2015 Master Thesis

Improving Chinese Native Language Identification by Cleaning Noisy Data and Adopting BM25

Date：February 1, 2016
Adviser： Professor HAYATO YAMANA

Department of Computer Science and Communications Engineering
Graduate School of Fundamental Science and Engineering
Waseda University
Student ID： 5114FG01-7

LAN WANG

Abstract

Native language identification (NLI), as a branch of author profiling, is a process by which an author's native language can be identified from essays written in the second language of the author. Not only can NLI be used to identify phishing sites and spam e-mails, but it can be also actively applied to Second Language Acquisition (SLA). Since Chinese, as an ideographic language, is growing in popularity, the demand of Chinese teaching services is increasing progressively. However, most of the current NLI studies focus on essays written in English. Therefore, research on Chinese NLI becomes indispensable. In this work, a supervised model is built to accomplish NLI based on a large Chinese learner corpus. In the NLI field, this is the first work to (1) eliminate noisy data automatically before the training phase, (2) explore skip-gram as informative features, and (3) employ a BM25 term weighting technique to score each feature. In addition, by dividing the dataset into training, tuning and test subset, evaluation is more robust than that in other Chinese NLI works. After adopting a hierarchical structure of linear support vector machine classifiers, a state-of-the-art accuracy of 75.3% is achieved by our proposed model.

Contents

Figure Contents

Table Contents

1. Introduction

There is no doubt that a Chinese native speaker can judge whether a person is a native speaker or not easily on the basis of his/her speaking patterns. Further, sometimes, it is easier to guess the mother tongue of a person on the basis of his/her pronunciation than the speaking patterns. A large number of studies have indicated that the same principle can be employed in text [1]. Therefore, it is possible to conclude that traits of different native language speakers will appear on their essays written in their second language.

Native language identification (NLI) is a task in which a writer's native language (hereafter, L1) can be recognized by his/her essays written in his/her second language (hereafter, L2). Concisely, it can be considered a classification task where machine-learning methods assign labels (native languages) to different objects (essays).

The applications of NLI are classified into two categories: security-related applications and second language acquisition (SLA)-related applications. For security-related applications, not only can NLI be applied to identify false information distributors on SNS, but also phishing sites or spam e-mails that usually consist of strange sentences that might be written by non-native persons [2]. The other applications are related to SLA. In recent years, an increasing number of people have shown interest in learning Chinese[1]. Through reviewing their Chinese writing styles, such as frequent expression and error patterns corresponding to each native language, we can provide instructive advice to language teachers so that they can teach and guide students in a more sophisticated way.

Since Chinese is an ideographic language, in which strokes and their combination need to be remembered exactly, it is more difficult to learn than an alphabetic language, such as English [3]. Moreover, the demand of Chinese teaching services is increasing progressively. However, most of the current NLI studies focus on essays written in English. Therefore, research on Chinese NLI becomes indispensable.

However, there are three major issues in the modern NLI research field. First, noisy data, whose essay length is very short or very long, in the training dataset will result in low performance. Second, features exploited are restricted to either lexical or syntactic features, which are not informative enough. Third, the different importance of each feature is not fully studied. In order to solve the three problems mentioned above, we

[1] http://www.cctv-america.com/2015/03/03/chinese-as-a-second-language-growing-in-popularity

propose a new supervised model integrating (1) automatic noisy data elimination for the training dataset, (2) 1-skip-bi-gram, which contains both lexical (word level) and syntactic (grammatical level) information as a feature, (3) a BM25 term weighting method, and (4) a hierarchical linear SVM classifier. In the experiment, we employ them on a large Chinese corpus, in which Chinese essays are written by students from ten countries, and evaluating with 10-fold cross validation by dividing the corpus into training, tuning and test subset.

The rest of this paper is structured as follows: Chapter 2 describes representative related works on NLI; Chapter 3 shows the corpus used in this work; Chapter 4 presents our proposed supervised model for NLI; Chapter 5 discusses the result of our experiments; Chapter 6 gives our conclusion and forecast for future work.

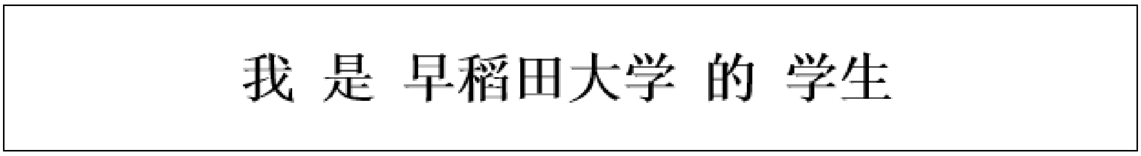
2. Related Works

In this chapter, we define the terminologies used in this work and introduce representative related works in NLI research field.

2.1 Terminology Definition

In this section, we define the terminologies employed in this work: n-gram and POS tags.

N-gram model, proposed by Claude Elwood Shannon [4], is widely used in natural language processing. We illustrate n-gram by an example of word-based Chinese sentence, as is shown in Figure 2.1.



我 是 早稻田大学 的 学生

Figure 2.1. Example of a Chinese word-based sentence

Character 1-gram, 2-gram, 3-gram and word 1-gram, 2-gram, 3-gram of the above Chinese sentence are listed in Table 2.1, respectively.

Table 2.1. Illustration of n-gram models

N-gram model	Example
Character 1-gram	我，是，早，稻，田，大，学，的，学，生
Character 2-gram	我是，是早，早稻，稻田，田大，大学，学的，的学，学生
Character 3-gram	我是早，是早稻，早稻田，稻田大，田大学，大学的，学的学，的学生
Word 1-gram	我，是，早稻田大学，的，学生
Word 2-gram	我是，是早稻田大学，早稻田大学的，的学生
Word 3-gram	我是早稻田大学，是早稻田大学的，早稻田大学的学生

Generally, character n-gram is based on a single Chinese character as a unit, while word n-gram lays a foundation on a single Chinese word. As for Part-of-Speech (hereafter, POS) tag n-gram, we first convert Chinese words into POS tags, and then construct n-grams. As an example, we transform the Chinese sentence in Figure 2.1 into POS tags, which is shown in Figure 2.2. Here, PN refers to Pronoun, VC represents verb, and so on. All the Chinese Penn Treebank POS tags are summarized in [5].

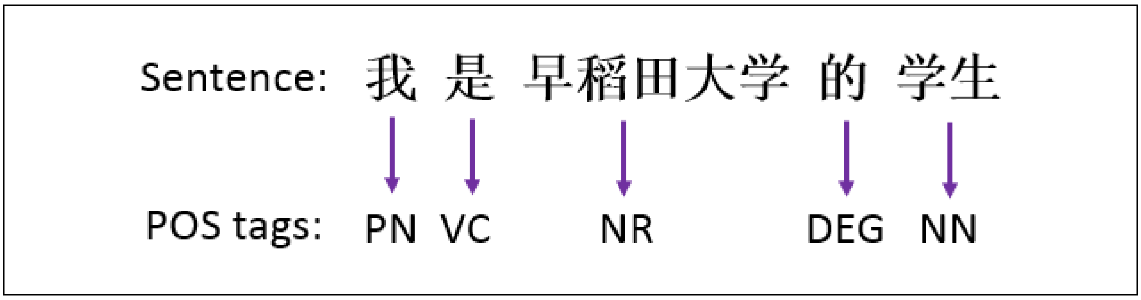


Figure 2.2. POS tags of sentence in Figure 2.1

2.2 Prototype of NLI System

Even though a number of representative related works will be introduced in Chapter 2, the prototype of those NLI systems are identical, which is described in Figure 2.3. As is shown, blue arrows and red arrows indicate training and test procedure, respectively. On one hand, in the training phase, training essays will be transformed to their vector space at first based on the selected features. After that, a supervised classifier will be trained. On the other hand, in the test phase, given vector spaces of test essays, we can employ the trained classifier to classify them. The supervised machine learning model proposed in this work is also based on Figure 2.3.

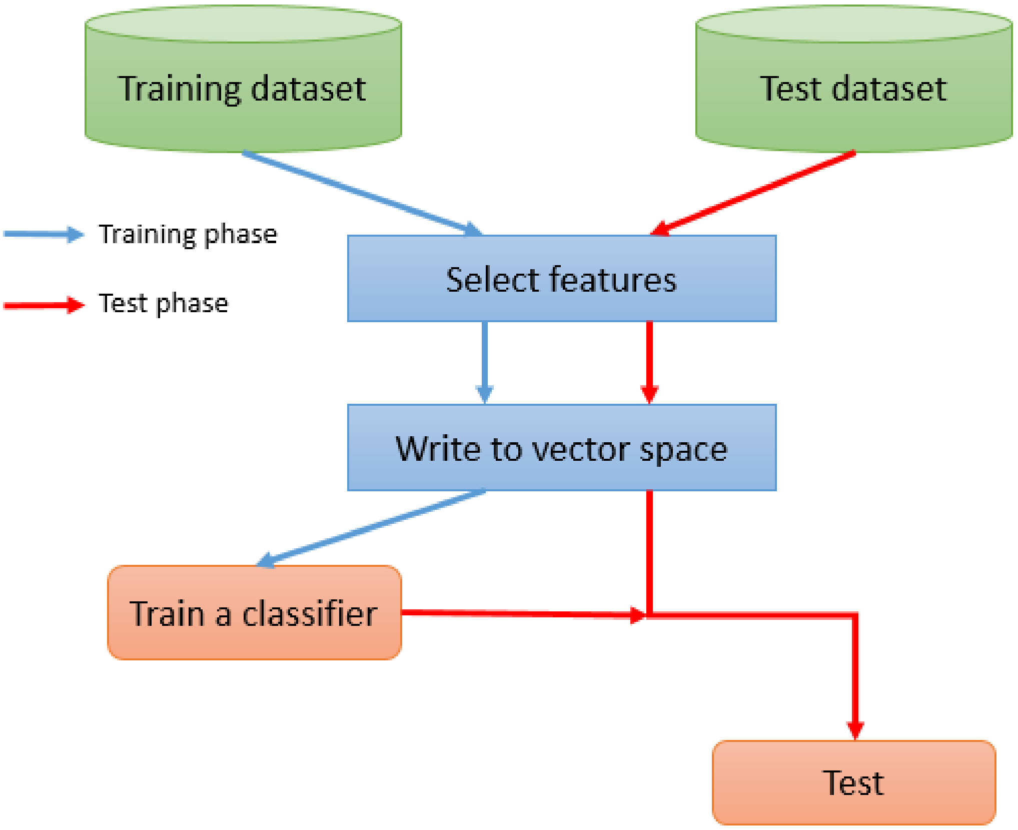


Figure 2.3. Prototype of NLI system

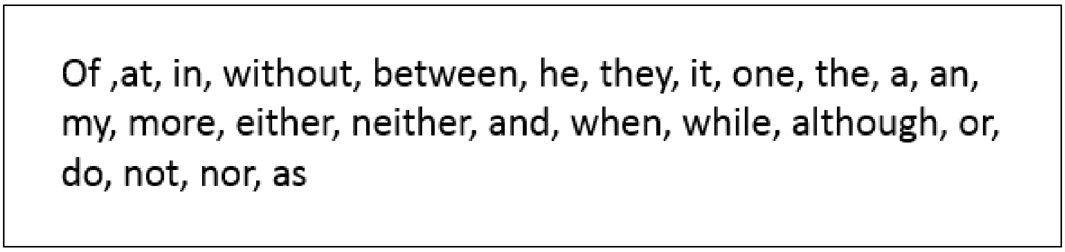
2.3 Research of Koppel et al. (2005)

NLI was first proposed by Koppel et al. [6] in 2005. They defined a “stylistic feature” set, which consists of function words[2], letter n-grams, errors and idiosyncrasies.

