A Meta Learning Approach to Grammatical Error Correction

Hongsuck Seo[1], Jonghoon Lee[1], Seokhwan Kim[2], Kyusong Lee[1]
Sechun Kang[1], Gary Geunbae Lee[1]
[1]Pohang University of Science and Technology
[2]Institute for Infocomm Research
{hsseo, jh21983}@postech.ac.kr, kims@i2r.a-star.edu.sg
{kyusonglee, freshboy, gblee}@postech.ac.kr

Abstract

We introduce a novel method for grammatical error correction with a number of small corpora. To make the best use of several corpora with different characteristics, we employ a meta-learning with several base classifiers trained on different corpora. This research focuses on a grammatical error correction task for article errors. A series of experiments is presented to show the effectiveness of the proposed approach on two different grammatical error tagged corpora.

1. Introduction

As language learning has drawn significant attention in the community, grammatical error correction (GEC), consequently, has attracted a fair amount of attention. Several organizations have built diverse resources including grammatical error (GE) tagged corpora.

Although there are some publicly released GE tagged corpora, it is still challenging to train a good GEC model due to the lack of large GE tagged learner corpus. The available GE tagged corpora are mostly small datasets having different characteristics depending on the development methods, e.g. spoken corpus vs. written corpus. This situation forced researchers to utilize native corpora rather than GE tagged learner corpora for the GEC task.

The native corpus approach consists of learning a model that predicts the correct form of an article given the surrounding context. Some researchers focused on mining better features from the linguistic and pedagogic knowledge, whereas others focused on testing different classification methods (Knight and Chandler, 1994; Minnen et al., 2000; Lee, 2004; Nagata et al., 2006; Han et al., 2006; De Felice, 2008).

Recently, a group of researchers introduced methods utilizing a GE tagged learner corpus to derive more accurate results (Han et al., 2010; Rozovskaya and Roth, 2010). Since the two approaches are closely related to each other, they can be informative to each other. For example, Dahlmeier and Ng (2011) proposed a method that combines a native corpus and a GE tagged learner corpus and it outperformed models trained with either a native or GE tagged learner corpus alone. However, methods which train a GEC model from various GE tagged corpora have received less focus.

In this paper, we present a novel approach to the GEC task using meta-learning. We focus mainly on article errors for two reasons. First, articles are one of the most significant sources of GE for the learners with various L1 backgrounds. Second, the effective features for article error correction are already well engineered allowing for quick analysis of the method. Our approach is distinguished from others by integrating the predictive models trained on several GE tagged learner corpora, rather than just one GE tagged corpus. Moreover, the framework is compatible to any classification technique. In this study, we also use a native corpus employing Dahlmeier and Ng's approach. We demonstrate the effectiveness of the proposed method against baseline models in article error correction tasks.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 328–332,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

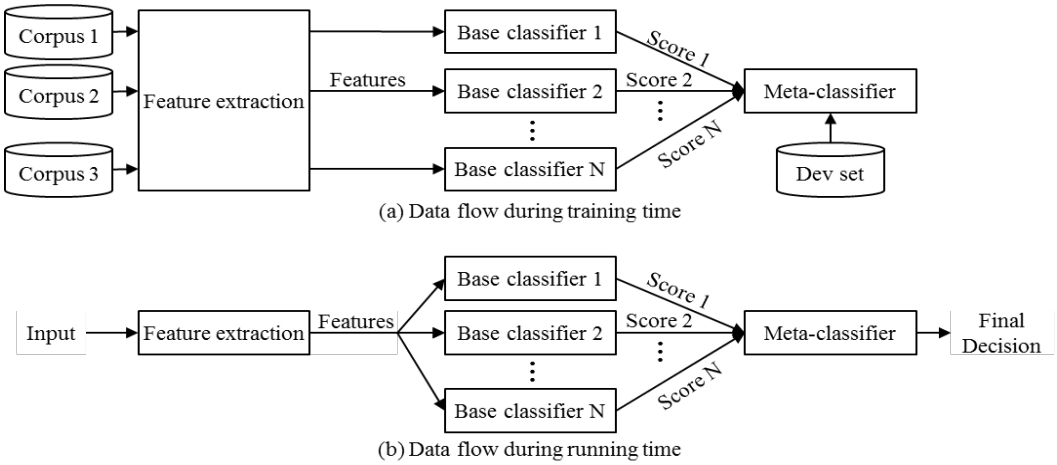

Figure 1: Overview of the proposed method

The remainder of this paper is organized as follows: Section 2 explains our proposed method. The experiments are presented in Section 3. Finally, Section 4 concludes the paper.

