Computational Linguistics and Chinese Language Processing
Vol. 16, No. 3-4, September/December 2011, pp. 41-60

Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction

Bor-Shen Lin* and Yi-Cong Chen*

Abstract

With the evolution of human lives and the spread of information, new things emerge quickly and new terms are created every day. Therefore, it is important for natural language processing systems to extract new words in progression with time. Due to the broad areas of applications, however, there might exist the mismatch of statistical characteristics between the training domain and the testing domain, which inevitably degrades the performance of word extraction. This paper proposes a scheme of word extraction in which histogram equalization for feature normalization is used. Through this scheme, the mismatch of the feature distributions due to different corpus sizes or changes of domain can be compensated for appropriately such that unknown word extraction becomes more reliable and applicable to novice domains.

The scheme was initially evaluated on the corpora announced in SIGHAN2. 68.43% and 71.40% F-measures for word identification, which correspond to 66.72%/32.94% and 75.99%/58.39% recall rates for IV/OOV, respectively, were achieved for the CKIP and the CUHK test sets, respectively, using four combined features with equalization. When applied to unknown word extraction for a novice domain, this scheme can identify such pronouns as “海角七號” (Cape No. 7, the name of a film), “蠟筆小新” (Crayon Shinchan, the name of a cartoon figure), “金融海嘯” (Financial Tsunami) and so on, which cannot be extracted reliably with rule-based approaches, although the approach appears not so good at identifying such terms as the names of humans, places, or organizations, for which the semantic structure is prominent. This scheme is complementary with the outcomes of two word segmentation systems, and is promising if other rule-based approaches could be further integrated.

* Department of Information Management, National Taiwan University of Science and Technology,
Tel: (886)-2-2703-1225 Fax: (886)-2-2737-6777
E-mail: bslin@cs.ntust.edu.tw; m9709104@mail.ntust.edu.tw

Keywords: Unknown Word Extraction, Word Identification, Machine Learning, Multilayer Perceptrons, Histogram Equalization.

1. Introduction

With the evolution of human lives and the accelerated spread of information, new words are created quickly as new things emerge every day. It is then necessary for natural language processing systems to identify and learn new words to progress with time. Chinese word segmentation systems, for example, typically utilize large dictionaries collected over a long period of time. No matter the size of the vocabulary for the dictionaries, it is hardly possible for them to include all of the words or phrases that have been invented so far in the extensive knowledge domains, not to mention to predict in advance new terms to appear in the future. Therefore, it is more practical for Chinese word segmentation systems to use dynamic dictionaries that can be updated quickly and frequently with the new words found in the corpora of the desired domains. Hence, unknown word extraction is actually essential for quite a few natural language processing systems. It is also useful for exploring hot or new terms for desired knowledge domains or internet communities.

The approaches to unknown word extraction can be roughly divided into two categories, rule-based approaches and statistical approaches. For rule-based approaches, semantic rules for specific types of words, such as the names of humans, places, and organizations, normally are specially designed (Sun *et al*., 1994). For statistical approaches, statistical features in corpora typically have been computed and used for the decision in the threshold test. Occurrence frequency, for example, is a widely used feature (Lu *et al.*, 2004). In such approaches, the threshold is often obtained heuristically and might depend highly on the corpus. In addition, statistical approaches and rule-based approaches can be combined. Some approaches have used statistical features obtained from the corpus and have designed rules for various types of unknown words based on these features, through which even the unknown words with low occurrence frequency can be extracted (Chen *et al*., 2002). For most of the approaches, the decision rules are obtained from the training corpus heuristically, and perhaps cannot be applied to the testing domain. Therefore, use of machine learning approaches with more general features is suggested in order to obtain the decision boundary by learning automatically. Liang, for example, proposed a tri-syllable filter for screening the word candidates and the artificial neural network with statistical features for the final decision (Liang *et al*., 2000). Nevertheless, the trained artificial neural network is not shown to be able to be applied to novice domains. Besides, Goh *et al*. made use of the character features (the POS and position) in support vector machine to extract new words (Goh *et al*., 2003).

To reduce the dependency of the word extraction scheme on the training corpus so that use in diverse or novice domains becomes possible, this paper utilizes the machine learning

approaches to combine the statistical features. Histogram equalization for statistical features was further introduced to compensate for the mismatch between the training and testing corpora that might come from the difference in corpus size or the change of the domain. It is then unnecessary to retrain the model parameters, and the extraction approach becomes more general for new domains. This scheme was first evaluated on SIGHAN2 corpora for traditional Chinese provided by Chinese Knowledge Information Processing Group (CKIP) and City University of Hong Kong (CUHK). When combing four heterogeneous statistical features, DLG, AV, Link, and PreC, and applying histogram equalization for DLG, the F-measures of 68.43% and 71.40% for within-domain CKIP corpus and cross-domain CUHK corpus, respectively, can be achieved. This scheme was finally used to explore unknown words in a novice domain of a news event. When compared with the words extracted by two word segmentation systems provided by CKIP and Institute of Computing Technology Chinese Academy of Science (ICTCAS), it was found that this approach is complementary with the other two. Such terms as "海角七號" (Cape No. 7, the name of a film), "蠟筆小新" (Crayon Shinchan, the name of a figure in a cartoon), "金融海嘯" (Financial Tsunami), and so on, with prominent statistical characteristics but less structure in semantics, can be extracted successfully by the proposed approach only. These terms are hard to identify using rule-based approaches because it is difficult to draw semantic rules from such terms. Without using semantic rules, however, this extraction approach seems less robust for extracting the names of humans, places, or organizations with prominent structure. This, however, could be overcome by integrating the proposed scheme with the rule-based approaches.

2. Statistical Features

Every sentence in a Chinese corpus contains a sequence of characters. If every combination of adjacent characters in a sentence must be considered as a word candidate, there would be huge number of word candidates where a large portion would be redundant. Therefore, every combination of adjacent characters, denoted as "character group" in this paper, needs to be screened first so the total number of word candidates can be reduced to a manageable size and the statistics could be computed. The occurrence count for each character group, *i.e.* the character n-gram, is computed and used as one of the screening criteria. Those character groups with length less than eight and with occurrence count more than or equal to five are accepted as word candidates. For each word candidate, the statistical features are computed as below.