1. Function words

   The word plays a grammatical role but has little lexical meaning themselves. Some examples of English function words are shown in Figure 2.4.

[2] http://www.psych.nyu.edu/pylkkanen/Neural_Bases/13_Function_Words.pdf



Of ,at, in, without, between, he, they, it, one, the, a, an, my, more, either, neither, and, when, while, although, or, do, not, nor, as

Figure 2.4. Examples of function words

2. Letter n-grams

   They are the same with character n-grams.

3. Errors and idiosyncrasies

   Orthography, syntax, neologisms, POS bi-grams (extracted from Brown Corpus[3]) errors are used in their experiment.

They trained linear support vector machines (SVMs) to classify English essays of International Corpus of Learner English (ICLE)[4] Ver.1 into five native languages (Czech, French, Bulgarian, Russian, and Spanish) and achieved 80.2% accuracy. They showed that linear SVMs are well suited to an NLI task.

2.4 Research of Bykh and Meurers (2012)

Even though NLI was initiated around 10 years back, notable works have been published in the recent years. Bykh and Meurers [7] defined the n-grams occurring in at least two essays as "recurring n-grams". It was the first work to take feature (n-grams) selection into account in the NLI field.

A recurring n-gram has two obvious advantages as follows:

1. Reducing noisy data caused by spelling errors.

2. Shortening the training time by dimensionality reduction.

After mapping essays into the vector space, they trained linear SVMs to classify

[3] http://clu.uni.no/icame/brown/bcm.html

[4] https://www.uclouvain.be/en-cecl-icle.html

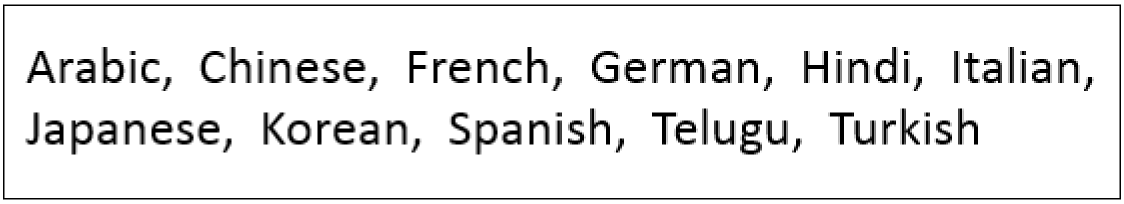
English essays of the ICLE Ver.2 dataset[5] into seven different native languages (Bulgarian, Chinese, Czech, French, Japanese, Russian, Spanish), achieving 89.7% accuracy.

Our proposed method is inspired by their paper, i.e., adopting recurring n-grams. In addition, since our feature vectors used for training are high-dimensional, we adopt n-grams appearing in more than 10 essays instead of at least two essays, which has a beneficial effect on selecting informative features.

2.5 Research of Gebre et al. (2013)

In order to improve accuracy, Gebre et al. [10] proposed a new perspective to tackle an NLI in the first NLI shared task [9] in 2013, i.e., adopting a term weighting technique before training. In their study, by scoring on the features that combine character n-grams (n=1~6), word n-rams (n=1, 2) and POS tags n-grams (n=1~4) with TF-IDF, they trained the above-mentioned linear SVMs to classify English essays in TOEFL11 dataset [8] and obtained 84.55% accuracy.

Contrary to ICLE dataset, which is used in [6] [7] facing the problem of topic bias, in TOEFL11 dataset, totally 12,100 English essays are evenly distributed by 11 native languages, as is shown in Figure 2.5. It is designed particularly for an English NLI task without topic bias compared with other corpora, such as ICLE.



Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish

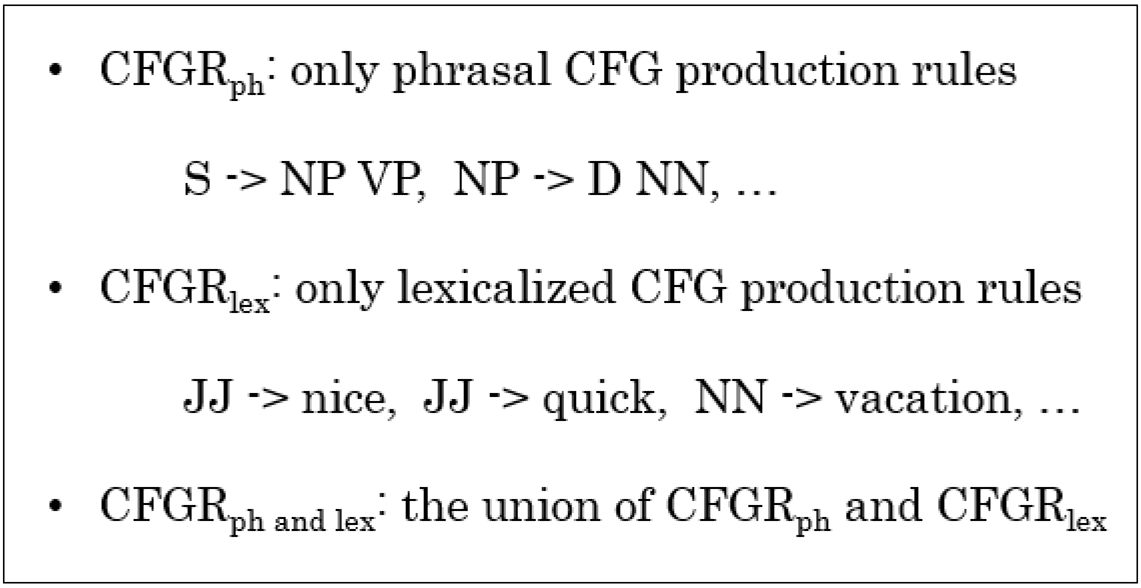
Figure 2.5. Native languages metadata in TOEFL11

We will extend this novel idea, which adopts the term weighting technique, that is, we take BM25 [11] [12] into consideration for achieving higher accuracy in our work.

[5] https://www.uclouvain.be/en-277586.html

2.6 Research of Bykh and Meurers (2014)

Bykh and Meurers [13] defined three new features according to context-free grammar-production rules (CFGR) as follows:



- CFGR_{ph}: only phrasal CFG production rules

  S -> NP VP, NP -> D NN, …

- CFGR_{lex}: only lexicalized CFG production rules

  JJ -> nice, JJ -> quick, NN -> vacation, …

- $\text{CFGR}_{\text{ph and lex}}$: the union of CFGR_{ph} and CFGR_{lex}

Figure 2.6. CFGR in [13]

They obtained 84.8% by using TOEFL11, higher than that reported in all works in [9].

Since CFGR, as is shown in Figure 2.6, indicates the syntactic grammar rules in a sentence, it is an informative feature in English NLI. Hence, we wonder whether CFGR can be used in Chinese NLI to obtain higher performance.

2.7 Research of Malmasi and Dras (2014)

Malmasi and Dras [14] first developed an NLI method for a Chinese dataset. Not only did their feature set adopt POS tag n-grams (n=1, 2, 3) and function words but they also used CFGR, which can capture the structure of syntactic grammar constructions. They applied their method to Chinese Learn Corpus (CLC) in which there are totally 3,216 essays consisting of 11 native languages, as is shown in Table 2.2. After 10-fold cross validation, they achieved 70.6% accuracy.

Table 2.2. Essay distribution of CLC [14]

Native Language	# of essays
Filipino	415
Indonesian	402
Thai	400
Laotian	366
Burmese	349
Korean	330
Khmer	294
Vietnamese	267
Japanese	180
Spanish	112
Mongolian	101

2.8 Summary

The details of the abovementioned related works are summarized in Table 2.3.

Table 2.3. **Summarization of related works**

Author	Dataset	# of L1	Classifier	Technique	Accuracy
Koppel et al. (2005) [6]	ICLE Ver.1	5	Linear SVM	Function words Character n-grams Spell errors	80%
Bykh et al. (2012) [7]	ICLE Ver.2	7	Linear SVM	Recurring n-grams	89.7%
Gebre et al. (2013) [10]	TOEFL11	11	Linear SVM	TF-IDF term weighting method	84.55%
Bykh et al. (2014) [13]	TOEFL11	11	Linear SVM	CFGR_{ph} and CFGR_{lex}	84.8%
Malmasi et al. (2014) [14]	CLC	11	Linear SVM	POS n-grams Chinese function words Chinese CFGR	70.6%

However, there are four problems to be solved in modern NLI research field.

a. Existing studies in the NLI domain do not take into account noisy essays, which lead to ineffective training.

b. Features in existing NLI studies are restricted to either lexical or syntactic features, which is not informative enough.

c. Term weighting schemes used for scoring features have not been investigated sufficiently.

d. The current CLC dataset [15] is not sufficiently large to reflect Chinese writing styles corresponding to each native language.

These abovementioned problems with corresponding prior works are listed in the Table 2.4.

Table 2.4. Summary of existing problems and prior works

Prior works Problems	[6]	[7]	[9]	[10]	[13]	[14]
a	✓	✓	✓	✓	✓	✓
b	✓	✓	✓	✓		✓
c	✓	✓			✓	✓
d						✓

3. JCLC Dataset

In our research, we adopt the Jinan Chinese Learner Corpus (JCLC) [16] instead of using the CLC dataset in [14]. JCLC is the first large scale corpus of Chinese L2 made available to research community. JCLC contains a total of 8,739 essays written by foreign students in an examination or as homework. Those students are learning Chinese at a variety of universities in China. The proficiency level of all essays are categorized based on Chinese study period and include: low (less than 1 year), medium (1-3 years) and high (more than 3 years). The proficiency level distribution is presented in Figure 3.1.

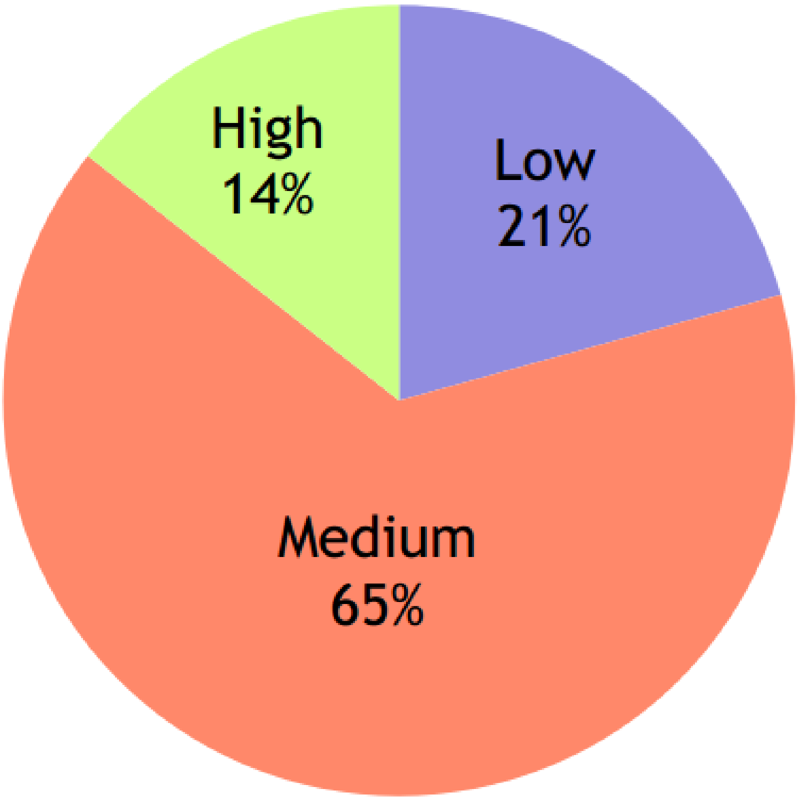


Figure 3.1. Proficiency distributions in JCLC

In addition, a histogram of essay lengths is shown in Figure 3.2. Since essays in this corpus are collected from various tasks, high variability in essay length[6] can be observed.

