Korean Journal of English Language and Linguistics, Vol 22, January 2022, pp. 19-39
DOI: 10.15738/kjell.22..202201.19

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Assessing Nativelikeness of Korean College Students' English Writing Using fastText

Hyesun Cho (Dankook University)

Received: Dec. 22, 2021
Revised: Jan. 24, 2022
Accepted: Jan. 30, 2022

Hyesun Cho
Associate Professor, Dept. of Education, Graduate School of Education, Dankook Univ.
hscho@dankook.ac.kr

ABSTRACT

Cho, Hyesun. 2021. Assessing nativelikeness of Korean college students' English writing using fastText. *Korean Journal of English Language and Linguistics* 21, 19-39.

Neural-network models have recently been used to assess nativelikeness of English sentences written by native or nonnative speakers. In this study, nativelikeness of Korean EFL college students' English writing is assessed using fastText, a neural-network text classifier using subword information. The training data consisted of English sentences from the corpora of native speakers of English and Korean EFL college students. The test sentences consisted of English writing assignments written by Korean EFL college students. fastText performed well for the task of binary classification into native and nonnative sentences, with high accuracy in less than a minute. The sentences that are classified as native with a high probability tend to have fewer grammatical as well as plausibility errors than those classified as nonnative. For the test sentences, correcting grammatical errors (involving articles, number, subject-verb agreement, voice) had weaker effects on the classification of the sentences than correcting plausibility errors (word choices), which conforms to the previous literature. This suggests that fastText is more sensitive to plausibility errors than grammaticality errors which requires knowledge on hierarchical syntactic structures.

KEYWORDS

English writing, nativelikeness, Korean EFL learners, fastText, deep learning, neural networks, plausibility, grammaticality

1. Introduction

1.1 Nativelikeness of L2 Writing

One goal of learning English as a second language is to achieve near native-like proficiency in English (Jung 2005, Nam 2011, Šišić 2016). Sentences that are nativelike are well-understood by English speakers, facilitating effective communication. Despite controversies over the concept of native speakers (Davies 1991), it is generally agreed that a native speaker model is necessary in language teaching (Cook 1999, Lee 2005). Nativelike English proficiency includes proficiency in various aspects of language: grammaticality, the vocabulary size and use, natural pronunciation, etc. Though it is hard to achieve native-like pronunciation after a critical period, it is possible to achieve native-like proficiency in lexicon and syntactic performance in later ages (Scovel 1988). Thus, grammar and vocabulary are the areas where adult learners, such as college students, can gain more nativelike proficiency through learning (cf. Morgan-Short et al. 2012).

For a sentence to be nativelike, it should be grammatical, but grammaticality alone does not ensure that the sentence is also nativelike (Lee et al. 2021, Nam, 2011, Park 2020, Park et al. 2019, 2020, Pawley and Syder 1983). Not all grammatically-permissible sentences sound natural to native speakers, which is one source of difficulties that English learners encounter. In addition to grammaticality, appropriate use of vocabularies, such as right collocations or multiword units, is essential. For example, the sentences such as 'there is a lot of rain falling', 'the rain is falling hard', and 'there is heavy rain' are all grammatical, but the most nativelike sentence is the last one (Nation 2008: 117). Likewise, Park et al. (2019) analyze nativelikeness of sentences in terms of grammaticality (well-formedness) and felicity (or plausibility, in Park et al. 2020)).

Plausibility largely depends on appropriate use of words and collocations. The ability to use multiword units is needed to say and write things like a native speaker (Nation 2008). Overall vocabulary size and lexical competence increase as the proficiency increases (Nation 2008, Zareva et al. 2005). According to Kim and Bae (2012), collocation knowledge is significantly correlated with writing quality. However, studies show that Korean adult learners have difficulties in collocation use due to L1 influence (Chang 2018). Korean speakers tend to use fewer amplifiers than native speakers due to their limited vocabulary (Lee 2006).

The level of EFL learners' proficiency can be easily seen in their academic writing. Academic writing is different from conversation or other more casual activities using L2 (Cummins 2005, Nam 2011). It can reveal the writers' English competence at their best because one usually goes through rounds of revisions and polishing. In academic writings of Korean college students, however, both grammatical and collocation errors are often found (Kim 2014, Kim 2018, Kim 2020). In Kim (2014), common grammatical errors by Korean college students involved verbs, adverbs, and articles. Kim (2018) examined low-level Korean college students' argumentative essays. The most common errors involved nouns, subject-verb agreement, sentence structure, and tense. Obviously, grammaticality is more of a concern for L2 learners compared with native speakers, though the native-speaker advantage may be limited because they also need to understand the genre of academic writing and its disciplinary knowledge to be a successful writer (Zhao 2017).

1.2 Literature Review: Assessing Nativelikeness of Sentences Using Deep Neural Networks

Native language identification (NLI) is a binary classification task where a model makes decisions on whether sentences are written by native speakers or L2 learners of the language (Goldin et al. 2018). Deep learning neural network models, a branch of machine learning algorithms using neural networks, have been adopted to assess

nativelikeness of second language learners' sentences in several recent studies: Park et al. (2019, 2020), Park (2020) and Lee et al. (2021). In these studies, native and nonnative (learner) corpora were used to train deep neural networks. The native corpus that Park et al. (2019) and Park (2020) used was the Corpus of Contemporary American English (COCA) (Davies 2008-), which has more than one billion words from eight genres. They used the texts from five of these genres. The learner corpora were the Yonsei English Learner Corpus (YELC) and the Gachon Learner Corpus (GLC). YELC is a corpus of English texts written by Korean college freshmen, containing over 1.08 million words (Rhee and Jung 2014). GLC is a corpus of English texts written by Korean college students, containing more than 2.5 million words (Carlstrom and Price 2012-2014). These studies showed that deep learning neural network models can be trained to classify native vs. nonnative sentences with high accuracy. Moreover, the neural-network models can find patterns that may elude human linguists' observations because they can learn patterns based on large data.

Park et al. (2019) used a recurrent neural network (RNN) model, which is widely used in natural language processing. Whereas one-directional ('feed-forward') neural-networks do not memorize previous information in the process of training, RNN models can store and utilize past information (i.e., previous words in a sentence), so they can deal with contextual information. For this reason, various RNN-based models are used for natural language processing (e.g., Cho et al. 2014). In Park et al. (2019), the RNN-based model achieved the validation accuracy of 93.9% for the task of classifying native and learner sentences.