2.1 Logarithm of Character N-Gram (*LogC*)

$$LogC(T_i) = log\left(C(T_i)\right) \quad (1)$$

T_i : the word candidate with index *i*.

$C(T_i)$: the occurrence count for the word candidate T_i.

Since words tend to appear repeatedly in the corpora, those word candidates with high occurrence count are more probable to be words. Nevertheless, there are often quite a few false alarms when occurrence count is the only decision feature.

2.2 Description Length Gain (*DLG*)

$$DLG(T_i) = L(X) - L(X[@ \rightarrow T_i]) \quad (2)$$

$$L(X) = -|X| \sum_{x \in V} p(x) \log_2 p(x)$$

X : all sentences in the corpus.

$X[@ \rightarrow T_i]$: all sentences in the corpus with T_i replaced as "@"

$L(\cdot)$: the entropy of the corpus.

$|X|$: the total number of characters in the corpus.

V: the set consisting of all characters in the corpus.

Description length gain was proposed by Kit *et al*. to measure the amount of information for every word candidate according to the degree of data compression (Kit *et al*., 1999). In Equation 2, $L(X)$ is the entropy of the corpus containing the word candidate T_i, while $L(X[@\rightarrow T_i])$ is the entropy of the corpus with T_i replaced by the token "@". Therefore, $DLG(T_i)$ indicates the entropy reduction due to the elimination of the word candidate T_i in the corpus, or equivalently the information gain of the corpus contributed by including the word candidate T_i. The more information a word candidate contributes, the higher the probability that it is a word.

2.3 Accessor Variety (*AV*)

$$AV(T_i) = min\{ L_{AV}(T_i), R_{AV}(T_i)\} \quad (3)$$

$L_{AV}(T_i)$: the number of different left-context characters for the candidate T_i

$R_{AV}(T_i)$: the number of different right-context characters for the candidate T_i

Access variety was proposed by Feng *et al*. to estimate the degree to which a character group occurs independently in the corpus (Feng *et al*., 2004). The access variety for a character group is evaluated by counting the number of different characters in its left or right context. If the access variety is high, it implies the character group is often used independently in diverse contexts and tends to be a word. On the contrary, low access variety implies that the character

group is often used together with specific characters, and thus tends to be a part of a word instead of being a word itself. Hence, the larger the access variety is, the more probable the character group is a word.

2.4 Logarithm of Total Links (*Link*)

The feature *LogC* defined in Eq. 1 considers the occurrence count of a word candidate but does not take its internal structure into account. Since the occurrence counts of partial character sequences for a word candidate (denoted as *links* here) might also provide some evidence in support of this candidate being a word, a novel feature for estimating such links is proposed as follows.

$$Link(T_i) = log\left(\sum_{k \le 1} C(S(T_i; k, l))\right) \tag{4}$$

$S(T_i;k,l)$: a partial character sequence of the word candidate T_i from position k through position l.

The word candidate "行政院長" (meaning *executive director*), for example, has the partial character sequences "行政," "行政院," "行政院長," "政院," "政院長," and "院長," in which the first three and the last one are also known words. The occurrence counts of these internal links can be accumulated, and the logarithm of the summation can be taken to obtain this feature.

2.5 Independence of Prefix Character(*PreC)*

In the Chinese language, some characters are frequently used and co-occur with other words as prefixes. The preposition "在" (meaning *at*), for example, might co-occur with the words "台北" (Taipei), "拍攝" (take a photo) or "學校" (school), and so on. Since such prefix characters are of high frequency, their combinations with other words (*e.g.* "在台北", "在拍攝" or "在學校") might also be of high frequency. This induces quite a few false alarms when only occurrence count is used for word extraction. To alleviate such problems, a novel feature is proposed here to measure the independence of the prefix character for a word candidate, which is defined as the average of the occurrence counts for all the character groups with the same prefix character.

$$\overline{C}(F) = \sum_{x \in S(F)} C(x_{1L}) \tag{5}$$

$$PreC(T_i) = \begin{cases} \dfrac{1}{|S(F)|}\overline{C}(F) & if\ |T_i| > 2 \\ C(T_i) & elsewhere \end{cases}$$

F: the prefix character of the word candidate T_i.

S(F): the set consisting of the character groups with the prefix character *F* and with length larger than two.

|*S(F)*|*:* the number of the character groups in the set *S(F)*.

x_{1L}: the partial sequence of a character group *x* after eliminating its prefix character *F*.

For the prefix character "在," the independence is computed according to the occurrence counts of those character groups whose first character is "在," such as "在台北," "在學校," and "在拍攝". If the average of these occurrence counts is high, it means this prefix character has high variety of context and should be separated from the other characters in a word candidate. In such a case, every word candidate with this prefix character is less probable to be a word. In other words, the higher the independence of the prefix character, the less probable that the candidate is a word.

2.6 Normalization

As the statistical features defined above are computed from the corpus, the dynamic range of the features for the training and the testing corpora might be different when the corpus is obtained from different domains and has a different size. Therefore, the statistical features need to be normalized before being used as the inputs of the classifier. In this paper, the following formula is utilized to normalize the features onto the range of 0 to 1.

$$F(v) = \frac{v - Min(y)}{Max(y) - Min(y)} \tag{6}$$

v: the input value of the feature.

y: the type of the feature.

Min(y): the minimum value of the feature *y*.

Max(y) : the maximum value of the feature *y*.

F(v) : the output value of the feature after normalization.

