Author's Native Language Identification from Web-Based Texts

Parham Tofighı, Cemal Köse, and Leila Rouka

***Abstract*—With the rapid growth of Internet technologies and applications, Text is still the most common Internet medium. Examples of this include social networking applications and web applications are also mostly text based. We developed a framework to determine an anonymous author's native language for short length, multi-genre such as the ones found in many Internet applications. In this framework, four types of feature sets (lexical, syntactic, structural, and content-specific features) are extracted and three machine learning algorithms (C4.5 decision tree, support vector machine and Naïve Bayes) are designed for author's native language identification based on the proposed features. To experiment this framework, we used English, Persian, Turkish and German online news texts. The experimental results showed that the proposed approach was able to identify author's native language in web-based texts with satisfactory accuracy of 70% to 80%. And Support vector machines outperformed the other two classification techniques in our experiments.**

***Index Terms*—Native language identification, web-based texts, stylometry, classification techniques**

I. Introduction

The rapid development and multiplication of Internet technologies and applications have created a new way to share information across time and space. Online social networking (such as Twitter, Facebook), e-commerce application (such as eBay, amazon), newsgroups, etc. are gaining more prominence. The Internet has usually utilized shorter forms of communication more easily than traditionally longer forms such as handwritten letters and essays. Authorship profiling, and may be concerned with determining an author's gender, Native Language, age or some other attributes. We are particularly interested in this paper is the Native Language of an author, where this is not the language that the text is written in. In this paper, we propose a framework for author's native language identification on Web-based texts. In this framework, four types of features that are identified in authorship-analysis research are extracted, and three major inductive learning techniques are used to build feature-based classification models to perform automated authorship profiling. Our framework address author Native Language identification from short Internet text, due to the following questions:

Manuscript received February 25, 2012; revised April 28,2012

The authors are with the Department of Computer Engineering, Faculty of Engineering, Karadeniz Technical University,61080 Trabzon, TURKEY (e-mail: parham.tofighi@gmail.com)

Can the author's native language identification be applied to Web-based texts?

Which types of writing-style features are effective for determining the author's Native Language?

Which classification techniques are effective for detecting the author's Native Language in online texts?

Most previous contributions on authorship attribution argued on determining the particular author from a set of candidate authors is possibly by looking at the documents that each author has written and matching a new document of unknown origin to a profile built of each author [1], [3], and [4]. New research direction grew out of the author profiling [5], and Author gender identification [2], [6]. Related work on Author's Native language identification only applied Writers' spelling and grammatical mistakes in traditional text (e.g. essays) that often influenced by patterns in their native language [7].

II. Author's Native Language Identification

In essence, Author's native language identification problem is a classification problem that can be developed as follows: Provided a set of texts in English from authors with different native languages, and assign a new anonymous text to detect an author's language class, namely. To implement this hypothesis a set of features that remains relatively constant for a large number of texts written by the authors of the same native language. Once the feature set has been chosen, a given text can be represented by an N-dimensional vector, where N is the total number of features. Given a set of pre-classified texts, we can apply many techniques to determine the category of a new vector created based on a new text.

Characteristics of Web-based texts .Compared with conventional objects of native language identification such as published articles [7], one challenge of native language identification of web-based texts and messages is the limited length of online texts. The short length of online texts may cause some identifying features in normal texts to be ineffective. Analysis of a corpus of tens of thousands of English webpages indicates significant differences in writing style and content between native English writers and non-native writers [9]. Such differences can be exploited to determine an unknown author's native language on the basis of a webpage's corpus. Cyber users distribute messages mainly in English over cyberspace. Non-native English authors write in a systematic manner, and corpus of writings often depends on the first language of the writer [8], [9].

III. A Framework for Author's Native Language Identification

Our proposed framework for author's native language identification process is divided into four steps.