Lee et al. (2021) classified Korean sentences written by native speakers and learners of Korean using a deep learning neural network model, employing KoBERT (Korean Bidirectional Encoder Representations from Transformers). KoBERT is a pre-trained language model by SKT T-Brain, trained with millions of Korean sentences from various sources such as Korean Wikipedia, newspapers, books. They tested the model with a native Korean corpus of college students' writing and an annotated learner corpus. The accuracy of the classification was 91%.

Despite the overall high accuracy of classification by the deep learning models, the classification sometimes diverges from human judgments. Park et al. (2019) and Lee et al. (2021) both found that native speakers' sentences may be judged as learner sentences, or *vice versa*, when the patterns in the test sentence rarely appear in the training data. They both found that the model is more sensitive to felicity errors rather than grammatical errors. For example, in Park et al. (2019), "I was satisfied with his performance" was classified as a learner sentence whereas "I satisfied with his performance" was classified as a native sentence. That is, the model did not recognize the grammatical error in the latter sentence as humans do. This is because such error patterns rarely exist in the learner corpus, according to the study. On the other hand, the model is good at finding grammatical collocation patterns and felicity errors. For example, a sentence containing "the same" is classified as a native sentence, but if there is no article before "same", then the sentence is classified as a learner sentence. The model is better at identifying felicity errors such as the position of adverbs, semantic prosody (e.g., **caused happiness* vs. *caused a disaster*), and semantic preference. Lee et al. (2021) reported similar findings. Even if there is no grammatical error, the deep learning model may classify a native sentence as a learner sentence due to the patterns that rarely exist in the training data.

1.3 Assessing Nativelikeness Using fastText

The present study explores the use of fastText for the task of classifying native vs. nonnative sentences. fastText is a third-party library for text representation and classification developed by Facebook AI Research Lab (Joulin,

Grave, Bojanowski, and Mikolov, 2017)[1]. It is well-known as a word-embedding algorithm, but it can also be used as a text classifier. It has one hidden layer, so it is a shallow, not deep, neural-network model. It has advantages in terms of processing time, accuracy, and accessibility. As its name suggests, it can process text-related tasks very fast, requiring a shorter running time than most deep learning models do, while its performance is comparable to or better than several other deep-learning based methods (Bojanowski et al. 2017, Joulin et al. 2017). In a sentiment analysis task, fastText showed a test accuracy of 92.5%, higher than other deep learning-based models on the same dataset (Joulin et al. 2017). In addition, it runs on a multicore CPU, whereas deep learning models usually require GPUs to speed up for large data sets. Therefore, it can be easily accessible by English-major students and researchers. With these advantages, the present study examines the following research questions:

Q1. How accurately does fastText make predictions on nativelikeness of sentences?
Q2. What are the characteristics of the sentences that are classified as native or nonnative by fastText, in terms of grammaticality and plausibility?
Q3. What are the similarities and differences in the nativelikeness judgments between fastText and deep neural networks?

To address these questions, I used fastText for the classification task of native and nonnative sentences and compare the sentences where fastText made correct or incorrect predictions. Finally, sentences with grammatical or plausibility errors are corrected and re-tested by fastText to examine whether the errors affected fastText's judgments on the nativelikeness of the sentences.

2. Research Method: Text Classification Using fastText

2.1 The fastText Library as a Text Classifier

fastText is a third-party library for efficient text representations and classification. In fastText, sentences are represented by subword units, or character *n*-grams, rather than words as units (Bojanowski et al. 2017). For example, when $n = 3$ (that is, trigram), the word *where* is represented by a bag (collection) of tri-grams such as *< wh, whe, her, ere, re >* ('<' is a boundary symbol). In addition, the *n*-grams for *n* greater than 3 (e.g., quadrigram) and the word itself (Bojanowski et al. 2017) are included. The sentences that are represented by subword units in this way are then fed to a text classifier to carry out classification. In previous models with words as a basic unit, unknown words and morphologically complex languages were hard to process. On the other hand, since fastText uses subword units, it can deal with unknown words or even typos that can be a characteristic of nonnative writings.

As mentioned, fastText can process a large amount of text data with high accuracy in a short time. Joulin et al. (2017) used fastText for classification tasks such as sentiment analysis and tag prediction. In a sentiment classification task, it achieved a higher accuracy than other character-based deep-learning models[2]. In these models, the running time for a single epoch[3] ranged from a few hours up to 5 days, but it took only 10 seconds for fastText

[1] https://fasttext.cc/

[2] These are the character-level convolutional model (Zhang and LeCun 2015) and the character-based convolution recurrent network (Xiao and Cho 2016), variants of convolutional neural networks (CNNs). It is known that the character-based convolutional models performed better than word-based models (Zhang et al. 2015).

[3] One cycle that a learning algorithm works through the entire training data

for the same task. In a tag-prediction task, fastText classified half a million sentences in less than a minute, with a higher precision than a compared classifier, Tagspace[4].

2.2 Datasets

2.2.1 Training and validation data: Native and non-native corpora

The training data consisted of native and nonnative sentences, collected from native and nonnative corpora. The native corpus was LOCNESS (the Louvain Corpus of Native English Essays) (Granger 1998), a corpus of native English essays written by British and American university students (total 324,304 words)[5]. The topics for British students include transport, boxing, parliamentary system, fox hunting, and French tradition and society. The topics for American students were more various, including capital punishment, teenagers, drinking age, football, great inventions and discoveries of the 20th century and their impact on people's lives, etc.

The non-native corpus was GLC (the Gachon Learner Corpus), consisting of English writing of Korean college students (Carlstrom and Price 2012-2014). The students wrote English texts between 100 and 150 words to 20 questions such as "What topics should people avoid during small talk? Why?", "Do you ever worry about using the Internet? Why or why not?". Training data in Park et al. (2019) consisted of texts from various genres, but in the present study the training and test data were limited to college students' writing, so that one can see the result when the text genre is controlled.