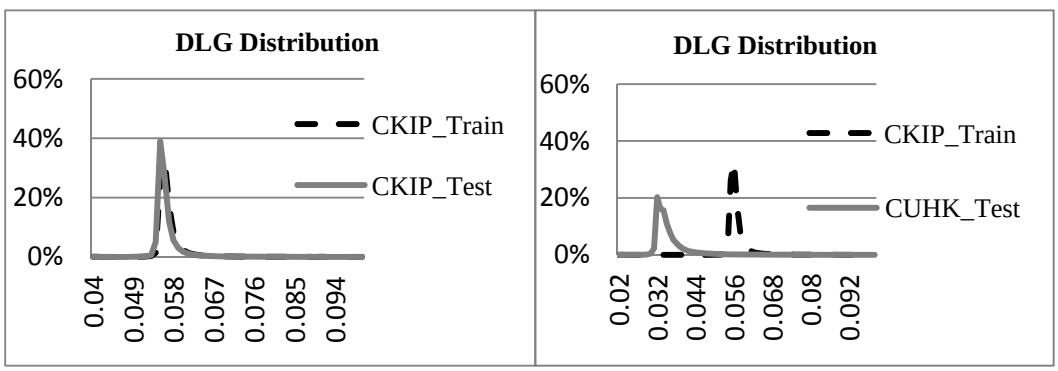
3. Word Extraction Method

3.1 Distribution of Statistical Features

Since the statistical features in this paper are obtained from the corpora, both the dynamic range and the distribution for the features might change. Although a normalization formula is introduced in Section 2.6 to deal with the problem, it is probably not sufficient for compensating for the mismatch of the feature distributions between the training and testing corpora, which often leads to performance degradation when the statistical approach is applied to new domains. In this section, we analyze how the histograms for the statistical features

might differ between various domains.

The SIGHAN2 corpora, provided by CKIP and CUHK, respectively, are used for analysis here. First, the CKIP corpus was randomly and equally divided into two sets, named as the CKIP_Train set for training and the CKIP_Test set for testing, respectively. CKIP_Test can be regarded as the within-domain test set. The corpus provided by CUHK is used as the cross-domain test set, and named as CUHK_Test set. The histograms of *DLG* feature for the CKIP_Train and CKIP_Test sets are depicted in Figure 1(a), while those for the CKIP_Train and CUHK_Test sets are shown in Figure 1(b). In Figure 1(a), it could be observed that, for corpora in the same domain with compatible sizes, the dynamic ranges of the DLG feature are very close, while the distributions still differ a little. It can also be noticed that, in Figure 1(b), the histograms for CUHK_Test set and the CKIP_Train set differ more prominently. Not only the dynamic ranges but also the shapes of the distributions differ for the two sets. If the classifier is trained with the CKIP_Train set and tested with the CUHK_Test set, the *DLG* feature appears useless without being further calibrated. More sophisticated normalization schemes will be discussed in the following section.



(a) Within-domain ***(b) Cross-domain***

Figure 1. DLG distribution for different corpora.

3.2 Advanced Normalization Schemes

When the mismatch between the training set and the testing set is significant, the classifier generally fails to classify the testing data reliably. Since we hope to use the classifier for word extraction to explore novice domains, such a problem is inevitable. To handle this problem, a typical normalization scheme, mean standard deviation weight (denoted as MSW here) was often used, as defined below.

$$X_d = M_d + \sigma_d \left(\frac{X_s - M_s}{\sigma_s} \right) \tag{7}$$

d : the destination domain.

s : the source domain.

M_d: the mean of the distribution for the destination domain.

M_s: the mean of the distribution for the source domain.

σ_d: the standard deviation of the distribution for destination domain.

σ_s: the standard deviation of the distribution for source domain.

X_s: the feature value obtained from the source domain.

X_d: the feature value for the destination domain.

Note that the source domain denotes the testing domain, while the destination domain denotes the training domain. This is because the classifier was trained with the training corpus, so the features for the testing corpus should be transformed back to the training domain to match the distribution of the training data as much as possible. MSW is a linear normalization scheme according to the distance between the feature value and the mean measured with the standard deviation. When the shapes of the distributions differ largely between the source and the destination domains, such a mismatch cannot be compensated for simply by linear shift or scaling, and MSW might not be effective enough.

Another normalization scheme, histogram equalization, denoted as HEQ here, was first introduced in image processing community and used for enhancing the contrast of an image (Hummel *et al.*, 1977; Efford 2000). As HEQ is a common technique for adjusting the statistics of the features via transformation, it can be used to compensate for the mismatch between different domains. This technique was successfully applied to such areas as speech or music processing for compensating for the mismatch of statistical features between the training and the testing domains (Ángel de la Torre *et al.*, 2005; Gallardo-Antolín *et al*. 2010). The transfer function of histogram equalization is described as follows.

$$X_d = P(X_s) \cdot (X_{MAX} - X_{MIN}) + X_{MIN} \tag{8}$$

X_s: the input feature from source domain.

X_d: the output feature of destination domain.

$P(X)$: the cumulative distribution function in the source domain.

$P_{EQ}(X)$: equalized cumulative distribution function in the destination domain.

X_{MAX}: the maximum value for the feature.

X_{MIN}: the minimum value for the feature.

Figure 2 illustrates how histogram equalization is performed. *P(X)* is the cumulative distribution function (CDF) of feature X in the source domain, as denoted by the solid curve, while $P_{EQ}(X)$ *is the equalized* cumulative distribution function in the destination domain, as denoted by the dashed line. The transfer function between the input feature X_s and the output feature X_d has to make the equality, $P(X_s) = P_{EQ}(X_d)$, hold, which leads to Equation 8. Since the heuristic cumulative distribution function of the output feature, $P_{EQ}(X_d)$, is desired to be linear, the corresponding probability density function, i.e. the histogram, needs to be uniform (equalized). Both HEQ and MSW have monotonic transfer functions, but the transfer function for HEQ could be nonlinear, and its output features in the destination domain will fall into the same dynamic range from X_{MIN} to X_{MAX} as the input features in the source domain.

