Robust Chinese Native Language Identification with Skip-gram

Lan WANG † Hayato YAMANA ‡

† Graduate School of Fundamental Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku Tokyo, Japan

‡ Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku Tokyo, Japan

‡ National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

E-mail: † wanglan0605@yama.info.waseda.ac.jp, ‡ yamana@yama.info.waseda.ac.jp

Abstract Native Language Identification (NLI) is a task in which an author's native language can be identified by his/her essays written in a second language. In this work, a supervised model is built to accomplish this task based on a Chinese learner corpus. This research is based on the authors' previous work, which highlighted the significance of noisy data elimination and term weighting techniques in pursuing higher accuracy. In this work, the innovation points include: (1) it is the first work to explore skip-gram as features in NLI field. (2) By dividing the dataset into training, tuning and test subset, evaluation is more robust than that in other Chinese NLI works. By employing a hierarchical structure of linear SVM classifiers, a state-of-the-art accuracy of 75.3% is achieved by our proposed model.

Keyword Text Mining, Author Profiling, Machine Learning

1. INTRODUCTION

There is no doubt that a Chinese native speaker can judge whether a person is a native speaker or not, easily based on his/ her speaking patterns. In addition, sometimes, it is easier to guess the mother tongue of a person from his/her pronunciation than his/her speaking patterns. A large number of studies have indicated that the same principle can be employed in texts [1]. Further, since native language is one of the authorship attributes, it is possible to estimate traits of different native language speakers based on their essays written in a 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 a second language (hereafter, L2). In a nutshell, it can be considered as a classification task where machine learning methods assign labels (native languages) to different objects (essays).

While NLI has been widely used in security, such as identifying false information distributors on SNS and detecting phishing websites [2], many applications can adopt Second Language Acquisition (SLA). In general, SLA can shed light on whether a person's L2 learning is influenced by his/her L1 background. Moreover, SLA investigates the difference between L2 writings (such as error patterns) of learners with different L1 background, which can be applied to provide language teachers with instructive advice to guide students in a more sophisticated way.

The motivation of this work is based upon three aspects. Firstly, as a large number of people have shown their interest in learning Chinese, the demand of Chinese teaching services is increasing progressively. Secondly, since Chinese is an ideographic language, in which strokes and their combination need to be remembered, it is more difficult to learn than an alphabetic language, such as English [3]. Thirdly, most of the current NLI studies focus on essays written in English. Therefore, research on Chinese NLI becomes indispensable.

However, there are two major issues in the modern NLI research field. First, features exploited are restricted to either lexical or syntactic features, which are not informative enough [7] [9] [10]. Second, it is not robust enough to adopt current evaluation methods of Chinese NLI on the ground; applying test data to tune parameters will reduce the applicability of the algorithm [12].

In order to solve the two problems mentioned above, besides automatic noisy data elimination and term weighting technique proposed in the authors' previous work, we integrate (1) skip-gram feature, which contains both lexical (word level) and syntactic (grammatical level) information, and (2) a more robust evaluation method of dividing the dataset into training, tuning and test subset.

The paper is structured as follows: Section 2 describes representative prior works related with NLI; Section 3 shows the corpus used in this work; Section 4 presents our proposed methodology for NLI; Section 5 discusses the evaluation of our experiments; Section 6 gives our conclusion and forecast for future work.

2. RELATED WORKS

NLI was first proposed by Koppel et al. [4] in 2005. Combining a feature set of character n-grams, POS bigrams, function words, and errors, they trained linear support vector machines (SVMs) to classify English essays of International Corpus of Learner English (ICLE)[1] Ver.1 to achieve 80.2% accuracy. They showed that linear SVMs are well suited to an NLI task.

Even though NLI was initiated around 10 years back, notable works have been published in the recent years. Bykh and Meurers [5] defined the n-grams occurring in at least two essays as "recurring n-grams." A recurring n-gram has the advantage to reduce noisy data caused by spelling errors. After mapping essays into the vector space, they trained linear SVMs to classify English essays of the ICLE Ver.2 dataset into seven different native languages, achieving 89.7% accuracy. Our proposed method is inspired by their paper, i.e., adopting another form of recurring n-grams. That is, in order to shorten the training time by dimensionality reduction, we adopt n-grams appearing in more than ten essays instead of at least two essays.

However, the ICLE dataset used in the above works faces the problem of topic bias. Thus, the TOEFL11 dataset [6] was published during the first NLI shared task [7] in 2013. It is designed particularly for an English NLI task without topic bias compared with other corpora, such as ICLE. In this task, in order to improve the accuracy, Gebre et al. [8] proposed a new perspective to tackle an NLI task, i.e., adopting a term weighting technique before training. In their study, by scoring on the features that combine n-gram characters/words/POS tags with TF-IDF, they obtained 84.55% accuracy with the above-mentioned linear SVMs.

