International Journal of Knowledge and Language Processing
Volume 3, Number 2, **2012**

www.ijklp.org
KLP International ©2012 ISSN 2191-2734
pp. 30-**38**

Domain-adaptive Chinese Word Segmentation

Liu Lezhong
Ludwig-Maximilian University Munich
Munich, German
lezhong@cis.uni-muenchen.de

Received March 2012; revised April 2012

ABSTRACT. *This paper presents a new approach to the central problem of domain-adaptive word segmentation in Chinese, i.e. the recognition of new words. Three heurisitic models are proposed here to this end. We will also present below an evaluation of our system and compare it to the performance of previous approaches found in the literature.*

Keywords: Domain Adaptive, Chinese Word Segmentation

1. **Introduction.** Developing a domain-adapting Chinese word segmenter is of great importance primarily because a segmenter is needed which can adapt automatically to processing texts from different domains. We applied the ICLCLASS segmenter to a chemical text, our test set, and evaluated its performance according to the PKU standard. Although the result reported in the First Sighan Bakeoff was 95.3% for the China Daily test set, the method only gave an F-Measure of 0.725% when applied to our domain-specific test set [11]. The poor result is due to the fact that the segmenter has not acquired a large amount of linguistic knowledge about the chemical text (mostly in the form of new words), as such knowledge was not derivable from the training corpus, in this case the People's Daily Corpus.

2. **Previous Work.** In [7], the authors mention the two problems of adapting the segmenter to different domains and to different Chinese word segmentation standards. Their approach performs quite well in terms of being able to adapt to the four different Chinese word segmentation standards, but their solution to domain-adaptation is debatable, as they do not explain how their approach can acquire new knowledge from domain-specific data given an annotated corpus taken from a more general domain. Furthermore, their test corpora do not contain domain-specific texts, hence the performance of their system remains unknown. After reviewing recent developments in Chinese word segmentation, such as in [2, 3, 6, 1, 9, 8, 10, 4 and [12], we established that most of the state of the art methods focus on acquiring ”positive” knowledge from a hand-annotated training corpus. Here the ”positive” knowledge refers to the Word Formation Rule as analyzed in Table 1 in section 3. This positive knowledge is then used to extract similar knowledge from an unseen corpus, for example measures such as the independent word probability (IWP) and the word formation

analogy (WFA)[13] are calculated. The IWP derived from the general corpus must be modified significantly for it to be used in a domain-specific corpus, such as a chemical text, because word formation, word combination and the compound formation rule differ between the two corpora, as shown in Table 1. A domain-adaptive segmenter should be able to handle this divergence, especially as it is unrealistic to expect a large number of chemical corpora to be manually tagged so that they can be used as training corpora. Furthermore the range of domains is limitless - one sentence or one paper can be considered a small domain.

3. **Insight.** Before we ask how we can acquire new knowledge from an unseen domain-specified corpus given an annotated general corpus, the following two questions must be answered: With respect to word segmentation, what new knowledge is contained in a domain-specific hand-annotated corpus compared to that in a general hand-annotated corpus (assuming their segmentation standard is the same)? What knowledge remains unchanged? After annotating the chemical text, we make the following assumption in Table 1.

The single-character-word distribution in Table 1 refers to the probability of a single-character-word occurring adjacent to a non single-character-word. The reason we assume the distribution of single-character-words remains unchanged is that most high frequency single-character-words in Chinese are prepositions, pronouns and auxiliary words.

Domain-specific and Unknown Word

As we have analyzed, what is required is to be able to tell the difference between domain and non-domain. One single article or sentence is a domain. Domains do not exist since the difference one domain and another domain is essentially the difference between a sentence and another sentence. The reason for needing to discuss the definition of a domain-specific corpus, is that we have developed a hand-annotated reference corpus, and we hope that we can learn the language knowledge from it, in order to apply this knowledge to a raw corpus. In this process we compare the developed reference corpus to the raw corpus intuitively. Those raw corpuses that are similar to the reference corpus are classified as non-domain, while those which are sufficiently different are classified as domain corpuses. This boundary between domain and non-domain is purely subjective. More precisely, sufficiently different means that there exist many new linguistical entities. These entities can be compounds, nominal phrases, arguments; all these entities are limitless, more particular, *to Chinese language they are unknown words*. If we use the positive knowledge to extract these entities, then no matter how good the language model is, it will be not sufficient.