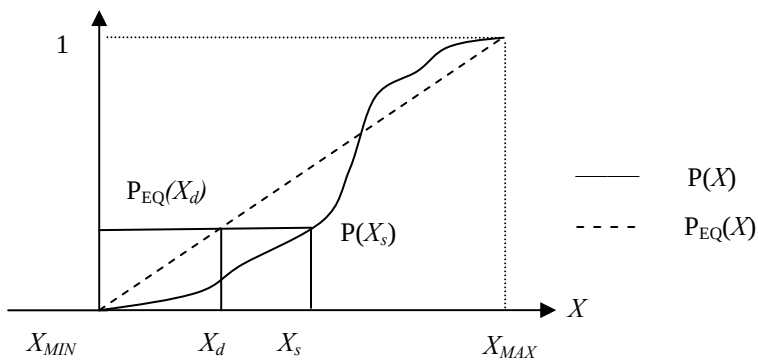


Figure 2. Histogram equalization.

When applying HEQ to word extraction, the cumulative distribution functions of the features for the training and testing domains need to be computed first, and are here denoted as $P_{TRAIN}(X)$ and $P_{TEST}(X)$, respectively. In the training phase, the features obtained from the training domain (with CDF $P_{TRAIN}(X)$) need to be transformed to the equalized domain according to Equation 8 so as to obtain the features for training. That is, the classifier is trained with the equalized features. In the classification phase, the features obtained from the testing domain (with CDF $P_{TEST}(X)$) also need to be transformed to the equalized domain, and the classifier then performs classification for equalized features.

It should be noted that, for either MSW or HEQ, such statistics as mean, standard deviation, and CDF need to be computed first so the transformation of the features can be performed accordingly. This imposes an extra limitation to batch mode for the statistical approaches since the testing corpus for computing statistics needs to be collected beforehand.

3.3 Classifier Based on Multilayer Perceptrons

The structures and rules for word formation in the Chinese language are so sophisticated that it is quite difficult to perform word identification based on a single feature. Occurrence count,

for example, is not a reliable enough feature because quite a few fragments (*e.g.* "是非常," meaning *is very*) occur frequently, but should not be regarded as words. If multiple decision features with complementary characteristics could be combined appropriately, better performance could be obtained in general. In this paper, a classifier based on multilayer perceptrons (MLP) is used for word verification. MLP is a machine learning approach based on nonlinear regression. In order to minimize the square errors, the gradient descent algorithm is applied, and the connection weights in the network are updated iteratively according to the errors propagated backwards till the estimation error converges. Figure 3 is the proposed word verification scheme for combining multiple features with an advanced normalization scheme. In this paper, the MLP classifier contains 3 layers of neurons and is trained 50000 times iteratively. The hidden layer contains five neurons, while the number of neurons for the input layer depends on how many features are used for training and testing. For every word candidate, the features *LogC*, *AV*, *Link*, *PreC*, and *DLG*, were computed with Equations 1 through 5 and normalized with Equation 6. The feature *DLG* was further processed with the advanced normalization scheme, HEQ or MSW, because of the significant difference between the histograms for the training and the testing data. In the selection module, features were combined to form the input vector $\boldsymbol{x}$ of the MLP classifier, whose output y is between 0 and 1. Through a threshold test on the output y, it can finally be decided whether the word candidate is accepted as a word. This word verification scheme, together with the screening process for word candidates as proposed in Section 2, can be used as a preprocessing stage for such NLP tasks as chunking or word segmentation to explore new terms quickly and efficiently for novice domains, such as news events or emerging communities.

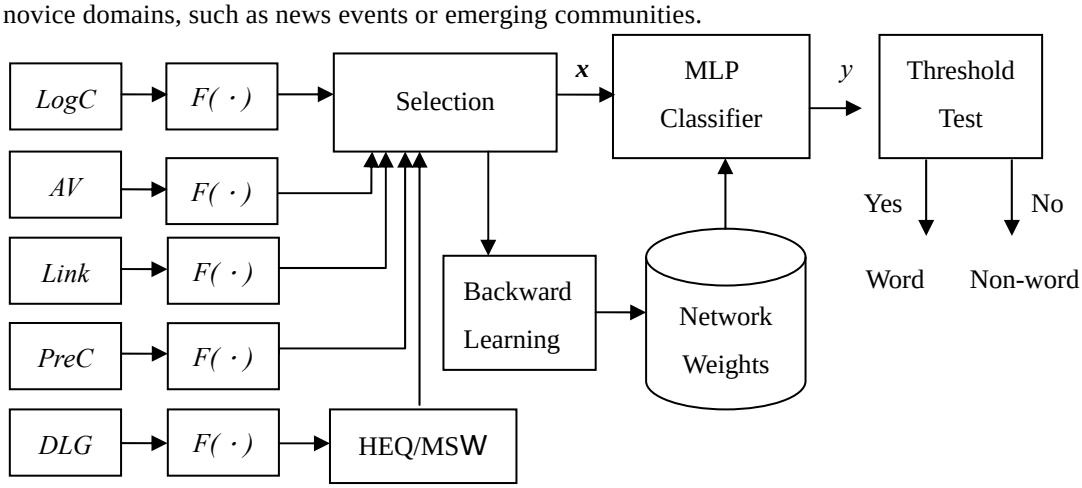


Figure 3. Word verification scheme based on MLP classifier.

4. Experiments and Analysis

In this section, the corpora CKIP_Train, CKIP_Test, and CUHK_Test, as described in Section 3.1, were used for experiments. They contain 361,691, 363,382, and 54,511 sentences and contain 222,446, 224,929, and 149,160 word candidates, respectively. In addition, the numbers of words for them are 33,429, 33,661, and 22,913, respectively, according to the sentences with word segmentation in every corpus. The segmentation results were used to label the ground truths for word verification, *i.e.*, whether a candidate can be a word or not in general without considering its usage context. That is, no information regarding to the local context of a candidate is tracked or used in word verification. The details of the experimental corpora are depicted in Table 1.

Table 1. Details of experimental corpora.

