Transliteration using a Phrase-based Statistical Machine Translation System to Re-score the Output of a Joint Multigram Model

Andrew Finch
NICT
3-5 Hikaridai
Keihanna Science City
619-0289 JAPAN
andrew.finch@nict.go.jp

Eiichiro Sumita
NICT
3-5 Hikaridai
Keihanna Science City
619-0289 JAPAN
eiichiro.sumita@nict.go.jp

Abstract

The system presented in this paper uses a combination of two techniques to directly transliterate from grapheme to grapheme. The technique makes no language specific assumptions, uses no dictionaries or explicit phonetic information; the process transforms sequences of tokens in the source language directly into to sequences of tokens in the target. All the language pairs in our experiments were transliterated by applying this technique in a single unified manner. The approach we take is that of hypothesis re-scoring to integrate the models of two state-of-the-art techniques: phrase-based statistical machine translation (SMT), and a joint multigram model. The joint multigram model was used to generate an *n*-best list of transliteration hypotheses that were re-scored using the models of the phrase-based SMT system. The both of the models' scores for each hypothesis were linearly interpolated to produce a final hypothesis score that was used to re-rank the hypotheses. In our experiments on development data, the combined system was able to outperform both of its component systems substantially.

1 Introduction

In statistical machine translation the re-scoring of hypotheses produced by a system with additional models that incorporate information not available to the original system has been shown to be an effective technique to improve system performance (Paul et al., 2006). Our approach uses a re-scoring technique to integrate the models of two transliteration systems that are each capable in their own right: a phrase-based statistical machine translation system (Koehn et al., 2003), and a joint multigram model (Deligne and Bimbot, 1995; Bisani and Ney, 2008).

In this work we treat the process of transliteration as a process of direct transduction from sequences of tokens in the source language to sequences of tokens in the target language with no modeling of the phonetics of either source or target language (Knight and Graehl, 1997). Taking this approach allows for a very general transliteration system to be built that does not require any language specific knowledge to be incorporated into the system (for some language pairs this may not be the best strategy since linguistic information can be used to overcome issues of data sparseness on smaller datasets).

2 Component Systems

For this shared task we chose to combine two systems through a process of re-scoring. The systems were selected because of their expected strong level of performance (SMT systems have been used successfully in the field, and joint multigram models have performed well both in grapheme to phoneme conversion and Arabic-English transliteration). Secondly, the joint multigram model relies on key features not present in the SMT system, that is the history of bilingual phrase pairs used to derive the target. For this reason we felt the systems would complement each other well. We now briefly describe the component systems.

2.1 Joint Multigram Model

The joint multigram approach proposed by (Deligne and Bimbot, 1995) has arisen as an extension of the use of variable-length *n*-grams (multigrams) in language modeling. In a joint multigram, the units in the model consist of multiple input and output symbols. (Bisani and Ney, 2008) refined the approach and applied to it grapheme-to-phoneme conversion, where its performance was shown to be comparable to state-of-the-art systems. The approach was later applied to Arabic-English transliteration (Deselaers et al., 2009) again with promising results.

Joint multigram models have the following characteristics:

- The symbols in the source and target are co-segmented

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 48–52,
Uppsala, Sweden, 16 July 2010.

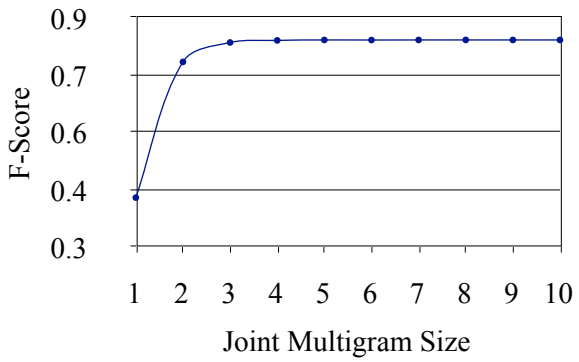

Figure 1: The effect on F-score by tuning with respect to joint multigram size

  - Maximum likelihood training using an EM algorithm (Deligne and Bimbot, 1995)

- The probability of sequences of joint multigrams is modeled using an *n*-gram model

In these respects the model can be viewed as a close relative of the joint source channel model proposed by (Li et al., 2004) for transliteration.

