Faculty of Mathematics and Computer Science
Adam Mickiewicz University in Poznań

Algorithms for automatic verification of grammatical correctness

Roman Grundkiewicz

Supervisor:
dr hab. Krzysztof Jassem
Co-supervisor:
dr Marcin Junczys-Dowmunt

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

December 2016

Abstract

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Declaration

I, AUTHOR NAME, declare that this thesis titled, 'THESIS TITLE' and the work presented in it are my own. I confirm that...

Signed: _____

Contents

Abstract i

Acknowledgements ii

Declaration iii

List of Tables vi

List of Figures vii

Abbreviations viii

1 Introduction 1
- 1.1 Motivations . . . 1
- 1.2 Goals and hypotheses . . . 2
- 1.3 The contributions of the thesis . . . 2
- 1.4 Thesis outline . . . 3
- 1.5 Publication notes . . . 3

2 Grammatical error correction 5
- 2.1 Grammatical errors . . . 5
- 2.2 Automatic error correction . . . 5
- 2.3 Difficulties in grammatical error correction . . . 5
- 2.4 Approaches to automatic grammatical error correction . . . 5
  - 2.4.1 Rule-based methods . . . 5
  - 2.4.2 Language modeling . . . 5
  - 2.4.3 Classification-based approach . . . 5
  - 2.4.4 Statistical machine translation . . . 5
  - 2.4.5 Combined approaches . . . 5
- 2.5 Correction methods for specific error types . . . 5
- 2.6 Shared tasks . . . 5
- 2.7 Summary . . . 5

3 Error corpora and monolingual data 6
- 3.1 Error corpora . . . 6
  - 3.1.1 Learner's corpora . . . 6

3.1.2 Artificial errors . . . 7
3.1.3 Text revision histories . . . 8
3.1.4 Social networks for language learners . . . 9
3.2 Monolingual data . . . 10
3.3 Creating WikEd Error Corpus . . . 11
3.3.1 Extracting edits from Wikipedia . . . 11
3.3.2 Collecting corrective edits . . . 12
3.3.3 Edition filtering . . . 13
3.3.4 Corpus format . . . 13
3.4 Corpus adaptation . . . 14
3.4.1 Error rates . . . 14
3.4.2 Error patterns . . . 15
3.4.3 Corpus adaptation using pattern selection . . . 15
3.5 Summary . . . 16

4 Evaluation metrics **18**
4.1 Difficulties in evaluation of GEC systems . . . 18
4.2 Evaluation metrics . . . 18
4.2.1 Standard metrics . . . 18
4.2.2 MaxMatch . . . 18
4.2.3 I-measure . . . 18
4.2.4 MT metrics . . . 18
4.3 Human evaluation of GEC systems . . . 18
4.4 Conclusions . . . 18

5 Grammatical error correction by statistical machine translation **19**

6 Classifier-based grammatical error correction **20**

7 Combination of SMT- and classifier-based approaches **21**

8 Conclusions **22**
8.1 Realization of thesis hypotheses . . . 22
8.2 Impact of thesis results . . . 22
8.3 Future work . . . 22

A Error types in the NUCLE corpora **23**

Bibliography **24**

List of Tables

3.1 Statistics of the NUCLE error corpora 7
3.2 Statistics of the Lang-8 Corpus. 9
3.3 Statistics of monolingual training data. 10
3.4 The most frequent edits in the WikEd corpus. 13
3.5 Comparison of error rates in parallel corpora. 15
3.6 The most frequent patterns extracted from the NUCLE corpus. 16
3.7 The comparison of the adapted WikEd 0.9 and Lang-8 NAIST corpora. 16

List of Figures

Abbreviations

AEE	Automatic Essay Evaluation
BLEU	BiLingual Evaluation Understudy
CALL	Computer-Assisted Language Learning
CSS	Context-Sensitive Spelling
EFL	English as a Foreign Language
ELL	English Language Learner
ESL	English as a Second Language
GEC	Grammatical Error Correction
L1	the first language of the writer
L2	the second language of the writer
LM	Language Model
M^2	MaxMatch metric
ML	Machine Learning
MT	Machine Translation
NMT	Neural Machine Translation
NLP	Natural Language Processing
NP	Noun Phrase
SMT	Statistical Machine Translation
SVA	Subject-Verb Agreement
POS	Part Of Speech
PBSMT	Phrase-Based Statistical Machine Translation
SER	Sentence Error Rate
WC	automatic Word Class
WER	Word Error Rate

Chapter 1

Introduction

This thesis considers the problem of automated Grammatical Error Correction (GEC) in texts written by learners of English. The main goal of our research was to develop efficient algorithms and methods that can verify and improve grammatical correctness of input text in a fully automatic manner.

1.1 Motivations

In recent years, automated grammatical error correction has grown in popularity as a part of Natural Language Processing (NLP) field of research. One reason for this is a number of possible practical applications. Simple spelling and grammar checking components exist in a number of applications, such as text processors, email clients and web browsers, facilitating writing of error-free texts. More sophisticated solutions are incorporated into systems for comprehensive proofreading and Computer-Assisted Language Learning (CALL). Moreover, fully automated text correction is used as a part of pre- or postprocessing in various NLP tasks, for example in optical character recognition, automatic speech recognition, and Machine Translation (MT). Despite the popularity, the automated grammatical error correction as a very difficult task is still far from being solved completely (Bryant and Ng, 2015).

In this thesis, we deal with errors made by English as a Second Language (ESL) learners. English is one of the most commonly-used languages; it is the third language on the world with the largest number of native speakers, and the first one according to the number of non-native speakers (Lewis et al., 2015). Also, most of the work in NLP field on automatic error correction has studied errors produced by second language (L2) learners. There are significantly more tools and resources, such as Part-Of-Speech (POS) taggers and parsers, developed for English than for other languages. Those allow researchers to produce more comparable results and to develop more advanced algorithms.