The essays in LOCNESS were all combined into one text file, and all the paragraphs were divided into sentences by making new lines at every end-punctuation symbol using regular expressions in a text editor. This resulted in 15,174 sentences. From here, sentences with less than three words and those containing French words were excluded, resulting in 14,893 sentences. GLC has more sentences, approximately 246,179, but to make the number of sentences balanced (as in Park (2020) and Lee et al. (2021)), the same number of sentences as LOCNESS (14,893) were selected from GLC. Native sentences in LOCNESS tended to be longer than nonnative sentences in GLC, so the model's nativelikeness judgments may be affected by sentence length, not just words themselves. To prevent this, sentence length (the number of words in a sentence) was considered when selecting sentences from GLC. For each sentence length, the same or similar number of sentences was randomly selected. For example, there were 8 sentences with a sentence length of 4 in LOCNESS. From GLC, then, only 8 sentences with the same sentence length (4) were randomly chosen and included in the training set. As a result, the average sentence length in the training data was 19.9 for native data and 19.3 for learner data. A total 29,786 sentences were labeled with native ('ENG') and nonnative ('KOR') labels. An example of training data is shown in Figure 1.

[4] A tag prediction model using a convolutional neural network (Weston et al. 2014)

[5] The composition of LOCNESS is as follows: argumentative essays written by American university students: 49,574 words (46%), literary-mixed essays written by American university students: 18,1826 words (5.8%), argumentative and literary essays written by British university students: 95,695 words (29.5%), British A-level argumentative essays: 60,209 words (18.5%)

```
__label__KOR     they drive so fast or drive closer or change often line or stop car suddenly or
threat other driver in different ways .
__label__ENG     points must be made accessible to the students in order for communication between
the students and the professor to take place .
__label__ENG     we can learn that there is not time like the present when approaching everyday
life .
__label__KOR     and they observe traffic signal when they through the narrow streets or there are no
cars on the street .
__label__KOR     but just a little ,  it ' s ok .
```

Figure 1. Example of the Training Data

The training data was then shuffled and split into training and validation data with an 8:2 ratio (23,829 and 5,957 sentences respectively), which is the most typical ratio in machine learning research (Goodfellow et al. 2016: 113).

2.2.2 Data preprocessing

The sentences in the training and validation data were lower-cased and special symbols, emoticons, and punctuation markers were all removed, as recommended in the fastText tutorial (also, see Seo and Shin 2020). LOCNESS contained sentences including French words, while Gachon Corpus contained sentences including words in Korean characters. These sentences were all removed so that the model may learn more meaningful patterns in English, rather than make the decisions based on the presence/absence of French words or Korean characters in a sentence.

2.2.3 The test set: Korean EFL college students' writing

The language model obtained by fastText was then used to classify nonnative sentences written by Korean college students. The test data was made up of writing assignments collected from three College English classes in a university in South Korea. Two of them were intermediate level, and one was beginner level. The students in the intermediate classes were asked to write a paragraph about the question, "The Internet has made our lives better. Do you agree or disagree?". The writing topic for the beginner class was "popular sports in Korea". From the two intermediate classes, 31 and 11 writing assignments were collected respectively, and from the beginner-level class, 32 writing assignments were collected. The number of sentences in the test data was 838 in total. The test data was preprocessed in the same way the training data was preprocessed. It was assumed that sentences with more errors will be more likely to be classified as nonnative and *vice versa*, following the previous studies (Lee et al. 2021, Park et al. 2019).

2.3 Training, Validation, and Testing

Like most other machine learning algorithms, the procedure of using fastText for text classification consists of three phases: training, validation, and testing[6]. To train the model, training data is fed to fastText, and the results are evaluated using validation data. In fastText, a model's performance is measured by precision and training loss. Precision is how precise the model's predictions are, calculated by (true positive)/(true positive + false positive)[7].

[6] https://fasttext.cc/docs/en/supervised-tutorial.html

[7] Precision is the proportion of correct labels among the corresponding labels predicted by the model (e.g., (actual native)/(predicted as native)). Models' performance is also commonly measured by 'accuracy', which is the proportion of

Training loss is the sum of the differences between the model's predictions and the actual labels, computed by a loss function. The parameters in the loss function are updated in the direction of minimizing the training loss. In doing so, there are parameter settings that users can change to control a model's performance, which are called hyperparameters (Goodfellow et al. 2016: 113). In fastText, hyperparameters include the number of epochs, learning rate, and word *n*-grams. An epoch means one cycle of working through the entire training set. Learning rate is the step size of updating gradient, which is related to how fast learning can occur. Word *n*-grams means the continuous sequences of *n* number of words (e.g., a bigram (i.e., $n = 2$) means a sequence of two words ('This is', 'an apple' are bigrams)). In the present study, training, validation, and testing were all done in the CLI (Command Line Interface) environment in Terminal on MacBook Pro (Intel Core i7). Running time was only a few seconds for each command line (See Appendix for command examples).

3. Results

3.1 The Best Model

fastText was trained with the training data and evaluated with the evaluation data, as described in Section 2, with hyperparameters varied as in Table 1. The hyperparameters include the number of epochs, learning rate, word *n*-grams. The precision and training loss values show the performance of the model under each condition.

Table 1. Hyperparameters Testing

	Epoch	Learning rate	*n*-gram	Precision*	Training loss
a	5	0.1	1	0.934	0.2082
b	10	0.1	1	0.937	0.1415
c	25	0.1	1	0.934	0.0796
d	10	1.0	1	0.934	0.131
e	10	0.1	2	0.942	0.098
f	10	0.1	3	0.937	0.108
g	**25**	**0.1**	**2**	**0.943**	**0.041**
h	30	0.1	2	0.942	0.032

* Precision = (true positive)/(true positive + false positive).

As precision increases, training loss decreases. Using bigram substantially improved precision from 0.937 to 0.943 ((c) vs. (g)) but using trigram did not ((e) vs. (f)). Changing the learning rate did not improve the result ((b) vs. (d)). The best performance (the highest precision) was found with the condition (g), with 25 epochs, a learning rate of 0.1, and bigram. The model returned 5823 correct labels and 314 incorrect labels for 5957 sentences in the validation set, so the validation accuracy is 97.75% (5823/5957).

3.2 Classification Results of the Validation Set

Table 2 shows example sentences from the validation set. The 'label' column shows the correct labels, the

correct labels from the entire labels given by the model. The fastText library provides precision, not accuracy.

'predict' column shows the labels predicted by the model. Thus, (a), (b), (g), and (h) are correct predictions, and (c), (d), (e), and (f) are incorrect predictions. The probability column shows the probability of the predicted labels given by the model. Probability indicates how confident the model is for the prediction (Park et al. 2019: 211).

