Statistical Translation with Scarce Resources: A South African Case Study

Kato Ronald, Etienne Barnard

Human Languague Technology Research Group,
University of Pretoria / Meraka Institute, Pretoria, South Africa
katoronald@tuks.co.za, ebarnard@up.ac.za

Abstract

Statistical machine translation techniques offer great promise for the development of automatic translation systems. However, the realization of this potential requires the availability of significant amounts of parallel bilingual texts. This paper reports on an attempt to reduce the amount of text that is required to obtain an acceptable translation system, through the use of active and semi-supervised learning. Systems were built using resources collected from South African government websites and the results evaluated using a standard automatic evaluation metric (BLEU). We show that significant improvements in translation quality can be achieved with very limited parallel corpora, and that both active learning and semi-supervised learning are useful in this context.

1. Introduction

In statistical machine translation the objective is to translate a sequence of words $f_1^J = f_1...f_j...f_J$ into a target word sequence $e_1^I = e_1...e_i...e_I$ by maximising the probability $P(e_1^I|f_1^J)$. This probability is usually factored into the translation model probability $P(f_1^J|e_1^I)$, which describes the correspondence between the words in the source and target languages and the language model probability $P(e_1^I)$, which describes the well formedness of the target sequence produced. These two probabilities are often modelled independently of one another. A detailed description of this process can be found in [1]. The translation model is trained from a bilingual sentence-aligned text corpus, while the language model is trained from a monolingual text corpus.

For such translation systems, the larger the available training corpus the better the performance and indeed for certain language pairs such resources exist in abundant quantities, for example, the Europarl corpus [2]. However, for new domains or new language pairs acquisition of a large high-quaility bilingual parallel corpus requires significant time and effort. We are therefore studying methods to better exploit existing training data with the prospect of building automatic translation systems for South African language pairs. One approach that has received much attention is to harvest the web for bilingual texts [3]. The STRAND architecture gathers web pages that are potential translations of one another, by looking for documents in one language which have links whose text contains the name of another language. For example, if an English web page had a link with the text "in isiZulu", the page linked to is treated as a candidate translation of the English page. A number of filtering techniques are used to verify this.

Another approach is co-training, a weakly supervised learning technique, where machine translations are used to create parallel corpora [4]. In co-training, implementations of multiple learners are used to label new examples and each implementation is retrained on some of the labeled examples created by other implementations. The intuition behind this approach is that an example which is easily annotated by one learner may be difficult for the others; therefore adding the confidently annotated example will provide information when the models are retrained.

In our work, which is similar in spirit, we investigate an active and semi-supervised approach to exploit the limited training data and bootstrap informative data to develop systems that produce acceptable translations. We study the benefits of incorporating human translations and machine translations to create training data and thus reduce the amount of time the human needs to spend translating sentences.

Section 2 presents an active and semi-supervised approach for machine translation. Section 3 reports on the language resources used. Section 4 gives a desciption of the phrase-based translation framework that we used. Experiments and results are presented in section 5 and a discussion and future directions are contained in section 6.

2. Combining Active and Semi-supervised Learning

Active learning algorithms offer the promise of building better classification systems with less annotated data, by giving the learner some control over the input on which it trains. There are two major approaches to active learning:

- In committee-based methods, a distinct set of classifiers is created using the small set of annotated examples. The unannotated instances whose classification differ most when presented to different classifiers are given to the labeler for annotation.

- In certainity-based methods, which are the focus of our work, an initial system is trained using a small set of annotated examples. The system then examines and labels the unannotated examples and determines the "confidence" or "certainty" of each of its predictions. The examples with the lowest certainity levels are presented to the labelers for manual annotation.

In semi-supervised learning algorithms, an initial classifier is created and used to label the unannotated examples. The most confident predictions are then selected from a pool of unannotated bilingual sentences, using a threshold. These samples and their annotations are added to the original data set, and the classifier is then re-trained. In such an approach the algorithm tries to fit both the machine-labeled data and the prior model.