Therefore we would like to propose a definition of *domain-specifity* based on this very property, i.e. on the coverage of L with respect to the terminology in T.

TABLE 1. Linguistic hypothesis.

From general corpus to domain-specific corpus	Linguistic hypothesis
Vocabulary	partly Changed
Segment Standard	Unchanged
Word Formation Rule	partly changed
Word Combination Rule	partly changed
Compound Formation Rule	partly changed
Single-character-word Distribution	Unchanged

Definition 1. *Let T be a manually segmented text, L be a lexicon. T is* domain-specific *(with respect to L) if the percentage of segments in T that are not in L is 7% or higher.*

The 7% boundary was obtained empirically, by studying the unknown segments rate of general newspaper text and comparing it to that of subject specific texts, e.g. chemical or medical scientific abstracts (each time using the same, general-purpose lexicon which had been created by extracting all multi-character sequences from the China Daily hand-annotated test corpus).

4. **System description.** Given an electronic text in Chinese, i.e. a string $t = t_1 \ldots t_n$ where each $t_i \in$ S and S is the alphabet of all Chinese letters, a segmentation algorithm is expected to provide a segmentation $s(t) = (s_1, \ldots, s_m)$ $(s_1 < s_2 < \ldots < s_m)$ where each s_j is the starting position of a new segment, i.e.

$$T(t) := \{(t_{s_j} \ldots t_{s_{(j+1)}-1}) : 1 \leq j < m\}$$

is the set of all word segments in t. Later we will also use the notion of *single character sequence*, by which we mean a series of consecutive segments of length 1, more precisely, a maximal subsequence $s_{j1}, \ldots, s_{jk}$ such that $s_{j1} + 1 = s_{j2}$, $s_{j2} + 1 = s_{j3}$ and so forth, and $k \geq 2$. Typically, when a segmentation algorithm does not recognize a certain multi-character word, it splits it into a single-character sequence, as though it consisted of several single character words. Thus single character sequences (although they sometimes represent the correct segmentation) are generally indicators of words unknown to the algorithm.

System Overview Our system firstly applies a maximum matching algorithm to tokenize the raw corpus. It assumes that all resulting single character tokens are unknown word candidates. Our system then uses three heuristic models to determine if these are in fact new words. The new words found are subsequently added to the original lexicon to produce an augmented lexicon. The maximum matching algorithm is then re-applied to the raw corpus using the augmented lexicon. This last step resolves any ambiguity arising from when two augmented lexicon entries overlap.

The input data provided to the algorithm consists of three files:
(i) A plain text file *T* to be segmented;
(ii) A reference corpus *R* of which the segment boundaries have been annotated by humans
(iii) A lexicon *L* of Chinese words (i.e. a list of n-grams of characters from S, each of which may be a word (but needs not be, depending on the context).

As we will see later, the algorithm described below works particularly well when *T* contains text specific to one or several domains, while *R* is a general-language corpus, such as the manually annotated China Daily Corpus (...), and *L* is also largely domain-independent. *L* will be used in the first step of the algorithm to create a preliminary segmentation into candidate words according to the lexicon; hence it is specifically the fact that *L* has not been optimized or enhanced to cover the domain-specific terminology of *T*, which characterizes best the typical situation when our algorithm should be used. And that is a situation which in fact occurs quite frequently, as building domain-specific lexical (and training corpora) is a costly and labor-intensive activity which people generally try to avoid.

