Identify Language Origin of Personal Names with Normalized Appearance Number of Web Pages

Jiali You[*2] *Yining Chen*[1] *Min Chu*[1] *Yong Zhao*[1] *Jinlin Wang*[2]

[1]Microsoft Research Asia, Beijing, China
[2]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]youjiali@mails.gucas.ac.cn, [1]{ynchen, minchu, yzhao}@microsoft.com, [2]wangjl@dsp.ac.cn

10.21437/Interspeech.2006-395

Abstract

Identifying the language origin of a personal name without context is interesting and useful in many areas. Morphological structure, which has long been considered as the main source of language origin information, is modeled by N-grams of letters or letter chunks. In this paper, we introduce a new information source, the appearance number of a name in web pages of different languages, for identifying its language origin. Since the distribution of web pages in various languages is not identical, and the state-of-the-art search engines can only provide the number of pages that contain the queried words, we propose a method to normalize the appearance number obtained from a search engine and use it as a new feature. When this new feature is used independently to identify language origin of names among four closely related languages (English, German, French, and Portuguese), the error rate is 26.9%, which is comparable to that of letter 4-gram features. When it is used together with the letter N-gram models, the error rate is reduced to 14.2%, which is about 43.2% error reduction, compared with the letter 4-gram based baseline model.

Index Terms: speech synthesis, letter-to-sound, language identification

1. Introduction

Identifying the language origin of a personal name without context is widely used in many areas such as speech synthesis, speech recognition, and name entity transliteration [1-6]. In this paper, we introduce a new feature, which comes from the appearance number in Word Wide Web (WWW) to represent how often a name is used in a language.

In speech synthesis and speech recognition, predicting the pronunciations of out-of-vocabulary words (the so-called letter-to-sound or LTS) is an important module. For normal words, the error rate of LTS is restricted to less than 5% for many languages. However, for proper names (most of them are personal names), the error rate will increase a lot [4]. Since many proper names are loanwords which may follow their original language rules. If we know the language origin of a proper name, it is helpful to find a correct pronunciation [1-3]. In name entity transliteration [6], names with pre-labeled language origin are needed for training the translation models.

To identify language origin without context, maximum posterior probability criterion is often adopted. With Bayesian formula, it is written as,

$$\begin{aligned} L^* &= \arg\max_l \{P(l \mid W)\} \\ &= \arg\max_l \left\{ \frac{P(W \mid l)P(l)}{P(W)} \right\} \\ &= \arg\max_l \{P(W \mid l)P(l)\} \end{aligned} \quad (1)$$

where, W is the given word, l is the possible language origin of W and L^* is the decision hypothesis. Since $P(l)$ is the prior probability of the application, the key for solving the problem is to estimate $P(W \mid l)$.

Conventionally, the morphological structure of a name is considered as the main information source of language origin. The morphological characteristics of a language are often extracted by N-gram models of letters or letter chunks from names used in the language [1-3, 5]. The likelihood for a word W with letter sequence $\{s_1, s_2, ..., s_t\}$ originating from language l, i.e. $P(W \mid l)$,.

$$\begin{aligned} P(W \mid l) &= P(s_1, s_2, ..., s_t \mid l) \\ &= P(s_1 \mid l)P(s_2 \mid s_1, l)..P(s_t \mid s_{t-1}, s_{t-2}, ..., s_1, l) \end{aligned} \quad (2)$$

If letter N-gram model is used, i.e. the appearance of a letter is assumed to depend only on the (N-1) preceding letters; equation (2) is rewritten to (3).

$$P(W \mid l) \approx P(s_1 \mid l)P(s_2 \mid s_1, l)..\prod_{i=N}^{t} P(s_i \mid s_{i-1}, ..., s_{i-N+1}, l) \quad (3)$$

The N-gram likelihood is relatively precise only when the N is close to the length of a word. However, because of the data sparse problem, models with large N cannot be estimated accurately. In our previous work [5], we proposed to combine the scores from multiple letter chunk N-grams with AdaBoost and the error rate is reduced to 22.5% which achieved about 18% error reduction. Although we have taken a better usage of the information of letter sequence in this task, the final accuracy is still lower than 80% in a four language task (English, German, French and Portuguese). The main reason is that the four languages are closely related to each other and sometimes cannot be distinguished by the word morphology itself. Therefore, other information is needed to further improve the performance. We consider how often a name is used in a

[*]This work is done when the first author visits Microsoft Research Asia as an intern.

language is an important information source of its language origin. We should estimate $P(W|l)$ from this aspect.