[6] Essay length used in this work is measured by the number of Chinese characters.

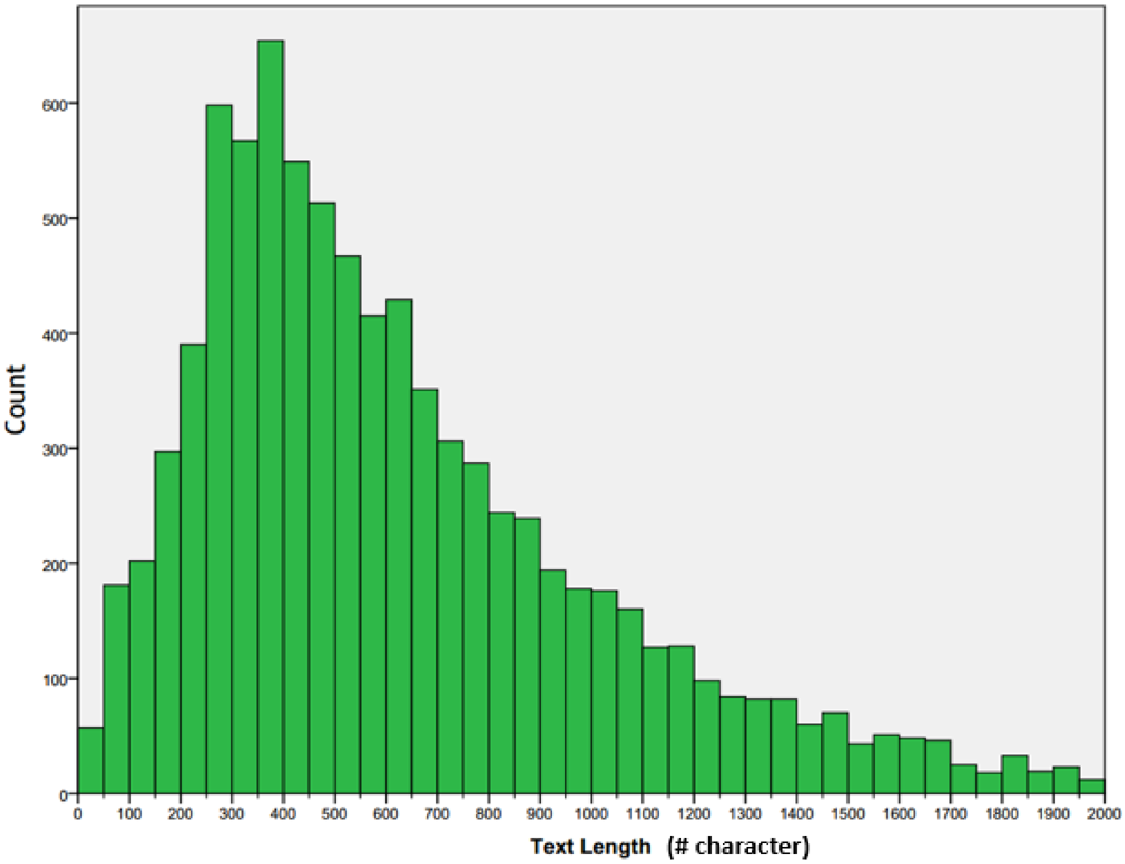


Figure 3.2. Histogram of essay lengths in JCLC dataset

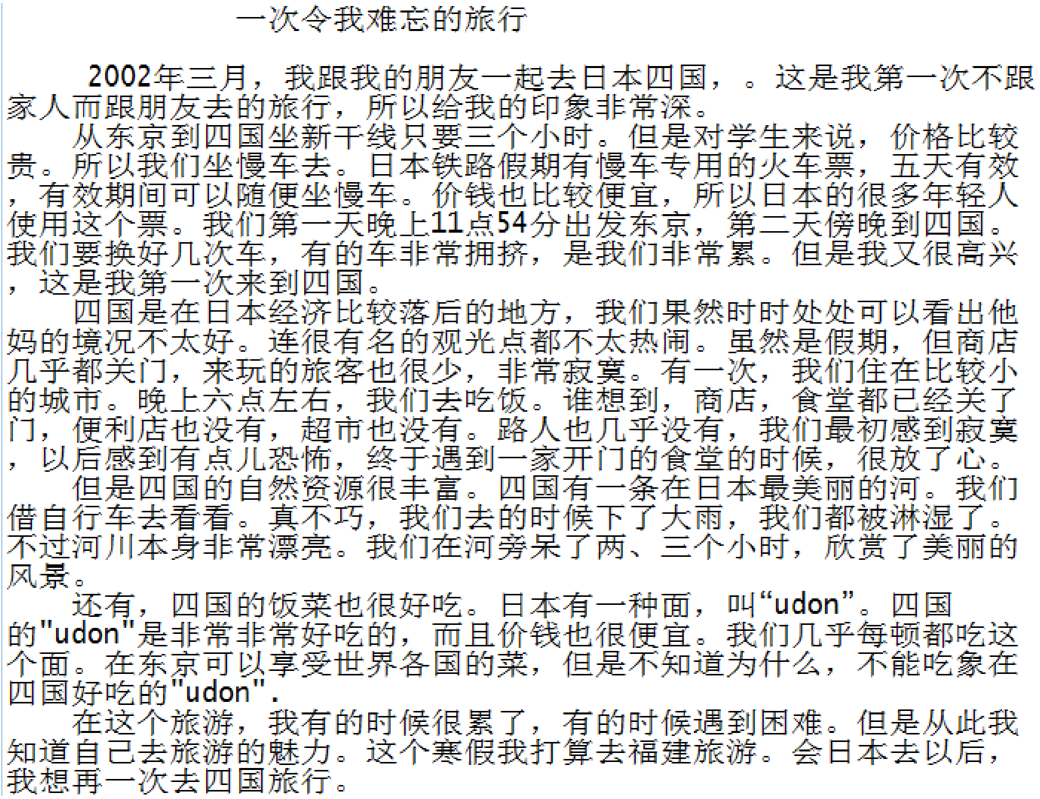
We extract all essays with explicit native language metadata in the dataset. The essay distribution in JCLC is listed in Table 3.1.

.

Table 3.1. Essay distribution of JCLC

Native Language	**# of essays**
Indonesian	3,381
Thai	1,307
Vietnamese	822
Korean	568
Burmese	410
Laotian	398
Khmer	329
Filipino	293
Japanese	270
Mongolian	119
Total	**7,897**

As a representative example, Figure 3.3 shows a Chinese essay whose native language label is Japanese.



一次令我难忘的旅行

2002年三月，我跟我的朋友一起去日本四国，。这是我第一次不跟家人而跟朋友去的旅行，所以给我的印象非常深。

从东京到四国坐新干线只要三个小时。但是对学生来说，价格比较贵。所以我们坐慢车去。日本铁路假期有慢车专用的火车票，五天有效，有效期间可以随便坐慢车。价钱也比较便宜，所以日本的很多年轻人使用这个票。我们第一天晚上11点54分出发东京，第二天傍晚到四国。我们要换好几次车，有的车非常拥挤，是我们非常累。但是我又很高兴，这是我第一次来到四国。

四国是在日本经济比较落后的地方，我们果然时时处处可以看出他妈的境况不太好。连很有名的观光点都不太热闹。虽然是假期，但商店几乎都关门，来玩的旅客也很少，非常寂寞。有一次，我们住在比较小的城市。晚上六点左右，我们去吃饭。谁想到，商店，食堂都已经关了门，便利店也没有，超市也没有。路人也几乎没有，我们最初感到寂寞，以后感到有点儿恐怖，终于遇到一家开门的食堂的时候，很放了心。

但是四国的自然资源很丰富。四国有一条在日本最美丽的河。我们借自行车去看看。真不巧，我们去的时候下了大雨，我们都被淋湿了。不过河川本身非常漂亮。我们在河旁呆了两、三个小时，欣赏了美丽的风景。

还有，四国的饭菜也很好吃。日本有一种面，叫“udon”。四国的"udon"是非常非常好吃的，而且价钱也很便宜。我们几乎每顿都吃这个面。在东京可以享受世界各国的菜，但是不知道为什么，不能吃象在四国好吃的"udon".

在这个旅游，我有的时候很累了，有的时候遇到困难。但是从此我知道自己去旅游的魅力。这个寒假我打算去福建旅游。会日本去以后，我想再一次去四国旅行。

Figure 3.3. Example of an essay in JCLC

Besides native language, metadata, such as age and gender, are included in this corpus, which is the same as the CLC corpus. In this work, only native language as informative label is used as in [14].

4. Methodology

The task of NLI can be regarded as a multiclass text classification. In our proposed method, four elements have been taken into consideration to achieve high accuracy: 1) How we should select effective data to extract features before training, 2) what kind of features we should adopt, 3) how we can estimate the importance of each feature, and 4) how we should construct machine learning models.

In order to solve the abovementioned problems, we adopt four techniques. For 1), a noisy data cleaning method based on the essay length is proposed. For 2), 1-skip-bi-gram feature which contains both lexical and syntactic information is exploited. For 3), a term weighting technique of BM25 is employed. For 4), a hierarchical structure of linear SVM classifiers is put forward in our method.