Corpus Name	Purpose	No. of Sentences	No. of Word Candidates	No. of Words
CKIP_Train	Training (Inside Test)	361,691	222,446	33,429
CKIP_Test	Within-domain Test	363,382	224,929	33,661
CUHK_Test	Cross-domain Test	54,511	149,160	22,913

First, the basic experiments of word verification were conducted using at least four types of features based on the architecture in Figure 3, but the advanced normalization scheme, MSW or HEQ, was not applied in this test. Here, CKIP_Train was used for training while CKIP_Train, CKIP_Test, and CUHK_Test were used for testing, respectively, and the results were shown in Table 2. In Table 2, ALL denotes all of the five features defined in Section 2 were used, while the others denote one of the five features was excluded. "No_LogC," for example, means that the feature *LogC* was excluded, so the other four features were used as the input features. It can be seen in Table 2 that nearly optimal F-measures of 60.09%, 60.03%, and 63.03% can be achieved for the three corpora, respectively, through combining the four features *DLG*, *AV*, *Link*, and *PreC* where *logC* is excluded (No_LogC). This implies the feature *LogC* appears to be redundant and relatively replaceable. This is partly because the occurrence count is included in the more informative feature, *Link*, defined by Eq. 4. Therefore, in later experiments, the feature *LogC* was not used. We can also find in Table 2 that the recall rates for OOV words were worse than the In-Vocabulary (IV) words for CKIP_Test and CUHK_Test sets. This is because the classifier is trained by IV words and non-words in CKIP_Train. We hope the word detector trained with more IV words can achieve more sufficient training and grasp the major stochastic characteristics of IV words such that it could detect novel terms with similar stochastic characteristics in novice domains. Of course, it is theoretically possible to build an OOV detector directly with machine learning

approaches. Nevertheless, the number of OOV words is much fewer than IV words such that the insufficient training often makes the classifier suffer from the over-fitting problem. For the same reason, the performance indices for word identification are mainly adopted instead of OOV detection in the development process.

Table 2. Verification performance by F-measure and R_{IV}/R_{OOV} for various testing corpora.

		CKIP_Train (inside test)	CKIP_Test (within-domain)	CUHK_Test (cross-domain)
F	No_LogC	60.09%	60.03%	63.03%
	No_DLG	57.11%	57.21%	62.59%
	No_AV	51.57%	51.74%	54.89%
	No_Link	48.15%	48.06%	42.62%
	No_PreC	53.39%	53.19%	56.76%
	ALL	59.74%	59.69%	63.91%
R_{IV}/R_{OOV}	No_LogC		65.75%/37.51%	74.84%/56.33%
	No_DLG		64.78%/34.99%	73.94%/55.11%
	No_AV		59.40%/41.35%	64.67%/52.22%
	No_Link		60.17%/31.97%	45.44%/41.46%
	No_PreC		72.20%/36.39%	67.60%/43.41%
	ALL		65.15%/35.42%	70.61%/56.18%

Then, the experiments were conducted with the advanced normalization scheme (MSW or HEQ) further applied individually. Figure 4 shows the verification performances with MSW (denoted as MSW), with HEQ (denoted as HEQ), and without MSW/HEQ (denoted as No_Equ), respectively. As can be seen in this figure, after applying HEQ for the feature DLG, the optimum F-measure rises significantly from 60.03% to 68.43%, but there is hardly any improvement achievable by applying MSW. The reason is probably that the dynamic ranges of *DLG* for the training and testing domains, as depicted in Figure 1(a), are almost the same, so MSW, which simply shifts and scales the features, becomes unhelpful. The shapes of the histograms in Figure 1(a), however, differ a little, so HEQ can help improve the performance through nonlinear transformation.

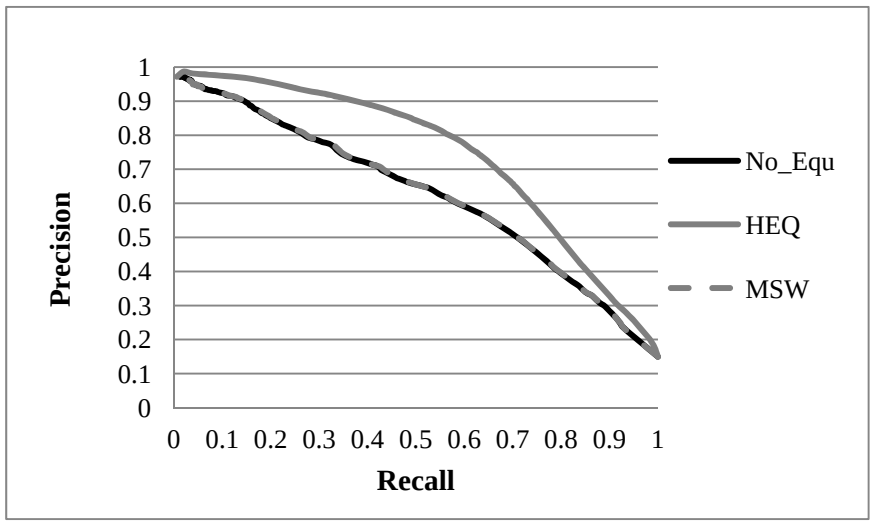


Figure 4. Performance improvements by MSW/HEQ for within-domain test.

Note here that only the performance of HEQs for *DLG* is depicted because the other features are not helpful when integrated with HEQ in our auxiliary experiments. For *Link* and *AV*, the histograms for the training and testing domains almost coincide such that it can hardly get benefits from equalization, while the extra nonlinear transformation of the feature might degrade the performance. This is similar to the robustness issue, where the robustness approaches for compensating for the mismatch between two environments usually incur the side effect of degrading the performance if the mismatch does not exist. In addition, for *PreC*, the histograms are very sparse and jerky with many zeros in bins, since many word candidates might share the independence of a prefix character. Therefore, it is not easy to model the cumulative distribution functions smoothly so HEQ can be well applied, and the results are not shown here.