Bykh and Meurers [9] defined three new features according to context-free grammar-production rules (CFGR) and obtained 84.8% by using TOEFL11, higher than those reported in all works in [7]. Since CFGR is an informative feature in English NLI, we wonder whether CFGR can be used in Chinese NLI to obtain higher performance.

Malmasi and Dras [10] first developed an NLI method for a Chinese dataset. Not only did their feature set involve POS-tag-n-grams and function words but they also exploited CFGR, which can capture the structure of syntactic grammar constructions. They applied their method to Chinese Learn Corpus (CLC) [11] and achieved 70.6% accuracy.

Wang et al. [12] first highlighted the significance of noisy data elimination and term weighting technique in pursuing high NLI accuracy. They put forward a model integrating both automatic noisy data elimination and BM25 [13] [14] term weighting method. After training hierarchical linear SVM classifiers to classify Chinese essays of Jinan Chinese Learner Corpus (JCLC) [15], 77.1% accuracy was gained.

However, there are two problems to be solved. To begin with, features in existing NLI studies are restricted to either lexical or syntactic features. Apart from this, by using test data to tune parameters, the evaluation method of current Chinese NLI [12] is not robust enough to verify the applicability of the algorithm.

3. DATASET

In order to compare the performance of our proposed model with the author's previous work directly, we continue adopting the Jinan Chinese Learner Corpus (JCLC) dataset as same as [12].

JCLC contains a total of 8,739 essays written by students in an examination or as homework. We extract all essays with explicit native language metadata in the dataset. The L1 distribution in JCLC is listed in Table 1. Besides native language, metadata, such as proficiency level and gender, are included in this corpus, however, only native language metadata is taken into consideration in our research.

Table 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**

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)

[1] http://www.uclouvain.be/en-cecl-icle.html

How we should select effective essays to extract features before training, 2) what kinds of features we should choose, 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 four problems, besides a noisy data cleaning method and a term weighting technique of BM25 proposed in the authors' previous work[12], 1-skip-bi-gram feature is employed in our method.

4.1 Noisy data elimination [12]

Figure 1 shows the length distribution of essays written by students of Myanmar in JCLC. As shown in Figure 1, we can see that there are some very short and very long essays. For example, a short essay consists of only one sentence. 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.

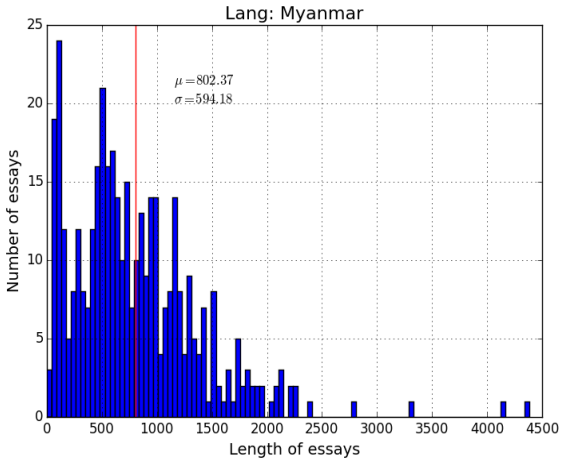

Figure 1. Essay length distribution of students from Indonesia

In order to select effective training essays by discarding both very long and very short ones, we still employ equation (1) in [12] to filter out such essays:

$$\mu_i - n_1 \times \sigma_i < essaylength < \mu_i + n_2 \times \sigma_i \qquad (1)$$

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 tuning accuracy can be obtained.

4.2 Features

4.2.1 Skip-gram

Although skip-gram is a technique commonly used in speech processing [16], it was first applied to model context in [17] and achieved high performance. In view of this, the usefulness of skip-grams is explored in our work. Figure 2 a) shows a Chinese sentence segmented with word-base. Figure 2 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 2. Example of a Chinese sentence with corresponding 1-skip-bi-gram

As is shown in Figure 2 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.

4.2.2 Feature combination

In addition to 1-skip-bi-gram, other lexical and syntactic features are adopted in our proposed method. Lexical features such as character n-gram, word n-gram, part-of-speech (POS) tag n-gram, and function words have been commonly utilized in NLI. As for syntactic features, we examine the context-free grammar production rules (CGPR), the same as what we explored in [12] .

The feature combination exploited is listed in Table 2. As is explained in Section 2, we adopt recurring n-grams, selecting n-grams that occur in more than ten essays as informative features.