If no manually segmented sample of *T* is available, so the segments unknown to the lexicon cannot be counted, applying a maximum matching algorithm to the text and counting the single character sequences is a good estimator of the domain-specifity as well:

Definition 2. *Let T be an unsegmented text, L be a lexicon. T is* domain-specific *(with respect to L) if the percentage of single character sequences left in T after applying a maximum matching algorithm to it, using L as its dictionary, is 15% or higher.*

5. **Proposed Method.** We suggest a method that consists roughly of three steps:
(1) Apply a maximum matching algorithm (MMA) to *T*, using a general-purpose lexicon *L* as its dictionary.
(2) Improve the lexicon based on the results of the first step, esp. by comparing single character sequences(SCS) left in it, to similar character sequences extracted from *R*. The improved lexicon is called *L'*.
(3) Re-apply MMA, using *L'*.

Thus, we basically transform *L* into a domain-adapted lexicon *L'* by adding some of the SCS' from the result of (1). The decision of whether or not to include an SCS, or a part of it, in the lexicon is formalized as a function

$$f : \Sigma^* \to \mathbb{N}$$

. where a return value of "0" indicates rejection, while any positive value *x* suggests that the segment-final part starting at position *x* should be included into the lexicon. Hence, if $c = (c_1, \ldots, c_k) \in \sum^*$ is a candidate SCS, and if $f(c) = x; n < k$,

$$c^{(x)} := (c_x, \ldots, c_k)$$

is accepted as a new entry of *L'*.

We consider three types of such functions: the "pure fragment filter" (PFF), the "iterative subfragment filter" (ISF), and the "n-gram filter" (NF). Each one is applied in turn, i.e. we use PFF, next (upon a positive result) we use ISF, then NF. Note that ISF and NF are only used if the previous filters returned $x > 0$ (otherwise the candidate is immediately accepted as an entry of *L'*).

5.1. **Pure fragment filter.** We suggest a method that consists roughly of three steps:

Given an SCS $c = (c_1, \ldots, c_k) \in \sum^*$ as input, the PFF looks it up in the hand-annotated corpus $R = (r_1, \ldots, r_N)$ and computes two sets:

$$\mathrm{Pos}(c) := \{(p,q) : (r_p, \ldots, r_q) = c, \text{ annotated as one continuous word in } R\}$$
$$\mathrm{Neg}(c) := \{(p,q) : (r_p, \ldots, r_q) = c, \text{ annotated as an SCS in } R\}$$

In other words, these are the sets of positive and negative examples for c as a candidate entry of L'. A general model of the PFF filter may be described as

$$\mathrm{PFF}_\gamma(c) := \begin{cases} 0 & \text{if } \frac{|\mathrm{Neg}(c)|}{0.001+|\mathrm{Pos}(c)|} > \gamma \\ 1 & \text{otherwise.} \end{cases}$$

for some $\gamma \in \mathbb{R}$. The simplest version of it is probably PFF_0, i.e. when a single negative example suffices to reject the candidate, and this is indeed the version that we used in our experiments. Thus, compared to other lexical decision filters proposed in the literature, and looking at its mathematical complexity, PFF_0 is extremely simple and can be summarized as "accept a word candidate unless there is one or more examples in R where it is split into single characters". Therefore, if this filter is superior to previously suggested methods, then this is because of the mere fact that negative examples are taken into account, rather than because of the superiority of the statistical model being used.

5.2. **Iterative subfragment filter**. When a candidate segment $c(x) = (c_x, \ldots, c_k)$ is wrongly accepted by PFF_0, i.e. if there are no negative examples of $c(x)$ in R, but it is not a word, one reason might be that $c(x)$ is so long that by coincidence its single character segments do not occur sequentially in this particular order anywhere in R, although they are separate words. Hence, if $(k-x+1) > 2$, we remove the first character c_x and apply PFF_0 and (iteratively) ISF to the remaining sequence $c(x+1)$. Thus ISF is formally defined as

$$\mathrm{ISF}(c) = \begin{cases} 1 & \text{if } \mathrm{PFF}_0(c) = 1 \\ 1 + \mathrm{ISF}(c^{(1)}) & \text{if } |c| > 2 \text{ and } \mathrm{ISF}(c^{(1)}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

