Chinese Word Auto-Confirmation Agent

Jia-Lin Tsai, Cheng-Lung Sung and Wen-Lian Hsu
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
{tsaijl,clsung,hsu}@iis.sinica.edu.tw

Abstract

In various Asian languages, including Chinese, there is no space between words in texts. Thus, most Chinese NLP systems must perform word-segmentation (sentence tokenization). However, successful word-segmentation depends on having a sufficiently large lexicon. On the average, about 3% of the words in text are not contained in a lexicon. Therefore, unknown word identification becomes a bottleneck for Chinese NLP systems.

In this paper, we present a Chinese word auto-confirmation (CWAC) agent. CWAC agent uses a hybrid approach that takes advantage of statistical and linguistic approaches. The task of a CWAC agent is to auto-confirm whether an n-gram input (n ≥ 2) is a Chinese word. We design our CWAC agent to satisfy two criteria: (1) a greater than 98% precision rate and a greater than 75% recall rate and (2) domain-independent performance (F-measure). These criteria assure our CWAC agents can work automatically without human intervention. Furthermore, by combining several CWAC agents designed based on different principles, we can construct a multi-CWAC agent through a building-block approach.

Three experiments are conducted in this study. The results demonstrate that, for n-gram frequency ≥ 4 in large corpus, our CWAC agent can satisfy the two criteria and achieve 97.82% precision, 77.11% recall, and 86.24% domain-independent F-measure. No existing systems can achieve such a high precision and domain-independent F-measure.

The proposed method is our first attempt for constructing a CWAC agent. We will continue develop other CWAC agents and integrating them into a multi-CWAC agent system.

Keywords: natural language processing, word segmentation, unknown word, agent

1. Introduction

For a human being, efficient word-segmentation (in Chinese) and word sense disambiguation (WSD) arise naturally while a sentence is understood. However, these problems are still difficult for the computer. One reason is that it is hard to create unseen knowledge in the computer from running texts [Dreyfus 1992]. Here, unseen knowledge refers to contextual meaning and unknown lexicon.

Generally, the task of unknown lexicon identification is to identify (1) unknown word (2) unknown word sense, (3) unknown part-of-speech (POS) of a word and (4) unknown word pronunciation. Unknown word identification (UWI) is the most essential step in dealing with unknown lexicons. However, UWI is still quite difficult for Chinese NLP. From [Lin *et al.* 1993, Chang *et al.* 1997, Lai *et al.* 2000, Chen *et al.* 2002 and Sun *et al.* 2002], the difficulty of Chinese UWI is caused by the following problems:

1. Just as in other Asian languages, Chinese sentences are composed of strings of characters that do not have blank spaces to mark word boundaries.
2. All Chinese characters can either be a morpheme or a word. Take the Chinese character 花 as an example. It can be either a free morpheme or a word.
3. Unknown words, which usually are compound words and proper names, are too numerous to list in a machine-readable dictionary (MRD).

To resolve these issues, statistical, linguistic and hybrid approaches have been developed and investigated. For statistical approaches, researchers use common statistical features, such as maximum entropy [Yu *et al.* 1998, Chieu *et al.* 2002], association strength [Smadja 1993, Dunnin 1993], mutual information [Florian *et al.* 1999, Church 2000], ambiguous matching [Chen & Liu 1992, Sproat *et al.* 1996], and multi-statistical features [Chang *et al.* 1997] for unknown word detection and extraction. For linguistic approaches, three major types of linguistic rules (knowledge): morphology, syntax, and semantics, are used to identify unknown words. Recently, one important trend of UWI follows a hybrid approach so as to take advantage of both merits of statistical and linguistic approaches. Statistical approaches are simple and efficient whereas linguistic approaches are effective in identifying low frequency unknown words [Chang *et al.* 1997, Chen *et al.* 2002].

Auto-detection and auto-confirmation are two basic steps in most UWI systems. Auto-detection is used to detect the possible n-grams candidates from running texts for a better focus, so that in the next auto-confirmation stage, these identification systems need only focus on the set of possible n-grams. In most cases, recall and precision rates are affected by auto-detection and auto-confirmation. Since trade-off would occur between recall and precision, deriving a hybrid approach with precision-recall optimization has become a major challenge [Chang *et al*. 1997, Chen *et al*. 2002].

In this paper, we introduce a Chinese word auto-confirmation (CWAC) agent, which uses a hybrid approach to effectively eliminate human intervention. A CWAC agent is an agent (program) that automatically confirms whether an n-gram input is a Chinese word. We design our CWAC agent to satisfy two criteria: (1) a greater than 98% precision rate and a greater than 75% recall rate and (2) domain-independent performance (F-measure). These criteria assure our CWAC agents can work automatically without human intervention. To our knowledge, no existing system has yet achieved the above criteria.

Furthermore, by combining several CWAC agents designed based on different principles, we can construct a multi-CWAC agent through a building-block approach and service-oriented architecture (such as web services [Graham *et al*. 2002]). Figure 1 illustrates one way of a multi-CWAC agent system combining three CWAC agents. If the number of identified words of a multi-CWAC agent is greater than that of its any single CWAC agent, we believe a multi-CWAC agent could be able to maintain the 98% precision rate and increase its recall rate by merely integrating with more CWAC agents.

