INNOVATION IN LANGUAGE LEARNING
International Conference

An Ensemble Classifier for Error Detection and Recommendation in the Use of Articles by Learners of English

KRISHNAMOORTHY Kiruthikaa (1), GAILLAT Thomas (2)

National University of Ireland, Galway, Ireland (1)
Insight Centre for Data analytics NUI Galway, Galway, Ireland (2)

Abstract

Learner English can be classified into multiple proficiency levels. To do so, texts written by learners are evaluated by native English speakers/teachers. This process can be time-consuming and could take a while before a learner gets some feedback. The aim of our research is to propose a method to leverage Machine Learning to predict errors and offer recommended alternatives as a feedback. The focus of this paper is to conduct error detection and correction on a, the, null articles as learners experience difficulties in choosing one of the forms. We propose an ensemble model built on syntactic and semantic features of texts. The Brown Corpus is a native written text collection used to build the model. REALEC, a learner corpus, is used in combination with the Brown Corpus as a Gold Standard to test the ability of the system to not only detect errors but also to validate the accuracy of its recommendations. Results show a 77% and 71% accuracy in the error prediction and correction experiments respectively, and thereby reduce the efforts of the human evaluator.

Keywords: *Article Microsystem, Learner English Evaluation, Multi-Layer annotations, Ensembles.*

1. Introduction

English is the globally adopted medium of communication. Therefore, many non-native speakers get enrolled in additional courses to learn the language. During the course, they are trained to hone their speaking, listening, writing and reading abilities. The writing tasks are graded by human evaluators who are proficient in the language. The evaluation consists of a typical error detection and correction procedure as a form of feedback to the learners. The time taken for a complete evaluation is proportional to the effort required. Therefore, it takes several days/weeks for the feedback to reach a class of enrolled learners. This phase may be intensive for the evaluators and causes anxiety among the learners. Hence, the aim of our system is to reduce this effort, and thereby make it a less time-consuming process benefiting both evaluators and learners.

Through this study we propose a machine learning based error prediction and recommendation system focusing on the articles - *a, the and null*, as they are foundational and most frequently used, and many learners lack clarity in this space. In this study, we build a classifier trained on features extracted from native English and test it on the learner English. In section 2 briefs about articles, section 3 describes the experiment conducted, section 4 describes its results.

2. Theoretical Background

Articles *a, the* and *zero/null* in English, a grammatical paradigm that functions as a microsystem, are the core focus of this paper. They are the most frequently occurring morphemes in English. The *zero article* is the most frequently used article, followed by *'the'* and then *'a'* [8]. The articles indicate the definiteness and specificity of a noun and are found preceding nouns.

The definite article, *the*, implies that the succeeding noun is referring to a specific entity. The definite article, preceding either a countable or uncountable noun, means that the referent of the noun phrase is assumed to be known to the speaker and the addressee [1].

The indefinite article, *a*, is used with singular, countable and a non-specific noun [1]. The *zero article* is found with non-specific or generic forms of uncountable and plural countable nouns [1].

In terms of acquisition of the microsystem, learners go through different stages. A sentence may look syntactically correct with or without the use of one of the articles. However, the paradigmatic correctness depends on the context of the text [3]. At the beginning of their learning process, most learners tend to avoid the use of articles [8], which is considered as an overuse of the *zero article*. At the later stage, they learn the necessity of an article and tend to use *'the'* for every noun. As they gain familiarity/fluency with the language, their article usage becomes accurate [8].

3. Experimental Setup

3.1 Proposed System Architecture

The generalized workflow of a grammatical microsystem evaluator comprises of four stages:

1. **Text Decomposition:** Extract the text fragments that are used to learn the microsystem.
2. **Feature Extraction:** For a given text fragment, a set of features that influence the usage of a microsystem is extracted.
3. **Feature Representation:** Convert the extracted features to vectors that can be used for building/running the machine learning model.
4. **Building/Running the model:** If the machine learning model is already built, the feature vector is run to determine the outcome. Otherwise, the feature vector along with the expected outcome is added to the dataset to train the model.

These steps are customized to evaluate the article microsystem as detailed in further sections.

3.2 Corpora

Corpora containing text features are the data used to build the machine learning model. The model learns the native usages and applies the knowledge acquired to evaluate incorrect instances. Hence, a native error-free corpus and a learner corpus with annotation errors are required. From the corpora we extract each occurrence of an article followed by a noun as a tuple in the dataset. For instance, the first two rows of Table 1 is extracted from *"Opponents made a similar proposal".*

The Brown corpus, built by native American writers, is the chosen error-free corpus [2].

The learner corpus chosen for this study is the REALEC [11], a corpus containing essays written by Russian learners of English. The corpus comprises of 3,400 essays and 838,000 word in total. It follows a hierarchical error annotation scheme in which each error is tagged along with the error-type, impact and correction performed by professors teaching Academic Writing in English.

3.3 Features

Texts are vast contextual content that need to be reduced to features that can be processed by a machine learning algorithm. In case of articles, the noun and its surrounding texts are features that determine the article usage. The following text features are extracted for this experiment.