2. Method

Our method predicts the type of article for a noun phrase within three classes: *null*, *definite*, and *indefinite*. A correction arises when the prediction disagrees with the observed article. The meta-learning technique is applied to this task to deal with multiple corpora obtained from different sources.

A meta-classifier decides the final output based on the intermediate results obtained from several base classifiers. Each base classifier is trained on a different corpus than are the other classifiers. In this work, the feature extraction processes used for the base classifiers are identical to each other for simplicity, although they need not necessarily be identical. The meta-classifier takes the output scores of the base classifiers as its input and is trained on the held-out development data (Figure 1a). During run time, the trained classifiers are organized in the same manner. For the given features, the base classifiers independently calculate the score, then the meta-classifier makes the final decision based on the scores (Figure 1b).

2.1. Meta-learning

Meta-learning is a sequential learning process following the output of other base learners (classifiers). Normally, different classifiers successfully predict results on different parts of the input space, so researchers have often tried to combine different classifiers together (Breiman, 1996; Cohen et al., 2007; Zhang, 2007; Aydın, 2009; Menahem et al., 2009). To capitalize on the strengths and compensate for the weaknesses of each classifier, we build a meta-learner that takes an input vector consisting of the outputs of the base classifiers. The performance of meta-learning can be improved using output probabilities for every class label from the base classifiers.

The meta-classifier for the proposed method consists of multiple linear classifiers. Each classifier takes an input vector consisting of the output scores of each base classifier and calculates a score for each type of article. The meta-classifier finally takes the class having the maximum score.

A common design of an ensemble is to train different base classifiers with the same dataset, but in this work one classification technique was used with different datasets each having different characteristics. Although only one classification method was used in this work, different methods each well-tuned to the individual corpora may be used to improve the performance.

We employed the meta-learning method to generate synergy among corpora with diverse characteristics. More specifically, it is shown by cross validation that meta-learning performs at a level that is comparable to the best base classifier (Dzeroski and Zenko, 2004).

2.2. Base Classifiers

In the meta-learning framework, the performance of the base classifiers is important because the improvement in base classification generally enha-

nces the overall performance. The base classifiers can be expected to become more informative as more data are provided. We followed the structural learning approach (Ando and Zhang, 2005), which trains a model from both a native corpus and a GE tagged corpus (Dahlmeire and Ng, 2011), to improve the base classifiers by the additional information extracted from a native corpus.

Structural learning is a technique which trains multiple classifiers with common structure. The common structure chooses the hypothesis space of each individual classifier and the individual classifiers are trained separately once the hypothesis space is determined. The common structure can be obtained from auxiliary problems which are closely related to the main problems.

A word selection problem is a task to predict the appropriate word given the surrounding context in a native corpus and is a closely related auxiliary problem of the GEC task. We can obtain the common structure from the article selection problem and use it for the correction problem.

In this work, all the base classifiers used the same least squares loss function for structural learning. We adopted the feature set investigated in De Felice (2008) for article error correction. We use the Stanford coreNLP toolkit[1] (Toutanova and Manning, 2000; Klein and Manning, 2003a; Klein and Manning, 2003b; Finkel et al, 2005) to extract the features.

2.3. Evaluation Metric

The effectiveness of the proposed method is evaluated in terms of accuracy, precision, recall, and F_1-score (Dahlmeire and Ng, 2011). Accuracy is the number of correct predictions divided by the total number of instances. Precision is the ratio of the suggested corrections that agree with the tagged answer to the total number of the suggested corrections whereas recall is the ratio of the suggested corrections that agree with the tagged answer to the total number of corrections in the corpus.

3. Experiments

3.1. Datasets

In this work we used a native corpus and two GE tagged corpora. For the native corpus, we used news data[2] which is a large English text extracted from news articles. The First Certificate in English exams in the Cambridge Learner Corpus [3] (hereafter, CLC-FCE; Yannakoudakis et al., 2011) and the Japanese Learner English corpus (Izumi et. al., 2005) were used for the GE tagged corpora.