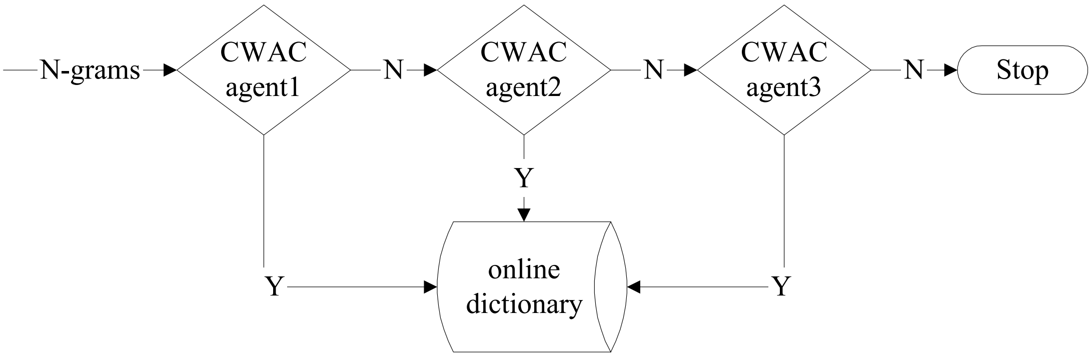

Figure 1. An illustration of a multi-CWAC agent system

This article is structured as follows. In Section 2, we will present a method for simulating a CWAC agent. Experimental results and analyses of the CWAC agent will be presented in section 3. Conclusion and future directions will be discussed in Section 4.

2. Development of the CWAC agent

The most frequent 50,000 words were selected from the CKIP lexicon (CKIP [1995]) to create the system dictionary. From this lexicon, we only use word and POS for our algorithm.

2.1 Major Processes of the CWAC Agent

A CWAC agent automatically identifies whether an n-gram input (or, say, n-char string) is a Chinese word. In this paper, an n-gram extractor is developed to extract all n-grams (n ≥ 2 and n-gram frequency ≥ 3) from test sentences as the n-gram input for our CWAC agent (see Figure 2). (Note that n-gram frequencies vary widely according to test sentences.)

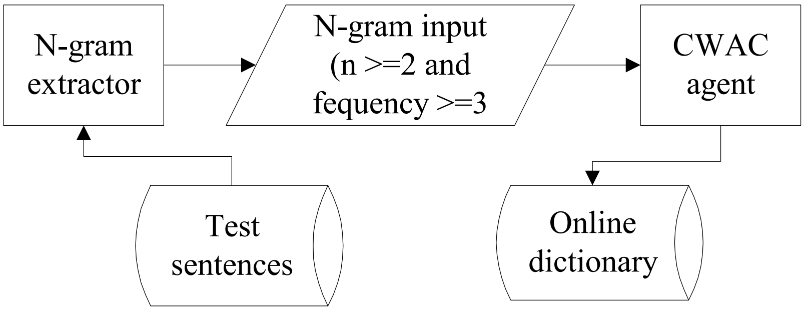

Figure 2. An illustration of n-gram extractor and CWAC agent

Figure 3 is the flow chart of the CWAC agent in which the major processes are labeled (1) to (6). The confirmation types, brief descriptions and examples of the CWAC agent, are given in Table 1. We apply linguistic approach, statistical approach and LFSL (linguistic first, statistical last) approach to develop the CWAC agent. Note in Figure 3, the processes (5) and (6) are statistical methods, and the remaining four processes are developed from linguistic knowledge. The LFSL approach means a combining process of a linguistic process (such as process 4) and a statistical process (such as process 5).

The details of these major processes are described below.

Process 1. System dictionary checking*:* If the n-gram input can be found in the system dictionary, it will be labeled **K0** (which means that the n-gram exists in the system dictionary). In Table 1, the n-gram 計程車 is a system word.

Process 2. Segmentation by system dictionary*:* In this stage, the n-gram input will be segmented by two strategies: left-to-right longest word first (LR-LWF), and right-to-left longest word first (RL-LWF). If LR-LWF and RL-LWF segmentations of the n-gram input are different, the CWAC agent will be triggered to compute the products of all word length for these