Table 2. Classification of Sentences in the Validation Set

	Sentences	Label	Predict	Probability
(a)	This would allow her to justify her actions internally as well as externally.	ENG	ENG	1
(b)	This he left mainly up to the church as he believed the church would promote social cooperation.	ENG	ENG	1
(c)	At the time, I played basketball.	ENG	KOR	0.99
(d)	My first time ever seeing weed was when I was a little kid.	ENG	KOR	0.99
(e)	These experiences are negative to their lives.	KOR	ENG	1
(f)	They were just very talented since they were very little boys or girls.	KOR	ENG	0.99
(g)	If a cashier gave me too much change, it makes feel good.	KOR	KOR	1
(h)	I see a red color that remind blood.	KOR	KOR	1

The native sentences in (c) and (d) are classified as nonnative. Classification of native sentences as nonnative is often observed in short sentences (cf. Park et al. 2019:210). According to Park et al (2019), short sentences do not have much information, so the classification of short sentences can be inaccurate.

The learner sentences in (e) and (f) are grammatical and classified as native. On the other hand, the learner sentences that are classified as learner sentences ((g) and (h)) contain clear grammatical errors. In (g), the object of *makes* is missing, and in (h), the verb in the relative clause does not agree with the antecedent (**that remind* for *that reminds*). However, nonnative sentences are sometimes classified as nonnative even if there is no grammatical error (e.g., *So I usually buy active clothes with vivid color; I think there are good and bad drivers in my city*).

3.3 Classification Results of the Test Data

The model obtained in 3.1 was used to classify the sentences in the test data. As can be seen in Table 3 and Figure 2, more sentences were classified as nonnative (577 nonnative, 261 native). A chi-squared test shows that the frequencies between the two groups are significantly different ($\chi^2(1) = 119.16$, $p < 0.0001$). The test accuracy is 68.9% (the proportion of the sentences classified as nonnative, given that the test data was all learner sentences).

Table 3. The Classification Frequency and Probability

	Number of sentences	Mean probability
Native ('ENG')	261 (31.1%)	0.897
Nonnative ('KOR')	577 (68.9%)	0.948

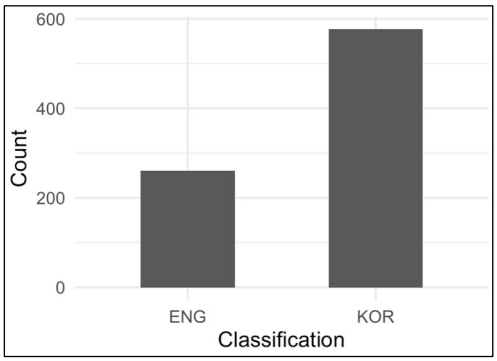

Figure 2. Number of Sentences in Each Class (ENG: Native, KOR: Nonnative)

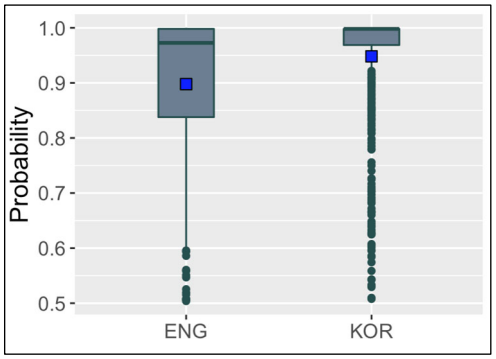

Figure 3. Probability of Classification as Native (ENG) or Nonnative (KOR)

The distribution of the probability of classification is shown in Figure 3. The mean probability was 94.8% for nonnative class and 89.7% for native class. The probability values were not normally distributed (W = 0.6, $p < 0.0005$, the Shapiro-Wilk normality test), so a Mann-Whitney U test was conducted to test whether the mean probability was significantly different. The result shows that the mean probability was significantly different between the groups (W = 51395, $p < 0.0005$). Given that probability can be interpreted as confidence on the label by the model (Park et al. 2019, 2020), fastText was more confident with giving nonnative labels for the test sentences than native labels. Since the test data was written by nonnative Korean speakers, this result is as expected.

3.4 The Sentences Classified as Native

In this section and the following section, the test sentences that are classified as native (Section 3.4) and those classified as nonnative (Section 3.5) are presented. It can be expected that the sentences with fewer errors are more likely to be classified as native, and *vice versa* (Park et al. 2019, 2021, Lee et al. 2020). We will see whether this is also observed in our results and what kinds of sentences were classified as native or nonnative by fastText.

The example sentences are randomly selected from the highest probabilities and the lowest probabilities. That is, the order of the test sentences was shuffled before the test, and then after obtaining the results they were sorted by their probabilities in descending order. The example sentences here are the first ten sentences with the highest probability (in Table 4), and the last ten sentences with the lowest probability (in Table 5).

Table 4 shows the sentences that are classified as native with the highest probabilities, in descending order. The errors column shows the errors in the sentences if any, with corrections put in the parentheses in *italic* fonts. Most

of the sentences are grammatical, and the observed grammatical errors seem rather minor. The errors involve missing an article (a, f), inserting an unnecessary article (h), incorrect adjective (Part-of-Speech error) (h) (**almost* for *most*), and sentence fragment (sentences without the main clause) (j).

Table 4. Sentences Classified as Native (High Probability)

	Native	Errors	Prob
a	He is now major league player.	article	1
b	I agree with this opinion.		1
c	There are many reasons to support this.		1
d	2 million people like football.		0.998
e	According to writing assignment guidelines, 2 million people are fans for football.		0.998
f	Actually soccer is more common name for us.	article	0.998
g	All the communication advantages and technology advancement should not be available.		0.998
h	Almost people like the football.	POS (*most*), article	0.998
i	Also, both men and women like football.		0.998
j	Although there are advantages and disadvantages both.	SF(~~*although*~~)	0.998

* POS: Part-of-speech error, SF: Sentence fragment error

In the case of sentence (i), the adverb *also* is placed in the beginning of the sentence. Park et al. (2019, 2020) considered this as plausibility error (Park et al. 2020:209). According to Park et al. (2019), Korean learners often place some adverbs at the beginning of the sentence, which may sound unnatural for native speakers. Although sentence (i) was classified as native here, there were 23 sentences in the test set that have sentence-initial *also* and out of these 23 sentences, only two sentences (8.6%) were classified as native. It thus seems likely that sentence-initial *also* increases the probability of being classified as nonnative by the model. This is a similar result to Park et al. (2019), where sentences with initial *also* were classified as nonnative. However, Park et al. (2020) found that native speakers judge this kind of sentences rather plausible (compared with other more serious errors). Native speakers use adverbs in the sentence-initial position in casual speech, so it seems somewhat acceptable.

