Evaluating performance of grammatical error detection to maximize learning effect

Ryo Nagata
Konan University
rnagata @ konan-u.ac.jp.

Kazuhide Nakatani
Konan University

Abstract

This paper proposes a method for evaluating grammatical error detection methods to maximize the learning effect obtained by grammatical error detection. To achieve this, this paper sets out the following two hypotheses — imperfect, rather than perfect, error detection maximizes learning effect; and precision-oriented error detection is better than a recall-oriented one in terms of learning effect. Experiments reveal that (i) precision-oriented error detection has a learning effect comparable to that of feedback by a human tutor, although the first hypothesis is not supported; (ii) precision-oriented error detection is better than recall-oriented in terms of learning effect; (iii) F-measure is not always the best way of evaluating error detection methods.

1 Introduction

To reduce the efforts taken to correct grammatical errors in English writing, there has been a great deal of work on grammatical error detection (Brockett et al., 2006; Chodorow and Leacock, 2000; Chodorow and Leacock, 2002; Han et al., 2004; Han et al., 2006; Izumi et al., 2003; Nagata et al., 2004; Nagata et al., 2005; Nagata et al., 2006). One of its promising applications is writing learning assistance by detecting errors and showing the results to the learner as feedback that he or she can use to rewrite his or her essay. Grammatical error detection has greatly improved in detection performance as well as in the types of the errors it is able to detect, including errors in articles, number, prepositions, and agreement.

In view of writing learning assistance, however, one important factor has been missing in the previous work. In the application to writing learning assistance, error detection methods should be evaluated by learning effect obtained by error detection. Nevertheless, they have been evaluated only by detection performance such as F-measure.

This brings up a further research question — are any of the previous methods effective as writing learning assistance? It is very important to answer this question because it is almost impossible to develop a perfect method. In other words, one has to use an imperfect method to assist learners no matter how much improvement is achieved. In practice, it is crucial to reveal the lower bound of detection performance that has a learning effect.

Related to this, one should discuss the following question. Most error detection methods are adjustable to be recall-oriented/precision-oriented by tuning their parameters. Despite this fact, no one has examined which is better in terms of learning effect — recall-oriented or precision-oriented? (hereafter, this problem will be referred to as the *recall-precision problem*). Chodorow and Leacock (2000) and Chodorow et al. (2007) argue that precision-oriented is better, but they do not give any concrete reason. This means that the recall-precision problem has not yet been solved.

Accordingly, this paper explores the relation between detection performance and learning effect. To do this, this paper sets out two hypotheses:

Hypothesis I : imperfect, rather than perfect, error detection maximizes learning effect

Hypothesis II : precision-oriented is better than recall-oriented in terms of learning effect

Hypothesis I contradicts the intuition that the better the detection performance is, the higher the learning effect is. To see the motivation for this,

Coling 2010: Poster Volume, pages 894–900,
Beijing, August 2010

suppose that we had a perfect method. It would detect all errors in a given essay with no false-positives. In that case, the learner would not have to find any errors by himself or herself. Neither would he or she have to examine the causes of the errors. In the worst case, they just copy the detection results. By contrast, with an imperfect method, he or she has to do these activities, which is expected to result in better learning effect. Besides, researchers, including Robb et al. (1986), Bitchener et al. (2005), and Ferris and Roberts (2001), report that the amount of feedback that learners receive does not necessarily correspond to the amount of learning effect. For instance, Robb et al. (1986) compared four types of feedback ((1) error detection and correction, (2) error detection and error type, (3) error detection, and (4) number of errors per line) and reported that (1), the most-detailed feedback, did not necessarily have the highest learning effect.

