Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–14,
Baltimore, Maryland, 26-27 July 2014.

present in English essays written by second language learners of English. Each participating team is given training data manually annotated with corrections of grammatical errors. The test data consists of new, blind test essays. Preprocessed test essays, which have been sentence-segmented and tokenized, are also made available to the participating teams. Each team is to submit its system output consisting of the automatically corrected essays, in sentence-segmented and tokenized form.

Grammatical errors consist of many different types, including articles or determiners, prepositions, noun form, verb form, subject-verb agreement, pronouns, word choice, sentence structure, punctuation, capitalization, etc. However, most prior published research on grammatical error correction only focuses on a small number of frequently occurring error types, such as article and preposition errors (Han et al., 2006; Gamon, 2010; Rozovskaya and Roth, 2010; Tetreault et al., 2010; Dahlmeier and Ng, 2011b). Article and preposition errors were also the only error types featured in the HOO 2012 shared task. Likewise, although all error types were included in the HOO 2011 shared task, almost all participating teams dealt with article and preposition errors only (besides spelling and punctuation errors). In the CoNLL-2013 shared task, the error types were extended to include five error types, comprising article or determiner, preposition, noun number, verb form, and subject-verb agreement. Other error types such as word choice errors (Dahlmeier and Ng, 2011a) were not dealt with.

In the CoNLL-2014 shared task, it was felt that the community is now ready to deal with *all* error types. Table 1 shows examples of the 28 error types in the CoNLL-2014 shared task.

Since there are 28 error types in our shared task compared to two in HOO 2012 and five in CoNLL-2013, there is a greater chance of encountering multiple, interacting errors in a sentence in our shared task. This increases the complexity of our shared task. To illustrate, consider the following sentence:

> Social *network plays* a role in providing and also filtering information.

The noun number error *networks* needs to be corrected (network → networks). This necessitates the correction of a subject-verb agreement error (plays → play). A pipeline system in which corrections for subject-verb agreement errors occur strictly before corrections for noun number errors would not be able to arrive at a fully corrected sentence for this example. The ability to correct multiple, interacting errors is thus necessary in our shared task. The recent work of Dahlmeier and Ng (2012a) and Wu and Ng (2013), for example, is designed to deal with multiple, interacting errors.

3 Data

This section describes the training and test data released to each participating team in our shared task.

3.1 Training Data

The training data provided in our shared task is the NUCLE corpus, the NUS Corpus of Learner English (Dahlmeier et al., 2013). As noted by (Leacock et al., 2010), the lack of a manually annotated and corrected corpus of English learner texts has been an impediment to progress in grammatical error correction, since it prevents comparative evaluations on a common benchmark test data set. NUCLE was created precisely to fill this void. It is a collection of 1,414 essays written by students at the National University of Singapore (NUS) who are non-native speakers of English. The essays were written in response to some prompts, and they cover a wide range of topics, such as environmental pollution, health care, etc. The grammatical errors in these essays have been hand-corrected by professional English instructors at NUS. For each grammatical error instance, the start and end character offsets of the erroneous text span are marked, and the error type and the correction string are provided. Manual annotation is carried out using a graphical user interface specifically built for this purpose. The error annotations are saved as stand-off annotations, in SGML format.

To illustrate, consider the following sentence at the start of the sixth paragraph of an essay:

> Nothing is *absolute* right or wrong.

There is a word form error (absolute → absolutely) in this sentence. The error annotation, also called *correction* or *edit*, in SGML format is shown in Figure 1. `start_par` (`end_par`) denotes the paragraph ID of the start (end) of the erroneous

Type	Description	Example
Vt	Verb tense	Medical technology during that time [**is** → was] not advanced enough to cure him.
Vm	Verb modal	Although the problem [**would** → may] not be serious, people [**would** → might] still be afraid.
V0	Missing verb	However, there are also a great number of people [**who** → who are] against this technology.
Vform	Verb form	A study in 2010 [**shown** → showed] that patients recover faster when surrounded by family members.
SVA	Subject-verb agreement	The benefits of disclosing genetic risk information [**outweighs** → outweigh] the costs.
ArtOrDet	Article or determiner	It is obvious to see that [**internet** → the internet] saves people time and also connects people globally.
Nn	Noun number	A carrier may consider not having any [**child** → children] after getting married.
Npos	Noun possessive	Someone should tell the [**carriers** → carrier's] relatives about the genetic problem.
Pform	Pronoun form	A couple should run a few tests to see if [**their** → they] have any genetic diseases beforehand.
Pref	Pronoun reference	It is everyone's duty to ensure that [**he or she** → they] undergo regular health checks.
Prep	Preposition	This essay will [**discuss about** → discuss] whether a carrier should tell his relatives or not.
Wci	Wrong collocation/idiom	Early examination is [**healthy** → advisable] and will cast away unwanted doubts.
Wa	Acronyms	After [**WOWII** → World War II], the population of China decreased rapidly.
Wform	Word form	The sense of [**guilty** → guilt] can be more than expected.
Wtone	Tone (formal/informal)	[**It's** → It is] our family and relatives that bring us up.
Srun	Run-on sentences, comma splices	The issue is highly [**debatable, a** → debatable. A] genetic risk could come from either side of the family.
Smod	Dangling modifiers	[**Undeniable,** → It is undeniable that] it becomes addictive when we spend more time socializing virtually.
Spar	Parallelism	We must pay attention to this information and [**assisting** → assist] those who are at risk.
Sfrag	Sentence fragment	**However, from the ethical point of view.**
Ssub	Subordinate clause	This is an issue [**needs** → that needs] to be addressed.
WOinc	Incorrect word order	[**Someone having what kind of disease** → What kind of disease someone has] is a matter of their own privacy.
WOadv	Incorrect adjective/ adverb order	In conclusion, [**personally I** → I personally] feel that it is important to tell one's family members.
Trans	Linking words/phrases	It is sometimes hard to find [**out** → out if] one has this disease.
Mec	Spelling, punctuation, capitalization, etc.	This knowledge [**maybe relavant** → may be relevant] to them.
Rloc−	Redundancy	It is up to the [**patient's own choice** → patient] to disclose information.
Cit	Citation	Poor citation practice.
Others	Other errors	An error that does not fit into any other category but can still be corrected.
Um	Unclear meaning	Genetic disease has a close relationship with the **born gene.** (i.e., no correction possible without further clarification.)

