Improving Domain-specific SMT for Low-resourced Languages using Data from Different Domains

Fathima Farhath, Pranavan Theivendiram, Surangika Ranathunga, Sanath Jayasena, Gihan Dias

Department of Computer Science and Engineering
University of Moratuwa
Katubedda 100400, Sri Lanka
{fathimafarhath, pranavan.11, surangika, sanath, gihan}@cse.mrt.ac.lk

Abstract

This paper evaluates the impact of different types of data sources in developing a domain-specific statistical machine translation (SMT) system for the domain of official government letters, between the low-resourced language pair Sinhala and Tamil. The baseline was built with a small in-domain parallel dataset containing official government letters. The translation system was evaluated with two different test data sets. Test data from the same sources as training and tuning gave a higher score due to over-fitting, while the test data from a different source resulted in a considerably lower score. With the motive to improve translation, more data was collected from, (a) different government sources other than official letters (pseudo in-domain), and (b) online sources such as blogs, news and wiki dumps (out-domain). Use of pseudo in-domain data showed an improvement for both the test sets as the language is formal and context was similar to that of the in-domain though the writing style varies. Out-domain data, however, did not give a positive impact, either in filtered or unfiltered forms, as the writing style was different and the context was much more general than that of the official government documents.

Keywords: domain-specific statistical machine translation, low-resourced languages, Sinhala, Tamil, Domain Adaptation

1. Background

Sri Lanka is a multi-ethnic country where Sinhala and Tamil are the official languages. However, only a small number of the population can communicate in both the languages. Therefore, to smoothly carry on the official government communication with the public, dissemination of information has to happen in both the languages. This requires the official government letters to be produced in both the languages. However, this process has become an extra burden to government institutions as they lack professional translators with bilingual knowledge. In order to reduce the burden on the translators to a certain extent, the alternative option is to use the assistance of *machine translation* (MT) in their translation workflow.

For constructing an MT system, prevailing methodologies are *statistical machine translation* (SMT) (Kohen, 2009) and *neural machine translation* (NMT) (Bahdanau, 2014). Though NMT is gradually establishing its root as a viable alternative to SMT ((Bojar et al., 2016), the applicability of it in a low-resource setup is questionable as it requires a large amount of parallel data to boost its gain (Koehn & Knowles, 2017). However, Sinhala and Tamil is a low-resource language pair with minimum linguistic resources and a very little prior research on machine translation between them. Therefore building an NMT system is not a fruitful option (Tennage et al. 2017) whereas an SMT system seems more feasible. However, since SMT systems are also data driven, this requires careful analysis in data selection and utilization in order to get the optimum out of available resources.

The aim of this research is to develop an effective SMT model that can be used to build a translation system for translating official government letters from Sinhala-to-Tamil and vice-versa. However, the size of the parallel dataset that contains official letters (collection contains letters from the Department of Education, Department of Official Languages, and from few regional administrative offices of the government) is relatively small (7,757 parallel sentences). Even collecting this small set of official letters in a ready-to-use form in both the languages required much effort and time due to many institutional problems and resource constraints.

When the system was evaluated using a test set (Test-1) that was randomly picked from the collection of letters from where the training and tuning data are also derived, the *bilingual evaluation understudy* (BLEU) scores (Papineni et al., 2002) of the translation were considerably high. However, when this same system was evaluated using a test set (Test-2) from a different set of letters, from which no data were included into training or tuning, the scores were significantly inferior. The exceptionally high scores in the former case are due to the over-fitting of the *language model* (LM) and *translation model* (TM) to the training data. Test-2 scored lower since a high number of *out of vocabulary* (OOV) words were present in the test data, and less context-related language flow was present in the model.

Since the in-domain parallel data described above was considerably small for this low-resourced domain-specific system, we explored the use of data from other sources that have some relevance and easier to access. Under this, a significant amount of new parallel data was gathered from documents produced by various government institutions. The writing style of these documents slightly differs from that of the letters from which the original in-domain data were derived. Yet, the new data added value to the vocabulary as they have a similar context and terminology to those in official government documents. In literature, this type of data is referred to as 'pseudo in-domain' data (Axelrod et al., 2011). We were able to collect about 15,000 pseudo in-domain parallel sentences, which would allow us to explore domain adaptation (Koehn and Schroeder, 2007).

