Selection of Discriminative Features for Translation Texts

Kuo-Ming TANG[1], *Chien-Kang HUANG*[1], *Chia-Ming LEE*[1]
(1) Department of Engineering Science and Ocean Engineering, National Taiwan University, Taipei, Taiwan (R.O.C.)
d965251013@ntu.edu.tw, ckhuang@ntu.edu.tw, trueming@gmail.com

ABSTRACT

Beginning in the first century AD, Buddhist texts underwent a series of translations during a period of nearly 1300 years. The identification of the translator, textual apocrypha, and translation style in Buddhist texts are always important issues. This study proposes an approach to find the most discriminative features that characterize the different Buddhist translation texts or other translation texts. We studied five different kinds of features that can be extracted from translation texts and exploited the F-score and SVM classifier to find the most discriminative features. Not only did we use the translated Buddhist texts, *Kalama Sutta*, for our experiment, but we also chose *The Canterbury Tales* to perform the same experiment and compare the results. According to our experiment results, the newly considered fifth-type features are very effective to identify translators. The selected features will be very useful for further studies of translator characteristics.

KEYWORDS : Feature selection, translator identification, F-score, SVM classifier

Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology, pages 49–64,
COLING 2012, Mumbai, December 2012.

1 Introduction

Translation is an activity to transform linguistic information to another language. Translation as a product that is a written text in a target-language (TL), which represents the result of a translation process, has been described and analysed by a comparison with the respective source-language (SL) text. The relation between the SL text and the TL text had been the object of the numerous and highly abstract models of equivalence (Koller 1978; 21983: 95; Ladmiral 1981: 393). In most cases, these models were prescriptive in nature and of very limited use for the practical translator. Problems in translating are caused at least as much by discrepancies in conceptual and textual grids as by discrepancies in languages.[1] Humankind has been engaged in transforming language for thousands of years, it affects the development of culture and language. "No language can exist unless it is steeped in the context of culture; and no culture can exist which does not have, at its center, the structure of natural language."[2] Translation activities can promote exchanges between different cultures and languages, and it is also very important in the spread and development of religion. For example, efforts to translate The Bible have occurred in Europe and around the world since it was first compiled during the fourth century AD. Translations of The Bible helped many countries to lay the foundation of language. In China, Buddhist texts also experienced a long history of translation during nearly 1300 years from the Eastern Han Dynasty to Song Dynasty (from 25 AD to 1297 AD). Different from the translation of texts from two other major global religions, Christianity and Islam, the translation and interpretation of Buddhist texts has been done with a very open attitude. Therefore, the identification of the translator, textual apocrypha, and translation style in Buddhist texts are particularly important.

This study tries to find the discriminative features in translations of Buddhist texts and other translated texts. Using these features we can set up a training model to identify the translator. Translator identification is a process of examining the characteristics of translation texts to distinguish who is the translator. Similar processes have been used in authorship identification, writing forensics, and similarity detection efforts to statistically analyze literary style. Most of the previous studies addressed the literary-style recognition and authorship analysis problems, which actually initiated this research domain of translation identification. The following sections present related works, the method and procedure, and the experimental evaluation.

2 Related Work

In early studies, researchers analyzed word usage of different authors to identify authors; however, the effectiveness of this approach is limited since word usage is highly dependent on the topic of the article. To achieve generic authorship identification in various applications, it need content free features. In early work, features such as sentence length and vocabulary richness (Yule, 1939 and 1944) were proposed. Later, Burrows (1987) used the high frequency words of occurrence of sets (typically 30 or 50) on *The Federalist Papers*. Holmes (1995) analyzed the use of shorter words. Such word-based and character-based features required intensive efforts in selecting the most appropriate set of words that best distinguished a given set

[1] Anton Popovič, 'The Concept of "Shift of Expression" in Translation Analysis' in James Holmes (ed.), *The Nature of Translation* (The Hague and Paris: Mouton, 1970).

[2] Robert Scholes, *Structuralism in Literature* (New Haven: Yale University Press, 1974), p. 10.

of authors (Holmes & Forsyth, 1995), and sometimes those features were not reliable discriminators when applied to a wide range of applications.

Few studies about the translator identification and textual apocrypha can be found in the past. However, there are many important results from the previous studies in authorship identification and literary-style recognition. The most convincing study in the field of authorship identification and literary-style recognition was conducted by Mosteller and Wallace in 1964. They studied the mystery of the authorship of *The Federalist Papers*, and their conclusion was generally accepted by historical scholars and became a milestone in this field.