6. **Evaluation.** The test sets included two test sets from the First Sighan bakeoff and one annotated domainspecific corpus dealing with chemistry. The chemical corpus was acquired from the company Alibaba and is over 50 MB in size. Due to time constraints, we only annotated 474k of it. Senior students of chemistry firstly annotate the text. After this, two Chinese linguistics students corrected the chemistry students' annotations according to the PKU segmentation standard. In the final stage, two computational linguistics students checked the annotated corpus again. As the percentage of unknown words in the chemcial corpus was 25.9%, it qualifies as a domain-specific corpus as per Definition 1. We ran our segmentation system and the ICTCLAS segmenter over the chemical corpus use the same word list mentioned above taken from the People's Daily newswire (Jan - Jun 1998).

The PK(Beijing University) test set was evaluated as an open test using the word list from the People's Daily newswire(Jan - Jun 1998); we used the word list from the training set as close test.

TABLE 2. Our system compare with ICTCLAS segmenter.

	Our system	ICTCLAS
Test Set	Chemical Text	Chemical Text
Size	474k	474k
OOV Rate	0.259	0.259
OOV Recall Rate	0.644	0.433
IV Recall Rate	0.843	0.927
Precision	0.799	0.653
Recall	0.792	0.799
F Measure	0.795	0.719
Test Set	AS(close)	AS(close)
Size	40k	40k
OOV Rate	0.022	0.022
OOV Recall Rate	0.160	0.178
IV Recall Rate	0.957	0.970
Precision	0.930	0.924
Recall	0.940	0.953
F Measure	0.935	0.938
Test Set	PK(open)	PK(open)
Size	56k	56k
OOV Rate	0.047	0.069
OOV Recall Rate	0.697	0.743
IV Recall Rate	0.962	0.980
Precision	0.952	0.957
Recall	0.951	0.963
F Measure	0.951	0.959
Test Set	PK(close)	PK(close)
Size	56k	56k
OOV Rate	0.069	0.069
OOV Recall Rate	0.592	0.724
IV Recall Rate	0.971	0.979
Precision	0.940	0.930
Recall	0.952	0.962
F Measure	0.940	0.951

The AS(Academia Sinica) test set was tested as a close test using the word list from the training set. An open test was not conducted on the AS due to the different segmentation standards between the PK and AS. We also constructed the Fragment Filter Model based on the training set from the AS corpus.

In the final step, we made use of the same PERL evaluation program used in the first Sighan bakeoff. The results of our evaluation are shown in Table 2 and Table 3.

The results above show that our segmentation system achieved almost the same level of performance compared to the ICTCLAS segmenter from ICT in two of test sets from the First Sighan bakeoff. Furthermore, our system performed exceptionally well when applied to the large domain-specific test set, particularly in terms of the OOV recall rate. In Table 2, the OOV recall rate of our segmenter is 0.644, whereas it is only 0.433 for the ICTCLAS segmenter. This result is particular pleasing as our test set, a real corpus, is 474k in size, and is much larger than the test sets used in the First Sighan bakeoff.

TABLE 3. Our system with different Models for Chemical text.

	FFM	FFM+PSFM	All Models
Size	474k	474k	474k
OOV Rate	0.259	0.259	0.259
OOV Recall Rate	0.633	0.645	0.644
IV Recall Rate	0.852	0.831	0.843
Precision	0.780	0.799	0.799
Recall	0.796	0.783	0.792
F Measure	0.788	0.791	0.795

Table 3 lists the results of the different models used. Since the OOV Rate is 25.9%, we can see exactly which model makes the most useful Model. We found the Fragment Filter model(FFM) and Pre-Suffix filter Model(PSFM) to be the most useful. The ”Chinese version” of the Levenshtein distance based on the naive Bayes law(CLD) actually decreases the OOV Recall Rate. This shows that new knowledge is acquired from the new domain-specific corpus mostly in using the FFM and PSFM.

