A Light Rule-based Approach to English Subject-Verb Agreement Errors on the Third Person Singular Forms

Yuzhu Wang[1,2], Hai Zhao[1,2 * †] and Dan Shi[3]
[1]Center for Brain-Like Computing and Machine Intelligence,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, 200240, China
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[3]LangYing NLP Research Institute,Shanghai LangYing Education Technology Co., Ltd.
hfut0830@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn,sweet@lyced.com

Abstract

Verb errors are one of the most common grammar errors made by non-native writers of English. This work especially focus on an important type of verb usage errors, subject-verb agreement for the third person singular forms, which has a high proportion in errors made by non-native English learners. Existing work has not given a satisfied solution for this task, in which those using supervised learning method usually fail to output good enough performance, and rule-based methods depend on advanced linguistic resources such as syntactic parsers. In this paper, we propose a rule-based method to detect and correct the concerned errors. The proposed method relies on a series of rules to automatically locate subject and predicate in four types of sentences. The evaluation shows that the proposed method gives state-of-the-art performance with quite limited linguistic resources.

1 Introduction

With the increasing number of people all over the world who study English as second language (ESL), grammatical errors in writing often occur due to cultural diversity, language habits, and education background. There has been a substantial and increasing need of using computational techniques to improve the writing ability for second language learners. In addition, such techniques and tools may help find latent writing errors in official documents as well. To meet the urgent need from ESL, a lot of works on natural language processing focus on the task of grammatical error detection and correction. Formally, it is a task of automatically detecting and correcting erroneous word usage and ill-formed grammatical constructions in text (Dahlmeier et al., 2012).

It is not a brand new task in natural language processing. However, it has been a challenging task for several reasons. First, many of these errors are context-sensitive so that errors cannot be detected and then corrected in an isolated way. Second, the relative frequency of errors is quite low: for a given type of mistake, an ESL writer will typically go wrong in only a small proportion of relevant language structures. For example, incorrect determiner usages usually occur in 5% to 10% of noun phrases in various annotated ESL corpora (Rozovskaya and Roth, 2011). Third, an ESL writer may make multiple mistakes in a single sentence, so that continuous errors are entangled, which let specific error locating and correction become more difficult.

* Correspondence author

† This work was partially supported by the National Natural Science Foundation of China (No. 61170114, and No. 61272248), the National Basic Research Program of China (No. 2013CB329401), the Science and Technology Commission of Shanghai Municipality (No. 13511500200), the European Union Seventh Framework Program (No. 247619), the Cai Yuanpei Program (CSC fund 201304490199 and 201304490171), and the art and science interdiscipline funds of Shanghai Jiao Tong University, No. 14X190040031, and the Key Project of National Society Science Foundation of China, No. 15-ZDA041.

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 345 - 353
Shanghai, China, October 30 - November 1, 2015

In recent decades, existing studies on this task have focused on errors in two typical word categories, article and preposition (Han et al., 2006; Felice and Pulman, 2008; Dahlmeier and Ng, 2011). However verb errors occur as often as article and preposition errors at least, though there are few works on verb related errors. Two reasons are speculated for why it is difficult to process verb mistakes. First, compared with articles and prepositions, verbs are more difficult to identify in text, as they can often be confused with other parts of speech (POS), and in fact many existing processing tools are known to make more errors on noisy ESL data (Nagata et al., 2011). Second, verbs are more complicated linguistically. For an English verb, it has five forms of inflections (see Table 1). Different forms imply different types of errors, even, one type of verb form may lead to multiple types of errors.

Form	**Example**
base(bare)	*speak*
base(infinitive)	*to speak*
third person singular	*speaks*
past	*spoke*
-ing participle	*speaking*
-ed participle	*spoken*

Table 1: Five forms of inflections of English verbs (Quirk et al., 1985), illustrated with the verb "*speak*". The base form is also used to construct the infinitive with "*to*".