Table 5 shows the sentences classified as native with the lowest probabilities. Since fastText uses the softmax function to compute the probability distribution of the classes, a probability higher than 0.5 is classified as native. The sentences here are those with probabilities just over 0.5, the ten lowest ones. This means that the model classified these sentences as native but less confidently. These sentences can be considered very close to learner sentences.

Table 5. Sentences Classified as Native (Low Probability)

	Native	Errors	Prob
a	Next, let's look at the benefits of education about the benefits of the internet.	repetition	0.559
b	It depends on their schedule.		0.551
c	Teenagers who are addicted to social media often lose their true self because they often live fake lives by pretending they 're living a perfect live but they are not.	POS (*life*)	0.547
d	We can see in a baseball game is baseball stadium.	preposition insertion(~~*in*~~), preposition (*in*), article missing (*a baseball~*)	0.525
e	Another advantage of the internet is the increased level of education.		0.520
f	Everybody likes baseball.		0.516
g	Because badminton plays in a stadium.	SF, voice (*is played*)	0.508
h	They are most of young people.	adverb (*mostly*)	0.508
i	As people who majored engineering has increased and accepting technologies from abroad, they have improved transportation.	preposition (*majored in*), agreement (*have*)	0.506
j	The internet has increased the amount of information that educators can discover and the amount of information that students can understand.	repetition	0.504

Compared with Table 4, the sentences in Table 5 seem to be longer. However, a linear regression result shows that the number of words in a sentence does not significantly affect the associated probability value ($t(836) = -1.6$, $p = 0.11$).

Overall, the sentences in Table 5 seem to have more errors than in those in Table 4. Sentence (a) has no grammatical errors, but the repetition of similar phrases makes it sound awkward. It is classified as native but with very low probability. In sentence (d), a preposition is incorrectly inserted (**see in a baseball game* for *see a baseball game*) and a verb is wrongly used where a preposition is needed (**is baseball stadium* for *in a baseball stadium*). Sentence (g) is a sentence fragment, and there is an error involving voice (*badminton is played*). Sentence (i) can be corrected as *they are mostly young people*, considering the context. Sentence (j) is grammatical but sounds awkward due to repetition. It can be simplified by avoiding repetition of the phrase, *the amount of information*. Overall, the sentences classified as native with lower probability seem to have more grammatical or plausibility errors than those classified as native with higher probability.

3.5 The Sentences Classified as Nonnative

Table 6 shows the sentences that are classified as nonnative with the highest probability (all probability of 1). This suggests that the model is highly confident that these sentences are nonnative. Compared with the sentences in Table 4, there are clearly more errors.

Table 6. Sentences Classified as Nonnative (High Probability)

	Native	Errors	Prob
a	Also, we play game and listen to music by internet.	number (*games*), article, preposition (*on the internet*)	1
b	Although the internet has a side effects like addiction, virus, and unsuitable information but it can be prevented if you use the internet carefully.	article/number (*side effects*), number (*viruses*), collocation (*inappropriate*), pronoun (*they*)	1
c	And when we don't know the word, we can use internet dictionary.	article (*an internet~*)/number (*dictionaries*)	1
d	As many people use the internet, a lot of informations from a lot of people who use internet are shared in internet.	number (*information*), article (*the*), article, preposition (*on the internet*)	1
e	Baseball famous sportsman is park in Korea.	*a famous baseball sportsman in Korea is park* (word order, article, placement of prepositional phrase)	1
f	Baseball is very popular in Korea.		1
g	Baseball is very popular in my country, south Korea.		1
h	Baseball is very popular in south Korea, Korea recently have a domed stadium.	*~Korea and Korea recently~* (Comma splice), *built* (verb tense)	1
i	Basketball is very popular in Korea.		1
j	Besides, when I get lost, I can watch the map on the internet.	word choice (*look up*)	1

We can see that the sentences classified as nonnative contain various kinds of errors and have more errors than those classified as native. Errors involve minor issues such as articles (a, b, c, d, e) and number (a, b, c, d). In (d), the uncountable noun *information* is in the plural form (**informations*). Other errors involve word order, word choice, and comma-splice. In (e), word order is wrong: the adjective *famous* is placed before a noun (**baseball famous*), and the prepositional phrase *in Korea* is placed at the end of the sentence instead of the end of the subject noun phrase, which would be more natural. Sentences (h) is a comma-splice (two clauses conjoined only by a comma without a conjunction) (Folse et al. 2020, Kim 2020) and there is a tense error: the adverb *recently* should be used with past or perfect tense, but it is used with the present tense. In (j), *look up the map* is a better collocation than **watch the map*. Also notice that the sentences that are grammatically correct and plausible (f, g, i) all contain the word *Korea*. The vocabulary may have affected the classification of these sentences as nonnative.

Table 7. Sentences Classified as Nonnative (Low Probability)

	Native	Errors	Prob
a	If you want to watch a basketball match, you can buy tickets from any tourist office and watch a basketball match in the stadium.		0.552
b	Jokgu is an important part of Korean culture.	article (*the*)	0.542
c	Although the internet has a negative effect on our lives, we agree that the internet has made our lives better because there are many more positive impacts on our lives.		0.542
d	Baseball's price of tickets is proper.	inanimate possessor (*the ticket price of baseball*), word choice (*reasonable*)	0.539
e	You can predict the future of robot industry and software industry industries and prepare for change.	redundant nouns (*robot and software industries*)	0.531
f	Online shopping is mostly done because everything we need are in one place, and we can do it out of the comfort of our own home.	agreement (*is*), preposition (*from*)	0.515
g	Fire is advantages and sometimes disadvantages.	verb choice (*has*)	0.511
h	In my opinion, first, sn s has made our lives better.		0.507
i	It can be enjoyed regardless of age, but it's usually enjoyed by older people.		0.505
j	It also contributed to technology and civilization development by inviting factory automation and technology advancement.		0.505

Table 7 shows the sentences that are classified as nonnative with the lowest probability, just over 0.5. This means that the model classified these sentences with lower confidence than those in Table 6. Compared with those in Table 6, the sentences show fewer grammatical errors. Errors involve simple ones such as the article in (b) and preposition in (f). Other errors include inanimate possessor in (d), repeated and redundant nouns in (e), and agreement in (f). In (d), the possessive ending (*'s*) is used for the inanimate noun (**baseball's*) which can be improved by using *of*-phrase (*the ticket price of baseball*). In (e), *industry* is repeated before and after *and*, and *industries* is redundant. In (f), the verb must agree with *everything* (**are* for *is*). There is also a verb choice error in (g) where the verb should be *has*, not *is*.