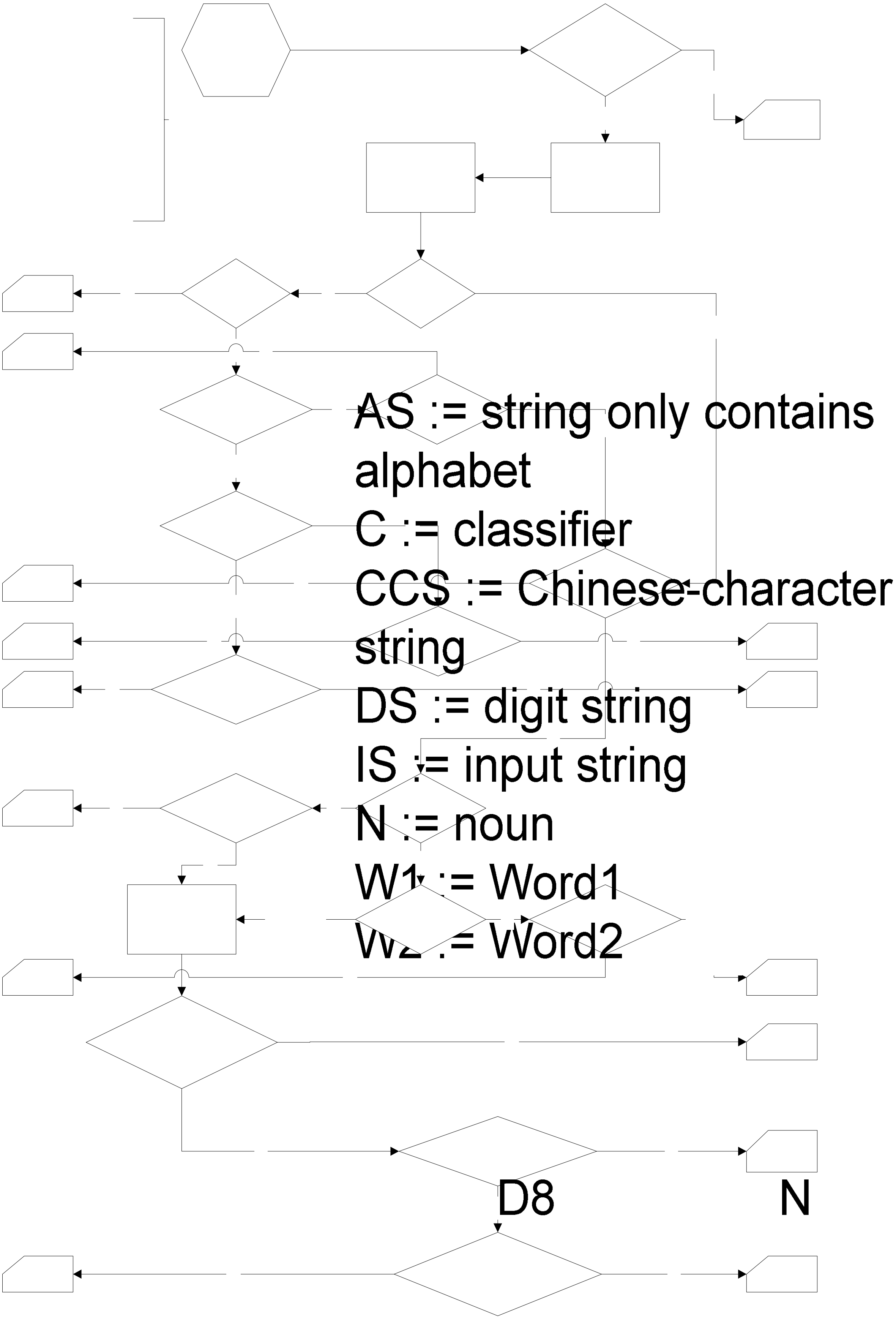

Figure 3. Flow chart for the CWAC agent

Table 1. Confirming results, types, descriptions and examples of the CWAC agent (The symbol / stands for word boundary according to system dictionary using RL-LWF)

Auto-Confirming Results	Types	Brief Descriptions	Examples	
			Input	Output
Word	K0	N-gram exists in system dictionary	計程車	計程車[1]
	K1	Both polysyllabic words exist in online dictionary	接駁公車	接駁/公車[1]
	K2	Two polysyllabic word compounds	食品公司	食品/公司[1]
	K3	Both first and last word of segmented N-gram are polysyllabic words and N ≧ 3	東港黑鮪魚	東港/黑/鮪魚[1]
	K4	Segmentation ambiguity is ≦ 50%	腸病毒	腸/病毒[1]
	K5	N-gram frequency exceeds 10	阿爾巴尼亞裔	阿/爾/巴/尼/亞裔[1]
Not Word	D1	Two polysyllabic word compounds with at least function word	問題一直	問題/一直[2]
	D2	N-gram contains function word	市場指出	市場/指出[2]
	D5	Segmentation ambiguity is > 50%	台北市立	台北市/立[2]
	D6	Suffix Chinese digit string	隊伍	隊/伍[1]
	D7	Digits suffix polysyllabic word	5 火鍋	5/火鍋[2]
	D8	N-gram is a classifier-noun phrase	名學生	名/學生[2]
	D9	N-gram includes unknown symbol	@公司	@/公司[2]
	D0	Unknown reason	[3]	[3]

[1] These n-grams were manually confirmed as *is-word* in this study
[2] These n-grams were manually confirmed as *non-word* in this study
[3] There were no auto-confirming types "D0" and "K0" in this study

segmentations. If both products are equal, the RL-LWF segmentation will be selected. Otherwise, the segmentation with the greatest product will be selected. According to our experiment, the segmentation precision of RL-LWF is, on the average, 1% greater than that of LR-LWF. Take n-gram input 將軍用的毛毯 as an example. Its LR-LWF and RL-LWF segmentations are 將軍/用/的/毛毯 and 將/軍用/的/毛毯, respectively. Since both products are equal (2x1x1x2=1x2x1x2), the selected segmentation output for this process is 將/軍用/的/毛毯 as it is the RL-LFW.