Table 1: The 28 error types in the shared task.

text span (paragraph ID starts from 0 by convention). `start_off` (`end_off`) denotes the character offset of the start (end) of the erroneous text span (again, character offset starts from 0 by convention). The error tag is `Wform`, and the correction string is `absolutely`.

The NUCLE corpus was first used in (Dahlmeier and Ng, 2011b), and has been publicly available for research purposes since June 2011[1]. All instances of grammatical errors are annotated in NUCLE.

To help participating teams in their preparation for the shared task, we also performed automatic preprocessing of the NUCLE corpus and released the preprocessed form of NUCLE. The preprocessing operations performed on the NUCLE essays include sentence segmentation and word tokenization using the NLTK toolkit (Bird et al., 2009), and part-of-speech (POS) tagging, constituency and dependency tree parsing using the Stanford parser (Klein and Manning, 2003; de Marneffe et al., 2006). The error annotations, which are originally at the character level, are then mapped to error annotations at the word token level. Error annotations at the word token level also facilitate scoring, as we will see in Section 4, since our scorer operates by matching tokens. Note that although we released our own preprocessed version of NUCLE, the participating teams were however free to perform their own preprocessing if they so preferred.

NUCLE release version 3.2 was used in the CoNLL-2014 shared task. In this version, 17 essays were removed from the first release of NUCLE since these essays were duplicates with multiple annotations. In addition, in order to facilitate the detection and correction of article/determiner errors and preposition errors, we performed some automatic mapping of error types in the original NUCLE corpus to arrive at release version 3.2. Ng et al. (2013) gives more details of how the mapping was carried out.

The statistics of the NUCLE corpus (release 3.2 version) are shown in Table 2. The distribution of errors among all error types is shown in Table 3.

While the NUCLE corpus is provided in our shared task, participating teams are free to not use NUCLE, or to use additional resources and tools in building their grammatical error correction systems, as long as these resources and tools are publicly available and not proprietary. For example, participating teams are free to use the Cambridge FCE corpus (Yannakoudakis et al., 2011; Nicholls, 2003) (the training data provided in HOO 2012 (Dale et al., 2012)) as additional training data.

[1]http://www.comp.nus.edu.sg/∼nlp/corpora.html

	Training data (NUCLE)	**Test data**
# essays	1,397	50
# sentences	57,151	1,312
# word tokens	1,161,567	30,144

Table 2: Statistics of training and test data.

3.2 Test Data

Similar to CoNLL-2013, 25 NUS students, who are non-native speakers of English, were recruited to write new essays to be used as blind test data in the shared task. Each student wrote two essays in response to the two prompts shown in Table 4, one essay per prompt. The first prompt was also used in the NUCLE training data, but the second prompt is entirely new and not used previously. As a result, 50 new test essays were collected. The statistics of the test essays are also shown in Table 2.

Error annotation on the test essays was carried out independently by two native speakers of English. One of them is a lecturer at the NUS Centre for English Language Communication, and the other is a freelance English linguist with extensive prior experience in error annotation of English learners' essays. The distribution of errors in the test essays among the error types is shown in Table 3. The test essays were then preprocessed in the same manner as the NUCLE corpus. The preprocessed test essays were released to the participating teams. Similar to CoNLL-2013, the test essays and their error annotations in the CoNLL-2014 shared task will be made freely available after the shared task.

4 Evaluation Metric and Scorer

A grammatical error correction system is evaluated by how well its proposed corrections or edits match the gold-standard edits. An essay is first sentence-segmented and tokenized before evaluation is carried out on the essay. To illustrate, consider the following tokenized sentence S written by an English learner:

```
<MISTAKE start_par="5" start_off="11" end_par="5" end_off="19">
<TYPE>Wform</TYPE>
<CORRECTION>absolutely</CORRECTION>
</MISTAKE>
```

Figure 1: An example error annotation.