4.1 Noisy Data Elimination

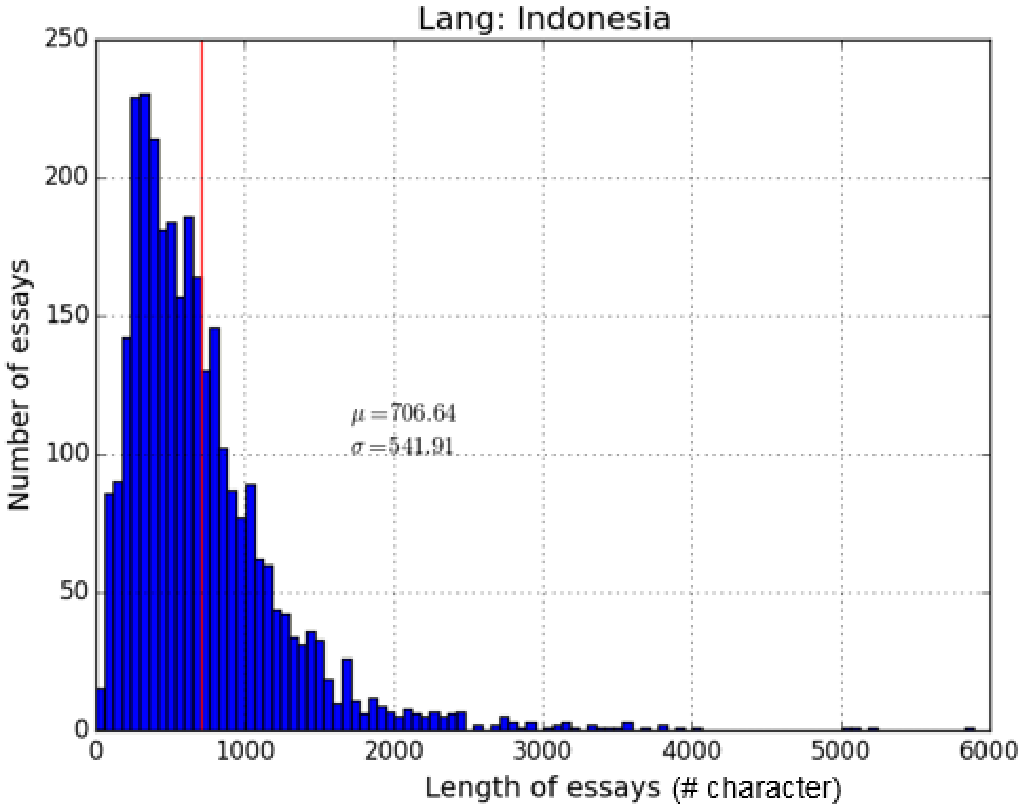


Figure 4.1. Essay length distribution of students from Indonesia

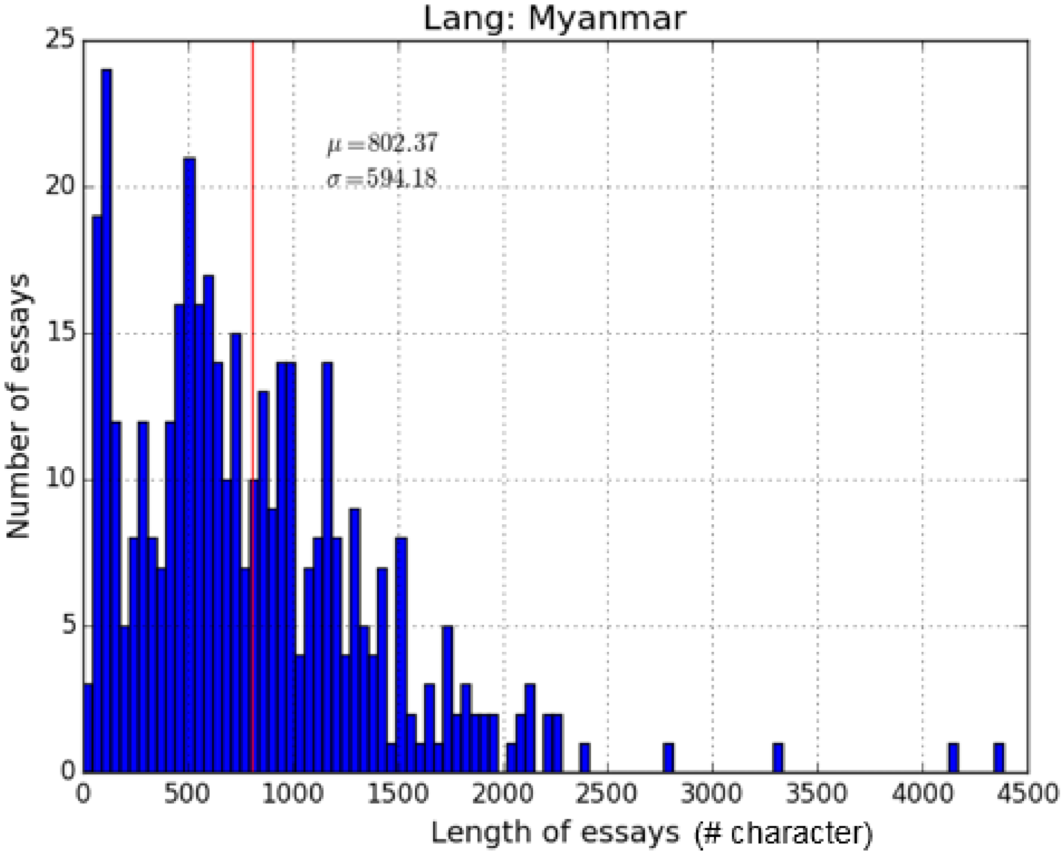
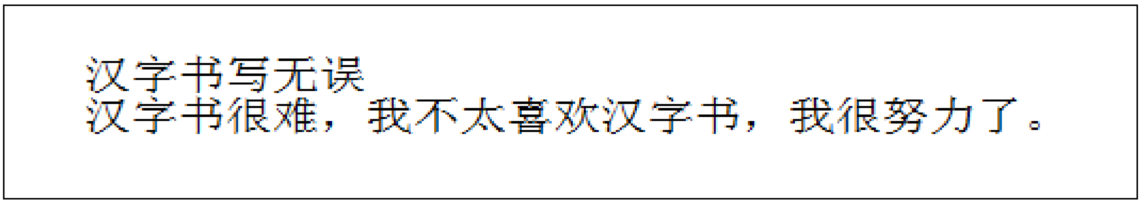


Figure 4.2. Essay length distribution of students from Myanmar

Figure 4.1 and Figure 4.2 show the length distribution of essays written by students of Indonesia and Myanmar in JCLC. As shown in Figure 4.1 and Figure 4.2, we can see that there are some very short and very long essays. For example, a short essay consists of only one sentence, as depicted in Figure 4.3. Such an essay has the tendency to consist of small L2-related characteristics because the essay is too short. Further, we assume that very long essays consist of the small characteristics of L2, because students who could write long essays might be the top level of L2 that results in the small appearance of the L2-related characteristics. These essays, i.e., both very short and very long essays, might not be suitable for the training dataset.



汉字书写无误
汉字书很难，我不太喜欢汉字书，我很努力了。

Figure 4.3. Example of a short essay

In order to select effective training essays by discarding both very long and very short ones, the following equation is employed to filter out such essays:

$$\mu_i - n_1 \times \sigma_i < essaylength < \mu_i + n_2 \times \sigma_i \qquad (4\text{-}1)$$

where i represents native language i , μ_i and σ_i denote the mean and the variance of the length of essays whose native language is i , respectively. n_1 and n_2 represent the parameters used for controlling the number of discarded essays.

Our goal in this step is to find a suitable pair (n_1, n_2) by which the highest accuracy can be obtained.

4.2 Feature Combination

Both the lexical and the syntactic features adopted in our proposed method are listed in Table 4.1. Lexical features such as character n-gram, word n-gram, and part-of-speech (POS) tag n-gram have been commonly utilized in NLI. As for syntactic features, in order to investigate the structure of grammatical rules used by students from different countries, we examine the context-free grammar production rules (CGPR) by using Stanford CoreNLP Ver. 3.5.2 [17]. As is shown in Figure 4.4, after parsing a Chinese sentence, production rules, each of which is a pair of directly connected predicate and word, are extracted to be used as one of the features. An example of the extracted rules is shown within the red frame in Figure 4.4.

Table 4.1. Our adopted features

Features
Character 1,2,3-gram
Word 1-gram
POS tag 1,2,3-gram
CFGR
Function words
1-skip-bi-gram

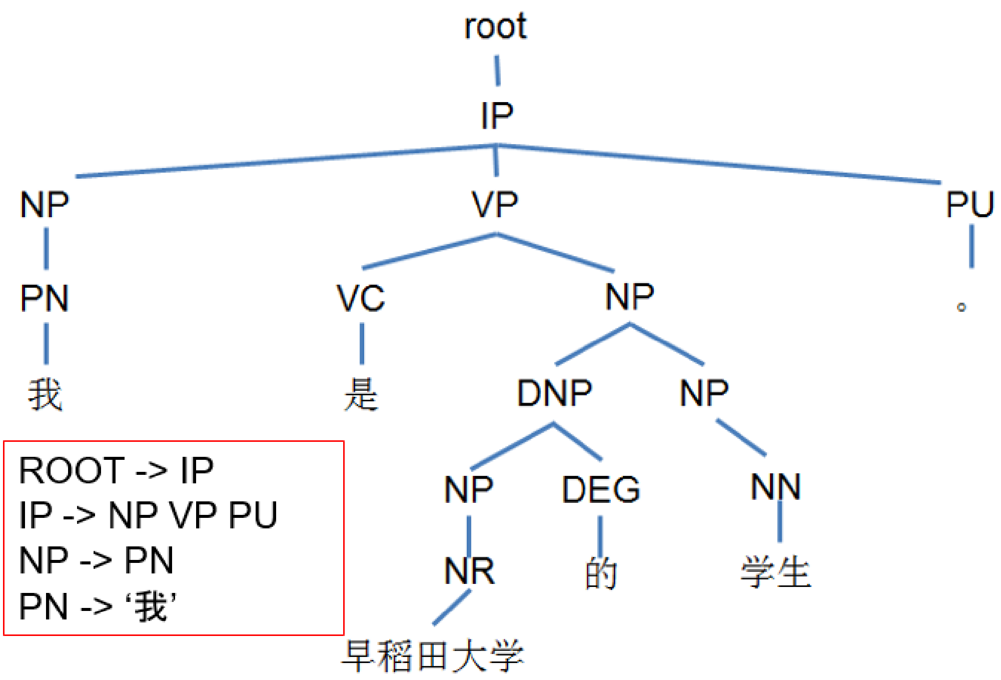
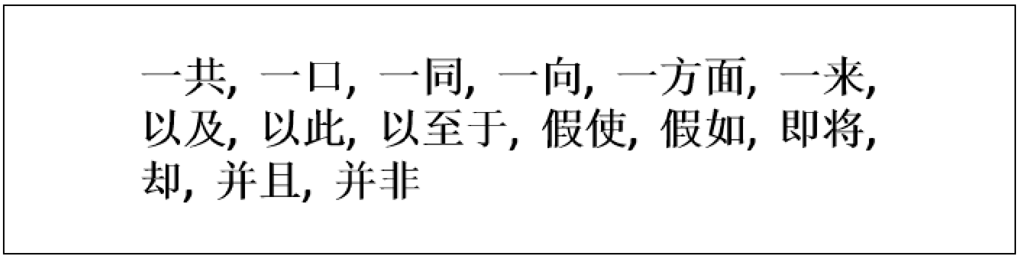


Figure 4.4. Example of production rules parsed from a sentence

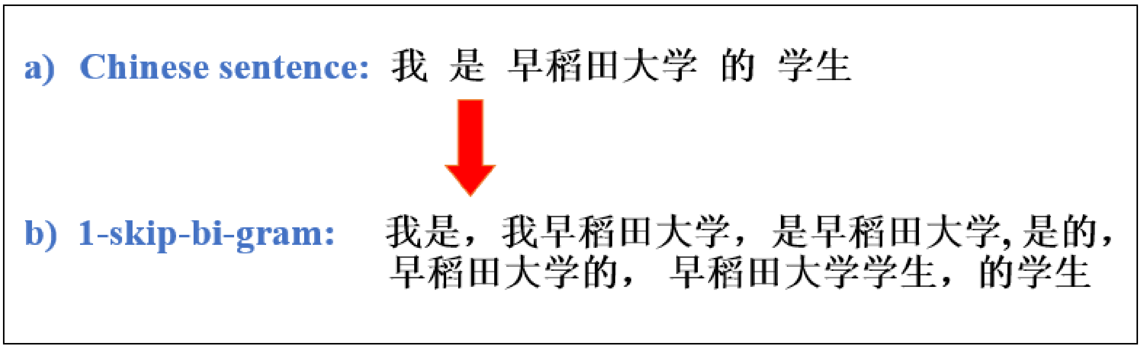
In addition, we also combine a list of 449 Chinese function words compiled by Malmasi and Dras [14] into our experiment. Examples of such Chinese function words are given in Figure 4.5.



一共, 一口, 一同, 一向, 一方面, 一来,
以及, 以此, 以至于, 假使, 假如, 即将,
却, 并且, 并非

Figure 4.5. Examples of Chinese function words

The last feature we adopt is skip-gram. Although skip-gram is a technique commonly used in speech processing [18], it was first applied to model context in [19] and achieved high performance. In view of this, the usefulness of skip-grams is explored in our work. Figure 4.6 a) shows a Chinese sentence segmented with word-base. Figure 4.6 b) presents the corresponding 1-skip-bi-grams of sentence in a).



a) Chinese sentence: 我 是 早稻田大学 的 学生
b) 1-skip-bi-gram: 我是，我早稻田大学，是早稻田大学, 是的，早稻田大学的， 早稻田大学学生，的学生

Figure 4.6. Example of a Chinese sentence with corresponding 1-skip-bi-gram

As is shown in Figure 4.6 b), in 1-skip-bi-grams, not only can adjacent tokens contain lexical information (the same with word bi-gram), but skipped tokens can also provide us with syntactic information (similar to grammatical dependency). 1-skip-bi-gram is liable to be a useful feature for Chinese NLI. Additionally, in k-skip-n-gram model, it is obvious to note that the skip-gram feature dimension will grow enormously as k and n increase, which will result in poor performance due to the generation of considerable useless n-grams. As a result, only 1-skip-bi-gram is taken into consideration in our proposed method

As is explained in section 2.4, we adopt recurring n-grams, selecting n-grams that occur in more than ten essays as informative features.

