Analyzing the Impact of Spelling Errors on POS-Tagging and Chunking in Learner English

Tomoya Mizumoto[1] and **Ryo Nagata**[2 1]
[1] RIKEN Center for Advanced Intelligence Project
[2] Konan University
tomoya.mizumoto@riken.jp, nagata-ijcnlp @hyogo-u.ac.jp

Abstract

Part-of-speech (POS) tagging and chunking have been used in tasks targeting learner English; however, to the best our knowledge, few studies have evaluated their performance and no studies have revealed the causes of POS-tagging/chunking errors in detail. Therefore, we investigate performance and analyze the causes of failure. We focus on spelling errors that occur frequently in learner English. We demonstrate that spelling errors reduced POS-tagging performance by 0.23% owing to spelling errors, and that a spell checker is not necessary for POS-tagging/chunking of learner English.

1 Introduction

Part-of-speech (POS) tagging and chunking have been essential components of Natural Language Processing (NLP) techniques that target learner English, such as grammatical error correction and automated essay scoring. In addition, they are frequently used to extract linguistic features relevant to the given task. For example, in the CoNLL-2014 Shared Task (Ng et al., 2014), 10 of the 12 teams used one or both POS-tagging and chunking to extract features for grammatical error correction.

They have also been used for linguistic analysis of learner English, particularly in corpus-based studies. Aarts and Granger (1998) explored characteristic POS patterns in learner English. Nagata and Whittaker (2013) demonstrated that POS sequences obtained by POS-tagging can be used to distinguish between mother tongue interferences effectively.

The heavy dependence on POS-tagging and chunking suggests that failures could degrade the performance of NLP systems and linguistic analyses (Han et al., 2006; Sukkarieh and Blackmore, 2009). For example, failure to recognize noun phrases in a sentence could lead to failure in correcting related errors in article use and noun number. More importantly, such failures make it more difficult to simply count the number of POSs and chunks, thereby causing inaccurate estimates of their distributions. Note that such estimates are often employed in linguistic analysis, including the above-mentioned studies.

Despite its importance in related tasks, we also note that few studies have focused on performance evaluations of POS-tagging and chunking. Only a few studies, including Nagata et al. (2011), Berzak et al. (2016) and Sakaguchi et al. (2012), have reported the performance of POS taggers in learner English and found a performance gap between native and learner English. However, none of those studies described the root causes of POS-tagging and chunking errors in detail. Detailed investigations would certainly improve performance, which in turn, would improve related tasks. Furthermore, to the best of our knowledge, no study has reported chunking performance when applied to learner English. [1]

Unknown words are a major cause of POS-tagging and chunking failures (Manning, 2011). In learner English, spelling errors, which occur frequently, are a major source of unknown words.

Spell checkers (e.g., Aspell) are used to correct spelling errors prior to POS-tagging and chunking. However, their effectiveness remains unclear.

Thus, we evaluate the extent to which spelling errors in learner English affect the POS tag-

[1] It appears that parsing doubles as chunking; however, chunking only considers a minimal phrase (non-recursive structures).

Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications, pages 54–58,
Taipei, Taiwan, December 1, 2017

ging and chunking performance. More precisely, we analyze the performance analysis of POS-tagging/chunking to determine (1) the extent to which performance is reduced due to spelling errors, (2) what types of spelling errors impact the performance, and (3) the effect of correcting spelling errors using a spell checker. Our analysis demonstrates that employing a spell checker is not required preliminary step of POS-tagging and chunking for NLP analysis of learner English.

2 Performance Analysis of Spelling Errors

Here, we explain how we analyzed POS tagging/chunking performance relative to spelling errors.

Extent of performance degradation due to spelling errors Spelling errors occur frequently in learner English. For example, the learner corpus used in (Flor et al., 2013) includes 3.4% spelling errors. Thus assuming that POS-tagging and chunking fails for all unknown words, performance would be reduced by 3.4% owing spelling errors. Realistically, performance does not drop a full 3.4% because POS-taggers and chunkers can infer POS/chunk from surrounding words. However, it is not clear how POS-tagging/chunking can correctly predict them. In contrast, if it is possible to estimate POSs/chunks of misspelled words from surrounding words, this has the potential to fail due to spelling errors. To investigate the extent to which performance is reduced due to spelling errors, we compared the results of POS-tagging and chunking on learner English without correcting spelling errors to results obtained by POS-tagging and chunking on learner English in which spelling errors were first corrected. In addition, we measured the effect of misspelled words had on them or their surrounding words by counting the number of correctly identified POSs/chunks.