The idea of combining active and semi-supervised learning has been used in text categorisation and automatic speech recognition for statistical language modelling [5][6]. Inspired by certainty-based methods and the prospect of incorporating unlabelled data in the learning process using a semi-supervised algorithm, we outline an approach that may be used to build machine translation systems for new language pairs. In this approach, an initial translation model is built using the available bilingual sentence-aligned corpus and a language model from the monolingual data. The system is then used to translate the monolingual text, and two new data sets are built using a threshold th for separation. One set contains the most confident translations while the other contains the least confident ones. The human is queried for the correct translations for the least confident case. A new translation model is built and used to translate the most confident set. This helps to reduce the amount of noise introduced in the system because of machine translation errors. The threshold is determined by the capacity of the human to produce new translations or by the performance of the current translation model. This approach is outlined in figure 1.

3. Language Resources

The parallel multilingual corpus used in our experiments was obtained from South African government web pages which are not copyrighted. The documents were preprocessed to remove markup tags, images and words that are repeated from one page to another but are not translated. Prior to using these resources in a machine translation system, it is important that they are aligned at the sentence level. Sentence alignment was done using an algorithm that uses word association and sentence length [7].

```
• Inputs: Bilingual corpus $L||EN$, Monolingual corpus $L_m$

• create initial translation model $L => EN$ (from parallel corpus $L||EN$) and language model from monolingual corpus

• Loop while translation improves as measured by word error rate or BLEU score.

    – For each sentence in $L_m$ create candidate pool of translations $en_1.....en_t$ by translating $l_1........l_t$ into English using $L => EN$
    – use certainty-based approach to determine the most informative translations $en_1....en_k$ and confident translations $en_j....en_t$ using a threshold $th$
    – Loop k times to obtain translation of $en_i....en_k$ from human (or oracle).
    – Train new translation model $L' => EN'$ by incorporating the human labels.
    – translate $l_j...l_t$ to obtain less noisy $en_j.....en_t$ using $L' => EN'$
    – Train new translation model $L'' => EN''$ by incorporating $en_j...en_t$

• output: $L'' => EN''$
```

Figure 1: Algorithm for combining active and semi-supervised learning for machine translation

Table 2 gives an overview of the corpus statistics. The number of words is from the English side of the corpus.

Table 1: *Multilingual corpus statistics.*

corpus	sentences	words
English-isiXhosa	3074	31213
English-isiZulu	3826	39881
English-Setswana	3408	36169
English-Afrikaans	2328	30815

In order to investigate the scenario with scarce training data, a small training corpus of 1 000 sentences was used. This was then augmented using the rest of the data set as "unlabelled" data, applying the algorithm in section 2.

4. Statistical Phrase-based Translation

We evaluated the performance of our approach using the phrased-based translation framework. In this framework the task is to find that English sentence **e** given a foreign sentence **f** for which $p(\mathbf{e}|\mathbf{f})$, the probability of an English sentence given a foreign language sentence is maximum. Rather than trying to estimate $p(\mathbf{e}|\mathbf{f})$ directly, we appeal to Bayes's rule to reformulate the translation probability for translating a foreign sentence **f** into English **e** as

$$\mathbf{e_{best}} = argmax_e p(\mathbf{e}|\mathbf{f}) = argmax_e p(\mathbf{f}|\mathbf{e}) p(\mathbf{e}) \quad (1)$$

This allows for a language model p(e) and a separate translation model $p(\mathbf{f}|\mathbf{e})$. During decoding the foreign input sentence **f** is segmented into a sequence of I phrases f_1^I. A uniform probability distribution over all possible segmentations is assumed. Each foreign phrase f_i in f_1^I is translated into an English phrase e_i. The English phrase may be reordered using a distortion probability distribution $d(a_i - b_{i-1})$, where a_i denotes the start position of the foreign phrase that was translated into the ith English phrase and b_{i-1} denotes the end position of the foreign phrase that was translated into the $(i-1)$th English phrase [8].

5. Experiments and Results

We evaluated our approach using the dataset summarized in table 1. We used a separate, consistent test set of 400 sentences to evaluate the translation quality of the system. We simulated the actions of the human in a scenario where we use a part of the remainder of the dataset to augement the training set, by simply extracting the available translations of those sentences selected for active learning. The number of sentences used for active and semi-supervised learning are shown in table 3.

