Fusion of Multiple Features and Supervised Learning for Chinese OOV Term Detection and POS Guessing

Yuejie Zhang, Lei Cen, Wei Wu, Cheng Jin, Xiangyang Xue
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai, China
{yjzhang, 082024072, 10210240122, jc, xyxue}@fudan.edu.cn

Abstract

In this paper, to support more precise Chinese Out-of-Vocabulary (OOV) term detection and Part-of-Speech (POS) guessing, a unified mechanism is proposed and formulated based on the fusion of multiple features and supervised learning. Besides all the traditional features, the new features for statistical information and global contexts are introduced, as well as some constraints and heuristic rules, which reveal the relationships among OOV term candidates. Our experiments on the Chinese corpora from both People's Daily and SIGHAN 2005 have achieved the consistent results, which are better than those acquired by pure rule-based or statistics-based models. From the experimental results for combining our model with Chinese monolingual retrieval on the data sets of TREC-9, it is found that the obvious improvement for the retrieval performance can also be obtained.

1 Introduction

In Information Retrieval (IR), most of users' queries and documents are generally composed of various terms, in which there are many Out-of-Vocabulary (OOV) terms like Named Entities (NEs), new words, terminologies and so on. The unidentified OOV terms may result in too many sequences of single characters in the segmented sentences and decrease the segmentation accuracy to a remarkable extent [Chen, 2003]. In addition, Part-of-Speech (POS) tagging is also an important task in query and document preprocessing of IR [Luo *et al*., 2003], but Chinese OOV term brings more difficulties for POS tagging [Lu, 2005]. The quality of OOV term detection and POS guessing directly influences the precision of querying more relevant information, and has become a crucial and challenging issue in IR.

This paper focuses on a hybrid pattern for Chinese OOV term detection, and emphasizes on the fusion of multiple features and supervised learning for OOV term acquisition. First, a series of feature representation forms are constructed to extract local, statistical and global features. After that, by using the statistical classification model and optimization processing based on the constraints and heuristic rules, the most possible label for each candidate string can be calculated to get OOV terms. Furthermore, a unified mechanism for Chinese OOV term detection and POS guessing is proposed to predict POS tags for the identified OOV terms. Our experiments on the Chinese corpora from both People's Daily and SIGHAN 2005 have achieved the consistent results, which are better than those acquired by pure rule-based or statistics-based models. From the experimental results for combining our model with Chinese monolingual retrieval on the data sets of TREC-9, it is found that the obvious improvement for the retrieval performance can also be obtained.

2 Related Research Work

In recent years, the methods for OOV term detection and POS guessing are taking a transition from rule-based to statistics-based, and then becoming a hybrid pattern combining statistical model and human knowledge [Ma *et al*., 2003].

Among various statistical models, Maximum Entropy (ME) model has many merits such as the flexibility of incorporating arbitrary features and has been proven to perform well on OOV term detection and POS guessing. However, ME model can only consider single state and then ME Markov Model (MEMM) was proposed to combine the probability of transferring between states [McCallum *et al*., 2000]. MEMM tends to get an optimal solution based on the optimization of each pair of states, which may cause the problem of label bias. As a result, Conditional Random Field (CRF) model was proposed to overcome such problems [Vail *et al*., 2007]. CRF model is widely used in many Natural Language Processing (NLP) tasks such as segmentation and parsing, and has become one of the best classifiers for NLP.

Another key factor that greatly affects the performance of OOV term detection and POS guessing is the feature selection [Smith *et al*., 2007]. Most of existing methods implement the evaluation through mining simple local features and heuristic rules. However, if only a certain document that an OOV term appears is considered, global information in the whole document set will be ignored. Meanwhile, a few designs look at special statistical features and linguistics knowledge about various OOV terms. Although there have been some attempts to combine the traditional features together, no reasonable explanation of the relationship between these features has been given. Therefore, simple local features and heuristic rules need to be expanded, and more local, statistical and global features and human heuristic knowledge should be fused to overcome defects with each other.

Little work has been done on the unified mechanism for Chinese OOV term detection and POS guessing [Nakagawa *et al*., 2006]. Goh *et al*. [2006] proposed a method for detecting Chinese unknown words and guessing their POS tags using contextual and internal component features with ME model, in which only local and no global contexts were used and just first and last character features were adopted. Such a unified mechanism could increase the parser coverage.