Process 3. Stop word checking**:* The segmentation output from ***Process 2 is referred to as *segmentation2*. In this stage, all words in *segmentation2* will be compared with the stop word list. There are three types of stop words: **begining**, **middle**, and **end**. The stop word list used in this study is given in Appendix A. These stop words were selected by native Chinese speakers according to those computed beginning, middle, and end single-character words with < 1% of being the beginning, middle, or end words of Hownet [Dong 1999], respectively. If the first and last words of *segmentation2* can be found on the list of begining and end stop words, they will be eliminated from the *segmentation2*. For those cases in which the word number of *segmentation2* is greater than 2, middle stop word checking will be triggered. If a middle word in *segmentation2* can be found in the middle stop word list, the n-gram input will be split into new strings at any matched stop word. These new strings will be sent to ***Process 1*** as new n-gram input. For example, *segmentation2* of the n-gram input 可怕的腸病毒” is 可怕/的/腸/病毒. Since there is a middle stop word “的” in this *segmentation2*, the new strings 可怕 and 腸病毒 will be sent to ***Process 1*** as new n-gram input.

Process 4. Part-of-Speech (POS) pattern checking**:* Once *segmentation2* has been processed by ***Process 3, the result is called *segmentation3*. If the word number of *segmentation3* is 2, POS pattern checking will be triggered. The CWAC agent will first generate all possible POS combinations of the two words using the system dictionary. If the number of generated POS combinations is one and that combination matches one of the POS patterns (**N/V**, **V/N**, **N/N**, **V/V**, **Adj/N**, **Adv/N**, **Adj/V**, **Adv/V**, **Adj/Adv**, **Adv/Adj**, **Adv/Adv** and **Adj/Adj**) the 2-word string will be tagged as a word and sent to ***Process 5***. This rule-based approach combines syntax knowledge and heuristic observation in order to identify compound words. For example, since the generated POS combination for *segmentation3* 食品/公司 is **N/N**, 食品公司 will be sent to ***Process 5***.

***Process 5. Segmentation ambiguity checking**:* This stage consists of 4 steps:

1) Thirty randomly selected sentences that include the n-gram input will be extracted from either a large scale or a fixed size corpus. For example, the Chinese sentence “人人做環保” is a selected sentence that includes the n-gram input “人人”. The details of large scale and fixed size corpus used in this study will be addressed on Subsection 3.2. (Note that the number of selected sentences may be less than thirty and may even be zero due to corpus sparseness.)
2) These selected sentences will be segmented using the system dictionary, and will be segmented by the RL-LWF and LR-LWF technique.
3) For each selected sentence, if the RL-LWF and LR-LWF segmentations are different, the sentence will be regarded as an ambiguous sentence. For example, the Chinese sentence “人人做環保” is not an ambiguous sentence.
4) Compute the ambiguous ratio of ambiguous sentences to selected sentences. If the ambiguous ratio is less than 50%, the n-gram input will be confirmed as word type **K1**, **K2** or **K4** by ***Process 5*** (see Fig. 3) ; other-

wise, it will be labeled **D1** or **D2**. According to our observation, the ambiguous ratios of non word n-grams usually are greater than 50%.

Process 6. Threshold value checking**:* In this stage, if the frequency of an n-gram input is greater or equal to 10, it will be labeled as word type **K5** by ***Process 6. According to our experiment, if we directly regard an n-gram input whose frequency is greater than or equal to a certain threshold value as a word, the trade-off frequency of 99% precision rate occurs at the threshold value 7.

3. Experiment Results

The objective of the following experiments is to investigate the performance of the CWAC agent. By this objective, in ***process1*** of the CWAC agent, if an n-gram input is found to be a system word, a temporary system dictionary will be generated. The temporary system is the original system dictionary without this n-gram input. In this case, the n-gram input will be sent to ***process2*** and the temporary system dictionary will be used as system dictionary in both ***process2*** and ***process5***.

Three experiments are conducted in this study. Their results and analysis are given in Sections 3.3, 3.4 and 3.5.

3.1 Notion of Word and Evaluation Equations

The definition of word is not unique in Chinese [Sciullo *et al.* 1987, Sproat *et al.*, 1996, Huang *et al.* 1997, Xia 2000]. As of our knowledge, the Segmentation Standard in China [Liu. *et al.* 1993] and the Segmentation Standard in Taiwan [CKIP 1996] are two of the most famous word-segmentation standards to Chinese. Since the Segmentation Guidelines for the Penn Chinese Treebank (3.0) [Xia 2000] has tried to accommodate the above famous segmentation standards in it, this segmentation was selected as our major guidelines for determining Chinese word. The notion of word in this study includes fixed-phrase words (such as 春夏秋冬, 你一句我一句, 奧林匹克運動會, etc.) , compounds (such as 腳踏車龍頭, 太陽眼鏡, etc.) and simple words (such as 房子, 老頭兒, 盤尼西林, etc.).

We use recall, precision and F-measure to evaluate the overall performance of the CWAC agent [Manning *et al.* 1999]. Precision, recall and F-measure are defined below. Note that the words in following equations (1) and (2) include new words and dictionary words.

recall = # of correctly identified words / # of words (1)

precision = # of correctly identified words / # of identified words (2)

F-measure = (2 × recall × precision) / (recall + precision) (3)

3.2 Large Scale Corpus and Fixed Size Corpus

In Section 2, we mentioned that the corpus used in ***process5*** of the CWAC agent can be large scale or fixed size. The description of a large scale and a fixed size corpus is given below.

