A Statistical Machine Translation Approach to Sinhala-Tamil Language Translation

Ruvan Weerasinghe
Department of Computation and Intelligent Systems
University of Colombo School of Computing
Colombo, Sri Lanka.
<arw@ucsc.cmb.ac.lk>

Abstract

Data-driven approaches to Machine Translation have come to the fore of Language Processing Research over the past decade. The relative success in terms of robustness of Example Based and Statistical approaches have given rise to a new optimism and an exploration of other data-driven approaches such as Maximum Entropy language modeling. Much of the work in the literature however, largely report on translation between languages within the European Family of languages. This research is an attempt to cross this language family divide in order to compare the performance of these techniques on Asian languages. In particular, this work reports on Statistical Machine Translation experiments carried out between language pairs of the three major languages of Sri Lanka: Sinhala, Tamil and English. Results indicate that current models perform significantly better for the Sinhala-Tamil pair than the English-Sinhala pair. This in turn appears to confirm the assertion that these techniques work better for languages that are not too distantly related to each other.

Keywords: Statistical Machine Translation, Language Modeling, Sinhala Language Processing, Tamil Language Processing.

1. Introduction

Machine processing of Natural (Human) Languages has a long tradition, benefiting from decades of manual and semi-automatic analysis by linguists, sociologists, psychologists and computer scientists among others. This cumulative effort has seen fruit in recent years in the form of publicly available online resources ranging from dictionaries to complete machine translation systems. The languages benefiting from such exhaustive treatment however tend to be restricted to the European Family, most notably English and French.

More recently, however, the feasibility of data driven approaches in the context of today's computing power holds out hope for the rest of us. These are those of us concerned with less studied languages, who have few or no linguistic resources to assist us, and for whom the cost of building these up from scratch is prohibitive.

1.1 Background

Sinhala is a language of around 13 million people living in Sri Lanka. It is not spoken in any other country, except of course by enclaves of migrants. Being a descendant of a spoken form (Pali) of the root Indic language, Sanskrit, it can be argued that it belongs to the large family of Indo-Aryan languages.

Tamil, the second most spoken language in Sri Lanka with about 15% of its 18 million people counting as native speakers, is also spoken by some 60 million Indian Tamils in Southern India. The dialects of Indian Tamil however differ significantly to those of Sri Lankan Tamil to cause difficulty in performing certain linguistics tasks [1]. Though originating in India, it does not share the Indo-Aryan heritage of Sinhala, but rather is one of the two main languages in another (unrelated) family known as Dravidian.

Over many centuries of co-existence especially within Sri Lanka, these two languages have come to share many of their structures and functions, thus causing them to be closer to each other than their family origins indicate. English, the link language of Sri Lanka, on the other hand, has had a relatively short co-existence with the other two languages, and though also belonging to the large Indo-European language family, has much less in common with Sinhala. Within the context of the past two decades of ethnic strife between the two primary lingua-cultures within Sri Lanka, any prospect that holds some promise of better understanding of each others language becomes a worthwhile task to aim at.

Towards this aim, this research is an exploration of the feasibility of a non-knowledge intensive approach to machine translation.

1.2 The Three Languages: Sinhala, Tamil and English

By far the most studied languages in terms of descriptive as well as computational models of linguistics have been the European ones – among them English receiving more attention than any other. From lexical resources, part-of-speech taggers, parsers and text alignment work done using these languages, fairly effective example-based and statistical machine translation efforts have also been made.

In contrast, electronic forms of Sinhala became only possible with the advent of word processing software less than 2 decades ago. Even so, being ahead of standardizing efforts, some of these tools became so widely used, that almost 2 years after the adoption of the ISO and Unicode standard for Sinhala (code page 0D80), there is still no effort to conform to it. The result is that any effort in collecting and collating electronic resources is severely hampered by a set of conflicting, often ill-defined, non-standard character encodings.

The e-readiness of the Tamil language lies somewhere in between the two extremes of those of English and Sinhala. Here the problem seems to be more the number of different standardization bodies involved, resulting in a plethora of different standards. TASCII was one of the earliest attempts and is an 8-bit Tamil encoding, while the newer Unicode standard is at code page 0B80. Owing to this, available electronic resources may be encoded in unpredictable ways, but tools for translating between non-standard encodings and the standard(s) exist.