Further, the same experiments were conducted for cross-domain testing data. That is, the CKIP_Train corpus was used for training while the CUHK_Test corpus was used for testing. The experimental results are shown in Figure 5. As can be observed in this figure, both MSW and HEQ can help improve the verification performance, but apparently the improvement for HEQ is more prominent. When comparing Figure 5 with Figure 4, it can be observed that the trend of MSW for the cross-domain test differs from that for the within-domain test. This is because, in Figure 1(b), the values of the cross-domain features become significantly smaller, which leads to rejections and degrades the verification performance. Such a problem can be alleviated slightly by MSW, which shifts the values, though the improvement is limited. HEQ, on the other hand, can adjust for the features of not only the dynamic range but also the shape of the distribution, so the improvement of the performance is more significant, with the F-measure increased from 63.03% to 71.40%.

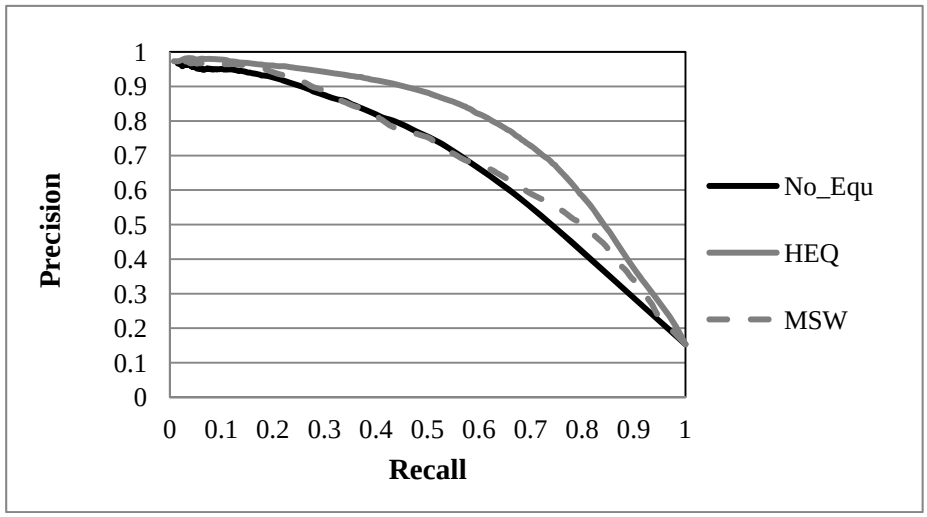


Figure 5. Performance improvements by MSW/HEQ for cross-domain test.

The above results show that HEQ can compensate for the mismatch of the *DLG* feature between the training and testing domains effectively, and the improvements achieved for cross-domain test and within-domain test are compatible. Besides, HEQ can improve the performance more significantly than MSW. This is because, the nonlinear equalization of HEQ works effectively even if the shapes of the distributions between different domains are quite different, but the equalization of MSW by shifting or scaling can work well only when the distributions have close shapes.

The above word verification scheme is further applied to the statistical classifier based on Gaussian mixture models (GMM) for comparison. That is, the MLP-based classifier in Fig. 3 is replaced with the GMM-based classifier. The F-measures for MLP and GMM classifiers with and without HEQ are shown in Table 3. As can be seen in this table, HEQ can also help to improve the verification performance for GMM-based classifier, but here MLP-based classifier achieves significantly better performance. GMM-based classifier is more sensitive to domain change, which can be compensated for appropriately through histogram equalization.

Table 3. F-measures and R_{IV}/R_{OOV} for MLP-based and GMM-based classifiers.

	CKIP_Train		CKIP_Test		CUHK_Test	
	No_HEQ	HEQ	No_HEQ	HEQ	No_HEQ	HEQ
MLP (F)	60.09%	68.70%	60.03%	68.43%	63.03%	71.40%
MLP (R_{IV}/R_{OOV})			65.75% 37.51%	66.72% 32.94%	74.84% 56.33%	75.99% 58.35%
GMM (F)	64.37%	64.65%	59.05%	64.44%	33.47%	66.34%

5. Unknown Word Extraction for Novice Domain

In this section, the word extraction scheme was applied to a novice domain. First, the corpus of the novice domain was collected from a news website in 2009 using the keyword of a news event, "八八水災" (the flood on August 8th). This corpus was used for test and denoted as UKW_Test. From 32,207 sentences in the corpus, 81,447 word candidates were extracted in accordance with the screening criteria in Section 2. Statistical features for these candidates were then computed and used as the input for the classifier trained with the corpus CKIP_Train. The system architecture is similar to that in Figure 3, but the threshold test for each candidate is not applied here. Instead, the classifier outputs between 0 and 1 for all the candidates sorted so as to obtain the 10,000 out of 81,447 candidates with the highest scores. The 10,000 candidates then were regarded as those words accepted by the classifier, while the other 71,477 candidates were regarded as rejected.

5.1 Labeling the Ground Truths

The main problem in applying word extraction in a novice domain is that there is no consensus about the definition of words for the Chinese language; therefore, it is difficult to decide the ground truth for every candidate. The CKIP corpus and CUHK corpus, for example, have different criteria for word defined by the organizations. Table 4 displays some sentences in which the definitions of some words are different. As can be observed in this table, "奶粉錢" (fee for milk) is regarded as a word in the CUHK corpus, but segmented into "奶粉"(milk) and "錢" (fee) in the CKIP corpus. Since such discrepancies exist, our strategy for labeling the ground truths here is to use two word segmentation systems to segment the test corpus, and accept a word candidate as a word if a consensus between the two systems can be reached. Those terms with discrepancy between the two systems then are inspected manually and labeled. Here, the two systems used in this paper are the web services of word segmentation provided by CKIP and ICTCAS, respectively.

Table 4. Examples of discrepancies between CKIP and HKCU.