(1) ***Large scale corpus***: In our experiment, texts are collected daily. Texts collected in most Chinese web sites can be used as a large scale corpus. Here, we select **OPENFIND** [OPENFIND], one of the most popular Chinese search engines, to act as a large scale corpus. If ***process 5*** of the CWAC agent is in large scale corpus mode, it will extract the first thirty matching sentences, including the n-gram input, from the **OPENFIND** search results.

(2) ***Fixed size corpus***: A fixed size corpus is one whose text collection is limited. Here, we use a collection of 14,164,511 Chinese sentences extracted from whole 2002 articles obtained from *United Daily News (UDN)* web site [UDN] as our fixed size corpus, called 2002 *UDN* corpus.

3.3 The First Experiment

The objective of the first experiment is to investigate whether our CWAC agent satisfies criterion 1: the precision rate should be greater than 98% and the recall greater than 75%.

First, we create a testing corpus, called 2001 *UDN* corpus, consisting of 4,539,624 Chinese sentences extracted from all 2001 articles on the *UDN* Web site. The testing corpus includes 10 categories: 地方(local), 股市(stock), 科技(science), 旅遊(travel), 消費(consuming), 財經(financial), 國際(world), 運動(sport), 醫藥(health) and 藝文(arts). For each category, we randomly select 10,000 sentences to form a test sentence set. We then extract all n-grams from each test sentence set. We then obtain 10 test n-gram sets. All of the extracted n-grams have been manually confirmed as three types: *is-word*, *unsure-word* or *non-word*. In this study, the average percentages of n-grams manually confirmed as *is-word*, *unsure-word*, and *non-word* are 78%, 2% and 20%, respectively. When we compute precision, recall and F-measure, all *unsure-word* n-grams are excluded. Table 2 shows the results of the

CWAC agent in large scale corpus mode. Table 3 shows the results of the CWAC agent in fixed size corpus mode.

Table 2. The first experimental results of the CWAC agent in large scale corpus mode

Large scale Corpus						
n-grams frequency ≥ 3				n-grams frequency ≥ 4		
Class	P	R	F	P	R	F
地方	97.72%	76.37%	85.74%	98.54%	76.27%	85.99%
股市	94.32%	74.40%	83.19%	95.32%	75.51%	84.26%
科技	96.51%	76.33%	85.24%	97.64%	76.54%	85.81%
旅遊	97.51%	77.80%	86.55%	98.13%	78.09%	86.97%
消費	97.85%	79.41%	87.67%	98.56%	78.72%	87.53%
財經	95.68%	74.63%	83.86%	97.32%	75.74%	85.18%
國際	96.41%	78.64%	86.62%	97.26%	78.36%	86.79%
運動	94.17%	78.99%	85.92%	95.08%	78.66%	86.10%
醫藥	96.80%	78.09%	86.44%	98.60%	76.85%	86.38%
藝文	96.94%	76.87%	85.75%	98.20%	76.44%	85.96%
Avg.	96.31%	77.18%	85.69%	97.82%	77.11%	86.24%

Table 3. The first experimental results of the CWAC agent in fixed size corpus mode

Fixed size Corpus						
n-grams frequency ≥ 3				n-grams frequency ≥ 4		
Class	P	R	F	P	R	F
地方	97.93%	73.46%	83.95%	98.37%	73.91%	84.41%
股市	95.76%	69.60%	80.61%	96.63%	70.30%	81.39%
科技	97.70%	69.01%	80.89%	98.15%	68.99%	81.03%
旅遊	97.95%	70.09%	81.71%	98.61%	70.49%	82.21%
消費	98.20%	74.76%	84.89%	98.79%	74.73%	85.09%
財經	97.02%	67.41%	79.55%	97.76%	68.56%	80.60%
國際	97.06%	73.56%	83.69%	97.81%	73.00%	83.60%
運動	95.77%	74.03%	83.51%	97.02%	74.96%	84.57%
醫藥	97.68%	71.72%	82.71%	98.26%	71.64%	82.87%
藝文	98.22%	70.20%	81.88%	99.02%	69.40%	81.61%
Avg.	97.32%	71.44%	82.39%	98.11%	71.61%	82.79%

As shown in Table 2, the CWAC agent in large scale corpus mode can achieve 96.31% and 97.82% precisions, 77.18% and 77.11% recalls and 85.69% and 86.24%

F-measures for n-gram frequencies of ≥ 3 and ≥ 4, respectively. Table 3 shows that the CWAC agent in fixed size corpus mode can achieve 97.32% and 98.11% precisions, 71.44 and 71.61% recalls and 82.39% and 82.79% F-measures.

The hypothesis tests of whether the CWAC agent satisfies criterion 1, **H1a** and **H1b**, for this experiment are given below. (One-tailed t-test, reject H_0 if its p-value > 0.05)

H1a. H_0: avg. precision ≤ 98%, H_1: avg. precision > 98%

H1b. H_0: avg. recall ≤ 77%, H_1: avg. recall > 77%

From Tables 2 and 3, we compute the p-values of **H1a** and **H1b** for four CWAC modes in Table 4. Table 4 shows that the CWAC agent passes both hypotheses **H1a** and **H1b** in large scale corpus mode with an n-gram frequency of ≥ 4.