7. **Discussion.** In our work, our strategy is not to search for the unknown words directly, but to determine the distribution and the environment of unknown word candidates. For example, in the FFM, the negative examples are all instances of fragments from natural text. In the PSFM, we concentrate on the single characters which occur next to unknown word candidates with high probability. The test set, i.e. the raw corpus, must therefore be at least 10K in length, not merely a single single sentence or short paragraph. The disadvantage of having a minimum test set size is that a short sentence can not be processed by our system. Nevertheless, our system has the advantage of being able to process the world web web easily.

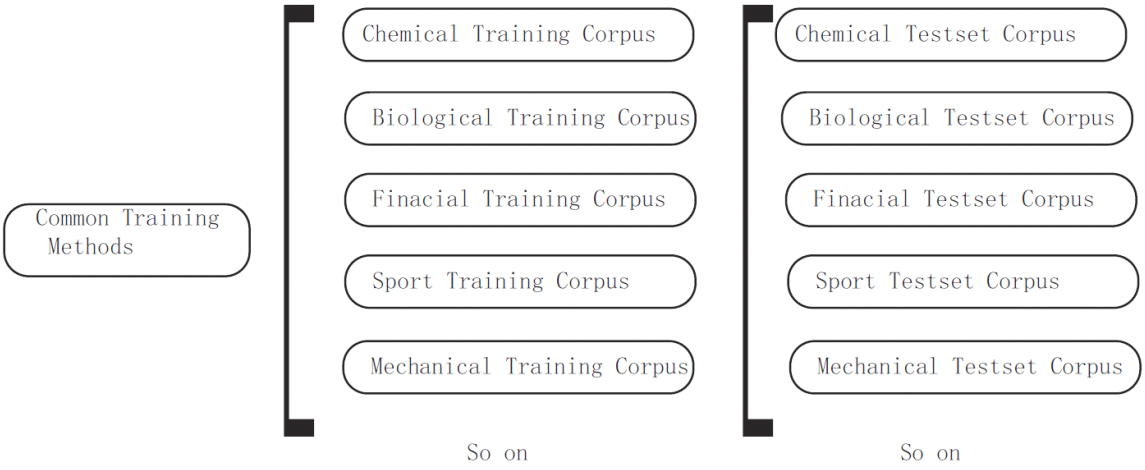

FIGURE 1. The common methods.

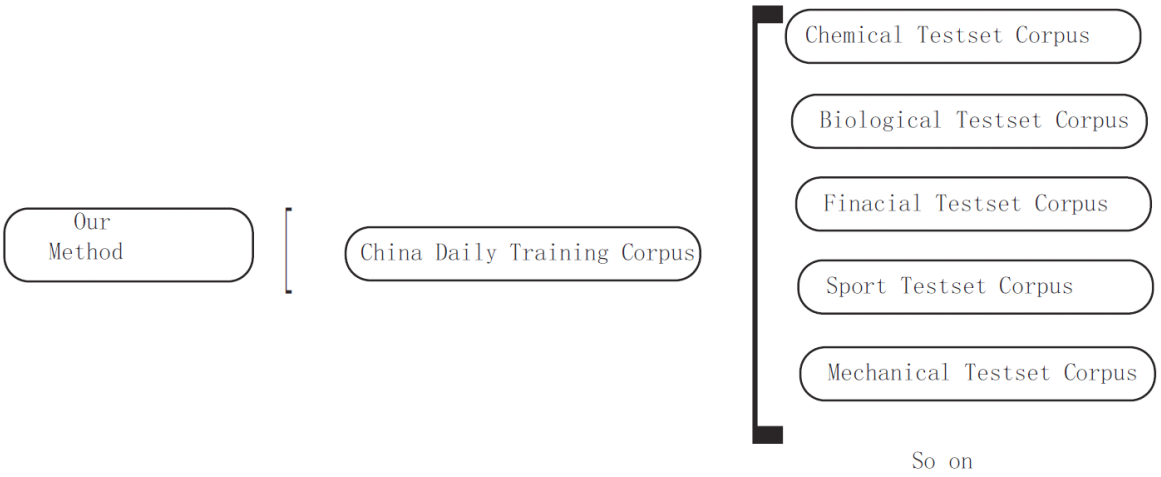

FIGURE 2. Our method.