Error type	Training data (NUCLE)	%	Test data (Annotator 1)	%	Test data (Annotator 2)	%
Vt	3,204	7.1%	133	5.5%	150	4.5%
Vm	431	1.0%	49	2.0%	37	1.1%
V0	414	0.9%	31	1.3%	37	1.1%
Vform	1,443	3.2%	132	5.5%	91	2.7%
SVA	1,524	3.4%	105	4.4%	154	4.6%
ArtOrDet	6,640	14.8%	332	13.9%	444	13.3%
Nn	3,768	8.4%	215	9.0%	228	6.8%
Npos	239	0.5%	19	0.8%	15	0.5%
Pform	186	0.4%	47	2.0%	18	0.5%
Pref	927	2.1%	96	4.0%	153	4.6%
Prep	2,413	5.4%	211	8.8%	390	11.7%
Wci	5,305	11.8%	340	14.2%	479	14.4%
Wa	50	0.1%	0	0.0%	1	0.0%
Wform	2,161	4.8%	77	3.2%	103	3.1%
Wtone	593	1.3%	9	0.4%	15	0.5%
Srun	873	1.9%	7	0.3%	26	0.8%
Smod	51	0.1%	0	0.0%	5	0.2%
Spar	519	1.2%	3	0.1%	24	0.7%
Sfrag	250	0.6%	13	0.5%	5	0.2%
Ssub	362	0.8%	68	2.8%	10	0.3%
WOinc	698	1.6%	22	0.9%	54	1.6%
WOadv	347	0.8%	12	0.5%	27	0.8%
Trans	1,377	3.1%	94	3.9%	79	2.4%
Mec	3,145	7.0%	231	9.6%	496	14.9%
Rloc−	4,703	10.5%	95	4.0%	199	6.0%
Cit	658	1.5%	0	0.0%	0	0.0%
Others	1,467	3.3%	44	1.8%	49	1.5%
Um	1,164	2.6%	12	0.5%	42	1.3%
All types	44,912	100.0%	2,397	100.0%	3,331	100.0%

Table 3: Error type distribution of the training and test data. The test data were annotated independently by two annotators.

ID	Prompt
1	"The decision to undergo genetic testing can only be made by the individual at risk for a disorder. Once a test has been conducted and the results are known, however, a new, family-related ethical dilemma is born: Should a carrier of a known genetic risk be obligated to tell his or her relatives?" Respond to the question above, supporting your argument with concrete examples.
2	While social media sites such as Twitter and Facebook can connect us closely to people in many parts of the world, some argue that the reduction in face-to-face human contact affects interpersonal skills. Explain the advantages and disadvantages of using social media in your daily life/society.

Table 4: The two prompts used for the test essays.

> There is no **a doubt** , tracking **system has** brought many benefits in this information age .

The set of gold-standard edits of a human annotator is $\mathbf{g}$ = {a doubt → doubt, system → systems, has → have}. Suppose the tokenized output sentence H of a grammatical error correction system given the above sentence is:

> There is no doubt , tracking system has brought many benefits in this information age .

That is, the set of system edits is $\mathbf{e}$ = {a doubt → doubt}. The performance of the grammatical error correction system is measured by how well the two sets $\mathbf{g}$ and $\mathbf{e}$ match, in the form of recall R, precision P, and $F_{0.5}$ measure: $R = 1/3, P = 1/1, F_{0.5} = (1 + 0.5^2) \times RP/(R + 0.5^2 \times P) = 5/7$.

More generally, given a set of n sentences, where $\mathbf{g}_i$ is the set of gold-standard edits for sentence i, and $\mathbf{e}_i$ is the set of system edits for sentence i, recall, precision, and $F_{0.5}$ are defined as follows:

$$R = \frac{\sum_{i=1}^{n} |\mathbf{g}_i \cap \mathbf{e}_i|}{\sum_{i=1}^{n} |\mathbf{g}_i|} \quad (1)$$

$$P = \frac{\sum_{i=1}^{n} |\mathbf{g}_i \cap \mathbf{e}_i|}{\sum_{i=1}^{n} |\mathbf{e}_i|} \quad (2)$$

$$F_{0.5} = \frac{(1 + 0.5^2) \times R \times P}{R + 0.5^2 \times P} \quad (3)$$

where the intersection between $\mathbf{g}_i$ and $\mathbf{e}_i$ for sentence i is defined as

$$\mathbf{g}_i \cap \mathbf{e}_i = \{e \in \mathbf{e}_i | \exists g \in \mathbf{g}_i, match(g, e)\} \quad (4)$$

Note that we have adopted $F_{0.5}$ as the evaluation metric in the CoNLL-2014 shared task instead of the standard F_1 used in CoNLL-2013. $F_{0.5}$ emphasizes precision twice as much as recall, while F_1 weighs precision and recall equally. When a grammar checker is put into actual use, it is important that its proposed corrections are highly accurate in order to gain user acceptance. Neglecting to propose a correction is not as bad as proposing an erroneous correction.

Similar to CoNLL-2013, we use the *MaxMatch* (M^2) scorer[2] (Dahlmeier and Ng, 2012b) as the official scorer in CoNLL-2014. The M^2 scorer[3] efficiently searches for a set of system edits that maximally matches the set of gold-standard edits specified by an annotator. It overcomes a limitation of the scorer used in HOO shared tasks, which can return an erroneous score since the system edits are computed deterministically by the HOO scorer without regard to the gold-standard edits.

