English to Sinhala Machine Translation: Towards Better information access for Sri Lankans

Jeevanthi Liyanapathirana
University of Colombo
School of Computing
Sri Lanka
juliyanapathirana@gmail.com

Ruvan Weerasinghe
University of Colombo
School of Computing
Sri Lanka
arw@ucsc.cmb.ac.lk

Abstract

Statistical Machine Translation is a well established data driven approach for automatic translation between different languages. However considerably few researches have taken place on Sri Lankan as well as Asian Languages. We research on possible English-Sinhala Translation using Statistical Machine Translation. Special attention is paid for the impact of parameter tuning on Sinhala Language. Results indicate that adequate parameter tuning to overcome structural differences existing in the language pair yields a satisfactory performance, providing valuable insight towards further research in this area.

1 Introduction

Machine Translation can be stated as an attempt to convert text from one source language to another target language using an automated procedure. The need of machine translation is visually evident day by day, with the need of overcoming the language barrier and communicating with different communities.

Machine Translation approaches range from rule based approaches to data driven approaches. Statistical Machine Translation (SMT) is a data driven approach, and is based upon statistical models which are build upon a bilingual corpora. Parameters of this models are derived based on the statistical properties of the corpora. SMT focuses on building the needed models and discovering and experimenting with different model parameters which improve the output obtained.

However, most experiments have been conducted among European languages. Efficient model generation and optimized parameter values have been calculated for those languages, leaving the way open to improvement of quality on those languages.

2 Background and Related Work

Sri Lanka is a country where three main languages are being spoken: Sinhala, Tamil and English. The deficiency observable is that most of the people who speak Sinhala or Tamil are not well aware of English: leading to barriers for information access, as well as ethnic group misunderstandings. Hence it would be a good approach to analyze how successful an application of Statistical Machine Translation be on English to Sinhala Language Translation.

Not much research has been conducted in this area regarding Sinhala Language. Weerasinghe (2004) has made a comparison in English-Sinhala translation and Tamil-Sinhala translation using SMT, arriving to the conclusion that languages which are closer in their origins do perform better in this approach. Developing an Asian-English translation is described by Ramanadan et al. (2006). An English-Tamil translator development is conducted by Germann (2001). Other attempts include translation between European languages. (Brown et al. (1990), Jones and Eisele (1983))

Section 3 and 4 would describe the basic SMT process and parameter tuning. Section 5 would describe the translation process. Experimental Setup would be explained in Section 6, followed by Results and Discussions in section 7 and the ending conclusion in Section 8.

3 The SMT Model

Given a source sentence f and target sentence e, the basic equation for getting the best possible target sentence is (Brown et al. (1983)):

$$P(e|f) = \frac{P(e)P(e|f)}{P(f)}$$

This provides the following estimation for get-

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

ting the maximum possible target output sentence,

$$P(e|f) = argmax(P(e)P(f|e))$$

The $p(e)$ component is the language model, which takes care of the fluency of the target sentence. $p(f|e)$ is the translation model, which proves the most possible translation for the given sentence. By decoding these two models gaining of a fluent as well as a correct output is expected.

Phrase Based Translation is also being used in place of word based translation. Here, the input text would be split to phrases, and translation would take place considering phrases.

4 The Log Linear Approach

Above decoding with two models has now been improved with added feature functions, along with which additional features would be added to the decoding stage to improve the translation process, as done by Rui and Fonollosa (2005).

$$P(e|f) = argmax(\lambda_m h_m(e, f))$$

h_m would be the system models: language model, translation model. However, the log linear model provides the ability to add up any feature function other than those two to improve the translation output. The weight assigned to each feature function is defined by λ_m.

5 Translation Process

Apart from data collection, the major components are the Language model, Translation Model and the Decoder.

5.1 Language Model

The language model in this case would be built in Sinhala. The experiments involved finding out the best smoothing technique (Chen and Goodman (1999)), experimenting backing off and interpolating models together, generating domain specific as well as a general language model and using perplexity to evaluate their quality. Smoothing techniques such as Good Turing Discounting, Natural Discounting, Kneyser Ney Discounting and Witten Bell Discounting has been used.

5.2 Translation Model

IBM models were generated as translation models and HMM models instead of IBM model 2 was also experimented to check their impact on Sinhala Language.

5.3 Decoder Model

Following Koehn et al. (2003), we expect to make use of phrase based translation rather than word based translation for our experiments, and intend to add additional feature functions to improve its performance.

With an aim to have one to many mappings between source and target words, bidirectional alignment translation models were generated. Different alignment strategies were experimented to check their impact on Sinhala Language: Intersection, Union, Grow-Diag-Final were generated between English and Sinhala. To cater the word order difference inherent in two languages, distortion models were generated to analyze their performance on our research.