For many major previous studies in literary-style recognition since the 1960s, lexical and syntactic features were most commonly used as the characteristics of literary-style. The most used approaches were statistical methods and machine-learning techniques. Few researchers have addressed multiple-language issues. These studies are summarized in Table 1.

2.1 Translation Identification

There are no scholarship paper can be found in Translation Identification or Translator Identification. But, the text style detection techniques to identify translator is very similar to Authorship Identification. This study refers to two major techniques for text style detection in Authorship Identification, statistical analysis and machine learning method. In early studies, most analytical tools used in authorship analysis were statistical univariate methods, such as Mosteller and Wallace (1964), Farringdon (1996), and Holmes (1998) . The advent of powerful computers instigated the extensive use of machine learning techniques in authorship analysis, such as Tweedie et al. (1998), Khmelev and Tweedie (2001), De Vel et al. (2001) and Argamon et al. (2003). In general, machine-learning methods achieved higher accuracy than did statistical methods. In Table 1, T1 denotes the use of the technique of statistical analysis and T2 denotes the use of the technique of machine learning.

2.2 Techniques in Identification

Due to the international nature of the Internet, it is important to study authorship identification in a multilingual context, but only Stamatatos et al. (1999 and 2001) conducted authorship identification with multiple languages, analyzing English and Greek newspaper articles. Peng, Schuurmans, Keselj, & Wang (2003) conducted experiments with Greek, English, and Chinese data to examine the performance of authorship attribution across multiple languages. In all three languages, the best accuracy achieved was 90%. However, the performance with Chinese writings was not as good as that with English writings, as shown in Table 1.

Our study is based on those previous studies and uses a machine-learning technique to recognize the translation-style of Buddhist texts. We also propose a framework for translator identification and literary-feature extraction. Machine learning methods have been used to establish an individual translator's translation-style vector-space-based model. According to the identification model, the identity of the translator of Buddhist texts can be clarified in the cases when the identity of the translator has previously been uncertain or unknown.

In order to find the more discriminative text-features, this study adopts an iteration of a feature extraction mechanism. The feature extractor can analyze and extract the text features in texts from the feature vector. After iterating the feature extraction method, the more discriminative text features are found.

Previous studies	Used type of features	Multilanguage	Authors	Training size (# of docs)
(Mosteller & Wallace, 1964)	T1	No	3	85
(Ledger & Merriam, 1994)	T1	No	2	N/A
(Merriam & Matthews, 1994)	T2	No	2	50
(Martindale & McKenzie, 1995)	T1+T2	No	3	85
(Mealand, 1995)	T1	No	1	N/A
(Holmes & Forsyth, 1995)	T1+T2	No	3	85
(Farringdon, 1996)	T1	No	N/A	N/A
(Baayen et al., 1996)	T1	No	2	2
(Tweedie et al., 1996)	T2	No	3	85
(Tweedie & Baayen, 1998)	T1	No	8	16
(Binongo & Smith, 1999)	T1	No	2	5
(Stamatatos et al., 1999)	T1	Yes	10	20
(De Vel et al., 2001)	T2	No	4	1259
(Stamatatos et al., 2001)	T1	Yes	10	300
(Khmelev & Tweedie, 2001)	T2	No	45	380
(Corney et al., 2002)	T2	No	N/A	N/A
(Baayen et al., 2002)	T1	No	8	72
(Peng et al., 2003)	T2	Yes	20	500
(Zheng et al., 2006)	T2	Yes	20	40

TABLE 1 – Previous studies in literary-style recognition and authorship identification. (T1 denotes the technique of statistical analysis and T2 denotes the technique of machine learning)

3 Method

In this study, we reduce the problem of translator identification to a classification problem. A learning classifier is able to learn based on a sample. Statistical methods are used to establish an individual translation-style vector-space-based model, such as Support Vector Machines (SVM), decision trees, etc. However, the focus of this study is not in the classification. The classification model just uses to extract the discriminative text features.

In order to find the more discriminative text features, this study adopts an iterative feature extract method using the F-score measure. The feature extractor can analyze and extract the text features in Buddhist texts, distinguishing them by the classification model. After the iteration of the feature extraction method, the more discriminative text features are found. The procedure for identifying translators by using feature extraction can be divided into three steps, as shown in Figure 1:

Step 1: Corpus Collection

In order to profile the translation styles of each translator and generate a translator identification model, in the first step we need to collect the translated Buddhist texts and a list of potential translators.