1.3 Scope

This work builds on Weerasinghe [6] which explored the application of Statistical Machine Translation between Sinhala and English by proposing a bootstrapping approach to build up the relevant resources. The results reported therein not being too promising, this research is an attempt to find out if the reasons are to do with Asian language characteristics or the 'linguistic distance' between the language pair chosen to be translated.

Towards this end, in this work we have undertaken the task of Statistical Machine Translation between

Sinhala and Tamil using the same Sinhala corpus and its translated Tamil version for the learning process. The models thus built would then permit a fair comparison in order to determine the similarity between English-Sinhala translation and Sinhala-Tamil translation.

2. Why Statistical Machine Translation?

Machine Translation as one of the major sub-disciplines Natural Language Processing and Computational Linguistics has as long a tradition as does its parent discipline. Much of the work in the field used a fundamentally knowledge-based approach to the problem, just as they did with Language Processing in general. It was becoming clear by this time that existing (knowledge-based) systems could only prove useful for 'toy' worlds and artificially controlled language constructs. Their scalability to the analysis or translation of naturally occurring speech or text was in serious question.

The mid 1980's saw somewhat of a resurgence in the use of data-driven and statistical approaches to Language Processing in general reflecting its increasing importance in the whole area of Artificial Intelligence. This could be attributed to its effectiveness in learning and adaptability over traditional knowledge-based approaches. By the early 1990's this revolution had also lead to new efforts into casting the Machine Translation problem also in terms of the data-driven paradigm.

While Example-Based Machine Translation (EBMT) is an example of this approach as applied to the area, the dominant theoretical foundation for such work derived from statistics and information theoretic ideas and came to be known as Statistical Machine Translation (SMT) in general. The most influential work in this regard came from work at IBM's research labs by Brown et.al. [2].

Knowledge-based approaches to Language Processing, in common with other AI-related tasks, have never been able to satisfactorily solve the knowledge acquisition bottleneck. In Language Processing this translates to spending much time and effort on agreeing on linguistic niceties in order to produce the large linguistic resources required by such an approach.

Data-driven approaches on the other hand require only fundamentally un-pre-processed ('raw' forms of) input such as corpora of naturally occurring text from which machine learning can be performed. This

makes Statistical Machine Translation an attractive alternative for translation between language pairs for which little or no electronic linguistic resources exist. Sinhala and Tamil are among the many world languages for which this is true.

2.1 The basic SMT model

Statistical Machine Translation is founded upon the assumptions of the Noisy Channel Model and Bayes Rule which help 'decompose' the complex probabilistic model that needs to be built for estimating the probability of a sentence in a source language (f) being translated into a particular target language sentence (e).

Using the notation common in the literature this decomposition can be stated as:
`P(e|f)=P(e)*P(f|e)/P(f)`

Since predicting in a statistical model corresponds to identifying the most likely translation, maximizing the above over all possible target sentences (e) gives the estimation:
argmaxeP(e|f)=argmaxeP(e) *P(f|e)

The main benefit gained by the above decomposition is that the burden of accuracy is moved away from the single probability distribution P(e|f) to two independent probabilities P(e) and P(f|e). The former is known as the 'language model' (for language e) while the latter is known as the 'translation model' (for predicting source sentences, f , from target sentences e).

While it would be impossible to estimate such a language model, the literature on using n-gram (mainly bi-gram and tri-gram) models for estimating sentence probabilities of a given language have matured over the past two decades. The estimation of the translation model would not be too difficult if machine readable dictionaries with frequency statistics were available. While this is impractical for even the most well studied languages, the dependence of such counts on the genre of the texts under consideration make it less than optimal.

This is where work carried out by Brown et. al. [2] at IBM stepped into providing a bootstrapping model building process. Beginning with the very simple word-for-word translation lexicon building models (IBM Models 1 and 2), this process constructs ever more sophisticated Models (3, 4 and 5) which account for more and more flexibility in the underlying assumptions (e.g. a single word in the source language may be translated by more than a single target word, and may appear in another part of the sentence).