In Chen *et al*. (2002), a word that occurs no less than three times in a document is a high frequency word; otherwise, it is a low frequency word. Since a low frequency word in a document could be a high frequency word in our test sentence sets, the results in Tables 2 and 3 can be regarded as an overall evaluation of UWI for low and high frequency words.

Table 4. The p-values of the hypothesis tests, **H1a** and **H1b**, for four CWAC modes

CWAC mode	P-value (**H1a**)	P-value (**H1b**)
Large scale & Frequency ≥ 3	0.0018 (accept H_0)	0.3927 (reject H_0)
Large scale & Frequency ≥ 4	0.1114 (reject H_0)	0.3842 (reject H_0)
Fixed size & Frequency ≥ 3	0.0023 (accept H_0)	0.0 (accept H_0)
Fixed size & Frequency ≥ 4	0.4306 (reject H_0)	0.0 (accept H_0)

In Chen *et al*. (2002), researchers try to use as much information as possible to identify unknown words in hybrid fashion. Their results have 88%, 84% and 89% precision rates; 67%, 82% and 68% recall rates; 76%, 83%, 78% F-measure rates on low, high, and low/high frequency unknown words, respectively.

3.3.1 A Comparative Study

Table 5 compares some of the famous works on UWI (here, the performance of our CWAC agent was computed solely against "new words" exclude words that are already in system dictionary). In Table 5, the system of [Chen *et al*. 2002] is one of the most famous hybrid approaches on unknown word extraction. Although Lai's system [Lai *et al*. 2000] achieves the best F-measure 88.45%, but their identifying target (including words and phrases) is different from conventional UWI system. Thus, Lai's result is not included in Table 5.

Table 5. Comparison of works on UWI

System	Method	Target	Test size	P	R	F
[Our CWAC]	Hybrid	n-gram word	100,000 sentences	94.32	74.50	83.25
[Chen *et al*. 2002]	Hybrid	n-gram word	100 documents	89	68	77.10
[Sun et al. 2002]	Statistical	name entity	MET2 (Chen *et al*. 1997)	77.89	86.09	81.79
[Chang *et al*. 1997]	Statistical	bi-gram word	1,000 sentences	72.39	82.83	76.38

3.4 Second Experiment

The objective of this experiment is to investigate whether the CWAC agent satisfies criterion 2: the F-measure should be domain-independent.

The hypothesis test **H2** for this experiment is given below. (Two-tailed t-test, reject H_0 if its p-value < 0.05)

H2. H_0: avg. F-measure = μ_0; H_1: avg. F-measure ≠ μ_0

Table 6 lists the p-values of **H2** for four CWAC modes. Table 6 shows that the CWAC agent passes H2 and satisfies criterion 2 in all four CWAC modes.

Table 6. The p-values of the hypothesis test **H2** for four CWAC modes

CWAC mode	μ_0 (F-measure)	P-value
Large scale & Frequency ≥ 3	86%	0.4898 (accept H_0)
Large scale & Frequency ≥ 4	86%	0.7466 (accept H_0)
Fixed size & Frequency ≥ 3	83%	0.2496 (accept H_0)
Fixed size & Frequency ≥ 4	83%	0.6190 (accept H_0)

Summing up the results of first and second experiments, we conclude that our method can be used as a CWAC agent in large scale corpus mode when an n-gram frequency is ≥ 4.

3.5 Third Experiment

The objective of this experiment is to investigate whether the precision of our

CWAC agent is corpus-independent (**Q1**) and whether its recall is corpus-dependent (**Q2**). We use large scale and fixed size corpus modes to test **Q1** and **Q2**.

The hypothesis tests, **H3a** and **H3b**, for this experiment are given below. (Two-tailed t-test, reject H_0 if its p-value < 0.05)

H3a.H_0: avg. precision of large scale (μ1) = avg. precision of fixed size (μ2)
H_1: avg. precision of large scale (μ1) ≠ avg. precision of fixed size (μ2)

H3b.H_0: avg. recall of large scale (μ3) = avg. recall of fixed size (μ4)
H_1: avg. recall of large scale (μ3) ≠ avg. recall of fixed size (μ4)

Table 7 lists the p-values of **H3a** and **H3b** for n-gram frequencies of ≥ 3 and ≥ 4. Table 7 shows that **H3a** is accepted at the 5% *significance* level. This shows that the precision of the CWAC agent is corpus-independent, since the average precisions of both corpus modes equal at the 5% level. On the other hand, **H3b** is rejected at the 5% *significance* level. This shows the recall is corpus-dependent, since the average recalls of both corpus modes are not equal at the 5% level.

Table 7. The p-values of the hypothesis tests, **H3a** and **H3b**, for two frequency modes

Frequency mode	P-value (H3a)	P-value (H3b)
Frequency ≥ 3	0.079392 (accept H_0)	0.0000107 (reject H_0)
Frequency ≥ 4	0.238017 (accept H_0)	0.0000045 (reject H_0)

Tables 8 and 9 were created to sum up the experimental results in Tables 2 and 3. Table 8 gives the comparison of the linguistic, statistic and LFSL approaches in this study. From Table 8, it shows that the CWAC agent using the technique of LFSL achieves the best optimization of precision-and-recall with the greatest F-measure. Table 9 is the overall experimental results of the CWAC agent for n-gram frequencies of ≥ 3 to ≥ 10. From Table 9, it indicates the precisions, recalls and F-measures of the CWAC agent are close for different n-gram frequency conditions.