China is a leading market for ESL. According to a rough statistics on essays written by Chinese students, verb related errors have given a percent as high as 15.6% among all grammatical errors, in which subject-verb agreement errors on the third person singular form cover 21.8%. Existing works paid little attention on such type of errors, or report unsatisfied performance (Rozovskaya et al., 2013). That is to say, errors made by Chinese students have a quite different type distribution from those by native English students, while existing computational approach cannot well meet the urgent requirement on grammatical error detection and correction. Furthermore, the previous approaches focus on machine learning that always needs a large scale of annotated data set available. However, being a machine learning task, grammatical error detection and correction is very difficult to receive satisfied performance as errors being negative samples has too low a portion in the entire text for learning (on average, 20 sentences can hold one error).

In this paper, to alleviate the drawbacks of existing work, we propose a full rule-based method to handle this sort of specific errors, without any requirement on annotated data. The rule model is built on the English grammar. As we avoid using high-level and time consuming support tools, typically, parser, only two lexicons and a part-of-speech (POS) tagger [1] (Toutanova et al., 2003) is adopted to provide necessary word category information. This makes our system can work with least linguistic resource compared to existing rule-based work.

The rest of this paper is organized as follows: Section 2 discusses a few related work. Section 3 gives detailed introduction about the proposed rule-based method. The experimental results will be presented and analyzed in Section 4, and the last section concludes this paper.

2 Related Work

Over the past few decades, there are many methods proposed for grammatical error detection and correction. Most of the efforts so far had been focused on article and preposition usage errors, as these were some of the most common mistakes among non-native English speakers (Dalgish, 1985; Leacock et al., 2010). These works were generally regarded as multiclass classification tasks (Izumi et al., 2003; Han et al., 2006; Felice and Pulman, 2008; Gamon et al., 2008; Tetreault et al., 2010; Rozovskaya and Roth, 2010b; Rozovskaya and Roth, 2011; Dahlmeier and Ng, 2011).

As for main techniques for the task, most methods can fall into two basic categories, machine learning based and rule-based. The use of machine learning methods to tackle this problem had shown a promising performance for specific error types. These methods were normally created based on a large corpus of well-formed native English texts (Tetreault and Chodorow, 2008; Tetreault et al., 2010) or annotated non-native data (Gamon, 2010;

[1] This POS tagger outputs a POS tag set as the same defined by Penn Treebank.

Han et al., 2010). Additionally, both generative and discriminative classifiers were widely used. Among them, Maximum Entropy (Rozovskaya and Roth, 2011; Sakaguchi et al., 2012; Quan et al., 2012) obtained a good result for preposition and article correction using a large feature set. Naive Bayes was also applied to recognize or correct the errors in speech or texts (Lynch et al., 2012). In addition, grammar rules and probabilistic language model were used as a simple but effective assistant for correction of spelling (Kantrowitz, 2003) and grammatical errors (Dahlmeier et al., 2012; Lynch et al., 2012; Quan et al., 2012; Rozovskaya et al., 2012).

As for rule-based method, (Rozovskaya et al., 2014) proposed a linguistically-motivated approach to verb error correction that made use of the notion of verb finiteness to identify triggers and types of mistakes, before using a statistical machine learning approach to correct these mistakes. In their approach, the knowledge of which mistakes should be corrected or of the mistake type was not required. But their model got a low recall.

Recently, researchers also made an attempt to integrate different methods. (Rozovskaya et al., 2013) presented a system that combined a set of statistical models, where each model specialized in correction one of the five type errors which were article, preposition, noun number , verb form and subject-verb agreement. Their article and preposition modules built on the elements of the systems described in (Rozovskaya and Roth, 2011).

(Gamon et al., 2009) mentioned a model for learning gerund/infinitive confusions and auxiliary verb presence/choice. (Lee and Seneff, 2008) proposed an approach based on pattern matching on trees combined with word n-gram counts for correcting agreement misuse and some types of verb form errors. However, they excluded tense mistakes. (Tajirei et al., 2012) considered only tense mistakes. In the above studies, it was assumed that the type of mistake that needs to be corrected is known, and irrelevant verb errors were excluded (Tajirei et al., 2012) addressed only tense mistakes and excluded from the evaluation other kinds of verb errors.

