Identifying the L1 of non-English writers

Zhaomin Xiao
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA
zhx36@pitt.edu

Abstract

I demonstrate one possible way to identify the native language of the non-native English authors. I extract various linguistics features for classification. Feature selection is based on logistic regression. The classification methods include SVM and logistic regression. The methods that yielded the highest accuracy are also reported.

1 Introduction

The task I focused is to identify the L1 of non-native English authors. More specifically, what I have done is to find the native language of TOEFL essay writer whose native language is one of 11 possible languages, Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, or Turkish.

Many researchers have done great researches in this area. Some of them focus on the stylistic features, some of them use many lexical features, and some of them are more interested in syntactic features. And of course, these features could be combined to form a more informative feature set. In my work, most of the features I use are syntactic features and lexical features. Besides of feature extraction, for classification methods, I used many traditional machine learning classification methods, like support vector machine, logistic regression, and linear support classifier. Most of previous work used SVM which brought the highest precision, recall, and F1 score at most times. In my experiment, logistic regression yielded the best result.

I address this task as a multi-class classification task. I describe my data in section 3 and classification methods in section 4. Like similar NLI tasks, the choice of feature is crucial. I discuss the features I have used in section 5. I report my results in section 6 and conclude with advices for future research.

2 Related Work

The Native Language Identification (NLI) task was introduced by Koppel et al (2005a; 2005b). Tsvetkov et al (2013) used various features, including lexical features, stylistic features, and syntactic features, and the final F1 score can be slightly above 81%. Besides of traditional machine learning classification methods, Sari et al (2017) also used neural network model and also obtained similar accuracy.

3 Data

There are many available dataset for NLI task. Besides the TOEFL11 corpus which is the very common for NLI task, there are also other datasets that we can use, such as the ASK Corpus (2013; 2006) which is designed for the learners of Norwegian, and the Jinan Chinese Learner Corpus (2015) which is designed for Chinese learners.

The dataset I use is the TOEFL11 corpus, which consists of TOEFL essays. Besides of essays, it also contains the English proficiency (low, medium, high) of author and the prompt (1 to 8) of essays. There are 9900 essays in training data, 1100 essays in development set, and 1100 essays in test set.

4 Model

The classification methods I use are mostly traditional machine learning classification methods. Empirically, SVM can yield the best result. While in my experiment, logistic regression with ℓ_2 penalty is the best methods. The best result is in 1. Visualized result are in Figure 1. Parameters of logistic regression are tuned using GridSearchCV

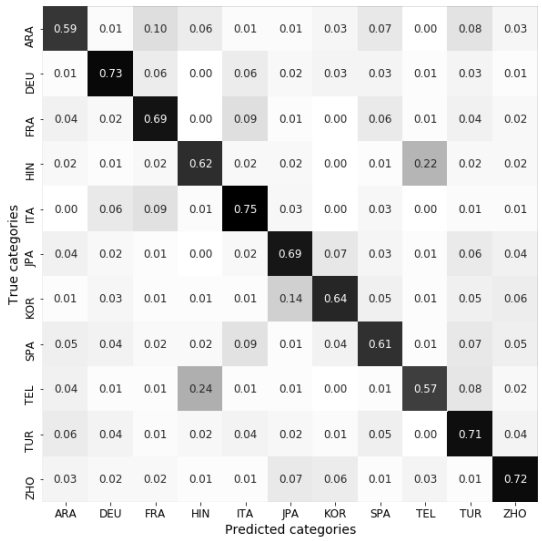

Figure 1: Visualized best result.

with 5-fold crossvalidation in `sklearn`. Also, ensemble methods like boosting and bagging works not well in this task.

5 Features

Like other NLP problems, feature is the crucial part. The most important features are described in 5.1. Other features are described in 5.2. Although features in 5.1 is much more important than the features in 5.2. The latter also contributes to the final result.

After extracting features, I also use logistic regression with ℓ_1 penalty to select feature. Feature selection significantly reduces the number of dimension of feature vector from more than 10000 to 248.

5.1 Main Features

The most important features are POS n-gram and word n-gram. These two features are stored in POS-based TF-IDF matrix and word-based TF-IDF matrix, respectively. The reason why I use TF-IDF matrix is that TF-IDF matrix can take the importance of each POS n-gram and each word n-gram, which can solve the problems caused by using only raw counts of n-gram.

POS n-gram The probability distribution of POS n-gram is skewed for author with some specific L1s. Some POS n-grams are very common for these authors. In my experiment, I use POS unigram, bigram, and trigram. N-grams with higher order are not informative enough to be included.

Word n-gram Features are extracted to count every word unigram, bigram and trigram.

