GEC into the future: Where are we going and how do we get there?

Keisuke Sakaguchi,[1] **Courtney Napoles**,[1] and **Joel Tetreault**[2]
[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Grammarly
{keisuke,napoles}@cs.jhu.edu, joel.tetreault@grammarly.com

Abstract

The field of grammatical error correction (GEC) has made tremendous bounds in the last ten years, but new questions and obstacles are revealing themselves. In this position paper, we discuss the issues that need to be addressed and provide recommendations for the field to continue to make progress, and propose a new shared task. We invite suggestions and critiques from the audience to make the new shared task a community-driven venture.

1 Introduction

In the field of grammatical error correction (GEC), the Helping Our Own shared tasks in 2011 (Dale and Kilgarriff, 2011) and 2012 (Dale et al., 2012), and then the CoNLL shared tasks of 2013 (Ng et al., 2013) and 2014 (Ng et al., 2014) marked a sea change. For the first time there were public datasets, most notably the NUS Corpus of Learner English (NUCLE; Dahlmeier et al., 2013), and evaluation metrics, of which the most commonly used to date is M^2 (Dahlmeier and Ng, 2012). This has allowed researchers from other fields, such as machine translation, to enter GEC more easily. It has also enabled new developments, with many papers published on metrics, new algorithms (most recently neural methods), and occasionally new datasets.

Even with the accelerated progress in GEC, problems yet remain in the field. The use of specific datasets may be GEC's worst enemy, as system and even evaluation metric development rely too heavily on the NUCLE test set. While probably one of the most important contributions to the field's development to date, the lack of publicly available alternatives has caused some over-optimization. Other issues have also gone undiscussed. For example, nearly all work that has been published in the NLP community has focused on standalone systems, and very few investigate their impact on downstream users, except, e.g., Nagata and Nakatani (2010); Chodorow et al. (2010).

In this short paper, we take stock of the current state of GEC (§2) and its limitations (§3), and outline where we believe the field should be five years from now (§4). We finish with a recommendation for a new *community-driven* shared task that will help the field progress even further (§5). We look forward to discussing this proposal with the community and to refine a shared task for 2018.

2 GEC: A Quick Retrospective

A complete retrospective is outside the scope of this paper and thus we focus on two key aspects of the field: For a more detailed review of the field, we refer the reader to Leacock et al. (2014).

2.1 Datasets

There are several error-annotated corpora, and for the purposes of this paper, we only focus on the most recent public datasets. The size and characteristics of each corpus is summarized in Table 1. The most frequently used corpus for GEC is NUCLE, which was the official dataset of the 2013 and 2014 shared tasks. It is a collection of essays written by students at the National University of Singapore (Dahlmeier et al., 2013). The test set and system results from the most recent shared task were released to the community (Ng et al., 2014), and have been the focus of recent work on automatic metrics (see §2.2). Additionally, this test set has been augmented with eight additional annotations from Bryant and Ng (2015) and eight from Sakaguchi et al. (2016).

The Cambridge Learner Corpus (CLC) contains a broader representation of native languages than the NUCLE, however only the First Certificate in

	Num.	Num.	Sents.	Err. type	Fluency	Err. span	Diverse	Diverse	Diverse	Native
Corpus	refs.	sent.	changed	labeled	edits	$>$1 sent.	proficiency	topic	L1	speakers
NUCLE	59k	2	38%	✓	(✗)	✓	✗	✗	✗	✗
FCE	34k	1	62%	✓	✗	✓	✓	✓	✓	✗
Lang-8	2.5M	$\geq$1	42%	✗	✓	✓	✓	✓	✓	✗
AESW	1.2M	1	39%	✗	✗	✓	✗	✗	✓	✓+✗
JFLEG	1.5k	4	86%	✗	✓	✗	✓	✓	✓	✗

Table 1: GEC corpora available for free (for research purposes) and desired properties, identified in §3.1. ✓ and ✗ indicate whether the corpus exhibits each property. Fluency edits for the NUCLE test set were added by Sakaguchi et al. (2016).

English (FCE) portion is publicly available (Yannakoudakis et al., 2011). The FCE is approximately the same size as NUCLE and was used for the 2012 shared task. However it has not been used to the same extent as NUCLE, presumably because it lacks multiple annotations and the 2012 shared task system outputs were not released.