2.2 Phrase-based SMT

It is possible to view the process of transliteration as a process of translation at the character level, without re-ordering. From this perspective it is possible to directly employ a phrase-based SMT system in the task of transliteration (Finch and Sumita, 2008; Rama and Gali, 2009). A phrase-based SMT system has the following characteristics:

- The symbols in the source and target are aligned one to many in both directions. Joint sequences of source and target symbols are heuristically extracted given these alignments
- Transliteration is performed using a log-linear model with weights tuned on development data
- The models include: a translation model (with 5 sub-models), and a target language model

The bilingual phrase-pairs are analogous to the joint multigrams, however the translation model of the SMT system doesn't use the context of previously translated phrase-pairs, instead relying on a target language model.

3 Experimental Conditions

3.1 SMT Decoder

In our experiments we used an in-house phrase-based statistical machine translation decoder called CleopATRa. This decoder operates on exactly the same principles as the publicly available MOSES decoder (Koehn et al., 2003). Our decoder was modified to be able to decode source sequences with reference to a target sequence; the decoding process being forced to generate the target. The decoder was also configured to combine scores of multiple derivations yielding the same target sequence. In this way the models in the decoder were used to derive scores used to re-score the *n*-best (we used *n*=20 for our experiments) hypotheses generated by the joint multigram model. The phrase-extraction process was symmetrized with respect to token order using the technique proposed in (Finch and Sumita, 2010). In order to adapt the SMT system to the task of transliteration, the decoder was constrained decode in a monotone manner, and furthermore during training, the phrase extraction process was constrained such that only phrases with monotone order were extracted in order to minimize the effects of errors in the word alignment process.

In a final step the scores from both systems were linearly interpolated to produce a single integrated hypothesis score. The hypotheses were then re-ranked according to this integrated score for the final submission.

3.2 Joint Multigram model

For the joint multigram system we used the publicly available Sequitur G2P grapheme-to-phoneme converter (Bisani and Ney, 2008). The system was used with its default settings, and pilot experiments were run on development data to determine appropriate settings for the maximum size of the multigrams. The results for the English-to-Japanese task are shown in Figure 1. As can be seen in the figure, the system rapidly improves to a near-optimal value with a maximum multigram size of 4. No improvement at all was observed for sizes over 7. We therefore chose a maximum multigram size of 8 for the experiments presented in this paper, and for the systems entered in the shared task.

3.3 Pre-processing

In order to reduce data sparseness we took the decision to work with data in only its lowercase form.

We chose not to perform any tokenization or phonetic mapping for any of the language pairs