5 Approaches

45 teams registered to participate in the shared task, out of which 13 teams submitted the output of their grammatical error correction systems. These teams are listed in Table 5. Each team is assigned a 3 to 4-letter team ID. In the remainder of this paper, we will use the assigned team ID to refer to a participating team. Every team submitted a system description paper (the only exception is the NARA team). Four of the 13 teams submitted their system output only after the deadline (they were given up to one week of extension). These four teams (IITB, IPN, PKU, and UFC) have an asterisk affixed after their team names in Table 5.

Each participating team in the CoNLL-2014 shared task tackled the error correction problem in a different way. A full list summarizing each

[2] http://www.comp.nus.edu.sg/∼nlp/software.html

[3] A few minor bugs were fixed in the M^2 scorer before it was used in the CoNLL-2014 shared task.

Team ID	Affiliation
AMU	Adam Mickiewicz University
CAMB	University of Cambridge
CUUI	Columbia University and the University of Illinois at Urbana-Champaign
IITB*	Indian Institute of Technology, Bombay
IPN*	Instituto Politécnico Nacional
NARA	Nara Institute of Science and Technology
NTHU	National Tsing Hua University
PKU*	Peking University
POST	Pohang University of Science and Technology
RAC	Research Institute for Artificial Intelligence, Romanian Academy
SJTU	Shanghai Jiao Tong University
UFC*	University of Franche-Comté
UMC	University of Macau

Table 5: The list of 13 participating teams. The teams that submitted their system output after the deadline have an asterisk affixed after their team names. NARA did not submit any system description paper.

team's approach can be found in Table 6. While machine-learnt classifiers for specific error types proved popular in last year's CoNLL-2013 shared task, since this year's task required the correction of all 28 error types, teams tended to prefer methods that could deal with all error types simultaneously. In fact, most teams built hybrid systems that made use of a combination of different approaches to identify and correct errors.

One of the most popular approaches to non-specific error type correction, incorporated to various extents in many teams' systems, was the Language Model (LM) based approach. Specifically, the probability of a learner n-gram is compared with the probability of a candidate corrected n-gram, and if the difference is greater than some threshold, an error was perceived to have been detected and a higher scoring replacement n-gram could be suggested. Some teams used this approach only to detect errors, e.g., IPN (Hernandez and Calvo, 2014), which could then be corrected by other methods, whilst other teams used other methods to detect errors first, and then made corrections based on the alternative highest n-gram probability score, e.g., RAC (Boroş et al., 2014). No single team used a uniquely LM-based solution and the LM approach was always a component in a hybrid system.

An alternative solution to correcting all errors was to use a phrase-based statistical machine translation (MT) system to "translate" learner English into correct English. Teams that followed the MT approach mainly differed in terms of their attitude toward tuning; CAMB (Felice et al., 2014) performed no tuning at all, IITB (Kunchukuttan et al., 2014) and UMC (Wang et al., 2014b) tuned $F_{0.5}$ using MERT, while AMU (Junczys-Dowmunt and Grundkiewicz, 2014) explored a variety of tuning options, ultimately tuning $F_{0.5}$ using a combination of kb-MIRA and MERT. No team used a syntax-based translation model, although UMC did include POS tags and morphology in a factored translation model.

With regard to correcting single error types, rule-based (RB) approaches were also common in most teams' systems. A possible reason for this is that some error types are more regular than others, and so in order to boost accuracy, simple rules can be written to make sure that, for example, the number of a subject agrees with the number of a verb. In contrast, it is a lot harder to write a rule to consistently correct Wci (wrong collocation/idiom) errors. As such, RB methods were often, but not always, used as a preliminary or supplementary stage in a larger hybrid system.

Finally, although there were fewer machine-learnt classifier (ML) approaches than last year, some teams still used various classifiers to correct specific error types. In fact, CUUI (Rozovskaya et al., 2014) *only* built classifiers for specific error types and did not attempt to tackle the whole range of errors. SJTU (Wang et al., 2014a) also preprocessed the training data into more precise error categories using rules (e.g., verb tense (Vt) errors might be subcategorized into present, past, or future tense etc.) and then built a single maximum entropy classifier to correct all error types. See Table 6 to find out which teams tackled which error types.

While every effort has been made to make clear which team used which approach to correct which set of error types, as there were more error types than last year, it was sometimes impractical to fit all this information into Table 6. For more information on the specific methods used to correct a specific error type, we must refer the reader to that team's CoNLL-2014 system description paper.

Table 6 also shows the linguistic features used by the participating teams, which include lexical features (i.e., words, collocations, n-grams), parts-of-speech (POS), constituency parses, and dependency parses.

While all teams in the shared task used the NUCLE corpus, they were also allowed to use additional external resources (both corpora and tools) so long as they were publicly available and not proprietary. Three teams also used last year's CoNLL-2013 test set as a development set in this year's CoNLL-2014 shared task. The external resources used by the teams are also listed in Table 6.

6 Results

All submitted system output was evaluated using the M^2 scorer, based on the error annotations provided by our annotators. The recall (R), precision (P), and $F_{0.5}$ measure of all teams are shown in Table 7. The performance of the teams varies greatly, from little more than five per cent to 37.33% for the top team.