3 Our Approach

Our approach requires two lexicons and a POS tagger as the basic linguistic resource to perform the task. As for the POS tagger, we use the POS tag set defined by Penn treebank. It has 36 POS tags, and each has a specific syntactic or even semantic role, which is shown in Table 2. The detailed roles of these POS tags will give basic criterion to locate subject and its predicate in a sentence.

As for lexicons, it is used to determine if a verb is in root form or not. To judge whether a verb has an agreement error, we build two dictionaries. One consists of $2,677$ original verbs which are extracted from Oxford Advanced Learner's Dictionary (Hornby et al., 2009). The other contains all $2,677$ verbs in the third person singular form. We find that there is not a word which exists in both dictionaries, so we can decide whether a verb is in the root form or in the third person form by checking the verb in which dictionary. Then the remaining job is to locate the subject and its predicate. Linguistically, subject and predicate can be either syntactic or semantic. The subject in syntax (grammar) and semantics may be the same in a few cases, but different in the others. For an interrogative sentence such as "*who are you?*", "*who*" is the true subject in grammar, however, what we always need is the semantic or nominal subject "*you*", so that we can check the agreement between "*you*" and its predicate "*are*". Throughout the entire paper, our rules and processing always take subject and its predicates as the semantic or nominal ones.

According to the different relative locations of subject and its predicate in sentences, we put all sentences into four categories, declarative, interrogative, subordinate and "*there be*" sentences. These sentence categories will be effectively determined through limited number of rules on specific punctuations and marker words. For declarative sentences, subject is before its predicate. For interrogative sentences, there is no fixed location relation between subjects and its predicates. For "*there be*" sentences, the nominal subject is after the predicate "*be*".

POS Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Table 2: Penn Treebank POS tag set

3.1 Declarative Sentences

For declarative sentences, predicate can be easily determined by searching for the first verb from the beginning of the sentence. Because most of the subjects are either nouns or pronouns, we continue to scan the sentence from beginning to the position of the predicate to confirm the subject. Except the case that the subject is "*I*" whose predicate must be "*am*", all the subjects can be divided into the third person singular and the non-third person singular. For noun, we regard the words with POS tag "*NN*" as the third person singular and the words with POS tag "*NNS*" as the non-third person singular. For pronoun, we collect two lists (see Table 3) to distinguish whether the subject is the third person singluar. Note that a person name can also be subject and we regard the name as the third person singular. We can utilize the POS tag "*NNP*" and "*NNPS*" to locate a person name. For this case, we continue to scan the sentence from the position of subject to find a verb.

Third Person Singular	Non Third Person Singular
He_PRP	*You_PRP*
he_PRP	*you_PRP*
She_PRP	*We_PRP*
she_PRP	*we_PRP*
It_PRP	*They_PRP*
it_PRP	*they_PRP*
That_DT	*These_DT*
that_WDT	*these_DT*
This_DT	*Those_DT*
this_DT	*those_DT*
That_WDT	*us_PRP*

Table 3: Pronouns of the third person and none third person (with POS tags)

With the above processes, we will still receive a wrong result for specific sentences with compound subject. For example, " *Tom and Jack come from America* .". So we need to add a rule to process these compound subjects. The desired subject can be determined by checking if it is after a word and POS tag combination, "*and_CC*", which means that the word is "*and*" as a conjunction for the case that the subject is determined to be third person.

Although we can deal with most of the simple sentences so far, there are also many sentences which can not be process according to these rules.

Firstly, for the sentences which have a modal verb before the predicate, the wanted verb must be in the original form no matter the subject is third person. We can identify this case by searching POS tag "*MD*" between the subject and the verb.