To support more precise Chinese OOV term detection and POS guessing, a unified hybrid pattern is formulated and trained to predict OOV terms and their POS tags based on multiple features and supervised learning, which combines statistical model with human heuristic knowledge and tries to gain the efficiency of statistical model (i.e., CRF model) and retain the quality of experts' knowledge. The representation forms of local, statistical and global feature are constructed under the consideration of the complex characteristics of Chinese OOV term. Some constraints are also introduced to reveal the relationships among OOV terms, and some rules are added to make optimization and filter wrongly identified OOV terms. All these ways can improve the precision of OOV term detection and POS guessing dramatically.

3 Feature Representation

Local Feature (LF) is constructed based on neighboring tokens and the token itself. There are two types of contextual information to be considered when extracting LFs, namely internal lexical and external contextual information.

(1#) **Component Information** – Aims to investigate special component information in OOV terms. If a candidate string contains some special characters like '·', '/' and '%', etc., the corresponding features are set as 1.

(2#) **Preceding and Succeeding Adjacent Word (PAW and SAW) Information** – Aims to utilize words proximate to the candidate string. *PAW* and *SAW* can be extracted separately as well as together. Given a character sequence "记者陈贻宁报道[*The reporter Chen Yining reported*]", the candidate string is "陈贻宁[*Chen Yining*]". Thus three features can be extracted for the class of Person Name (PRN), i.e., (PAW_{-1}=记者[*reporter*], *PRN*), (SAW_{+1}=报道[*report*], *PRN*) and (PAW_{-1}=记者[*reporter*] and SAW_{+1}=报道[*report*], *PRN*).

(3#) **First and Last Character (FC and LC) Information** – Aims to take a scrutiny of inner information of OOV term. Consider the example in (2#), ("陈[*Chen*]", *PRN*) and ("宁[*ning*]", *PRN*) can be extracted as *FC* and *LC* respectively.

(4#) **Lexicon Candidate Information** – To overcome the problem of data sparseness, the corresponding features are set as 1 for all the terms in the lexicon.

(5#) **OOV Information** – Aims to consider whether the candidate string is a potential OOV term. Strings that do not exist in Chinese Basic Dictionary are regarded as candidate strings, and then the corresponding features are set as 1.

(6#) **NE Dictionary Information** – Aims to utilize the related dictionary information for three important classes of OOV terms, i.e., Person Name (PRN), Location Name (LCN) and Organization Name (OGN), in order to overcome the lack of training data. If a candidate string exists in the NE Dictionary, then its corresponding feature is set as 1.

(7#) **Affix Information** – Aims to extract affix information contained in OOV terms, in order to provide more contextual information for detection and POS guessing. Hence the affix lists for NE and new term are constructed respectively.

(8#) **POS Information of PAW and SAW** – Aims to utilize the POS information of words proximate to the candidate string. Similar to (2#), the POS tags of *PAW* and *SAW* can be extracted separately as well as together.

Statistical Feature (SF) is the corpus-based feature and built based on the statistical information from the training corpus. SFs are used to explore the statistical measure for the internal cohesion degree inside a candidate string to judge the possibility of this candidate string being an OOV term.

(1#) **In-Term Probability (ITP)** – This is a characteristic of Chinese morpheme and refers to the probability that a morpheme appears in terms. It aims to evaluate how a morpheme is used to form a term, and is defined as:

$$ITP(c) = \frac{N(InTerm(c))}{N(c) + N(InTerm(c))} \qquad (1)$$

where $N(InTerm(c))$ is the number of occurrences of the morpheme c in terms of the corpus and $N(c)$ is the number of occurrences of c as an independent term. If the product of *ITP* values for a candidate string is larger, this string will be more likely to be an OOV term with a specific POS tag.

(2#) **Morphological Productivity (MP)** – This is a reliability measure value proposed by Baayen [1989]. It is related to the Good-Turing estimate and defined as $MP(c)=n_1(c)/N(c)$, where $n_1(c)$ is the number of types of the construction, such as the number of unique token types with the morpheme "们[*men* (Pinyin)]" in their last positions; $N(c)$ is the total number of tokens of a particular construction in the corpus, such as the total number of tokens with "们[*men* (Pinyin)]" in their last positions. A larger *MP* value indicates a higher probability that the component parts of a sequence of single characters appear to be an OOV term with a specific POS tag.