An ideal estimation of $P(W|l)$ is to divide appearance number of W in language l by the sum of numbers of all words used in l, as given in (4).

$$P(W|l) = \frac{C(W|l)}{\sum_i C(W_i|l)} \quad (4)$$

However, it is very difficult to achieve a reliable estimation because of the lack of large enough text corpus for many languages. In recent years, with the widely spread of internet access and the fast increase of web contents, the abundant texts in multifarious languages have been used for various language researches [7] such as spelling check, word translation disambiguation [8], and addressing data sparseness for language modeling [9]. We find that almost every name we ever met can be found from Internet. Therefore the appearance numbers of a name in web pages of different languages are a good cue for the language origin of the word.

In this paper, we propose to use such numbers obtained by online search as new features for the task. However, since the web resource is a dynamic system and the distribution of contents in different languages is not uniform, the appearance number has to be normalized in order to get stable performance. Because the normalized appearance number of a word and letter N-gram score of it are two independent information sources that describe the language origin of the name, much better results are achieved when they are used together.

We introduce the approach for calculating the normalized appearance number by online search in Section 2. The method for combining the new features with letter N-gram scores is presented in Section 3. In Section 4, the new feature is evaluated individually as well as jointly with the letter 4-gram scores. The final conclusions are given in Section 5.

2. Normalized appearance number obtained by online search

For detecting the language origin of a name with online search results, a straightforward idea is to use the raw appearance numbers of the name in all languages as the features. A name is more likely to belong to the language that has the highest number. However, the result of this naïve method is very bad. There are at least two factors that we have to consider.

2.1. Two affecting factors

First, the web resource is a dynamic system and it is changing all the time. On one hand, some studies [10] show that the known Internet is growing by more than 10 million new, static pages each day. On the other hand, it is not a surprise that if a web site with 1 million pages disappears some day. At the same time, search engines update their database automatically all the time. As the result, the appearance number of a word obtained by queries at different time is normally not a constant. It will change from time to time.

Second, the total amount of available contents in different languages is not uniform and not stable as well. Figure 1 shows the distribution of web pages in several languages provided by [11]. These numbers are obtained in year 2000 and they change with time. Since the number of English web pages is much larger than those of French and German pages, a lager appearance number in English does not necessarily mean a proportionally higher likelihood for the word originated from English. For example: "Hertzberg" is obviously a German name, however, it has a higher appearance number in English web pages than in German.

In order to solve the two problems, the prior probabilities in web of the candidate languages should be considered. However, it is hard to get exact prior probability of each language. In fact, the prior probability changes all the time. Thus, we tried to set up a reference system for queries in different languages and at different times.

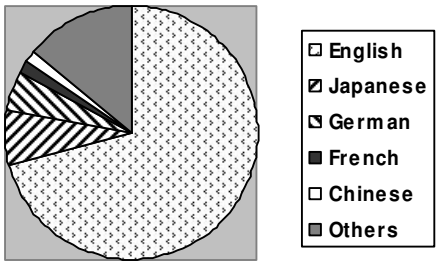

Figure 1. *The distribution of number of pages in different languages (quoted from [11]).*

2.2. Normalization of appearance number

To estimate $P(W|l)$ according to equation (4), we need to get $C(W|l)$ for all words used in all languages we consider. Yet, it is very difficult. With most current search engines, we are not able to obtain $C(W|l)$. Instead, we can only get $C(N_W|l)$, which is the number of pages the word W appears in. Since the word W may appear multiple times in a page, and a page may contains multiple words, $\frac{C(N_W|l)}{\sum_i C(N_{W_i}|l)}$ is not a good approximation of $P(W|l)$. So, $\frac{C(N_W|l)}{C(l)}$ is used instead, where $C(l)$ is the number of web pages available for each language.

Although some statistical organizations provide $C(l)$ periodically, their numbers cannot reflect the rapid changes of the internet structures. We have to get an estimation of $C(l)$ on the fly. In most languages, there are some function words, such as "the", "a", and "an" in English, "die", "das" and "der" in German, almost evenly distributed in all web pages of the related language. Therefore, the number of pages contain those words are roughly proportional to the total number of pages in the language. Then, $C(l)$ can be approximated by $C(N_{w_f}|l)$, the appearance number of such a function word, N_{w_f}. To get a more precise rate, we set a function words list for each language and search them on the fly. The largest page number is used. Then, $P(W|l)$ is estimated by equation (5)

$$P(W|l) \approx \frac{C(N_W|l)}{C(N_{w_f}|l)} \quad (5)$$

If we put equation (5) into equation (1), we get (6).

$$L^* = \arg\max_l \left\{ \frac{C(N_W \mid l)}{C(N_{w_f} \mid l)} P(l) \right\} \quad (6)$$

From this equation, we can see that if we assume page number distribution of different languages is equal or similar to the application prior distribution $P(l)$, $\frac{P(l)}{C(N_{w_f} \mid l)}$ should be a constant and the equation (6) becomes,

$$L^* = \arg\max_l \{C(N_W \mid l)\} \quad (7)$$

It is the raw appearance numbers mentioned at the beginning of this section. This means the selecting the best raw appearance numbers is only a special case of this normalization approach. In the real world, $\frac{P(l)}{C(N_{w_f} \mid l)}$ is normally not a constant. In this paper, we assume the language distribution in developing corpus is similar to the real application and use it to approximate $P(l)$.