One of the objectives of this research is to maintain a simplicity and clarity of created models and to make them language-independent to the extent possible. Recently, the state-of-the-art GEC systems to be able to detect and correct the largest possible number of errors usually combine various algorithms and approaches developed separately for specific error types (Rozovskaya and Roth, 2014). This results in a high degree of complexity and makes the results more difficult to reproduce. We believe, in accordance with Occam's Razor, that between two models that solve the same problem on similar levels of quality the simpler one is to be preferred. Thus, we avoid creating heuristic rules and using advanced NLP tools.

For the above reasons, we have chosen a Statistical Machine Translation (SMT) approach to create a baseline system. Factorized phrasal-based SMT model allows incorporating new algorithms, or even other approaches, in a clear and straightforward way at various levels of processing. Furthermore, the approach where grammatical error correction is considered as some kind of machine translation, in this case from erroneous English to correct English, was turned out to be underresearched.

Our motivation was to research the aforementioned areas in the field of grammatical error correction.

1.2 Goals and hypotheses

The main goal of this thesis is to develop new efficient algorithms for automatic detection and correction of grammatical and usage errors produced by English learners in texts written in natural language.

The main hypothesis of the thesis can be formulated as:

The combination of supervised classification models and phrase-based statistical machine translation models that is applied to automated grammatical error correction will improve the correction performance that can be achieved from the two models separately.

To verify the above hypothesis the following tasks have been defined:

1. Collecting examples of naturally-occurring grammatical and usage errors needed as training data for statistical data-driven approaches.
2. Choosing the most adequate automated evaluation metric that correlate with human judgements.
3. Building an efficient automated grammatical error correction system based on phrase-based SMT approach.
4. Developing a general classification model for various grammatical error types that can be further incorporated into phrase-based SMT model.
5. Incorporating the classification algorithms into phrase-based SMT model for automated grammatical error correction.

All these tasks are completed and described in the thesis.

1.3 The contributions of the thesis

This thesis makes a number of contributions to the field of automated grammatical error correction at both theoretical and engineering levels.

First, we provide a review of the historical and current state-of-the-art methods in the field. The review is presented from the point of view of different approaches to automated error detection and correction. This part of the thesis may serve useful information for someone new in the area.

Second, we develop a method for automatic extraction of potential errors from text edition histories. Both, the WikEd Error Corpus and tools developed to build it, are made publicly available. We also present, how a noisy corpus of corrective edits containing various types of edits can be adapted to a specific topic via error pattern selection.

The thesis also present the first large-scale human evaluation of automated grammatical error correction systems. We use the produced system rankings to evaluate standard metrics for GEC in terms of correlation with human judgment and then we show that the currently commonly used metric M^2 is generally good correlated. The developed tools and collected data are made publicly available and should be useful for other researchers.

Next, we reinvestigate the classifier-based approach to automated grammatical error correction. We show that the state-of-the-art results for article and preposition error correction can be achieved with significant simplification of features. This is also the first attempt when automatic word classes are applied to the task as a possible substitution of part-of-speech tags. We also propose a simple but effective method for detecting places in a sentence where words could have been omitted and an algorithm for generation of confusion sets from error corpus.

Finally, we evaluate a number of methods that combine the classification and SMT-based GEC systems, including pipelining and word lattices. We also incorporate the discriminative classifier into the generative SMT model as a new feature function.

1.4 Thesis outline

The structure of the thesis is as follows. In Chapter 2 we introduce the automated grammatical error correction as a field of natural language processing. The scope of the work is established and the state-of-the-art approaches are described. Chapter **??** is a description of theoretical and mathematical background of various supervised machine learning techniques that are used through the rest of the thesis.

Next, in Chapter 3, we introduce used data sets, including monolingual and error corpora. In this chapter, we also present a method for extracting naturally-occurring error examples from text revision histories. The choice of evaluation metrics and the results of a large-scale human evaluation of GEC systems are described in Chapter 4.

In Chapter 5, I describe the baseline GEC system based on generative phrase-based statistical machine translation models. Chapter 6, in turn, describes experiments on the application of discriminative classification algorithms to the correction of various error types. The combination of classification algorithms and phrase-based SMT models is studied in Chapter 7. Both are described: the combination of results produced individually by investigated approaches and more advanced incorporation of one approach into another via a feature function.

I conclude in Chapter 8 by summarizing the key achievements and results, and establish outstanding plans for future work.

1.5 Publication notes

Some parts of this dissertation based on the following publications:

- Grundkiewicz, R. (2013a). Automatic extraction of Polish language errors from text edition history. In *Text, Speech, and Dialogue — 16th International Conference, TSD*

2013, volume 8082 of *Lecture Notes in Computer Science*, pages 129–136, Plzen, Czech. Springer Berlin Heidelberg

- Grundkiewicz, R. (2013b). Errano: a tool for semi-automatic annotation of language errors. In *Proceedings of the 6th Language & Technology Conference*, pages 309–313, Poznan, Poland

- Junczys-Dowmunt, M. and Grundkiewicz, R. (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*, pages 25–33, Baltimore, Maryland. Association for Computational Linguistics

- Grundkiewicz, R. and Junczys-Dowmunt, M. (2014). The WikEd error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In Przepiórkowski, A. and Ogrodniczuk, M., editors, *Advances in Natural Language Processing — Lecture Notes in Computer Science*, volume 8686, pages 478–490. Springer

- Grundkiewicz, R., Junczys-Dowmunt, M., and Gillian, E. (2015). Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics

- Grundkiewicz, R. and Junczys-Dowmunt, M. (2015). Grammatical error correction with (almost) no linguistic knowledge. In *Proceedings of the 7th Language & Technology Conference*, pages 240–245, Poznan, Poland

- Junczys-Dowmunt, M. and Grundkiewicz, R. (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*, pages 1546–1556, Austin, Texas. Association for Computational Linguistics

Chapter 2

Grammatical error correction

2.1 Grammatical errors

2.2 Automatic error correction

2.3 Difficulties in grammatical error correction

2.4 Approaches to automatic grammatical error correction

2.4.1 Rule-based methods

2.4.2 Language modeling

2.4.3 Classification-based approach

2.4.4 Statistical machine translation

2.4.5 Combined approaches

2.5 Correction methods for specific error types

2.6 Shared tasks

2.7 Summary

Chapter 3

Error corpora and monolingual data

Although some types of errors, for instance subject-verb mistakes, can be corrected using heuristic rules, others, like article or preposition errors, are difficult to correct without substantial amounts of corpus-based information (Leacock et al., 2010). This is especially true for data-driven approaches, such as supervised classification (Cahill et al., 2013) and statistical machine translation (Mizumoto et al., 2012). Parallel data are also essential for the correction training paradigm (Rozovskaya and Roth, 2014).