Secondly, there are often many compound sentences in statement. For example,
1. "*He likes apple but she like orange .*"
2. "*She will name him whatever she want to .*"
3. "*I love her because she give me life .*"
4. "*As we all know , human can not live without water .*"

For these cases, we divide the sentences into two parts and handle the rest part as declarative sentence recursively. For sentences like example 1-3, we build a list which consists of the words called *separate word* (see Table 4). We split the sentences by means of finding the *separate word*. For the sentences like example 4, the comma mark is used as the splitting boundary. We can utilize the words called *guided word* (see Table 5) to identify this type of sentences.

and_CC	*but_CC*
so_RB	*or_CC*
because_IN	*nor_CC*
whatever_WDT	*whatever_WPT*
whether_IN	*what_WP*
why_WRB	*where_WRB*
when_WRB	*how_WRB*
whose_WPS	*that_IN*
before_IN	*if_IN*
wherever_WPT	

Table 4: The *separate words* (with POS tags)

As_IN	*If_IN*
Although_IN	*When_WRB*
So_RB far_RB as_IN	

Table 5: The *guided words* (with POS tags)

However, for sentences that were led by a prepositional phrase, the rules proposed above can not correctly deal with. Here are two examples:
1. "*In my view, they are right .*"
2. "*In the morning , the dogs are running on the road .*"

We will regard the "*view*" and "*morning*" as subject according to the existing rules. But the true subjects are "*they*" and "*dogs*". So if there is "*In_IN*" before the noun, we will abandon the noun and regard the rest of the sentence as a new sentence for processsing.

3.2 Interrogative Sentences

In English grammar, questions mainly contain four categories. They are general question, alternative question, special question and tag question. Here are four examples:
1. "*Are you student ?*"
2. "*Can you speak Chinese or English ?*"
3. "*Who are you ?*"
4. "*They work hard , don't they ?*"

As in general predicate is before subject in most interrogative sentences, we scan the sentence from the beginning and regard the first verb as the predicate according to POS tag "*VB*". Then we continue to scan the sentence until the subject is found. The rules are the same as those proposed for declarative sentences.

Note that a tag question consists of two parts, a declarative sentence and a general question in abbreviation form. So we must divide the disjunctive question into two parts and process the first part as declarative sentence. Note that the fourth symbol from the end is a comma in all tag questions. We will make a full use of this mark to effectively divide a tag question.

There are also a few sentences that deserve our attention. For instance,
1. "*Whose jeans are they ?*"
2. "*How many boys are there ?*".

We can find that subject is in front of predicate in these sentences, so we can simply regard these sentence as declarative sentences. These types of sentences can be found by checking if they start from words like "*Whose_JJ*", "*How_WRB many_JJ*" and "*How_WRB much_RB*".

3.3 Subordinate Clause

So far, we have considered most of simple sentences. But there are many compound sentences with subordinate clause in real expression. We

furthermore divide the sentences with subordinate clause into five categories. Here are five examples:
1. "*The girl who is speaking now comes from Japan .*"
2. "*He gives me a gift which is very beautiful .*"
3. "*What she wants is a lovely doll .*"
4. "*The club will give whoever wins the competition a prize .*"
5. "*She will give him whatever he wants to .*"

For the first and second categories, we need pay attention to the conjunctions "*who*", "*which*" and "*that*". But the positions of the conjunctions are different in first and second categories. For the sentence like example 1, we check whether there is a conjunction between subject and predicate. If we find the conjunctions, we regard both the first and the second verbs as the predicate with the same subject.

For the second category, we check whether there is a conjunction after the predicate. If the conjunction is found, we will scan the sentence from the position of the conjunction to the position of predicate to find the subject of subordinate clause. The rules and treatments used to find the subject are the same as those proposed for declarative sentence. At last we scan the sentence from the position of conjunction to the end to find the predicate of subordinate clause.

For a sentence as example 3, we check whether the sentence begins with "*What_WP*" or "*Whether_IN*". If it is, we regard the second verb as the predicate of the subordinate clause and consider the subject of the subordinate clause as the third person. If we find "*whoever_WP*" after the verb in a sentence, we will scan the sentence from the position of "*whoever_WP*" to the end to find the second predicate and consider its subject as the third person.