Many additional features derived likewise were integrated with the simple language model and translation model ,via the log linear approach. These additional features are integrated in Moses (2007) toolkit. Thus, the additional feature functions used would be:

- Phrase Translation Probability $P(f|e)$
- Inverse Phrase Traslation Probability
- Lexical Probability
- Inverse Lexical Probability
- Distortion Model Probabilities
- Word Penalty $\omega^{length(e)}$
- Phrase Penalty (Constant 2.718)

The decoding equation is

$$e_{best} = argmax(P(F|E)P(LM)\omega^{length(e)})$$

The phrase translation probability component $p(F|E)$ would now be a compact translation model with all features integrated and weighted appropriately, which would be combined with the Language Model probability and word penalty.

To address the fact that not all words would be included in phrase pairs, integrating word based translation with phrase based translation was also experimented.

5.4 Minimum Error Rate Training (MERT)

After all the decoders are generated, MERT provides a way to optimize the weights that has been given to the feature functions in the decoding equation. MERT does this with the help of a development set , where a source language corpus with its reference translation would be used to provide the optimum weight setting for the decoder. In order to accomplish that, once the decoders were generated, MERT procedure was conducted. This was to check what the impact of MERT procedure on a Sinhala output would be.

6 Experimental Setup

6.1 Data collection and Data Preprocessing

For the language model, Sinhala Data files were gathered from the WSWS web site (www.wsws.org) (political domain), and the other was the UCSC/LTRL Beta Corpus which had data from several domains. For the Translation Model, the main data source was the WSWS web site. This contained documents in both English and Sinhala language. Data Preprocessing was conducted using script files to clean and align the data extracted from web pages.

6.2 Tools Used

6.2.1 Language Model

SRILM toolkit with its language model specific tools was used for this purpose.

6.2.2 Translation Model

IBM models including HMM model was generated using GIZA++.

6.2.3 Decoder

Moses, a beam search decoder providing ability to include additional feature functions was used for this purpose.

6.2.4 Evaluation

BLEU Metric by Papineni et al. (2001) was chosen as the evaluation metric for our experiments.

7 Results and Discussion

7.1 Language Model

The statistics of data gathered for the language models from different domains are as in Table 1. The best smoothing method was determined by calculating perplexity of relevant test sets from each domain (Table 2). Both combining data of a specific domain together to build a large LM as well as interpolating small LM s were experimented (Table 3), from which we concluded that Unmodified Kneyser-Ney Discounting along with interpolation is the best to be used with Sinhala data.

Domain	Sentences	Unique Words
Political	17570	44652
Feature	24655	70690
Foreign	4975	24999
Other	2317	9378

Table 1: Language Model Data Statistics.

	Political	Feature	Foreign	Foreign
KN	604.46	825.61	744.11	601
UKN	584.63	763.37	712.88	582.25
UKN+INT	**572.23**	**723.31**	**689.99**	**579.62**
WB	649.02	814.46	766.61	664.13
WB+INT	651.53	799.53	767.62	662.45

Table 2: Perplexities against In-Domain Test Sets

Corpus	Interpolation Weight			
	0.0	0.4	0.5	0.8
Mixed LMs	572.23	**563.35**	569.81	596.62
LM from Combined Corpus	603.93			

Table 3: Mixing and Interpolating LMS

7.2 Translation Model and Decoding

Tabel 4 shows the statistics for parallel data gathered for building the translation model. The whole corpus was divided into training and test data sets of 50, 100 sentences accordingly.

Different alignment strategies and reordering strategies together were experimented with test sets of 50 sentences (Table 5). The reordering strategies used are msd-bd-fe, msd-bd-f, msd-fe, msd-f in these experiments.

The best configuration (7.0756) was then experimented with varying distortion limits for translations and dropping unknown words, which eventually increased the Bleu score by around 1 point (Table 6) (to around 8.11). The above found best configuration (grow-diag-final,msd-bdf, distortion limit 20, drop unknowns) was then used with the tuning set of 2000 sentences for MERT, to get optimum weights for different features. The test set performance was again experimented with the gained optimized weights.(Table 7).

	Sentences	Unique Words
WSWS Corpus	10000	32863

Table 4: Translational Model Data

Type	msd-bd-f	msd-bd-fe	msd-fe	msd-f
grow-diag-final	**7.0756**	6.9899	6.8215	6.9147
grow-final	6.8838	6.8808	6.7891	6.8244
union	6.4264	6.3856	6.6708	6.3936
grow-diag-final	6.8530	6.9815	6.8705	7.0341

Table 5: BLEU Values against alignment and reordering strategies

DL	0	4	8	12	20	20du	25du
BLEU	3.76	6.93	7.03	7.21	7.33	**8.11**	8.05

Table 6: BLEU values against different distortion values and dropping unknown words

7.3 Improving Phrase with Lexical Probabilities

In order to reduce the risk of having too many unknown words during decoding, word based translation was integrated along with phrase based translation approach. A very low weight was allocated to this new feature (the value given whenever a word based translation is used - in this case we will name it as lexical penalty.) so as to give priority to phrase probabilities. Table 8 shows the resulting Bleu score for 50 sentences, and then the same thing was experimented with best distortion limits and unknown words dropped (Table 9) .

With this, the performance reached almost 14 in BLEU score (grow-final, msd-bd-f, distortion limit 20 and unknown words dropped). This shows that lexical translation probabilities add a considerable effect by adding itself as a certain word based support for the phrase based translation.