(3#) **Frequency (Freq)** – An important characteristic for OOV term with a specific POS tag is the repeatability. An OOV term usually appears more than once in a document or in more than one document, such as "世博会[*World EXPO*]", "禽流感[*bird flu*]" and "苏丹红[*tonyred*]". Assume a token appears *n* times in a corpus, its *Freq* value can be gotten by dividing *n* by the total number of tokens in the corpus.

(4#) **POS Transition Probability** – As an OOV term is usually a common noun, verb or adjective, and the POS tagging is context-sensitive, the credibility values for the OOV term and its POS tag could be evaluated based on the existing POS sequences. For the current candidate string W_0 with the context window-size 2, the POS transition probability $P(W_{-1}, W_{+1})$ is used as another statistical contextual feature, in which W_{-1} and W_{+1} denote the adjacent words before and after the candidate string respectively.

Global Feature (GF) is extracted from other occurrences of the same or similar tokens in the whole corpus. The common case is that the OOV terms in the previous parts of documents often occur with the same or similar forms in the latter parts. The contextual information from the same and other documents may play an important role in determining and tagging the final OOV term. Consider the following three sentences:

Sentence (1) – "来自**北京大学[*Peking University*]**、**清华大学[*Tsinghua University*]**等 200 所院校现场接受考生的咨询。"
Sentence (2) – "许多内地优秀学生舍弃**北京大学[*Peking University*]**与**清华大学[*Tsinghua University*]**而转投香港高校。"
Sentence (3) – "在如火如荼的高考招生季节，人们开始替**北大[*Peking University*]**、**清华[*Tsinghua*]**担忧。"

These sentences are from the same document and ranked by their positions. The OOV terms in (1) are easy to be identified and tagged. Those in (2) are also not difficult to process if the result of (1) is utilized. Those in (3) seem to be hard to process since both internal and external features provide limited useful information. However, if the contextual information in the same document ((1) and (2)) can be used, the problem is easily solved. Therefore, GFs need to be constructed to make full use of such global information.

(1#) **Other Occurrences with the Same Form** – Aims to provide information of the identified and tagged OOV terms with the same form in the previous parts of documents. **Dynamic Word List (DWL)** is constructed to store such terms. This feature is set as 1 when a candidate string appears in DWL. Thinking of the previous example, "北京大学[*Peking University*]" and "清华大学[*Tsinghua University*]" are stored in DWL after they are identified and tagged in (1), which can help the processing of candidate strings in (2).

(2#) **Other Occurrences for Affix** – Aims to provide the affix information of the identified and tagged OOV terms. For example, the suffix of "清华大学[*Tsinghua University*]" supplies the helpful contextual information for "清华[*Tsinghua*]" in (3). This feature indicates whether a candidate string can share the affix information of the terms in DWL.

(3#) **Abbreviation Form** – Aims to provide the abbreviation information of the identified and tagged OOV terms. In (3), as "北大[*Peking University*]" is the abbreviation form of "北京大学[*Peking University*]", such information contributes to the detection and tagging for "北大[*Peking University*]".

4 Chinese OOV Term Detection

4.1 Reliability Evaluation and Optimization

The strings, formed by consecutive single characters or the specific string combination operation for the segmentation result, become the OOV term candidates. Based on the contexts of each candidate, the reliability for each OOV term class is computed and the class with the highest value is chosen as the detection result[1]. Given a candidate string "贝尔格莱德[*Belgrade*]", its context x=(W_{-1}, "首都[*capital*]"), a possible class y="*LCN*", then the matching features are searched in the feature base. By using the feature weight λ_j and the normalized factor $Z(x)$ generated by the statistical model training, each conditional probability is computed as:

$$p(y\,|\,x)=\frac{1}{Z(x)}\prod_{j=1}^{k}\lambda_j^{\,f_j(x,y)} \qquad (2)$$

$p(y|x)$ becomes the original reliability value of the OOV term class y for the candidate string x under the current contexts.

To further optimize the initial reliability, it's necessary to apply some heuristic human knowledge in **Rule Base (RB)**. If the context of a candidate string matches the rules in RB, the reliability value would be augmented. Such rules in RB have the larger weights than the features in the feature base.