All of the corpora described above have been annotated with spans of text containing an error and assigned an error code. Unlike these, the Lang-8 Learner Corpora Corpus of Learner English (Tajiri et al., 2012) is a parallel set of original and corrected sentences from `lang-8.com`, an online community of language learners who post text that is corrected by other users. It is also the largest public GEC corpora, with more than 2 million English sentences.[1] Another large corpus currently available was released for the first Automatic Evaluation of Scientific Writing shared task (AESW; Daudaravicius et al., 2016). Unlike the other corpora, it contains scientific writing by native and non-native English speakers, corrected by professional editors. Because the writers are highly proficient, there is a lower diversity of errors than the other corpora. More than half of the errors are related to punctuation (Flickinger et al., 2016), which compose less than 7% of NUCLE errors.

Finally, the JHU FLuency-Extended GUG corpus (JFLEG) is a small dataset for tuning and evaluating GEC systems. 1.5k sentences are taken from the GUG corpus (Heilman et al., 2014), which labels sentences with an ordinal grammaticality score. In JFLEG, each sentence is corrected four times for grammaticality and *fluency* (Sakaguchi et al., 2016).

2.2 Evaluation

Precision, recall, and F-score have been used to evaluate GEC systems that correct targeted error types. Three additional evaluation metrics have been proposed for GEC: MaxMatch (M^2; Dahlmeier and Ng, 2012), I-measure (Felice and Briscoe, 2015), and GLEU (Napoles et al., 2015). The first two metrics compare the changes made in the output to error-coded spans of the reference corrections. M^2 was the metric used for the 2013 and 2014 CoNLL GEC shared tasks (Ng et al., 2013, 2014). It captures word- and phrase-level edits by building an edit lattice and calculating an F-score over the lattice. I-measure (IM) is based on token-level alignment-based accuracy among the source, hypothesis, and gold-standard. IM considers the distinction between "do-nothing (already grammatical) baseline" and systems that only propose wrong corrections (i.e., make the source sentence worse). Unlike these two approaches, GLEU does not need error-coded references (Napoles et al., 2015). Based on BLEU (Papineni et al., 2002), it computes n-gram precision of the system output against reference sentences, and additionally penalizes n-grams in the hypothesis that should have been corrected but failed.

3 Limitations

3.1 Problems with Datasets

As we saw in the previous section, the majority of the commonly used datasets are limited to students, specifically college-level ESL writers. To date, the overwhelmingly majority of publications benchmark on NUCLE, save for a few exceptions such as Cahill et al. (2013) and Rei and Yannakoudakis (2016) which means that research efforts are becoming over-optimized for one set. This lack of diversity means that it is not clear how systems perform on other genres under different training conditions. We should look to the parsing community as a warning sign. For well over a decade, the field was heavily focused on improving parsing accuracy on the Penn Treebank (Marcus et al., 1993), but robustness was greatly improved with the advent of Ontonotes (Hovy et al., 2006) and the Google Web Treebank (Petrov and

[1] Because of noise and implementation differences in sentence extraction, the size varies from 2–2.5 million sentences.

	GLEU	IM [-100,	M^2 [0, 100]		
System	[0,100]	100]	P	R	$F_{0.5}$
“a”	0.2	0.0	28.4	31.3	28.9
“a a”	0.6	0.0	28.7	31.8	29.3
“a a a”	1.6	0.0	28.7	32.0	29.4
Source	57.4	0.0	100.0	0.0	0.0
CAMB14	64.3	-5.3	39.7	30.1	37.3
CUUI14	64.6	-2.2	41.8	24.9	36.8
AMU14	64.6	-2.5	41.6	21.4	35.0
Src>Game	✓	✗	✓	✗	✗
Src<Sys	✓	✗	✗	✓	✓

Table 2: Metric scores of three artificially contrived systems (Game), input source sentences (Src), and top 3 system outputs (Sys) on CoNLL14 data. The bottom two rows show whether each metric scores the systems better than Game or worse than Source. Humans judge all systems be better than over Source.

McDonald, 2012).

Another issue is training data size. The sister field of machine translation (MT) usually has datasets in the orders of millions or even tens of millions of sentence pairs. The largest GEC datasets barely approach that figure, with 2.5 million sentences at a maximum, a number which includes sentences that were not corrected.