Step 2: Feature Extraction

Based on the classification model, the feature extractor analyzes and extracts the features in Buddhist texts. An iteration of feature extraction occurs using the F-score measure to find the more discriminative features. After feature extraction, each unstructured text is represented as a vector of the translation-style features.

Step 3: Model Validation

As done in a typical classifier learning process, the Buddhist text 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 might lead to models with different predictive powers. The other subset is called the testing set, which is used to cross-validate the prediction power of the translator-identification model generated by the classification model. If the performance of the classifier is verified by the testing set, it can be used to identify the new translations. An iterative training and testing process might be needed to develop a good translator-prediction model.

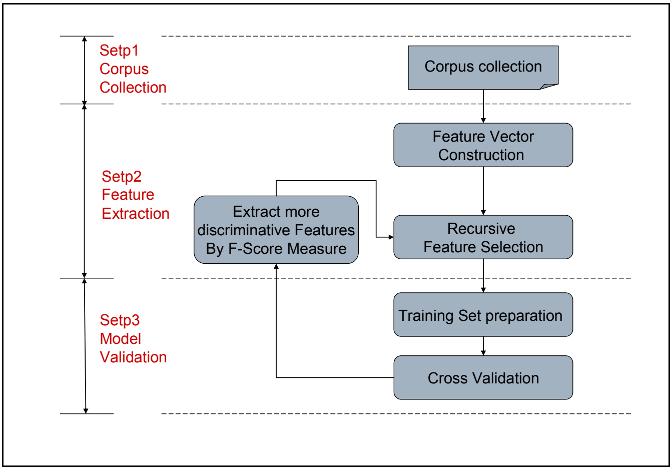

FIGURE 1 – Procedure of translator identification and feature extraction

3.1 Corpus Collection

In "Linguistic Aspects of Translation," Roman Jakobson (1960) distinguishes three types of translation:

1. Intralingual translation or rewording (an interpretation of verbal signs by means of other signs in the same language).
2. Interlingual translation or translation proper (an interpretation of verbal signs by means of some other language).
3. Intersemiotic translation or transmutation (an interpretation of verbal signs by means of signs of nonverbal sign systems).

In the message collection, two important texts must be collected: the original Buddhist texts and the texts of translators.

The main purpose of this study is to identify the translator of Buddhist texts. However, in order to do so, the identification model must be very versatile. Therefore, another literary work—a collection of English tales—was chosen as a testing corpus. For our studies, we used two kinds of translation texts as sample corpora: the *Kalama Sutta* as the Buddhist text and *The Canterbury Tales* as the collection of English tales. Each of these texts are well-known and versions of each have been translated by different translators. There is a difference between these two literary works. The Buddhist text, *Kalama Sutta*, was translated from different languages. However, the English tales, *The Canterbury Tales,* were written in the same language but at different time periods. The background of these two translation texts and their translators is described below.

3.1.1 Buddhist Text

Kalama Sutta is one of training corpora for Buddhist texts in this study. The original sutta is the Pali version, and it was translated into English. The sutta starts off by describing how the Buddha passes through the village of Kesaputta and is greeted by its inhabitants, the Kalamas of the title. They ask for his advice; they say that many wandering holy men and ascetics pass through the village, expounding their teachings and criticizing the teachings of others. So whose teachings should they follow? He delivers in response a sermon that serves as an entry point to the Buddhadhamma for those unconvinced by mere spectacular revelation.

Buddha proceeds to list the criteria by which any sensible person can decide which teachings to accept as true. He tells the Kalamas not to believe religious teachings just because they are claimed to be true or even through the application of various methods or techniques. Direct knowledge grounded in one's own experience can be called upon. He advises that the words of the wise should be heeded and taken into account. Not, in other words, passive acceptance but, rather, constant questioning and personal testing to identify those truths that you are able to demonstrate to yourself actually reduce your own stress or misery.

Two important translators who had translated the Kalama Sutta into English were Thānissaro Bhikkhu (born 1949) and Bodhi Bhikkhu (born 1944). This study used their translations of Buddhist texts to generate a translation-style identification model and find the more discriminative features of their translations.

