Grammatical error correction using neural machine translation

Zheng Yuan and **Ted Briscoe**
The ALTA Institute
Computer Laboratory
University of Cambridge
{zy249,ejb}@cam.ac.uk

Abstract

This paper presents the first study using neural machine translation (NMT) for grammatical error correction (GEC). We propose a two-step approach to handle the rare word problem in NMT, which has been proved to be useful and effective for the GEC task. Our best NMT-based system trained on the CLC outperforms our SMT-based system when testing on the publicly available FCE test set. The same system achieves an $F_{0.5}$ score of 39.90% on the CoNLL-2014 shared task test set, outperforming the state-of-the-art and demonstrating that the NMT-based GEC system generalises effectively.

1 Introduction

Grammatical error correction (GEC) is the task of detecting and correcting grammatical errors in text written by non-native English writers. Unlike building machine learning classifiers for specific error types (e.g. determiner or preposition errors) (Tetreault and Chodorow, 2008; Rozovskaya and Roth, 2011; Dahlmeier and Ng, 2011), the idea of 'translating' a grammatically incorrect sentence into a correct one has been proposed to handle all error types simultaneously (Felice et al., 2014; Junczys-Dowmunt and Grundkiewicz, 2014). Statistical machine translation (SMT) has been successfully used for GEC, as demonstrated by the top-performing systems in the CoNLL-2014 shared task (Ng et al., 2014).

Recently, several neural machine translation (NMT) models have been developed with promising results (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2014). Unlike SMT, which consists of components that are trained separately and combined during decoding (i.e. the translation model and language model) (Koehn, 2010), NMT learns a single large neural network which inputs a sentence and outputs a translation. NMT is appealing for GEC as it may be possible to correct erroneous word phrases and sentences that have not been seen in the training set more effectively (Luong et al., 2015). NMT-based systems thus may help ameliorate the lack of large error-annotated learner corpora for GEC.

However, NMT models typically limit vocabulary size on both source and target sides due to the complexity of training (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015; Jean et al., 2015). Therefore, they are unable to translate rare words, and out-of-vocabulary (OOV) words are replaced with *UNK* symbol. This problem is more serious for GEC as non-native text contains not only rare words (e.g. proper nouns), but also misspelled words (i.e. spelling errors). By replacing all the OOV words with the same *UNK* symbol, useful information is discarded, resulting in systems that are not able to correct misspelled words or even keep some of the error-free original words, as in the following examples (OOV words are underlined):

Original sentence

*... I am **goign** to make a plan ...*

System hypothesis

... I am UNK to make a plan ...

Gold standard

Proceedings of NAACL-HLT 2016, pages 380–386,
San Diego, California, June 12-17, 2016.

*... I am **going** to make a plan ...*

Original sentence

I suggest you visit first the cathedral of " Le Seu d'Mrgell " because it is the most emblematic building in the area .

System hypothesis

I suggest you visit first the cathedral of " Le UNK UNK " because it is the most UNK building in the area .

Gold standard

I suggest you visit first the cathedral of " Le Seu d'Mrgell " because it is the most emblematic building in the area . (unchanged)

Inspired by the work of Luong et al. (2015), we propose a similar but much simpler two-step approach to address the rare word problem: rather than annotating the training data with alignment information, we apply unsupervised alignment models to find the sources of the words in the target sentence. Once we know the source words that are responsible for the unknown target words, a word level translation model learnt from parallel sentences is used to translate these source words.

This paper makes the following contributions. First, we present the first study using NMT for GEC, outperforming the state-of-the-art. Second, we propose a two-step approach to address the rare word problem in NMT for GEC, which we show yields a substantial improvement. Finally, we report results on two well-known publicly available test sets that can be used for cross-system comparisons.

2 Neural machine translation

NMT systems apply the so-called *encoder-decoder* mechanism proposed by Cho et al. (2014) and Sutskever et al. (2014). An encoder reads and encodes an entire source sentence $\mathbf{x} = (x_1, x_2, ..., x_T)$ into a vector c:

$$c = q(h_1, h_2, ..., h_T) \quad (1)$$

where a hidden state h_t at time t is defined as:

$$h_t = f(x_t, h_{t-1}) \quad (2)$$

A decoder then outputs a translation $\mathbf{y} = (y_1, y_2, ..., y_{T'})$ by predicting the next word y_t based on the encoded vector c and all the previously predicted words $\{y_1, y_2, ..., y_{t-1}\}$:

$$p(\mathbf{y}) = \prod_{t=1}^{T'} p(y_t|\{y_1, y_2, ..., y_{t-1}\}, c) = \prod_{t=1}^{T'} g(y_{t-1}, s_t, c) \quad (3)$$

where s_t is the hidden state of the decoder.