***Rule* 1**: If $W_{-1} \in$ {numeral}, then W_0 is not a PRN.

***Rule* 2**: If W_{-1}='、', then the OOV term class of W_0 is the same as that of W_{-2}.

According to *Rule* 1, in "一项荣誉[*an honor*]", the candidate string "项荣誉[*Xiang Rongyu*]" is not likely to be identified as a PRN. In Chinese, the punctuation '、' indicates a coordinate relationship between two terms, thus an OOV term following '、' must have the same OOV term class as the adjacent term before '、'. Given "副主席张思卿[*Zhang Siqing*]、白立忱[*Bai Lichen*]、…", "张思卿[*Zhang Siqing*]" could be easily identified as a PRN, while "白立忱[*Bai Lichen*]" could not. It is because the latter lacks strong contextual information for neither its internal lexical nor adjacent features. This frequent problem can be solved by *Rule* 2.

4.2 Constraints for Candidate Selection

Although the selected candidate strings are most likely OOV terms, the search space is too large if all the possible sequences are enumerated as candidates. To determine candidates more precisely, two constraints are introduced to make optimal selection. Unlike the features, the constraints suggest the relationships among candidates rather than their own inner information. When candidates have some relationships, the constraints can boost the overall precision for detection.

Constraint 1. Mutually Exclusive Terms (Cons1) Based on our observation, more than one candidate string may be extracted from a sentence. Considering the segmentation result "他[*he*]/ 不[*doesn't*]/ 希望[*want*]/ 做[*to be*]/ 房[*house*]/ 奴[*slave*]/ 。", "做房[*make house*]", "房奴[*house slave*]" and "做房奴[*to be a house slave*]" are extracted as the candidate strings, but only "房奴[*house slave*]" is true. Such strings are called *Mutually Exclusive Terms* and Cons1 is used to solve this problem. The assumption is that the string with the highest reliability is regarded as a candidate.

Constraint 2. A Complement to Mutually Exclusive Terms (Cons2) This constraint is an exception to solve the problem of mutually exclusive terms. Considering "世[*shi* (Pinyin)]/ 博[*bo* (Pinyin)]/ 会[*hui* (Pinyin)]/", "世博[*World EXPO*]", "博会[*bohui* (Pinyin)]" and "世博会[*World EXPO*]" are extracted as the candidate strings. The *ITP*, *MP* and *Freq* values of the former two strings are higher than those of the last one. Every feature excluding contextual information may prefer "世博[*World EXPO*]" or "博会[*bohui* (Pinyin)]" as an OOV term rather than "世博会[*World EXPO*]". Therefore a hypothesis is proposed as Cons2. If the difference between the frequency values of two mutually exclusive terms is smaller than a preset threshold γ (empirically set as 2), the longer one is preferred to be a candidate.

4.3 Rules for Post-Filtering

To filter the identified OOV terms with errors, the characteristics of Chinese morpheme can be utilized through adding some expert knowledge. Morpheme is the minimum semantic unit to form a word. Some morphemes are often used

[1] In this paper, we divide the OOV term into five classes for detection, i.e., PRN, LCN, OGN, new term and other category.

alone and viewed as simple words, while others belonging to a certain POS class may have no practical conceptual meaning. Thus some additional wordlists are constructed.

List 1 is a list of morphemes that are not used productively to form multi-morphemic words, including monosyllabic empty words which have been fully grammaticalized, configuration morphemes, monosyllabic adverbs and pronouns, such as "的[*de* (Pinyin)]", "很[*very*]", "是[*be*]", "这[*this*]" and "我[*I*]". **List 2** is a list of morphemes with weak ability to form words, including monosyllabic empty words which cannot be fully grammaticalized, such as "比[*than*]", "用[*use*]" and "像[*like*]". **List 3** is a list of specific morphemes, including monosyllabic prepositions, conjunctions and particles. All the lists above are considered as **Filtering Rule Set (FRS)**, including 53 morphemes. The results from the statistical module will be further filtered by using these lists.

5 POS Guessing for Chinese OOV Term

A straightforward formulation of POS guessing for Chinese OOV term without definite class is to learn a mapping function *f*: *OOV_term->pos*, where *OOV_term* denotes the identified OOV terms to be labeled and *pos* is a finite set of POS tags. This is a multi-category classification problem, but has some practical difficulties. The training samples are ($term_1$, pos_1), …, ($term_i$, pos_i), …, ($term_n$, pos_n), $term_i \in OOV\_term$, $pos_i \in pos$. However, there are 47 POS tags (according to "*Specification for Corpus Processing at Peking University*"), which may make the classifier unlearnable from the limited samples. Moreover, the POS of an OOV term should be also used in some contextual information, which contradicts with the fact that *f* is a function of only $term_i$ but not pos_i.

