The Spoken/Written Language Classification of English Sentences with Bilingual Information

Kuan Li[1,*], Zhongyang Xiong[1], Yufang Zhang[1], Xiaohua Liu[2], Ming Zhou[2], and Guanghua Zhang[1]

[1] College of Computer Science, Chongqing University, Chongqing, China
sloweater@163.com, {zyxiong,zhangyf}@cqu.edu.cn, Guanghua0420@gmail.com
[2] Microsoft Research Asia, Beijing, China
{xiaoliu,mingzhou}@microsoft.com

Abstract. To alleviate the problem with Chinese being poor at telling the difference between spoken and written English which is important for learning and using the language, we propose to classify English sentences with bilingual information into the two categories automatically. Based on the text categorization technology, we explore a variety of features, including words, statistics and their combinations, and find that a classification accuracy nearly 95% can be achieved in the open test through Chinese characters + sentence length + average syllable number, or other similar combinations.

Keywords: Text categorization, Sentence classification, Spoken and written language, Bilingual sentences.

1 Introduction

Back in the early twentieth century, foreign academics began the studies on spoken and written English. Spoken English (or Colloquialism) refers to the expressions employed in conversational or informal language but not in formal speech or formal writing [1]. Colloquialisms include words (such as *gonna* and *wanna*), phrases (such as *old as the hills*, *raining cats and dogs* and *dead as a doornail*) and aphorisms (such as *There's more than one way to skin a cat*) [1].

Professor Chafe indicates that fragmentation and involvement are related to the spoken language, while integration and detachment are associated with the written [2]. Due to very little time for utterance planning and deep involvement of communicators in the conversational context, the spoken sentences are simple and short, one or a few of which express a fragmental idea unit. In contrast, writers tend to integrate more information into an idea unit and detach the language from specific conversational context by using a variety of devices such as clauses, the passive voice or nominalization. [2]

[*] This work had been done while the author was visiting Microsoft Research Asia.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 370–377, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Statistics online show that in 2006 about 300 million Chinese people were learning English, over one third of which were students, and the number was growing quickly. Gui and Yang [3] point out that most learners of English as a Second Language (ESL) write as native English speakers talk. The English learners from Chinese universities show a strong colloquial tendency in their written, which isn't improved significantly by more time spent on learning [4].

Sample sentences are very useful for English learners. Today we can mine millions of English-Chinese bilingual sentence pairs from web to build a tremendous sample sentence corpus. If we can further label them with spoken or written English with a relatively high accuracy, learners may hopefully get the ability to tell between spoken and written through reading many labeled samples besides theories. It would be an excellent complement to the mainstream English teaching.

The idea above seems promising thanks to the fast development of the text categorization (TC) in the recent years. The research on TC has long focused on the document classification (DC), such as the spam detection in e-mails [5], the news classification [6][7], etc. Recently, some researchers have moved their focus onto the sentence classification (SC) and achieved some results, such as the Chinese question classification [8], the classification of sentences in e-mails [9], in legal documents [10], in the abstracts of medical literatures [11], with different class sets.

To alleviate the problem with Chinese being poor at telling the difference between spoken and written English which is important for learning and using the language, we propose to classify English-Chinese bilingual sentence pairs mined from web into the two categories, spoken and written English, automatically. We focus on exploring a variety of features to find a feature group performing adequately in our experimental environment and to help English learners in practice.

2 Problem Formulation and Method

2.1 Problem Description and Data Set

Our mission is to classify a sentence into spoken or written English. A sentence s refers to an English sentence with its bilingual counterpart, like "*Nice to meet you! / 很高兴见到你!* ". The class set is $C=\{Spoken, Written\}$. We put the mission as a classical supervised machine learning problem. Given a few labeled sentences $S=\{(s_1, c_1), (s_2, c_2), \ldots, (s_n, c_n) \mid c_{1\ldots n} \in C\}$ as a training set, a target function $f:S \rightarrow C$ [12] is learned on it, which is called the training phase. In the predicting phase, given an s as the argument, f outputs its label $c \in C$.

Due to the lack of standard data sets for our mission, we constructed the data set by our own. The training set contains about 20,000 bilingual sentence pairs in movie lines mined from web as the spoken part, and about 25,000 sample sentences (bilingual) from authorized English-Chinese dictionaries like "*A bond is a promissory note, usually issued for a specified amount. / 债券是一种期票，通常以一定数额发行。*" as the written part. We randomly sampled hundreds of bilingual pairs from

another big set automatically mined from web, and manually labeled 800 ones (400 spoken + 400 written) for the open test.