In addition, a large amount of monolingual data in both Sinhala and Tamil was gathered from online sources such

as blogs, news, and wiki dumps. Since the writing style of these differs from that of official letters and the topics of interest deviate from that of official government letters, we consider this as out-domain data. In literature, when developing domain-adapted systems, using out–domain data is recommended in a filtered form, where sentences are removed from the corpus based on perplexity-based measure differences (Moore and Lewis, 2010). We experimented with out-domain data in our system, by using them in unfiltered (raw form) as well as in filtered form.

Our experimental results indicate that adding pseudo in-domain data to parallel as well to monolingual data contributes to an increase in translation accuracy over both the test sets, while the contribution of the out-domain monolingual data was not significant neither in the filtered form nor unfiltered form.

The rest of the paper is organized as follows. The second section relates on literature on MT, Sinhala – Tamil MT, and commonly used approaches on domain adaptation. Section three describes the data. The fourth section elaborates on the experiments carried out. In section five, we present the evaluation and analysis based on the experiments and results. And finally, the sixth chapter presents the conclusion of the research along with future work.

2. Related Work

2.1 Machine Translation

Today, machine translation is a mature natural language processing task. Successful translation systems have been produced for many language pairs in the world, European languages in particular. These have enabled the use of machine translation in professional translation workflows to enhance the productivity in domains such as medical, automobile, manufacturing and legal (Kohen, 2009).

For more than a decade, SMT has been the fundamental technique for machine translation. However, recently NMT has invaded this position by producing better outputs than SMT (Bahdanau et al. 2014). Both the methodologies are data-driven. Yet, in NMT, the higher performance gain is achieved with greater amounts of parallel data and more computing power (Koehn & Knowles, 2017). This makes NMT less practicable for a low-resource setup.

2.2 Sinhala –Tamil Machine Translation

Sinhala language falls under the family of Indo-Aryan languages, while Tamil belongs to the Dravidian family. When comparing with English, the differences between Sinhala and Tamil are less, where the default sentence structure in both the languages is comprised of subject-object-verb. Yet these two languages are highly inflectional and many differences exist in their syntactic structures.

Both Sinhala and Tamil lack quality linguistic resources (Weerasinghe, 2003). Minimal research is done in building MT systems for this language pair. An initial *phrase-based statistical machine translation* (PBSMT) feasibility study for this language pair was done by Weerasinghe (2003) on 5000+ parallel sentences. To date, the highest BLEU score for the open domain PBSMT is 10.1 for Sinhala to Tamil (Pushpananda et al., 2014), and 13.11 for Tamil to Sinhala (Pushpananda et al., 2015).

The initial NMT approach for this language pair was by Tennage et al. (2017) with a total size of 23,000+ of parallel data (includes training, validation and testing), for the domain of official government documents. The results revealed that the quality of the output of NMT is much worse compared to that of SMT.

.

2.3 Domain Adaptation for Domain-specific SMT

Writing style of a language differs along the genre (e.g. blogs, scientific writing, and legal documents). Moreover, the literal meaning of words and the flow of the language highly depend on it. SMT systems developed for open-domain are not capable of addressing these domain specific variations, as they are trained using general data. The best way to build a domain specific SMT system is to develop a SMT system solely with a large amount of in-domain data. Yet, finding such an amount of in-domain data is practically infeasible in the context of many languges. Domain adaptation (Kohen, 2009) could be used in such situations.

Two major approaches are used in domain adaptation:

1. **Using an open-domain system to fine tune into a specific domain:**

Koehn and Schroeder (2007) suggest the use of cross-domain adaptation. Here, a considerably small amount of in-domain data is being exploited over a considerably large amount of out-domain data using a linear interpolation technique.

Foster and Kuhn (2007) used the concept of mixture modeling (McLachlan and Peel, 2004) to develop dynamic domain adaptation. Here, for different domains, adaptation was done using a cross-domain technique. By analyzing the input text, a mixture model is generated based on an unsupervised clustering method and mixture weights are estimated dynamically. This is an extended version of Koehn and Schroeder's (2007) system, as they cater domain adaptation for multiple domains in one system in a dynamic manner.

Civera and Juan (2007) use mixture modeling in domain adaptation to enhance the word alignments by intervening the alignment process to generate topic-dependent word alignment over general alignment. Yet they doubt on the applicability of this technique as the performance of SMT depends on many factors.

2. **Data Filtration techniques to extract data from open-domain corpus that are similar to the in-domain data**

In order to guarantee that data is from a same or similar domain, different filtration techniques are used in collecting and filtering open-domain monolingual data (Eck et al., 2004), as well as parallel data (Hildebrand et al., 2005).