5.1. Software

We used freely-available software for all our experiments. In particular, the main training and decoding tasks were performed as follows:

- For training the translation models: GIZA++, an open source implementation of the IBM word alignment models;
- For training to produce translation models at the phrase level: Pharaoh, a state-of-the-art phrased based decoder [9].
- To train language models, the SRI Language modelling toolkit was used [10].
- Finally, Pharaoh was again employed to find the best translation.

5.2. Evaluation Measures

In recent years, a number of methods have been proposed to automatically evaluate machine-translation quality by comparing hypothesized translations with reference translations. These evaluation measures try to approximate human assessment and often achieve a great degree of correlation with human subjective evaluation of fluency and adequacy. One such metric is BLEU, which we used to evaluate machine translation quality [11]. This criterion computes the geometric mean of the precision of n-grams of various lengths between a hypothesis and a set of reference translations. A BLEU score of 0.0 corresponds to a system output that is very different from all of the reference translations, whereas a score of 1.0 corresponds to system output that is identical to some patchwork combination of the reference translations. In this paper we used a single reference translation, since only one translation per language was available to us.

5.3. Results

Table 2 summarizes the main results of our experiments. As expected, we are able to improve the accuracy of our systems in all languages when we apply an active and semi-supervised learning approach. It can be seen that the best performing pair has an increase of 0.12 in the BLEU score. This is a significant improvement; although the resulting BLEU scores are still too low for practical applications, they suggest that such a level can be attained with a manageable amount of data.

These improvements can be attributed to the ability of the algorithm to deal with some core problems when translating in a resource constrained environment. These include:

- coping with morphology. The statistical framework used treats words with the same root morpheme as completely different. A system that knows how to translate *car* will not translate *cars* unless it also sees it during training.
- increased word coverage. This approach increases the number of correctly translated words that the system had not seen during initial training.
- improvement in grammar. The system learns better translation patterns for words since it encounters these words in different contexts with subsequent training.

To further assess the value of active learning, and to distinguish the contributions of the active and semi-supervised components, we have performed two additional experiments. In the first we compared active learning and random selection of sentences for translation by using the English-Setswana parallel corpus. We started with an initial baseline model trained from 1000 sentence pairs, and incrementally trained the model using sentences sampled from the rest of the corpus. The results are compared with those obtained using active learning in figure 2. This shows that significant improvements can be achieved with less training data by using the active learning approach as opposed to random selection - active

Table 2: *Results of a combined active and semi-supervised learning algorithm for various language pairs.*

Translation pair	baseline	active-semi-supervised
English-isiXhosa	0.1795	0.2329
English-isiZulu	0.2004	0.2931
English-Setswana	0.2074	0.3271
English-Afrikaans	0.1193	0.1999

Table 3: *Number of sentences used for active and semi-supervised learning parts of the algorithm.*

Translation pair	active	semisupervised
English-isiXhosa	1279	241
English-isiZulu	1463	250
English-Setswana	1648	360
English-Afrikaans	731	197

learning reduces the amount of training data required to achieve a given level of performance by about a third.

In the second experiment, we evaluated the benefits of semi-supervised learning alone. We started from the baseline model and incrementaly trained the model using machine translations of sentences sampled from the rest of the corpus. The results are shown in figure 3. Since semi-supervised learning adds no hand-labelled data to the training set, it is not surprising that it should be less accurate than active learning. It is nevertheless encouraging to see that semi-supervised learning does in fact produce significant improvements by itself.

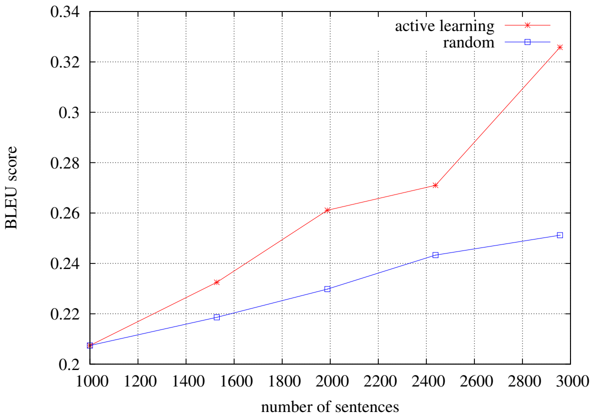

Figure 2: *BLEU scores obtained using certainity based active learning and random selection of sentences for translation.*