3.1.2 Canterbury Tales

The Canterbury Tales is a collection of stories written in Middle English by Geoffrey Chaucer at the end of the 14th century. The tales were mostly written in verse, although some are in prose, and they are told as part of a story-telling contest by a group of pilgrims as they travelled together on a journey from Southwark to the shrine of Saint Thomas Becket at the Canterbury Cathedral. The prize for this contest was a free meal at the Tabard Inn at Southwark on their return.

Following a long list of works written earlier in his career, including Troilus and Criseyde, House of Fame, and Parliament of Fowls, the Canterbury Tales was Chaucer's *magnum opus*. He uses the tales and the descriptions of the characters to paint an ironic and critical portrait of contemporary English society and particularly of the Church. Structurally, the collection bears the influence of The Decameron, which Chaucer is said to have come across during his first diplomatic mission to Italy in 1372. However, Chaucer peoples his tales with “sondry folk” rather than Boccaccio's fleeing nobles.

A modernised version or translation was published by A. S. Kline in 2007 that retained Chaucer's rhyme scheme and remained close to the original, but eliminated archaisms that would require explanatory notes. Another version was translated and edited by Gerard NeCastro in 2007. Both of these versions are written in modern English, translated from Middle English.

3.2 Feature Extraction

Most previous studies addressed the authorship identification problem, which actually initiated this research domain. Table 3 summarizes major studies in authorship identification since the 1960s. Lexical and syntactic features were most commonly used. Statistical approaches were extensively used in the past, but more applications of machine learning techniques have been observed recently.

3.2.1 Feature type

Lexical features can be further divided into character-based and word-based features. In our research, we included character-based lexical features used in de Vel (2000), Forsyth and Holmes (1996), and Ledger and Merriam (1994), vocabulary-richness features in Tweedie and Baayen (1998), and word-length-frequency features used in Mendenhall (1887) and de Vel et al. (2000).

Syntactic features, including function words, punctuation, and parts of speech, can capture an author’s writing style at the sentence level. The discriminating power of syntactic features is derived from people’s different habits of organizing sentences.

Structural features represent the way an author organizes the layout of a piece of writing. De Vel (2000) introduced several structural features specifically for e-mail. Because e-mail contains many general structural features, we adopted those features applicable for online texts. In addition, we added features, such as paragraph indentation and signature-related features. In total, we adopted 14 structured features, including 10 features from de Vel (2000) and four newly proposed features.

Content-specific features are important discriminating features. The selection of such features is dependent on specific application domains.

Translation features include the simplification feature and explicit features, as shown in Table 2.

Features	Label	Content
Lexical features	F1	Average word/sentence length, Vocabulary richness
Syntactic Features	F2	Frequency of function words, Use of punctuation
Structural Features	F3	Paragraph length, Indentation
Content-specific Features	F4	Frequency of keywords
Translation Features	F5	Simplification and explicit features

TABLE 2 – Features of Authorship Identification

3.2.2 Iterative Selection

This paper used the F-score as a feature filter to find the more discriminative features. It is a simple technique that measures the discrimination of two sets of real numbers. Given training vectors x_k, $k = 1,\ldots, m$, if the number of positive and negative instances are n_+ and $n_–$, respectively, then the F-score of the ith feature is defined as follows (Y.-W. Chen, 2005):

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

- $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the average of the *i*th feature of the whole, positive, and negative data sets, respectively.
- $x_{k,i}^{(+)}$ is the *i*th feature of the *k*th positive instance
- $x_{k,i}^{(-)}$ is the *i*th feature of the *k*th negative instance

The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative.

There are five steps in this F-score measure:
1. Calculate the F-score of every feature.
2. Pick some possible thresholds to remove low and high F-scores.
3. For each threshold, do the following: Drop features with an F-score below this threshold. Randomly split the training data into Xtrain and Xvalid. Let Xtrain be the new training data. Use the SVM procedure to obtain a predictor; use the predictor to predict Xvalid. Repeat the steps above five times, and then calculate the average validation error.
4. Choose the threshold with the lowest average validation error.
5. Eliminate features with an F-score below the selected threshold.

After the execution of above steps, apply the SVM procedure again.