2.2 Sentence Representation

We use Vector Space Model (VSM) to represent the sentence space. If only words are put as features, a sentence can be represented as an N-dimensional vector $<w_1, w_2, \ldots, w_N>$, where N represents the number of different words in the training set, and w_k is the weight of the k^{th} word in the sentence. We follow Khoo to use 1/0 as the weight value [9] representing if the word appears in the sentence. Other features than words will be also introduced into sentence vectors, which will be discussed in Section 3.

2.3 Classification Algorithm

Support Vector Machines (SVM) solves some critical problems for machine learning, such as the small sample, nonlinear, high dimension and local minima, etc. [14] The SVM classifier does well in [9-12], and it is essentially a two-class classifier [13], suitable for our mission. Through survey, we chose LIBLINEAR [15] for our experiments. The performance of its linear SVM classifier meets our requirements.

2.4 Evaluation Metrics

We use the classification accuracy A to evaluate the overall performance of the two-class classification, as defined below:

$$A = \frac{TP_1 + TP_2}{TP_1 + FP_1 + TP_2 + FP_2} \tag{1}$$

Here TP_1, TP_2, FP_1, FP_2 represent the number of true positive on Class 1/2, false positive on Class 1/2 respectively.

For each class, we use precision P, recall R and $F1$ to evaluate the performance, as defined below (taking Class 1 as example):

$$P_1 = \frac{TP_1}{TP_1 + FP_1} \tag{2}$$

$$R_1 = \frac{TP_1}{TP_1 + FN_1} \tag{3}$$

$$F1_1 = \frac{2 \times P_1 \times R_1}{P_1 + R_1} \tag{4}$$

Here TP_1, FP_1 represent the same as above, and FN_1 represents the number of false negative on Class 1.

3 Sentence Features

3.1 English Words and Chinese Characters

Following the reported DC and SC experiments, we use English words (EW) and Chinese characters (CC) as features to encode the class information. For "*Nice to meet you! / 很高兴见到你！*", the features are:

- EW (lowercased): *nice, to, meet, you;*
- CC: *很, 高, 兴, 见, 到, 你.*

3.2 Statistic Information of Sentences

Sentence Length. According to [2], most spoken sentences are simpler and shorter than written ones, which inspires us to adopt sentence length (SL) as a feature. In practice, we put the number of the words in an English sentence as its SL.

Average Syllable Number. According to [1-2], the spoken language often occurs in a conversation, easy to speak and understand, so syllables of its most words could be less than those of big words in the written. Therefore we introduce average syllable numbers (ASN) as a feature, as defined below:

$$ASN = \frac{\sum_{k=1}^{M} sw_k}{M} \tag{5}$$

Here M represents the number of the words in an English sentence, and sw_k represents the number of the k^{th} word syllables gotten from a syllable dictionary. If the j^{th} word doesn't exist in the dictionary, we assign sw_j an approximate value, one third of the word letter number.

Flesch–Kincaid Grade Level [16]. The Flesch–Kincaid grade level (*F-K*) is designed to indicate the readability of a piece of English text. The lower its value is, the easier the text is to understand. As mentioned above, the spoken language is easier to understand than the written, which means that their *F-Ks* could be different. Therefore we introduce *F-K* as a feature. The formula to calculate *F-K* is defined below:

$$F-K = 0.39 \times \frac{Total\ Words}{Total\ Sentences} + 11.8 \times \frac{Total\ Syllables}{Total\ Words} - 15.59 \tag{6}$$

Here *Total Words, Total Sentences, Total Syllables* represent the numbers of all the words, sentences, syllables of the text respectively. For a single English sentence, *Total Words* is *SL* defined above, *Total Sentences* is 1, and *Total Syllables* divided by *Total Words* is *ASN* defined above. Therefore the formula can be transformed into the following one:

$$F-K = 0.39 \times SL + 11.8 \times ASN - 15.59 \tag{7}$$

So, *F-K* is a weighted sum of *SL* and *ASN* when used for a sentence.

4 Experiment Results and Analysis

To verify the features discussed above and to find feature combinations performing well, we conduct the following three experiments, in all of which there are close and open tests. In the close test, we adopt 10-fold cross-validation [17]. Meanwhile, for the practical application, we pay more attention to the open one, and use the evaluation metrics *P, R, F1* only in it.