For the last category, we divide the sentence into two parts by locating the word "*whatever_WP*" and handle both parts as declarative sentences.

3.4 *"There be"* Sentences

The semantic subject of "*there be*" sentence is the first noun right after the verb "*be*". Note that sentences like "*Here is five questions to be answered.*" also can be regard as "*there be*" sentences. All these types of sentences can be identified by searching the leading words *"There_EX"* and "*Here_RB*".

3.5 Additional Rules

Although most of the sentences can be processed by the proposed rules now, there are still some very special cases that can not be handled. Moreover, the outputs of POS tagger are not exact completely. So we give a few additional rules to strengthen the model.

Firstly, the words like "*Chinese*" are third person when they mean a language, otherwise, they are not. We call these words *language words*. We observe that when the *language word* means *language*, there is always a word "*language*" in the sentence. So we check whether there is "*language*" in the sentence that contains a *language word*. If we find "*language*", we will compulsively modify the corresponding word with the an updated POS tag "*NN*". Otherwise, we change the word with the an updated POS tag "*NNS*". There is also a situation that the subject is a gerund sometimes. We know that the gerund can not be a predicate by itself. So we change all the gerunds with the POS tag "*NN*". Table 6 shows additional rules to fortify the model.

3.6 Correction

Because there is not a word in both original form and third person form and one verb only has one third person form, we build a mapping dictionary to map a word from its root form to the third person singular form. Each word that is detected as error can be restored by searching this mapping dictionary.

4 Result

We select 300 sentences with agreement errors and $3,000$ correct sentences from essays written by Chinese students as the test data. This data set is provided by Shanghai LangYing Education Technology Co., Ltd.. The results are evaluated by the metrics, precision P, recall R of error detection and correction, and their harmonic average $F1$ score (Table 7). As Lee model (Lee and Seneff, 2008) can process subject-verb agreement errors well, we compare their results with ours on the same test data set[2].

[2]As (Lee and Seneff, 2008) do not release their data set and system implementation, we have accurately re-implement their system to make this comparison.

The case need to be handled	**The rules**
If there is "*Not only*".	Abandon all the words before "*also*"
If there is "*I think*".	Check whether "*I think*" is wrong then abandon "*I think*".
If there is "*percent of*".	Abandon "*percent of*".
If there is "*a lot of* ".	Abandon "*a lot of* ".
If there is "*a number of*".	Abandon "*a number of*".

Table 6: Additional rules

The comparison in Table 7 shows that our model outperforms Lee model by 6.7% in terms of F1 score. In addition, the results of Lee model were achieved by adopting advanced parse tree, while we use no more than POS tags.

We also show the result of Rozovskaya model (Rozovskaya et al., 2014) and UIUC model (Rozovskaya et al., 2013) (see Table 8 and 9). Our model is significantly better than theirs for subject-verb agreement errors though their model can deal with various types of errors. However, it is worth noting that their test data sets are different for all existing works and ours. Therefore, we compare their results only for reference.

5 Conclusion

Verb errors are commonly made by ESL writers but difficult to process. Subject-verb agreement errors on the third person singular form cover 21.8% of the verb errors according to statistics from a typical ESL group. Previous works paid little attention on such type of errors, and report unsatisfied performance. Using quite limited linguistic resources, we develop a rule-based approach that gives state-of-the-art performance on detecting and correcting the subject-verb agreement errors.

Model		**P**	**R**	**F1**
Our Model	Identification	85.0	81.7	83.3
	Correction	85.0	81.7	83.3
Lee Model	Identification	82.3	71.6	76.6
	Correction	82.3	71.6	76.6

Table 7: Results

Models	**P**	**R**	**F1**
Scores on the original annotations			
Articles	48	11	18
+Prepositions	48	12	19
+Noun number	48	21	29
+Subject-verb agr	48	22	30
+Verb form(All)	46	23	31
Scores based on the revised annotations			
All	62	32	42

Table 9: Results of the UIUC model

References

A S Hornby, Sally Wehmeier and Michael Ashby. 2009. *Oxford Advanced Learner's Dictionary* . Oxford University Press, Oxford, England.