3. Combining morphological information and Web information

In the previous section, we propose a new feature obtained by online search for identifying the language origins of personal names. The normalized appearance number is a rough approximation of $P(W \mid l)$ and it reflects how often a name is used in a language. However, a name appeared in a webpage written in one language may not belong to this language. For example, famous persons, like scientists or stars, may very often appear in all language pages. For this kind problem, analyzing morphology of the name may be much helpful.

In our previous work [5], letter or letter chunk N-gram models are used to approximate $P(W \mid l)$, which captures the morphological structure of a language. The two types of features describe the language origin of a name from different aspects. Therefore, we use them together to get a better accuracy.

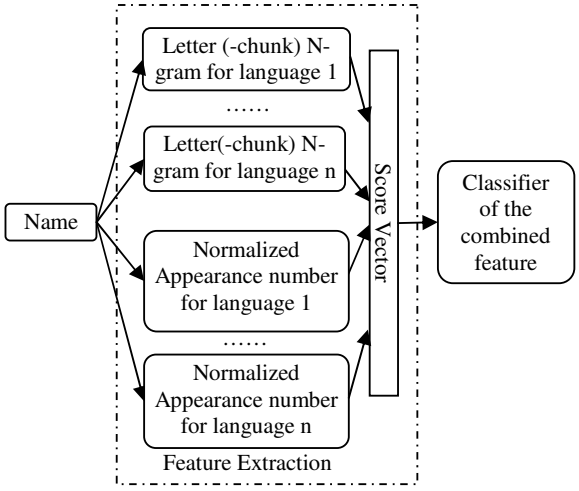

Figure 2. *The framework for combining normalized appearance number with N-gram scores.*

The combining framework is similar to that proposed in [5]. The main difference lies in that modules for calculating normalized appearance number are added and they are parallel to letter (-chunk) N-gram models, as shown in Figure 2. A given new name is scored by multiple letter (-chunk) N-gram models of all candidate languages. At the same time, the name and the representative function words are sent to a search engine to get the appearance numbers in all these languages. Then the likelihoods from N-gram models and the normalized appearance numbers form a feature vector, representing the language origin of the name. They are sent to the classifier of combined feature. In this work, AdaBoost is used for its good performance in combining weak classifiers [12,13,5].

4. Evaluation and discussion

4.1. Data

Four languages, English, German, French and Portuguese are considered in our experiments. Table 1 lists the size of the four name lexicons we have. In each lexicon, 80% items are used for training letter N-gram models. 6.7% are kept for training AdaBoost classifier. Another 6.7% are used as the developing set for tuning parameters and the remaining 6.6% is kept for testing.

Table 1. *The size of personal name lexicons*.

English	German	French	Portuguese
25,436	15,144	11,494	8,956

4.2. Evaluation of the normalization approach

To validate the normalization approach for appearance number, the normalized number is compared with the raw one. The criterion of decision is just selecting the language with the highest feature. The results are shown in Figure 3. The error rate of using the raw number directly is 37.1% and it is reduced to 26.9%, when the normalized number is used. The result of latter one is comparable with that of using letter 4-gram models, which shows that the normalization algorithm works well and it has similar performance to the letter 4-gram models.

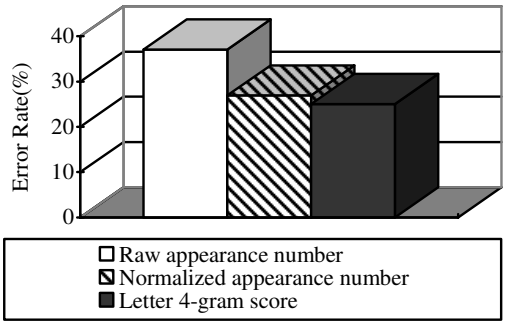

Figure 3. *The error rate using different features*

When comparing the errors of the normalized appearance number with those of the letter 4-gram model, we find that their distributions are much different. Figure 4 shows the error distribution in English. Less than 1/3 errors are overlapped.