Intuitively, once the translation model performs its task of predicting a set of possible (good and bad) candidate translations for a particular source sentence, the (target) language model will calculate the probability of such sentences being acceptable in the language in order to select the best translation. It is this 'sharing of the burden of accuracy' between the two models that has been at the heart of the relative success of the SMT approach.

2.2 The SMT process

The SMT process at its very heart requires the compilation of a bi-lingual corpus. The corpus in addition needs to be 'sentence aligned': each sentence in the target language must have an identified equivalent source language sentence to which it must be aligned in some way. While this process can be performed manually current research has promising results on automating this process.

The complete SMT process involves (a) the building of a target language model, (b) the construction of the translation model, (c) the decoding process and (d) the process of scoring resultant translations.

While a bi-lingual parallel corpus is the primary and only resource that is needed for applying the SMT process to the language pair concerned, there is no theoretical necessity for the (target) language model, P(e) to be constructed just from its portion in the bi-lingual corpus. It is common practice to augment this with an expanded target language model in order to improve the overall model.

The CMU-Cambridge Toolkit† [4] can be used to build such a target language model from a minimally tagged monolingual plain text corpus based on n-gram statistics.

The second component of the SMT process is the building of the translation model as outlined in section 2.1. The IBM models described have been more recently improved in terms of efficiency and made available in the public domain by Al-Onaizan et.al. [3]. This GIZA system provides a set of tools that facilitate the building of translation models from scratch‡.

---

† Downloadable from svr-www.eng.cam.ac.uk/ ~prc14/toolkit.html
‡ Downloadable from www.clsp.jhu.edu/ws99/

Once the translation and language models have been constructed from the training data, the combined model needs to be applied to new test data in order to determine the outcome of the translation process. Since the maximization concerned requires exploring an entire word trellis, a decoding approach is required to determine the highest scoring sentence hypothesis. For this process we used the publicly available ISI-Rewrite decoder with various different parameters and smoothing methods in order to arrive at the best possible scheme for the respective language pair.

While this completes the entire SMT process, the evaluation of the output produced by the system requires a metric that can be applied to any given language pair in order to be able to compare results of different approaches. While many such metrics have been used, they have mostly been based on human judgment and thus not reproducible exactly.

Papineni et. al. [5] suggested a completely automatic metric they referred to as BLUE which is based on the Word Error Rate (WER) metric popular in Speech Recognition. This metric scores between 0 and 1 for any potential translation of a sentence by comparing it to (possibly multiple) professionally translated 'reference translations'.

3. Sinhala – Tamil SMT

While not requiring large hand-crafted linguistic resources agreed on by linguistic experts, SMT does require a reasonably large parallel bi-lingual corpus whose translations are fairly faithful to its purpose. So, for instance, many obvious sources turn out not to be good candidates for the compilation of such a corpus.

In the Sinhala-English case, as reported in [6], it was found that a major newspaper of Sri Lanka publishing in both languages has two different reporters for each event who report news in very different ways. Even in the context of articles translated from one to the other, it has been found that in most cases, translators have been given full freedom to 'interpret' such articles afresh and so end up with translations that have quite different sentence and even paragraph structures. All such candidate sources turn out to be unsuitable for the SMT task. Weerasinghe [6] identified a website (www.wsws.org) as a possible source for the compilation of a Sinhala-English parallel corpus and we here discovered a superset of this corpus as a suitable candidate for a trilingual Sinhala-Tamil-English parallel corpus.

A set of WSWS articles available on the site during 2002, containing translations of English articles into Sinhala and Tamil, were selected to form a small tri-lingual parallel corpus for this research. This consists of news items and articles related to politics and culture in Sri Lanka.

3.1 Basic Processing

The fundamental task of sentence boundary detection was performed employing a semi-automatic approach. In this scheme, a basic heuristic was first applied to identify sentence boundaries and those situations that were exceptions to the heuristic identified. These were then simply added to an 'exceptions list' and the process repeated. This process proved adequate to provide accurate sentence boundary detection for the Sinhala and Tamil corpora.

