A Learner Corpus-based Approach to Verb Suggestion for ESL

Yu Sawai **Mamoru Komachi*** **Yuji Matsumoto**
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
{yu-s, komachi, matsu}@is.naist.jp

Abstract

We propose a verb suggestion method which uses candidate sets and domain adaptation to incorporate error patterns produced by ESL learners. The candidate sets are constructed from a large scale learner corpus to cover various error patterns made by learners. Furthermore, the model is trained using both a native corpus and the learner corpus via a domain adaptation technique. Experiments on two learner corpora show that the candidate sets increase the coverage of error patterns and domain adaptation improves the performance for verb suggestion.

1 Introduction

In this study, we address verb selection errors in the writing of English learners. Selecting the right verb based on the context of a sentence is difficult for the learners of English as a Second Language (ESL). This error type is one of the most common errors in various learner corpora ranging from elementary to proficient levels[1].

> *They **?connect/communicate** with other businessmen and do their jobs with the help of computers.*[2]

This sentence is grammatically acceptable with either verb. However, native speakers of English would less likely use "connect", which means "forming a relationship (with other businessmen)", whereas "communicate" means "exchanging information or ideas", which is what the sentence is trying to convey.

Previous work on verb selection usually treats the task as a multi-class classification problem (Wu et al., 2010; Wang and Hirst, 2010; Liu et al., 2010; Liu et al., 2011). In this formalization, it is important to restrict verbs by a **candidate set** because verb vocabulary is more numerous than other classes, such as determiners. Candidate sets for verb selection are often extracted from thesauri and/or round-trip translations. However, these resources may not cover certain error patterns found in actual learner corpora, and suffer from low-coverage. Furthermore, all the existing classifier models are trained only using a native corpus, which may not be adequate for correcting learner errors.

In this paper, we propose to use error patterns in ESL writing for verb suggestion task by using candidate sets and a domain adaptation technique. First, to increase the coverage, candidate sets are extracted from a large scale learner corpus derived from a language learning website. Second, a domain adaptation technique is applied to the model to fill the gap between two domains: native corpus and ESL corpus. Experiments are carried out on publicly available learner corpora, the Cambridge Learner Corpus First Certificate of English dataset (CLC-FCE) and the Konan JIEM corpus (KJ). The results show that the proposed candidate sets improve the coverage, compared to the baseline candidate sets derived from the WordNet and a round-trip translation table. Domain adaptation also boosts the suggestion performance.

To our knowledge, this is the first work for verb suggestion that uses (1) a learner corpus as a source of candidate sets and (2) the domain adaptation technique to take learner errors into account.

*Now at Tokyo Metropolitan University.

[1]For example, in the CLC-FCE dataset, the replacement error of verbs is the third most common out of 75 error types. In the KJ corpus, lexical choice of verb is the sixth most common out of 47 error types.

[2]This sentence is taken from the CLC-FCE dataset.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 708–713,
Sofia, Bulgaria, August 4-9 2013.

2 Verb Suggestion Considering Error Patterns

The proposed verb suggestion system follows the standard approach in related tasks (Rozovskaya and Roth, 2011; Wu et al., 2010), where the candidate selection is formalized as a multi-class classification problem with predefined candidate sets.

2.1 Candidate Sets

For reflecting tendency of learner errors to the candidate sets, we use a large scale corpus obtained from learners' writing on an SNS (Social Networking Service), *Lang-8*[3]. An advantage of using the learner corpus from such website is the size of annotated portion (Mizumoto et al., 2011). This SNS has over 1 million manually annotated English sentences written by ESL learners. We have collected the learner writings on the site, and released the dataset for research purpose[4].

First, we performed POS tagging for the dataset using the treebank POS tagger in the NLTK toolkit 2.10. Second, we extracted the correction pairs which have "VB*" tag. The set of correction pairs given an incorrect verb is considered as a candidate set for the verb.

We then performed the following preprocessing for the dataset because we focus on lexical selection of verbs:

- Lemmatize verbs to reduce data sparseness.
- Remove non-English verbs using WordNet.
- Remove incorrect verbs which occur only once in the dataset.

The target verbs are limited to the 500 most common verbs in the CLC-FCE corpus[5]. Therefore, verbs that do not appear in the target list are not included in the candidate sets. The topmost 500 verbs cover almost 90 percent of the vocabulary of verbs in the CLC-FCE corpus[6].