1) **Text Collection.** Collecting a suitable corpus of web-based texts to profile the writing styles.
2) **Feature Extraction.** A feature set includes **writing**-style characteristics to discriminating different native language authors. Feature Extraction is extracting feature set from texts automatically.
3) **Classification model.** Generating a classification model by dividing collection set into two subsets.
4) **Native language identification.** After designing classification model, it can be used to **predict** the native language of unknown texts

A. Text Collection

Most previous studies on native language identification traditionally used International Corpus of Learner English [10], which was assembled for precise purpose of studying the English writing of non-native English from variety of countries. For the purpose of this paper, we inspected all available Web-based texts. Among different types of Web-based texts, personal email and chat messages often involve privacy issues and are difficult to collect. Furthermore, collecting texts with known authors' native language is troublesome. Therefore, publicly available news agencies texts in English written by native English, Persian, Turkish and German authors are selected and collected as the Dataset in this study. We collected 150 texts for each language.

B. Feature Extraction

Authorship attribution has its root in stylometry; the use of an extended set of features could improve the scalability of writing-style analysis by enabling greater discriminatory ability across larger sets of classes. The feature set may significantly influence the performance of author's native language identification. Our feature set includes four types of features: lexical, syntactic, structural, and content-specific features as shown in Table 1. Since automatic spell-checker applied in most of Web-based application; idiosyncratic features include misspelling, and other usage anomalies ignored in this work.

Lexical Features can be divided into character based and word-based features. In our research, we included character-based lexical features used in [8], vocabulary richness features in [11], and word-length frequency features used in [13]. In total, we adopted 64 lexical features to discriminating authors with different native languages in our texts shown in Table 1.

Syntactic Features, involving author's writing style at the sentence level. The discriminating power of syntactic features is derived from people's different habits of organizing sentences. Syntactic features include common punctuation (such as comma, colon, etc.) and Function words. Function words are important distinguishing features for online texts. Function words (or grammatical words) are words that have little lexical meanings or have ambiguous meanings, but instead serve to reveal grammatical relationships with other words within a sentence, or specify the mood of the author. Function words are useful for authorship attribution [13]. It is logical that such words might also be useful for native language identification since particular function words are likely to be used more or less frequently by native authors and non-native authors, depending on the presence or absence of those words in the given language. A good example is the word the, which is typically used less frequently by native speakers of languages. We used 308 syntactic features in author's native language identification process.

Structural Features People have various habits when organizing texts, Structural features represent the way an author organizes the layout of a piece of writing introduced several structural features. These features, such as paragraph length and use of greetings, can be strong authorial evidence of author with different native language's writing styles. This is more prominent in online documents, which have less content information but more flexible structures or richer stylistic information. We used 13 structure-related features as listed in Table I.

Content-specific Features are comprised of important keywords and phrases on certain topics [17] such a word n-grams [11].Using n-grams to develop author's profile has proven to be a successful method of translating a corpus of texts into a set of models for authors [12]. In the case of native language attribution, the models are generated as n-gram distributions; the best matching author is decided by finding the most frequent Bi-grams and Tri-grams in a document and the frequency with which they occur. In order to extract the most important features we selected phrases with Tem frequency (TF) bigger than ten.

TABLE I: Proposed Feature sets

Lexical features
Character-based features
1. Total number of characters (C)
2. Total number of letters (a-z)/C
3. Total number of upper characters/C
4. Total number of digital characters/C
5. Total number of white-space characters/C
6. Total number of tab space characters/C
7 –29. Number of special characters (%, &, etc.) /C (23 features)
Word-based features
30. Total numbers of words (N)
31. Average lengths per word (in characters)
32. Vocabulary richness (total different words /N)
33. Words longer than 6 characters /N
34. Total number of short words (1-3 characters)/N
35. Hapax legomena/N
36. Hapax dislegomena/N
37. Yule's K measure
38. Simpson's D measure
39. Sichel's S measure
40. Honore's R measure
41. Brunet's W measure
42. Entropy measure
43. The number of net abbreviation /N
44– 64. Word length frequency distribution/N (20 features)
Syntactic features
65. Number of single quotes (') /C
66. Number of commas (,) /C
67. Number of periods (.) /C
68. Number of colons (:) /C
69. Number of semi-colons (;) /C
70. Number of question marks (?) /C
71. Number of exclamation marks (!)/C
72. Number of ellipsis (. . .) /C
Function words