3.3 Classification Model

This study used a support vector machine (SVM) method as a classification technology. SVM is a set of related supervised learning methods that analyze data and recognize patterns, which can be used for classification and regression analysis. As in a typical classifier learning process, the translation of texts 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 might lead to models with different predictive powers. The other subset is called the testing set, which is used to validate the prediction power of the translator-identification model generated by the classification model. If the performance of the classifier is verified by the testing set, it can even be used to identify a new translator. An iterative training and testing process might be needed to develop a good translator-prediction model. This paper uses LIBSVM as an SVM tool. LIBSVM is a set of an integrated software. Components of LIBSVM have different functions: C-SVC and nu-SVC are used for support vector classification, epsilon-SVR and nu-SVR are used for regression, and one-class SVM is used for distribution estimation. It supports multi-class classification.

3.4 Training Set

A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e. the class labels) and several attributes (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) that predicts the target values of the test data given only the test data attributes.

3.5 Cross Validation.

This study uses LIBSVM to find two parameters for an RBF kernel: C and γ automatically. It is not known beforehand which C and γ are best for a given problem; consequently some kind of model selection (parameter search) must be done. The goal is to identify a good set of parameters (C; γ) so that the classifier can accurately predict unknown data (i.e. testing data). It is important to note that it might not be useful to achieve high levels of training accuracy (i.e. a classifier that accurately predicts training data whose class labels are indeed known). A common strategy is to separate the dataset into two parts, of which one is considered unknown. An improved version of this procedure is known as cross-validation. In v-fold cross-validation, we divide the training set into v subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining v-1 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data that are correctly classified.

4 Experiments

4.1 Experimental Design

To examine different features and techniques, we designed several translation identification tasks. First, four feature sets were created. In this case, we use F1, F2, F3, and F4 to denote lexical, syntactic, structural, and content-specific features, respectively. The first feature set contained lexical features (F1) only. Syntactic features were added to F1 to form the second feature set (F1+F2). Structural features were added to form the third feature set (F1+F2+F3). The fourth and fifth feature sets contained four types (F1+F2+F3+F4) and five types (F1+F2+F3+F4+F5) of features, respectively. We chose this incremental method in this order because it represents the evolutionary sequence of style features, and we intended to examine the effect of adding relatively new features to existing ones. Second, we adopted SVM classifiers as the classifiers. A 5-fold cross-validation was used to estimate the accuracy of the classification model.

For this study, we used the Buddhist text corpus, Kalama Sutta, and the English tales corpus, The Canterbury Tales. The basic information about these two corpora is shown in Table 3.

4.2 Experimental Results

Using the SVM classifier, we found that the maximum validation accuracy of Lexical Features (F1) was 68.42% and 100% in the *Kalama Sutta* and *The Canterbury Tales* texts, respectively. The maximum validation accuracy of Syntactic Features (F2) was 86.84% and 100%, respectively. The maximum validation accuracy of Structural Features (F3) was 68.42% and 55.26%, respectively. The maximum validation accuracy of Content-specific Features (F4) was 92.01% and 78.94%, respectively. The maximum validation accuracy of Translation Features (F5) was 89.47% and 100%, respectively. Details are shown in Tables 4 and 5 and in Figure 2.

Item	Kalama Sutta	The Canterbury Tales
Corpus size	38 samples	38 samples
File size	0.6M	1.4M
Number of Samples	38 paragraphs	38 paragraphs
Number of Words	37772	3201248
Size of Vocabulary	798	22875
Average bytes per sample	1069.9	906012.1
Average characters per sample	534.9	45306.1
Training/Testing	5-fold cross validation	5-fold cross validation
Dimensions of feature vector	69	69
Type of classifiers	SVM	SVM

TABLE 3 – Basic information of the translation corpora

Feature sets	Feature sizes	Extracted features sizes	Maximum validation accuracy
F1	14	2	68.42%
F2	81	80	86.84%
F3	6	2	68.42%
F4	231	231	92.10%
F5	28	13	89.47%
F1+F2	95	45	86.84%
F1+F2+F3	101	23	92.10%
F1+F2+F3+F4	332	81	97.36%
F1+F2+F3+F4+F5	360	81	97.36%

TABLE 4 – Maximum validation accuracy for different features of *Kalama Sutta*

Feature sets	Feature sizes	Extracted features sizes	Maximum validation accuracy
F1	14	2	100%
F2	81	80	100%
F3	6	2	55.26%
F4	231	231	78.94%
F5	28	13	100%
F1+F2	95	45	100%
F1+F2+F3	101	23	100%
F1+F2+F3+F4	332	64	100%
F1+F2+F3+F4+F5	360	71	100%

TABLE 5 – Maximum validation accuracy for different features of *The Canterbury Tales*