The average number of candidates in a set is 20.3[7]. Note that the number of candidates varies across each target verb[8].

[3] http://lang-8.com

[4] Further details can be found at http://cl.naist.jp/nldata/lang-8/. Candidate sets will also be available at the same URL.

[5] They are extracted from all "VB" tagged tokens, and they contain 1,292 unique verbs after removing non-English words.

[6] This number excludes "be".

[7] In this paper, we limit the maximum number of candidates in each set to 50.

[8] For instance, the candidate set for "get" has 315 correction pairs, whereas "refund" has only 4.

2.2 Suggestion Model

The verb suggestion model consists of multi-class classifiers for each target verb; and based on the classifiers' output, it suggests alternative verbs. Instances are in a fill-in-the-blank format, where the labels are verbs. Features in this format are extracted from the surrounding context of a verb. When testing on the learner corpus, the model suggests a ranking of the possible verbs for the blank corresponding to a given context. Note that unlike the fill-in-the-blank task, the candidate sets and domain adaptation can be applied to this task to take the original word into account.

The model is trained on a huge native corpus, namely the ukWaC corpus, because the data-size of learner corpora is limited compared to native corpora. It is then adapted to the target domain, i.e., learner writing. In our experiment, the Lang-8 corpus is used as the target domain corpus, since we assume that it shares the same characteristics with the CLC-FCE and the KJ corpora used for testing.

2.3 Domain Adaptation

To adapt the models to the learner corpus, we employ a domain adaptation technique to emphasize the importance of learner domain information. Although there are many studies on domain adaptation, we chose to use **Feature Augmentation** technique introduced by (Daumé III, 2007) for its simplicity. Recently, (Imamura et al., 2012) proposed to apply this method to grammatical error correction for writings of Japanese learners and confirmed that this is more effective for correcting learner errors than simply adding the target domain instances.

In this study, the source domain is the native writing, and the target domain is the ESL writing. Our motivation is to use the ESL corpus together with the huge native corpus to employ both an advantage of the size of training data and the ESL writing specific features.

In this method, adapting a model to another model is achieved by extending the feature space. Given a feature vector of F dimensions as $\boldsymbol{x} \in \mathbb{R}^F (F > 0)$, using simple mapping, the augmented feature vectors for source and target domains are obtained as follows,

$$\text{Source domain: } < \boldsymbol{x}_S, \boldsymbol{x}_S, \boldsymbol{0} > \quad (1)$$

$$\text{Target domain: } < \boldsymbol{x}_T, \boldsymbol{0}, \boldsymbol{x}_T > \quad (2)$$

where **0** denotes a zero-vector of F dimensions. The three partitions mean a common, a source-specific, and a target-specific feature space. When testing on the ESL corpora, the target-specific features are emphasized.

2.4 Features

In previous work, various features were used: lexical and POS n-grams, dependencies, and arguments in the verb context. (Liu et al., 2011) has shown that shallow parse features, such as lexical n-grams and chunks, work well in realistic settings, in which the input sentence may not be correctly parsed. Considering this, we use shallow parse features as context features for robustness.

The features include lexical and POS n-grams, and lexical head words of the nearest NPs, and clustering features of these head words. An example of extracted features is shown in Table 2.4. Note that those features are also used when extracting examples from the target domain dataset (the learner domain corpus). As shown in Table 2.4, the n-gram features are 3-gram and extracted from ± 2 context window. The nearest NP's head features are divided into two (Left, Right).

The additional clustering features are used for reducing sparseness, because the NP's head words are usually proper nouns. To create the word clusters, we employ Brown clustering, a hierarchical clustering algorithm proposed by (Brown et al., 1992). The structure of clusters is a complete binary tree, in which each node is represented as a bit-string. By varying the length of the prefix of bit-string, it is possible to change the granularity of cluster representation. As illustrated in Table 2.4, we use the clustering features with three levels of granularity: 256, 128, and 64 dimensions. We used Percy Liang's implementation[9] to create 256 dimensional model from the ukWaC corpus, which is used as the native corpus.

Feature	Example
n-grams (surface)	they-*V*-with <S>-they-*V* *V*-with-other
n-grams (POS)	PRP-*V*-IN <S>-PRP-*V* *V*-IN-JJ
NP head (Left, Right)	L_they, L_PRP R_businessmen, R_NNS
NP head cluster (Left, Right)	L_01110001, L_0111000, L_011100 R_11011001, R_1101100, R_110110

(e.g., *They (communicate) with other businessmen and do their jobs with the help of computers.*)
"<S>" denotes the beginning of the sentence, "*V*" denotes the blanked out verb.