Instead of predicting the POS of an OOV term, a new formulation is suggested to predict how likely an OOV term matches a candidate POS tag. It has the mapping function *g*: *OOV_term*pos->R*, where *OOV_term*pos* is a set of pairs <$term_i$, pos_i> and $R \in [-1, +1]$ is a real value indicating the association degree between $term_i$ and pos_i, with 1 denoting match and -1 mismatch. The training sample ($term_1$, pos_1), …, ($term_i$, pos_i), …, ($term_n$, pos_n) need to be restructured to fit into *g*. Each one is rewritten as a series of triples (<$term_i$, pos_{i1}>, r_{i1}), …, (<$term_i$, pos_{ij}>, r_{ij}), …, (<$term_i$, pos_{in}>, r_{in}).

$$r_{ij} = \begin{cases} -1 & pos_{ij} \neq pos_i \\ +1 & pos_{ij} = pos_i \end{cases} \qquad (3)$$

where $\{pos_{ij}\}$ is the POS tag set, and r_{ij} is a credibility which indicates whether pos_{ij} is the correct POS of $term_i$ or not.

By transforming a multi-category classification into a binary classification problem, the statistical modeling difficulties can be solved. The mapping function *g* can be reliably learned from the sufficient training data, and all kinds of features can be easily represented and utilized.

6 Experiment and Analysis

6.1 Data Set and Evaluation Metrics

The training corpus contains the data from People's Daily of the first half of 1998, and is labeled with POS tags according to Chinese Text POS Tag Set provided by Peking University (PKU). One test set is randomly selected from the raw texts of People's Daily of the second half of 1998, and another one from the PKU corpus in SIGHAN 2005. They are two balanced test sets covering several different domains, such as politics, economy, entertainment and sports.

Four parameters are used in the performance evaluation. The precision for OOV term detection (P_{OOV}) aims at measuring the ability of rejecting the incorrect OOV terms. It is calculated as the percentage of the correctly identified OOV terms in all the identified OOV terms. The recall for OOV term detection (R_{OOV}) aims at measuring the ability of detecting the correct OOV terms. It is calculated as the percentage of the correctly identified OOV terms in all the OOV terms contained in the test set. The F-measure for OOV term detection (F_{OOV}) is a weighted combination of P_{OOV} and R_{OOV}.

$$F_{OOV} = \frac{(\beta^2 + 1) \times P_{OOV} \times R_{OOV}}{(\beta \times P_{OOV}) + R_{OOV}} \qquad (4)$$

where β is the relative weight and usually set as 1. The precision for POS guessing (P_{POS}) aims at measuring the ability of assigning the correct POS tags for the identified OOV terms. It is defined as the percentage of the correctly tagged OOV terms in all the correctly identified OOV terms.

6.2 Experiments on Detection and POS Guessing

The evaluation about the whole performance is based on six patterns: (1) *Base+LF*–based on the statistical model and LFs, the baseline performance can be obtained; (2) *Base+LF+SF*–based on the baseline model, SFs are integrated together; (3) *Base+LF+SF+GF*–through integrating the baseline model and SFs, GFs are introduced; (4) *Base+LF+SF+GF+Cons*–by fusing the statistical model and multiple features, two constraints are added; (5) *Base+LF+SF+GF+Cons+RB*–aims to explore the effect of heuristic rules; (6) *Base+LF+SF+GF+Cons+RB+FRS*–aims to investigate the effect of FRS on filtering the identified OOV terms. (1)-(3) are used in both OOV term detection and POS guessing, and (4)-(6) are only used in OOV term detection. For the statistical classification model, CRF model is selected for its good integration of various constraints and better compatibility to OOV term detection and POS guessing. The experimental results on two test sets are shown in Table 1.