Types of spelling errors There are various types of spelling errors in learner English (Sakaguchi et al., 2012). The most common type of spelling error is a typographical error (e.g., *studing/studying). In learner English, other types of errors include homophones (e.g., *see/sea), confusion (e.g., *form/from), splits (e.g., *home town/hometown), merges (e.g., *airconditioner/air conditioner), inflections (e.g., *program/programming) and derivations (e.g., *smell/smelly).[2] Some spelling errors, such as typographical and merge errors, result in unknown words, whereas others, such as homophones and split errors, are known words. For unknown words, it is possible to predict POSs/chunks from surrounding words, whereas for known words (e.g., homophone errors), POS-tagging/chunking fails. We use specific examples to investigate what types of spelling errors impact the performance of POS-tagging.

Some spelling errors have effective information that helps determine POSs. For example, for the above typographical error (i.e., *studing/studying), it may be possible to predict the corresponding POS as a "gerund or present participle verb" based on the suffix "ing." We also consider the effectiveness of prefix and suffix (i.e., affix) information in determining the corresponding POS for misspelled words. For this investigation, we compared POS-tagging systems both with and without affix information.

Effects of a spell checker Some previous studies into grammatical error correction investigated using a spell checker in a preprocessing step to reduce the negative impact of spelling errors. However, as noted above, little is known about the performance of POS-tagging and chunking for misspelled words and their surrounding words. Therefore, the effectiveness of a spell checker in a preprocessing step on POS-tagging and chunking for learner English remains unclear. Spell checkers can correct some errors, particularly unknown word errors; thus, POS-tagging and chunking have the potential to predict correct tags. We therefore examined the effect of a spell checker has on POS-tagging and chunking performance by comparing results obtained with and without the use of a spell checker.

3 Experimental Setup

To evaluate the performance of POS-tagging and chunking, we used the Konan-JIEM (KJ) corpus (Nagata et al., 2011) , which consists of 3,260 sentences and 30,517 tokens. Note that the essays in the KJ corpus were written by Japanese university students. The number of spelling errors targeted in this paper was 654 (i.e., 2.1% of all words).

We used a proprietary dataset comprising English teaching materials for reading comprehen-

[2]Note that we do not address split and merge errors.

#TP	#FP	#FN	Precision	Recall	F-score
409	197	120	67.49	77.32	72.07

Table 1: Performance of spelling error correction

sion for Japanese students. We annotated this dataset with POS tags and chunks to train a model for POS-tagging and chunking. This corpus consists of 16,375 sentences and 213,017 tokens, and does not contain grammatical errors. We also used sections 0-18 of the Penn TreeBank only to train the model for POS-tagging.

We formulated the POS-tagging and chunking as a sequence labeling problem. We used a conditional random field (CRF) (Lafferty et al., 2001) for sequence labeling and CRF++[3] with default parameters as a CRF tool. The features used for POS-tagging were based on the widely used features employed in Ratnaparkhi (1996). These features consist of surface, original form, presence of specific characters (e.g., numbers, uppercase, and symbols), and prefix and suffix (i.e., affix) information. In addition to (Ratnaparkhi, 1996), we used the original forms of words as features. For the chunking task, we also employed generally used features in this case from Sha and Pereira (2003). These features were based on surface, the original form of the words and POSs. These features are used in which tools are commonly used for grammatical error correction tasks.

We also developed a spell checker for our experiments. We constructed the spell checker based on a noisy channel model to capture the influence of spelling errors originating via the mother tongue. Table 1 summarizes the spelling correction performance of the spell checker on the KJ corpus. As can be seen, better performance results is demonstrated compared to Sakaguchi et al. (2012). In most previous research into grammatical error correction, a spell checker is used in a pipeline. Therefore, we used this pipeline method and treated spelling correction and POS-tagging and chunking as cascading problems.

For our evaluation metrics, we used accuracy (number of correct tokens / number of tokens in the corpus). In addition, we counted the number of correct tokens identified despite spelling errors, as well as their preceding and succeeding tokens, to observe the effect of spelling errors had on their surrounding words.