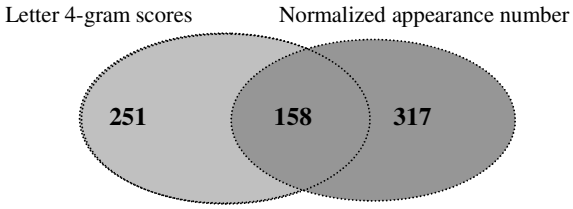

Figure 4. *The error distributions for using different features.*

4.3. Evaluation of the combined feature

When the normalized appearance numbers are used together with letter 4-gram scores, the error rate is further reduced to 14.2%. As shown in Figure 5, if it is compared with the error rate of letter 4-gram models, the combined approach results in 43.2% error reduction. If compared with error rate of normalized appearance number, the reduction is 47.2%.

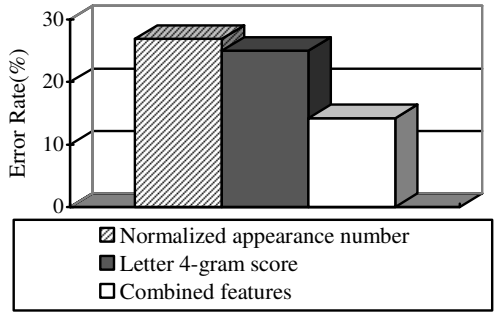

Figure 5. *Error rates of combined features and the individual features.*

By analyzing the results, we find many errors in using one type of features are corrected by the other type of features. For example, "Berghoff" is a typical German name, and it is incorrectly classified to English when the normalized appearance number is used. However, since it contains typical German morphemes "berg" and "hoff", the likelihood of German is very high when it is scored by letter 4-gram models. When the two features are combined, it is classified correctly. "Youri" is a French name, yet, often gets high likelihood of originating from English by letter 4-gram models. Since it frequently appears in French web pages, by considering the appearance number in conjunction with letter 4-grams, it is correctly classified.

5. Conclusion

Identifying the language origin of personal names is important for many applications. However, this task is difficult because names may be short and the letter sequence sometimes does not contain enough information to reveal their language origin. In this paper, we propose a new feature, the normalized appearance number, which is obtained by querying a search engine on the fly. When the new feature is used to identify the language origin of names directly, its performance is rather close to that of using morphological information. When the two types of features are combined with AdaBoost classifiers, more than 40 percent errors are removed. Such big improvements are achieved by leveraging two independent information sources.

In the future, language identification will be integrated into the framework of letter-to-sound task to help finding suitable pronunciation for personal names.

6. Acknowledgements

The author would like to thank Shiun-Zu Kuo for supporting the N-gram training and decoding tools, Lie Lu for supporting the AdaBoost tools.

7. References

[1] Llitjos, A. F. and Black, A. W., "Knowledge of language origin improves pronunciation accuracy of proper names", Eurospeech Proc., 1919-1922, 2001.

[2] Lewis, S., McGrath, K., and Reuppel, J., "Language identification and language specific letter-to-sound rules", Colorado Research in Linguistics, 17(1):1-8, 2004.

[3] Vitale, T., "An Algorithm for High Accuracy Name Pronunciation by Parametric Speech Synthesizer", Computational Linguistics, 17(1):257-276, 1991.

[4] Black, A.W., Lenzo, K., and Pagel, V., "Issues in Building General Letter to Sound Rules", 3rd ESCA Workshop on Speech Synthesis Proc., 77-80, 1998.

[5] Chen Y. N., et al., "Identifying language origin of person names with N-grams of different units", ICASSP Proc., 729-732, 2006.

[6] Huang, F., "Cluster-specific Name Transliteration", HLT-EMNLP Proc., 435-442, 2005.

[7] Kilgarriff, A. and Grefenstette, G., "Introduction to the Special Issue on the Web as Corpus", Computational Linguistics 29(3):333-348, 2003.

[8] Cheng, P. J., et al., "Translating unknown queries with web corpora for cross-language information retrieval", ACM-SIGIR Proc., 146-153, 2004.

[9] Cheng, P. J., et al., "Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora", ACL Proc., 535-542, 2004.

[10] http://www.metamend.com/internet-growth.html .

[11] Xu, J.L., "Multilingual search on the World Wide Web", NICSS-33 Proc., 2000.

[12] Freund, Y. and Schapire, R. E., "A decision-theoretic generalization of online learning and an application to boosting", Computer and System Sciences, 5(1): 119-139, 1997.

[13] Guruswami, V. and Sahai, A., "Multiclass learning, boosting, and error-correcting codes", 12th annual conference on computational learning theory Proc., 145-155, 1999.