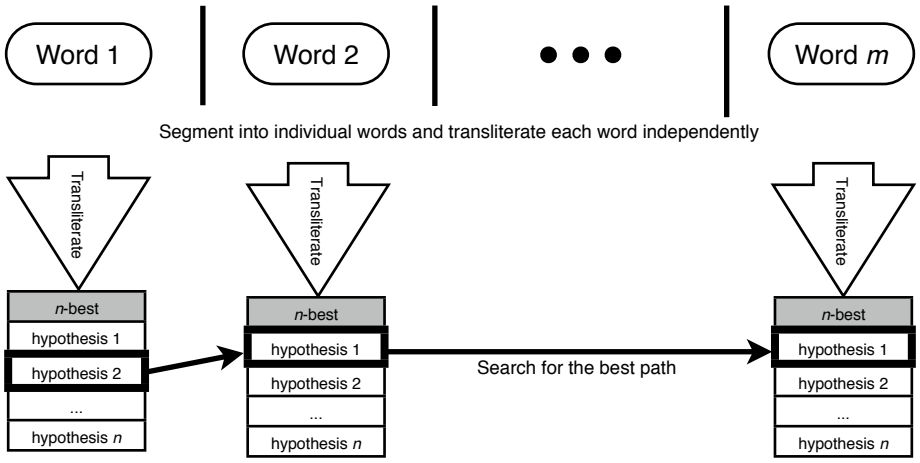

Figure 2: The transliteration process for multi-word sequences

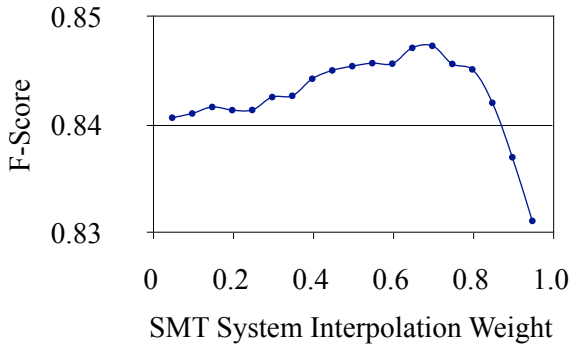

Figure 3: The effect on the F-score of the integrated system by tuning with respect to the SMT system's interpolation weight

in the shared task. We adopted this approach because:

- It allowed us to have a single unified approach for all language pairs
- It was in the spirit of the shared task, as it did not require extra knowledge outside of the supplied corpora

3.4 Handling Multi-Word Sequences

The data for some languages contained some multi-word sequences. To handle these we had to consider the following strategies:

- Introduce a <space> token into the sequence, and treat it as one long character sequence to transliterate; or
- Segment the word sequences into individual words and transliterate these independently, combining the *n*-best hypothesis lists for all the individual words in the sequence into a single output sequence.

We adopted both approaches: for those multi-word sequences where the number of words in the source and target matched, the latter approach was taken; for those where the numbers of source and target words differed, the former approach was taken. The decoding process for multi-word sequences is shown in Figure 2. During recombination, the score for the target word sequence was calculated as the product of the scores of each hypothesis for each word. Therefore a search over all combinations of hypotheses is required. In almost all cases we were able to perform a full search. For the rare long word sequences in the data, a beam search strategy was adopted.

3.5 Building the Models

For the final submissions, all systems were trained on the union of the training data and development data. It was felt that the training set was sufficiently small that the inclusion of the development data into the training set would yield a reasonable boost in performance by increasing the coverage of the systems. All tunable parameters were tuned on development data using systems built using only the training data. Under the assumption that these parameters would perform well in the systems trained on the combined development/training corpora, these tuned parameters were transferred directly to the systems trained on all available data.

3.6 Parameter Tuning

The SMT systems were tuned using the minimum error rate training procedure introduced in (Och, 2003). For convenience, we used BLEU as a proxy for the various metrics used in the shared task evaluation. The BLEU score is commonly used to evaluate the performance of

Language Pair	Accuracy in top-1	Mean F-score	MRR	MAP_{ref}
English → Thai	0.412	0.883	0.550	0.412
Thai → English	0.397	0.873	0.525	0.397
English → Hindi	0.445	0.884	0.574	0.445
English → Tamil	0.390	0.887	0.522	0.390
English → Kannada	0.371	0.871	0.506	0.371
English → Japanese	0.378	0.783	0.510	0.378
Arabic → English	0.403	0.891	0.512	0.327
English → Bangla	0.412	0.883	0.550	0.412

Table 1: The results of our system in the official evaluation on the test data on all performance metrics.

machine translation systems and is a function of the geometric mean of *n*-gram precision. The use of BLEU score as a proxy has been shown to be a reasonable strategy for the metrics used in these experiments (Finch and Sumita, 2009). Nonetheless, it is reasonable to assume that one would be able to improve the performance in a particular evaluation metric by doing minimum error rate training specifically for that metric. The interpolation weight was tuned by a grid search to find the value that gave the maximal f-score (according to the official f-score evaluation metric for the shared task) on the development data, the process for English-Japanese is shown in Figure 3.