Hypothesis II concerns the recall-precision problem. If a limited number of errors are detected with high precision (i.e., precision-oriented), learners have to carefully read their own essay to find the rest of the errors by examining whether their writing is correct or not, using several sources of information including (i) the information that can be obtained from the detected errors, which is useful for finding undetected errors similar to the detected ones; (ii) their knowledge on English grammar and writing, and (iii) dictionaries and textbooks. We believe that learning activities, especially learning from similar instances, have a favorable learning effect. By contrast, in a recall-oriented setting, these activities relatively decrease. Instead, learners focus on judging whether given detection results are correct or not. Besides, learning from similar instances is likely not to work well because a recall-oriented setting frequently makes false-positives.

This paper proposes a method for testing the two hypotheses in Sect. 2. It conducts experiments based on the method in Sect. 3. It discusses the experimental results in Sect. 4.

2 Method

We conducted a pre-experiment where ten subjects participated and wrote 5.6 essays on average. We used the obtained data to design the method.

2.1 Target Errors

To obtain general conclusions, one has to test **Hypothesis I** and **Hypothesis II** against a variety of errors and also a variety of error detection methods. However, it would not be reasonable or feasible to do this from the beginning.

Considering this, this paper targets errors in articles and number. The reasons for selecting these are that (a) articles and number are difficult for learners of English (Izumi et al., 2003; Nagata et al., 2005), and (b) there has been a great deal of work on the detection of these errors.

2.2 Error detection method

Among the previous methods for detecting errors in articles and number, this paper selects Nagata et al. (2006)'s method that detects errors in articles and number based on countability prediction. It has been shown to be effective in the detection of errors in articles and number (Nagata et al., 2005; Nagata et al., 2006). It also has the favorable property that it can be adjusted to be recall-oriented or precision-oriented by setting a threshold for the probability used in countability prediction. This subsection briefly describes Nagata et al. (2006)'s method (See Nagata et al. (2006) for the details).

The method, first, automatically generates training instances for countability prediction. Instances of each noun that head their noun phrase (NP) are collected from a corpus with their surrounding words. Then, the collected instances are tagged with their countability by a set of hand-coded rules. The resulting tagged instances are used as training data for countability prediction.

Decision lists (Yarowsky, 1995) are used to predict countability. Tree types of contextual cue are used as features: (i) words in the NP that the target noun heads; (ii) three words to the left of the NP; (iii) three words to its right. The log-likelihood ratio (Yarowsky, 1995) decides in which order rules in a decision list are applied to the target noun in countability prediction. It is the log ratio of the probabilities of the target noun being count and non-count when one of the features appears in its context. To predict countability in error detection, each rule in the decision list is tested on the target noun in the sorted order until the first applicable one is found. The prediction is made by the first applicable one.

After countability prediction, errors in articles and number are detected by using a set of rules. For example, if the noun in question is plural and predicted to be non-count, then it is an error. Similarly, the noun in question has no article and is singular and is predicted to be count, then it is an error.

The balance of recall and precision in error detection can be adjusted by setting a certain threshold to the probabilities used to calculate the log-likelihood ratio[1]. If the probability of the applied rule in countability prediction is lower than a certain threshed, error detection is blocked. Namely, the higher the threshed is, the more precision-oriented the detection is.

2.3 Learning Activity

The proposed method is based on a learning activity consisting of essay writing, error detection, and rewriting. Table 1 shows the flow of the learning activity. In Step 1, an essay topic is assigned to learners. In Step 2, they have time to think about what to write with a piece of white paper for preparation (e.g., to summarize his or her ideas). In Step 3, they write an essay on a blog system in which the error detection method (Nagata et al., 2005) is implemented. This system allows them to write, submit, and rewrite their essays (though it does not allow them to access the others' essays or their own previous essays). They are not allowed to use any dictionary or textbook in this step. They are required to write ten sentences or more. In Step 4, the system detects errors in each essay. It displays each essay of which errors are indicated in red to the corresponding learner. Although the detection itself takes only a few seconds, five minutes are assigned to this step for two purposes: to take a short break for learners and to remove time differences between learners. Finally, in Step 5, learners rewrite their essay using the given feedback. Here, they are allowed to use a dictionary (Konishi and Minamide, 2007) and an A4 paper that briefly explains article and number usage, which was made based on grammar books (Hirota, 1992; Iizuka and Hagino, 1997). They are informed that the feedback may contain false-positives and false-negatives.