Data filtration is the process where a given set of data is being processed to remove the less similar sentences from an out-domain corpus with reference to the given in-domain corpus.

One of the measures used for filtration is perplexity (Kohen, 2009). Different techniques are being developed based on this concept to filter parallel as well as monolingual data. This concept is used in SMT domain adaptation with the motive to reduce the influence of highly deviating or less similar sentences.

Gao et al. (2002) suggest the use of a simple perplexity metric of sentences based on the in-domain LM to filter off the sentences that have a perplexity higher than a threshold. Moore and Lewis (2010) convey the idea of using the cross-entropy difference between the in-domain and out-domain LMs as the measure for filtration. They point out that this methodology works better in reducing the perplexity than Gao et al.'s (2002) method. Both these techniques can be applied for parallel as well as monolingual data. Axelrod et al. (2011) suggest the use of bilingual cross-entropy differences, which can only be used for filtering parallel data.

Though these techniques are based on minimizing the perplexity, improvement in the SMT translation quality is not assured (Moore and Lewis 2010, Axelrod et al., 2011), as the behavior of SMT systems depends on multiple factors.

3. Data Set

Gathered data was categorized into three, namely, in-domain, pseudo in-domain and out-domain, based on the context and writing styles.

Data gathered from official letters (e.g., from the Department of Education etc.,) was considered as in-domain.

Since the size of in-domain data was small, additional data was gathered from other government sources such as annual reports, parliament order papers, circulars, and establishment codes. Though these were from government institutions, the writing style was different from letters described above (e.g. the parliament order papers were more like question and answer form), thus these were categorized under pseudo in-domain. A reasonable amount of pseudo in-domain parallel data with respect to the in-domain data was collected. Some source documents of in-domain and pseudo in-domain were hard copies in a single language (i.e., either the Tamil or Sinhala version of the document), while some were soft copies in PDF format. The single-language source documents were manually translated and typed. Data from PDF documents were extracted using a custom developed tool. Parallel data was created by using the sentence alignment tool created by Hameed et al. (2017). To make sure that there are no duplicates in the training, tuning and testing sets, duplicate sentences were removed using a custom script. In addition, we collected some monolingual Tamil sentences of this category, where the sources were annual reports.

Other easily accessible data sources were from the web, (such as articles from blogs, news and wiki dumps), and other free sources. This out-domain data was collected from some freely available sources (Ramasamy et al., 2012, Goldhahn et al., 2012) as well as by web crawling. Yet, the context with respect to official government letters, was quite different. Therefore, these were categorized as out-domain data. However, it was possible only to gather monolingual data under this category. Since we had a comparatively larger amount of out-domain data w.r.t. in-domain data, with the motive to use the data that is more relevant to the context, filtration is done based on perplexity measure (Moore and Lewis, 2010). Here, the extraction of relevance sentences (w.r.t the in-domain corpus) was done based on a threshold value. The difference between the perplexities of the sentences on the in-domain based LM and out-domain based LM is considered. Sentences with this difference greater than the threshold value are considered. The threshold value is dynamically set based on the tuning set perplexities. And this filteration process is iterated multiple times.

Two test sets were prepared for evaluations. One set was a set of sentences randomly picked from the collection from where the training and tuning data were derived (Test-1). The other test set sample (Test-2) was from official letters of an office of a university; no data from these letters were used in training or tuning sets. The average sentence lengths of Test -1 were 10.95, 9.90 and Test-2 were 13.94, 11.21 for Sinhala and Tamil, respectively.

Statistics on the parallel data, Tamil monolingual data and Sinhala monolingual data are shown in Table-1, Table-2, and Table-3, respectively. Moreover, in Table-2 and Table-3, in column 4, the perplexity (lower the perplexity higher the similarity between the reference and sample data (Jelinek et al., 1977)) values of each monolingual data source calculated with respect to the tuning set are listed.