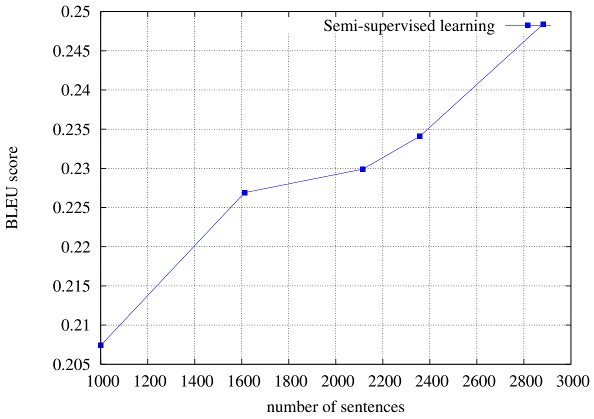

Figure 3: *BLEU scores obtained using semi-supervised learning alone.*

6. Discussion and Future Directions

We have presented an approach that combines active and semi-supervised learning for machine translation. The aim is to reduce the number of bilingual parallel sentence pairs by selectively sampling a subset of the untranslated sentences and exploiting the unselected ones as well. We have shown that active learning can yield significant improvements when compared to random selection of sentences. In our investigation, semi-supervised learning showed comparable performance improvements to random selection of manually transcribed sentences. This can clearly not continue to hold as the number of training sentences increase, since the manual transcriptions keep adding novel translated words and phrases to the system, which are not available in semi-supervised learning. It is therefore best to view the additional improvements obtained with semi-supervised learning as an added bonus - with active learning as the mainstay of improved translation when resources are scarce.

In conclusion, we have considered two main questions. *Which sentences should we give to a human to translate? and what do we do with the remaining untranslated sentences?* Although we must caveat our conclusions with the fact that they are based on a small data set consisting of only 400 test sentences, there are strong indications that we can use this method to build more robust systems for South African language pairs that have scarce data.

Our future research includes investigating approaches that incorporate more knowledge into the translation process to obtain more fluent translations. We plan to study syntax-based approaches to machine translation since these could produce better linguistically correct output. Such a system could learn better the correct syntax and morphology in the target language for example noun-phrase agreement and subject-verb agreement. Further more we plan to investigate the benefits of merging phrase-based and syntax-based systems to build a hybrid machine translation system. One would hope that such a system would increase the fluency and adequacy of machine translations in a resource constrained environment.

7. References

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

[2] Koehn, P, Europarl: A parallel corpus for statistical machine translation. School of informatics, University of Edinburgh, Scotland. http://www.statmt.org/europarl/ (13th october 2006)

[3] Resnik, P., Smith, N.A, The Web as a parallel corpus. Computational Linguistics 2003, 29(3):349-380,

[4] Callison-Burch, C., Osborne M., Bootstrapping parallel corpora. In: Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Associatioin of Computational Linguistics (HLT-NAACL) 2002.

[5] Ghani, R., Combining labelled and unlabelled data for multiclass text categorization. In: Proc. Internal. Conf. on Machine Learning (ICML), 2002, Sydney, Australia

[6] Gokhan, T., Hakkani-Tur, D., Schapire., R.E. Combining active and sem-supervised learning for spoken language understanding. Speech Communication 45 (2005) 171-186.

[7] Moore, R., Fast and accurate sentence alignment of a Bilingual corpus. Machine Translation: From Research to Real User. in Proc. 5th conference of Machine Translation in the Americas, 2002.

[8] Koehn, P., Och, F.J., Marcu, D., Statistical Phrased-Based Translaton, In: Proceedings of HLT-NAACL, Main Papers, 2003, pp. 48-54. Edmonton.

[9] Koehn, P., Pharaoh: A beam search decoder for phrased-based statistical machine translation, University of Southern California, Information Sciences Institute,2003. http://www.isi.edu/publications/licensed-sw/pharaoh/ (13th october 2006)

[10] A.Stolcke. SRILM- an extensible language modelling toolkit. In Proc. international Conf. on Spoken Language Processing,2002 , Denver, Colorado.

[11] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-j (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report.