Table 1 summarizes the strengths and weaknesses of the most commonly used GEC corpora across different properties ranging from size to diversity in native language (L1). The most notable weakness across corpora is the lack of multiple reference corrections. NUCLE contains two corrections per sentence and JFLEG 4. M^2 and GLEU scores increase with more references but at a diminishing rate (Bryant and Ng, 2015; Sakaguchi et al., 2016). Further investigation is warranted to determine what an ideal number of references is, given the trade off between cost and reliability. Some corpora contain little diversity in proficiency, topic, and/or native language of the writers (namely NUCLE and AESW), however AESW is the only corpus to contain sentences by native English speakers.

3.2 Problems with Evaluation

The 2014 CoNLL shared task has enabled, for the first time, the development of evaluation metrics. These metrics are evaluated by comparing their ranking of the shared task systems with the ranking done by human annotators. Sakaguchi et al. (2016) showed that GLEU could rank systems closer to a human ranking than M^2 and IM, and a higher correlation could be found when combining GLEU with a reference-less fluency metric (Napoles et al., 2017). However, it is important to take these results with a grain of salt—all benchmarking of the metrics was done with the CoNLL 2014 systems and data, and it remains to be seen if this ranking would hold on other, larger datasets.

Another issue with the metrics is the number of references available for comparison. As in machine translation, the more references (human-generated gold-standard corrections) one has, the better one can evaluate a system. The CoNLL 2014 test set has 18 references annotated, but one can find examples where a system produces a correction which is not reflected in the references. This gets more complicated when human raters feel it is necessary to rewrite a sentence.

A third issue is that no metric directly measures meaning preservation. This means that a system could produce a more fluent version of the original but accidentally change one word, and that could change the meaning of the whole sentence. For example, if a system accidentally corrected *documentary* to *document* in “The documentary gave a nice summary of global warming.” By current metrics, that error would have the same penalty as a minor spelling mistake.

Finally, the most commonly used GEC metric, M^2, has a serious weakness, which has been noted in earlier papers (Felice and Briscoe, 2015; Sakaguchi et al., 2016; Bryant et al., 2017). The phrasal alignments under-penalize a sequence of incorrect tokens, and to illustrate how troubling this is, we tested a series of dummy systems, where each system produces the same sentence regardless of input (the sentences produced by each system are *a*, *a a*, and *a a a*). Table 2 shows their scores on the CoNLL 2014 test set evaluated on the official NUCLE references (without alternatives), compared to the top 3 systems in the shared task, CAMB14 (Felice et al., 2014), CUUI14 (Rozovskaya et al., 2014), and AMU14 (Junczys-Dowmunt and Grundkiewicz, 2014). The reader will notice that GLEU and IM score these sentences at or near zero, however according to M^2, the dummy system that only returns the string “a a” scores higher than 7/13 systems participating in the 2014 Shared Task. The IM score is also problematic in that the gamed sentences have the same score as the source.

		Metric score (rank)		
		GLEU	IM	M^2
System	Sentence	[0,100]	[-100,100]	[0,100]
Source	In both advertisements is said that these tooth pastes will make your teeth briliant and brighter .	15.7 (4)	0.0 (4)	0.0 (5)
Reference	□ Both advertisements □ say that the toothpaste will make your teeth brilliant and brighter .	50.7 (1)	17.4 (3)	65.2 (3)
AMU16 & NUS16	In both advertisements is said that these tooth pastes will make your teeth briliant and brighter .	15.7 (4)	0.0 (4)	0.0 (5)
CAMB14	In both advertisements is said that these tooth pastes will make your teeth brilliant and brighter .	35.5 (3)	100.0 (1)	83.3 (1)
CAMB16	In both advertisements it is said that these tooth problems will make your teeth brilliant and brighter .	39.5 (2)	56.1 (2)	71.4 (2)
Dummy	a a a .	2.9 (5)	-47.7 (5)	52.6 (4)

Table 3: An original source sentence and candidate corrections, along with the score of each sentence from different metrics. Changed or inserted spans are underlined and □ indicates deletions.

4 Looking into the Future

In this section we outline our recommendations for how the field should develop.

4.1 Data

As the world's communication is not limited to college-level essays, it is important that we have datasets which better represent as much breadth as possible. Ideally, datasets should span different genres (such as emails, blog posts, and official documents) and include content from both native and non-native speakers, as well as from different proficiency levels. All of these changes will enable the field to better assess how we are helping *more* of the world's writers under different conditions, and also enable one to test adaptation between domains.