4.3 Term Weighting Method of BM25

Thus far, most of the works have adopted term frequency (TF) [6] [7] [13] [14] or term frequency–inverse document frequency (TF-IDF) [10] as a term weighting technique, when mapping essays into a vector space. Typical TF and TF-IDF calculations are shown in (4-2) and (4-3), respectively.

$$w_{ij} = tf_{ij} \tag{4-2}$$

$$w_{ij} = tf_{ij} \times \log \frac{|D|}{df_i} \tag{4-3}$$

where w_{ij} represents term t_i 's weighting in document j . tf_{ij} shows the term frequency of term t_i in document j . df_i shows the document frequency of term t_i , i.e., the number of documents that consist of term t_i . $|D|$ represents the total number of Chinese essays.

From Weinert [20], we know that people would like to remember regular phrases when learning a foreign language. Nevertheless, in case a learner's level is high, he/she will use different phrases to express the same meaning. On the other hand, if the learner's level is low, he/she will use almost the same phrases, which will lead to a linear increase in term weighting when adopting TF. Here, assume that two Korean students with different Chinese levels are writing essays where the frequency difference of the same word is dramatically large. In such a case, a contradiction will occur. For targeting expert students, the term weight should be large; however, for targeting beginner students, the term weight should be small. Nonetheless, we must use the same term weighting for the same word.

Thus, it is essential to fill the gap. Gebre et al. [10] make use of log arithmetic to modify

TF.

$$w_{ij} = \log(tf_{ij} + 1) \times \log \frac{|D|}{df_i} \tag{4-4}$$

According to [12], BM25 has a positive effect on dampening the term frequency and weakening a single term's impact. Hence, we examine the application of BM25, which is defined as (4-5).

$$w_{ij} = \frac{tf_{ij} \times (k_1 + 1)}{tf_{ij} + k_1 \times \left(1 - b + b \times \frac{len(d_j)}{avgdl}\right)} \times \log \frac{|D|}{df_i} \tag{4-5}$$

where $avgdl$ denotes the average document length and $len(d_j)$ represents the length of document j . According to Robertson and Zaragoza [11], high performance can be obtained under the condition $1.2 \le k_1 \le 2$ and $b = 0.75$. In our proposed method, we assigned 1.2 to k_1 and 0.75 to b .

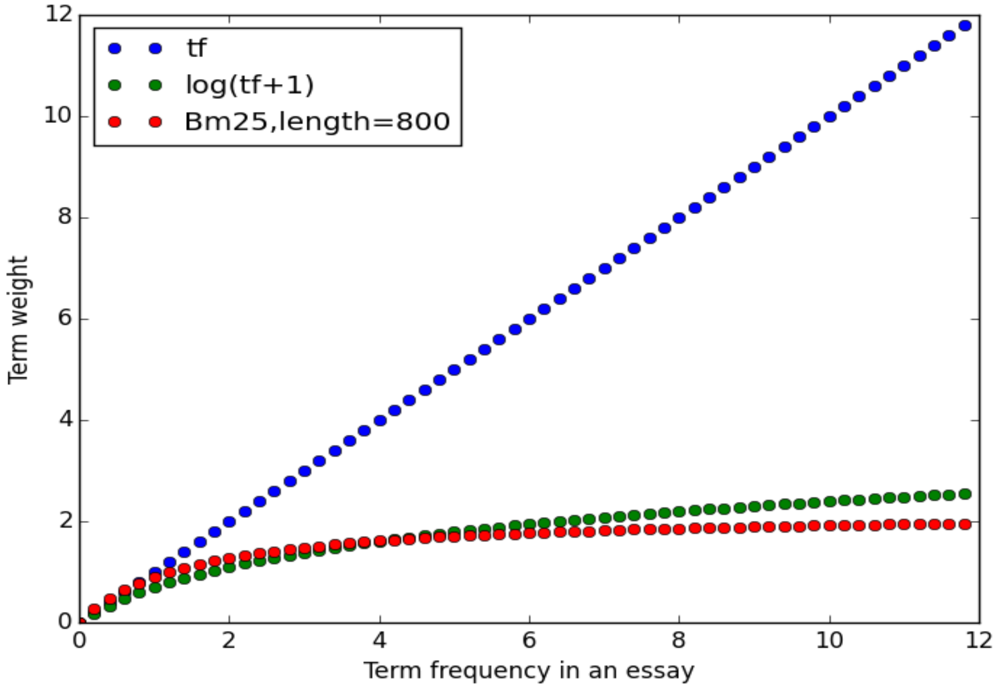


Figure 4.7. Term weight increase of tf, log(tf + 1), BM25

As shown in Figure 4.7, in the case of BM25, even if the term frequency is large, the term weight is small compared with TF-IDF. In other words, even when a term occurs rarely, it can be assigned a high weight.

4.4 Hierarchical Classifiers

Most of the works on the first NLI shared task [9] verified that the application of linear SVMs to NLI can achieve higher performance than that of the other machine learning methods. As in previous works, we continue exploiting the application of linear SVMs to Chinese NLI.

In the TOEFL11 dataset, English essays are evenly distributed by 11 native languages. In contrast, in the JCLC dataset, the number of essays corresponding to each native language is considerably different from each other. As is shown in Table 3.1, since Chinese essays written by Indonesians form more than 40% of the dataset, it is more feasible to identify the Indonesian label first and then, conduct a flat classification of the remaining essays. On the basis of this idea, we put forward a new hierarchical structure of linear SVM classifiers, as shown in Figure 4.8.

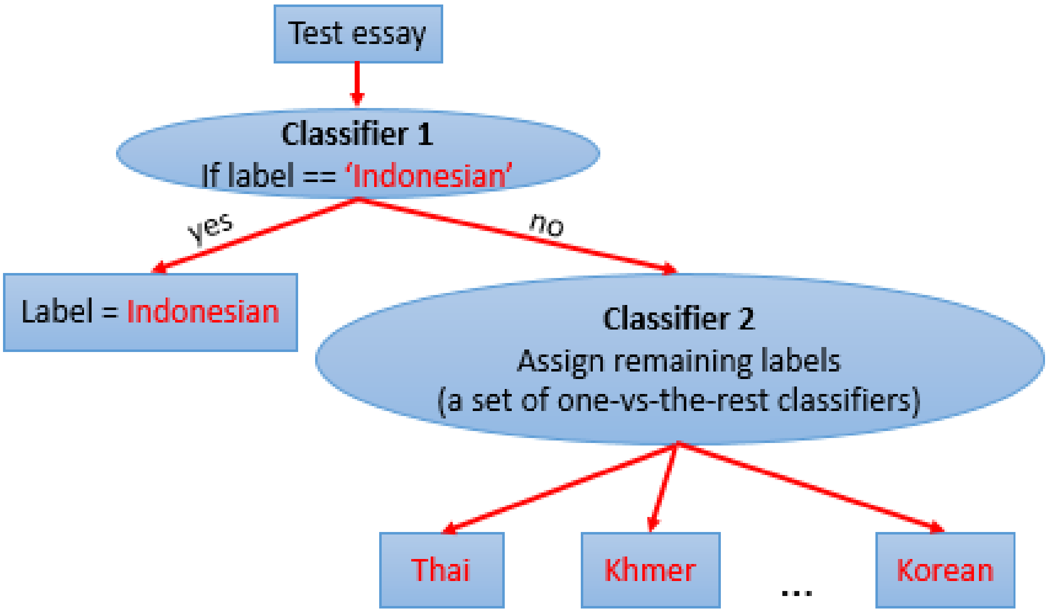


Figure 4.8. Structure of our hierarchical SVM classifiers

In our classification, we determine whether a test essay belongs to the label "Indonesian" by using classifier 1 first. If so, then the classification is complete. Otherwise, continue to identify which label should be assigned among the remaining nine native languages, such as Thai, Korean by using classifier 2 (one-vs-the-rest). Here, one-vs-the-rest method chooses the label (native language) which classifies the test essay with greatest margin.

5. Results and Discussion

5.1 Evaluation Method

We run the 10-fold cross validation experiment: the JCLC dataset is divided randomly into ten subsets of equal size, nine of which are used for training and the tenth subset is used for testing. This process will be repeated ten times with each subset being held out for test exactly once. However, in order to evaluate the robustness of our model, besides test dataset, we divide the remaining data into training dataset and tuning dataset at a ratio of 4:1 in each round of 10-fold cross validation, as is shown in Figure 5.1.

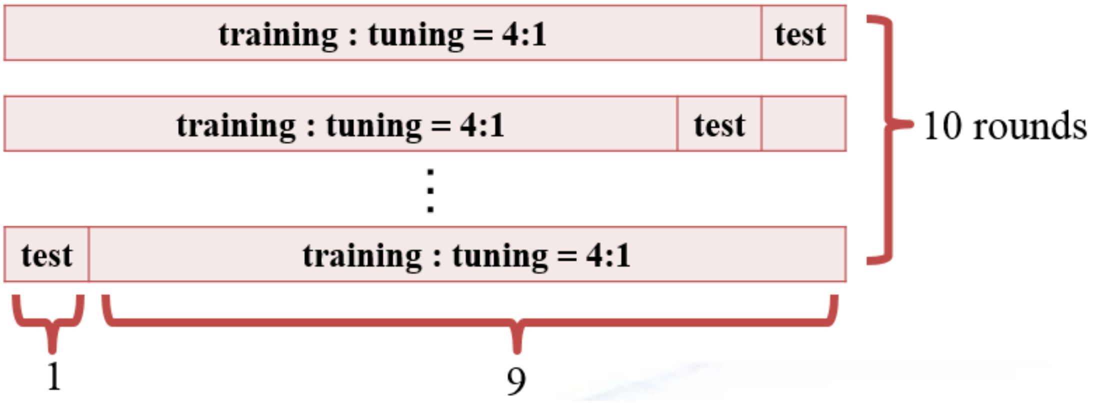


Figure 5.1. 10-fold cross validation in this work

In this case, our proposed model is trained with varying (n_1, n_2) and then test on the tuning dataset. After tuning phase, parameters will be fixed so that they can be applied to the machine learning model for test particularly.

5.2 Measurement

In this work, overall accuracy is employed as measurement, as same as prior works in Chapter 2. Overall accuracy is calculated as the following equation:

$$accuracy = \frac{\sum_{i=1}^{10} ta_i}{|D|} \qquad (5\text{-}1)$$

where, ta_i is the accuracy whose native language label is i . $|D|$ represents the total number of Chinese essays in the test dataset. We calculate the overall accuracy in each round of 10-fold cross validation and then adopt the average of the 10-fold results as our final accuracy.

5.3 Choosing Parameters n1 and n2

In parameter tuning phase, classifiers are trained with varying pair of (n_1, n_2) where =0, 1, 2, 3 and =0, 1, 2, 3, respectively. After training, we test the trained model with the tuning dataset to select best parameters. Table 5.1 summarizes the candidate pairs of (n_1, n_2) in each round of 10-fold cross validation (left column), by which highest tuning accuracy can be obtained (right column).