In this chapter, we discuss the types and sources of data used to build grammatical error correction systems. We review selected error corpora in Section 3.1, and monolingual data in Section 3.2. Section 3.3 describes a language-independent method of edition mining from Wikipedia's revision histories which lead to the building of WikEd Error Corpus (Grundkiewicz and Junczys-Dowmunt, 2014). Finally, in Section 3.4 we compare error corpora used in the experiments presented in the next chapters, and introduce a corpus adaptation method.

3.1 Error corpora

Three main approaches to gathering error corpora are present in literature: manual annotation of learners' writings, artificial errors generation within well-formed sentences, and the extraction of errors and their corrections from text edit histories. A fourth possibility are social networks for language learners.

3.1.1 Learner's corpora

Compared to multilingual translation corpora which today are plentiful or can be easily collected[1], genuine error corpora are not easy to come by. As noted by Leacock et al. (2010), even if large quantities of students' writings are produced and corrected every day, only a small number of them is archived in electronic form.

Most of the available error-annotated corpora has been created from ESL learners' writings. The most popular publicly-available data set nowadays, served as a standard resource for empirical approaches to grammatical error correction, is **the NUS Corpus of Learner English** (Dahlmeier et al., 2013) (NUCLE). The corpus was used as training data in two CoNLL GEC shared tasks in 2013 and 2014 (Ng et al., 2014, 2013) and in a number of succeeding researches, for example: Felice and Yuan (2014), Grundkiewicz and Junczys-Dowmunt (2014), Mizumoto

[1] http://www.statmt.org/moses/?n=Moses.LinksToCorpora

Corpus	Sentences	Tokens	Annotators
NUCLE	57.15 K	1.15 M	1
CoNLL-2013 Test Set	1.38 K	29.07 K	1
CoNLL-2014 Test Set	1.31 K	30.11 K	2
GEC-10 Test Set	1.31 K	30.11 K	10

TABLE 3.1: Statistics of the NUCLE error corpora

and Matsumoto (2016), Rozovskaya and Roth (2014), Yuan and Briscoe (2016). NUCLE consists of 1,414 essays written by Singaporean students who are non-native speakers of English. The essays cover topics, such as environmental pollution, health care, etc. The corpus includes 57,151 sentences in total. Grammatical errors in these sentences have been manually corrected by professional English teachers and annotated with one of the 27 predefined error type. The error types are presented in Appendix A.

Another 50 essays, collected and annotated similarly as NUCLE, were used in both CoNLL GEC shared tasks (Ng et al., 2014, 2013) as blind test data. The CoNLL-2013 test set has been annotated by one annotator, the CoNLL-2014 by two human annotators. The former data set contains 1,381 sentences, the latter 1,312. Bryant and Ng (2015) extended the CoNLL-2014 test set with additional annotations from two to ten annotators. Statistics of the NUCLE corpora are presented in Table 3.1. Tokens are provided for the source (erroneous) side.

Other publicly-available ESL learner corpora are: the dataset of CLC FCE scripts (Yannakoudakis et al., 2011)[2] (FCE-CLC) extracted from the proprietary Cambridge Learner Corpus (Nicholls, 2003) (CLC), and the International Corpus Network of Asian Learners of English[3] (ICNALE). They are usually small and do not provide referenced test sets. NUCLE is a notable exception, but for machine learning approaches even the ca. 50,000 sentences are a rather small resource.

There is much more publicly available unannotated ESL corpora[4], however, they have limited application to data-driven GEC approaches.

3.1.2 Artificial errors

One proposed solution to overcome data sparseness is the creation of artificial data. In the case of artificial error corpora, grammatical errors are introduced by random substitutions, insertions, or deletions. New errors can be also generated according to the frequency distribution observed in seed corpora.

Izumi et al. (2003) generate artificial errors assuming uniform distribution to target article mistakes made by Japanese ESL learners. Brockett et al. (2006) introduce 14 mass/count noun errors that pose problems to Chinese ESL learners with hand-constructed rules which have been used to train statistical machine translation system. Wagner et al. (2007) produce ungrammatical sentences with four types of errors: context-sensitive spelling errors, agreement errors, errors involving a missing word and errors involving an extra word. Error generation was based on an error analysis carried out on a corpus formed by roughly 1,000 error-annotated sentences. Lee and Seneff (2008) created an artificial corpus of verb form errors. A tool for the

[2] http://ilexir.co.uk/applications/clc-fce-dataset/
[3] http://language.sakura.ne.jp/icnale/
[4] https://www.uclouvain.be/en-cecl-lcworld.html

production of artificial errors that imitate genuine errors from two data sets: a grammatical corpus and a list of naturally-occuring errors, has been introduced by Foster and Andersen (2009), and used to generate the corpus for the task of grammatical error detection. Yuan and Felice (2013) extracted lexical and part-of-speech patterns for five types of errors from NUCLE and applied them to well-formed sentences. The work of Felice and Yuan (2014) extends this experiments by using more linguistic information to derive generation probabilities and create artificial data sets.

Artificial errors can be useful not only for building artificial error corpora, but also for increasing error rate in training data sets to help data-driven methods to spot less frequent errors (Cahill et al., 2013, Rozovskaya and Roth, 2010).

Researchers reported different, often contrary, influence of artificial error corpora into a GEC system performance. Artificial errors can increase the recall at the cost of precision (Rozovskaya and Roth, 2010) or vice versa (Felice and Yuan, 2014).