The CLD(Chinese Levensthein Distance) is calculated using negative knowledge. The reason for this is because in any domain-specific text, the negative examples are always negative; the positive examples, however, are not always positive. As discussed in the introduction, the positive knowledge is not reliable, because the word formation rule has changed. Therefore, the China Daily Corpus is the only training corpus required by our method; By contrast, the other methods that use the knowledge of the word formation rule require a training set for each different domain they are applied to, as illustrated in the Figure 1 and Figure 2.

We further argue, that the methods that use the knowledge of the word formation rule, are unrealistic due to two main reasons:

1. The domain is limitless. Therefore the word formation rule changes continuously.
2. Human resources are limited, as we cannot annotate a large enough corpus for every domain.

To ensure that the evaluation is adequate, the proposed new method must be compared with the existing methods mentioned above. This comparison should be conducted in the following way:

1. Use the same non-domain-specific training text and the same non-domain-specific text set.
2. Use the same domain-specific training text and the same domain-specific text set.

These two methods are, however, not applicable to domain-adaptive methods, such as the one we are proposing.

One may argue that is in unfair to train the WIP using the non-domain-specific text set, and then apply it to the domain-specific test set, as we do not train the WIP in the domain-specific test set. Our method cannot be evaluated in the usual way; A comparison based on conventional mentions would be like comparing an orange with an apple.

REFERENCES

[1] Ando, Rie Kubota and Lilian Lee, Mostly-Unsupervised Statistical Segmentation of Japanese Kanji Sequences. Natural Language Engineering, 9(2),pp.127-149, 2003.

[2] Zhenxing Wang, Changning Huang and Jingbo Zhu, Which Performs Better on In-Vocabulary Word Segmentation: Based on Word or Character? In:*Sixth SIGHAN Workshop on Chinese Language Processing*. 2008.

[3] Gao Jianfeng, Mu Li and Chang-Ning Huang, Improved source-channel model for Chinese word segmentation. In:*ACL2003*. 2003.

[4] Chooi-Ling GOH , Masayuki ASAHARA and Yuji MATSUMOTO, 2004. Pruning False UnknownWords to Improve Chinese Word Segmentation. In Sighan workshop. 2004.

[5] Hua-ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang, and Hong-Kui Yue, Chinese Lexical Analysis Using Hierarchical Hidden Markov Model. In:*Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp.63-70, July 2003, Sapporo, Japan*. 2003.

[6] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong and Qun Liu, HHMM-based Chinese Lexical Analyzer ICTCLAS. In:*Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp.184-187, July 2003, Sapporo, Japan*. 2003.

[7] Jianfeng Gao, Andi Wu, Mu Li, Chang-Ning Huang, Hongqiao Li, Xinsong Xia, Haowei Qin, Adaptive Chinese word segmentation. In:*ACL2004*. 2004.

[8] Jin Hu Huang and David Powers, Chinese Word Segmentation based on Contextual Entropy. Pacific Asia Conference on Language, Information and Computation. 2003.

[9] Nianwen Xue, Chinese Word Segmentation as Character Tagging. International Journal of Computational Linguistics and Chinese Language Processing, 8(1). 2003.

[10] Peng,Fuchun and Schuurmans, Dale, Self-supervised Chinese Word Segmentation. The 4th Internation Symposium on Intelligent Data Analysis(IDA2001), September, 2001, Lisbon, Portugal. 2001.

[11] Sproat, Richard and Tom Emerson, The first international Chinese word segmentation bakeoff. In:*SIGHAN 2003*. 2003.

[12] Sproat, Richard and Tom Emerson, Corpus-based methods in Chinese morphology and phonology. In:*COLING 2002*. 2002.

[13] Wu Andi and Zixin Jiang, Statistically-enhanced new word identification in a rule-based Chinese system. In:*Proc of the 2rd ACL Chinese Processing Workshop*. 2000.