We extracted noun phrases from each corpus by parsing the text of the respective corpora. (1) We parsed the native corpus from the beginning until approximately a million noun phrases are extracted. (2) About 90k noun phrases containing ~3,300 mistakes in article usage were extracted from the entire CLC-FCE corpus, and (3) about 30k noun phrases containing ~2,500 mistakes were extracted from the JLE corpus.

The extracted noun phrases were used for our training and test data. We hold out 10% of the data for the test. We applied 20% under-sampling to the training instances that do not have any errors to alleviate data imbalance in the training set.

We emphasize the fact that the two learner corpora differ from each other in three aspects. The first aspect is the styles of the texts: the CLC is literary whereas the JLE is colloquial. The second is the error rate: about 3.5% for CLC-FCE and 8.5% for JLE. Finally, the third is the distribution of L1 languages of the learners: the learners of the CLC corpus have various L1 backgrounds whereas the learners of the JLE consist of only Japanese. These experiments demonstrate the effectiveness of the proposed method relying on the diversity of the corpora.

The native corpus was used to find the common structure using structural learning and two GE tagged learner corpora are used to train the base classifiers by structural learning with the common structure obtained from the news corpus.

We trained three classifiers for comparison; (1) the classifier (INTEG) trained with the integrated training set of the two GE tagged corpora, and two base classifiers used for the ensemble: (2) the base classifier (CB) trained only with the CLC-FCE and (3) the other base classifier (JB) trained with the JLE.

3.2. Results

The accuracy obtained from the word selection task with the news corpus was 76.10%. Upon

[1] http://nlp.stanford.edu/software/corenlp.shtml

[2] http://www.statmt.org/wmt09/translation-task.html

[3] http://www.ilexir.com/

Model	Acc.	Prec.	Rec.	F_1
INTEG	73.37	4.69	**72.39**	8.82
CB	77.20	5.39	71.17	10.03
Proposed	**86.99**	**6.17**	45.77	**10.88**

Table 1: Best results for GEC task on CLC-FCE test set.

Model	Acc.	Prec.	Rec.	F_1
INTEG	78.87	14.88	85.47	25.35
JB	78.02	14.49	**86.32**	24.82
Proposed	**89.61**	**19.28**	46.60	**27.27**

Table 2: Best results for GEC task on JLE test set.

Model	Acc.	Prec.	Rec.	F_1
INTEG	74.64	6.84	**77.86**	12.58
Proposed	**87.50**	**8.61**	46.12	**14.52**

Table 3: Best results for GEC task on the integrated set of CLC-FCE and JLE test sets.

obtaining the parameters of the word selection task, the structural parameter Θ was calculated by singular value decomposition and was used for the structural learning of the main GEC task.

We used three different test data sets: the CLC-FCE, the JLE and an integrated test set of the two. The accuracy (Acc.) and the precision (Prec.) of the INTEG was poorer than CB on the CLC-FCE test set (Table 1), whereas INTEG outperformed JB on the JLE test (Table 2).

Some instances extracted from the CLC-FCE corpus have similar characteristics to the instances from the JLE corpus. This overlap of instances affected the performance in both positive and negative ways. Prediction of instances similar to those in the JLE was enhanced. Consequently, INTEG model demonstrated better accuracy and precision for the JLE test set. Unfortunately, for the CLC test set, the instances resulted in lower accuracy and precision.

The proposed model is able to alleviate this model bias due to similar instances observed in the INTEG model. The accuracy of the proposed model consistently increased by over 10% for all three data sets. The relative performance gain in terms of F1-score (F_1) was 15% on the integrated set. This performance gain stems from the over 25% relative improvement of the precision (Table 1, 2 and 3).

We believe the improvement comes from the contribution of reconfirming procedures performed by the meta-classifier. When the prediction of the two base classifiers conflicts with each other, the meta-classifier tends to choose the one with a higher confidence score; this choice improves the accuracy and precision because known features generate a higher confidence whereas unseen or less-weighted features generate a lower score.