Table 5.1. Result of tuning parameters

(n_1,n_2)	**Best tuning accuracy**
(2,3)(3,3)	0.7563
(2,3)(3,3)	0.7542
(2,3)(3,3)	0.7549
(3,2)	0.7584
(2,3)(3,3)	0.7563
(2,3)(3,3)	0.7668
(2,3)(3,3)	0.7598
(2,1)(2,3)(3,1)(3,3)	0.7521
(2,3)(3,3)	0.7500
(2,3)(3,3)	0.7437

Even though candidate pairs of (n_1, n_2) are obtained, the next issue is, in a multiple candidates case, which pair of (n_1, n_2) we should choose. In this step, a pair of (n_1, n_2) is selected at random and is applied to the proposed model for test. Table 5.2 presents the test result of 10-fold cross validation. As is shown, the left column is the randomly selected pair of (n_1, n_2), with its corresponding test accuracy in the right column. Finally, after averaging all of them, an accuracy of 0.753 is achieved by our proposed model.

Table 5.2. Test result with randomly selected parameters

(n_1,n_2)	**Test accuracy**
(3,3)	0.7465
(2,3)	0.7452
(3,3)	0.7693
(3,2)	0.7262
(2,3)	0.7465
(2,3)	0.7769
(3,3)	0.7579
(2,3)	0.7452
(3,3)	0.7490
(3,3)	0.7630
Final score	**0.753**

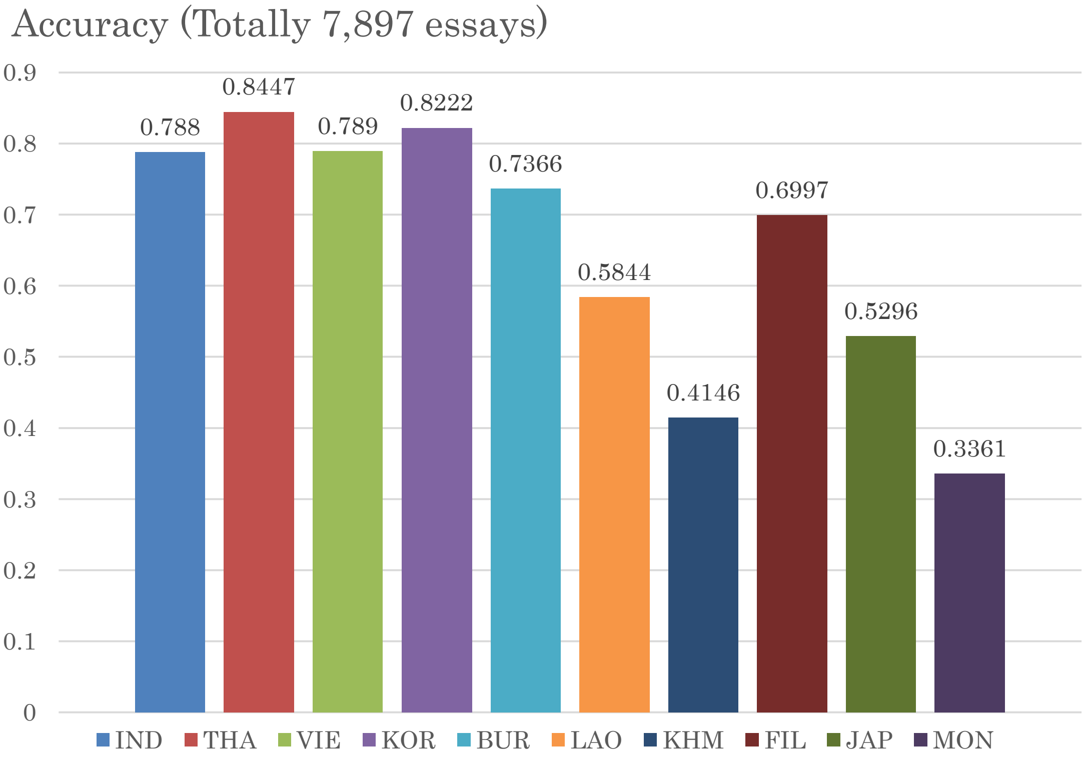


Figure 5.2. Accuracy of each native language

Figure 5.2 shows the accuracy of each native language in our experiment. As is shown, essays written by Thai achieves the best accuracy of 0.8447, whereas essays written by Mongolian gains the poorest performance of 0.3361 accuracy. Such low accuracy arises from the fact that there is not sufficient training data so that effective training cannot be carried out.

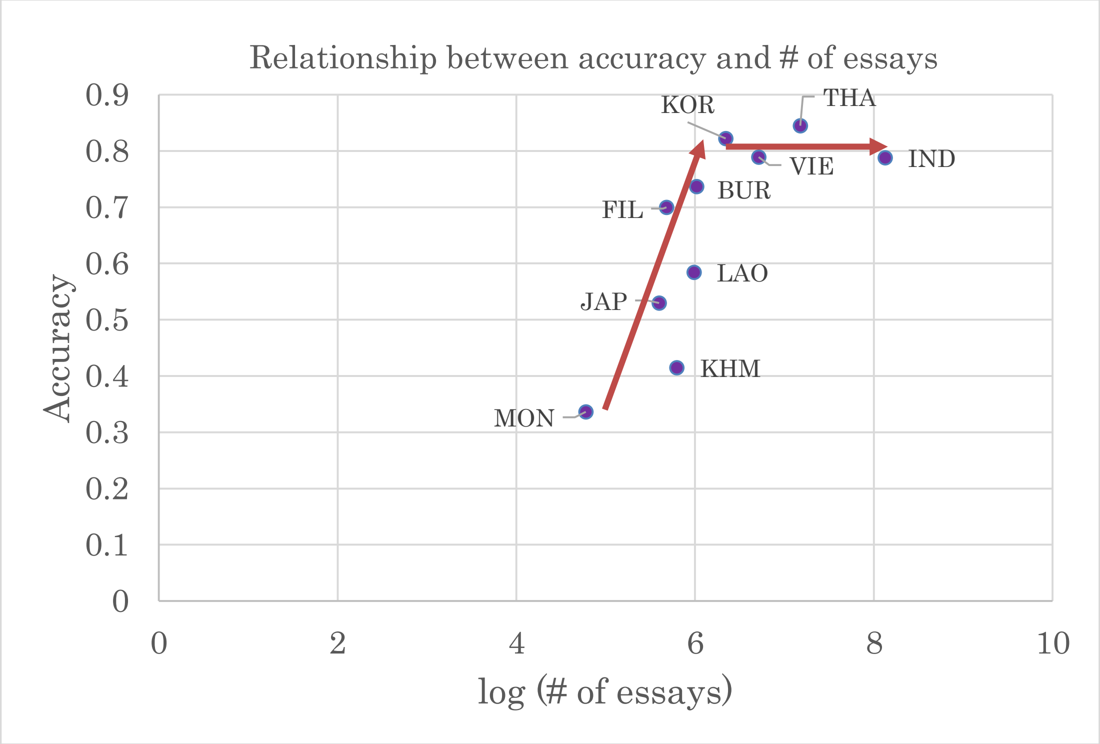


Figure 5.3. Relationship between accuracy and log (# of training essays)

Figure 5.3 illustrates the relationship between accuracy and the number of essays. In order to demonstrate the pattern intuitively, we calculate the logarithm of the number of essays as x-axis. As is shown, when $\log(\text{\# of essays}) \leq 6.3$, i.e., the number of essays is smaller than or around 550, there seems a liner upward trend that the accuracy increases in direct proportion to the number of essays. When $\log(\text{\# of essays}) > 6.3$, the accuracy seems to reach a limit around 0.8 with a constant forward trend.

Further, we confirmed the accuracy without cleaning noisy training essays, as is shown in Figure 5.4. Using BM25 and our hierarchical linear SVM classifiers, we find that cleaning noisy essays in advance led to a higher accuracy of 0.014 than that obtained in the case without cleaning.

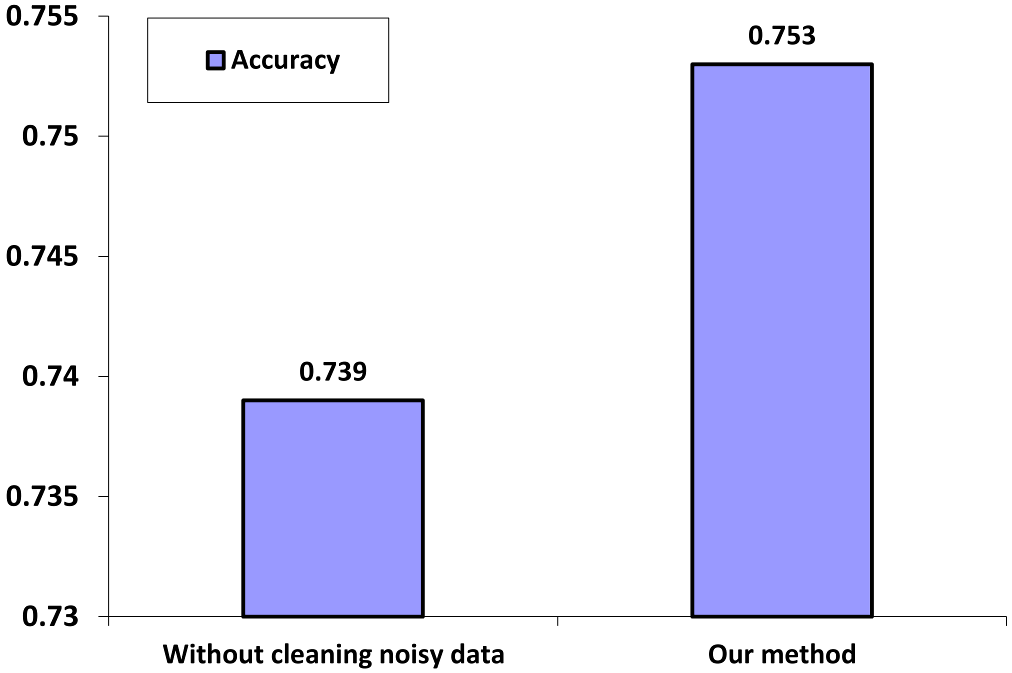


Figure 5.4. Accuracy with and without cleaning noisy data

5.4 Comparison of Term Weighting Methods

Next, we compared the accuracy of four conditions: TF shown as formula (4-2), TF-IDF shown as formula (4-3), log-arithmetic TF-IDF shown as formula (4-4), and BM25 shown as formula (4-5). Here, BM25 is our proposed method.

Figure 5.5 illustrates the accuracy of different term weighting methods adopting our proposed hierarchical linear SVM classifiers. As we can see, BM25 achieves the highest accuracy of 0.753 which outperforms other methods markedly. The result indicates that assigning a higher weight to high-frequency terms will not always lead to ideal performance.