[3] https://taku910.github.io/crfpp/

Method	Accuracy
Baseline	93.97 (92.71)
Base+Aff	95.31 (93.93)
Base+Checker	94.21 (93.13)
Base+Aff+Checker	95.37 (94.05)
Base+Aff+Gold	**95.54** (**94.16**)

Table 2: Results of POS-tagging. Accuracies of POS-tagging trained on Penn TreeBank are shown in parentheses.

Method	# of s_i correct	# of s_{i-1} correct	# of s_{i+1} correct
Baseline	344	540	590
Base+Aff	465	542	598
Base+Aff+Gold	528	547	596

Table 3: Results of POS tagging for misspelled words and their surrounding words. s_i indicates a misspelled word.

4 POS-tagging Experiments

We conducted POS-tagging experiments to investigate the question introduced in Section 2. We prepared the following five methods:

1. A POS-tagging system trained with surface, original form, and presence of particular character features (**Baseline**)
2. A system with prefix and suffix (affix) features added to the *Baseline* (**Base+Aff**)
3. The *Baseline* POS-tagging system with a spell checker (**Base+Checker**)
4. The *Base+Aff* POS-tagging system with a spell checker (**Base+Aff+Checker**)
5. The *Base+Aff* POS-tagging system without a spell checker, i.e., errors were corrected manually (**Base+Aff+Gold**)

Table 2 summarizes the experimental results for POS-tagging. The results show the same tendency for POS-tagging trained on in-house data and POS-tagging trained on Penn TreeBank, i.e., Base+Aff+Gold > Base+Aff+Checker > Base+Aff > Base+Checker > Baseline. Therefore, to simplifying analysis, we used results obtained with the in-house data. First, we compared *Base+Aff* to *Base+Aff+Gold* to determine the influence of spelling errors. *Base+Aff+Gold* achieved a 0.23% improvement over *Base+Aff*. From this, we conclude that the POS-tagging performance dropped 0.23% due to spelling errors.

This also indicates that an ideal spell checker does have a positive impact on POS-tagging.

We also observed that *Base+Aff* demonstrated 1.3% higher accuracy compared to *Baseline*. Similarly, *Base+Aff* showed higher accuracy than that of *Base+Checker*. These results indicate that affix information is important to assigning corresponding POSs in learner English. Furthermore, there was only a difference of only 0.06% between *Base+Aff* and *Base+Aff+Checker*, thereby demonstrating that a spell checker is not necessary and that it is sufficient to assign POSs using affix information.

Table 3 shows the number of correct POSs identified for misspelled and surrounding words. As can be seen by comparing the *Baseline* to *Base+Aff+Gold*, the number of correct POSs for misspelled words increased. In contrast, for the number of correct POSs identified for surrounding words, there was nearly no difference, implying that spelling errors do not influence the accuracy of estimating the POSs of their surrounding words.

Types of spelling errors that affect performance We first compared *Baseline* to *Base+Aff* to observe spelling errors that can be corrected with affix information. The numbers of correct POSs for *Baseline* and *Base+Aff* were 344 and 465, respectively. Therefore, by using affix information, we could identify the correct POS for approximately 120 misspelled words. Two examples in which the *Baseline* failed in POS-tagging but *Base+Aff* succeeded are shown in the following.

(1) a. Winter is decolated/***Verb, past*** ...
b. Accoding/***Verb, gerund*** to ...

Here, the POS-tagger was able to assign correct POSs to misspelled words using affix information. Both *decolated* (**decorated)* and *Accoding* (**According*) were inferred via the *ed* and *ing* suffixes, respectively.

Next, we analyzed the output of *Base+Aff* and *Base+Aff+Gold* to identify spelling errors that make it difficult to predict POS-tags. The number of POSs that *Base+Aff* failed to identify in POS-tagging but *Base+Aff+Gold* identified successfully was 105. We divided these 105 errors into five types according to the cause of the failure. The most frequent cause (54 instances) was unknown words from spelling errors (e.g., evey). The remaining causes of failure were as follows: 20 errors in which a POS was predicted based on affix features (e.g., whiting), 17 errors due to different words (e.g., thought→though), 10 errors in which the POS was predicted based on the presence of uppercase characters (e.g., Exsample), and three errors caused by romanized Japanese words.

Method	Accuracy
Baseline	94.38
Base+Checker	94.41
Base+Gold	**94.58**

Table 4: Chunking results

Method	# of s_i correct	# of s_{i-1} correct	# of s_{i+1} correct
Baseline	532	504	565
Base+Gold	566	519	570

Table 5: Results of chunking involving misspelled words, as well as corresponding preceding and succeeding words.