4.1 Words as Features

Table 1 shows the experiment results of using English words (EW), Chinese characters (CC) and EW+CC as features. In the open test, CC performs best, while EW does far from expected. In the close test, the three kinds of features do better than in the open one, especially EW.

The results indicate that data is more consistent with each other inside the training set than with that from the testing set, which is not surprising due to the lack of manual labeling for the training set. There must be spoken sentences in the samples from dictionaries. Many spoken words were introduced into the written part when we constructed the training set. The classifier trained on it labels the spoken with the written in the open test. That is why the spoken recall is much less than the precision, and the written is in reverse.

From the perspective of Chinese characters, the training data is more consistent with the testing data, because the Chinese part of a sample in dictionaries is relatively formal no matter its English counterpart is the written or not. Therefore the classifier trained with CC performs better.

Table 1. The experiment results of words as features (%)

(C: close test, O: open test, S: spoken, W: written)

Feature(s)	A(C)	A(O)	F1(S)	R(S)	P(S)	F1(W)	R(W)	P(W)
CC	93.11	**90.75**	**90.19**	**85.00**	**96.05**	**91.25**	**96.50**	**86.55**
EW	94.39	81.25	79.67	73.50	86.98	82.60	89.00	77.06
CC+EW	**95.57**	87.38	86.37	80.00	93.84	88.24	94.75	82.57

4.2 Statistics as Features

Table 2 shows the experiment results of using sentence length (SL), average syllable number (ASN), Flesch–Kincaid Grade Level (F-K) and their combinations as features. We can see that SL or ASN doesn't perform well alone. SL can't distinguish some short written sentences from the spoken, while ASN can't distinguish long written ones if their simple words are so many to lower their ASN too much. However, they are complements to each other, so we need their combinations as features. There are two ways to combine them:

- A weighted sum of SL and ASN: F-K alone
- Vector: 2 or 3- dimension feature vector consisting of SL, ASN or F-K.

The combinations improve the performance significantly as the last five lines show in Table 2. Interestingly, the open test results of the last four lines are exactly the same, and their A(O) is the highest in this experiment, which means SL+ASN and the other three similar combinations are almost the same good in performance in our data sets.

Table 2. The experiment results of statistics as features (%)

(C: close test, O: open test, S: spoken, W: written)

Feature(s)	A(C)	A(O)	F1(S)	R(S)	P(S)	F1(W)	R(W)	P(W)
SL	82.17	78.00	80.31	89.75	72.67	75.07	66.25	86.60
ASN	67.97	81.75	79.50	70.75	90.71	83.56	92.75	76.02
F-K	79.07	88.88	88.30	84.00	**93.07**	89.39	**93.75**	85.42
SL+ASN	**86.89**	**89.88**	**89.96**	**90.75**	89.19	**89.79**	89.00	**90.59**
SL+ F-K	**86.89**	**89.88**	**89.96**	**90.75**	89.19	**89.79**	89.00	**90.59**
ASN+ F-K	86.88	**89.88**	**89.96**	**90.75**	89.19	**89.79**	89.00	**90.59**
SL+ASN+ F-K	**86.89**	**89.88**	**89.96**	**90.75**	89.19	**89.79**	89.00	**90.59**

4.3 Combinations of Words and Statistics

In this experiment, we combine CC performing best in 4.1 with the last four feature groups performing best in 4.2 to do the mission. The results in Table 3 show that their performances are the same in the open test, and more importantly, their A(O) is higher than the best results in 4.1 and 4.2 by 4~5%, which means a best classification accuracy (94.88%) in our data set can be achieved through CC+SL+ASN or other similar combinations.