Source	S	W (Sinhala)	W (Tamil)
In-domain	6,428	80,849	73,066
Pseudo in-domain	15,645	237,498	197,271
Tuning	1,000	12,740	11,544
Test-1	340	3,724	3,368
Test-2	340	4,740	3,810

S: # sentences, W: # words

Table 1: Sources of parallel data

Source	S	W (Tamil)	Perplexity
In-domain	6,428	80,849	214.6239
Pseudo in-domain	76,692	788,544	415.5571
Out-domain	1,525,966	21,348,157	2210.4860
Filtered out-domain	14,682	178,840	1814.9250

Table 2: Tamil monolingual data

Source	S	W (Sinhala)	Perplexity
In-domain	6,428	73,066	87.4207
Pseudo in-domain	15,646	237,498	604.4787
Out-domain	4,735,658	72,531,342	918.3833
Filtered out-domain	159,597	2,865,591	518.1778

Table 3: Sinhala monolingual data

4. Experiments

The experimental setup was built using the *Moses* (Kohen et al., 2007) PBSMT system. As the publicly available tokenizers did not work for this pair, we used our tailor-made tokenizer for Sinhala as well as Tamil. Parallel data was filtered using standard *Moses* filtrations to remove misaligned sentences and sentence pairs with high length ratio differences.

To generate the word alignment, Giza++ (Och and Ney, 2003), was used with 'grow-diag-final-and' as the symmetrization heuristic and 'msd-bidirectional-fe' as the reordering technique.

'Good Turing' was used as the smoothing technique for the phrase table score smoothing. In addition to phrase translation score; lexical translation scores, word and phrase penalties, and linear distortion were used as features in the TM, which are commonly used features (Kohen, 2009).

An LM of order 5 (5-gram) was created using *SRILM* (Stolcke, 2002). For the setup with multiple data sources, by experimenting on different configurations (single model with all data, log-linear interpolation of multiple LM and linear interpolation), log-linear interpolation of multiple LM was chosen as it gave the best scores. Here individual LMs were created for each type of source, and were used as individual sub modules under LM by giving individual weights. At the time of tuning, these weights were adjusted (based on the relevance to the tuning set translation).

XenC (Rousseau, 2013) was used to do the filtration over the out-domain data. Two separate LMs were created for in-domain and out-domain using SRILM. These LMs were used in calculating the perplexity difference of each sentence between both the models. This difference between the perplexities is used in determining the eligibility of a sentence to be filtered out (a sentence with this value higher than the threshold are filtered out).

Cube pruning techniques available in Moses were used with a stack size of 5,000, and a maximum phrase length of 5.

The feature weights of each model were tuned using *Minimum Error Rate Training* (MERT) (Och, 2003) on 100 best translations of 1000 sentences / phrases.

The baseline was set up only with the in-domain data (letters) as the TM and LM sources. The pseudo in-domain data was added to TM and LM in a step-wise approach and evaluated. On top of this configuration, the out-domain monolingual data was added in unfiltered and filtered forms one by one separately, and the impact was evaluated. Same set of experiments were carried on in either direction of the language pair.

5. Evaluation & Analysis

Performance of each setup was evaluated using two separate test sets, each containing 340 unique sentences/phrases based on BLEU.

Results for the five different experimental setups for both the test sets are tabulated in Table 4 and Table 5 for Sinhala-to-Tamil and Tamil-to-Sinhala, respectively. The BLEU columns show the BLEU scores and the OOV columns show number of OOV in the translated output which is the unique word count of untranslated words.

Setup A is the baseline where the TM and LM are built using only the in-domain data. In setup B, the pseudo in-domain parallel data is added to the TM of the setup A (baseline). Setup C is extended by adding an extra LM to setup B where this new LM is built using the target side of the pseudo in-domain. In setup D, another LM built using the unfiltered out-domain data is added to the setup C. The LM built with the filtered out-domain data added to setup C is represented by setup E.

In addition, another two sets of results are included into Table 4 and Table 5 as Setup F and setup G.

- Setup F: This is the experimental results reported by Tennage et al. (2018) for NMT for the same data set as setup C. Yet they experiment only using one test set (Test-1) and they have not reported on OOVs (so cells in the last 3 columns are left blank in Table 4 and Table 5 for setup F).
- Setup G: This shows the scores calculated on the output of the Google Translate for both of the test sets that are used in our experiments. Since Google Translate drops the OOVs in its output (in some cases, the entire sentences are dropped) OOV cells are left blank for this setup.

Based on BLEU sores (refer columns 3 and 4 in Tables 4 and 5), the baseline SMT system scores are higher than NMT (setup F) and Google Translate (setup G).