Table 1: Flow of learning activity

Procedure	Min
1. Learner is assigned an essay topic	–
2. Learner prepares for writing	5
3. Learner writes an essay	35
4. System detects errors in the essay	5
5. Learner rewrites the essay	15

2.4 How to Measure Learning Effect

Before discussing how to measure learning effect, one has to define the ability to write English. Considering that this paper aims at the evaluation of error detection, it is reasonable to define the ability as the degree of error occurrence (that is, the fewer errors, the better). To measure this, this paper uses error rate, which is defined by

$$e = \frac{\text{Number of target errors in Step 3} + 1}{\text{Number of NPs in Step 3} + 1}. \quad (1)$$

Ones ("+1") are added to the numerator and denominator for a mathematical reason that will be clear shortly. The addition also has the advantage that it can evaluate a longer essay to be better when no errors occur.

Having defined ability, it is natural to measure learning effect by a decrease in the error rate. Simply, it is estimated by applying the linear regression to the number of instances of learning and the corresponding error rates.

Having said this, this paper applies an exponential regression instead of the linear regression. There are two reasons for this. The first is that it becomes more difficult to decrease the error rate as it decreases (in other words, it becomes more difficulty to improve one's ability as one improves). The other is that the error rate is expected to asymptotically decrease to zero as learning proceeds. The exponential regression is defined by

$$e = \exp\{a(t + b)\} \quad (2)$$

[1] Setting a threshold to the probability is equivalent to setting a threshold to the log-likelihood and both has the same effect on the balance of recall and precision. However, we use the former because it is intuitive and easy to set a threshold

where t, a, and b denote the number of instances of learning, decrease in the error rate (learning effect), and the ability before the learning starts, respectively. The parameters a and b can be estimated from experimental data by least squares.

To examine **Hypothesis I** and **Hypothesis II**, the learning effect parameter a must be estimated for several error detection conditions. To do this, detection performance (recall, precision, and F-measure) is first defined. Recall and precision is defined by

$$r = \frac{\text{Number of errors correctly detected}}{\text{Number of errors}} \tag{3}$$

and

$$p = \frac{\text{Number of errors correctly detected}}{\text{Number of errors detected}}, \tag{4}$$

respectively. Using recall and precision, F-measure is defined by

$$f = \frac{2rp}{r+p}. \tag{5}$$

With these, this paper compares four conditions. In the first condition, the system detects no error at all. Thus, it plays a role as a baseline. The second and third conditions are recall-oriented and precision-oriented, respectively. The threshold that maximized F-measure, which was 0.60, was computed by applying the error detection method to the essays obtained in the pre-experiment (increasing the threshold from 0 to 1, 0.05 at a time). This was selected as the recall-oriented condition. Then, the threshold for the precision-oriented condition was determined to be 0.90 so that its precision became higher. The final condition corresponds to the perfect error detection. Because it was impossible to implement such error detection, a native speaker of English took this part. Hereafter, the four conditions will be referred to as **No-feedback**, **Recall-oriented**, **Precision-oriented**, and **Human**.

3 Experiments

As subjects, 26 Japanese college students (first to fourth grade) participated in the experiments. These 26 subjects were assigned to each condition as follows: **Human**: 6; **Recall-oriented**: 7; **Precision-oriented**: 7; **No-feedback** 6:.