Corpus	**Pattern (CRF-based)**	**OOV Term Category**	P_{OOV}	R_{OOV}	F_{OOV}	P_{POS}
People's Daily	***Base+LF***	*OOV Term*	0.5570	0.7393	0.6353	0.7089
	Base+LF+SF	*OOV Term*	0.5920	0.7825	0.6740	0.7976
	Base+LF+SF+GF	*OOV Term*	0.5991	0.8028	0.6862	**0.9092**
	Base+LF+SF+GF+Cons	*OOV Term*	0.6293	0.7777	0.6957	-
	Base+LF+SF+GF+Cons+RB	*OOV Term*	0.6498	0.7634	0.7020	-
	Base+LF+SF+GF+ Cons+RB+FRS	*NE*	0.8955	0.8647	0.8798	-
		New Term	0.6546	0.7408	0.6950	-
		OOV Term	**0.6691**	**0.7589**	**0.7112**	-
***SIGHAN* 2005**	***Base+LF+SF+GF***	*OOV Term*	-	-	-	0.8406
	Base+LF+SF+GF+ Cons+RB+FRS	*OOV Term*	0.6476	0.7493	0.6947	-

Table 1. Results for CRF-based hybrid model.

It can be seen from Table 1 that in comparison with the baseline model, the performance for OOV term detection could be improved and promoted to a great degree by adding statistical features, global features and all kinds of human heuristic knowledge in turn. Especially for two main classes of Chinese OOV term, NE and new term, the experimental results exhibit the better detection performance. It can also be observed that the fusion mechanism of multiple features and supervised learning is very beneficial to the performance of POS guessing for the correctly identified OOV terms.

6.3 Experiments on Other Statistical Model

To make comparison between different statistical models, the ME-based hybrid model is also implemented on the same data sets. The experimental results are shown in Table 2.

Corpus	Pattern (ME-based)	OOV Term Category	P_{OOV}	R_{OOV}	F_{OOV}	P_{POS}
People's Daily	***Base+LF***	*OOV Term*	0.5174	0.6168	0.5627	0.6757
	Base+LF+SF	*OOV Term*	0.5533	0.6547	0.5997	0.7378
	Base+LF+SF+GF	*OOV Term*	0.5648	0.6945	0.6230	**0.8443**
	Base+LF+SF+GF+Cons	*OOV Term*	0.5885	0.6826	0.6321	-
	Base+LF+SF+GF+Cons+RB	*OOV Term*	0.6059	0.6944	0.6471	-
	Base+LF+SF+GF+ Cons+RB+FRS	*NE*	0.8447	0.8670	0.8557	-
		New Term	0.6093	0.6902	0.6472	-
		OOV Term	**0.6135**	**0.7181**	**0.6617**	-
***SIGHAN* 2005**	*Base+LF+SF+GF*	*OOV Term*	-	-	-	0.7995
	Base+LF+SF+GF+ Cons+RB+FRS	*OOV Term*	0.6067	0.6744	0.6388	-

Table 2. Results for ME-based hybrid model.

It can be viewed from Table 2 that the ME-based hybrid model has the similar case to the CRF-based one. Through the comparison between these two different models, it can be found that CRF model is superior to ME model and exactly more suitable for solving the problem of sequence labeling for Chinese OOV term detection and POS guessing.

6.4 Experiments on Parameter Setting

To show the effect of different training set size, we divide the training data into 10 pieces and compare the performance rising speeds for *Base+LF+SF+GF+Cons+RB+FRS*. The experimental results are shown in Figure 1. It can be observed from Figure 1 that the performance of our model rises with the increasing size of training data, but the rising speed becomes slower with more and more training data. Thus we can conclude that the better performance could be obtained by using an appropriate amount of training data.

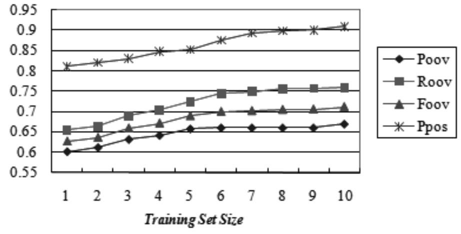

Figure 1. Results for training set size.

For the further filtering of features from the candidate feature base, we maintain the features with the frequency not lower than a preset threshold and discard others. As shown in Figure 2, when the threshold is set as 5, the steep learning curves and better performance can be acquired.

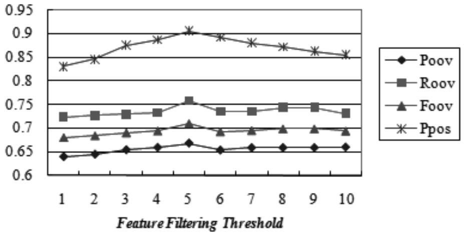

Figure 2. Results for feature filtering threshold.