Table 8. Comparison of the linguistic, statistical and LFSL approaches results

N-grams frequency	Approach[1]	Precision (large, fixed)[2]	Recall (large, fixed)	F-measure (large, fixed)
≥ 3	Linguistic (L)	92.44%, 93.71%	67.41%, 48.96%	77.96%, 64.31%
≥ 3	Statistical (S)	89.15%, 100.00%	4.67%, 3.39%	8.88%, 6.56%
≥ 3	LFSL	96.72%, 97.43%	98.27%, 97.24%	97.49%, 97.34%

[1] The linguistic approaches include auto-confirmation types K3, D6, D7, D8 and D9; the statistical approaches include auto-confirmation types K1, K5, D1 and D5; the LFSL (linguistic approach first, statistical approach last) approaches include auto-confirmation types K2, K4 as shown in Fig. 3

[2] "large" means large scale corpus mode and "fixed" means fixed size corpus mode

Table 9. Overall experiment results

N-grams frequency	# of n-grams	Precision (large, fixed)[1]	Recall (large, fixed)	F-measure (large, fixed)
≥ 3	70502	96.31%, 97.32%	77.18%, 71.44%	85.69%, 82.39%
≥ 4	49500	97.82%, 98.11%	77.11%, 71.61%	86.24%, 82.79%
≥ 5	38179	97.49%, 98.52%	77.11%, 71.78%	86.11%, 83.05%
≥ 6	31382	97.64%, 98.76%	76.78%, 71.78%	85.96%, 83.14%
≥ 7	26185	97.77%, 99.00%	76.50%, 71.52%	85.84%, 83.05%
≥ 8	22573	97.86%, 99.11%	76.23%, 71.48%	85.70%, 83.06%
≥ 9	19473	97.84%, 99.16%	75.60%, 70.99%	85.29%, 82.74%
≥ 10	17048	97.72%, 99.17%	75.26%, 70.96%	85.03%, 82.73%

[1] "large" means large scale corpus mode and "fixed" means fixed size corpus mode

4. Conclusion and Directions for Future Research

UWI is the most important problem in handling unknown lexicons in NLP systems. A lexicon consists of words, POSs, word senses and word pronunciations. As shown in [Lin *et al.* 1993, Chang *et al.* 1997, Lai *et al.* 2000, Chen *et al.* 2002 and Sun *et al.* 2002], UWI is still a very difficult task for Chinese NLP systems. One important trend toward resolving unknown word problems is to follow a hybrid approach by combining the advantages of statistical and linguistic approaches. One of the most critical issues in identifying unknown words is to overcome the problem of precision-and-recall trade-off.

In this paper, we create a CWAC agent adopting a hybrid method to auto-confirm n-gram input. Our experiment shows that the LFSL (linguistic approach first, statistical approach last) approach achieves the best precision-and-recall optimization. Our results demonstrate that, for n-gram frequency ≥ 4 in large corpus mode, our CWAC agent can achieve 97.82% precision, 77.11% recall, and 86.24% F-measure. Thus, it satisfies that two criteria. Moreover, we discover that the use of large scale corpus in this method increases recall but not precision. On the other hand, we find that the precision of using either a large scale corpus or a fixed size corpus is not statistical significantly different at the 5% level.

This method is our first attempt to create a CWAC agent. We have also considered a building-block approach to construct a multi-CWAC agent. We believe a multi-CWAC agent could be able to maintain the 98% precision rate and increase recall rate by integrating more CWAC agents.

In the future, we will continue addressing agent-oriented and service-oriented approaches for handling unknown lexicons, such as unknown word POS auto-tagging agent and unknown word-sense auto-determining agent. Furthermore, the method to achieve corpus-independent recall will also be considered.

5. Acknowledgements

We are grateful to the our colleagues in the Intelligent Agent Systems Lab., Gladys Hsieh, Yi-Fan Chang and Yin-Ru Chen, who helped us to manually-confirm n-grams and create tools for this study.

References

Chen, K.J. and W.Y. Ma, "Unknown Word Extraction for Chinese Documents," *Proceedings of 19th COLING 2002*, Taipei, pp.169-175

Chieu, H.L. and H.T. Ng, "Named Entity Recognition: A Maximum Entropy Approach Using Global Information," *Proceedings of 19th COLING 2002*, Taipei, pp.190-196

Chang, J.S. and K.Y. Su, "An Unsupervised Iterative Method for Chinese New Lexicon Extraction," *International Journal of Computational Linguistics & Chinese language Processing*, 1997

Chen, K.J. and S.H. Liu, "Word Identification for mandarin Chinese Sentences," *Proceedings of 14th COLING*, pp. 101-107

Church, K.W., "Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p*p," *Proceedings of 18th COLING 2000*, pp.180-186

CKIP (Chinese Knowledge Information processing Group), Technical *Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Taiwan, Taipei, Academia Sinica, 1995. http://godel.iis.sinica.edu.tw/CKIP/r_content.html

CKIP (Chinese Knowledge Information processing Group), *A study of Chinese Word Boundaries and Segmentation Standard for Information processing (in Chinese)*. Technical Report, Taiwan, Taipei, Academia Sinica, 1996.