Automatic sentence alignment proved to be unsuccessful as reported in [6] and hence a manual alignment of sentences was carried out.

After cleaning up the texts and manual alignment, a total of 4064 sentences of Sinhala and Tamil were marked up in accordance with TEI-Lite guidelines. This amounted to a Sinhala corpus of 65k words and a parallel Tamil corpus of 46k words.

3.2 Language Modeling

Owing to the lack of lemmatizers, taggers etc. for Sinhala and Tamil, all language processing done used raw words and were based on statistical information gleaned from the respective 'half' of the bi-lingual corpus. The CMU-Cambridge Statistical Language Modeling Toolkit (version 2) was used to build n-gram language models using both the Sinhala and Tamil 'halves' of the corpus independently.

Table 1 shows some statistics of the resulting language models with respect to a small test corpus extracted of new articles on the WSWS site. The perplexity figure for Sinhala and Tamil is higher than for English as reported in [6]. However, in both cases larger test sets produced higher percentages of out of vocabulary (unknown) words indicating that the basic corpus size needs enhancing.

For the purpose of building better language models needed for the statistical translation process a

monolingual Sinhala (or Tamil as the case may be) corpus needs to be extracted from the same domain.

Description	Sinhala Corpus	Tamil Corpus
Testset	2667 words	
Perplexity	509.38 (9 bits)	
# 3-grams	349 (13.09%)	
# 2-grams	696 (26.10%)	
# 1-grams	1622 (60.82%)	
# unseen words	426 (13.77%)	

Table 1. Perplexities and other statistics for the Sinhala and Tamil WSWS corpora

3.3 Translation model

As discussed in section 3.1, the Sinhala and Tamil bitext was first sentence aligned using semi-automatic means. The resultant parallel corpus contained 4064 sentence pairs in the two languages.
In order to construct the Sinhala-Tamil translation model we first used the public domain GIZA toolkit and then also experimented with the newer GIZA++ system of Och et. al. [7].

Many different training schemes were experimented with including the use of a HMM step in place of IBM model 2 using GIZA++. Many different smoothing techniques were also applied in order to arrive at an optimal translation model.

3.4 How Using the SMT model

The ISI-Rewrite decoder was used to generate translations for a set of 162 Tamil test sentences taken from new news items appearing in the WSWS site. Rewrite was used to select the best 3 translations for visual evaluation purposes. In order to test the system however, only the topmost translation returned was used.

Table 2 shows an example fragment of this text and its translated Sinhala equivalent. As can be seen, the intelligibility itself is affected by the unseen Tamil words which are simply translated by transliteration. What is however more interesting is the evaluation of

the these translations against the human translation retrieved from the WSWS website.

Source Tamil Text
2001 @p b r v `Ý  22 @k ÄN m  v `m @p r Ë Ða `N v p l v `h #Þ m t v J p `û  i Ø Ý p W í  n #ñ  n Ë Ẉ l < e n  × `N w m #ñ v r n @Y Ù w ^n Ì L h ` @J Ô s A ì [`n v l a W s ì k r g #Ï m s >h ` o Ò Ó́y v r   @g n ñ @B .

Target Sinhala Translation
22 පෙබරවාරි 2001 වාජ්පායිට ආන්ව බටහිර බෙංගාලයේ රිව්ජිපිබ්බූඑ් දෑම්මට බලයේ සිට වඔට්බුඒ්ජුඔිඟුද කරුවන් සිඔබ්තුබුඹුඔ්කුද සෝලිය, ප්‍රාන්ත ඌඩ්විතිබ්ඔිඳු ද් සහා බෙංගාලයේ විය. පලාතේ සහ සිපිඅයි(එම්) පස්දෙනෙක් සිපිඅයි(එම්) ඒඳු ඊතිඳි ශක්තිමත් ක්‍රියාකාරී තන්‍තුයේ ලිඟිල්

Table 2. An example translation of a Tamil test sentence into Sinhala

4. Results & evaluation

In order to carry out an unbiased evaluation of the accuracy of the Sinhala-Tamil translations achieved with respect to the state-of-the-art, we used an implementation of the IBM BLEU score generator.

