Towards Sinhala Tamil Machine Translation

Randil Pushpananda[#1], Ruvan Weerasinghe[#2], Mahesan Niranjan[*3]
[#]*Language Technology Research Laboratory*
University of Colombo School of Computing, Sri Lanka
[*]*School of Electronics and Computer Science*
University of Southampton, Highfield, Southampton SO17 1BJ, UK

[1]rpn@ucsc.lk, [2]arw@ucsc.lk, [3]M.Niranjan@Southampton.ac.uk

***Keywords*— Statistical Machine Translation, Sinhala Language Processing, Tamil Language Processing, Language Modeling**

INTRODUCTION

Statistical Machine Translation is a well established data-driven approach to translate source language text to target language text using statistical methods using bilingually aligned corpora. However there has been little research in this area for less well-resourced languages such as Sinhala and Tamil.

The focus of this research is to investigate how translation performance varies with the amount of parallel training data in order to find out the minimum needed to develop a baseline machine translation system for the Sinhala-Tamil language pair.

Experiments and Results

This section presents the main steps followed in building the statistical models needed for the translator. A well-known open source phrase based statistical machine translation system (MOSES) was used to build the language and translation models for Sinhala-Tamil, Tamil-Sinhala, French-English and German-English language pairs. The BLEU (Bilingual Evaluation Understudy) evaluation metric was used as the evaluation metric of the translation systems.

Due to the lack of resources and sparseness of the available parallel data in the Sinhala-Tamil language pair, preparing a parallel corpus has proved extremely difficult. However, we were able to extract 1006 Sinhala-Tamil parallel sentences from the Sinhala translation of the book *An introduction to Spoken Tamil* written by J. W. Gair et al. French-English and German-English parallel sentences were extracted from the Europarl corpus. Since the sentence lengths of the Sinhala-Tamil parallel corpus were all between 1 and 15, a similar length restriction was placed on the French-English and German-English parallel corpora used in this experiment. All corpora were divided randomly into testing, tuning and training sets such that 53 parallel sentences were set apart for testing, 53 for tuning and rest for training. From the parallel training sentences 400, 600 and 800 random sample sentences were extracted for three experiments with Sinhala-Tamil, French-English and German-English language pairs for building the translation model. The same parallel sentences extracted for the Sinhala-Tamil pair was also used for building a Tamil-Sinhala translation model.

The next step was to collect monolingual corpora for each target language to build their language models. A Sinhala monolingual corpus was extracted from UCSC-LTRL 10 million word corpus, a Tamil monolingual corpus from a 3 million word news corpus and an English corpus from the Europarl parallel corpus.

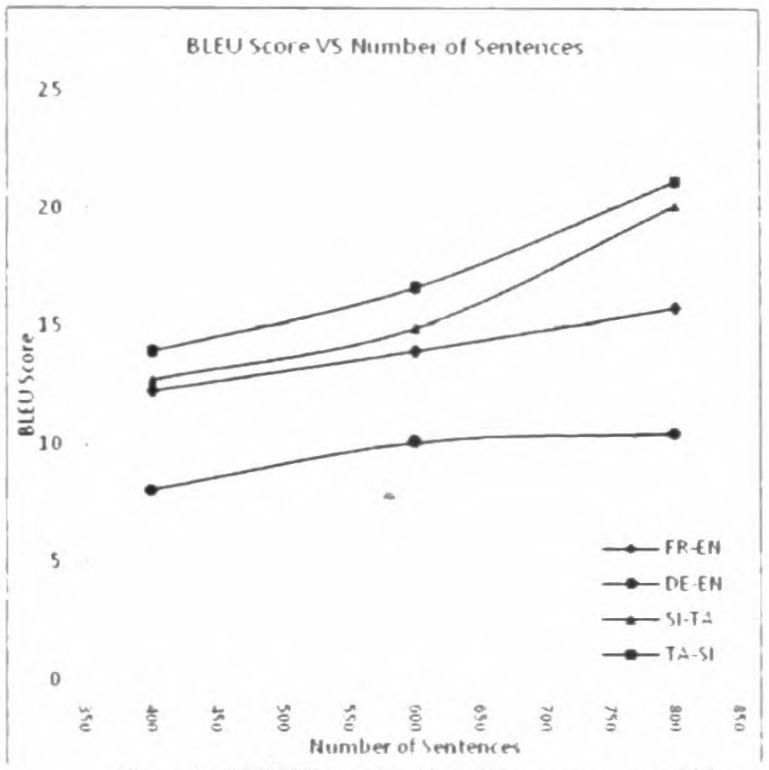

Figure1: BLEU Score Vs No of Sentences up to 800

Figure 1 shows how the BLEU Score values vary against the number of parallel sentences up to 800 for each language pair after training, tuning and testing processes. According to Figure 1, we can see that the BLEU performance measure values of all translations increase as expected when the data size increases from 400 and 600 to 800. More importantly for our experiment, we observe that translation quality results of the Sinhala-Tamil and Tamil-Sinhala language pairs improve at a higher rate than those of the French-English and German-English language pairs with this limited amount of data. The close alignment between Sinhala and Tamil word order and sentence structure could be the main reason for this improved performance.

By comparing the Sinhala-Tamil and Tamil-Sinhala translation results with French-English and German-English language pair translations, we have positive evidence for expecting statistical machine translation to perform well for translating between the Sinhala and Tamil languages. Using this approach, we plan to implement a system capable of producing acceptable translations between Sinhala and Tamil for use by the wider community.