4 Results

The results of our experiments are shown in Table 1. These results are the official shared task evaluation results on the test data, and the scores for all of the evaluation metrics are shown in the table. The reader is referred to the workshop white paper (Li et al., 2010) for details of the evaluation metrics. The system achieved a high level of performance on most of the language pairs. Comparing the individual systems to each other, and to the integrated system, the joint multigram system outperformed the phrase-based SMT system. In experiments run on the English-to-Japanese katakana task, the joint multigram system in isolation achieved an F-score of 0.837 on development data, whereas the SMT system in isolation achieved an F-score of 0.824. When integrated the models of the systems complemented each other well, and on the same English-Japanese task the integrated system achieved an F-score of 0.843.

We feel that for some language pairs, most notably Arabic-English where a large difference existed between our system and the top-ranked system, there is much room for improvement. One of the strengths in terms of the utility of our approach is that it is free from dependence on the linguistic characteristics of the languages being processed. This property makes it generally applicable, but due to the limited amounts of data available for the shared task, we believe that in order to progress, a language-dependent approach will be required.

5 Conclusion

We applied a system that integrated two state-of-the-art systems through a process of re-scoring, to the NEWS 2010 Workshop shared task on transliteration generation. Our systems gave a strong performance on the shared task test set, and our experiments show the integrated system was able to outperform both of its component systems. In future work we would like to depart from the direct grapheme-to-grapheme approach taken here and address the problem of how best to represent the source and target sequences by either analyzing their symbols further, or agglomerating them. We would also like to investigate the use of co-segmentation schemes that do not rely on maximum likelihood training to overcome the issues inherent in this technique.

Acknowledgements

The results presented in this paper draw on the following data sets. For English-Japanese and Arabic-English, the reader is referred to the CJK website: http://www.cjk.org. For English-Hindi, English-Tamil, and English-Kannada, and English-Bangla the data sets originated from the work of Kumaran and Kellner, 2007.

References

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer, 1991. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 263-311.

Sabine Deligne, and Frédéric Bimbot, 1995. Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams. *In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, Detroit, MI, USA, pp. 169–172.

Maximilian Bisani and Hermann Ney, 2008. Joint-Sequence Models for Grapheme-to-Phoneme Conversion. Speech Communication, Volume 50, Issue 5, Pages 434-451.

Thomas Deselaers, Sasa Hasan, Oliver Bender, and Hermann Ney, 2009. A Deep Learning Approach to Machine Transliteration. *In Proceedings of the EACL 2009 Workshop on Statistical Machine Translation (WMT09)*, Athens, Greece.

Andrew Finch and Eiichiro Sumita, 2008. Phrase-based machine transliteration. *In Proceedings of WTCAST'08*, pages 13-18.

Andrew Finch and Eiichiro Sumita, 2009. Transliteration by Bidirectional Statistical Machine Translation, *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore.

Andrew Finch and Eiichiro Sumita, 2010. Exploiting Directional Asymmetry in Phrase-table Generation for Statistical Machine Translation, *In Proceedings of NLP2010*, Tokyo, Japan.

Kevin Knight and Jonathan Graehl, 1997. Machine Transliteration. *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 128-135, Somerset, New Jersey.

Philipp Koehn, Franz Josef Och, and Daniel Marcu, 2003. Statistical Phrase-Based Translation. *In Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003)*, Edmonton, Canada.

Franz Josef Och, 2003. Minimum error rate training for statistical machine translation, *Proceedings of the ACL*.

A Kumaran and Tobias Kellner, 2007. A generic framework for machine transliteration, *Proc. of the 30th SIGIR*.

Haizhou Li, Min Zhang, Jian Su, 2004. A joint source channel model for machine transliteration, *Proc. of the 42nd ACL*.

Haizhou Li, A Kumaran, Min Zhang and Vladimir Pervouchine, 2010. Whitepaper of NEWS 2010 Shared Task on Transliteration Generation. In *Proc. of ACL2010 Named Entities Workshop*.

Michael Paul, Eiichiro Sumita and Seiichi Yamamoto, 2006, Multiple Translation-Engine-based Hypotheses and Edit-Distance-based Rescoring for a Greedy Decoder for Statistical Machine Translation, *Information and Media Technologies*, Vol. 1, No. 1, pp.446-460 .

Taraka Rama and Karthik Gali, 2009. Modeling machine transliteration as a phrase based statistical machine translation problem, *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore.