Recognizing English Learners' Native Language from Their Writings

Baoli LI
Department of Computer Science
Henan University of Technology
1 Lotus Street, High & New Technology Industrial Development Zone
Zhengzhou, China, 450001
csblli@gmail.com

Abstract

Native Language Identification (NLI), which tries to identify the native language (L1) of a second language learner based on their writings, is helpful for advancing second language learning and authorship profiling in forensic linguistics. With the availability of relevant data resources, much work has been done to explore the native language of a foreign language learner. In this report, we present our system for the first shared task in Native Language Identification (NLI). We use a linear SVM classifier and explore features of words, word and character n-grams, style, and metadata. Our official system achieves accuracy of 0.773, which ranks it 18th among the 29 teams in the closed track.

1 Introduction

Native Language Identification (NLI) (Ahn, 2011; Kochmar, 2011), which tries to identify the native language (L1) of a second language learner based on their writings, is expected to be helpful for advancing second language learning and authorship profiling in forensic linguistics. With the availability of relevant data resources, much work has been done to explore the effective way to identify the native language of a foreign language learner (Koppel et al., 2005; Wong et al., 2011; Brooke and Hirst, 2012a, 2012b; Bykh and Meurers, 2012; Crossley and McNamara, 2012; Jarvis et al., 2012; Jarvis and Paquot, 2012; Tofighi et al., 2012; Torney et al. 2012).

To evaluate different techniques and approaches to Native Language Identification with the same setting, the first shared task in Native Language Identification (NLI) was organized by researchers from Nuance Communications and Educational Testing Service (Tetreault et al., 2013). A larger and more reliable data set, TOEFL11 (Blanchard et al., 2013), was used in this open evaluation.

This paper reports our NLI2013 shared task system that we built at the Department of Computer Science, Henan University of Technology, China. To be involved in this evaluation, we would like to obtain a more thorough knowledge of the research on native language identification and its state-of-the-art, as we may focus on authorship attribution (Koppel et al., 2008) problems in the near future.

The NLI2013 shared task is framed as a supervised text classification problem where the set of native languages (L1s), i.e. categories, is known, which includes Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. A system is given a large part of the TOEFL11 dataset for training a detection model, and then makes predictions on the test writing samples.

Inspired by our experience of dealing with different text classification problems, we decide to employ a linear support vector machine (SVM) in our NLI2013 system. We plan to take this system as a starting point, and may explore other complex classifiers in the future. Although in-depth syntac-

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 119–123,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

tic features may be helpful for this kind of tasks (Bergsma et al., 2012; Wong and Dras, 2011; Swanson and Charniak, 2012; Wong et al., 2012), we decide to explore the effectiveness of the traditional word and character features, as well as style features, in our system. We would like to verify on the first open available large dataset whether these traditional features work and how good they are.

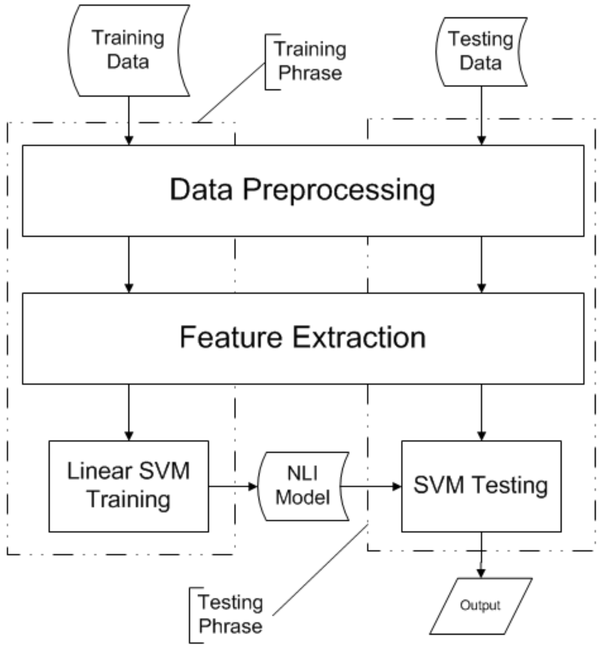

Figure 1. System Architecture.

We submitted four runs with different feature sets. The run with all the features achieved the best accuracy of 0.773, which ranks our system 18th among the 29 systems in the closed track.

In the rest of this paper we describe the detail of our system and analyze the results. Section 2 gives the overview of our system, while Section 3 discusses the various features in-depth. We present our experiments and discussions in Section 4, and conclude in Section 5.

2 System Description

Figure 1 gives the architecture of our NLI2013 system, which takes machine learning framework. At the training stage, annotated data is first processed through preprocessing and feature extraction, then fed to the classifier learning module, and we can finally obtain a NLI model. At the testing stage, each test sample goes through the same preprocessing and feature extraction modules, and is assigned a category with the learned NLI model.

Data Preprocessing: this module aims at transforming the original data into a suitable format for the system, e.g. inserting the category information into the individual writing sample and attaching metadata to essays.

Feature Extraction: this module tries to obtain all the useful features from the original data. We considered features like: word, word n-gram, character n-gram, style, and available metadata.

Linear SVM training and testing: these two modules are the key components. The training module takes the transformed digitalized vectors as input, and train an effective NLI model, where the testing module just applies the learned model on the testing data. As linear support vector machines (SVM) achieves quite good performance on a lot of text classification problems, we use this general machine learning algorithm in our NLI2013 system. The excellent SVM implementation, Libsvm (Chang and Lin, 2011), was incorporated in our system and TFIDF is used to derive the feature values in vectors. Then, we turn to focus on what features are effective for native language identification. We explore words, word n-grams, character n-grams, style, and metadata features in the system.

3 Features

In this section, we explain what kind of features we used in our NLI2013 system.

3.1 Word and Word n-gram

The initial feature set is words or tokens in the dataset. As the dataset is tokenized and sentence/paragraph split, we simply use space to delimit the text and get individual tokens. We remove rare features that appear only once in the training dataset. Words or tokens are transformed to lowercase.

Word n-grams are combined by consecutive words or tokens. They are expecting to capture some syntactic characteristics of writing samples. Two special tokens, “BOS” and “EOS”, which indicate “Beginning” and “Ending”, are attached at the two ends of a sentence. We considered word 2-grams and word 3-grams in our system.

3.2 Character n-gram

We assume sub-word features like prefix and suffix are useful for detecting the learners' native languages. To simplify the process rather than employing a complex morphological analyzer, we consider character n-grams as another important feature set. The n-grams are extracted from each sentence by regarding the whole sentence as a large word / string and replacing the delimited symbol (i.e. white space) with a special uppercase character 'S'. As what we did in getting word n-grams, we attached two special character "B" and "E" at the two ends of a sentence. Character 2-grams, 3-grams, 4-grams, and 5-grams are used in our system.

3.3 Style

We would like to explore whether the traditional style features are helpful for this task as those features are widely used in authorship attribution. We include the following style features:

- __PARA__: a paragraph in an essay;
- __SENT__: a sentence in an essay;
- PARASENTLEN=NN: a paragraph of NN sentences long;
- SENTWDLEN=NN: a sentence of 4*NN words long;
- WDCL=NN: a word of NN characters long;

3.4 Other

As the TOEFL11 dataset includes two metadata for each essay, English language proficiency level (high, medium, or low) and Prompt ID, we include them as additional features in our system.

4 Experiments and Results

4.1 Dataset

The dataset of the NLI2013 shared task contains 12,100 English essays from the Test of English as a Foreign Language (TOEFL). Educational Testing Service (ETS) published the dataset through the LDC with the motivation to create a larger and more reliable data set for researchers to conduct Native Language Identification experiments on. This dataset, henceforth TOEFL11, comprises 11 native languages (L1s) with 1,000 essays per language. The 11 covered native languages are: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. In addition, each essay in the TOEFL11 is marked with an English language proficiency level (high, medium, or low) based on the judgments of human assessment specialists. The essays are usually 300 to 400 words long. 9,900 essays of this set are chosen as the training data, 1,100 are for development and the rest 1,100 as test data.

Runs	**HAUTCS-1**	**HAUTCS-2**	**HAUTCS-3**	**HAUTCS-4**
Accuracy	**0.773**	**0.758**	**0.76**	**0.756**
ARA	0.731[1]	0.703	0.703	0.71
CHI	0.82	0.794	0.794	0.782
FRE	0.806	0.788	0.786	0.783
GER	**0.897**	**0.899**	**0.899**	**0.867**
HIN	*0.686*	0.688	0.694	0.707
ITA	0.83	0.84	0.844	0.844
JPN	0.832	0.792	0.798	0.81
KOR	0.763	0.764	0.768	0.727
SPA	0.703	*0.651*	*0.651*	*0.65*
TEL	0.702	0.702	0.702	0.751
TUR	0.736	0.715	0.716	0.698

Table 1. Official results of our system.

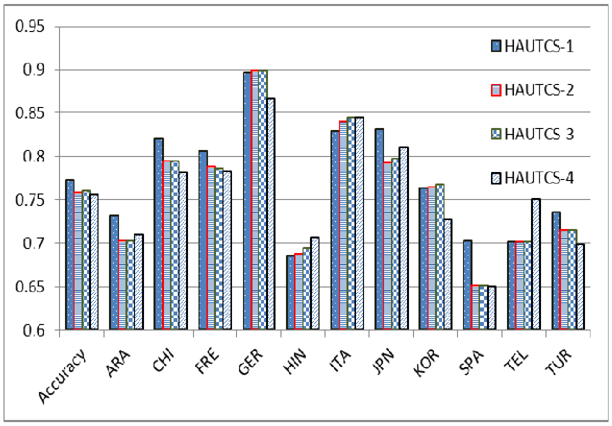

Figure 2. Performance of our official runs.

4.2 Official Results

Accuracy, which measures the percentage of how many essays are correctly detected, is used as the main evaluation metric in the NLI2013 shared task.

Table 1 gives the official results of our system on the evaluation data. We submitted four runs with different feature sets:

HAUTCS-1: all the features, which include words, word 2-grams, word 3-grams, **character 2-grams, character 3-grams, character 4-grams,**

[1] This number, as well as others in the cells from this row to the bottom, is value of F-1 measure for each language.

character 5-grams, style, and other metadata features;

HAUTCS-2: uses words, word 2-grams, word 3-grams, **style**, and other metadata features;

HAUTCS-3: uses words, **word 2-grams, word 3-grams**, and other metadata features;

HAUTCS-4: uses words or tokens and other metadata features.

For the runs HAUTCS-2, HAUTCS-3, and HAUTCS-4, we combined the development and training data for learning the identification model, where for the HAUTCS-1, it's a pity that we forgot to include the development data for training the model.

Our best run (HAUTCS-1) achieved the overall accuracy (0.773). The system performs best on the German category, but poorest on the Hindi category, as can be easily seen on figure 2.

Analyzing the four runs' performance showing on figure 2, we observe: word features are quite effective for Telugu and Hindi categories, but not powerful enough for others; word n-grams are helpful for languages Chinese, French, German, Korean, and Turkish, but useless for others; Style features only boost a little for French; Character n-grams work for Arabic, Chinese, French, Japanese, Spanish, and Turkish; Spanish category prefers character n-grams, where Telugu category likes word features. As different features have different effects on different languages, a better NLI system is expected to use different features for different languages.

After the evaluation, we experimented with the same setting as the HAUTCS-1 run, but included both training and development data for learning the NLI model. We got accuracy 0.781 on the new released test data, which has the same format with paragraph split as the training and development data.

As we include style features like how many paragraphs in an essay, the old test data, which removed the paragraph delimiters (i.e. single blank lines), may be not good for our trained model. Therefore, we did experiments with the new test data. Unfortunately, the accuracy 0.772 is a little poorer than that we obtained with the old test data. It seems that the simple style features are not effective in this task. As shown in table 1, HAUTCS-2 performs poorer than HAUTCS-3, which helps us derive the same conclusion.

4.3 Additional Experiments

We did 10-fold cross validation on the training and development data with the same setting as the HAUTCS-1 run. The data splitting is given by the organizers. Accuracies of the 10 runs are show in table 2. The overall accuracy 0.799 is better than that on the test data.

Fold	**1**	**2**	**3**	**4**	**5**
Accuracy	0.802	0.795	0.81	0.791	0.79
Fold	**6**	**7**	**8**	**9**	**10**
Accuracy	0.805	0.789	0.803	0.798	0.805

Table 2. Results of 10-fold cross validation on the training and development data.

To check how metadata features work, we did another run HAUTCS-5, which uses only words as features. This run got the same overall accuracy 0.756 on the old test data as HAUTCS-4 did, which demonstrates that those metadata features may not provide much useful information for native language identification.

5 Conclusion and Future Work

In this paper, we report our system for the NLI2013 shared task, which automatically detecting the native language of a foreign English learner from her/his writing sample. The system was built on a machine learning framework with traditional features including words, word n-grams, character n-grams, and writing styles. Character n-grams are simple but quite effective.

We plan to explore syntactic features in the future, and other machine learning algorithms, e.g. ECOC (Li and Vogel, 2010), also deserve further experiments. As we discussed in section 4, we are also interested in designing a framework to use different features for different categories.

Acknowledgments

This work was supported by the Henan Provincial Research Program on Fundamental and Cutting-Edge Technologies (No. 112300410007), and the High-level Talent Foundation of Henan University of Technology (No. 2012BS027). Experiments were performed on the Amazon Elastic Compute Cloud.

References

Ahn, C. S. 2011. Automatically Detecting Authors' Native Language. Master's thesis, Naval Postgraduate School, Monterey, CA.

Bergsma, S., Post, M., and Yarowsky, D. 2012. Stylometric analysis of scientific articles. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 327–337, Montréal, Canada. Association for Computational Linguistics.

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

Brooke, J. and Hirst, G. 2012a. Measuring interlanguage: Native language identification with l1-influence metrics. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), pages 779–784, Istanbul, Turkey

Brooke, J. and Hirst, G. 2012b. Robust, Lexicalized Native Language Identification. In Proceedings of COLING 2012, pages 391-408, Mumbai, India.

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

Chang, C.-C. and Lin C.-J. 2011. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:3:27:1-27.

Crossley, S. A. and McNamara, D. 2012. Detecting the First Language of Second Language Writers Using Automated Indices of Cohesion, Lexical Sophistication, Syntactic Complexity and Conceptual Knowledge. In Jarvis, S. and Crossley, S. A., editors, Approaching Language Transfer through Text Classification, pages 106-126. Multilingual Matters.

Jarvis, S., Castañeda-Jiménez, G., and Nielsen, R. 2012. Detecting L2 Writers' L1s on the Basis of Their Lexical Styles. In Jarvis, S. and Crossley, S. A., editors, Approaching Language Transfer through Text Classification, pages 34-70. Multilingual Matters.

Jarvis, S. and Paquot, M. 2012. Exploring the Role of n-Grams in L1 Identification. In Jarvis, S. and Crossley, S. A., editors, Approaching Language Transfer through Text Classification, pages 71-105. Multilingual Matters.

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. Determining an author’s native language by mining a text for errors. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 624–628, Chicago, IL. ACM.

Koppel, M., Schler, J., and Argamon, S. 2008. Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 60(1):9–26.

Li, B., and Vogel, C. 2010. Improving Multiclass Text Classification with Error-Correcting Output Coding and Sub-class Partitions. In Proceedings of the 23rd Canadian Conference on Artificial Intelligence, pages 4-15, Ottawa, Canada.

Swanson, B. and Charniak, E. 2012. Native language detection with tree substitution grammars. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 193–197, Jeju Island, Korea.

Tetreault, J., Blanchard, D., and Cahill, A. 2013. A report on the first native language identification shared task. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. Atlanta, GA, USA.

Tofighi, P.; Köse, C.; and Rouka, L. 2012. Author’s native language identification from web-based texts. International Journal of Computer and Communication Engineering. 1(1):47-50

Torney, R.; Vamplew, P.; and Yearwood, J. 2012. Using psycholinguistic features for profiling first language of authors. Journal of the American Society for Information Science and Technology. 63(6):1256-1269.

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, pages 1600–1610, Edinburgh, Scotland, UK.

Wong, S.-M. J., Dras, M., and Johnson, M. 2012. Exploring Adaptor Grammars for Native Language Identification. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 699–709, Jeju Island, Korea.