Effect of spelling correction by spell checker We analyzed spelling errors where POS-tagging failed in the system with affix information but the system with the spell checker succeeded. The number of spelling errors that were correctly assigned to POSs with the spell checker was 74, whereas the number of spelling errors incorrectly assigned a POS was 49. The system with the spell checker correctly assigned a POS to the following:

(2) a. pepole/***Noun, singular*** → people/***Noun, plural***
b. tow/***Noun, singular*** apples→ two/***Numeral*** apples

These examples show cases in which spelling errors were corrected by the spell checker. As mentioned priviously, these spelling errors are examples of words in which POS-tagging failed due to unknown words. Examples in which POS-tagging with the spell checker failed involved the spell checker changing misspelled words to different but incorrect words (e.g., *tero* → *to* (correct is *terrorist*), *tittle* → *little* (correct is *title*)).

5 Chunking Experiments

As with the POS-tagging experiments, we performed chunking experiments on learner English. As described in Section 1, we examined the performance of chunking in learner English for the first time. We compared the following three systems: (1) a system using the features presented in Section 3 (**Baseline**), (2) a baseline chunking

system with spell checking (**Base+Checker**), and (3) a baseline chunking system with no spelling errors, i.e., spelling errors were corrected manually (**Base+Gold**). We used POSs that were automatically assigned by the POS-tagger[4] to train our chunking model.

The experimental results on chunking are summarized in Table 4. As can be seen by comparing *Baseline* to *Base+Checker*, there was only a 0.03% difference, which has no statistical significance; thus, the spell checker had nearly no practical effect. Comparing *Baseline* to *Base+Gold*, there was a difference of 0.2% which is statistically significant even though it is only a small difference. Thus, we conclude here that an ideal spell checker has a positive effect on chunking. However, since chunking uses POSs identified by the POS-tagger as its features, it was assumed that POS-tagging errors would directly affect chunking. Table 5 shows the number of correctly identified chunks for misspelled and surrounding words. As with POS-tagging, the number of correctly identified chunks for misspelled words increased, whereas there was nearly no difference in the number of correctly identified chunks for surrounding words.

6 Conclusions

In this paper, we have investigated the performance of POS-tagging and chunking in learner English. The primary cause of failures in POS-tagging and chunking is well known to be unknown words; thus, we focused our investigation on spelling errors, which are the primary sources of unknown words. Furthermore, we have demonstrated the performance of chunking in learner English for, to the best of our knowledge, the first time. From our experiments, we conclude that POS-tagging performance dropped 0.23% due to spelling errors. Furthermore a spell checker is not necessary for POS-tagging, and it is sufficient to assign POS-tags using affix information.

References

Jan Aarts and Sylviane Granger. 1998. Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In Sylviane Granger, editor, *Learner English on Computer*, pages 132–141. Addison Wesley Longman: London and New York.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. In *Proceedings of ACL*, pages 737–746.

Michael Flor, Yoko Futagi, Melissa Lopez, and Matthew Mulholland. 2013. Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus. In *the Second Learner Corpus Research Conference*.

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.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282–289.

Christopher Manning. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *Proceedings of CICLing*, pages 171–189.

Ryo Nagata and Edward Whittaker. 2013. Reconstructing an Indo-European Family Tree from Non-native English Texts. In *Proceedings of ACL*, pages 1137–1147.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a Manually Error-tagged and shallow-parsed corpus. In *Proceedings of ACL-HLT*, pages 1210–1219.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of CoNLL Shared Task*, pages 1–14.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of EMNLP*, pages 133–142.

Keisuke Sakaguchi, Tomoya Mizumoto, Mamoru Komachi, and Yuji Matsumoto. 2012. Joint English Spelling Error Correction and POS Tagging for Language Learners Writing. In *Proceedings of COLING*, pages 2357–2374.

Fei Sha and Fernando Pereira. 2003. Shallow Parsing with Conditional Random Fields. In *Proceedings of HLT-NAACL*, pages 134–141.

Jana Z. Sukkarieh and John Blackmore. 2009. c-rater: Automatic Content Scoring for Short Constructed Responses. In *Proceedings of FLAIRS*, pages 290–295.

[4]The POS-tagger was trained with all features.