Table 3. The experiment results of combinations of words and statistics (%)

(C: close test, O: open test, S: spoken, W: written)

Feature(s)	A(C)	A(O)	F1(S)	R(S)	P(S)	F1(W)	R(W)	P(W)
CC+SL+ASN	93.81	**94.88**	**94.65**	**90.75**	**98.91**	**95.08**	**99.00**	**91.45**
CC+SL+ F-K	93.82	**94.88**	**94.65**	**90.75**	**98.91**	**95.08**	**99.00**	**91.45**
CC+ASN+ F-K	93.81	**94.88**	**94.65**	**90.75**	**98.91**	**95.08**	**99.00**	**91.45**
CC+SL+ASN+ F-K	**93.83**	**94.88**	**94.65**	**90.75**	**98.91**	**95.08**	**99.00**	**91.45**

We randomly selected some sentences classified by CC+SL+ASN, manually checked them and found that most of the results are convincing and useful. For example, "*She has an elegant style. / 她具有优雅的风格。*" is labeled as the written even if it is short, which is believed reasonable due to its complete structure and no colloquialism word. Another sentence "*I'm up to my ears in work. / 我工作忙得不可开交。*" is labeled as the spoken, which is also convincing due to its "*I'm*" and "*up to my ears*".

A kind of obvious error made by our system is to classify some titles and organization names into the spoken, like "*Study on Multifunctional Teaching DPTV Platform / 多功能DPTV教学演示平台的研制*", "*State Bureau of Machine Building Industry / 国家机械工业局*", etc. We manually checked our training set and found that there are barely any titles or organization names in the written part. Meanwhile, their grammatical structures are not complete. Therefore it is difficult for the classifier to label them all correctly. However, since they are easy to recognize for English learners, this type of error doesn't affect the practical application of our research much.

5 Conclusions and Future Work

English learners in China suffer from being poor at telling the difference between spoken and written English. We propose to classify English sentences with bilingual information into the two categories automatically. The experiments show that the best classification accuracy nearly 95% can be achieved in the open test through Chinese characters + sentence length + average syllable number, or other similar combinations.

Since our training set is far from perfect, which is mentioned in 4.1, and several feature combinations achieve the same scores in 4.2 and 4.3, we are going to:

- Overcome the size and quality problems of the training set with semi-supervised learning technologies;
- Build a bigger testing set and design more experiments to find out which is the best feature combination.

Acknowledgements. We thank the anonymous reviewers for their valuable comments. We also thank Long Jiang, Shidou Jiao, Shiquan Yang and Wei Li from MSRA NLC group for their great support to our research.

References

1. Colloquialism, `http://en.wikipedia.org/wiki/Colloquialism`
2. Chafe, W.L.: Integration and Involvement in Speaking, Writing, and Oral Literature. In: Norwood, D.T. (ed.) Spoken and Written Language: Exploring Orality and Literacy, ABLEX Pub. Corp., New Jersey (1982)
3. Gui, S., Yang, H.: Chinese Learner English Corpus. Shanghai Foreign Language Education Press, Shanghai (2002)
4. Liu, X.: The Colloquial Tendency in the Written English of English Learners from Chinese Universities. Journal of Technology College Education 24(1) (2005)
5. Drucker, H., Wu, D., Vapnik, V.N.: Support Vector Machines for Spam Categorization. IEEE Transactions on Neural Network 10(5), 1048–1054 (1999)
6. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of ICML 1997: The 14th International Conference on Machine Learning, Nashville, US, pp. 412–420 (1997)

7. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
8. Jia, K., Chen, K., Fan, X., Zhang, Y.: Chinese Question Classification Based on Ensemble Learning. In: Proceedings of SNPD 2007: The 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Qingdao, China, vol. 3, pp. 342–347 (2007)
9. Khoo, A., Marom, Y., Albrecht, D.: Experiments with Sentence Classification. In: Proceedings of the 2006 Australasian Language Technology Workshop (ALTW 2006), pp. 18–25 (2006)
10. Hachey, B., Grover, C.: Sequence Modelling for Sentence Classification in a Legal Summarisation System. In: Proceedings of SAC 2005: The 2005 ACM Symposium on Applied Computing, New Mexico, US, pp. 292–296 (2005)
11. Yamamoto, Y., Takagi, T.: A Sentence Classification System for Multi Biomedical Literature Summarization. In: Proceedings of the 21st International Conference on Data Engineering, ICDE 2005 (2005)
12. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
13. Chen, X., Chen, Y., Wang, L., Li, R., Hu, Y.: Text Categorization Based on Classification Rules Tree by Frequent Patterns. Journal of Software 17(5), 1017–1025 (2006)
14. Ren, S., Fu, Y., Li, X., Zhuang, Z.: Feature Selection Based on Classes Margin. Journal of Software 19(4), 842–850 (2008)
15. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9, 1871–1874 (2008), Software available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`
16. Flesch-Kincaid Readability Test, `http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test`
17. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2001)