Alla Rozovskaya and Dan Roth. 2010b. *Training paradigms for correcting errors in grammar and usage*. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 154-162.

Alla Rozovskaya and Dan Roth. 2011. *Annotating ESL errors: Challenges and rewards*. In Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications, pp. 28-36

Alla Rozovskaya, Dan Roth and Srikumar Vivek. 2014. *Correcting Grammatical Verb Errors* . In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 358-367

Alla Rozovskaya, Kaiwei Chang, Mark Sammons and Dan Roth. 2013. *The University of Illinois System in the CoNLL-2013 Shared Task* . In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pp. 13-19

Alla Rozovskaya, Mark Sammons and Dan Roth. 2012. *The UI system in the HOO 2012 shared task on error correction* . In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 272C280.

Chang-Ning Huang and Hai Zhao. 2006. *Which Is Essential for Chinese Word Segmentation: Character versus Word* . In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20), pp. 1-12

Error type	Correction			Identification		
	P	R	F1	P	R	F1
Agreement	90.62	9.70	17.52	90.62	9.70	17.52
Tense	60.51	7.47	13.31	86.63	10.70	19.06
Form	81.83	16.34	27.24	83.47	16.67	27.79
Total	71.94	10.24	17.94	85.81	12.22	21.20

Table 8: Results of Rozovskaya model

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

Daniel Dahlmeier and Hwee Tou Ng. 2011. *Grammatical Error Correction with Alternating Structure Optimization*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 915-923

Daniel Dahlmeier, Hwee Tou Ng and Eric JUn Feng Ng. 2012. *NUS at the HOO 2012 Shared Task* . In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 216-224

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi and Hitoshi Isahara. 2003. *Automated Grammatical Error Detection for Language Learners* . In Proceedings of 41st Annual Meeting of the Association for Computational Linguistics, pp. 145-148

G. Dalgish. 1985. *Computer-assisted ESL research* . CALICO Journal, 2(2)

Gerard Lynch, Erwan Moreau and Carl Vogel. 2012. *A Naive Bayes classifier for automatic correction of preposition and determiner errors in ESL text* . In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 257C262.

Hai Zhao and Chunyu Kit. 2007. *Incorporating Global Information into Supervised Learning for Chinese Word Segmentation*. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING-2007), pp. 66-74

Hai Zhao, Chang-Ning Huang and Mu Li. 2006. *An Improved Chinese Word Segmentation System with Conditional Random Field*. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5), pp. 162-165

Hai Zhao, Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. *Multilingual Dependency Learning: A Huge Feature Engineering Method to Semantic Dependency Parsing* . In Proceedings of Thirteenth Conference on Computational Natural Language Learning, pp. 55-60

Hai Zhao, Xiaotian Zhang, and Chunyu Kit. 2013. *Integrative Semantic Dependency Parsing via Efficient Large-scale Feature Selection*. Journal of Artificial Intelligence Research, Volume 46:203-233

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. *Cross Language Dependency Parsing using a Bilingual Lexicon* . Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 55-63

Hai Zhao. 2009. *Character-Level Dependencies in Chinese: Usefulness and Learning* . In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 879-887

Jian Zhang, Hai Zhao, Liqing Zhang, and Baoliang Lu. 2011. *An Empirical Comparative Study on Two Large-Scale Hierarchical Text Classification Approaches*. International Journal Computer Processing of Oriental Language, pp. 309-326

Jingyi Zhang and Hai Zhao. 2013 *Improving Function Word Alignment with Frequency and Syntactic Information*. In Proceedings of International Joint Conference on Artificial Intelligence-2013, pp. 2211-2217

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

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, pp. 353-358

John Lee and Stephanie Seneff. 2008. *Correcting Misuse of Verb Forms*. In Proceedings 46th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies, pp. 175-182

Junhui Li, Guodong Zhou, Hai Zhao, Qiaoming Zhu, and Peide Qian. 2009. *Improving Nominal SRL in Chinese Language with Verbal SRL Information and*

Automatic Predicate Recognition. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1280-1288

Keisuke Sakaguchi, Yuta Hayashibe, Shuhei Kondo, Lis Kanashiro, Tomoya Mizumoto, Mamoru Komachi and Yuji Matsumoto. 2012. *NAIST at the HOO 2012 Shared Task* . In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 281C288.