Different neural network models have been proposed, for example, Kalchbrenner and Blunsom (2013) proposed a hybrid of a recurrent neural network (RNN) and a convolutional neural network, Sutskever et al. (2014) used a Long Short-Term Memory (LSTM) model, Cho et al. (2014) proposed a similar but simpler gated RNN model, and Bahdanau et al. (2014) introduced an attentional-based architecture.

In this work, we use the *RNNsearch* model of Bahdanau et al. (Bahdanau et al., 2014), which contains a bidirectional RNN as an encoder and an attention-based decoder. The bidirectional RNN encoder has a forward and a backward RNN. The forward RNN reads the source sentence from the first word to the last, and the backward RNN reads the source sentence in reverse order. By doing this, it captures both historical and future information. The attention-based model allows the decoder to focus on the most relevant information in the source sentence, rather than remembering the entire source sentence.

3 Handling rare words

The rare word problem in NMT has been noticed by (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015; Jean et al., 2015). Jean et al. (2015) proposed a method based on importance sampling that uses a very large target vocabulary without increasing training complexity. However, no matter how large the target vocabulary size is, there are still OOV words. We also notice that in GEC, the source side vocabulary size is much larger than that of the target side as there are many incorrect words in the source (e.g. spelling mistakes and word form errors) (see Section 4.1). Luong et al. (2015) introduced three new annotation strategies to annotate the training data, so that unknown words in the output can be

traced back to their origins. The training data was first re-annotated using the output of a word alignment algorithm. NMT systems were then built using this new data. Finally, information about the OOV words in the target sentence and their corresponding words in the source sentence was extracted from the NMT systems and used in a post-processing step to translate these OOV words using a dictionary.

We propose a similar two-step approach: 1) aligning the unknown words (i.e. *UNK* tokens) in the target sentence to their origins in the source sentence with an unsupervised aligner; 2) building a word level translation model to translate those words in a post-processing step. In order to locate the source words that are responsible for the unknown target words, we apply unsupervised aligners directly and use only the NMT model output instead of first re-annotating training data, and then building new NMT models using this newly annotated data as proposed by Luong et al. (2015). Our approach is much simpler as we avoid re-annotating any data and train only one NMT model. Due to the nature of error correction (i.e. both source and target sentences are in the same language), most words translate as themselves, and errors are often similar to their correct forms. Thus, unsupervised aligners can be successfully used to align the unknown target words. Two automatic alignment tools are used: GIZA++ (Och and Ney, 2003) and METEOR (Banerjee and Lavie, 2005). GIZA++ is an implementation of IBM Models 1-5 (Brown et al., 1993) and a Hidden-Markov alignment model (HMM) (Vogel et al., 1996), which can align two sentences from any languages. Unlike GIZA++, METEOR aligns two sentences from the same language. The latest METEOR 1.5 only supports a few languages, and English is one of them. METEOR identifies not only words with exact matches, but also words with identical stems, synonyms, and unigram paraphrases. This is useful for GEC as it can deal with word form, noun number, and verb form corrections that share identical stems, as well as word choice corrections with synonyms or unigram paraphrases. To build a word level translation model for translating the source words that are responsible for the target unknown words, we need word-aligned data. The IBM Models are used to learn word alignment from parallel sentences.

4 Experiments

4.1 Dataset

We use the publicly available FCE dataset (Yannakoudakis et al., 2011), which is a part of the Cambridge Learner Corpus (CLC) (Nicholls, 2003). The FCE dataset contains 1,244 scripts produced by learners taking the First Certificate in English (FCE) examination between 2000 and 2001. The texts have been manually annotated by linguists using a taxonomy of approximately 80 error types. The publicly available FCE dataset contains about 30,995 pairs of parallel sentences for training (approx. 496,567 tokens on the target side) and about 2,691 pairs of parallel sentences for testing (approx. 41,986 tokens on the target side). Since the FCE training set is too small to build good MT systems, we add training examples extracted from the CLC. Overall, there are 1,965,727 pairs of parallel sentences in our training set. The source side contains 28,823,615 words with 248,028 unique words, and the target side contains 29,219,128 words with 143,852 unique words. As we can see, the source side vocabulary size is much larger than that of the target side. Training and test data is pre-processed using RASP (Briscoe et al., 2006).