When comparing the result of each test set, there is a drastic difference between them as well as in number of OOV. The first set (Test-1) had an abnormally high range of BLEU scores, while the second set (Test-2) had a much lower range of BLEU scores. This can be explained based on the nature of the test data. Test-1 was a subset from where the training and tuning data were derived, while Test-2 does not intersect with the type of sentences that training and tuning data contains. These contrasting results are because of over-fitting of Test-1. Moreover, there is a score difference noticed for Google Translate as well. This may be due to the average length difference as well as source-target length ratio differences of test sets (Average sentence length of Test-1 is lesser than Test-2's. Source-target sentence length different ratio is higher in Test-2 than Test- 1. This makes translating Test-2 more complex than Test-1).

Based on the results for the TM enhancement, by adding the pseudo in-domain parallel data to the system, all four test results (two in either direction) showed a noticeable increase in the BLEU score. As more parallel data was added, the vocabulary of the system increases and improves the word alignment. This improves the TM, which helps to reduce the OOV in both directions (refer to the last two columns in Table 4 and Table 5). In all 4 test cases, the number of unknown words has reduced in Setup B, with respect to Setup A.

Adding pseudo in-domain data to LM results in a slight improvement in the BLEU scores. Here**,** the score was affected by the improvement in the LM, as the system

learns more on language flow (e.g., out of available options, which translation will suit more in a certain context), and more on inflection options. For example, the word "අවසන් වසර" (/awasan wasara/) in Sinhala means 'final year'. However, the word "අවසන්" (/awasan/) alone can mean 'last', 'final', 'ending', or 'finish'. With a small LM (in-domain only), the system translated this Sinhala word into the Tamil word phrase "நிறைவு ஆண்டு" (/niraiwu aandu/), where the word "நிறைவு" (/niraiwu/) bears the meanings 'ending' or 'finish', which is not appropriate. When pseudo in-domain LM is added, this word got translated to the correct term "இறுதி ஆண்டு" (/iruthi aandu/), where the meaning is 'final year'.

However, the integration of out-domain data reduced the BLEU scores in both filtered and unfiltered cases in either direction. Moreover, the result of filtered data was inferior to that of unfiltered data for all the test cases though the perplexity of the filtered data was lower than that of unfiltered data (refer column 4 in Table 2 and Table 3).

Source	Setup	**BLEU**		**OOV**	
		Test-1	Test-2	Test-1	Test-2
Baseline	A	36.32	7.55	375	645
A + PIS :TM	B	**37.01**	9.13	285	505
B + PIS : LM	C	**37.01**	9.17	285	505
C + UOD:LM	D	36.26	**9.18**	285	505
C + FOD: LM	E	36.06	9.17	285	505
NMT	F	6.78	-	-	-
Google	G	14.49	7.38	-	-

PIS: pseudo in-domain, UOD: unfiltered out-domain, FOD: filtered out-domain

Table 4: Experimental results for Sinhala-to-Tamil

Source	Setup	**BLEU**		**OOV**	
		Test-1	Test-2	Test-1	Test-2
Baseline	A	45.19	12.46	555	937
A + PIS :TM	B	**46.64**	13.15	438	760
B + PIS : LM	C	46.30	**13.17**	438	760
C + UOD:LM	D	46.01	12.46	438	760
C + FOD: LM	E	45.98	11.70	438	760
NMT	F	6.84	-	-	-
Google	G	14.96	7.69	-	-

Table 5: Experimental results for Tamil-to-Sinhala

This kind of behavior has been mentioned by Moore and Lewis (2010) as well. One reason they mentioned is that as the perplexity reduces, more weight is given to the out-domain LM, though still the writing style is drastically different. This can be the reason even in our experiments as well. For example, the phrase "சிக்கலான நிலையை உருவாக்கியுள்ளது." (/sikkalaana nilaiai uruwahiyulathu/) means 'has created a problematic situation', where the word "சிக்கலான" (/sikkalaana/) has meant to be 'problematic' though it can also take the meaning as 'complex', 'issue', or 'conflict'. Without out-domain data, the phrase is translated as "ගැටළු සහගත වී ඇත" (/gatalu sahagatha wee atha/), which is the correct translation. However, when the out-domain data is added, system translates it as "සංකීර්ණ වී ඇත" (/sankeerna wee atha/), which means 'has become complex' which is not the proper translation according to the context.

Perplexity values of the Tamil out-domain corpus (2210.4860) are comparatively higher than that of Sinhala out-domain corpus (918.3833). Tamil out-domain corpus had more blog articles where the writing style was more informal with more colloquial style data. Sinhala out-domain corpus had more news articles where the writing style is less colloquial. This could be the reason for this high variation in perplexity values.

