The Application of Machine Transliteration Techniques to Spelling Correction

Keiko Taguchi
Doshisha University

Andrew Finch
NICT

Seiichi Yamamoto
Doshisha University

Eiichiro Sumita
NICT

{ dun0153@mail4.doshisha.ac.jp, andrew.finch@nict.go.jp, seyamamo@mail.doshisha.ac.jp, eiichiro.sumita@nict.go.jp }

Abstract

This paper extends existing work on spelling correction using statistical machine translation by incorporating techniques that have proved valuable in the related field of machine transliteration. We investigate training the models using a non-parametric Bayesian aligner, alternative translation model features, a language model trained to bias the decoding process towards producing words from a dictionary, and the integration of a joint source-channel model into the set of log-linear models. Our experiments show that all of the enhancements we propose can match or improve the accuracy over a respectable baseline phrase-based statistical machine translation system. Furthermore, the Bayesian aligner gave rise to considerably more compact models and the proposed language model results in a more efficient decoding process by eliminating partial hypotheses that cannot lead to useful results from the search graph.

1 Introduction

The task of spelling correction is a sequence-to-sequence transduction process. The process is mostly monotonic, with typically little re-ordering apart from the occasional transposition of pairs of characters. Since this re-ordering usually happens on a local scale, it is an application well-suited to the tools and apparatus of statistical machine translation (SMT). Previous work has studied the direct application of phrase-based statistical machine translation (PBSMT) techniques to spelling correction [7], and the results reported were competitive to several state-of-the-art baselines. In this paper we revisit this work, with an eye to leveraging some of the adaptations that have been made to the PBSMT framework in order to handle the somewhat similar monotonic sequence transduction task of transliteration.

2 Related Work

Traditionally, spelling correction algorithms have relied on edit distance [14]. However in recent years, statistical methods based on alignment have allowed for more complex edits involving multiple characters. In [1], a noisy channel model was proposed which forms the basis of much of the related work on the field, including work in this paper. Our research also builds upon the work presented in [7] that directly employs the apparatus of PBSMT to the task of spelling correction. Both of these methods transduce between sequences using many-to-many block edit operations. The method of [1] in effect uses a single integration of the EM algorithm to calculate the block edit probability. However, the EM algorithm is well-known to cause overfitting in many-to-many alignment, unless steps are taken to curb this effect. In [7] the block edits (known as phrase pairs) are derived heuristically from two one-to-many alignments. This approach has proven effective in many applications, but has a tendency to generate large set of possible block edits. This is due to the fact that the block edit extraction heuristic extract all possible pairs that are consistent with the alignments.

Our approach relies on a Bayesian method that learns a stochastic block edit distance [10] between erroneous and correct spellings. A many-to-many alignment is made between source and target sequences, and our models are built directly from this alignment. The technique offers many advantages over competitive approaches in that it has a tendency to align without overfitting, is symmetrical and gives rise to models with few parameters. A similar approach in the field of transliteration generation was reported in [2] and their method was shown to outperform models derived from a PBSMT approach that used GIZA++ for alignment [12] and the grow-diag-final-and heuristic for phrase-pair extraction. We omit the details of the Bayesian alignment process for brevity here; the reader is referred to [4, 5] for a detailed description of the technique.

3 Contributions

This paper makes the following contributions:

1. We propose a stochastic block edit distance-based model of spelling correction;

2. We propose an alternative set of features to be used in the PBSMT translation model, commensurate with the Bayesian alignment method used;
3. We propose the use of a simple maximum likelihood n-gram language model for spelling correction;
4. We investigated the use of a joint source-channel model in spelling correction.

3.1 Stochastic block edit distance-based model

As mentioned earlier, we used a Bayesian non-parametric aligner to perform a many-to-many alignment between source (miss-spellings) and target (correct) sequences. Currently although the aligner can learn (block) null alignments on both sides, we do not use this feature, and the study of this extension remains future research. At the end of the alignment process, the corpus will be force-aligned; that is, each block of source characters will be aligned to a block of target characters. The alignment process is perfectly symmetrical. Given this alignment of corpus we extract a phrase-table for use in the PBSMT decoder. The phrase-table consists of the set of many-to-many alignments induced by the aligner, and we describe two methods for obtaining features for the translation model for them in the next section.

3.2 Translation Features

Our first approach was to mimic the features typically used in a PBSMT phrase table, i.e.: $p(s|t)$, $p(t|s)$ where s and t are source and target phrases (character sequences in our case), and two lexical weighting functions. In our spelling correction model we chose not to use the lexical weighting functions since the character-level models we are building are less sparse than word-level models. The conditional probabilities were calculated using maximum likelihood estimation from the aligned corpus.