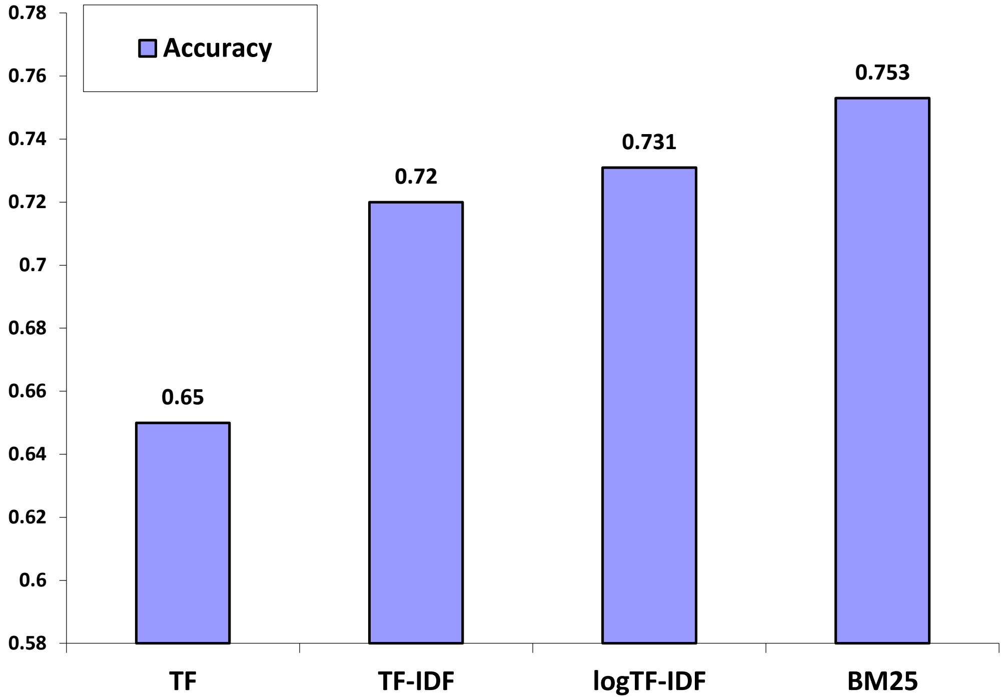


Figure 5.5. Accuracy of different term weighting methods

5.5 Linear SVMs and Hierarchical Classifiers

In this section, we compare the performance of linear SVMs and our proposed hierarchical linear SVM classifiers, both employing BM25 as the term weighting technique. As shown in Figure 5.6, we can see that our proposed hierarchical linear SVM classifier model exceeds the one-layer linear SVMs around 0.011. That is, if we know the native language distribution of all training essays in advance and consider employing our hierarchical structure of classifiers, we can achieve higher performance.

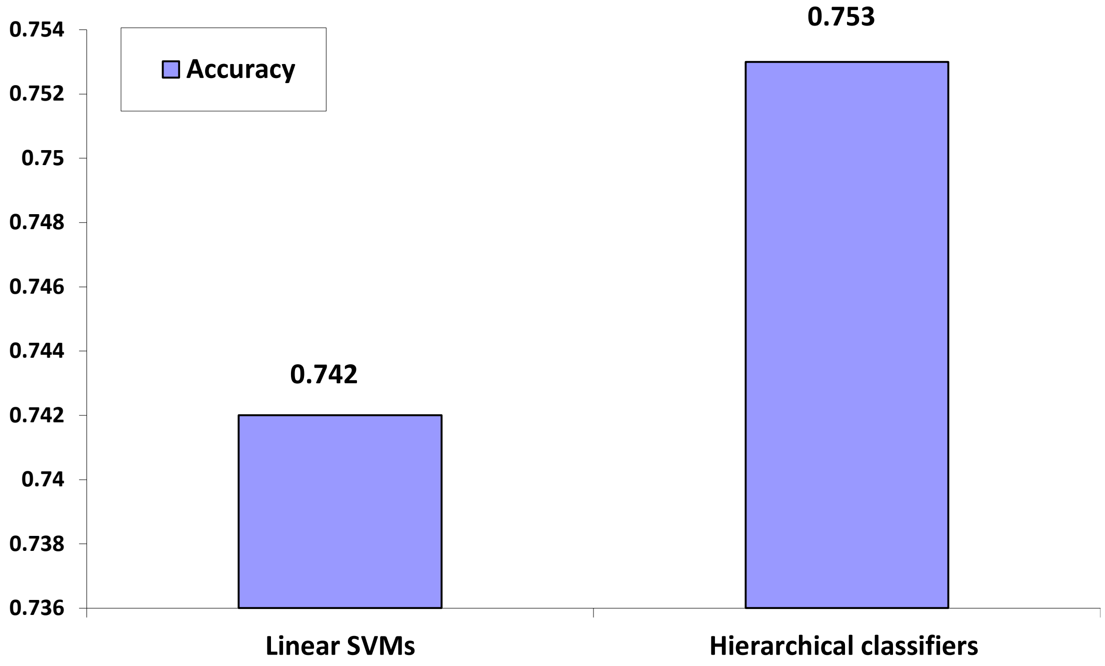


Figure 5.6. Accuracy of Linear SVMs and hierarchical classifiers

5.6 Comparison with Malmasi and Dra’s Method

Malmasi and Dras [14] developed the first application of Chinese NLI and gained 70.6% accuracy on the CLC corpus. We implemented their algorithm with the JCLC corpus used in our work and obtained an accuracy of 0.653. We compared it with that of our proposed supervised model shown in Figure 5.7. As is illustrated in Figure 5.7, our proposed method outperforms the baseline by 10%.

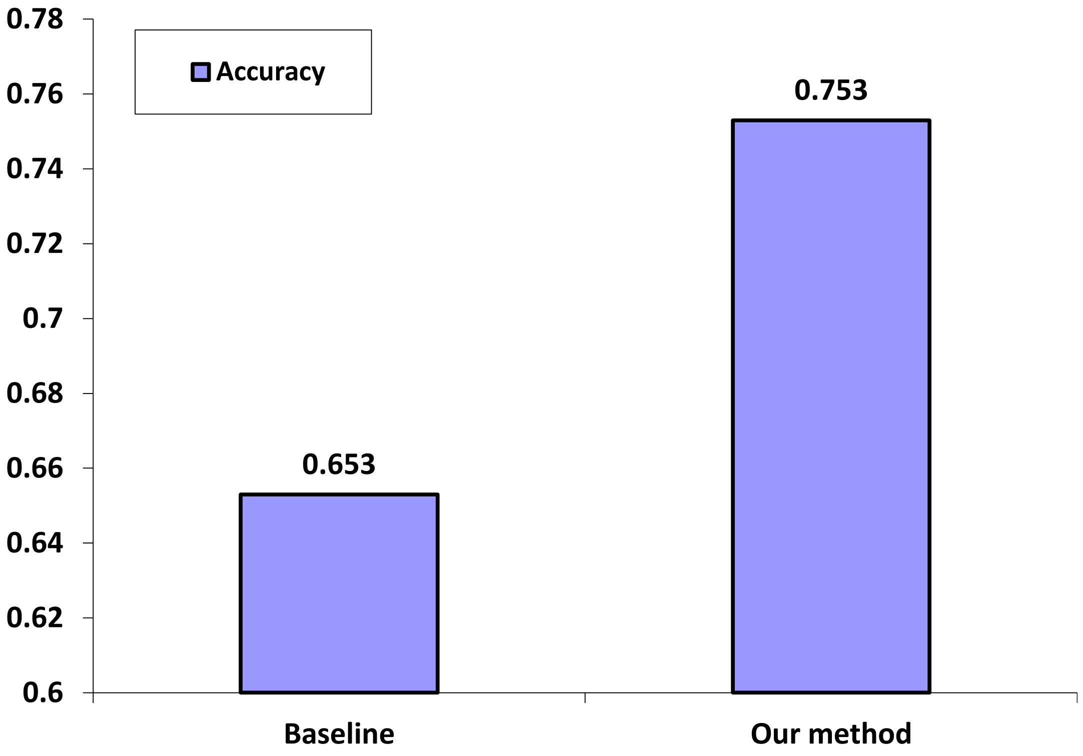


Figure 5.7. Accuracy compared with baseline

5.7 Error analysis

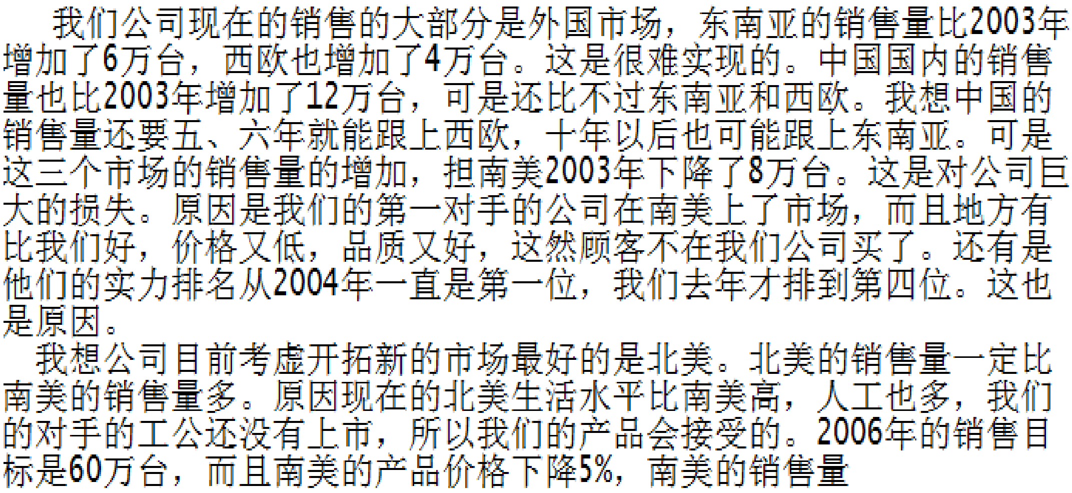
Table 5.3 and Table 5.4 illustrate the summation confusion matrix and percentage confusion matrix for each native language by 10-fold cross validation, respectively. After analyzing all the misclassified essays, we find some interesting patterns. To start with, as is stated in Section 5.3, low accuracy of L1 classes, such as Mongolian, results from insufficiency of training data. In the second place, most of mispredicted essays fall into Indonesian or Thai class. It is probably because training essays with Indonesian label and Thai label account for a large proportion, which will lead to relatively ineffective training for other native language labels. Finally, we find another though-provoking error pattern: Chinese essays written by Japanese and Koreans are prone to be mislabeled with each other. As an example, Figure 5.8 shows a Chinese essay written by a person whose L1 is Japanese, whereas this essay is assigned with Korean label by our proposed model. Contrarily, Figure 5.9 describes a Chinese essay written by a Korean is mislabeled with Japanese. In total, Japanese is classified as Korean 54 times, which takes up 20% of all Japanese essays. And Korean is classified as Japanese 38 times, which accounts for 6.69% of all Korean essays.

Table 5.3. Confusion matrix for each L1 by 10-fold cross validation

		Predicted L1									
		IND	THA	VIE	KOR	BUR	LAO	KHM	FIL	JAP	MON
Original L1	IND	2662	397	68	76	57	33	58	11	13	3
	THA	153	1104	13	20	6	5	2	3	1	0
	VIE	50	95	647	10	1	2	2	9	4	0
	KOR	31	24	6	467	1	0	1	0	38	0
	BUR	23	51	7	10	302	3	13	0	1	0
	LAO	32	88	6	3	18	232	12	2	0	4
	KHM	45	80	12	2	21	25	136	1	3	3
	FIL	48	36	1	0	0	1	2	205	0	0
	JAP	26	41	6	54	0	0	0	0	143	0
	MON	7	15	3	2	17	8	27	0	0	40