The nature of grammatical error correction is such that multiple, different corrections are often acceptable. In order to allow the participating teams to raise their disagreement with the original gold-standard annotations provided by the annotators, and not understate the performance of the teams, we allow the teams to submit their proposed alternative answers. This was also the practice adopted in HOO 2011, HOO 2012, and CoNLL-2013. Specifically, after the teams submitted their system output and the error annotations on the test essays were released, we allowed the teams to propose alternative answers (gold-standard edits), to be submitted within four days after the initial error annotations were released.

Team ID	Precision	Recall	$F_{0.5}$
CAMB	39.71	30.10	37.33
CUUI	41.78	24.88	36.79
AMU	41.62	21.40	35.01
POST	34.51	21.73	30.88
NTHU	35.08	18.85	29.92
RAC	33.14	14.99	26.68
UMC	31.27	14.46	25.37
PKU*	32.21	13.65	25.32
NARA	21.57	29.38	22.78
SJTU	30.11	5.10	15.19
UFC*	70.00	1.72	7.84
IPN*	11.28	2.85	7.09
IITB*	30.77	1.39	5.90

Table 7: Scores (in %) *without* alternative answers. The teams that submitted their system output after the deadline have an asterisk affixed after their team names.

The same annotators who provided the error annotations on the test essays also judged the alternative answers proposed by the teams, to ensure consistency. In all, three teams (CAMB, CUUI, UMC) submitted alternative answers.

The same submitted system output was then evaluated using the M^2 scorer, with the original annotations augmented with the alternative answers. Table 8 shows the recall (R), precision (P), and $F_{0.5}$ measure of all teams under this new evaluation setting.

The $F_{0.5}$ measure of every team improves when evaluated with alternative answers. Not surprisingly, the teams which submitted alternative answers tend to show the greatest improvements in their $F_{0.5}$ measure. Overall, the CUUI team (Rozovskaya et al., 2014) achieves the best $F_{0.5}$ measure when evaluated with alternative answers, and the CAMB team (Felice et al., 2014) achieves the best $F_{0.5}$ measure when evaluated without alternative answers.

For future research which uses the test data of the CoNLL-2014 shared task, we recommend that evaluation be carried out in the setting that does *not* use alternative answers, to ensure a fairer evaluation. This is because the scores of the teams which submitted alternative answers tend to be higher in a biased way when evaluated with alternative answers.

We are also interested in the analysis of the system performance for each of the error types.

Team	Error	Approach	Description of Approach	Linguistic Features	External Resources
AMU	All	MT	Phrase-based translation optimized for F-score using a combination of kb-MIRA and MERT with augmented language models and task-specific features.	Lexical	Wikipedia, CommonCrawl, Lang-8
CAMB	All	RB/LM/MT	Pipeline: Rule-based → LM ranking → Untuned SMT → LM ranking → Type filtering	Lexical, POS	Cambridge "Write and Improve" SAT system, Cambridge Learner Corpus, CoNLL-2013 Test Set, First Certificate in English corpus, English Vocabulary Profile corpus, Microsoft Web LM
CUUI	ArtOrDet, Mec, Nn, Prep, SVA, Vform, Vt, Wform, Wtone	ML	Different combinations of averaged perceptron, naïve Bayes, and pattern-based learning trained on different data sets for different error types.	Lexical, POS, lemma, shallow parse, dependency parse	CoNLL-2013 Test Set, Google Web 1T
IITB	All	MT/ML	Phrase-based translation optimized for F-score using MERT and supplemented with additional RB modules for SVA errors and ML modules for Nn and ArtOrDet.	Lexical, shallow parse	None
IPN	All except Prep	LM/RB	Low LM score trigrams are identified as errors which are subsequently corrected by rules.	Lexical, lemma, dependency parse	Wikipedia
NTHU	ArtOrDet, Nn, Prep, "Prep+Verb", Spelling and Commas, SVA, Wform	RB/LM/MT	External resources correct spelling errors while a conditional random field model corrects comma errors. SVA errors corrected using a RB approach. All other errors corrected by means of a language model. Interacting errors corrected using an MT system.	Lexical, POS, dependency parse	Aspell, GingerIt, Academic Word List, British National Corpus, Google Web 1T, Google Books Syntactic N-Grams, English Gigaword
PKU	All	LM/ML	A LM is used to find the highest scoring variant of a word with a common stem while maximum entropy classifiers deal with articles and prepositions.	Lexical, POS, stem	Gigaword, Apache Lucene Spellchecker
POST	All	LM/RB	N-gram-based approach finds unlikely n-gram "frames" which are then corrected via high scoring LM alternatives. Rule-based methods then improve the results for certain error types.	Lexical, POS, dependency parse, constituency parse	Google Web 1T, CoNLL-2013 Test Set, PyEnchant Spellchecking Library
RAC	See Footnote[a]	RB/LM	Rule-based methods are used to detect errors which can then be corrected based on LM scores.	Lexical, POS, lemma, shallow parse	Google Web 1T, News CRAWL (2007 – 2012), Europarl, UN French-English Corpus, News Commentary, Wikipedia, LanguageTool.org
SJTU	All	RB/ML	Rule-based system generates more detailed error categories which are then used to train a single maximum entropy model.	Lexical, POS, lemma, dependency parse	None
UFC	SVA, Vform, Wform	RB	Mismatched POS tags generated by two different taggers are treated as errors which are then corrected by rules.	POS	Nodebox English Linguistics Library
UMC	All	MT	Factored translation model using modified POS tags and morphology as features.	Lexical, POS, prefix, suffix, stem	WMT2014 Monolingual Data

Table 6: Profile of the participating teams. The *Error* column lists the error types tackled by a team if not all were corrected. The *Approach* column lists the type of approach used, where LM denotes a Language Modeling based approach, ML a Machine Learning classifier based approach, MT a statistical Machine Translation approach, and RB a Rule-Based approach.