Since the alignment model we use is symmetric and is a generative model that generates phrase-pair by phrase-pair using the joint probability of s and t, we created a set of two translation features using probabilities from the generative model: the first feature was the probability of the phrase-pair given by the Dirchlet process model, and the second was the value of the base measure.

3.3 Maximum Likelihood N-gram Model (MLLM)

In the original work applying PBSMT to spelling correction [7], a back-off target language model is employed. Their approach then proceeds to generate a large set of spelling correction hypotheses, many of which are non-words. In order to produce the final list of correction candidates, a large n-best list (in their experiments an n-best list size of 500 was used) was filtered using a dictionary of words to remove the non-words. One potential issue with this approach is that it is possible for the search space to become overwhelmed by hypotheses that will not lead to a useful output.

We therefore propose to overcome this issue by incorporating the filtering into the decoding process itself by changing the nature of the language model used. A back-off language model reserves a certain proportion of its probability mass for n-grams that were not observed in the training process. In our method we train a model that uses maximum likelihood to train the n-gram probabilities of only those n-grams that occur. This language model is trained on a dictionary of words, and will (intentionally) overfit the dictionary. During the decoding process should any n-gram be hypothesized that did not occur in the training dictionary, the hypothesis containing it will receive a zero (or in our implementation, a near-zero) probability, and its search state will be unlikely to be advanced. This approach does not guarantee the output of the decoder consists only of words, but ensures that any words that can be generated will be generated in preference to pursuing hypothesis that will lead to non-words in the output. The benefits of this approach are potentially a more efficient decoding process, and a better search leading to a greater proportion of, and number of words in the output.

3.4 Joint Source-Channel Model

In machine transliteration joint source-channel models [9] have proven to be highly effective. Particularly relevant to this work is the approach used in [2, 3] in which joint source-channel models were introduced into a phrase-based machine transliteration system similar in architecture to the systems in this paper. The model proved to be a key component in their system which achieved state-of-the-art performance in the NEWS shared evaluations. We therefore introduced this type of model, built directly from the Bayesian alignment using standard language modeling tools (in our case the SRILM toolkit [13]). Surprisingly, we did not see any improvement in performance when this model was added as a feature into the log-linear model of the PBSMT system. We report this negative result here, as we expected this model to give rise to a substantial improvement. We believe that reason why the model proved ineffective may be due to both the size and the nature of spelling correction data. The data set size is small (as it typically is in transliteration) but also most of the edit operations are simple identity substitutions, with the edits representing the corrected parts of the words making up only a tiny part of the corpus. We believe that it is likely that joint source-channel models can give rise to noticeable improvements with larger data set sizes.

System	1-best	5-best	10-best
Baseline	45.8 (2.9)	56.2 (3.4)	59.6 (3.4)
Baseline MLLM	42.3 (2.6)	61.0 (2.9)	64.0 (2.7)
Bayesian	45.4 (2.2)	55.4 (3.0)	59.2 (3.2)
Bayesian GIZA MLLM	42.2 (2.7)	60.9 (2.5)	64.5 (2.4)
Bayesian MLLM	42.1 (2.9)	61.2 (2.8)	64.8 (2.7)

Table 1: N-best Spelling Correction Accuracy (standard error in parentheses). The n-best accuracies are averages over 10-folds.

4 Experiments

4.1 Data

For this paper, a corpus consisting of English spelling errors made by Japanese language speakers was used. We used the Atsuo-Henry corpus which was used in the work of [7]. The corpus consists of 4,874 spelling errors paired with their corrections.

Due to the small quantity of available data, we chose to run 10 experiments by 10-fold jackknifing of the corpus. The data was split such that none of the corrected words in the test set appeared in the training and development sets. This was to avoid any bias towards generating these target words in the models. The data was divided into training, development and test sets in approximately the following proportions: train 80% (4000 pairs), development 10% (450 pairs) and test 10 % (450 pairs). The exact numbers in these splits depended on the fold.