1. **POS Tags:** A Penn Treebank tagset consisting of 36 POS tags and 12 tags for symbols is used to tag the text. These tags are the primary features of the text [7].
2. **Named Entity Recognition (NER):** To add more details to the nouns, we perform NER to categorize a proper noun or entity as a person, location, organization etc [9].
3. **Countability:** Nouns can be classified as countable and uncountable based on their finiteness. The Google N-gram service takes a n-gram as input and returns the frequency of the n-gram after searching in all Google Books published between a range of years. From this service, we compute the frequency of a noun preceded by the "many" and then, by "much". If the former is the greatest, then the noun is countable (represented by "CNT"), else considered uncountable ("UNCNT").
4. **Anaphora:** When two words co-refer, then it implies the presence of an anaphoric link. This is identified by determining the presence of synonyms to a noun [6].

For example, "*She saw a blue car outside her house. The vehicle was punctured. ".* Here, the vehicle refers to the car mentioned in the previous sentence.
This is represented by 'CXT' aside the tags.

3.4 Features

A snippet containing nouns and corresponding articles from the Brown Corpus is tabulated in Table1.

Table 1. Example of data from the Brown Corpus

Article	Text	Feature
zero article	Opponents	NNS+CNT
a	similar proposal	JJ NN+CNT
the	election	NN+CNT+CXT

A snippet of the Gold Standard Dataset that is used to test the system is tabulated in Table 2. First is a row from the REALEC. The next is a row from Brown Corpus, hence the error column is *<none>.*

Table 2. Sample of the Gold Standard dataset

Text	Feature	Error	Expected Outcome
Sharp line	JJ+NN+CNT	Zero article	a
Jury	NN+CNT+CXT	<none>	the

3.5 Ensemble Model

This section describes the machine learning model, validation and evaluation methodologies adopted in this study.

Building the model: A Majority Voting Ensemble of Naive Bayes, SVM and ME classifiers to learn and classify the features presented is built [10]. Individually, these classifiers reported an accuracy of 74%, 78% and 75% respectively, when tested with a sample of the Brown corpus. However, an ensemble of these three algorithms resulted in an 88% accuracy when tested with the same sample. Evidently, the ensemble model makes more accurate predictions [4][5] and hence it is used as the supervised learning technique to build the model to learn the article microsystem.

Validation: The Brown Corpus dataset is used to perform k-fold cross validation to validate the fitness of the model. This approach flags problems like overfitting and gives indications of how the model will generalize to an independent dataset. For this experiment,
we chose 10 as the arbitrary value for k. Throughout the 10 iterations, the model resulted in an accuracy ranging from 85% to 90%.

Evaluation: Prior to error correction, error detection performance is crucial. Post gaining confidence from this experiment, we conduct the error correction experiment, wherein, every correction suggested by the system is validated against the expected outcome. Metrics such as precision, recall, accuracy and F1 score are computed to quantify the performance.

4. Results and Discussion

Table 3 tabulates the results of the pre-requisite error prediction experiment. It shows that 85% of the correct usages were marked correct by the system and 72% of the errors were identified.

Table 3. Error Detection: Confusion Matrix

Predicted Outcome		
	Correct	Incorrect
Correct	2125	375
Incorrect	750	1750

Table 4 tabulates the error correction performance. It is understood that there are 3,500 instances of each class, indicating that the dataset is balanced. 68%, 72% and 71% of the instances expecting article *a, the* and *null* respectively have been classified correctly.

Table 4. Error Correction: Confusion Matrix

Predicted Outcome			
	a	the	Zero article
a	2396	35	1069
the	210	2520	770
Zero article	903	115	2482

Table 5 tabulates the values of the evaluation metrics computed.

Table 5. Evaluation metrics

Measure	Value
Accuracy	0.704
Recall	0.66
Precision	0.68
F1-Score	0.67

The results state that 71% of the system's decisions coincided with those of a human annotator. Note, the REALEC corpus has instances where learners used the incorrect noun which misguided the classifier at the cost of its accuracy. Hence, these results are subject to variations in performance depending upon the types of errors in the corpus.

References

[1] Biber, D., Johanson, S., Leech, G., Conrad, S., Finegan, E.: Longman Grammar of Spoken and Written English. Longman, Harlow (1999)
[2] Bird, S., Loper, E.: Nltk: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. p. 31. Association for Computational Linguistics (2004)
[3] De Felice, R., Pulman, S.G.: A classifier-based approach to preposition and determiner error correction in l2 English. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 169–176. Association for Computational Linguistics (2008)
[4] Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer, Berlin, Heidelberg (2000)
[5] Dong, Y.S., Han, K.S.: A comparison of several ensemble methods for text categorization. In: Services Computing, 2004.(SCC 2004). Proceedings. 2004 IEEE International Conference on. pp. 419–422. IEEE (2004)
[6] Fellbaum, C.: A semantic network of english verbs. WordNet: An electronic lexical database 3, 153–178 (1998)
[7] Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: The penn treebank. Computational linguistics 19(2), 313–33 (1993)
[8] Master, P.: The English article system: Acquisition, function. System 25(2), 215– 232 (1997)
[9] Mohit, B.: Named entity recognition. In: Natural language processing of semitic languages, pp. 221–245. Springer, Berlin, Heidelberg (2014)
[10] Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
[11] Vinogradova, O.: The role and applications of expert error annotation in a corpus of English learner texts. Computational Linguisitics and Intellectual Technologies. Proceedings of Dialog 2016 15, 740–751 (2016)