4.2 Evaluation

System performance is evaluated using three automatic evaluation metrics: I-measure (Felice and Briscoe, 2015), M^2 Scorer (Dahlmeier and Ng, 2012) and GLEU (Napoles et al., 2015). In the I-measure, an Improvement (I) score is computed by comparing system performance with that of a baseline which leaves the original text uncorrected (i.e. the source). The M^2 Scorer was the official scorer in the CoNLL shared tasks (Ng et al., 2013; Ng et al., 2014), with $F_{0.5}$ being the reported metric in the 2014 edition. GLEU is a simple variant of BLEU (Papineni et al., 2002), which shows better correlation with human judgments on the CoNLL-2014 shared task test set.

4.3 SMT baseline

Following previous work (e.g. Brockett et al. (2006), Yuan and Felice (2013)), we build a phrase-based SMT error correction system as the baseline. Pialign (Neubig et al., 2011) is used to create a phrase

translation table. In addition to default features, we add character-level Levenshtein distance to each mapping in the phrase table as proposed by Felice et al. (2014). Decoding is performed using Moses (Koehn et al., 2007). The language model used during decoding is built from the corrected sentences in the learner corpus, to make sure that the final system outputs fluent English sentences. The IRSTLM Toolkit (Federico et al., 2008) is used to buid a 5-gram language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995).

4.4 NMT training details

Our training procedure and hyper-parameters for the NMT system are similar to those used by Bahdanau et al. (2014). We train models with sentences of length up to 100 words, which covers about 99.96% of all the training examples. In terms of vocabulary size, we limit the target vocabulary size to 30K, and experiment with three different source vocabulary sizes: 30K, 50K and 80K.[1] Each model is trained for approximately 5 days using a Tesla K20 GPU.

The output sentences from the NMT systems are aligned with their source sentences using GIZA++. In addition, alignment information learnt by METEOR is used by GIZA++ during aligning. All the *UNK* tokens in the output sentences are replaced with the translation of the source words that are responsible for those *UNK* tokens. The translation is performed using a word level model learnt from IBM Model 4.

4.5 Results

From the results in Table 1, we can see that NMT-based systems alone are not able to achieve comparable results to an SMT-based system. It is probably because of the rare word problem, as increasing the source side vocabulary size helps. The performance of the best NMT system alone (*NMT 80K-30K*), without replacing *UNK* tokens, is still worse than the SMT baseline. When we replace the *UNK* tokens in the NMT output, using GIZA++ for unknown word alignment improves the system performance for all three NMT systems in all three evaluation metrics. We can see that our proposed approach is more useful for NMT systems trained on a small source side vocabulary (e.g. 30K) than a large vocabulary (e.g. 50K, 80K). The larger the vocabulary size, the smaller the gain after replacing *UNK* tokens. The introduction of the METEOR alignment information to GIZA++ yields further improvements. Our best system (*NMT 30K-30K + GIZA++ + METEOR*) achieves an $F_{0.5}$ score of 53.49%, an I score of 3.94%, and a GLEU score of 71.16%, outperforming the SMT baseline in all three evaluation metrics.

System	GLEU	$F_{0.5}$ (M^2)	I-measure
Source	60.39	0	0
SMT baseline	70.15	52.90	2.87
NMT-based systems			
NMT 30K-30K	69.04	46.10	-1.30
+ GIZA++	70.89	52.79	3.89
+ METEOR	**71.16**	**53.49**	**3.94**
NMT 50K-30K	68.95	46.78	-1.14
+ GIZA++	70.31	52.02	2.86
+ METEOR	70.40	52.35	2.89
NMT 80K-30K	70.02	49.17	-1.04
+ GIZA++	71.18	53.48	2.40
+ METEOR	71.18	53.49	2.41

Table 1: System performance on the FCE test set (in percentages). The results of our best system are marked in **bold**.

Comparing the output of the SMT baseline with that of the NMT system reveals that there are some learner errors which are missed by the SMT system but are captured by the NMT system. One possible reason is that the phrase-based SMT system is trained on surface forms and therefore unaware of syntactic structure. In order to make a correction, it has to have seen the exact correction rule in the training data. Since the NMT system does not rely on any correction rules, in theory, it should be able to make any changes as long as it has seen the words in the training data. For example:

Original sentence

*There are **kidnaps** everywhere and not all of the **family** can afford the ransom ...*

SMT hypothesis

*There are **kidnaps** everywhere and not all of the **families** can afford the ransom ...*

NMT hypothesis

[1] Preliminary experiments show that increasing the source side vocabulary size is more useful than target side.

*There are **kidnappings** everywhere and not all of the **families** can afford the ransom ...*

Gold standard

*There are **kidnappings** everywhere and not all of the **families** can afford the ransom ...*

The SMT system fails to correct the word form error as the correction rule (*kidnaps → kidnappings*) is not in the SMT phrase table learnt from the training data. Since these two words (*kidnaps* and *kidnappings*) have been seen in the training data, the NMT system corrects this error successfully.