Admittedly, artificial error generation is an efficient and economic way to increase the size of training datasets, but there are drawbacks. The diversification of errors in such corpora can be lower due to small set of real seed data. For specific error types, especially for open-class errors, it may be difficult to create descriptive patterns that can be applied to well-formed sentences (Felice and Yuan, 2014). It is easier to replicate errors within confusion sets, such as articles or prepositions (Rozovskaya and Roth, 2010). Also, errors involving a redundant words usually require specific methods, for example, detecting spaces preceding noun phrases as potential places where an article or determiner could be incorrectly used. Furthermore, it has been reported that artificial data can be less suited for evaluation purposes (Zesch, 2012).

3.1.3 Text revision histories

An alternative solution for gathering error corpora consists in the extraction of errors from text revision histories. The most frequently used are Wikipedia revisions.

Miłkowski (2008) proposes the construction of error corpora from text revision histories based on the hypothesis that the majority of frequent minor edits are the corrections of spelling, grammar, style and usage mistakes. This hypothesis, although very accurate, does not yield the expected result in the form of a wide range of error types, e.g. inflectional errors, because they are rarely repeated. Grundkiewicz (2013a) built a Polish corpus consists of errors automatically extracted from Wikipedia revisions without this restriction. To distinguish error corrections from unwanted edits and to determine error categories, hand-written rules were used.

Wikipedia revisions have been used for the creation of naturally-occurring corrections and sentence paraphrase corpora (Max and Wisniewski, 2010), evaluation of statistical and knowledge-based measures of contextual fitness on the task of real-word spell checking (Zesch, 2012) and correction of preposition errors (Cahill et al., 2013).

Cahill et al. (2013) confirm that data from Wikipedia is useful for both, training a grammatical error correction system and creating artificial data. They shows that models trained with Wikipedia data perform well across diversify test sets representing variety of preposition error distribution. This research focuses only on prepositions. In Section 3.3 we present a method for extracting any error type and we perform experiments on a much larger scale in Chapter 5.

The main advantage of Wikipedia-extracted data sets is their size, but there are also disadvantages, for instance Wikipedia's encyclopedic style and an abundance of vandalism.

Corpus	Sentences	Tokens
Lang-8 NAIST	2,567,969	28,506,540
Lang-8 WEB	3,733,116	51,070,389
Mizumoto et al. (2011)	391,699	n/a
Yoshimoto et al. (2013)	1,217,124	n/a
Susanto et al. (2014)	1,114,139	12,945,666
Mizumoto and Matsumoto (2016)	1,069,127	n/a
Yuan and Briscoe (2016)	n/a	28,823,615
Rozovskaya and Roth (2016)	n/a	ca. 48,000,000

TABLE 3.2: Statistics of the Lang-8 Corpus.

3.1.4 Social networks for language learners

Probably the best resource for language errors has made a recent appearance in the form of social networks for language learners, an example being *Lang-8*[5]. Learners with different native languages correct each others texts based on their own native-language skills.

Mizumoto et al. (2011) published a list of learners' corpora[6] that were scraped from the social language learning site Lang-8 called **the Lang-8 Learner Corpora**. Pairs of learner sentence and corrected sentence has been extracted from Lang-8 data making the use of HTML tags that represents user's annotations. Each sentence may be corrected by more than one corrector, have more than one correction and may contain additional inline comments, usually added in a less structured manner. Version 1.0 of the Lang-8 corpus is free for academic purposes.

We collected all entries from "Lang-8 Learner Corpora v1.0"[7] with English as the learned language, we do not care about the native language of the user. Only entries for which at least one sentence has been corrected were taken into account. Sentences without corrections from such entries were treated as error-free and mirrored on the target side of the corpus. We did not remove alternative versions of the correction for a single sentence. Eventually, we obtain a corpus of 2,567,969 sentence pairs with 28,506,540 tokens on the uncorrected source side (see Table 3.2). We call this resource *Lang-8 NAIST*.

In order to further investigate the effect of adding even greater parallel resources, we scraped Lang-8 for additional entries[8]. We manage to nearly double the size of the corpus to 3,733,116 sentences with 51,259,679 tokens on the source side. In Table 3.2, this joint resource is labeled as *Lang-8 WEB*.

The Lang-8 corpus has been used by other researchers to build general GEC systems for English, for example: Mizumoto et al. (2011), Mizumoto and Matsumoto (2016), Rozovskaya and Roth (2016), Susanto et al. (2014), Yoshimoto et al. (2013), Yuan and Briscoe (2016), as well as systems tackling specific error types (Cahill et al., 2013, Sawai et al., 2013, Tajiri et al., 2012). The comparison of various versions of Lang-8 corpus is presented in Table 3.2. Different sizes come from different techniques for filtering out noisy sentences. Besides, Mizumoto et al. (2011) made the use of sentences written only by Japanese ESL learners. Rozovskaya and Roth (2016) used our Lang-8 WEB corpus.

[5] http://lang-8.com
[6] http://cl.naist.jp/nldata/lang-8
[7] The corpus has been released in December, 2012.
[8] Additional data were scrapped in March, 2014.

Corpus	Sentences	Tokens
Wikipedia	213.08 M	3.37 G
Common Crawl	59.13 G	975.63 G

Table 3.3: Statistics of monolingual training data.

Compared to learner error corpora, data on social network services are not annotated in a well-organized manner and thus may contain unrestricted comments or annotations. This, and the fact that the data are automatically processed, causes that the data contains some noise, which has to be taken into account. However, the cost of corpus creation is much cheaper.

3.2 Monolingual data

Monolingual data are more widely available and much larger in size than annotated data. In GEC, monolingual corpora are usually used as a source of error-free texts (even that this assumption is usually not true) providing statistics about the correct usage of the language. Common examples of application are language modeling (Hdez. and Calvo, 2014, Junczys-Dowmunt and Grundkiewicz, 2014, Yi et al., 2008) or calculation of word probability distribution for generative models (Rozovskaya and Roth, 2014).