Table 2: Essay topics used in the experiments

No.	Topic
1	University life
2	Summer vacation
3	Gardening
4	My hobby
5	My frightening experience
6	Reading
7	My home town
8	Traveling
9	My favorite thing
10	Cooking

The number of learning activities was ten. Essay Topics for each learning activity is shown in Table 2 They were selected based on a writing textbook (Okihara, 1985). The experiments were conducted from Oct. 2008 to Dec. 2008. The subjects basically did the learning activity twice a week on average. Some of them could not finish the ten-essays assignment during this term. Subjects who did not do the learning activity eight or more times were excluded from the experiments. As a result, 22 subjects were valid in the end (**Human**: 4; **Recall-oriented**: 7; **Precision-oriented**: 6; **No-feedback**: 5).

Figure 1 shows the experimental results. It shows the plots of Eq. (2) where a is calculated by averaging the estimated values of a over each condition (**No-feedback**: $a = -0.024$; **Recall-oriented**: $a = -0.015$; **Precision-oriented**: $a = -0.038$; **Human**: $a = -0.046$). The value of b is set to 0 for the purpose of comparison.

4 Discussion

Although **Hypothesis I** is not supported, the experimental results reveal that **Precision-oriented** has a learning effect comparable to **Human**. A concrete example makes this clearer. **Precision-oriented** takes 18 instances of learning to decrease the error rate 32%, which is the average of the subjects at the beginning, by half. This is very near the 16 instances of **Human**. By contrast, **No-feedback** takes nearly double that (29 times), and **Recall-oriented** far more (47 times).

From these results, it follows that one should

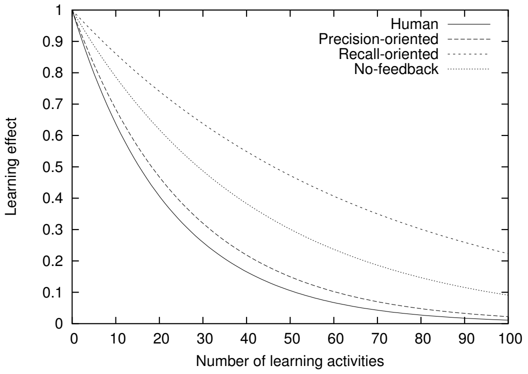

Figure 1: Experimental results

use precision-oriented error detection for writing learning assistance in a circumstance where feedback by a human tutor is not fully available (e.g., writing classroom consisting of a number of students). According to Burstein et al. (1998), the best way to improve one's writing skills is (i) to write, (ii) to receive feedback from a human tutor, (iii) to revise based on the feedback, and then repeat the whole process as often as possible. However, it is almost impossible to practice this in a writing classroom consisting of a number of students. In such circumstances, this can be done by using precision-oriented error detection. At the end, learners may have their essays corrected by a human tutor, which guarantees the quality of feedback, still reducing the efforts of human tutors.

At the same time, it should be emphasized that this is not a general but a limited conclusion because the experiments involve limited target errors and a limited number of subjects. In different conditions (e.g., setting a higher threshold), **Precision-oriented** may outperform **Human**, meaning that **Hypothesis I** is not conclusively rejected.

The experimental results support **Hypothesis II** as we expected. The learning effect of **Recall-oriented** is even less than **No-feedback**. A possible reason for this is that false-positives, which **Recall-oriented** frequently makes, confused the subjects. By contrast, **Precision-oriented** achieved better learning effect because it detected a few errors with a high precision. To be precise, **Recall-oriented** achieved a precision of 0.60 with a recall 0.31 whereas a precision of 0.72 with a recall of 0.25 in **Precision-oriented**. Besides, the fact that **Recall-oriented** detects errors more frequently with less precision (that is, the number of false-positives is higher) might make learners feel as if the precision is lower than is actually. This might have discouraged the subjects in **Recall-oriented** from learning.

These results suggest interesting findings from another point of view. In the past, overall performance of error detection has often been evaluated by F-measure, which considers both recall and precision. Following this convention, one comes to the conclusion that **Recall-oriented** ($F = 0.41$) is superior to **Precision-oriented** ($F = 0.37$). Contrary to this, the experimental results favor **Precision-oriented** over **Recall-oriented** in terms of learning effect. This suggest that F-measure is not always the best method of evaluation.