73–373. Frequencies of function words (300 features)
Structural Features
374. Total number of lines
375. Total number of sentences (S)
376. Total number of paragraphs
377. Average number of sentences per paragraph
378. Average number of words per paragraph
379. Average number of characters per paragraph
380. Average number of words per sentence
381. Number of sentences beginning with upper case /S
382. Number of sentences beginning with lower case /S
383. Number of blank lines/total number of lines
384. Average length of non-blank line
385. Number of greeting words
386. Number of farewell words

Feature Extraction. The extraction phase includes extraction of all features formulated in Feature sets section. Authors' writing-style features need to be extracted from the unstructured text and then feature extractor produced a 386 dimension vector to represent the value of static features then we added dynamic features (Content-specific Features).

C. Classification Model

We used three classification techniques which are widely used and powerful classifiers: C4.5 decision tree [13], [14], Naïve Bayes [6] and Support vector machine (SVM) [15], [16]. As in a standard classifier learning process, the online message collection is divided into two subsets. One subset, called the training set, is used to train the classification model. The classification techniques applied in this process may lead to models with various predictive powers. The other subset is testing set, which is used to validate the prediction power of the author's native language identification model. If the performance of the classifier is confirmed by the testing set, it can be used to identify the native language of lately texts supplied. A repetitive training and testing process may be needed to obtain a good author's native language prediction model.

D. Native Language Identification

After the author's native language identification model is developed; it can be used to predict the unknown web-based texts' native language. The result of author's native language identification can improve the researcher on the different set of Web-based texts and various non-native English authors.

IV. Experimental Results

We conducted several experiments to examine the performance of different classification techniques, the impact of different types of feature, and the significance of the proposed feature set. To assess the prediction, we utilized the accuracy measure, which has been commonly adopted in data mining field. Accuracy indicates the overall prediction of a particular classifier, which is defined as in Equation 1:

$$Classification\ Accuracy = \frac{Number\ of\ Correctly\ Classified\ Texts}{Total\ number\ of\ Texts}$$

Equation (1)

A. Comparison of Classification Techniques

To compare the performances of classification techniques (Naïve Bayes, C4.5, and SVM) on the accuracy of author's native language identification we applied classifiers separately, using different size of texts per class and all four types of feature. The results shown in Fig. 1 indicate that SVM outperforms the other two methods. Figure 1 also shows that the performance of techniques does not change significantly with the size of dataset. The best classification result produced by the SVM classifier was the accuracy of 86.44 by 150 texts in dataset.

B. Impact of Features Types

To investigate the significance of proposed feature types on the accuracy of author's native language identification for each classification techniques, we assessed contribution of feature types in native language identification as shown in Fig. 2. Lexical and Syntactic features showed the good discriminating capability in native language identification. Although we used several Structural features in our feature set, it has a significance contribution in classification results. It reveals that authors with same native language have a consistent writing patterns were reflected in the structural features.