One of the most popular source of English data nowadays is **English Wikipedia**[9]. The content of Wikipedia is moderated, mostly by native-speakers, and the article quality is constantly improved, which allow to percieve Wikipedia as error-free. On the other hand, the encyclopaedic style differ significantly from the general style of ESL writings.

We extracted all text from the English Wikipedia[10]. The total number of ca. 213.08 million of sentences is collected. The statistics are presented in Table 3.3.

The second widely used monolingual dataset is the **Common Crawl** data[11] being a publicly available crawl of the web. The main advantage of this data is its size and also the diversity of topics. Crawled texts can be diverse quality, but it has been shown that web-scale language models can improve the performance of a number of NLP tasks, such as machine translation (Buck et al., 2014).

We use data made-available by Buck et al. (2014). We filter the English resources with cross-entropy filtering (Moore and Lewis, 2010) using the corrected NUCLE corpus as seed data. This resulted in roughly 300GB of compressed text consists of ca. 59.13 billion of sentences.

Other large monolingual corpora used for grammatical error correction systems are English Gigaword (Parker et al., 2011), Google Web 1T 5-gram Corpus (Brants and Franz, 2006), or the British National Corpus (BNC)[12].

[9] https://en.wikipedia.org/
[10] Wikipedia database dump from December 2nd, 2013: http://dumps.wikimedia.org/enwiki/20131202/
[11] http://commoncrawl.org/
[12] http://www.natcorp.ox.ac.uk/corpus/

3.3 Creating WikEd Error Corpus

In this section, we introduce the process of building the WikEd Error Corpus: the largest corpus of corrective edits available for the English language[13].

In contrast to other works that use Wikipedia to build various NLP resources (Cahill et al., 2013, Max and Wisniewski, 2010, Zesch, 2012), we processed the entire English Wikipedia revision history[14] and gathered ca. 56 million sentences with annotated edits. The corpus is also not limited to certain types of errors (e.g. real-word spelling errors or preposition errors) and does not exclude specific corrections, such as repetitions or omissions of words, and corrections that refers only to punctuation or case modification as in the case of the work of Max and Wisniewski's (2010). Possible application of the created corpus include, but are not limited to, sentence paraphrasing, spelling correction, grammar correction, etc.

This corpus consists of edited sentences extracted from Wikipedia revisions, and as such inherits the user-friendly CC BY-SA 3.0 license of the original resource. Both, the WikEd Error Corpus and the tools used to produce it have been made available for unrestricted download[15].

3.3.1 Extracting edits from Wikipedia

Wikipedia dumps with complete edit histories are provided in XML format[16]. Similarly to Max and Wisniewski (2010), we iterate over each two adjacent revisions of every Wikipedia page, including articles, user pages, discussions, and help pages. To minimize the number of unwanted vandalism, we skip revisions and preceding revisions if comments contain suggestions of reversions, e.g. *reverting after (. . .)*, *remove vandalism*, *undo vandal's edits*, *delete stupid joke*, etc. This is done by hand-written rules involving regular expressions.

Next, we remove XML and Wikitext Markup Language annotations[17] from each article version and split texts into sentences with the NLTK toolkit[18]. Pairs of edited sentences are identified with the Longest Common Subsequence algorithm (LCS) (Maier, 1978) that is applied at token level. Editions consisting of additions or deletions of full paragraphs are disregarded.

Two edited sentences (we will refer to two corresponding edited fragments as sentences, even if they are not well-formed) s_i and s_j are collected if they meet several surface conditions:

- the sentence length is between 2 and 120 tokens,
- the length difference is less than 5 tokens,
- the relative token-based edit distance $\mathrm{ed}(s_i, s_j)$ with respect to the shorter sentence is smaller than 0.3.

The threshold values in the above restrictions were chosen experientially. The relative token-based edit distance is defined as:

$$\mathrm{ed}(s_i, s_j) = \frac{\mathrm{dist}(s_i, s_j)\,\min(|s_i|, |s_j|)}{\log_b \min(|s_i|, |s_j|)},$$

[13] This section presents an extended work initially published in the following papers: Grundkiewicz (2013a) and Grundkiewicz and Junczys-Dowmunt (2014).

[14] Wikipedia database dump from January 2nd, 2014: http://dumps.wikimedia.org/enwiki/20140102/

[15] http://romang.home.amu.edu.pl/wiked/wiked.html

[16] http://dumps.wikimedia.org/

[17] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[18] http://nltk.org/

where $\text{dist}(s_i, s_j)$ is the token-based Levenshtein edit distance (Levenshtein, 1966), $|s|$ is the length of the sentence s in tokens, and the logarithm base b is empirically set to 20. This formula implies that the longer the sentence is, the more edits are allowed, but it prevents the acceptance of too many edits for long sentences.

3.3.2 Collecting corrective edits

At this stage, 12,130,508 pairs of edited sentences from the English version of Wikipedia have been collected. The most useful edits include:

- spelling error corrections:
  You can use rsync to donload (⇒ download) the database .,
- grammatical error corrections:
  There is (⇒ are) also a (⇒) two computer games based on the movie .,
- stylistic changes:
  Predictably , the (⇒ The) game ended predictably (⇒) when she crashed her Escalade. . . ,
- sentence rewordings and paraphrases:
  These anarchists argue against (⇒ oppose the) regulation of corporations .,
- encyclopaedic style adjustments:
  A local education authority (⇒ Local Education Authority) (LEA) is the part of a council in England or Wales.

The WikEd corpus contains also less useful edits for grammatical error correction task, e.g.:

- time reference changes:
  The Kiwi Party is (⇒ was) a New Zealand political party formed in 2007 .,
- information supplements:
  Aphrodite is the Greek goddess of love (⇒ , sex) and beauty .,
- numeric information updates:
  In May 2003 (⇒ August 2004) this percentage increased to 62 (⇒ 67)% .,
- item additions/deletions to/from bulleted lists:
  Famous Bronxites include (⇒ Regis Philbin ,) Carl Reiner , Danny Aiello. . . ,
- amendments of broken MediaWiki's markups:
  The bipyramids are the [[dual polyhedron — (⇒) dual polyhedra [[(⇒) of the prisms .,
- changes made by vandals:
  David Zuckerman is a writer and producer (⇒ poopface) for television shows.