Although the proposed model introduced a tradeoff between precision and recall (Rec.), this tradeoff was tolerable in order to improve the overall F1-score. Since GEC is a task where false alarm is critical, obtaining high precision is very important. The low precision on the whole experiments is due to the data imbalance. Instances in the dataset are mostly not erroneous, e.g., only 3.5% of erroneous instances for the CLC corpus. The standard for correct prediction is also very strict and does not allow multiple answers. Performance can be evaluated in a more realistic way by applying a softer standard, e.g., by evaluating manually.

4. Conclusion

We have presented a novel approach to grammatical error correction by building a meta-classifier using multiple GE tagged corpora with different characteristics in various aspects. The experiments showed that building a meta-classifier overcomes the interference that occurs when training with a set of heterogeneous corpora. The proposed method also outperforms the base classifier themselves tested on the same class of test set as the training set with which the base classifiers are trained. A better automatic evaluation metric would be needed as further research.

Acknowledgments

Industrial Strategic technology development program, 10035252, development of dialog-based spontaneous speech interface technology on mobile platform, funded by the Ministry of Knowledge Economy (MKE, Korea).

References

R.K. Ando and T. Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6, pp. 1817-1853.

U. Aydın, S. Murat, Olcay T Yıldız, A. Ethem, 2009, Incremental construction of classifier and discriminant ensembles, *Information Science*, 179 (9), pp. 144-152.

L. Breiman, 1996, Bagging predictors, *Machine Learning*, pp. 123–140.

S. Cohen, L. Rokach, O. Maimon, 2007, Decision tree instance space decomposition with grouped gain-ratio, *Information Science*, 177 (17), pp. 3592–3612.

D. Dahlmeier, H. T. Ng, 2011, Grammatical error correction with alternating structure optimization, In *Proceedings of the 49th Annual Meeting of the ACL-HLT 2011,* pp. 915-923.

R. De Felice. 2008. *Automatic Error Detection in Non-native English*. Ph.D. thesis, University of Oxford.

S. Dzeroski, B. Zenko, 2004, Is combining classifiers with stacking better than selecting the best one?, *Machine Learning*, 54 (3), pp. 255–273.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the ACL*, pp. 363-370.

N.R. Han, M. Chodorow, and C. Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(02), pp. 115-129.

N.R. Han, J. Tetreault, S.H. Lee, and J.Y. Ha. 2010. Using an error-annotated learner corpus to develop an ESL/EFL error correction system. In *Proceedings of LREC*.

D. Klein and C.D. Manning. 2003a. Accurate unlexicalized parsing. In *Proceedings of ACL*, pp. 423-430.

D. Klein and C.D. Manning. 2003b. Fast exact inference with a factored model for natural language processing. *Advances in Neural Information Processing Systems (NIPS 2002)*, 15, pp. 3-10.

K. Knight and I. Chander. 1994. Automated postediting of documents. In *Proceedings of AAAI*, pp. 779-784.

J. Lee. 2004. Automatic article restoration. In *Proceedings of HLT-NAACL*, pp. 31-36.

R. Nagata, A. Kawai, K. Morihiro, and N. Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of COLING-ACL*, pp. 241--248.

A. Mariko, 2007, Grammatical errors across proficiency levels in L2 spoken and written English, *The Economic Journal of Takasaki City University of Economics*, 49 (3, 4), pp. 117-129.

E. Menahem, L. Rokach, Y. Elovici, 2009, Troika-An imporoved stacking schema for classification tasks, *Information Science*, 179 (24), pp. 4097-4122.

G. Minnen, F. Bond, and A. Copestake. 2000. Memory-based learning for article generation. In *Proceedings of CoNLL*, pp. 43-48.

E. Izumi, K. Uchimoto, H. Isahara, 2005, Error annotation for corpus of Japanese learner English, In *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora*, pp. 71-80.

A. Rozovskaya and D. Roth. 2010. Training paradigms for correcting errors in grammar and usage. In *Proceedings of HLT-NAACL*, pp. 154-162.

K. Toutanova and C. D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *In Proceedings of the Joint SIGDAT Conference on EMNLP/VLC-2000*, pp. 63-70.

H.Yannakoudakis, T. Briscoe, B. Medlock, 2011, A new dataset and method for automatically grading ESOL texts, In *Proceedings of ACL*, pp. 180-189.

G. P. Zhang, 2007, A neural network ensemble method with jittered training data for time series forecasting, *Information Sciences: An International Journal*, 177 (23), pp. 5329–5346.