The next experiment was varying the weight/value allocated whenever a word based translation (lexical penalty) occured, to check whether it would impact the Bleu score. Table 10 clearly shows that the value given for the lexical penalty has played a considerable role in the Bleu value, resulting in a final Bleu score of 15.06. This shows that integrating word based translation along with phrase based translation with certain lexical penalties do impact the translation performance in a very positive manner, though the current analysis is not enough to successfully point out that this specific value would be good to be used as the lexical penalty value in English to Sinhala Translation.

Tuning Set size	Original BLEU	After MERT
2000	8.11	**9.26**

Table 7: Impact of MERT on BLEU

	grow-diag-final	grow-final	union
msd-bd-fe	11.6959	6.9899	11.2667
msd-bd-f	12.3110	11.5597	6.7891
msd-fe	11.6907	10.7907	10.7641
msd-f	11.3744	10.7183	10.5767

Table 8: BLEU for configurations with added lexical probability

8 Conclusion and Future Work

The purpose of this research was to find out how English to Sinhala Translation can be accomplished using SMT techniques. Alignment models, reordering models, integrating phrase models with word models and distortion limits together built up a considerably significant increase in performance. All these add up to the conclusion that the structural differences inherent in these two languages can be overcome to a certain extent via appropriate parameter tuning. This is specially visible via the increase in Bleu score upon varying reordering models and distortion limits. Previous research (Weerasinghe , 2004) turned out to provide a Bleu score of around 2-6, where as this research yielded a Bleu score of around 15 , symbolizing an impressive insight on future work in this area.

Possible future work might involve using larger corpora for Sinhala Language. Another possible research would be to find out better phrase extraction heuristics for the phrase based translation model, so that the decoding process relies on the word based model as less as possible. This would make the translation process be relying more on phrase based translation, which would eventually

	Grow		grow-final		union	
	dl20	+du	dl20	+du	dl20	+du
msd-bd-fe	12.73	13.55	13.36	13.66	12.39	12.31
msd-bd-f	13.27	13.69	13.75	**13.99**	12.23	12.32
msd-fe	12.63	13.14	12.99	13.23	11.08	11.29
msd-f	12.43	12.91	13.35	13.18	11.10	11.74

Table 9: BLEU Score for phrase tables with added lexical probabilities : DL20

Corpus	Lexical Penalty Value			
	0.001	0.005	0.01	0.2
Test Set 50 Configuration	13.99	14.49	**15.06**	13.84
Test Set 100 Configuration	11.16	**11.76**	11.62	11.43

Table 10: Bleu Scores for best decoder configuration with varying lexical penalty : DL20

result in better and fluent translation outputs.

To conclude with, this research can be stated as a promising start towards better information access for rural communities via English to Sinhala Machine Translation. This resesarch can be conducted to improve the performance even more by finding out other possible features which can be added to this process. In a more general point of view, this research can also be stated as an extensive research on machine translation between structurally dissimilar languages.

Acknowledgment

Authors would like to acknowledge the Language Technology Research Laboratory of University of Colombo School of Computing, Sri Lanka for providing human and physical resources to carry out this research.

References

Brown, P. F., Cocke, J., Della-Pietra, S. A., Della-Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer R. L. and Roossin, P.S. 1990. A Statistical Approach To Machine Translation. *Computational Linguists*.

Brown, P.F., Della-Pietra, S. A., Della-Pietra, V. J. and Mercer, R. L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation *Computational Linguistics*.

Chen, S. and Goodman, J. 1999. An empirical study of smoothing techniques for language modelling. *Computer Speech and Language*,

Germann, U. 2001. Building a Statistical Machine Translation System from Scratch: How Much Bang Can We Expect for the Buck? *Proceedings of the Data-Driven MT Workshop of ACL-01, Toulouse. France*.

Jones,D. and Eisele, A. 2006. Phrase-based Statistical Machine Translation between English and Welsh *International Conference on Language Resources and Evaluation, 5th SALTMIL Workshop on Minority Languages: Strategies for developing machine translation for minority languages, Genoa, Italy*.

Koehn, P. , Och, F. J.and Marcu, D. 2003. Statistical Phrase Based Translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada*.

Moses. 2007. Moses - a factored phrase based beam-search decoder for machine translation. *http://www.statmt.org/moses/*,(cited 2011.01.10).

Papineni, K. A., Roukos, S., Ward, T. and Zhu, W.J. 2001. Bleu : A method of automatic evaluation for machine translation. *Publications Manual*. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Ramanadan,A. , Bhattacharyya, P., Sasikumar, M. and Shah, R. 2006. Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU.

Rui, M. and Fonollosa, A. 2005. Improving Phrase-Based Statistical Translation by modifying phrase extraction and including several features *Proceedings of the ACL Workshop on Building and Using Parallel Texts.Ann Arbor:June 2005.Association for Computational Linguistics*.

Weerasinghe, A. R. 2004. A Statistical Machine Translation Approach to Sinhala-Tamil Language Translation. *SCALLA*.