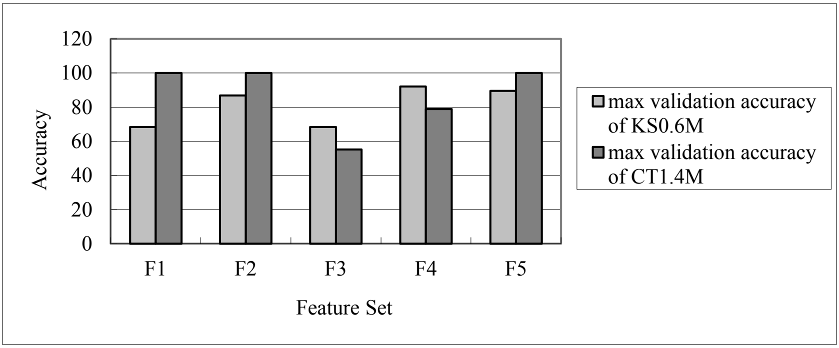

FIGURE 2 – Line chart of the maximum validation accuracy for different features in SVM

As seen in Tables 5 and 6, the best maximum validation accuracy for *Kalama Sutta* is 92.10% in Content-specific Features (F4). However, the maximum validation accuracy for *The Canterbury Tales* is 100% in Lexical Features (F1), Syntactic Features (F2) and Translation Features (F5). For *Kalama Sutta*, although the features F4 performance is relatively good, features F2 and F5 continue to be essential features. For *The Canterbury Tales*, features F4 and F3 are not good, relatively.

There are five combinations of feature sets: F1 (Lexical Features), F1+F2 (Lexical and Syntactic Features), F1+F2+F3 (Lexical, Syntactic, and Structural Features), F1+F2+F3+F4 (Lexical, Syntactic, Structural, and Content-specific Features), and all five features, F1+F2+F3+F4+F5 (Lexical, Syntactic, Structural, Content-specific, and Translation Features). The size of F1 is 14, maximum validation accuracy is 68.42% for *Kalama Sutta*; F1+F2 is 86.84%; and F1+F2+F3 is 92.10%. Both of the maximum validation accuracy values of Features F1+F2+F3+F4 and F1+F2+F3+F4+F5 are 97.36% for *Kalama Sutta*. However, the maximum validation accuracy of all feature combinations is 100% in CT1.4M.

For *Kalama Sutta*, the best maximum validation accuracy from Table 5 is 92.10% in Content-specific Features (F4). Also from Table 5, both of the maximum validation accuracy values of Features F1+F2+F3+F4 and F1+F2+F3+F4+F5 are 97.36%. It can be seen that the Content-specific Features (F4) dominate the results of maximum validation accuracy among all features of the *Kalama Sutta*. However, in *The Canterbury Tales*, the dominant features are Lexical Features (F1), Syntactic Features (F2), and Translation Features (F5), as shown in Table 6. The comparison of maximum validation accuracy values between *Kalama Sutta* and *The Canterbury Tales* is shown in Figure3.

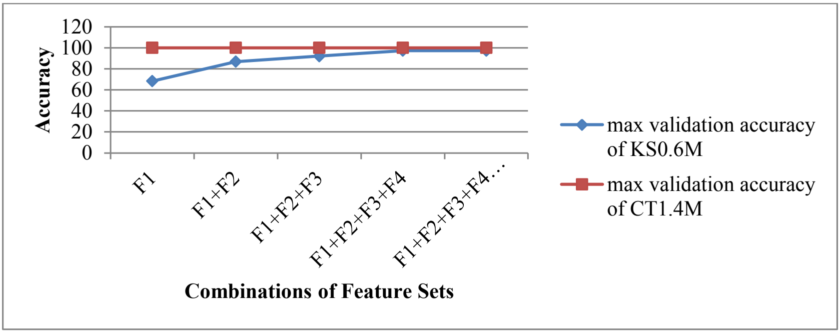

FIGURE 3 – Comparison of maximum validation accuracy values for different feature sets

This study used the F-score measure as a feature filter to find the more discriminative features in different combinations of feature sets. Details are shown in Figure 4.