Table 1: Example of extracted features as the fill-in-the-blank form.

3 Experiments

Performance of verb suggestion is evaluated on two error-tagged learner corpora: CLC-FCE and KJ. In the experiments, we assume that the target verb and its context for suggestion are already given.

For the experiment on the CLC-FCE dataset, the targets are all words tagged with "RV" (replacement error of verbs). We assume that all the verb selection errors are covered with this error tag. All error tagged parts with nested correction or multi-word expressions are excluded. The resulting number of "true" targets is 1,083, which amounts to 4% of all verbs. Therefore the dataset is highly skewed to correct usages, though this setting expresses well the reality of ESL writing, as shown in (Chodorow et al., 2012).

We carried out experiments with a variety of resources used for creating candidate sets.

- **WordNet**
  Candidates are retrieved from the synsets and verbs sharing the same hypernyms in the WordNet 3.0.
- **LearnerSmall**
  Candidates are retrieved from following learner corpora: NUS corpus of learner English (NUCLE), Konan-JIEM (KJ), and NICT Japanese learner English (JLE) corpus.
- **Roundtrip**
  Candidates are collected by performing "round-trip" translation, which is similar to (Bannard and Callison-Burch, 2005) [10].
- **WordNet+Roundtrip**
  A combination of the thesaurus-based and the translation table-based candidate sets, similar to (Liu et al., 2010) and (Liu et al., 2011).
- **Lang-8**
  The proposed candidate sets obtained from a large scale learner corpus.
- **Lang-8+DA**
  Lang-8 candidate sets with domain adapta-

[9] https://github.com/percyliang/brown-cluster

[10] Our roundtrip translation lexicons are built using a subset of the WIT^3 corpus (Cettolo et al., 2012), which is available at http://wit3.fbk.eu.

Settings	Candidates/set (Avg.)
WordNet	14.8
LearnerSmall	5.1
Roundtrip	50
Roundtrip (En-Ja-En)	50
WordNet+Roundtrip	50
Lang-8	20.3

Table 2: Comparison of candidate set size for each setting.

tion via feature augmentation.

Table 3 shows a comparison of the average number of candidates in each setting. In all configurations above, the parameters of the models underlying the system are identical. We used a L2-regularized generalized linear model with log-loss function via Scikit-learn ver. 0.13.

Inter-corpus Evaluation

We also evaluate the suggestion performance on the KJ corpus. The corpus contains diary-style writing by Japanese university students. The proficiency of the learners ranges from elementary to intermediate, so it is lower than that of the CLC-FCE learners. The targets are all verbs tagged with "v_lxc" (lexical selection error of verbs).

To see the effect of L1 on the verb suggestion task, we added an alternative setting for the Roundtrip using only the English-Japanese and Japanese-English round-trip translation tables (En-Ja-En). For the experiment on this test-corpus, the LearnerSmall is not included.

Datasets

The ukWaC web-corpus (Ferraresi et al., 2008) is used as a native corpus for training the suggestion model. Although this corpus consists of over 40 million sentences, 20,000 randomly selected sentences are used for each verb[11].

The Lang-8 learner corpus is used for domain adaptation of the model in the Lang-8+DA configuration. The portion of data is the same as that used for constructing candidate sets.

Metrics

Mean Reciprocal Rank (MRR) is used for evaluating the performance of alternative suggestions. The mean reciprocal rank is calculated by taking the average of the reciprocal ranks for each instance. Given r_gold_i as the position of the gold correction candidate in the suggestion list S_i for i-th checkpoint, the reciprocal rank RR_i is defined as,

$$RR_i = \begin{cases} \frac{1}{r\_gold_i} & (gold_i \in S_i) \\ 0 & (\text{otherwise}) \end{cases} \quad (3)$$

[11] e.g., a classifier with a candidate set containing 50 verbs is trained with 1 million sentences in total.

4 Results

Tables 5 and 5 show the results of suggestion performance on the CLC-FCE dataset and the KJ corpus, respectively. In both cases, the Lang-8 and its domain adaptation variant outperformed the others. The coverage of error patterns in the tables is the percentage of the cases where the suggestion list includes the gold correction. Generally, the suggestion performance and the coverage improve as the size of the candidate sets increases.