4.2 Training

The baseline system was implemented using the MOSES decoder and accompanying experimental framework [8]. The joint source-channel model experiments were conducted using the OCTAVIAN decoder, a similar phrase-based decoder to MOSES. The models were tuned using the MERT procedure [11]. We tuned to the BLEU score in our experiments, rather than the final evaluation metric, but we believe although not optimal, this is a reasonable proxy for the final metric as was reported in transliteration generation [2]. We trained two types of language model. In the baseline system, we trained a 5-gram language model using Witten-Bell smoothing on a corpus of words from the CMU pronunciation dictionary (as in the experiments in [7]) using the SRILM toolkit. These words were weighted by frequency (the frequency counts came from the English Gigaword corpus [6]), as this type of weighting was shown to have a large impact on correction performance [7]. In the MLLM system, we trained a weighted language model on the same data as the baseline language model using an in-house script to calculate the language model from weighted n-gram counts output by the ngram-count tool in the SRI language modeling toolkit. A 34-gram language model was trained in order to ensure that all words in the dictionary could be covered within the span of a single n-gram.

4.3 Results

4.3.1 Bayesian Alignment

Table 1 shows the results when using a non-parametric Bayesian aligner to produce the translation model for the PBSMT system. In all of the experimental conditions the system trained using the Bayesian aligner give rise to similar n-best accuracy.

On average the phrase table derived from the Bayesian aligner was only 15% of the size (in terms of the number of phrase-pairs) of that produced using GIZA++ and the grow-diag-final-and heuristic.

4.3.2 Translation Features

The last two rows of Table 1 shows the results when using the proposed set of joint-probability translation features in the translation model for the PBSMT system, relative to a baseline system that uses a more typical pair of conditional probabilities. Both of the systems used models trained from the non-parametric Bayesian aligner. In 2 out of 3 of the experimental conditions the system that used the proposed translation features give rise to slightly higher n-best accuracy.

4.3.3 Maximum Likelihood Language Model

Table 1 shows the results when using the MLLM relative to various baselines that used a 5-gram back-off language model. In all of the experimental conditions except for 1-best accuracy, the MLLM system give rise to substantially higher n-best accuracy.

5 Conclusion

This paper has proposed and investigated a number of improvements inspired by work in the field of machine transliteration, to a spelling correction system built using a phrase-based statistical machine translation framework. Our experiments show that using a non-parametric Bayesian aligner to build the models gives rise to a similar level of performance to the GIZA++ method in terms of

n-best accuracy, whilst producing a translation model that is considerably smaller in size. Furthermore, this alignment technique opens the door for the use of a joint source-channel model and also null alignments. Although our experiments currently failed to show any benefit from using a joint source-channel model, we believe it may become an important factor when larger data sets are used. We also proposed a simple set of translation features based on the joint probability of source and target character sequences which gave rise to a small increase in performance relative to similar features based on conditional probabilities. Finally, we have shown that using a language model trained to assign probability mass only to n-grams which occur in a dictionary of words can bias the decoding process towards the production of words in the dictionary, leading to both more efficient decoding and more accurate spelling correction.

References

[1] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. pages 286--293, 2000.

[2] A. Finch, P. Dixon, and E. Sumita. Integrating models derived from non-parametric bayesian co-segmentation into a statistical machine transliteration system. In Proceedings of the Named Entities Workshop, pages 23--27, Chiang Mai, Thailand, Nov 2011. Asian Federation of Natural Language Processing.

[3] A. Finch, P. Dixon, and E. Sumita. Rescoring a phrase-based machine transliteration system with recurrent neural network language models. In Proceedings of the 4th Named Entity Workshop (NEWS) 2012, pages 47--51, Jeju, Korea, July 2012. Association for Computational Linguistics.

[4] A. Finch and E. Sumita. A Bayesian Model of Bilingual Segmentation for Transliteration. In M. Federico, I. Lane, M. Paul, and F. Yvon, editors, Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT), pages 259--266, 2010.

[5] T. Fukunishi, A. M. Finch, E. Sumita, and S. Yamamoto. A bayesian alignment approach to transliteration mining. ACM Transactions on Asian Language Information Processing (TALIP), 12(3), 2013.

[6] D. Graff. English Gigaword. 2003.

[7] D. Hovermale. Erron: a phrase-based machine translation approach to customized spelling correction. PhD thesis, The Ohio State University, 2011.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowa, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007): demo and poster sessions, pages 177--180, Prague, Czeck Republic, June 2007.

[9] H. Li, M. Zhang, and J. Su. A joint source-channel model for machine transliteration. In ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 159, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[10] K. Nakatani, A. Finch, K. Tanaka-Ishii, and E. Sumita. 確率的ブロック編集距離. In Proceedings of NLP 2014, Sapporo, Japan, 2014.

[11] F. J. Och. Minimum error rate training for statistical machine translation. In Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003), Sapporo, Japan, 2003.

[12] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19--51, 2003.

[13] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of the International Conference on Spoken Language Processing, volume 2, pages 901--904, Denver, 2002.

[14] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. Journal of the ACM (JACM), 21(1):168--173, 1974.