To summarize 3.4 and 3.5, the sentences that are classified as native with high probability have much fewer errors than those classified as nonnative with high probability (Table 4 vs. Table 6). The sentences that are classified with low probability in either native or nonnative category do not show very noticeable such differences.

4. The Effects of Correcting Errors

The sentences in Sections 3.4-3.5 that contain errors were corrected to examine the effects of correction on the classification by fastText. For this, the following errors were corrected: articles, number (singular or plural form), prepositions, subject-verb agreement, and word choice. These are the types of errors that are frequently found in Korean college students' English writing (Kim 2018, Kim 2020). Sentences were minimally corrected, mostly one at a time, instead of rewriting the whole sentences. This is to examine the effect of each correction independently. In all the examples in this section, the first sentences (sentence (a)s in (1)-(11)) are the original student sentence, and the rest are corrected sentences. Corrected parts are underlined. For each sentence, its predicted category and

the probability are shown in parentheses.

4.1 Articles and Number

Correcting articles and number does not make any significant improvements, as shown in (1)-(3). In (1b), the article *a* is removed; in (1c), *virus* is changed to plural form (*viruses*). None of these made any improvement in the result. The same is observed in (2). Adding an article as in (2b) or changing the noun *dictionary* to the plural form as in (2c) does not make any significant changes in the result. In (3), the uncountable noun (**informations*) is corrected to the singular form, but there is no change in the result.

(1) a. Although the internet has a side effects like addiction, virus and unsuitable information but it can be prevented if you use the internet carefully. (Nonnative, 1)
b. Although the internet has side effects like addiction, virus and unsuitable information but it can be prevented if you use the internet carefully. (Nonnative, 1)
c. Although the internet has side effects like addiction, viruses and unsuitable information but it can be prevented if you use the internet carefully. (Nonnative, 0.99)

(2) a. And when we don't know the word, we can use internet dictionary (Nonnative, 1)
b. And when we don't know the word, we can use an internet dictionary (Nonnative, 0.99)
c. And when we don't know the word, we can use internet dictionaries (Nonnative, 1)

(3) a. As many people use the internet, a lot of informations from a lot of people who use Internet are shared in internet. (Nonnative, 1)
b. As many people use the internet, a lot of information from a lot of people who use Internet are shared in internet. (Nonnative, 1)

On the other hand, in Park et al. (2019), the presence/absence of the article affected the classification. The sentence without the article 'the' before 'same' was classified as nonnative, indicating that the collocation pattern of 'the same' is recognized by the model.

4.2 Sentence Fragments and Voice

Errors involving sentence fragments (sentences without the main clause) (Folse et al. 2020, Kim 2020) showed mixed results, depending on the conjunction. In sentence (4a), a sentence fragment with *although*, is classified as native with a probability of 1. Removing the conjunction as in (4b) does not much change the result.

(4) a. Although there are advantages and disadvantages both. (Native, 1)
b. There are advantages and disadvantages both. (Native 0.99)

(5) a. Because badminton plays in a stadium. (Native, 0.51)
b. Badminton plays in a stadium. (Native, 0.99)
c. Because badminton is played in a stadium. (Nonnative, 0.93)
d. Badminton is played in a stadium. (Native, 0.95)

On the other hand, sentence fragments with the conjunction *because* show different results. In (5), the original student sentence (5a) is a sentence fragment with the conjunction *because*. This is classified as native, but with a very low probability of only 0.51, nearly nonnative. Removing the conjunction, as in (5b), substantially increases the probability of being native from 0.51 to 0.99. Thus, we can see that a sentence fragment sounds nonnative if the conjunction is *because*. Korean speakers make errors using a sentence fragment with *because* (Kim 2020: 145). On the other hand, a sentence fragment with *although* is rare, so the model cannot differentiate (4a) and (4b), due to the absence of such errors in the training data. Lack of negative evidence in the training data has been pointed out as one source of misclassification by deep neural networks (Park et al. 2019: 217).

Furthermore, the verb form in (5a) is changed to active voice, as in (5c), leaving the conjunction as is. Comparing (5a) and (5c), correcting voice changes the category of the sentence to nonnative, which is an unexpected result. Further removing the conjunction from (5c), as in (5d), changes the category back to native. This reaffirms that, for the model, a sentence fragment with the conjunction *because* is a strong cue for a nonnative sentence, but voice is not.

That is, the ungrammatical sentence (5b) and the grammatical sentence (5d) are not properly differentiated: both are classified as native with high probability. This result shows that the model does not detect the voice error properly, as is also the case in Park et al. (2019). According to Park et al. (2019), the English education in Korea emphasized teaching passive verb forms, so the learners do not often make such errors, and therefore, they are underrepresented in the training data, providing not enough information for the model to make an accurate prediction when it comes to passive verb forms.

4.3 Subject-verb Agreement with Intervening Modifiers

The model did not make correct predictions when there is a subject-verb agreement error, especially when the subject head noun and the verb are separated by a modifying phrase. A similar issue has also been reported in long short-term memory (LSTM) neural networks (Linzen, Dupoux and Goldberg 2016).

In the student sentence (6a), the subject is plural (*people*) whereas the verb is singular (*has*). Despite the grammatical error, the sentence is classified as native, though the probability is low. Correcting the error by changing the verb, as in (6b) reverses the classification to nonnative. Adding the preposition *in* does not make a significant change. Sentence (d) is classified as native, where the preposition is correctly added, but the subject-verb agreement is violated. Comparing (6c) and (6d), it is evident that the model looks at the sequential linear order only, not syntactic dependencies. The immediately preceding word (*engineering*) is a singular form, so the singular verb form is considered correct and more nativelike, though the correct subject head noun is plural.

(6) a. As people who majored engineering has increased and accepting technologies from abroad, they have improved transportation. (Native, 0.506)
b. As people who majored engineering have increased and accepting technologies from abroad, they have improved transportation. (Nonnative, 0.746)
c. As people who majored in engineering have increased and accepting technologies from abroad, they have improved transportation. (Nonnative, 0.630)
d. As people who majored in engineering has increased and accepting technologies from abroad, they have improved transportation. (Native, 0.625)

This result suggests that fastText does not capture the subject-verb agreement when there is an intervening modifier, as in other neural networks (Linzen et al. 2016).