To conclude this section, let us discuss some problems with the proposed method that the experiments have revealed. To obtain more general conclusions, the amount of experimental data should be increased. However, it appeared to be difficult for the subjects to do the learning activity more than ten times; some subjects might have got bored with repeating the same learning activities. This is the problem that has to be solved in its actual use in learning assistance. Another problem is that detection performance tends to decrease relative to the original as learning proceeds because subjects improve (for instance, $F = 0.44$ for the first half and $F = 0.38$ for the last half in **Recall-oriented**). In order to investigate the relation between detection performance and learning effect more deeply, one should take this fact into consideration.

5 Conclusions

This paper tested the two hypotheses — imperfect, rather than perfect, error detection maximizes learning effect; and precision-oriented error detection is better than a recall-oriented one in terms of learning effect. The experiments revealed the interesting findings that precision-oriented error detection has learning effect similar to that of feedback by a human tutor, although the first hypothesis was not supported. Considering the findings, this paper has come to the conclusion that one should use precision-oriented error detection to assist writing learning in a circumstance where feedback by human tutors is not fully available. By contrast, the experiments supported the second hypothesis. They also showed that F-measure was not always the best way of evaluation.

In future work, we will expand the experiments in terms of both the number of subjects and target errors, such as errors in preposition, to obtain more general conclusions. The essays which are collected and error-annotated[2] in the experiments are available as a learner corpus for research and education purposes. Those who are interested in the learner corpus should contact the author.

References

Bitchener, John, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14(3):191–205.

Brockett, Chris, William B. Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proc. of 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia, July.

Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary D. Harris. 1998. Automated scoring using a hybrid feature identification technique. In *Proc. of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 206–210.

Chodorow, Martin and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proc. of 1st Meeting of the North America Chapter of ACL*, pages 140–147.

Chodorow, Martin and Claudia Leacock. 2002. Techniques for detecting syntactic errors in text. In *IEICE Technical Report (TL2002-39)*, pages 37–41.

Chodorow, Martin, Joel R. Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proc. of 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30.

Ferris, Dana and Barrie Roberts. 2001. Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, 10(3):161–184.

Han, Na-Rae, Martin Chodorow, and Claudia Leacock. 2004. Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. In *Proc. of 4th International Conference on Language Resources and Evaluation*, pages 1625–1628.

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

Hirota, Shigeaki. 1992. *Mastery (in Japanese)*. Kirihara Shoten, Tokyo.

Iizuka, Shigeru and Satoshi Hagino. 1997. *Prestige*. Buneido, Tokyo.

Izumi, Emi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *Proc. of 41st Annual Meeting of ACL*, pages 145–148.

Konishi, Tomoshichi and Kosei Minamide. 2007. *Genious English-Japanese dictionary, 4th ed.* Taishukan, Tokyo.

Nagata, Ryo, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2004. Recognizing article errors based on the three head words. In *Proc. of Cognition and Exploratory Learning in Digital Age*, pages 184–191.

Nagata, Ryo, Takahiro Wakana, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2005. Detecting article errors based on the mass count distinction. In *Proc. of 2nd International Joint Conference on Natural Language Processing*, pages 815–826.

Nagata, Ryo, Astuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proc. of 44th Annual Meeting of ACL*, pages 241–248.

Okihara, Katsuaki. 1985. *English writing (in Japanese)*. Taishukan, Tokyo.

Robb, Thomas, Steven Ross, and Ian Shortreed. 1986. Salience of feedback on error and its effect on EFL writing quality. *TESOL QUARTERY*, 20(1):83–93.

[2] Including not only errors in articles and number but also other types of error.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of 33rd Annual Meeting of ACL*, pages 189–196.