Words		Original Sentences	
CKIP	CUHK	CKIP	CUHK
奶粉 錢	奶粉錢	**奶粉錢**也有點需要	爲了賺**奶粉錢**和教育基金
別 無 選擇	別無選擇	那自然**別無選擇**	除此**別無選擇**
混 日子	混日子	懶懶散散的**混日子**	以做肉串**混日子**
身 陷	身陷	則可能**身陷**其中無法自拔	**身陷**逃兵醜聞的韓星宋承憲
紐約 市長	紐約市長	**紐約市長**魯迪	朱利安尼當上**紐約市長**後

The UKW_Test corpus is first segmented with the two systems, respectively, and the vocabulary set of the segmented words for each system is generated. Each of the 81,447 word candidates obtained previously was checked to see if it was included in the vocabulary set. For CKIP and ICTCAS systems there were 11,290 and 10,642 candidates included in the two vocabulary sets, respectively, as depicted in Figure 6(a). This means 11,290 candidates were accepted as words by the CKIP system while 10,642 candidates were accepted by the ICTCAS system. The intersection of the two sets, containing 9,802 word candidates, were confirmed as words and used to label the basic ground truths since each of them was agreed upon by both systems. After the ground truths were labeled, the 10,000 words previously extracted with our approach can be compared with the 9,802 words in the intersection, *i.e.* the set of referenced answers. When the two sets were compared, it could be found 6,577 words were successfully identified in our approach, which corresponds to 67.09% recall rate provided the 9,802 words are used as target. The other 3,423 extracted words were accepted by our approach but not contained in the referenced answers. After manual inspection, 1,179 out of the 3,423 extracted words were labeled as acceptable answers while the others (2,224) were regarded as non-words, as shown in Figure 6(b). As a consequence, totally 7,756 (6,577 + 1,179) out of the 10,000 extracted words are either equal to the referenced answers or regarded acceptable, which corresponds to 77.56% precision rate. Such results are compatible with the F-measure for the CUHK_Test corpus obtained in Section 4.

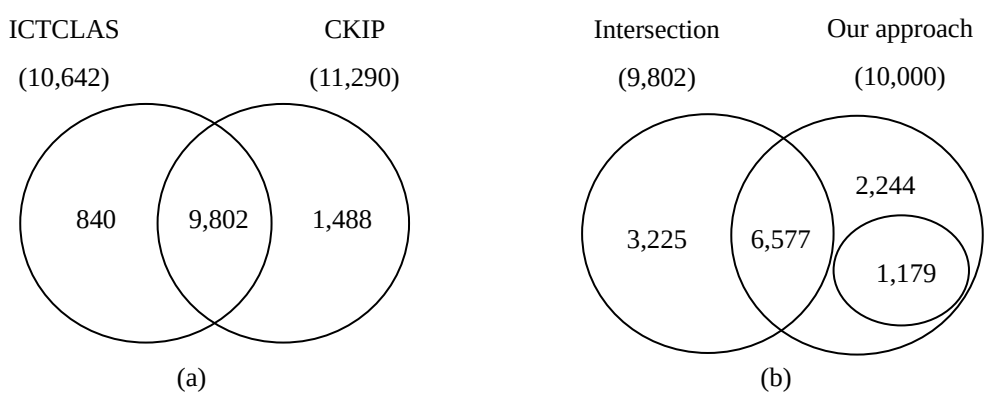


Figure 6. Ground truth labeling and test result.

5.2 Analysis of Unknown Word Extraction

Since the words extracted by our approach are different from those extracted from the two word-segmentation systems, the unknown words extracted by all of these approaches are compared in this section. Note that the unknown words are defined as those words unseen in

the training corpus CKIP_Train. Table 5 shows the number of unknown words for every approach after removing those already appearing in the training corpus from the set of words accepted by that approach. Out of the 7,756 labeled words depicted in the previous section, 1,486 unknown words were obtained by our approach, while 2,404 and 1,477 unknown words were obtained by the CKIP system and the ICTCAS system, respectively. Note that the "words" (or extracted terms) for the CKIP/ICTCAS systems are based their respective results of word segmentation instead of consentient ground truths, since, for new domains with novice terms, it is not easy to reach consensus among all systems. This does not matter, however, since our concern here is how many extra terms outside the training corpus each system can extract and how they differ. In addition, for our approach, the number of extracted words can be controlled by adjusting the verification threshold, and set as 10,000 here because this is compatible with the number of words obtained by the CKIP/ICTCAS systems and manageable for laborious manual inspection. If more unknown words are desired, this can also be accomplished by simply lowering the threshold so as to accept more words in word verification.

Table 5. Numbers of unknown words for all approaches.

Approach of unknown word extraction	No. of extracted words (accepted)	No. of unknown words
This paper	10,000	1,486
CKIP System	11,290	2,402
ICTCAS System	10,642	1,477

Figure 7 further displays how the unknown words obtained by our approach differ from those obtained by the two systems. As can be seen in this figure, there were 522 common unknown words, while 897, 316, and 530 mutually exclusive unknown words can be obtained by the CKIP system, the ICTCAS system, and our approach, respectively. This implies that there are quite a few words for which consensus among the approaches cannot be reached. Some examples in the mutually exclusive results for these approaches are shown in Table 6 for illustration. Table 6(a), for example, lists some words that were extracted by our approach but not by the other two. As can be seen in Table 6(a), some hot or novice words, such as "海角七號" (Cape No. 7, the name of a film), "蠟筆小新" (Crayon Shinchan, the name of a figure in a cartoon), "金融海嘯"(Financial Tsunami), "批踢踢"(PTT, the name of a web site), can be successfully extracted only by our approach. These words are popular pronouns whose patterns are very dynamic and do not have apparent semantic structure. Thus, it is difficult to extract these words using semantic rules only. Nevertheless, since they have prominent stochastic characteristics, they can be extracted more reliably by the machine learning

approach with HEQ using multiple statistical features including the novel features proposed in this paper. The capability of identifying such words is quite crucial for exploring novice domains.

