Simple Yet Powerful Native Language Identification on TOEFL11

Ching-Yi Wu
University of Texas at Dallas
800 W Campbell Rd
Richardson, TX, USA
cxw120631@utdallas.edu

Po-Hsiang Lai
Emerging Technology Lab
Samsung R&D - Dallas
1301 Lookout Drive
Plano, TX, USA
s.lai@samsung.com

Yang Liu **Vincent Ng**
University of Texas at Dallas
800 W Campbell Rd
Richardson, TX, USA
yangl@hlt.utdallas.edu
vince@hlt.utdallas.edu

Abstract

Native language identification (NLI) is the task to determine the native language of the author based on an essay written in a second language. NLI is often treated as a classification problem. In this paper, we use the TOEFL11 data set which consists of more data, in terms of the amount of essays and languages, and less biased across prompts, i.e., topics, of essays. We demonstrate that even using word level n-grams as features, and support vector machine (SVM) as a classifier can yield nearly 80% accuracy. We observe that the accuracy of a binary-based word level n-gram representation (~80%) is much better than the performance of a frequency-based word level n-gram representation (~20%). Notably, comparable results can be achieved without removing punctuation marks, suggesting a very simple baseline system for NLI.

1 Introduction

Native language identification (NLI) is an emerging field in the natural language processing community and machine learning community (Koppel et al., 2005; Blanchard et al., 2013). It is a task to identify the native language (L1) of an author based on his/her texts written in a second language. The application of NLI can bring many benefits, such as providing a learner adaptive feedback of their writing errors based on the native language for educational purposes (Koppel et al., 2005; Blanchard et al., 2013).

NLI can be viewed as a classification problem. In a classification problem, a classifier is first trained using a set of training examples. Each training example is represented as a set of features, along with a class label. After a classifier is trained, the classifier is evaluated using a testing set (Murphy, 2012). Good data representation often yields a better classification performance (Murphy, 2012). Often time, the simpler representations might produce better performance. In this work, we demonstrate that a binary-based word level n-gram representation yields much better performance than a frequency-based word level n-gram representation. In addition, we observed that removing punctuation marks in an essay does not make too much difference in a classification performance.

The contributions of this paper are to demonstrate the usefulness of a binary-based word level n-gram representation, and a very simple baseline system without the need of removing punctuation marks and stop words.

This paper is organized as the following. In Section 2, we present related literatures. TOEFL11 data set is introduced in Section 3. In Section 4, our features and system design are described. The results are presented in Section 5, followed by conclusion in Section 6.

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 152–156,
Atlanta, Georgia, June 13 2013.

2 Related Work

The work by Koppel et al. (2005) is the first study to investigate native language identification. They use the International Corpus of Learner English (ICLE). They set up this task as a classification problem studied in machine learning community. They use three types of features: function words, character n-gram, errors and idiosyncrasies, e.g. spelling and grammatical errors. For errors and idiosyncrasies, they used Microsoft Office Word to detect those errors. Their features were evaluated on a subset of the ICLE corpus, including essays sampled from five native languages (Russian, Czech, Bulgarian, French and Spanish) with 10-fold cross validation. They achieve an accuracy of 80.2% by combining all of the features and using a support vector machine as the classification algorithm. In addition, Tsur and Rappoport (2007) show that using character n-gram only on the ICLE can yield an accuracy of 66%.

The work from Kochmar (2011) identifies an author's native language using error analysis. She suggests that writers with different native languages generate different grammatical error patterns. Instead of using ICLE, this work uses a different corpus, English learner essays from the Cambridge Learner Corpus. She uses SVM on manually annotated spelling and grammatical errors along with lexical features.

Most of the systems described in NLI literature reach good performance in predicting an author's native language, using character n-gram and part of speech n-gram as features (Blanchard et al., 2013). In recent years, various studies have started to look into complex features in order to improve the performance. Wong and Dras (2009) use contrastive analysis, a systematic analysis of structural similarities and differences in a pair of languages. A writer's native language influences the target language they aim to learn. They explore the impact of three English as Second Language (ESL) error types, subject-verb disagreement, noun-number disagreement and determiner errors, and use a subset of ICLE with 7 languages. However, although the determiner error feature seems useful, when it is combined with a baseline model of lexical features, the classification performance is not significantly improved (Wong and Dras, 2009).

Wong and Dras (2011) use complex features such as production rules from two parsers and reranking features into the classification framework, incorporating lexical features of Koppel et al. (2005). They achieve a classification performance of 81.71% on the 7-native-languages NLI, slightly better than 80.2% accuracy of the original Koppel et al. (2005).

Note that although the International Corpus of Learner English (ICLE) is used in most of the NLI studies, ICLE has been known to have fewer essays, and a skewed distribution toward topics of essays (Blanchard et al., 2013). In addition, even though there are 16 native languages in ICLE, as each language has different numbers of essays, most work often uses different subsets of 7 native languages, which makes comparison harder across different studies (Blanchard et al., 2013). The NLI shared task 2013 provides a new data set, namely the TOEFL11 (Blanchard et al., 2013), which addresses these issues. As previously discussed, complex features do not necessarily improve classification accuracy. In this work, we use TOEFL11 to investigate the classification performance using simple word n-gram based features.