4.4 Word Choice

Replacing words have the greatest effect on the result, though there are some mixed results. Changing words did not make any change in (7) and (8), whereas it made categorical changes in (9) and (10). In (7), *unsuitable* is replaced with *inappropriate*, which is a more frequent collocation (frequency according to google search: *unsuitable information* (17,500), *inappropriate information* (4.32 million)). Both sentences are classified as nonnative with a probability of near 1. In (8), changing the verb to *look up* did not change the result. This could be because the learning data was not big enough to learn these collocations.

(7) a. Although the internet has side effects like addiction, viruses and unsuitable information but it can be prevented if you use the internet carefully. (Nonnative, 0.99)
b. Although the internet has side effects like addiction, viruses and inappropriate information but it can be prevented if you use the internet carefully. (Nonnative, 0.99)

(8) a. Besides, when I get lost, I can watch the map on the internet. (Nonnative, 1)
b. Besides, when I get lost, I can look up the map on the internet. (Nonnative, 1)

The student sentence (9a) is corrected to (9b) by changing the inanimate possessor, but it is still classified as nonnative with even a higher probability. However, changing the adjective *proper* in (9b) to *reasonable* as in (9c) reversed the classification as native with a high probability.

(9) a. Baseball's price of tickets is proper. (Nonnative, 0.543)
b. The ticket price for baseball is proper. (Nonnative, 0.658)
c. The ticket price for baseball is reasonable. (Native, 0.890)

(10) a. Online shopping is mostly done because everything we need are in one place, and we can do it out of the comfort of our own home. (Nonnative, 0.533)
b. Online shopping is mostly done because everything we need is in one place, and we can do it out of the comfort of our own home. (Native, 0.542)
c. Online shopping is mostly done because everything we need is in one place, and we can do it from the comfort of our own home. (Native, 0.599)

In (10), the plural verb (*are*) in (10a) is changed to the singular verb (*is*) in (10b), which made the sentence classified as native. It is unlikely that the model recognized the long-distant subject head *everything* (considering (6)). Instead, it is more likely that the immediately preceding word *need* is wrongly considered as a singular form and so the following singular verb is considered correct by the model. In addition, changing the preposition as in (10c) slightly increased the probability of sounding native.

(11) a. Fire is advantages and sometimes disadvantages. (Nonnative, 0.531)
b. Fire has advantages and sometimes disadvantages. (Native, 0.978)

In (11), changing the verb had a very clear effect on the classification. Replacing the verb *is* in (11a) with *has* as in (11b) resulted in a category change, with the probability of near 1. All in all, despite some mixed results, it is

evident that replacing words have far greater impacts on the classification than correcting articles, number, or subject-verb agreement with intervening modifiers. This indicates that the model is more sensitive to the choice of the words rather than grammatical errors. This result conforms to the previous literature on deep neural network models (Park et al. 2019, 2020).

5. Discussion and Conclusion

In this paper, I explored the performance of fastText, a text classifier that has a shallow neural-network architecture based on sub-word information. Overall, fastText performed well for the task of binary classification of native and nonnative sentences. Regarding the research question 1, fastText achieved a validation accuracy of 97.75% and a test accuracy of 68.9%. The validation accuracy is somewhat higher than in the previous research (Park et al. 2019, Lee et al. 2021). However, a direct comparison of the accuracy values is not meaningful because the training data was different in size and homogeneity. In particular, whereas Park et al. (2019) used the corpora from various genres, the training data in the present study was domain-specific (student writing) and the size was smaller.

Regarding the research question 2, the sentences classified as native with a high probability had fewer errors than those classified as nonnative with a high probability. The sentences classified with a low probability (just over 0.5) did not have clear differences between native and nonnative categories in terms of errors. In addition, the probability of classification was higher when the model classified sentences as nonnative, which means that the model was more confident with the nonnative category, and less confident with the native category, when the test data was nonnative sentences.

Regarding the research question 3, experiments with correcting errors revealed that fastText tends to be better at spotting word-choice errors (plausibility error) rather than grammatical errors involving articles, number, passive verb form. Such results are similar to the findings in Park et al. (2019), where, for example, incorrect passive verb forms are classified as native. Both models are more sensitive to word choices rather than grammaticality, so probably this is a common characteristic of neural-network models. Grammaticality often depends on the hierarchical syntactic structure rather than flat, linear word ordering (e.g., subject-verb agreement), whereas word collocation is a matter of the linear combination of words. Thus, grammatical errors seem harder to detect by neural network models that are dependent on the linear ordering of words.

A limitation of the present study is that the training data was small (23,829 sentences), compared to those in other similar research: ca. 610K sentences (Lee et al. 2021), 586K sentences (Park et al. 2019). The genre was also limited to college students' writing, so it is difficult to generalize our findings to other genres. Another general limitation is that fastText learns from linear ordering of words, rather than hierarchical structure. So, it cannot learn hierarchical syntactic structures such as subject-verb agreement with an intervening modifier, which is a common problem in deep learning models in the previous research (e.g., Linzen et al. 2016, Park et al. 2019).

In future research, we may consider classifying sentences into three categories. For example, Harust et al. (2020) built a neural-network classifier for the task of native-like expression identification (NLEI). Their classifier is designed to return three labels: *native*, *neutral*, and *L2*, instead of just two, native/nonnative categories, in order to identify distinctively nativelike sentences among the sentences written by native speakers. Excluding neutral sentences, their classifier picked out combinations of very basic words that are used by native speakers but not by L2 speakers (e.g., *to be a hard sell*, *sit out a bit*). In our present study, the sentences that are classified with a low probability, close to 0.5, could actually have neutral properties – sentences that can be produced by both native

and nonnative speakers. From a pedagogical point of view, these 'neutral' sentences may not be very informative and have less priority in teaching. On the other hand, the expressions that are mostly used by native speakers but not by L2 speakers are more worth spending class time on. The state-of-the-art neural-network technology and other similar ICT technologies can help find native-like patterns in a sentence, which will make second language learning more efficient.

References

Bojanowski, P., E. Grave, A. Joulin and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics,* 5. 135–146. doi: https://doi.org/10.1162/tacl_a_00051

Carlstrom, B. and N. Price. 2012-2014. The Gachon Learner Corpus. Available online at http://koreanlearnercorpusblog.blogspot.kr/p/corpus.html.