5.2 Additional Features

Besides of main features above, I have also tried to use other features. These features are mostly from linguistic intuition. The improvements yielded by adding these features are not significant.

Document Length The length of document in terms of the number of tokens is highly correlated with the author's L1.

Word Length Some authors are more likely to use long words while others are more interested in short words. The average length of tokens can be considered as one feature.

Stop Words Stop words are the most common words in one language. The set of stop words I use is part of corpora in NLTK. The feature is the number of stop words.

Punctuation Different languages use punctuation differently.

5.3 Discarded Features

I also discard some features as follows. Although adding these features will improve model's performance intuitively, the performance are worse than before add these features. Among these discarded features, Lemma n-gram and character n-gram are overlap with word n-gram so that using all of these three features is not necessary. In my experiment, using any one of these three features along leads to similar result.

Passive Verbs The ratio of the number of passive verbs to the total number of verbs could be a significant indictor of L1.

Lemma n-gram Unlike word n-gram, lemma n-gram can emphasize the base form of words, not distracted by grammatical constraint.

Spelling Error Both of the number of spelling errors and the degree of diversity of spelling errors are related to identify author's L1.

Char n-gram Like POS n-gram, the degree of skewness of probability distribution of char n-gram is also a possible indictor of L1.

Function Words Function words are words that have little lexical meaning but express grammatical relationships among other words within a sentence. I use a list of 34 function words.

L1	Precision (%)	Recall (%)	F1 (%)
ARA	59	59	59
DEU	73	73	73
FRA	69	69	69
HIN	62	62	62
ITA	63	75	69
JPN	67	69	68
KOR	76	64	69
SPA	67	61	64
TEL	66	57	61
TUR	58	71	64
ZHO	70	72	71

Table 1: Best Result on training set of different L1. Parameters are tuned based on GridSearch with 5-fold cross-validation.

6 Results

The full model that I used to classify test set combines that features in section 5.1. Using these features, the F1 score on combined set, consisting of training set and development set, is 67%, on test set is 69%. Table 1 lists the precision, recall and F1 score for each L1. Among these 11 L1s, the predictions of Hindi and Telugu are very close to each other. This makes sense, since many Hindi and Telugu speakers are native speakers of Indian English. People with the same native language have similar language usage preference.

The discussion about this result can also be divided into three parts, based on the types of features. The first part is syntax, among the main features in 5.1, the POS n-gram is the feature that has the most significant effect on the result. The reason for that is obvious. For example, there is no article in Chinese, then Chinese speakers always tends to use fewer articles than people whose L1 has article. This will lead to a skewed distribution of POS n-gram in their essays.

Secondly, from the aspect of usage of vocabulary, word n-gram is very informative. Since Arabic has no indefinite article, some researchers (2013) speculate that Arabic speakers tend to use *alot* instead of *a lot*. Another interesting thing is, among these kinds of n-gram, the confusions of *l* and *r* and the frequent usage of hyphen are common in Japanese essays and German essays, respectively. People speaking Indian English are likely to use *then*. Thus, these specific n-grams are very informative since only people with specific L1 could use it.

Lastly, different language speakers tend to write essays in different length. In average, each of Arabic's essays has approximately 300 tokens which is just little above the length requirement of TOEFL essay, which is 300. But each of German's essays has approximately 390 tokens. Thus, the difference between the length of essays in terms of the number of tokens can provide us much information.

7 Future

And for feature extraction, besides of the features in this paper, other feature like passive verbs, which is also an indicator of L1, might also helpful for NLI task. But these linguistic features are not very reliable. They are mostly coming from intuition. As far as I am concerned, focusing more on IR-based techniques might be a good choice.

Speaking of classification methods, recent deep learning techniques like sequence to sequence model and long short-term memory can also yield good prediction, although the theory behind it is not very clear. Also, training a deep neural network is very time-consuming compared to the traditional machine learning methods.

References

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005a. *Automatically determining an anonymous authors native language*. Springer.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005b. Determining an author's native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Chicago, IL, pages 624–628.

Yunita Sari, Muhammad Fatchurrahman, and Meisyarah Dwiastuti. 2017. A shallow neural network for native language identification with character n-grams. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*. ACL, Denmark.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ask corpus: A language learner corpus of norwegian as a second language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. ELRA, Italy, pages 1821–1824.

Kari Tenfjord, Paul Meurer, and Silje Ragnhildstveit. 2013. Norsk andrespr akskorpus - a corpus of norwegian as a second language. In *Learner Corpus Research Conference*. LCA, Norway.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. 2013. Identifying the l1 of non-native writers: the cmu-haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, Atlanta, GA, pages 279–287.

Maolin Wang, Shervin Malmasi, and Mingxuan Huang. 2015. The jinan chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, Denver, CO, pages 118–123.