6. Conclusion

This paper presented a domain-specific SMT system for the low-resourced language pair, Sinhala and Tamil. Evaluations were carried on analyzing the impact of different data types in improving the system for domain-specific SMT, as the available in-domain data was minimal and coming from few government institutions. Results show that the use of pseudo in-domain data gave positive results in TM and a less significant improvement for LM. However, the use of out-domain monolingual data did not improve the performance in unfiltered or filtered form, while the filtered data resulted in inferior results to unfiltered data. Further, results obtained with an NMT system as well as from Google Translate highlight the effectiveness of our SMT work.

This conveys the message that for our domain of consideration, data from sources that have informal writing such as news and blogs will not add value. Furthermore, the results reveal that the system requires quality data in higher quantities from diverse subject matters and sources (e.g., numerous government institutions), to perform better.

In future, with more data, we plan to experiment on dynamic model adaptation based on the context of the letters, as suggested by Foster and Kuhn (2007). We hope that this will help to get a fine-tuned system that can be used for better translation of letters from different government institutions based on the context.

7. Acknowledgement

The authors of this paper would like to extend their gratitude to the Senate Research Council (SRC) Grant of University of Moratuwa and the Department of Official Languages (DOL) for the funding provided for the research, and the individuals who contributed in data gathering and corpus preparation.

8. Bibliographical References

Axelrod, A., He, X. and Gao, J., (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*, pages 355-362. Association for Computational Linguistics.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V., Monz, C. and Negri, M., (2016). Findings of the 2016 Conference on Machine Translation. In *ACL 2016 first conference on machine translation (WMT16)* pp. 131-198. The Association for Computational Linguistics.

Civera, J., & Juan, A. (2007). Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177-180. Association for Computational Linguistics.

Eck, M., Vogel, S., & Waibel, A. (2004). Language Model Adaptation for Statistical Machine Translation Based on Information Retrieval. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 327–330.

Foster, G., & Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the second workshop on statistical machine translation*, pages 128-135. Association for Computational Linguistics.

Gao, J., Goodman, J., Li, M., & Lee, K. F. (2002). Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1), 3-33.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 759-765.

Hameed, R. A., Pathirennehelage, N., Ihalapathirana, A., Mohamed, M. Z., Ranathunga, S., Jayasena, S., Dias, G., & Fernando, S. (2016). Automatic Creation of a Sentence Aligned Sinhala-Tamil Parallel Corpus. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)*, pages 124-132

Hildebrand, A. S., Eck, M., Vogel, S., & Waibel, A. (2005). Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of European Association for Machine Translation (EAMT)*, Vol. 2005, pages 133-142.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1), S63-S63.

Koehn, P., & Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, pages 224-227. 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., 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 (ACL)*, pages 177-180. Association for Computational Linguistics.

Koehn, P. (2009). Statistical machine translation. Cambridge University Press.

Koehn, P., & Knowles, R., (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

McLachlan, G., & Peel, D. (2004). Finite mixture models. John Wiley & Sons.

Moore, R. C., & Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference short papers*, pages 220-224. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Volume 1 pages 160-167. Association for Computational Linguistics.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311-318. Association for Computational Linguistics.

Pushpananda, R., Weerasinghe, R., & Niranjan, M. (2014). Sinhala-Tamil Machine Translation: Towards better Translation Quality. In *Australasian Language Technology Association Workshop 2015*, pages 123-133.

Pushpananda, R., Weerasinghe, R., & Niranjan, M. (2015, April). Statistical Machine Translation from and into Morphologically Rich and Low Resourced Languages. In *Conference on Computational Linguistics and Natural Language Processing (CICLing)*, pages 545-556. Springer.

Ramasamy, L., Bojar, O., & Žabokrtský, Z. (2012). Morphological processing for English-Tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pages 113-122.

Rousseau, A. (2013). Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100(1):73–82.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*, pages 901– 904.

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., Ranathunga, S., Jayasena, S., and Dias, G. (2017). Neural Machine Translation for Sinhala and Tamil Languages. In *Proceeding of 21st International Conference on Asian Language Processing (IALP)*.

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., and Ranathunga, S. (2018). Handling Rare Word Problem using Synthetic Training Data for Sinhala and Tamil Neural Machine Translation. *11th International Conference on Language Resources and Evaluation (LREC)*.

Weerasinghe, R. (2003). A statistical machine translation approach to Sinhala-Tamil language translation. *Towards an ICT enabled Society*, pages 136-141.