Tuning a Grammar Correction System for Increased Precision

Anoop Kunchukuttan[*], **Sriram Chaudhury**[†], **Pushpak Bhattacharyya**[*]
[*] Department of Computer Science and Engineering, IIT Bombay, India
{anoopk,pb}@cse.iitb.ac.in

[†] Crimson Interactive Pvt. Limited, Mumbai, India
Sriram.Chaudhury@crimsoni.com

Abstract

In this paper, we propose two enhancements to a statistical machine translation based approach to grammar correction for correcting all error categories. First, we propose tuning the SMT systems to optimize a metric more suited to the grammar correction task (F-β score) rather than the traditional BLEU metric used for tuning language translation tasks. Since the F-β score favours higher precision, tuning to this score can potentially improve precision. While the results do not indicate improvement due to tuning with the new metric, we believe this could be due to the small number of grammatical errors in the tuning corpus and further investigation is required to answer the question conclusively. We also explore the combination of custom-engineered grammar correction techniques, which are targeted to specific error categories, with the SMT based method. Our simple ensemble methods yield improvements in recall but decrease the precision. Tuning the custom-built techniques can help in increasing the overall accuracy also.

1 Introduction

Grammatical Error Correction (GEC) is an interesting and challenging problem and the existing methods that attempt to solve this problem take recourse to deep linguistic and statistical analysis. In general, GEC may partly assist in solving natural language processing (NLP) tasks like Machine Translation, Natural Language Generation etc. However, a more evident application of GEC is in building automated grammar checkers thereby non-native speakers of a language. The goal is to have automated tools to help non-native speakers to generate good content by correcting grammatical errors made by them.

The CoNLL-2013 Shared Task (Ng et al., 2013) was focussed towards correcting some of the most frequent categories of grammatical errors. In contrast, the CoNLL-2014 Shared Task (Ng et al., 2014) set the goal of correcting all grammatical errors in the text. For correcting specific error categories, custom methods are generally developed, which exploit deep knowledge of the problem to perform the correction (Han et al., 2006; Kunchukuttan et al., 2013; De Felice and Pulman, 2008). These methods are generally the state-of-the-art for the concerned error categories, but a lot of engineering and research effort is required for correcting each error category. So, the custom development approach is infeasible for correcting a large number of error categories.

Hence, for correction of all the error categories, generic methods have been investigated - generally using language models or statistical machine translation (SMT) systems. The language model based method (Lee and Seneff, 2006; Kao et al., 2013) scores sentences based on a language model or count ratios of n-grams obtained from a large native text corpus. But this method still needs a candidate generation mechanism for each error category. On the other hand, the SMT based method (Brockett et al., 2006) formulates the grammar correction problem as a problem of translation of incorrect sentences to correct sentences. SMT provides a natural unsupervised method for identifying candidate corrections in the form of the translation model, and a method for scoring them with a variety of measures including the language model score. However, the SMT method requires a lot of parallel non-native learner corpora. In addition, the machinery in phrase based SMT is optimized towards solving the language translation problem. Therefore, the community has explored approaches to adapt the

Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pages 60–64,
Baltimore, Maryland, 26-27 July 2014.

SMT method for grammar correction (Buys and van der Merwe, 2013; Yuan and Felice, 2013). These include use of factored SMT, syntax based SMT, pruning of the phrase table, disabling or reordering, etc. The generic SMT approach has performed badly as compared to the specific custom made approaches (Yuan and Felice, 2013).

Our system also builds upon the SMT methods and tries to address the above mentioned lacunae in two ways:

- Tuning the SMT model to a metric suitable for grammar correction (i.e.*F*-β metric), instead of the BLEU metric.
- Combination of custom-engineered methods and SMT based methods, by using classifier based for some error categories.

Section 2 describes our method for tuning the SMT system to optimize the F-β metric. Section 3 explains the combination of classifier based method with the SMT method. Section 4 lists our experimental setup. Section 5 analyzes the results of our experiments.