The total number of edits is 16,013,830 among which 3,273,862 (20.44%) are deletions and 4,829,019 (30.16%) insertions. The most frequently occurring edits are presented in Table 3.4. The `sub(·,·)` stands for word(s) substitutions, `del(·)` for deletions, and `ins(·)` for insertions.

Edits	Freq.	Edits	Freq.	Edits	Freq.
`ins(")`	667,098	`ins(a)`	45,870	`ins(and)`	28,518
`ins(,)`	348,341	`ins(')`	41,473	`del(of)`	26,257
`del(")`	226,854	`del(.)`	41,161	`sub(a,an)`	24,626
`del(,)`	158,324	`sub(is,was)`	40,062	`ins(was)`	23,670
`ins(.)`	138,322	`sub(',")`	37,236	`del(])`	22,443
`del('s)`	80,669	`del(')`	36,051	`sub(was,is)`	21,372
`ins(the)`	79,708	`del())`	34,401	`ins(()`	20,079
`del(the)`	61,999	`del(persons)`	33,773	`del(a)`	19,615
`ins())`	60,852	`ins(The)`	32,819	`ins(in)`	18,651
`ins(< br >)`	51,802	`sub(it,its)`	31,171	`ins(is)`	18,647

TABLE 3.4: The most frequent edits in the WikEd corpus.

3.3.3 Edition filtering

Sentences with potentially unwanted edits, e.g. updates of bulleted list, amendments of MediaWiki markup, and vandalism can be effectively filtered out using simple heuristic rules. For example, all pairs of sentences s_i and s_j that satisfy the following conditions can be disregarded:

- Either the sentence s_i or s_j contains a vulgar word (determined by the list of vulgarisms) or a very long sequence of character with no spaces (e.g. produced by random keystrokes).
- Any of the sentences s_i or s_j contains fragments of XML or Wiki markup, e.g. `<ref>`, `<br>` or `[http:`.
- Edits concern changes in dates or numerical values only.
- The only edit made consists of removing a full stop or semicolon at the end of the sentence s_i.
- The ratio of non-words tokens in s_j to word tokens is higher than a given threshold (we used 0.5).

In the end, 1,775,880 (14.63%) pairs of sentences are marked as potentially harmful, but not removed from the WikEd Corpus. For instance, vandalized entries may be useful for various tasks by themselves.

3.3.4 Corpus format

It was our intention to release the WikEd Error Corpus in a machine-friendly format. We chose a representation based on GNU `wdiff` output[19] extended by comments including meta-data. For example, for a sentence *This page lists links about ancient philosophy.* with the following two edits: insertion of *some* at third position and substitution of *about* with *to*, the WikEd entry corresponds to:

```
This page lists {+some+} links [-about-] {+to+} ancient philosophy.
```

Meta-data consists of:

[19] https://www.gnu.org/software/wdiff/manual/wdiff.html#wdiff

- the revision id, accompanying comment, and revision timestamp,
- the title and id of the edited Wikipedia page,
- the name of the contributor or IP address if it is an anonymous edition.

All sentences preserve the chronological order of the original revisions.

3.4 Corpus adaptation

As showed by other researchers, it is relatively simple to extract specific error types tailored for a particular task, for example the real-word error correction (Max and Wisniewski, 2010) or prepositional error correction (Cahill et al., 2013). We will show that it is also possible to adapt the entire WikEd Error Corpus using the existing ESL corpus.

In this section we discuss the differences between data sets introduced earlier in respect to the measures based on edit frequencies. Next, we present a method for corpus adaptation using error pattern selection and apply it to filter edits from WikEd corpus.

3.4.1 Error rates

Learner error corpora differ each other due to several reasons: essays are written by learners with different L1 and language proficiency; texts cover various topics and style; the annotation schemas among scientific centres building error corpora may be different, or even annotators' preferences may influence for the same data sets (Bryant and Ng, 2015).

Word Error Rate (WER) is a common metric of the performance of speech recognition or machine translation systems. Even that it is not used as an evaluation metric in GEC (reasons for that will be discussed in Chapter 4), it can be used to characterize error corpora. WER is defined as:

$$WER = \frac{S + D + I}{N},$$

where S is the number of words substituted by an annotator, D is the number of deleted words, and I is the number of inserted words. N is the total number of words in the corrected sentence.

Sentence Error Rate (SER) defines fraction of sentences that contain one or more edits, i.e.:

$$SER = \frac{|\text{number of sentences with one or more edits}|}{M}.$$

M is the total number of sentences in data set.

Error rates, including distribution of substitutions, deletions and insertions is presented in Table 3.5. Results were calculated on tokenized and true cased sentences, which corresponds to preprocessing that will be performed in the subsequent experiments. This affected typographic edits and edits concerned changes in letter case. Results for data sets annotated by two or more annotators have been averaged.

The NUCLE corpus contains less edits that test sets from both CoNLL shared tasks, what results in lower WER and SER. Lang-8 corpus has higher error frequency than NUCLE. High SER in WikEd comes from the fact that we extracted only sentences containing one or more edits. However, the lower WER in respect to high SER indicates shorter edits which can be

Corpus	Sub.	Del.	Ins.	WER	SER
NUCLE	0.5646	0.2221	0.2134	0.0609	0.3760
CoNLL-2013 Test Set	0.6304	0.2172	0.1525	0.1476	0.8088
CoNLL-2014 Test Set	0.6341	0.1860	0.1799	0.1137	0.7466
GEC-10 Test Set	0.6352	0.1766	0.1882	0.1427	0.8453
Lang-8 NAIST	0.5593	0.1359	0.3048	0.2241	0.5820
Lang-8 WEB	0.5690	0.1421	0.2890	0.2095	0.8356
WikEd	0.6089	0.1645	0.2266	0.0819	1.0000

TABLE 3.5: Comparison of error rates in parallel corpora.

more typical for errors made by native-speakers. More than a half edits in each corpus are substitutions, but Lang-8 corpora contain more insertions, which may be due to the annotator comments enclosed in corrected sentence.