Table 2. Our adopted features

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

4.3 Term weighting method of BM25 [12]

Wang et al. [12] first demonstrated that term weighting method of BM25 played a significant role in pursuing higher NLI accuracy. In this work, we continue utilizing the application of BM25, which is defined as equation (2).

$$w_{ij} = \frac{tf_{ij} \times (k_1 + 1)}{tf_{ij} + k_1 \times \left(1 - b + b \times \frac{len(d_j)}{avgdl}\right)} \times \log \frac{|D|}{df_i} \quad (2)$$

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. *avgdl* denotes the average document length and $len(d_j)$ represents the length of document j. According to Robertson and Zaragoza [13], 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.

4.4 Hierarchical classifiers [12]

Most of the works on the first NLI shared task [7] verified that the application of linear SVMs to NLI can achieve higher performance than that of the other machine learning methods. In [12], the authors put forward a new hierarchical structure of linear SVM classifiers to classify Chinese essays in JCLC dataset and yielded higher accuracy.

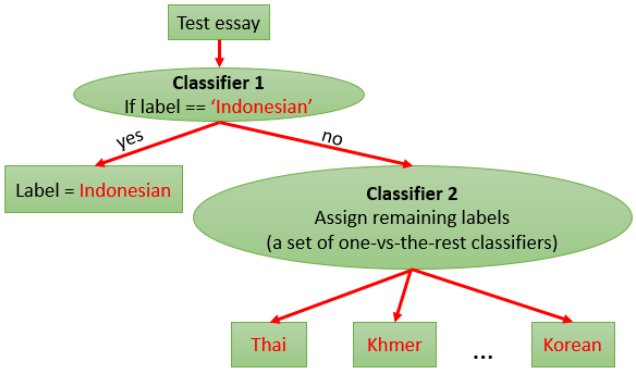

Figure 3. Structure of our hierarchical linear SVM classifiers

In the JCLC dataset, the number of essays corresponding to each native language is considerably different from each other. For this reason, we continue exploiting the application of hierarchical linear SVMs to Chinese NLI. The structure of our hierarchical linear SVM classifiers is shown in Figure 3. 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, Khmer, and 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. EXPERIMENT AND EVALUATION

5.1 Evaluation method

In the authors' previous work [12], we ran 10-fold cross validation experiment: the JCLC dataset was divided randomly into ten subsets of equal size, nine of which were used for training with varying n_1 and n_2, and the tenth was used for testing. This process was repeated ten times with each subset being held out for test exactly once. However, in our previous paper [12], test data were used for tuning parameters that results in unfairness of the evaluation.

Therefore, in order to execute fair evaluation, besides test dataset, we divide the remaining data into training dataset and tuning dataset at a ratio of 4:1 while carrying out 10-fold cross validation, as is shown in Figure 4. In this case, after tuning phase, parameters will be fixed so that they can be applied to the machine learning model for test particularly.

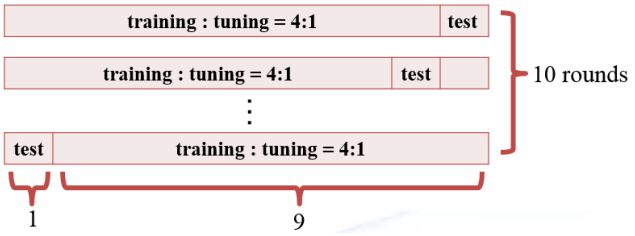

Figure 4. 10-fold cross validation in this work

5.2 Tuning parameters n_1 and n_2

In parameter tuning phase, classifiers are trained with varying pair of (n_1 , n_2) where n_1 =0, 1, 2, 3 and n_2 =0, 1, 2, 3, respectively. After training, we test the trained model with the tuning dataset to select best parameters. Table 3 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 3. 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

5.3 Test result

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 *4* 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 4. Test result of 10-fold cross validation

(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**

5.4 Analysis

Figure 5 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.