3 Data

In this work, we use TOEFL11 as our corpus. TOEFL11 is a new data set for NLI (Blanchard et al., 2013). There are 11 native languages, including Arabic (ARA), Chinese (CHI), French (French), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR). Authors write essays based on 8 different topics in English. There are 1,100 essays for each language, and sampled from 8 different topics, i.e., prompts. Each essay is also annotated with an English proficiency level (low/medium/high) determined by assessment specialists. Among 12,100 essays, there are 9,900 essays in the training set, 1,100 essays in the development set, i.e., validation set in machine learning, and 1,100 essays in the testing set. In the training set and the development set, there are equal numbers of essays from each of the 11 native languages. By using TOEFL11, it makes our analysis less biased toward a specific topic of essays (Blanchard et al., 2013).

4 NIL System Design

In this section, we describe our NLI system, the features, and the classifier we use.

4.1 Data Preprocessing

Each essay is tokenized, and then capitalizations are removed. Note that we did not remove English stop words, which might be useful to discriminate the native language for a writer. For example, function words, which belong to stop words, such as '*the*', '*at*', '*which*', have been proven to be effective to distinguish native language for writers (Koppel et al., 2005). There are two settings: either punctuation marks are removed or kept. When punctuation marks are kept, they are viewed the same as word in constructing n-grams. For example, in the sentence "NLI is fun.", "fun ." is viewed as a bigram.

4.2 Features

In our system, word level n-grams are used to represent an essay. Previous studies have shown that word level n-grams are useful in determining the native language of a writer (Bykh and Meurers, 2012). One reasonable hypothesis is that non-native English writers with the same native languages tend to choose more similar words to express the same or similar concepts. In addition, the combination of a sequence of words might also be affected by the different native language of writers. Therefore, word n-gram is useful to distinguish the native language of a writer. Even though some previous studies have looked into using word level n-grams as features, how to use word level n-grams has not been explored too much yet on TOEFL11 corpus. To our knowledge, the most recent study by Blanchard et al. (2013) started to research the effect of different forms of word level n-gram representations.

There could be many ways to represent an essay by word level n-grams. One possible representation of an essay is to use the frequency of a specific word n-gram, i.e., the number of times a specific word n-gram appears in an essay divided by the number of times all word n-grams appear in an essay. In this representation, an essay is a vector whose elements are the frequency of different word n-grams in the essay. Another possible representation is to use binary representation, i.e., 1 indicates this word n-gram is in this essay, 0 indicates this word n-gram is not in this essay. One interesting question to ask is:

Which representation can be more informative to distinguish the native language of writers of essays?

Here we compare the performance of a frequency-based word level n-gram representation and a binary-based word level n-gram representation. We included all word level n-grams in the training set, without any frequency cutoff. For both binary-based and frequency-based representations, we run the experiments on the two settings: punctuation marks are either removed or kept.

In addition to word level n-grams, since TOEFL11 also consists of English proficiency levels evaluated by assessment experts, we also included it to test whether this feature might improve the classification performance. All of the features used in our system are summarized in Table 1. Besides each feature described above, we have also combined different features to test whether various combinations of features might improve the accuracy performance. Here, we simply aggregated different features, for example, all word level unigrams, combined with all word level bigrams.

4.3 Classifier

Previous literatures have used various methods such as Naïve Bayse, logistic regression and support vector machine on NLI problem. As it has been shown that when representing an essay in order to perform a classification task, it often results in an essay being represented in a very high dimensional space. Since support vector machine (SVM) is known to be adaptive when the feature dimension is high, we chose SVM as our classification algorithm. We also compared the results from Naïve Bayse for an experimental purpose and found that SVM is better. We use SVM-Light for our system (Joachims, 1999). We then train our SVM classifier on the training set (n=9900), and test the trained classifier on the testing set (n=1100).

5 Results and Discussions

5.1 Results

Table 1 and Table 2 show the accuracies on the testing set for the different feature sets, when punctuation marks are removed or kept respectively. As the results demonstrated, the accuracies of word level bigram are better than unigram using a binary-based representation. When combining word level unigram and bigram, the accuracy is improved in a binary-based representation. This is consistent when punctuations are either removed or kept. This observation is consistent with the existing NLI literatures: when combining word n-grams, it seems to improve the accuracy of the classifier, compared with a word n-gram alone. But we do not observe too much difference when punctuation marks are removed or kept, using both unigram and bigram. In fact, including punctuation marks lead to high accuracies in many scenarios, especially in unigram in a frequency-based representation, suggesting the usage of punctuation marks varies across native languages.