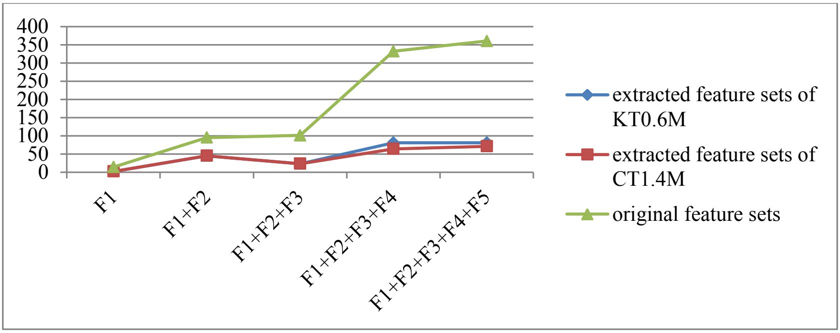

FIGURE 4 – Comparison of the number of extracted features for different feature sets

As shown in Tables 4 and 5 and in Figure 4, the number of F1 features is 14, the number of discriminative features is 2 extracted by the iteration filter in both *Kalama Sutta* and *The Canterbury Tales*. The number of F1+F2 features is 95, and the number of discriminative features extracted by the iteration filter in both corpora is 45. The number of F1+F2+F3 feature is 101, and the number of discriminative features extracted by the iteration filter in both corpora is only 23. The number of F1+F2+F3+F4 feature is 332, and the number of discriminative features extracted by the iteration filter in *Kalama Sutta* is 81. However, the number of discriminative features extracted by the iteration filter in *The Canterbury Tales* is only 64. The number of F1+F2+F3+F4+F5 feature is 360, and the number of discriminative features extracted by the iteration filter in *Kalama Sutta* is also 81 (the same as Feature F1+F2+F3+F4). And, the number of discriminative features extracted by the iteration filter in *The Canterbury Tales* is 71.

5 Conclusion

From the results of our experiment, the Content-specific Features (F4) dominate the results of maximum validation accuracy among all features in *Kalama Sutta*. However, in *The Canterbury Tales*, the dominant features are Lexical Features (F1), Syntactic Features (F2) and Translation Features (F5), as shown in Tables 5 and 6.

Additionally, as seen in Tables 5 and 6 and in Figure 4, the number of discriminative features extracted by the iteration filter in *Kalama Sutta* is 81. Also, the number of discriminative features extracted by the iteration filter in *The Canterbury Tales* is 71. It means that fewer features can effectively discriminate the features in translation texts. The number of discriminative features for *Kalama Sutta* and *The Canterbury Tales* are compared for each feature set in Table 6.

Corpus	Features sizes	F1	F2	F3	F4	F5
Kalama Sutta	81/360	0/14	12/80 (14.8%)	0/6	69/230 (85.2%)	0/28
The Canterbury Tales	71/360	2/14 (2.8%)	25/80 (35.2%)	0/6	41/230 (57.7%)	3/28 (4.3%)

TABLE 6 – Comparison of discriminative features.

The F4 feature set (Content-specific Features) has a great impact in Kalama Sutta. There are 69 discriminative features selected from all 230 F4 features (about 85.2%). However, it has less impact in *The Canterbury Tales*; only 41 discriminative features were selected from all 230 F4 features (which still accounts for 57.7%).

There are 81 more discriminative features extracted from feature sets F2 and F4. We can identify these features in *Kalama Sutta.* The content of the F4 feature set (Content-specific Features) in *Kalama Sutta* can be divided into Proper Noun and Adjective, and further analyzed. In the same way, we can also use 71 more discriminative features extracted from F1, F2, F4, and F5 to identify the translation text of The Canterbury Tales.

In future, using the discriminative features, we can develop an authorship-identification model to be used in the prediction of the authorship of unknown translation texts. The result of authorship identification will help the investigator focus his or her efforts on a small set of texts and authors. More formally, a support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), because in general the larger the margin, the lower the generalization error of the classifier.

Acknowledgments

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this research under Contract No. NSC 99-2410-H-002-203-MY2.

Reference

Anton Popovič. 1970. “The Concept of “Shift of Expression” in Translation Analysis’”. *Shift of Expression-in Translation Analysis", James S. Holmes et al*: 78–87.

Argamon, S., M. Koppel, J. Fine & A. R Shimoni. 2003. “Gender, genre, and writing style in formal written texts”. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-* 23 (3): 321–346.

Argamon, S., M. Šarić 及 S. S Stein. 2003. “Style mining of electronic messages for multiple authorship discrimination: first results”. in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 475–480.

Baayen, H., H. van Halteren, A. Neijt & F. Tweedie. 2002. “An experiment in authorship attribution”. in *6th JADT*.