2 Tuning SMT system for F-β score

We model our grammar correction system as a phrase based SMT system which translates grammatically incorrect sentences to grammatically correct sentences. The phrase based SMT system selects the best translation for a source sentence by searching for a candidate translation which maximizes the score defined by the maximum entropy model for phrase based SMT defined below:

$$P(\mathbf{e}, \mathbf{a}|\mathbf{f}) = \exp \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{a}, \mathbf{f})$$

where,
h_i: feature function for the i^{th} feature. These are generally features like the phrase/lexical translation probability, language model score, etc.
λ_i: the weight parameter for the i^{th} feature.

The weight parameters (λ_i) define the relative weights given to each feature. These parameter weights are learnt during a process referred to as *tuning*. During tuning, a search over the parameter space is done to identify the parameter values which maximize a measure of translation quality over a held-out dataset (referred to as the *tuning* set). One of the most widely used metrics for tuning is the BLEU score (Papineni et al., 2002), tuned using the Minimum Error Rate Training (MERT) algorithm (Och, 2003). Since BLEU is a form of weighted precision, along with a brevity penalty to factor in recall, it is suitable in the language translation scenario, where fidelity of the translation is an important in evaluation of the translation. Tuning to BLEU ensures that the parameter weights are set such that the fidelity of translations is high.

However, ensuring fidelity is not the major challenge in grammar correction since the meaning of most input sentences is clear and most don't have any grammatical errors. The metric to be tuned must ensure that weights are learnt such that the features most relevant to correcting the grammar errors are given due importance and that the tuning focuses on the grammatically incorrect parts of the sentences. The F-β score, as defined for the CoNLL shared task, is the most obvious metric to measure the accuracy of grammar correction on the tuning set. We choose the F-β metric as a score to be optimized using MERT for the SMT based grammar correction model. By choosing an appropriate value of β, it is possible to tune the system to favour increased recall/precision or a balance of both.

3 Integrating SMT based and error-category specific systems

As discussed in Section 1, the generic SMT based correction based systems are inferior in their correction capabilities compared to the error-category specific correction systems which have been custom engineered for the task. A reasonable solution to make optimum use of both the approaches is to develop custom modules for correcting high impact and the most frequent error categories, while relying on the SMT method for correcting other error categories. We experiment with two approaches for integrating the SMT based and error-category specific systems, and compare both with the baseline SMT approach:

- Correct all error categories using the SMT method, followed by correction using the custom modules.
- Correct only the error categories not handled by the custom modules using the SMT method, followed by correction using the custom modules.

The error categories for which we built custom modules are noun number, determiner and subject-verb agreement (SVA) errors. These errors are amongst the most common errors made by non-native speakers. The noun number and determiner errors are corrected using the classification model proposed by Rozovskaya and Roth (2013), where the label space is a cross-product of the label spaces of the possible noun number and determiners. We use the feature-set proposed by Kunchukuttan et al. (2013). SVA correction is done using a prioritized, conditional rule based system described by Kunchukuttan et al. (2013).

4 Experimental Setup

We used the NUCLE Corpus v3.1 to build a phrase based SMT system for grammar correction. The NUCLE Corpus contains 28 error categories, whose details are documented in Dahlmeier et al. (2013). We split the corpus into training, tuning and test sets are shown in Table 1.

Set	Document Count	Sentence Count
train	1330	54284
tune	20	854
test	47	2013

Table 1: Details of data split for SMT training

The phrase based system was trained using the *Moses*[1] system, with the *grow-diag-final-and* heuristic for extracting phrases and the *msd-bidirectional-fe* model for lexicalized reordering. We tuned the trained models using Minimum Error Rate Training (MERT) with default parameters (100 best list, max 25 iterations). Instead of BLEU, the tuning metric was the F-0.5 metric. We trained 5-gram language models on all the sentences from NUCLE corpus using the Kneser-Ney smoothing algorithm with *SRILM* [2].

The classifier for noun number and article correction is a Maximum Entropy model trained on the NUCLE v2.2 corpus using the MALLET toolkit. Details about the resources and tools used for feature extraction are documented in Kunchukuttan et al. (2013).