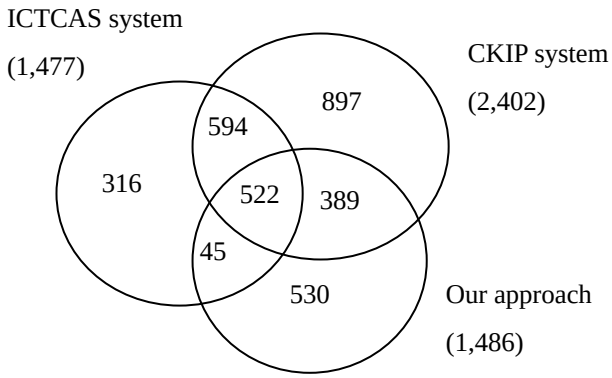


Figure 7. Comparison for differences of unknown words.

Although the proposed approach is distinguished in extracting some hot or novice words, it is more vulnerable than the other two in extracting those terms whose patterns are more static with specific structure, such as “蘇縣長” (Su, the head of the county), “經發局” (the bureau of economic development) or “光林村” (Kuang-lin village), as can be seen in Table 6. Many of these belong to the types of humans, places, organizations, numbers, and so on, which can be more readily extracted with semantic rules. Therefore, it is promising to integrate this approach with other rule-based ones such that they can complement each other.

Table 6. Mutually exclusive unknown words for three approaches.

(a) Our approach

海角 7 號	功夫灌籃
小巨蛋	批踢踢
佳暮英雄	綠豆椪
蠟筆小新	焦糖哥哥
紙教堂	龍眼乾
語音信箱	金融海嘯
那瑪夏鄉	劍湖山

(b)ICTCAS system

陳添勝	新發大橋
林政助	二手衣
夢工場	簡志忠
南迴公路	消費券
光林村	梅山鄉
總執行長	泰武村
義賣品	馬總統

(c) CKIP system

救難隊	凱達格蘭
平安米	秀姑巒溪
馬政府	監察院長
蘇縣長	正大光明
秋節禮品	毀於一旦
頂呱呱	副駕駛
張瑞賢	經發局

6. Conclusion

This paper proposes a more reliable word extraction scheme by combining multiple statistical features based on machine learning approaches. Since the formation and the structure for Chinese words are sophisticated, it is generally not robust enough to extract words simply according to single feature. This paper combined four features of the word candidates with diverse statistical characteristics to achieve the optimal performance, including the *DLG* that conveys the information for entropy gain with respect to the corpus, the *AV* for the usage context, the *Link* for the evidences of the internal structure, and the *PreC* for the independence of the prefix character. This scheme was initially verified on the CKIP corpus announced in SIGHAN2, and the performance of F-measure at 60.03% was achieved for within-domain test.

This scheme was further applied to the study of statistical mismatch problem between the training the testing domains. The difference of corpus size and the change of domain might lead to difference of dynamic range or distribution for the features, which inevitably degrades the verification performance. Histogram equalization proposed in this paper can compensate for the mismatch of DLG features effectively; thus, it is unnecessary to rebuild the training data for every desired testing domain or to worry about the incompatibility of the feature distributions due to different sizes of corpora. When this scheme of word extraction was evaluated on the within-domain test corpus provided by CKIP and cross-domain test corpus by CUHK, the F-measures can be improved from 60.03% and 63.03% to 68.43% and 71.40%, respectively, by equalization.

Finally, this scheme was used to explore a novice domain for a news event of the flood in Taiwan on Aug. 8th 2009. We proposed a strategy of labeling the ground truths for novice domains according to the consensus between two word segmentation systems. Experimental results show that this scheme can successfully identify some pronouns with prominent statistical tendency but without apparent semantic structure, which cannot be reliably identified with rule-based approaches. This scheme, however, is less robust for extracting those terms whose patterns are static with prominent semantic structure, since it is based on the statistical features instead of the semantic rules. Due to the functional complementation, it is promising to integrate this scheme with other rule-based approaches.

References

Sun, M.S., Huang, C. N., Gao, H.Y., & Fang, J. (1994). Identifying Chinese Name in Unrestricted Texts. *Journal of Chinese Language and Computing*, 4(2), 113-122.

Lu, X. Q., Zhang, L., & Hu, J. F. (2005). Statistical Substring Reduction in Linear Time. *Lecture Notes in Computer Science*, 3248, 320-327.

Zhao, H., & Kit, C. Y. (2008). An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. *Proceedings of*

The 3nd International Joint Conference on Natural Language Processing(IJCNLP), 9-16.

Chen, K. J., & Ma, W. Y. (2002). Unknown Word Extraction for Chinese Documents. *Proceedings of The 19nd International Conference on Computational Linguistics (COLING)*, 169-175.

梁婷, 葉大榮 (2000). 應用構詞法則與類神經網路於中文新詞萃取. *Proceedings of Research on Computational Linguistics Conference XIII (ROCLING)*, 21-40.

Goh, C. L., Asahara, M., & Matsumoto, Y. (2003). Chinese Unknown Word Identification Using Character-based Tagging and Chunking. *Proceedings of The 41nd Annual Meeting on Association for Computational Linguistics*, 2, 197-200.

Kit, C. Y., & Wilks, T.(1999). Unsupervised Learning of Word Boundary with Description Length Gain. *Proceedings of Workshop On Computational Natural Language Learning CoNLL*.

Feng, H., Chen, K., Deng, X., & Zheng, W. (2004) Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1), 75-93.

Ji, L., Sum, M., Lu, Q., Li, W., & Chen, Y. (2009). Chinese Terminology Extraction Using Window-Based Contextual Information. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, 62-74.

Hummel, R. (1977). Image Enhancement by Histogram Transformation. *Computer Graphics and Image Processing*, 6, 184-195.

Efford, N. (2000). *Digital Image Processing: A Practical Introduction Using Java*. Pearson Education Limited.

De La Torre, Á., Peinado, A. M., Segura, J. C., Pérez-Córdoba, J. L., Benítez, M. C., & Rubio, A. J. (2005). Histogram Equalization of Speech Representation for Robust Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3), 355-366.

Gallardo-Antolín, A., & Montero, J. M. (2010). Histogram Equalization-Based Features for Speech, Music, and Song Discrimination. *IEEE Signal Processing Letters*, 17(7), 659-662.