3.4.2 Error patterns

The frequency of edits built on token level highly depend on a word frequency. We propose a simple method of edit generalization using regular expressions, resulting in extraction of **error patterns**. Error patterns can be used to measure the similarity between two error corpora and to adapt a general error corpus into a reference data set.

For each pair of uncorrected and corrected sentences from error corpus, we compute a sequence of deletions and insertions with the LCS algorithm (Maier, 1978) that transform the source sentence into the target sentence. Adjacent deleted words are concatenated to form a phrase deletion (`del(·)`), adjacent inserted words result in a phrase insertion (`ins(·)`). A deleted phrase followed directly by a phrase insertion is interpreted as a phrase substitution (`sub(·,·)`). Substitutions are generalized if they consist of common substrings, again determined by the LCS algorithm, that are equal to or longer than three characters. We encode generalizations by the regular expression `(\w{3,})` and a back-reference, e.g. `\1`.

Patterns can contain multi-word strings, e.g. `sub((\w{3,}) is,\1s are)` models a case of subject-verb agreement. Sometimes, more than one generalization within an edit is possible, e.g. `sub((\w{3,})-(\w{3,}),\1\2)`. Table 3.6 contains some of the most frequent patterns extracted from NUCLE for all 27 error types. The table includes also the most frequent error categories matching the pattern.

3.4.3 Corpus adaptation using pattern selection

As mentioned before, the WikEd Error Corpus is not an ESL error corpus and may contain a very different type of errors from those made by language learners. We try to mitigate this by selecting errors that resemble mistakes from NUCLE, other errors are replaced by their corrections.

First, we extracted error patterns from NUCLE. Frequency threshold is defined at 5, patterns that occur less often are discarded. In the end 666 patterns remain.

Next, we perform the same computation for sentence pairs from WikEd. Edits that result in patterns from our list are not modified and remain in the data, for all other edits, the selected correction is applied to the source sentence. Error types not covered by the patterns thus

Pattern	Count	Categories with count
sub((\w{3,}),\1s)	2864	Nn(2188) SVA(395) Wform(146)
ins(the)	2494	ArtOrDet(2424)
del(the)	1772	ArtOrDet(1696)
sub((\w{3,})s,\1)	1317	Nn(651) SVA(263) Wform(141) Rloc-(92)
ins(,)	971	Mec(733) Srun(196)
ins(a)	679	ArtOrDet(646)
sub((\w{3,}),\1d)	300	Vt(112) Vform(105) Wform(62)
del(,)	266	Mec(175) Rloc-(83)
sub((\w{3,}),\1ed)	252	Vt(138) Vform(75) Wform(29)
ins(an)	246	ArtOrDet(234)
del(of)	222	Prep(202)
sub(is,are)	219	SVA(198)
del(.)	205	Rloc-(135) Mec(60)
sub((\w{3,})d,\1)	202	Vt(109) Wform(46) Vform(28) Rloc-(11)

TABLE 3.6: The most frequent patterns extracted from the NUCLE corpus.

Corpus	Sentences	Tokens	Edits	WER	SER
NUCLE	57,151	1,161,567	42,423	0.0609	0.3760
Lang-8 NAIST	2,567,964	28,506,516	3,408,834	0.2241	0.5820
+Select	2,567,964	34,351,819	1,066,690		0.2815
WikEd 0.9	12,130,508	292,570,716	16,013,830	0.0819	1.0000
+Select	12,130,508	294,965,241	5,327,293		0.3262

TABLE 3.7: The comparison of the adapted WikEd 0.9 and Lang-8 NAIST corpora.

disappear. Noise like vandalism is either removed or reduced to identical sentences on both sides for the training corpus. In both cases this cannot harm our systems. Eventually, 3,957,547 (32,62%) sentences remain that still contain edit pattern. We keep all sentences with surviving errors and randomly select sentences without edits to be kept as well. The final parallel corpus consists of 4,703,353 sentence pairs.

Two versions of WikEd are used in our experiments: the error selected corpus which is labeled *WikEd+Select* and a second version consisting of the same sentences but with all errors present (a proper subset of the unprocessed WikEd), this version is denoted as *WikEd*.

Error selection can be also applied to Lang-8 NAIST. The comparison of the adapted and unadapted WikEd and Lang-8 corpora is presented in Table 3.7.

We will evaluate the corpus adaptation method experimentally in Chapter 5.

3.5 Summary

Both error corpora and monolingual data are crucial to create effective grammatical error correction systems using data-driven approaches. In this chapter we described the types of error corpora used in automated ESL grammatical error correction: annotated ESL learner writings, artificially generated errors, corrections scrapped from the social network sites, and edits extracted from text revisions. We discussed their advantages and disadvantages, and compare

them using word error rate and sentence error rate metrics. The sources of large-scale monolingual data have also be shown.

We introduced the WikEd Error Corpus — a new large resource with possible applications to sentence paraphrasing, spelling correction, and grammar correction — and shown how it can be adapted to the task of grammatical error correction using seed corpus via error pattern selection.

Last but not least, in this chapter we defined the datasets that we will use in our experiments.

Chapter 4

Evaluation metrics

4.1 Difficulties in evaluation of GEC systems

4.2 Evaluation metrics

4.2.1 Standard metrics

4.2.2 MaxMatch

4.2.3 I-measure

4.2.4 MT metrics

4.3 Human evaluation of GEC systems

4.4 Conclusions

Chapter 5

Grammatical error correction by statistical machine translation

Chapter 6

Classifier-based grammatical error correction

Chapter 7

Combination of SMT- and classifier-based approaches

Chapter 8

Conclusions

8.1 Realization of thesis hypotheses

8.2 Impact of thesis results

8.3 Future work

Appendix A

Error types in the NUCLE corpora

Bibliography

Brants, T. and Franz, A. (2006). {Web 1T 5-gram Version 1}.