6.5 Experiments on Chinese Monolingual IR

To explore the usefulness of our OOV term detection and POS guessing model for Chinese monolingual IR, four runs are carried out on the Chinese topic set (25 topics) and corpus (127,938 documents) from TREC-9 based on our own Lucene-based Chinese monolingual IR system. (1) *C-C_LongIR*1–using *long query* (terms in both title and description fields); (2) *C-C_LongIR*2–using *long query* and our unified mechanism for query and document preprocessing; (3) *C-C_ShortIR*1–using *short query* (only terms in the title field); (4) *C-C_ShortIR*2–using *short query* and our unified mechanism. The Precision-Recall curves and Median Average Precision (MAP) values are shown in Figure 3.

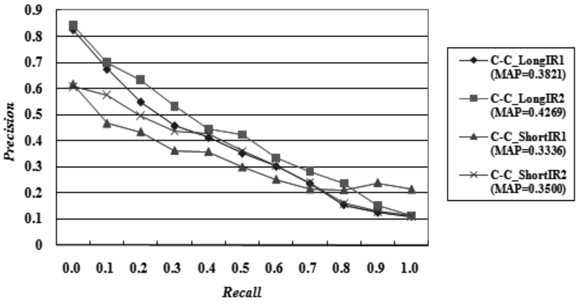

Figure 3. Results for Chinese monolingual IR.

It can be seen from Figure 3 that the best run is *C-C_LongIR*2 and its results exceed those by another run *C-C_LongIR*1 based on long query. By adopting our unified mechanism for OOV term detection and POS guessing, the Chinese monolingual IR for long query has gained the significant improvement on the whole retrieval performance. Compared with the traditional query and document preprocessing, such a combination manner is exactly a better way for extracting more precise query terms, document indexing terms and their related POS information. Additionally, through comparing the results for the other two runs *C-C_ShortIR*1 and *C-C_ShortIR*2 based on short query, it can be further confirmed that our unified mechanism is also able to support Chinese IR for short query effectively.

6.6 Comparison and Analysis

From all the experimental results above, it can be seen that information from a sentence is insufficient to detect and tag an OOV term correctly. The statistical information and global contexts from the whole document or corpus are available. The high performance for OOV term detection and POS guessing can be achieved through combining multiple features. Meanwhile, human knowledge can not only reduce the search space, but also increase the detection performance significantly. The same conclusions on different test sets show the consistence of our model on different data sources.

On the other hand, the method of Goh *et al*. [2006] is the most analogous with our approach and implemented by using the same data as ours. The experimental results are listed in Table 3, which reflect the difference of power between these two methods. Compared with Goh *et al*.'s method, ours can not only improve P_{OOV}, R_{OOV} and F_{OOV} to 66.91%, 75.89% and 71.12% respectively, but also improve P_{POS} to 90.92%.

Corpus	OOV Term Category	Method	P_{OOV}	R_{OOV}	F_{OOV}	P_{POS}
People's Daily	*OOV Term*	*Goh et al.'s* (2006)	0.5678	0.6449	0.6039	0.7800
		Our Unified Mechanism	**0.6691**	**0.7589**	**0.7112**	**0.9092**

Table 3. Results for Goh *et al.*'s [2006] method.

