Means for Crime Topic. Journal of Theoretical and Applied Information Technology (JATIT). 44(2): 19-26