5 Discussions

Although the expert-annotated learner corpora contain candidates which are more reliable than a web-crawled Lang-8 corpus, the Lang-8 setting performed better as shown in Table 5. This can be explained by the broader coverage by the Lang-8 candidate sets than that of the LearnerSmall. Similarly, the WordNet performed the worst because it contains only synonym-like candidates. We can conclude that, for the verb suggestion task, the coverage (recall) of candidate sets is more important than the quality (precision).

We see little influence of learners' L1 in the results of Table 5, since the Roundtrip performed better than the Roundtrip (En-Ja-En). As already mentioned, the number of error patterns contained in the candidate sets seems to have more importance than the quality.

As shown in Tables 5 and 5, a positive effect of domain adaptation technique appeared in both test-corpora. In the case of the CLC-FCE, 280 out of 624 suggestions were improved compared to the setting without domain adaptation. For instance, confusions between synonyms such as "**?live/stay**", "**?say/tell**", and "**?solve/resolve**" are improved, because sentences containing these confusions appear more frequently in the Lang-8 corpus. Although the number of test-cases for the KJ corpus is smaller than the CLC-FCE, we can see the improvements for 33 out of 66 sug-

Settings	MRR	Coverage
WordNet	0.066	14.0 %
LearnerSmall	0.128	23.5 %
Roundtrip	0.185	48.1 %
WordNet+Roundtrip	0.173	48.1 %
Lang-8	0.220	57.6 %
Lang-8+DA	**0.269***	57.6 %

The value marked with the asterisk indicates statistically significant improvement over the baselines, where $p < 0.05$ bootstrap test.

Table 3: Suggestion performance on the CLC-FCE dataset.

Settings	MRR	Coverage
WordNet	0.044	5.0 %
Roundtrip	0.241	53.8 %
Roundtrip (En-Ja-En)	0.188	38.8 %
WordNet+Roundtrip	0.162	53.8 %
Lang-8	0.253	68.9 %
Lang-8+DA	**0.412***	68.9 %

The value marked with the asterisk indicates statistically significant improvement over the baselines, except "Roundtrip", where $p < 0.05$ bootstrap test.

Table 4: Suggestion performance on the KJ corpus.

gestions. The improvements appeared for frequent confusions of Japanese ESL learners such as "**?see/watch**" and "**?tell/teach**".

Comparing the results of the Lang-8+DA on both test-corpora, the domain adaptation technique worked more effectively on the KJ corpus than on the CLC-FCE. This can be explained by the fact that the style of writing of the additional data, i.e., the Lang-8 corpus, is closer to KJ than it is to CLC-FCE. More precisely, unlike the examination-type writing style of CLC-FCE, the KJ corpus consists of diary writing similar in style to the Lang-8 corpus, and it expresses more closely the proficiency of the learners.

We think that the next step is to refine the suggestion models, since we currently take a simple fill-in-the-blank approach. As future work, we plan to extend the models as follows: (1) use both incorrect and correct sentences in learner corpora for training, and (2) employ ESL writing specific features such as learners' L1 for domain adaptation.

Acknowledgments

We thank YangYang Xi of Lang-8, Inc. for kindly allowing us to use the Lang-8 learner corpus. We also thank the anonymous reviewers for their insightful comments. This work was partially supported by Microsoft Research CORE Project.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 597–604.

Peter F Brown, Vincent J Della Pietra, Peter V DeSouza, Jenifer C Lai, Robert L Mercer, and Vincent J Della Pietra. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479, December.

M Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 28–30.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in Evaluating Grammatical Error Detection Systems. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling2012)*, pages 611–628.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Adriano Ferraresi, Eros Zanchetta, and Marco Baroni. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*, pages 45–54.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 388–392.

Xiaohua Liu, Bo Han, Kuan Li, Stephan Hyeonjun Stiller, and Ming Zhou. 2010. SRL-based verb selection for ESL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1076.

Xiaohua Liu, Bo Han, and Ming Zhou. 2011. Correcting verb selection errors for ESL with the perceptron. In *12th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 411–423.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 147–155.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 924–933.

Tong Wang and Graeme Hirst. 2010. Near-synonym lexical choice in latent semantic space. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1182–1190.

Jian-Cheng Wu, Yu-Chia Chang, Teruko Mitamura, and Jason S Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics Short Papers*, pages 115–119.