Binongo, J. N.G, & M. W. A. Smith. 1999. “The application of principal component analysis to stylometry”. *Literary and Linguistic Computing* 14 (4): 445.

Burrows, J. F., & J. F. Burrows. 1987. *Computation into criticism: A study of Jane Austen’s novels and an experiment in method*. Clarendon Press Oxford. http://www.getcited.org/pub/102535837.

Carney, D. M, & R. H Nguyen. 2002. *Method and apparatus for adjusting an interval of polling a network printer based on changes in working status of the network printer*. Google Patents.

Chen, Y. W, & C. J Lin. 2006. “Combining SVMs with various feature selection strategies”. *Feature Extraction*: 315–324.

Farringdon, J. M, A. Q. Morton, M. G Farringdon & M. D. Baker. 1996. *Analysing for Authorship: A Guide to the Cusum Technique*. University of Wales Press, Cardiff.

Grzybek, P. 2006. “History and methodology of word length studies”. *Contributions to the Science of Text and Language*: 15–90.

Holmes, D. I, & R. S Forsyth. 1995. “The Federalist revisited: New directions in authorship attribution”. *Literary and Linguistic Computing* 10 (2): 111.

Holmes, D. I. 1998. “The evolution of stylometry in humanities scholarship”. *Literary and linguistic computing* 13 (3): 111–117.

Jakobson, R. 1960. “Closing statement: Linguistics and poetics”. *Style in language* 350: 377.

Khmelev, D. V, & F. J Tweedie. 2001. “Using Markov Chains for Identification of Writer”. *Literary and linguistic computing* 16 (3): 299.

Koppel, M., & J. Schler. 2003. “Exploiting stylistic idiosyncrasies for authorship attribution”. in *Proceedings of IJCAI*, 3:69–72.

Koppel, M., J. Schler, S. Argamon & E. Messeri. 2006. “Authorship attribution with thousands of candidate authors”. in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 659–660.

Ledger, G., & T. Merriam. 1994. “Shakespeare, Fletcher, and the two noble kinsmen”. *Literary and Linguistic Computing* 9 (3): 235.

Martindale, C., & D. McKenzie. 1995. “On the utility of content analysis in author attribution: The Federalist”. *Computers and the Humanities* 29 (4): 259–270.

McCarthy, P. M, G. A Lewis, D. F Dufty & D. S McNamara. 2006. “Analyzing writing styles with Coh-Metrix”. in *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)*, 764–769.

Mealand, D. L. 1995. “Correspondence analysis of Luke”. *Literary and linguistic computing* 10 (3): 171.

Merriam, T. V.N, & R. A.J Matthews. 1994. “Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe”. *Literary and Linguistic Computing* 9 (1): 1.

Mosteller, F., & D. Wallace. 1964. “Inference and disputed authorship: The Federalist」.

Oberlander, J., & S. Nowson. 2006. “Whose thumb is it anyway?: classifying author personality from weblog text”. in *Proceedings of the COLING/ACL on Main conference poster sessions*, 627–634.

Peng, F., D. Schuurmans, S. Wang & V. Keselj. 2003. “Language independent authorship attribution using character level language models”. in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, 267–274.

Pruscha, H. 1998. “Statistical models for vocabulary and text length with an application to the NT corpus”. *Literary and linguistic computing* 13 (4): 195.

Stamatatos, E., N. Fakotakis & G. Kokkinakis. 1999. “Automatic authorship attribution”. in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 158–164.

Stamatatos, E. 2000. “Text genre detection using common word frequencies”. in *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 808–814.

Stamatatos, E. 2001. “Computer-based authorship attribution without lexical measures”. *Computers and the Humanities* 35 (2): 193–214.

Tweedie, F. J, & R. H Baayen. 1998. “How variable may a constant be? Measures of lexical richness in perspective”. *Computers and the Humanities* 32 (5): 323–352.

De Vel, O., A. Anderson, M. Corney & G. Mohay. 2001. “Mining e-mail content for author identification forensics”. *ACM Sigmod Record* 30 (4): 55–64.

Yule, G. U. 1939. “On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship”. *Biometrika* 30 (3/4): 363–390.

Zheng, R., J. Li, H. Chen & Z. Huang. 2006. “A framework for authorship identification of online messages: Writing-style features and classification techniques”. *Journal of the American Society for Information Science and Technology* 57 (3): 378–393.