[a] The RAC team uses rules to correct error types that differ from the 28 official error types. They include: "the correction of the verb tense especially in time clauses, the use of the short infinitive after modals, the position of frequency adverbs in a sentence, subject-verb agreement, word order in interrogative sentences, punctuation accompanying certain lexical elements, the use of articles, of correlatives, etc."

Type	AMU	CAMB	CUUI	IITB	IPN	NARA	NTHU	PKU	POST	RAC	SJTU	UFC	UMC
Vt	10.66	19.12	3.79	1.74	0.88	14.18	10.61	12.30	3.76	26.19	4.17	0.00	14.84
Vm	10.81	22.58	0.00	0.00	0.00	29.03	0.00	3.23	0.00	35.90	0.00	0.00	6.45
V0	17.86	25.00	0.00	0.00	0.00	36.67	0.00	0.00	0.00	0.00	0.00	0.00	25.93
Vform	22.76	24.37	21.43	1.85	4.63	27.62	24.30	25.64	1.89	27.35	3.67	0.95	14.68
SVA	24.30	31.36	70.34	1.06	14.14	27.50	62.67	17.31	20.56	30.36	14.85	28.70	14.41
ArtOrDet	15.52	49.48	58.85	0.68	0.33	50.89	33.63	8.20	54.45	0.63	12.54	0.00	24.05
Nn	58.74	54.11	56.10	4.49	10.36	57.32	46.76	41.78	55.60	36.45	10.11	0.00	17.03
Npos	14.29	7.69	4.76	0.00	0.00	20.00	0.00	0.00	0.00	4.76	4.55	0.00	5.26
Pform	22.22	22.58	7.14	0.00	0.00	14.81	16.13	12.00	0.00	3.70	0.00	0.00	17.24
Pref	9.33	19.35	1.32	0.00	0.00	10.00	1.20	1.35	1.32	0.00	0.00	0.00	12.05
Prep	18.41	38.26	15.45	2.12	0.00	29.72	19.42	0.00	2.28	0.00	7.92	0.00	14.55
Wci	12.00	9.17	0.94	0.36	0.35	7.55	0.63	1.65	1.27	0.34	0.00	0.00	3.23
Wa	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wform	45.56	45.05	17.24	4.05	2.60	39.08	14.81	25.88	6.49	11.25	1.39	1.30	16.46
Wtone	81.82	36.36	36.36	0.00	0.00	14.29	0.00	0.00	28.57	0.00	16.67	0.00	44.44
Srun	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Smod	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Spar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.00	0.00	0.00	0.00
Sfrag	0.00	0.00	0.00	0.00	0.00	20.00	0.00	0.00	0.00	0.00	25.00	0.00	25.00
Ssub	7.89	14.63	0.00	0.00	2.27	15.38	0.00	0.00	9.52	2.38	2.27	0.00	6.98
WOinc	0.00	3.03	0.00	3.57	0.00	3.03	0.00	0.00	0.00	0.00	0.00	0.00	6.67
WOadv	0.00	47.62	0.00	12.50	0.00	43.75	0.00	0.00	0.00	0.00	0.00	0.00	44.44
Trans	13.43	21.43	2.86	1.43	0.00	11.25	1.41	1.52	2.67	0.00	0.00	0.00	12.16
Mec	29.35	28.75	15.79	1.02	4.33	36.69	6.67	30.28	36.61	43.51	0.51	0.00	16.80
Rloc−	5.41	20.16	7.76	0.00	5.56	18.64	9.68	10.48	9.26	9.09	2.50	0.00	15.84
Cit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Others	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.12	0.00	0.00	0.00
Um	7.69	9.09	0.00	0.00	0.00	4.00	0.00	15.79	8.70	8.33	0.00	0.00	0.00

Table 9: Recall (in %) for each error type *without* alternative answers, indicating how well each team performs against a particular error type.