[1] http://www.statmt.org/moses/
[2] http://goo.gl/4wfLVw

5 Results and Analysis

Table 2 shows the results on the development set for different experimental configurations generated by varying the tuning metrics, and the method of combining the SMT model and custom correction modules. Table 3 shows the same results on the official CoNLL 2014 dataset without alternative answers.

5.1 Effect of tuning with F-0.5 score

We observe that both precision and recall drop sharply when the SMT model is tuned with the F-0.5 metric (system S2), as compared to tuning with the traditional BLEU metric (system S1). We observe that system S2 proposes very few corrections (82) as compared to system S1 (188), which contributes to the low recall of system S2. There are very few errors in the tuning set (202) which may not be sufficient to reliably tune the system to the F-0.5 score. It would be worth investigating the effect of number of errors in the tuning set on the accuracy of the system.

5.2 Effect of integrating the SMT and custom modules

Comparing the results of systems S1, S3 and S5, it is clear that using the SMT method alone gives the highest F-0.5 score. However, the recall is higher for systems which use the custom modules for some error categories. The recall is highest when custom modules as well as SMT method are used for the high impact error categories. The above observation is a consequence of the fact that the custom modules have higher recall for certain error categories compared to the SMT method. The lower precision of custom modules is due to the large number of false positives. If the custom modules are optimized for higher precision, then the overall ensemble can also achieve higher precision and consequently higher F-0.5 score. Thus, the integration of SMT method and custom modules can be beneficial in improving the overall accuracy of the SMT system.

6 Conclusion

We explored two approaches to adapting the SMT method for the problem of grammatical correction. Tuning the SMT system to the F-β metric did not improve performance over the BLEU-based tuning. However, we plan to further investigate to understand the reasons for this behaviour. We

Id	SMT Data	Custom Modules	Tuning Metric	%P	%R	%F-0.5
S1	All errors	No	BLEU	62.23	11.53	33.12
S2		No	F-0.5	55.32	5.13	18.71
S3		Yes	BLEU	10.99	26.33	12.44
S4		Yes	F-0.5	9.80	22.98	11.07
S5	All errors, except Nn, ArtOrDet, SVA	Yes	BLEU	10.15	23.96	11.47

Table 2: Experimental Results for various configurations on the development set

Id	SMT Data	Custom Modules	Tuning Metric	%P	%R	%F-0.5
S1	All errors	No	BLEU	38.81	4.15	14.53
S2		No	F-0.5	30.77	1.39	5.90
S3		Yes	BLEU	29.02	17.98	25.85
S4		Yes	F-0.5	28.23	16.72	24.81
S5	All errors, except Nn, ArtOrDet, SVA	Yes	BLEU	28.67	17.29	25.34

Table 3: Experimental Results for various configurations on the CoNLL-2014 test set without alternatives

also plan to explore tuning for recall and other alternative metrics which could be useful in some scenarios. An ensemble of the SMT method and custom methods for some high impact error categories was shown to increase the recall of the system, and with proper optimization of the system can also improve the overall accuracy of the correction system.

References

Chris Brockett, William B Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*.

Jan Buys and Brink van der Merwe. 2013. A Tree Transducer Model for Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better Evaluation for Grammatical Error Correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The NUS Corpus of Learner English. In *To appear in Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*.

Rachele De Felice and Stephen G Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*.

Ting-hui Kao, Yu-wei Chang, Hsun-wen Chiu, Tzu-Hsi Yen, Joanne Boisson, Jian-cheng Wu, and Jason S. Chang. 2013. CoNLL-2013 Shared Task: Grammatical Error Correction NTHU System Description. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*.

Anoop Kunchukuttan, Ritesh Shah, and Pushpak Bhattacharyya. 2013. IITB System for CoNLL 2013 Shared Task: A Hybrid Approach to Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*.

J. Lee and S. Seneff. 2006. Automatic grammar correction for second-language learners. In *Proceedings of Interspeech*, pages 1978–1981.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of*

the Eighteenth Conference on Computational Natural Language Learning.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*.

A. Rozovskaya and D. Roth. 2013. Joint Learning and Inference for Grammatical Error Correction. In *EMNLP*.

Zheng Yuan and Mariano Felice. 2013. Constrained Grammatical Error Correction using Statistical Machine Translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*.