Through the analysis for the identified Chinese OOV terms and their guessed POS tags with failure, it can be found that there are some typical errors shown as follows. (1) Conflicts between OOV terms and existing words – For example, "李[*Li*]/ 洪亮[*Hongliang*/*orotund*]/" and "王[*Wang*]/ 大方[*Dafang*/*generous*]/". This type of error is because the detection process may be triggered by OOV character strings, but not by existing words. One solution is to apply the detection process to all the possible character strings in a certain window-size. (2) No separation between consecutive OOV terms – For example, "李学军董芳忠[*Li Xuejun Dong Fangzhong*]". This type of error might be avoided by considering more information about the particular characteristics for Chinese word building and the general length for Chinese OOV term. (3) Global features may reduce precision – If an OOV term was falsely identified or tagged for the first time and then stored in DWL, it may trigger the chain false effect in the latter classification processing. (4) Fuzzy bound of OOV term – Current means used to choose candidate strings may cause the problem that a few unreasonable candidate strings are too long with many characteristics of OOV term. For example, "美国将会[*USA will*]" has the prefix of "美国[*USA*]" and the suffix of "会[*committee*]", which are the characteristics of OGN. Splitting such term into several parts and training them separately might solve this problem. (5) Lack of available features – In some contexts, either local features or global features are limited. For example, "博卡萨[*Bokassa*]/ 上台后[*came into power*]/" almost provides no specific information that can be used to detect such an OOV term. Actually this is the most stubborn problem in OOV term detection. (6) Uncommon OOV terms with low frequency – For example, "宝应拾屯[*Baoyingshi Village*]" and "尖草坪区[*Jianchaoping District*]". Such terms cannot be correctly identified as OOV terms. Now there are still no good ideas to solve this difficult problem. (7) Uncommon OOV terms with long length – For example, "辩证唯物论者[*dialectic materialist*]" and "五个一工程[*Five Best Works*]". Such OOV terms with the longer length cannot be identified correctly either. Usually the average length of Chinese term is about 2.4 characters, thus only the strings with the length less than 4 characters are considered as candidate strings in OOV term detection. Therefore, the length of a candidate string should be expanded to a more suitable range. (8) Uncommon OOV terms with abbreviation form – For some abbreviated OOV terms, such as "党政军[*party*, *government and army*]" and "燃眉之急[*extremely urgency*]", it is very difficult to adopt appropriate features to guess their correct POS tags. Generally the average length for abbreviated OOV terms is more than 3 characters. However, there are still some abbreviated OOV terms with only 2 characters, such as "中共[*Communist Party of China*]" and "非典[*SARS*]". For sure, there are no good solutions to tag such terms correctly.

7 Conclusions

In this paper, a supervised-learning-based unified mechanism is established for Chinese OOV term detection and POS guessing. The new features for statistical information and global contexts are proposed, as well as some constraints and heuristic rules, which reveal the relationships among OOV term candidates. At present, OOV term detection and POS guessing are viewed as the post processing of word segmentation and OOV term detection respectively. Next OOV term detection will be integrated into word segmentor, so that more efficient statistical information and contextual features can be investigated. Meanwhile, combining OOV term detection and POS guessing to make them benefit from each other will also be our research focus in the future.

Acknowledgments

This work is supported by Shanghai Natural Science Fund (No. 09ZR1403000), 973 Program of China (No. 2010CB327906), National Natural Science Fund of China (No. 60873178), Shanghai Leading Academic Discipline Project (No. B114) and Shanghai Municipal R&D Foundation (No. 08dz1500109).

References

[Baayen, 1989] R.H. Baayen. A Corpus-based Approach to Morphological Productivity. *Statistical Analysis and Psycholinguistic Interpretation*, Ph.D. Thesis, 1989.

[Chen *et al*., 2003] A.T. Chen. Chinese Word Segmentation Using Minimal Linguistic Knowledge. In *Proc. of SIGHAN 2003*, pages 148-151, 2003.

[Goh *et al*., 2006] C.L. Goh, M. Asahara, and Y. Matsumoto. Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing. *Journal of Chinese Language and Computing*, 16(4):185-206, 2006.

[Lu, 2005] X.F. Lu. Hybrid Methods for POS Guessing of Chinese Unknown Words. In *Proc. of ACL 2005 Student Research Workshop*, pages 1-6, 2005.

[Luo *et al*., 2003] S.F. Luo and M.S. Sun. Chinese Word Extraction based on the Internal Associative Strength of Character Strings. *Journal of Chinese Information Processing*, 17(1):9-14, 2003.

[Ma *et al*., 2003] W.Y. Ma and K.J. Chen. A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. In *Proc. of SIGHAN 2003*, pages 31-38, 2003.

[McCallum *et al*., 2000] A. McCallum, D. Freitag, and F. Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proc. of ICML 2000*, pages 591-598, 2000.

[Nakagawa *et al*., 2006] T. Nakagawa and Y. Matsumoto. Guessing Parts-of-Speech of Unknown Words Using Global Information. In *Proc. of ACL 2006*, pages 705-712, 2006.

[Smith *et al*., 2007] N.A. Smith, D.L. Vail, and J.D. Lafferty. Computationally Efficient M-Estimation of Log-Linear Structure Models. In *Proc. of ACL 2007*, pages 752-759, 2007.

[Vail *et al*., 2007] D.L. Vail, M.M. Veloso, and J.D. Lafferty. Conditional Random Fields for Activity Recognition. In *Proc. of AAMAS 2007*, pages 1331-1338, 2007.