Chang, Y. 2018. Features of lexical collocations in L2 writing: A case of Korean adult learners of English. *English Teaching* 73(2), 3-36.

Cho, K., B. Merrienboer, C. Gehre, F. Bougares, H. Schwenk, H. Schwenk and Y. Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. CoRR,abs/1406.1078,2014. URL http://arxiv.org/abs/1406.1078

Cook, V. 1999. Going beyond the Native Speaker in Language Teaching. *TESOL Quarterly*, 33(2), 185-209, http://www.viviancook.uk/Writings/Papers/NS1999.htm

Cummins, J. 2005. Language proficiency, bilingualism and academic achievement. In P. A. Richard-Amato and M. A. Snow, eds., *Academic Success for English Learners: Strategies for K-12 Mainstream Teachers,* 76-86. White Plains, NY: Pearson.

Davies, A. 1991. *The Native Speaker in Applied Linguistics*. Edinburgh: Edinburgh University Press.

Davies, M. 2008-. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Folse, K. S., A. Muchmore-Vokoun and E.V. tri Solomon. 2020. *College Writing 2: Great Paragraphs*. National Geographic Learning.

Goldin, G., E. Rabinovich and S. Wintner. 2018. Native language identification with user generated content. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3591–3601.

Goodfellow, I., Y. Bengio and A. Courville. 2016. *Deep Learning*. The MIT Press.

Granger, S. 1998. The computer learner corpus: a versatile new source of data for SLA research. In S. Granger, ed., *Learner English on Computer*, 3-18, London: Longman.

Harust, O., Y. Murawaki and S. Kurohashi. 2020. Native-like expression identification by contrasting native and proficient second language speakers. *Proceedings of the 28th International Conference on Computational Linguistics*, 5843–5854.

Nam, D. 2011. Native-speakerness: A case study of two Korean-English bilinguals. *Studies in Foreign Language Education* 25(2), 171-193.

Joulin, A., E. Grave, P. Bojanowski and T. Mikolov. 2017. Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* 2, Short Papers, 427-431.

Jung, W. 2005. Attitudes of Korean EFL learners towards varieties of English. *English Teaching* 60(4), 239-260.

Kim, Y. J. 2014. A study of learner's errors in verbs and articles by two different levels of college students, *The Journal of English Language and Literature* 19(3), 139-164

Kim, J. 2020. A corpus-based analysis of syntactic/semantic errors in university level EFL learners' writing. *Lingua Humanities* 22(2), 135-156.

Kim, S. 2018. Grammatical and lexical errors in Korean college students' writing at a low level of proficiency. *Journal of the Korea English Education Society* 17(3), 77-92.

Lee, J. J. 2005. The native speaker: An achievable model? *Asian EFL Journal* 7(2), 152-163.

Lee, J., J. Kim and H. Kim. 2021. A study on the judgment of nativelikeness of Korean learner corpus by deep learning language model. *Language and Culture* 17(1), 155-177.

Lee, S. 2006. A corpus-based analysis of Korean EFL learners' use of amplifier collocations, *English Teaching* 61(1), 3-17.

Linzen, T., E. Dupoux and Y. Goldberg. 2016. Assessing the ability of LSTM to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4, 521-535.

Morgan-Short, K., I. Finger, S. Grey and M. T. Ullman. 2012. Second language processing shows increased native-like neural responses after months of no exposure. *PLoS ONE* 7(3), e32974. https://doi.org/10.1371/journal.pone.0032974

Nation, I.S.P. 2008. *Teaching Vocabulary: Strategies and Techniques*. Boston: Heinle ELT.

Park, K. 2020. *A Comparison of Nativeness Judgments by the Deep Learning Systems and Humans on Korean EFL Learner Sentences*. Master's Thesis, Korea University.

Park, K., S. You and S. Song. 2019. Using the deep learning techniques for understanding the nativelikeness of Korean EFL learners. *Language Facts and Perspectives* 48, 195-227.

Park, K., S. You and S. Song. 2020. Not yet as native as native speakers: Comparing deep learning predictions and human judgments. *English Language and Linguistics* 26(1), 199-228.

Pawley, A. and F. H. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt, eds., *Language and Communication*, 191-226. New York: Longman.

Rhee, S.-C. and C. K. Jung. 2014. Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *The Journal of the Korea Contents Association* 14(11), 1019–1029. https://doi.org/10.5392/JKCA.2014.14.11.1019

Scovel, T. 1988. *A Time to Speak: A Psycholinguistic Inquiry into Critical Period for Human Speech*. New York: Harper and Row.

Seo, H. and J. Shin. 2020. Data preprocessing and transformation in the sentiment analysis using a deep learning technique. *Korean Journal of English Language and Linguistics* 20, 42-63.

Šišić, E. 2016. *EFL Learners' Attitudes Towards Native-Like Proficiency as an Achievement Target*. Graduation Thesis, University of Zagreb.

Weston, J., S. Chopra and K. Adams. 2014. #tagspace: Semantic embeddings from hashtags. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1822-1827.

Xiao, Y. and K. Cho. 2016. Efficient character-level document classification by combining convolution and recurrent layers. Available at https://arxiv.org/abs/1602.00367

Zareva, A., P. Schwanenflugel and Y. Nikolova 2005. Relationship between lexical competence and language proficiency. *Studies in Second Language Acquisition* 27(4), 567-595.

Zhang, X. and Y. LeCun. 2015. Text understanding from scratch. Available at http://arxiv.org/abs/1502.01710

Zhang, X., J. Zhao and Y. LeCun. 2015. Character-level convolutional networks for text classification. *Proceedings of the 28th International Conference on Neural Information Processing Systems* 1, 649–657.

Zhao, J. 2017. Native speaker advantage in academic writing?: Conjunctive realizations in EAP writing by four groups of writers. *Ampersand* 4, 47-57.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

Appendix

fastText Commands (Examples)

Training data file name: train.txt
Validation data file name: valid.txt
Test data file name: test.txt
Test results are saved in test.predict.txt
Name of the language model: my.model
Number of epochs: 25
Word N-grams: 2

Training:
~% fasttext supervised -input ./train.txt -output my.model -epoch 25 -wordNgrams 2
Validation:
~% fasttext test my.model.bin valid.txt
Testing:
~% fasttext predict-prob my.model.bin test.txt > test.predict.txt