Table 5.4. Percentage confusion matrix of table 5.3

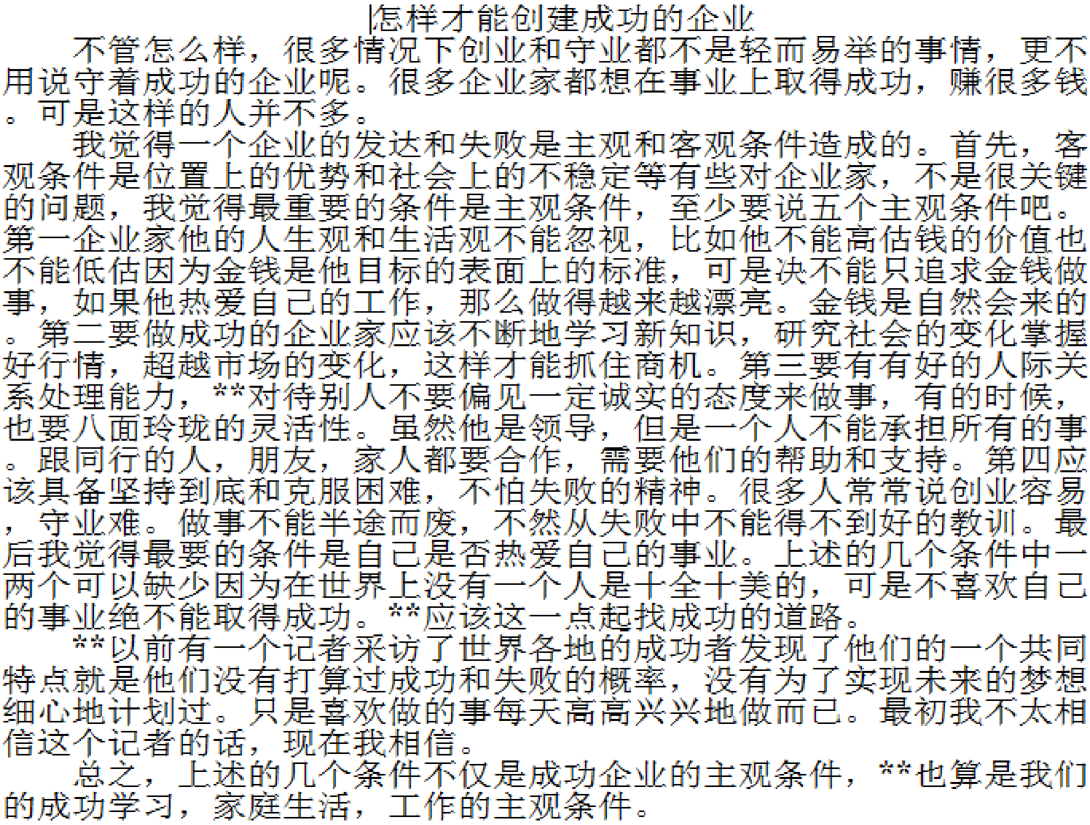
		Predicted L1									
		IND	THA	VIE	KOR	BUR	LAO	KHM	FIL	JAP	MON
Original L1	IND	0.788	0.1175	0.0201	0.0225	0.0169	0.0098	0.0172	0.0033	0.0004	0.0001
	THA	0.1171	0.8447	0.0099	0.0153	0.0046	0.0038	0.0015	0.0023	0.0008	0
	VIE	0.061	0.1159	0.789	0.0122	0.0012	0.0024	0.0024	0.011	0.0049	0
	KOR	0.0546	0.0423	0.0106	0.8222	0.0018	0	0.0018	0	0.0669	0
	BUR	0.0561	0.1244	0.0171	0.0244	0.7366	0.0073	0.0317	0	0.0024	0
	LAO	0.0806	0.2217	0.0151	0.0076	0.0453	0.5844	0.0302	0.005	0	0.0101
	KHM	0.1372	0.2439	0.0366	0.0061	0.064	0.0762	0.4146	0.003	0.0091	0.0091
	FIL	0.1638	0.1229	0.0034	0	0	0.0034	0.0068	0.6997	0	0
	JAP	0.0963	0.1519	0.0222	0.2	0	0	0	0	0.5296	0
	MON	0.0588	0.1261	0.0252	0.0168	0.1429	0.0672	0.2269	0	0	0.3361



我们公司现在的销售的大部分是外国市场，东南亚的销售量比2003年增加了6万台，西欧也增加了4万台。这是很难实现的。中国国内的销售量也比2003年增加了12万台，可是还比不过东南亚和西欧。我想中国的销售量还要五、六年就能跟上西欧，十年以后也可能跟上东南亚。可是这三个市场的销售量的增加，担南美2003年下降了8万台。这是对公司巨大的损失。原因是我们的第一对手的公司在南美上了市场，而且地方有比我们好，价格又低，品质又好，这然顾客不在我们公司买了。还有是他们的实力排名从2004年一直是第一位，我们去年才排到第四位。这也是原因。

我想公司目前考虑开拓新的市场最好的是北美。北美的销售量一定比南美的销售量多。原因现在的北美生活水平比南美高，人工也多，我们的对手的工公还没有上市，所以我们的产品会接受的。2006年的销售目标是60万台，而且南美的产品价格下降5%，南美的销售量

Figure 5.8. Example of Japanese L1 mislabeled with Korean L1



怎样才能创建成功的企业

不管怎么样，很多情况下创业和守业都不是轻而易举的事情，更不用说守着成功的企业呢。很多企业家都想在事业上取得成功，赚很多钱。可是这样的人并不多。

我觉得一个企业的发达和失败是主观和客观条件造成的。首先，客观条件是位置上的优势和社会上的不稳定等有些对企业家，不是很关键的问题，我觉得最重要的条件是主观条件，至少要说五个主观条件吧。第一企业家他的人生观和生活观不能忽视，比如他不能高估钱的价值也不能低估因为金钱是他目标的表面上的标准，可是决不能只追求金钱做事，如果他热爱自己的工作，那么做得越来越漂亮。金钱是自然会来的。第二要做成功的企业家应该不断地学习新知识，研究社会的变化掌握好行情，超越市场的变化，这样才能抓住商机。第三要有有好的人际关系处理能力，**对待别人不要偏见一定诚实的态度来做事，有的时候，也要八面玲珑的灵活性。虽然他是领导，但是一个人不能承担所有的事。跟同行的人，朋友，家人都要合作，需要他们的帮助和支持。第四应该具备坚持到底和克服困难，不怕失败的精神。很多人常常说创业容易，守业难。做事不能半途而废，不然从失败中不能得不到好的教训。最后我觉得最要的条件是自己是否热爱自己的事业。上述的几个条件中一两个可以缺少因为在世界上没有一个人是十全十美的，可是不喜欢自己的事业绝不能取得成功。**应该这一点起找成功的道路。

**以前有一个记者采访了世界各地的成功者发现了他们的一个共同特点就是他们没有打算过成功和失败的概率，没有为了实现未来的梦想细心地计划过。只是喜欢做的事每天高高兴兴地做而已。最初我不太相信这个记者的话，现在我相信。

总之，上述的几个条件不仅是成功企业的主观条件，**也算是我们的成功学习，家庭生活，工作的主观条件。

Figure 5.9. Example of Korean L1 mislabeled with Japanese L1

According to [21], Japanese and Korean language have similar lexical, syntactic, morphological language patterns. Moreover, as is stated in Chapter 1, L1 patterns of the learner have a potential influence on L2 acquisition. Based upon the above two standpoints, we may draw a conclusion that L2 writings of Japanese and Korean are difficult to distinguish due to the L1 similar transfer effects. Consequently, the essence of misidentification between Korean and Japanese L2 can be clearly interpreted.

6. Conclusion

In conclusion, in this work, we propose a new Chinese NLI method and achieved the state-of-the-art accuracy of 75.3% with the JCLC corpus, which can contribute to (1) providing instructive advice to second language education, and (2) identification of false information distributors on SNS or phishing sites. Above all, our proposed model is cross lingual, which can also be applied to accomplish English NLI.

Not only is our supervised model the first work to automatically eliminate noisy data before the training phase, but we also employ BM25 to assign an effective weight to each feature. In addition, by integrating skip-gram as an informative feature and constructing a hierarchical structure of linear SVM classifiers, our work outperforms the baseline by 10% with a robust evaluation method.

In the future, in the light of the high-dimensional feature vector in this task, an efficient feature selection method needs to be considered.

Acknowledgement

I would like to gratefully acknowledge a variety of people.

To begin with, I am deeply indebted to Professor Yamana, my advisor, who is always like the beacon lighting up my research road ahead. Owing to his patient guidance and valuable comment, this thesis is completed smoothly. It is my great honor to study under his supervision.

Apart from this, I would like to express deepest thanks to members of Yamana lab, for their kind endless help, generous advice and support during my 2-year master course life.

Publications

- Lan Wang, Masahiro Tanaka and Hayato Yamana, “What is your Mother Tongue?: Improving Chinese Native Language Identification by Cleaning Noisy Data and Adopting BM25,” Proc. of IEEE Int’l Conf. on Big Data Analysis, 2016.
- Lan Wang, Hayato Yamana, “Robust Chinese Native Language Identification with Skip-gram,” DEIM, 2016.

Reference

[1] Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow, “Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification,” Proc. of 24th COLING, pp. 2585-2602, 2012.

[2] Ria Perkins, “Native Language Identification (NLID) for Forensic Authorship Analysis of Weblogs,” New Threats and Countermeasures in Digital Crime and Cyber Terrorism. IGI Global, pp. 213-234, 2015.

[3] Bernd H. Schmitt, Yigang Pan, and Nader T. Tavassoli, “Language and Consumer Memory: The Impact of Linguistic Differences between Chinese and English,” Journal of Consumer Research, vol. 21, No. 3, pp. 419-431, 1994.

[4] Shannon Claude Elwood, Warren Weaver, and Richard E. Blahut, “The mathematical theory of communication,” Vol. 117. Urbana: University of Illinois press, 1949.

[5] Fei Xia, The Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0), Technical Reports, 2000.

[6] Moshe Koppel, Jonathan Schler, and Kr Zigdon, “Determining an Author's Native Language by Mining a Text for Errors,” Proc. of 11th KDD, pp. 624-628, 2005.

[7] Serhiy Bykh and Detmar Meurers, “Native Language Identification Using Recurring N-grams - Investigating Abstraction and Domain Dependence,” Proc. of 24th COLING, pp. 425-440, 2012.

[8] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow, “TOEFL11: A Corpus of Non-Native English,” Technical report, Educational Testing Service, 2013.

[9] Joel Tetreault, Daniel Blanchard, and Aoife Cahill, “A Report on the First Native Language Identification Shared Task,” Proc. of 8th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 48-57, 2013.

[10] Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg and Tom Heskes, “Improving Native Language Identification with TF-IDF Weighting,” Proc. of 8th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 216-223, 2013.

[11] Stephen Robertson and Hugo Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” J. of Foundations and Trends in Information Retrieval, vol. 3, No. 4, pp. 333-389, 2009.

[12] John S. Whissell and Charles L. A. Clarke, “Improving Document Clustering Using Okapi BM25 Feature Weighting,” J. of Information Retrieval, vol. 14, No. 5, pp. 466-

487, 2011.

[13] Serhiy Bykh and Detmar Meurers, “Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization,” Proc. of 25th COLING, pp. 1962-1973, 2014.

[14] Shervin Malmasi and Mark Dras, ”Chinese Native Language Identification,” Proc. 14th Conf. of the European Chapter of the Association for Computational Linguistics, pp. 95-99, 2014.

[15] Maolin Wang, Qi Gong, Jie Kuang, and Ziyu Xiong, “The Development of a Chinese Learner Corpus,” Proc. of Int’l Conf. on Speech Database and Assessments (Oriental COCOSDA), pp. 1-6, 2012.

[16] Maolin Wang, Shervin Malmasi, and Mingxuan Huang, “The Jinan Chinese Learner Corpus,” Proc. of 10th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 118-123, 2015.

[17] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” Proc. of 52nd Ann. Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60, 2014.

[18] Manhung Siu and Mari Ostendorf, “Variable n-grams and extensions for conversational speech language modelling,” IEEE Transactions on Speech and Audio Processing, vol. 8, No. 1, pp. 63-75, 2000.

[19] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks, ”A Closer Look at Skip-gram Modelling,” Proc. of the Fifth International Conference on Language Resources and Evaluation, pp. 1222–1225, 2006.

[20] Regina Weinert, “The Role of Formulaic Language in Second Language Acquisition: A Review,” Applied Linguistics, vol. 16 (2), pp. 180-205, 1995.

[21] Jasone Cenoz, Btitta Hufeisen and Ulrike Jessner, “Cross-Linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives,” pp. 149-167, 2001.