Kristina Toutanova, Dan Klein, Christopher Manning and Yoram Singer. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency* . In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp. 252-259

Li Quan, Oleksandr Kolomiyets and Marie-Francine Moens. 2012. *KU Leuven at HOO-2012: a hybrid approach to detection and correction of determiner and preposition errors in non-native English text* . In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 263C271.

Mark Kantrowitz. 2003. *Method and apparatus for analyzing affect and emotion in text* . Patent No. 6,622,140.

Michael Gamon, Claudia Leacock, Chris Brockett, William B. Dolan, Jianfeng Gao, Dmitriy Belenko and Alexandre Klementiev. 2009. *Using statistical techniques and web search to correct ESL errors* . CALICO Journal, Special Issue on Automatic Analysis of Learner Language, 26(3):491C511.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko and Lucy Vanderwende. 2008. *Using contextual speller techniques and language modeling for ESL error correction* . In Proceedings of third International Joint Conference on Natural Language Processing Proceedings of the Conference

Michael Gamon. 2010. *Using mostly native data to correct errors in learners writing: a meta-classifier approach* . In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 163C171

Na-Rae Han, Joel Tetreault, Soo-Hwa Lee and Jin-Young Ha. 2010. *Using an error-annotated learner corpus to develop an ESL/EFL error correction system* . In Proceedings of LREC, pp. 763C770

Na-Rae Han, Martin Chodorow and Claudia Leacock. 2012. *Detecting errors in English article usage by non-native speakers* . Journal of Natural Language Engineering, pp. 115-129

Rachele De Felice and Stephen G. Puluman. 2008. *A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English*. In Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008), pp. 169-176

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language* . Longman, NewYork

Rui Wang, Hai Zhao, Baoliang Lu, Masao Utiyama, and Eiichiro Sumita. 2015 *Bilingual Continuous-Space Language Model Growing for Statistical Machine Translation,*. IEEE/ACM Transactions on Audio, Speech, and Languange Processing, Vol.23(7): 1209-1220

Rui Wang, Hai Zhao, Baoliang Lu, Masao Utiyama, and Eiichro Sumita. 2014. *Neural Network Based Bilingual Language Model Growing for Statistical Machine Translation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 189-195

Rui Wang, Masao Utiyama, Isao Goto, Eiichro Sumita, HaiZhao, and Baoliang Lu. 2013. *Converting Continuous-Space Language Models into N-gram Language Models for Statistical Machine Translation* . In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 845-850

Ryo Nagata, Edward Whittaker and Vera Sheinman. 2011. *Creating a manually error-tagged and shallow-parsed learner corpus* . In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1210-1219

Toshikazu Tajiri, Mamoru Komachi and Yuji Matsumoto. 2012. *Tense and aspect error correction for esl learners using global context* . In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 198-202

Xiaolin Wang, Hai Zhao, and Baoliang Lu. 2013 *Labeled Alignment for Recognizing Textual Entailment*. International Joint Conference on Natural Language Processing, pp. 605-613

Xuezhe Ma and Hai Zhao. 2012. *Fourth-Order Dependency Parsing*. In Proceedings of the 24th International Conference on Computational Linguistics, pp. 8-15

Zhongye Jia , Peilu Wang, and Hai Zhao. 2013. *Grammatical Error Correction as Multiclass Classification with Single Model*. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pp.74-81

Zhongye Jia and Hai Zhao. 2014. *A Joint Graph Model for Pinyin-to-Chinese Conversion with Typo Correction*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics , pp. 1512-1523