Type	AMU	CAMB	CUUI	IITB	IPN	NARA	NTHU	PKU	POST	RAC	SJTU	UFC	UMC
Vt	11.61	20.00	5.79	1.90	0.98	16.18	12.90	14.16	3.31	29.17	4.59	0.00	17.60
Vm	11.11	23.33	0.00	0.00	0.00	29.03	0.00	3.33	0.00	39.47	0.00	0.00	7.69
V0	19.23	29.63	0.00	0.00	0.00	38.71	0.00	0.00	0.00	0.00	0.00	0.00	30.77
Vform	23.93	27.42	21.05	1.92	4.85	29.09	24.07	26.79	2.83	26.96	3.77	0.98	15.32
SVA	25.00	33.90	72.41	1.11	14.74	28.57	63.76	17.82	22.86	32.43	15.46	30.09	14.95
ArtOrDet	18.75	54.74	67.38	1.81	0.36	54.42	37.96	9.65	59.41	0.66	14.63	0.00	33.42
Nn	62.14	62.03	65.53	4.91	12.29	62.69	52.89	51.01	64.14	42.67	11.93	0.00	22.22
Npos	23.33	40.00	4.35	0.00	0.00	29.17	0.00	0.00	0.00	9.52	4.55	0.00	13.64
Pform	22.22	23.33	7.69	0.00	0.00	14.81	17.86	12.50	0.00	4.00	0.00	0.00	22.22
Pref	9.59	18.56	1.32	0.00	0.00	9.80	1.25	1.33	1.37	0.00	0.00	0.00	11.11
Prep	18.41	38.63	18.22	2.21	0.00	30.28	20.42	0.00	2.25	0.00	8.95	0.00	16.98
Wci	15.26	15.18	0.96	0.79	0.38	8.05	1.33	3.17	1.94	0.36	0.00	0.00	9.57
Wa	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wform	45.45	46.59	21.11	2.90	2.67	40.91	15.58	27.38	6.49	12.50	1.47	1.37	17.11
Wtone	88.24	38.46	52.63	0.00	0.00	12.50	0.00	0.00	50.00	0.00	33.33	0.00	55.56
Srun	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Smod	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Spar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.00	0.00	0.00	0.00
Sfrag	0.00	0.00	0.00	0.00	0.00	16.67	0.00	0.00	0.00	0.00	25.00	0.00	20.00
Ssub	7.89	14.29	0.00	0.00	2.33	15.38	0.00	0.00	9.76	2.44	2.33	0.00	6.98
WOinc	0.00	3.45	0.00	4.00	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	7.14
WOadv	0.00	50.00	0.00	16.67	0.00	44.44	0.00	0.00	0.00	0.00	0.00	0.00	50.00
Trans	14.52	22.39	3.08	1.67	0.00	11.84	1.56	1.64	2.82	0.00	0.00	0.00	20.78
Mec	31.56	30.67	17.47	1.13	4.79	37.28	7.17	31.69	37.88	45.82	1.10	0.00	22.31
Rloc−	5.45	26.47	7.38	0.00	5.62	21.43	11.34	12.38	11.82	10.00	3.66	0.00	29.66
Cit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Others	0.00	3.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00	0.00	0.00
Um	7.69	9.09	0.00	0.00	0.00	4.35	0.00	15.00	4.55	8.70	0.00	0.00	0.00

Table 10: Recall (in %) for each error type *with* alternative answers, indicating how well each team performs against a particular error type.

Team ID	Precision	Recall	$F_{0.5}$
CUUI	52.44	29.89	45.57
CAMB	46.70	34.30	43.55
AMU	45.68	23.78	38.58
POST	41.28	25.59	36.77
UMC	43.17	19.72	34.88
NTHU	38.34	21.12	32.97
PKU*	36.64	15.96	29.10
RAC	35.63	16.73	29.06
NARA	23.83	31.95	25.11
SJTU	32.95	5.95	17.28
UFC*	72.00	1.90	8.60
IPN*	11.66	3.17	7.59
IITB*	34.07	1.66	6.94

Table 8: Scores (in %) *with* alternative answers. The teams that submitted their system output after the deadline have an asterisk affixed after their team names.

Computing the recall of an error type is straightforward as the error type of each gold-standard edit is provided. Conversely, computing the precision of each of the 28 error types is difficult as the error type of each system edit is not available since the submitted system output only contains corrected sentences with no indication of the error type of the system edits. Predicting the error type out of the 28 types for a particular system edit not found in gold-standard annotation can be tricky and error-prone. Therefore, we decided to compute the per-type performance based on recall. The recall scores when distinguished by error type are shown in Tables 9 and 10.

6.1 Cross Annotator Comparison

To measure the agreement between our two annotators, we computed Cohen's Kappa coefficient (Cohen, 1960) for identification, which measures the extent to which annotators agreed which words needed correction and which did not, regardless of the error type or correction. We obtained a Kappa coefficient value of 0.43, indicating moderate agreement (since it falls between 0.40 and 0.60). While this may seem low, it is worth pointing out that the Kappa coefficient does not take into account the fact that there is often more than one valid way to correct a sentence.

In addition to computing the performance of each team against the gold standard annotations of both annotators with and without alternative annotations, we also had an opportunity to compare the performance of each team's system against each annotator individually.

A recent concern is that there can be a high degree of variability between individual annotators which can dramatically affect a system's output score. For example, in a much simplified error correction task concerning only the correction of prepositions, Tetreault and Chodorow (2008) showed an actual difference of 10% precision and 5% recall between two annotators. Table 11 hence shows the precision (P), recall (R), and $F_{0.5}$ scores for *all* error types against the gold standard annotations of each CoNLL-2014 annotator individually.

The results show that there can indeed be a high amount of disagreement between two annotators, the most noticeable being precision in the UFC system: precision was 70% for Annotator 2 but only 28% for Annotator 1. This 42% difference is, however, likely to be an extreme case, and most teams show little more than 10% variation in precision and 5% variation in $F_{0.5}$. Recall remained fairly constant between annotators. 10% is still a large margin however, and these results reinforce the idea that error correction systems should be judged against the gold-standard annotations of multiple annotators.