4.2 Evaluation

We envision evaluation metrics which check that corrections are not only grammatically valid, but also check that the corrections are native-sounding and preserve the original meaning or intent of the writer. Future metrics should be easy to compute and be interpretable. For instance, a range between -1 and 1 may be preferred (like IM uses), since it is possible a suggested set of corrections could produce a sentence which is *worse* than the original. If multiple references are used, metrics should assign credit to corrections which match different references in different places, assuming the outcome is overall coherent. In addition, most (if not all) evaluation schemes to date have focused on the sentence as the minimal unit. It would be good to take the entire document into account and allow for more global rewrites, such as consistent tense.

Ultimately, a metric should say whether or not a system has attained the same level of performance as a human judge. One way of doing this is through a GEC Turing Test, where system outputs are blindly judged alongside human corrections of the same sentences. If human adjudicators think the system outputs are indistinguishable in quality from the human corrections (for example, given a set of criteria such as being good corrections, meaning preserving and native-sounding) then that is a very strong signal that GEC has attained human-level performance.

To illustrate the shortcomings of current metrics, Table 3 contains a JFLEG sentence corrected by current leading systems (AMU16 (Junczys-Dowmunt and Grundkiewicz, 2016); NUS16 (Chollampatt et al., 2016); CAMB16 (Yuan and Briscoe, 2016)) and the automatic metric scores.[2] Notice that the CAMB16 sentence, which changes *tooth pastes* → *tooth problems*, is ranked the highest system output by GLEU and the second highest by IM and M^2. All metrics score it higher than the unchanged Source sentence. Another issues evidenced in the table is that IM and M^2 score the imperfect correction (CAMB14) as better than Reference; and according to M^2, the Dummy output is better than Source.

We believe that the GEC field should take

[2] All metrics run with default settings. Reference is evaluated against the other 3 references; other sentences are evaluated against all 4 references.

notes from the Workshop on Machine Translation (WMT) (Bojar et al., 2016). There the participants in the evaluation shared tasks are also responsible for contributing system ranking judgments. This makes the whole effort more community-driven and takes the pressure off one group from having to supply all annotations.

4.3 Consensus on Goals and Applications

As a corollary to data and metrics, the end-goal of GEC also needs to be refined within the community. Initial approaches to GEC seemed to focus on providing feedback to English language learners where specific error types would be targeted and feedback would be given in terms of detection or possible corrections. The work was also motivated by concurrent work in using NLP for automatic essay scoring (for example, Attali and Burstein (2006)). Chodorow et al. (2012) noted several other applications for GEC: improving overall writing quality for both native and non-native writers, assistive language learning, and applications within NLP (such as post-editing in MT). More recently the field has drifted to "whole sentence GEC" using statistical or neural MT approaches. In this situation, the writer simply gets a complete rewrite of their sentence, which may be useful as an instructional tool in some circumstances, but not all.

There is no consensus on what the focus application(s) should be. Which application determines which methods and which evaluation metrics one uses. For example, if one wants to provide feedback to language learners, then a high-precision, interpretable method is preferred. Conversely, if the application is simply to automatically clean up one's writing without any feedback, then a whole sentence approach may be preferred. Very few papers delve into error detection and correction for goals other than whole-sentence error correction or targeted feedback for ESL writers. Datasets and metrics should be created with a specific goal in mind. Thus, the field should reassess what are the goals and how we evaluate with respect to these goals.

5 Proposal for a New Shared Task

We believe it is time for another shared task in the field, this one designed with consideration the field should be several years from now. The CoNLL shared tasks were instrumental in unifying the field with a common benchmark corpus and metric, and the AESW shared task provided data from a new domain to evaluate on. We recommend the following:

- **Data**: A new corpus for training and evaluation that spans different genres. We have already begun collecting conversational data from native and non-native writers and from genres other than essays, such as emails. Our aim is to construct a corpus larger than the NUCLE to support the development of data-hungry methods such as neural MT.
- **Annotation**: The data is corrected for fluency with crowdsourcing as in Sakaguchi et al. (2016) which is a cheap and efficient way of collecting annotations of reasonable quality. Error types can be automatically tagged using a method such as that described in Bryant et al. (2017)
- **Metric Evaluation**: Borrowing from the WMT community, the shared task should also be a venue to improve automatic GEC evaluation. Participants will provide judgments on system rankings.