Features	Performance of Binary Word n-gram Representation	Performance of Freq. Word n-gram Representation
word unigram	70.91%	25.36%
word bigram	76.00%	17.64%
word unigram and word bigram	79.73%	23.36%

Table 1 Accuracy of Different Feature Sets, without Punctuation Marks

Features	Performance of Binary Word n-gram Representation	Performance of Freq. Word n-gram Representation
word unigram	70.18%	30.00%
word bigram	77.09%	18.73%
word unigram and word bigram	79.45%	28.73%

Table 2 Accuracy of Different Feature Sets, with Punctuation Marks

Table 3 shows the confusion matrix of classification performance, using unigram and bigram, in a binary-based representation when punctuation marks are removed. We observe that some of native languages, such as German, Italian, and Chinese, lead to better classification accuracy than for Korean, Spanish, and Arabic.

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR	Precision	Recall	F-measure
ARA	**75**	1	5	3	1	3	1	1	3	4	3	78.9	75.0	76.9
CHI	3	**86**	0	0	1	0	5	4	0	0	1	81.9	86.0	83.9
FRE	1	1	**79**	7	3	4	2	0	1	0	2	77.5	79.0	78.2
GER	3	1	2	**87**	1	1	1	0	2	0	2	79.8	87.0	83.3
HIN	1	2	1	2	**77**	0	0	0	5	10	2	74.0	77.0	75.5
ITA	0	0	6	4	0	**85**	0	0	3	0	2	83.3	85.0	84.2
JPN	2	2	1	0	0	1	**86**	3	2	0	3	77.5	86.0	81.5
KOR	0	8	2	1	1	0	14	**72**	1	1	0	82.8	72.0	77.0
SPA	4	0	6	3	4	6	1	1	**70**	1	4	78.7	70.0	74.1
TEL	1	0	0	1	15	0	0	0	0	**82**	1	83.7	82.0	82.8
TUR	5	4	0	1	1	2	1	6	2	0	**78**	79.6	78.0	78.8

Average Performance: 79.7%. Precision, Recall, F-measures are in %.

Table 3 Confusion Matrix on Testing Set

5.2 Binary Based of Word N-Gram Representation

We observe that the accuracy of a binary-based word level n-gram representation in our system is significantly better than a frequency-based representation. This is similar to the result reported by Blanchard et al., (2013) in TOEFL11 corpus. The differences between their system and ours are that the system developed by Blanchard et al., (2013) used logistic regression with L1-regularzation, instead of SVM and they did not remove all punctuation marks and special characters.

This might imply that a frequency-based word n-gram representation do not capture the characteristics of the data. This might be because the data resides in a high dimension space, and the frequencies of word level n-grams would be skewed. In a future study, one might investigate a better representation form and other complex features that have a stronger interpretative power of the data.

5.3 Effects of Proficiency Level

In our results, we have included English proficiency level (low/medium/high) as a feature provided by assessment experts. However, we did not find a strong improvement in accuracies, for example, 79.13% using a binary-based word level n-grams when punctuation marks removed. We think this might be because only one feature will

not dramatically change the accuracies. This may be due to the fact word n-grams have already contributed a large amount of features.

6 Conclusion

In this paper, we used a new data set, TOEFL11 to investigate NLI. In the most existing literatures, ICLE corpus was used. However, ICLE has fewer data and is known to be biased to topics of essays. The newly released corpus, TOEFL11 addresses these two drawbacks, which is useful for NLI community. Support vector machine (SVM) was used as a classifier in our system. We have demonstrated that a binary-based word level n-gram representation has resulted in a significantly better performance compared to a frequency-based n-gram representation. We observed that there is not much difference in classification accuracies when punctuation removed or kept, when combining both unigram and bigram. Interestingly, a frequency-based word unigram with punctuation marks outperforms than the case without punctuation marks, suggesting the potential of utilizing punctuation marks in NLI. In addition, English proficiency level has also been included in our feature set, but did not yield a significant improvement in accuracy. As most of the essays are represented in a high dimension space using word level n-grams, we are looking into feature selection to reduce dimensionality and how to represent those features in order to improve accuracy, as well as other features.

References

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. 2013. *TOEFL11: A Corpus of Non-Native English*. Educational Testing Service.

Bykh, S. and Meurers, D. 2012. *Native Language Identification using Recurring n-grams - Investigating Abstraction and Domain Dependence*. In Proceedings of COLING 2012, 425-440, Mumbai, India. The COLING 2012 Organizing Committee.

Joachims, T. 1999. *Making large-Scale SVM Learning Practical*. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press.

Kochmar, E. 2011. *Identification of a writer's native language by error analysis*. Master's thesis, University of Cambridge.

Koppel, M., Schler, J., and Zigdon, K. 2005. *Automatically determining an anonymous author's native language*. In ISI, 209–217.

Murphy, K. P. 2012. Machine learning: a probabilistic perspective. MIT Press.

Tsur, O. and Rappoport, A. 2007. *Using classifier features for studying the effect of native language on the choice of written second language words*. In Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, 9–16, Prague, Czech Republic. Association for Computational Linguistics.

Wong, S.-M. J. and Dras, M. 2009. *Contrastive analysis and native language identification*. In Proceedings of the Australasian Language Technology Association Workshop 2009, 53–61, Sydney, Australia.

Wong, S.-M. J. and Dras, M. 2011. *Exploiting parse structures for native language identification.* In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 1600–1610, Edinburgh, Scotland, UK. Association for Computational Linguistics.