Table 12 additionally shows how each annotator compares against each other; i.e., what score Annotator 1 gets if Annotator 2 was the gold standard (part (a) of Table 12) and vice versa (part (b)).

The low $F_{0.5}$ scores of 45.36% and 38.54% represent an upper bound for system performance on this data set and again emphasize the difficulty of the task. The low human $F_{0.5}$ scores imply that there are many ways to correct a sentence.

7 Conclusions

The CoNLL-2014 shared task saw the participation of 13 teams worldwide to evaluate their grammatical error correction systems on a common test set, using a common evaluation metric and scorer. The best systems in the shared task achieve an $F_{0.5}$ score of 37.33% when it is scored without alternative answers, and 45.57% with alternative answers. There is still much room for improvement in the accuracy of grammatical error correction systems. The evaluation data sets and scorer used in our shared task serve as a benchmark for

Team ID	Annotator 1			Annotator 2		
	P	**R**	$\mathbf{F}_{0.5}$	**P**	**R**	$\mathbf{F}_{0.5}$
AMU	27.30	13.55	22.69	35.49	12.90	26.29
CAMB	24.96	19.62	23.67	35.22	20.29	30.70
CUUI	26.05	15.60	22.97	36.91	16.37	29.51
IITB	23.33	0.88	3.82	24.18	0.66	2.99
IPN	5.80	1.25	3.36	9.62	1.51	4.63
NARA	13.54	19.20	14.38	18.74	19.69	18.92
NTHU	22.19	11.38	18.64	31.48	11.79	23.60
PKU	21.53	8.36	16.37	27.47	7.72	18.17
POST	22.39	13.89	19.94	29.53	13.42	23.81
RAC	19.68	8.28	15.43	28.52	8.80	19.70
SJTU	21.08	3.09	9.75	24.64	2.59	9.12
UFC	28.00	0.59	2.70	70.00	1.06	4.98
UMC	20.41	8.78	16.14	26.63	8.38	18.55

Table 11: Performance (in %) for each team's output scored against the annotations of a single annotator.

P	**R**	$\mathbf{F}_{0.5}$
50.47	32.29	45.36

(a)

P	**R**	$\mathbf{F}_{0.5}$
37.14	45.38	38.54

(b)

Table 12: Performance (in %) for output of one gold standard annotation scored against the other gold standard annotation: (a) The score of Annotator 1 if Annotator 2 was the gold standard, (b) The score of Annotator 2 if Annotator 1 was the gold standard.

future research on grammatical error correction[4].

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. We thank our two annotators Mark Brooke and Diane Nicholls who provided the gold-standard annotations.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Tiberiu Boroş, Stefan Daniel Dumitrescu, Adrian Zafiu, Dan Tufiş, Verginica Mititelu Barbu, and Paul Ionuţ Văduva. 2014. RACAI GEC – a hybrid approach to grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

[4]http://www.comp.nus.edu.sg/∼nlp/conll14st.html

Daniel Dahlmeier and Hwee Tou Ng. 2011a. Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Daniel Dahlmeier and Hwee Tou Ng. 2011b. 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. 2012a. A beam-search decoder for grammatical error correction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 568–578.

Daniel Dahlmeier and Hwee Tou Ng. 2012b. 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 Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249.

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 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 54–62.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth Conference on Language Resources and Evaluation*, pages 449–454.

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*.

Michael Gamon. 2010. Using mostly native data to correct errors in learners' writing: A meta-classifier approach. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 163–171.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

S. David Hernandez and Hiram Calvo. 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*.

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*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.

Anoop Kunchukuttan, Sriram Chaudhury, and Pushpak Bhattacharyya. 2014. Tuning a grammar correction system for increased precision. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Publishers.

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.

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.

Alla Rozovskaya and Dan Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 961–970.

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*.

Joel R. Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *COLING Workshop on Human Judgments in Computational Linguistics*, Manchester, UK.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358.

Peilu Wang, Zhongye Jia, and Hai Zhao. 2014a. Grammatical error detection and correction using a single maximum entropy model. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*.

Yiming Wang, Longyue Wang, Derek F. Wong, Lidia S. Chao, Xiaodong Zeng, and Yi Lu. 2014b. Factored statistical machine translation for grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*.

Yuanbin Wu and Hwee Tou Ng. 2013. Grammatical error correction using integer linear programming. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1456–1465.

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*, pages 180–189.

CoNLL. The SIGNLL Conference on Computational Natural Language Learning. Previous shared tasks. 2019. Cross-Framework Meaning Representation Parsing. English. Proceedings. 2018. Universal Morphological Reinflection.Â English. Proceedings. 2014. Grammatical Error Correction. English. Proceedings. 2013. Grammatical Error Correction. English. Proceedings. 2012. Modelling Multilingual Unrestricted Coreference in OntoNotes. Share this article. Click here to see the statistics on "Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task" . Cited by 25 articles. Show/hide lists. 2014. The conll-2014 shared task on grammatical error correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pages 1â€“14. Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and elt.Â 2017. Grammatical error correction in non-native english. Technical report, University of Cambridge, Computer Laboratory. Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented archi-tecture with unlabeled data.