We invite discussion from the community and seek others to help contribute data, annotations and other resources to make this a community-driven event. Our goal is to host a shared task in 2018. We believe that this type of collaboration has made the WMT evaluations a success, and will similarly benefit GEC. We have set up a public mailing list where others can post their comments and suggestions: `https://groups.google.com/forum/#!forum/gec-sharedtask`.

6 Conclusions

The goal of this paper is to laud the progress that the GEC field has made, but also highlight the limitations that must be addressed for the field to grow further. The reliance on a few narrow datasets is problematic as it has a major impact on system development and metric development, as well as robustness when applying these approaches in the real world. Our concern is that unless data and metrics are improved, it will be hard to assess the value of new algorithms optimized for a small set of datasets and metrics. We list a recommendation for a new shared task to fuel discussion offline as well as at the BEA12 Workshop in Copenhagen.[3]

[3] `https://www.cs.rochester.edu/~tetreaul/emnlp-bea12.html`

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment* 4(3).

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W16-2301.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Beijing, China, pages 697–707. http://www.aclweb.org/anthology/P15-1068.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust systems for preposition error correction using wikipedia revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 507–517. http://www.aclweb.org/anthology/N13-1055.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 611–628. http://www.aclweb.org/anthology/C12-1038.

Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. The utility of grammatical error detection systems for English language learners: Feedback and assessment. *Language Testing* 27(3).

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1901–1911. https://aclweb.org/anthology/D16-1195.

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*. Association for Computational Linguistics, Montréal, Canada, 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 Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Atlanta, Georgia, pages 22–31. http://www.aclweb.org/anthology/W13-1703.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. Hoo 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, Montréal, Canada, pages 54–62. http://www.aclweb.org/anthology/W12-2006.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The hoo 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*. Association for Computational Linguistics, Nancy, France, pages 242–249. http://www.aclweb.org/anthology/W11-2838.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A report on the automatic evaluation of scientific writing shared task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, San Diego, CA, pages 53–62. http://www.aclweb.org/anthology/W16-0506.

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*. Association for Computational Linguistics, Denver, CO.

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 Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, Baltimore, Maryland, pages 15–24. http://www.aclweb.org/anthology/W14-1702.

Dan Flickinger, Michael Goodman, and Woodley Packard. 2016. Uw-stanford system description for aesw 2016 shared task on grammatical error detection. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, San Diego, CA, pages 105–111. http://www.aclweb.org/anthology/W16-0511.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Baltimore, Maryland, pages 174–180. http://www.aclweb.org/anthology/P14-2029.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, pages 57–60.

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 Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, Baltimore, Maryland, pages 25–33. http://www.aclweb.org/anthology/W14-1703.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1546–1556. https://aclweb.org/anthology/D16-1161.

C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners, Second Edition*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. https://books.google.com/books?id=bi0QAwAAQBAJ.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics* 19(2):313–330.

Ryo Nagata and Kazuhide Nakatani. 2010. Evaluating performance of grammatical error detection to maximize learning effect. In *Coling 2010: Posters*. Coling 2010 Organizing Committee, Beijing, China, pages 894–900. http://www.aclweb.org/anthology/C10-2103.

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*. Association for Computational Linguistics, Beijing, China, pages 588–593. http://www.aclweb.org/anthology/P15-2097.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 229–234. http://www.aclweb.org/anthology/E17-2037.

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: Shared Task*. Association for Computational Linguistics, Baltimore, Maryland, pages 1–14.

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*. Association for Computational Linguistics, Sofia, Bulgaria, pages 1–12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *SANCL*.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1181–1191. http://www.aclweb.org/anthology/P16-1112.

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 Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, Baltimore, Maryland, pages 34–42. http://www.aclweb.org/anthology/W14-1704.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics* 4:169–182.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jeju Island, Korea, pages 198–202. http://www.aclweb.org/anthology/P12-2039.

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*. Association for Computational Linguistics, Portland, Oregon, USA, pages 180–189. http://www.aclweb.org/anthology/P11-1019.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 380–386. http://www.aclweb.org/anthology/N16-1042.