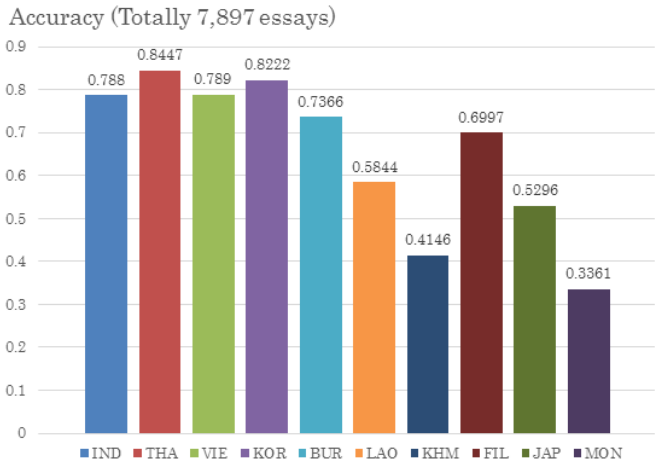

Figure 5. Accuracy of each native language

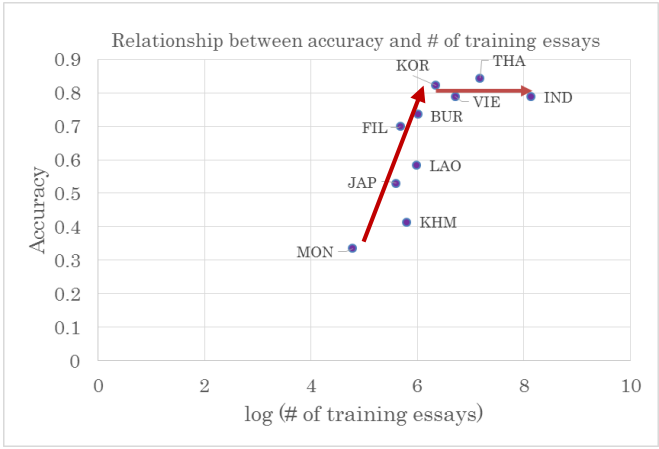

Figure 6. Relationship between accuracy and # of training essays

Figure 6 illustrates the relationship between accuracy and the number of training essays. In order to demonstrate the pattern intuitively, we calculate the logarithm of the number of training essays as x-axis. As is shown, when $x \leq 6.3$, 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 $x > 6.3$, the accuracy seems to reach a limit around 0.8 with a constant forward trend.

5.5 Comparison with the authors' previous work

Besides 70.6% accuracy in [10], a supervised model based on JCLC corpus was built in the authors' previous work [12] and 77.1% accuracy was achieved. As baselines, we implemented those two algorithms with the evaluation method proposed in this work and obtained an accuracy of 0.653 and 0.748. We compare it with that of our proposed supervised model shown in Figure 7. As is illustrated in Figure 7, our proposed method outperforms the baselines

by 0.5% and 10%.

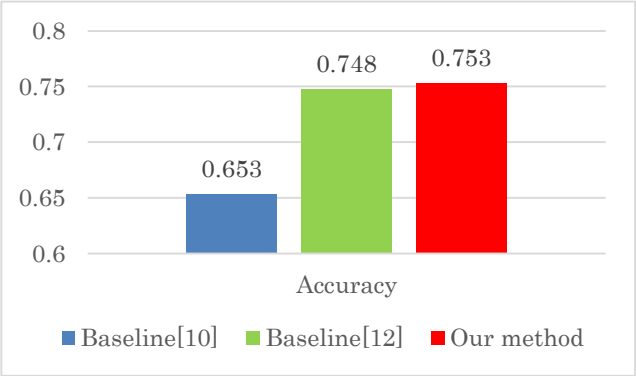

Figure 7. Accuracy compared with baseline

6. CONCLUSION AND FUTURE WORK

In conclusion, in this paper, we put forward a new Chinese NLI method based on JCLC dataset and achieved the state-of-the-art accuracy of 75.3% with a robust evaluation method, which can contribute to second language education and identification of phishing. Above all, our proposed model is cross lingual, which can also be applied to accomplish English NLI.

On a basis of our previous work, which highlighted the significance of noisy data elimination and term weighting techniques, in this work, not only is our supervised model the first work to verify skip-gram is an informative feature, but we also propose a robust evaluation method to validate the applicability of the algorithm.

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

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,” J. of Consumer Research, vol. 21 (3), pp.419-431, 1994.

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

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

[6] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill and Martin Chodorow, “TOEFL11: A Corpus of Non-Native English,” ETS RR-13-24, 2013.

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

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

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

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

[11] 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.

[12] 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.

[13] Stephen Robertson and Hugo Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” J. of Foundations and Trends in Information Retrieval, vol. 3 (4), pp. 333-389, 2009.

[14] John S. Whissell and Charles L. A. Clarke, “Improving Document Clustering Using Okapi BM25 Feature Weighting,” J. of Information Retrieval, vol. 14 (5), pp.466-487, 2011.

[15] 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.

[16] Manhung Siu and Mari Ostendorf, “Variable n-grams and extensions for conversational speech language modelling,” IEEE Trans. on Speech and Audio Processing, 8:63-75, 2000.

[17] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks, ”A Closer Look at Skip-gram Modelling,” Proc. of the 5th Int’l Conf. on Language Resources and Evaluation, pp.1222–1225, 2006.