Dreyfus, H.L., *What computers still can't do: a critique of artificial reason*, Cambridge, Mass. : MIT Press, 1992

Dong, Z. and Q. Dong, Hownet, 1999, http://www.keenage.com/

Dunnin, T., "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, n 1., 1993

Florian, R. and D. Yarowsky, "Dynamic nonlocal language modeling via hierarchical topic-based adaptation," Proceedings of ACL99, 1999, pp. 167-174

Graham, S., S. Simeonov, T. Boubez, D. Davis, G. Daniels, Y. Nakamura and R. Neyama, *Building Web Services With Java*, Pearson Education, 2002

Huang, C.R., Chen, K.C., Chen, F.Y. and Chang, L.L., "Segmentation Standard for Chinese natural language Processing," Computational Linguistics and Chinese Language Processing, 2(2), Aug., 1997, pp.47-62

Lai, Y.S. and Wu, C.H., "Unknown Word and Phrase Extraction Using a Phrase-Like-Unit-based Likelihood Ratio," *International Journal of Computer Processing Oriental Language*, 13(1), pp.83-95

Lin, M.Y., T.H. Chiang and K.Y. Su, "A preliminary Study on Unkown Word Problem in Chinese Word Segmentation," *Proceedings of ROCLING VI*, 1993, pp. 119-137

Manning, C.D. and Schuetze, H., *Fundations of Statistical Natural Language Processing*, MIT Press, 1999, pp.534-538

OPENFIND, OPENDFIN Chinese Search Web Site, http://www.openfind.com.tw/

Sciullo, A.M.D. and Williams, E., *On the Definition of Word*, MIT press, 1987

Smadja, F., "Retrieving Collocations from Text: Xtract," *Computational Linguistics*, 22(1)

Sproat, R., C. Shih, W. Gale and N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22(3), 1996, pp.377-404

Sun, J., J. Gao, L. Zhang, M. Zhou and C. Huang, "Chinese Named Entity Identification Using Class-based Language Model," *Proceedings of 19th COLING 2002*, Taipei, pp.967-973

UDN, On-Line United Daily News , http://udnnews.com/NEWS/

Xia, F., *The Segmentation Guidelines for the Penn Chinese Treebank (3.0)*, October 17, 2000

Yu, S., S. Bai, and P. Wu, "Description of the Kent Ridge Digital Labs System Used for MUC-7," *Proceedings of the 7th Message Understanding Conference*, 1998

Appendix A. Stop Words List

I. Begining stop word list

/兒/呀/嗎/吧/呢/呼/了/是/你/我/他/又/等/既/或/有/到/去/在/爲/
/及/和/與/之/的/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/
/未/能/將/此/可/與/到/向/以/用/乃/入/又/下/久/乎/者/小/已/才/
/互/仍/勿/太/欠/且/乎/去/只/必/再/吁/多/好/如/早/而/至/行/但/
/別/即/吧/呀/更/沒/矣/並/和/呢/或/所/則/卻/哉/很/後/怎/既/甚/
/皆/相/若/唷/哼/哩/唉/哦/啊/得/都/最/喂/喔/喳/喲/等/著/嗎/嗨/
/嗚/嗡/愈/跟/較/過/嘛/嘎/嘟/嘻/嘿/噓/噗/罷/噹/噯/還/雖/嚕/

II. Middle stop word list

/可/已/各/被/到/等/既/但/且/而/並/同/又/爲/是/有/或/及/和/與/
/之/的/在/的/在/以/已/將/與/和/是/及/也/或/之/於/由/都/並/卻/
/且/只/則/但/又/才/仍/該/各/其/有/時/前/後/上/中/下/再/更/不/
/很/最/多/非/稍/否/至/了/吧/嗎/但/因/爲/而/且/就/對/雖/裡/裏/
/等/要/把/到/去/給/打/做/作/個/你/妳/我/他/她/它/們/這/那/此/
/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/未/能/將/此/可/
/與/到/向/以/用/乃/入/又/下/久/乎/者/已/互/仍/勿/欠/且/乎/去/
/只/必/再/吁/多/好/如/早/而/至/但/別/即/吧/呀/更/沒/矣/並/呢/
/或/所/則/卻/哉/很/後/怎/既/甚/皆/相/若/唷/哼/哩/唉/哦/啊/得/
/都/最/喂/喔/喳/喲/等/著/嗎/嗨/嗚/嗡/愈/跟/較/過/嘛/嘎/嘟/嘻/
/嘿/噓/噗/罷/噹/噯/還/雖/嚕/

III. End stop word list

/等/及/與/的/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/未/
/能/將/此/可/會/與/到/向/以/用/乃/入/又/下/久/乎/者/小/已/才/
/互/仍/勿/太/欠/且/乎/去/只/必/再/吁/多/好/如/早/而/至/行/但/
/別/即/吧/呀/更/沒/矣/並/和/呢/或/所/則/卻/哉/很/後/怎/既/甚/
/皆/相/若/唷/哼/哩/唉/哦/啊/得/都/最/喂/喔/喳/喲/等/著/嗎/嗨/
/嗚/嗡/愈/跟/較/過/嘛/嘎/嘟/嘻/嘿/噓/噗/罷/噹/噯/還/雖/嚕/