After testing with multiple translation models, we achieved a best BLEU score of 0.1362 for this task. A straight comparison of this result with the work reported in [6] gives cause for hope. In that work, the best BLEU score achieved for English-Sinhala translations was just 0.0618.

Papineni [5] shows that the BLEU score is sensitive to the number of (expert) reference translations. The most common numbers of reference translations for which scores are quoted in the literature are 2 and 4.

In this work, we had access only to a single translation which was taken to be the reference translation. From [5] it can be gauged that BLEU scores of translations compared with 4 reference translations can be 1.35 times as high as those with 2 reference translations.

Assuming a similar ratio in score differentials between 2 and 1 reference translation(s), the above scores correspond to a Sinhala-Tamil translation BLEU score of 0.185 and a English-Sinhala

translation BLEU score of 0.084 with 2 reference translations.

These scores can be compared with the machine translation scores reported in [5] of 0.0527, 0.0829 and 0.0930 and the human translation scores of 0.1934 and 0.2571 – all on 2 reference translations.

Among the obstacles remaining between the current system and more intelligible output translations include (a) the limited size of the corpora highlighted by the high perplexity of the Sinhala and Tamil language models, and (b) the long-distance 'movement' of mutually translated words and phrases not captured in current translation models.

In order to address (a), current efforts are underway to extract a larger Sinhala-Tamil parallel corpus as well as larger mono-lingual corpora in each of the two languages concerned. In order to address (b), serious consideration would need to be given to the underlying assumptions of the IBM models and other data-driven techniques pursued where appropriate.

Further work is also planned to combine the information obtainable from all three languages Sinhala, Tamil and English in order to arrive at more accurate models between any two of them.

5. Conclusions and further work

It is apparent from the above results that further work is needed in SMT as a whole to produce intelligible translations. One of the limitations in this particular application of the process is the size of the parallel corpus used for learning.

The dimensions of the Sinhala-Tamil corpus used however being very similar to that of the Sinhala-English corpus reported on in [6], some comparison of the two experiments is warranted.

It is clear from the perplexities of both the Sinhala and Tamil corpora used that their language models are deficient. Despite this however, the Sinhala-Tamil SMT process consistently produced BLEU scores significantly higher than those for English-Sinhala translation reported in [6]. In fact, in most cases, the former score was about twice that of the latter. Closer examination of the type of errors generated in the English-Sinhala case, suggest that sentence structure accounts for many of the incorrect translations. On the contrary, in the Sinhala-Tamil case, sentence structures are much more predictable from each other. This may offer a clue as to the reason for the better performance of the Sinhala-Tamil case.

Further, the above results also point to a possible link between the relative success of SMT for linguistically related language pairs such as English and French as reported in the SMT literature. As such the results of this work contribute to the general body of SMT work by suggesting that SMT between linguistically closely related language pairs perform significantly better than that between linguistically less related language pairs.

This research also provides some reasons for optimism of the general SMT approach for solving the translation problem among Sri Lanka's two native languages, Sinhala and Tamil.

References:

[1] 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 (2001)

[2] Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J. and Mercer, R. L.: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19(2) (1993) 263-311.*

[3] Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D.: *Statistical Machine Translation, Final Report*, JHU Workshop 1999. Technical Report, CLSP/JHU (1999)

[4] Clarkson, P.R. and Rosenfield, R.: Statistical Language Modeling using the CMU-Cambridge Toolkit, *Proceedings ESCA Eurospeech*, Rhodes, Greece (1997)

[5] Papineni, K., Roukos, S., Ward, T. and Zhu, W. BLEU – a method for automatic evaluation of machine translation. In *Proceedings of the Association of Computational Linguistics (2002).*

[6] Weerasinghe, A.R. Bootstrapping the lexicon building process for machine translation between 'new' languages. *Proceedings of the Association of Machine Translation in the Americas Conference (AMTA), 2002.*

[7] Och, F.J., Tillmann, C. and Ney, H. Improved alignment models for statistical machine translation. In *Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing (EMNLP),* Maryland, 1999.