Brockett, C., Dolan, W. B., and Gamon, M. (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*, pages 249–256, Stroudsburg, USA. Association for Computational Linguistics.

Bryant, C. and Ng, H. T. (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 (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.

Buck, C., Heafield, K., and van Ooyen, B. (2014). N-gram counts and language models from the Common Crawl. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3579–3584, Reykjavík, Iceland.

Cahill, A., Madnani, N., Tetreault, J., and Napolitano, D. (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*, pages 507–517, Atlanta, Georgia. Association for Computational Linguistics.

Dahlmeier, D., Ng, H. T., and Wu, S. M. (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*, pages 22–31.

Felice, M. and Yuan, Z. (2014). Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–126, Gothenburg, Sweden. Association for Computational Linguistics.

Foster, J. and Andersen, Ø. E. (2009). Generrate: generating errors for use in grammatical error detection. In *Proceedings of the fourth workshop on innovative use of nlp for building educational applications*, pages 82–90. Association for Computational Linguistics.

Grundkiewicz, R. (2013a). Automatic extraction of Polish language errors from text edition history. In *Text, Speech, and Dialogue — 16th International Conference, TSD 2013*, volume 8082 of *Lecture Notes in Computer Science*, pages 129–136, Plzen, Czech. Springer Berlin Heidelberg.

Grundkiewicz, R. (2013b). Errano: a tool for semi-automatic annotation of language errors. In *Proceedings of the 6th Language & Technology Conference*, pages 309–313, Poznan, Poland.

Grundkiewicz, R. and Junczys-Dowmunt, M. (2014). The WikEd error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In Przepiórkowski, A. and Ogrodniczuk, M., editors, *Advances in Natural Language Processing — Lecture Notes in Computer Science*, volume 8686, pages 478–490. Springer.

Grundkiewicz, R. and Junczys-Dowmunt, M. (2015). Grammatical error correction with (almost) no linguistic knowledge. In *Proceedings of the 7th Language & Technology Conference*, pages 240–245, Poznan, Poland.

Grundkiewicz, R., Junczys-Dowmunt, M., and Gillian, E. (2015). Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Hdez., S. D. and Calvo, H. (2014). Conll 2014 shared task: Grammatical error correction with a syntactic n-gram language model from a big corpora. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 53–59, Baltimore, Maryland. Association for Computational Linguistics.

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., and Isahara, H. (2003). Automatic error detection in the japanese learners' english spoken data. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 145–148. Association for Computational Linguistics.

Junczys-Dowmunt, M. and Grundkiewicz, R. (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*, pages 25–33, Baltimore, Maryland. Association for Computational Linguistics.

Junczys-Dowmunt, M. and Grundkiewicz, R. (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*, pages 1546–1556, Austin, Texas. Association for Computational Linguistics.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, 3(1):1–134.

Lee, J. and Seneff, S. (2008). Correcting misuse of verb forms. In *Proceedings of ACL-08: HLT*, pages 174–182, Columbus, Ohio. Association for Computational Linguistics.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2015). *Ethnologue: Languages of the world*, volume 18.

Maier, D. (1978). The Complexity of Some Problems on Subsequences and Supersequences. *J. ACM*, 25(2):322–336.

Max, A. and Wisniewski, G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In *Proceedings of LREC*.

Miłkowski, M. (2008). Automated Building of Error Corpora of Polish. In *Corpus Linguistics, Computer Tools, and Applications — State of the Art*, pages 631–639. Peter Lang.

Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M., and Matsumoto, Y. (2012). The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012*, pages 863–872.

Mizumoto, T., Komachi, M., Nagata, M., and Matsumoto, Y. (2011). Mining revision log of language learning SNS for automated japanese error correction of second language learners. In *The 5th International Joint Conference on Natural Language Processing*, pages 147–155.

Mizumoto, T. and Matsumoto, Y. (2016). Discriminative reranking for grammatical error correction with statistical machine translation. In *Proceedings of NAACL-HLT*, pages 1133–1138.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (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, Sofia, Bulgaria. Association for Computational Linguistics.

Nicholls, D. (2003). The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition, linguistic data consortium. Technical report, Technical Report. Linguistic Data Consortium, Philadelphia.

Rozovskaya, A. and Roth, D. (2010). Training paradigms for correcting errors in grammar and usage. In *North American Chapter of the Association for Computational Linguistics*.

Rozovskaya, A. and Roth, D. (2014). Building a state-of-the-art grammatical error correction system.

Rozovskaya, A. and Roth, D. (2016). Grammatical error correction: Machine translation and classifiers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2205–2215, Berlin, Germany. Association for Computational Linguistics.

Sawai, Y., Komachi, M., and Matsumoto, Y. (2013). A learner corpus-based approach to verb suggestion for esl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 708–713, Sofia, Bulgaria. Association for Computational Linguistics.

Susanto, R. H., Phandi, P., and Ng, H. T. (2014). System combination for grammatical error correction. pages 951–962.

Tajiri, T., Komachi, M., and Matsumoto, Y. (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: Short Papers - Volume 2*, ACL '12, pages 198–202, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wagner, J., Foster, J., and van Genabith, J. (2007). A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 112–121. Association for Computational Linguistics.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (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-Volume 1*, pages 180–189. Association for Computational Linguistics.

Yi, X., Gao, J., and Dolan, W. B. (2008). A web-based english proofing system for english as a second language users. In *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pages 619–624.

Yoshimoto, I., Kose, T., Mitsuzawa, K., Sakaguchi, K., Mizumoto, T., Hayashibe, Y., Komachi, M., and Matsumoto, Y. (2013). Naist at 2013 conll grammatical error correction shared task. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 26–33, Sofia, Bulgaria. Association for Computational Linguistics.

Yuan, Z. and Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of NAACL-HLT*, pages 380–386.

Yuan, Z. and Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 52–61, Sofia, Bulgaria. Association for Computational Linguistics.

Zesch, T. (2012). Measuring Contextual Fitness Using Error Contexts Extracted from the Wikipedia Revision History. In *Proceedings of EACL*, pages 529–538.