5 CoNLL-2014 shared task

The CoNLL-2014 shared task on grammatical error correction required participating systems to correct all errors present in learner English text. The official training and test data comes from the National University of Singapore Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013). $F_{0.5}$ was adopted as the evaluation metric, as reported by the M^2 Scorer. In order to test how well our system generalises, we apply our best system trained on the CLC to the CoNLL-2014 shared task test data directly without adding the NUCLE training data or tuning for the NUCLE. The state-of-the-art $F_{0.5}$ score was reported by Susanto et al. (2014) after the shared task. By combining the outputs from two classification-based systems and two SMT-based systems, they achieved an $F_{0.5}$ score of 39.39%. Results of the uncorrected baseline, our best NMT-based system, Susanto et al. (2014)'s system and the top three systems in the shared task are presented in Table 2. We can see that our NMT-based system outperforms the top three teams, achieving the highest $F_{0.5}$, I and GLEU scores. It also outperforms the state-of-the-art combined system from Susanto et al. (2014). Our system achieves the best $F_{0.5}$ score of 39.90% even though it is not trained on the NUCLE data. This result shows that our system generalises well to other datasets. We expect these results might be further improved by retokenising the test data to be consistent with the tokenisation of the CLC.[2]

[2]The NUCLE data was preprocessed using the NLTK toolkit, whereas the CLC was tokenised with RASP.

System	GLEU	$F_{0.5}$ (M^2)	I-measure
Source	64.19	0	0
Our NMT-based system			
30K-30K + GIZA++ + METEOR	65.59	39.90	-3.11
Top 3 systems in CoNLL-2014			
CAMB (Felice et al., 2014)	64.32	37.33	-5.58
CUUI (Rozovskaya et al., 2014)	64.64	36.79	-3.91
AMU (Junczys-Dowmunt and Grundkiewicz, 2014)	64.56	35.01	-3.31
State-of-the-art			
Susanto et al. (2014)	n/a	39.39	n/a

Table 2: System performance on the CoNLL-2014 test set without alternative answers (in percentages).

6 Conclusions

We have shown that NMT can be successfully applied to GEC once we address the rare word problem. Our proposed two-step approach for *UNK* replacement has been proved to be effective, and to provide a substantial improvement. We have developed an NMT-based system that generalises well to another dataset. Our NMT system achieves an $F_{0.5}$ score of 53.49%, an I score of 3.94%, and a GLEU score of 71.16% on the publicly available FCE test set, outperforming an SMT-based system in all three metrics. When testing on the official CoNLL-2014 test set without alternative answers, our system achieves an $F_{0.5}$ score of 39.90%, outperforming the current state-of-the-art. In future work, we would like to explore other ways to address the rare word problem in NMT-based GEC, such as incorporating the soft-alignment information generated by the attention-based decoder, or using character-based models instead of word-based ones.

Acknowledgements

We would like to thank Cambridge English Language Assessment and Cambridge University Press for granting us access to the CLC for research purposes as well as the anonymous reviewers for their comments and suggestions. We acknowledge NVIDIA for an Academic Hardware Grant. This work also used the Wilkes GPU cluster at the University of Cambridge High Performance Computing Service, provided by Dell Inc., NVIDIA and Mellanox, and part funded by STFC with industrial sponsorship from Rolls Royce and Mitsubishi Heavy Industries.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80.

Chris Brockett, William B. Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the COLING/ACL 2006*, pages 249–256.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 915–923.

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: Human Language Technologies*, pages 568–572.

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 *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, pages 1618–1621.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 578–587.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task*, pages 15–24.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1–10.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task*, pages 25–33.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 181–184.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation. In *Proceedings of the ACL-IJCNLP*, pages 11–19.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 588–593.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 632–641.

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: Shared Task*, pages 1–12.

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 18th Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Diane Nicholls. 2003. The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 572–581.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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 of the Association for Computational Linguistics*, pages 311–318.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 924–933.

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. 2014. The Illinois-Columbia System in the CoNLL-2014 Shared Task. In *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task*, pages 34–42.

Hendy Raymond Susanto, Peter Phandi, and Tou Hwee Ng. 2014. System combination for grammatical error correction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 951–962.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Joel R. Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 865–872.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational linguistics*, volume 2, pages 836–841.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.

Zheng Yuan and Mariano Felice. 2013. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the 17th Conference on Computational Natural Language Learning: Shared Task*, pages 52–61.