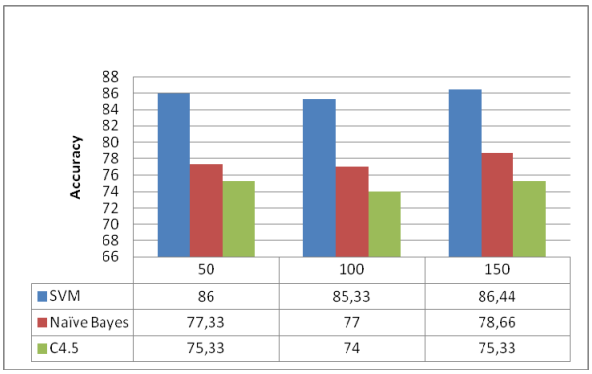

Fig. 1. Accuracy comparison of different classifers

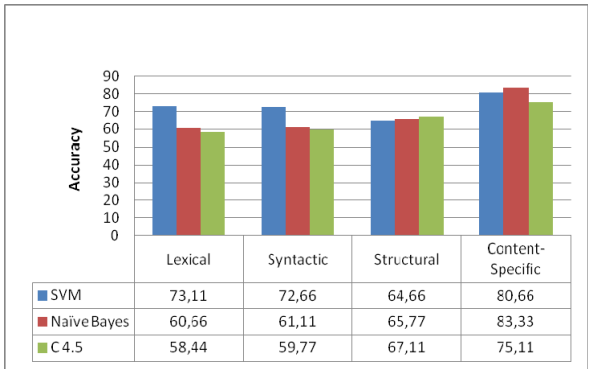

Fig. 2. Accuracy for different feature sets

And different numbers of texts in data set by using 100 texts in data set

V. Conclusion

The experimental results showed that the proposed approach is able to identify the author's native language from Web-based texts. Lexical features and content-specific features showed particular discriminating capabilities for native language identification from online messages. SVM outperformed Naïve Bayes and C4.5 significantly for the author's native language identification role. We believe that the proposed framework has the capability to aid in tracing authors' native language identity in cyberspace. In the

future we will investigate about how to identify the optimal set of features for author's native language identification of Web-based texts. Other future research directions are, handling more different candidate native languages, and using much shorter texts (i.e. Facebook messages and Twitter tweets) for discriminating author's native language.

REFERENCES

[1] A. Abbasi and H. Chen, "Visualizing authorship for identification," *IEEE International Conference on Intelligence and Security Informatics*, pp. 60–71, 2006.

[2] C. Köse, Ö. Özyurtve, and C. İkibaş, *a Comparison of Textual Data Mining Methods for Sex Identification in Chat Conversations, Springer: Lecture Notes in Computer Science*, pp. 638-643, 2008.

[3] J. Li, R. Zheng, and H. Chen, from fingerprint to writeprint. Commun, ACM, vol. 49, no. 4, pp. 76–82, 2006.

[4] A. Orebaugh and J. Allnutt, Classification of Instant Messaging Communications for Forensics Analysis, *The International Journal of Forensic Computer Science*, pp. 22–28, 2009.

[5] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text. Common," ACM, vol. 52, no. 2, pp. 119–123, 2009.

[6] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," in the *IEEE Symposium on Computational Intelligence and Data Mining Conference*, 2009.

[7] M. Koppel, J. Schler, and K. Zigdon, Determining an Author's Native Language by Mining a Text for Errors (short paper), in *Proc. of KDD,* Chicago IL, 2005.

[8] J. Lee and S. Seneff. "An analysis of grammatical errors in non-native speech in English," In *Proc. of the 2008 Spoken Language Technology Workshop*. 2008.

[9] R. J. Tetreault and M. Chodorow, *Examining the Use of Region Web Counts for ESL Error Detection*.

[10] S. Granger, E. Dagneaux, and F. Meunier, "the International Corpus of Learner English. Handbook and CD-ROM," *Louvain-la-Neuve. Presses Universitaires de Louvain.*2002.

[11] F. J. Tweedie and R. H. Baayen, "how variable may a constant be? Measures of lexical richness in perspective." *Computers and the Humanities*, vol. 32, pp-323–352. 1998.

[12] T. C. Mendenhall, *The characteristic curves of composition. Science*, vol. 11, no. 11, pp. 237–249. 1887.

[13] R. H. Baayen, V. H. Halteren, and F. J. Tweedie, "outside the cave of shadows: Using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing*, vol. 2, pp. 110–120. 1996.

[14] T. G. Dietterich, H. Hild, and G. Bakiri, "A comparative study of ID3 and backpropagation for English text-to-speech mapping," In *Proc. of the Seventh International Conference on Machine Learning*, pp. 24–31, 1990.

[15] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc of the 5th Annual ACM Workshop on Computational Learning Theory*, ACM Press, pp. 144–152. 1992.

[16] S. S.Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649. 2001.

[17] De Vel, O. Mining, E-mail authorship, Paper presented at the Workshop on Text Mining, *ACM International Conference on Knowledge Discovery and Data Mining*, 2000.