Statistical Augmentation of a Chinese Machine-Readable Dictionary

Pascale Fung
Columbia University
Computer Science Department
New York, NY 10027
USA
pascale@cs.columbia.edu

Dekai Wu
HKUST
Department of Computer Science
University of Science & Technology
Clear Water Bay, Hong Kong
dekai@cs.ust.hk

Abstract

We describe a method of using statistically-collected Chinese character groups from a corpus to augment a Chinese dictionary. The method is particularly useful for extracting domain-specific and regional words not readily available in machine-readable dictionaries. Output was evaluated both using human evaluators and against a previously available dictionary. We also evaluated performance improvement in automatic Chinese tokenization. Results show that our method outputs legitimate words, acronymic constructions, idioms, names and titles, as well as technical compounds, many of which were lacking from the original dictionary.

1 Introduction

Finding new lexical entries for Chinese is hampered by a particularly obscure distinction between characters, morphemes, words, and compounds. Even in Indo-European text where words can be separated by spaces, no absolute criteria are known for deciding whether a collocation constitutes a compound word. Chinese defies such distinctions yet more strongly. Characters in Chinese (and in some other Asian languages such as Japanese and Korean) are not separated by spaces to delimit words; nor do characters give morphological hints as to word boundaries. Each single character carries a meaning and can be ambiguous; most are many-way polysemous or homonymous.

Some characteristics of Chinese words are nonetheless clear. A word in Chinese is usually a bigram (two character word), a unigram, a trigram, or a 4-gram. Function words are often unigrams, and n-grams with $n > 4$ usually are specific idioms. According to the *Frequency Dictionary of Modern Chinese* (FDMC 1986), among the top 9000 most frequent words, 26.7% are unigrams, 69.8% are bigrams, 2.7% are trigrams, 0.007% 4-grams, and 0.0002% 5-grams. Another study (Liu 1987) showed that in general, 75% of Chinese words are bigrams, 14% trigrams, 6% n-grams with $n > 3$.

Inadequate dictionaries have become the major bottleneck to Chinese natural language processing. Broad coverage is even more essential than with Indo-European languages, because not even the most basic lexicosyntactic analysis can proceed without first identifying the word boundaries. Thus a significant number of models for *tokenizing* or *segmenting* Chinese have recently been proposed, using either rule-based or statistical methods (Chiang *et al.* 1992; Lin *et al.* 1992; Chang & Chen 1993; Lin *et al.* 1993; Wu & Tseng 1993; Sproat *et al.* 1994). But all of these approaches rely primarily upon dictionary lookup of the potential segments; in spite of experimental heuristics for handling unknown words in the input text, accuracy is seriously degraded when dictionary entries are missing.

Tokenization problems are aggravated by text in specialized domains. Such documents typically contain a high percentage of technical or regional terms that are not found in the tokenizer's dictionary (machine-readable Chinese dictionaries for specialized domains are not readily available). Most effective tokenizers have domain-specific words added manually to the dictionary. Such manual strategies are too tedious and inefficient in general.

This paper discusses a fully automatic statistical tool that extracts words from an untokenized Chinese text, creating new dictionary entries. In addition, it is desirable to identify regional and domain-specific technical terms that are likely to appear repeatedly in a large corpus. We extended and re-targeted a tool originally designed for extracting English compounds and collocations, Xtract, to find words in Chinese. We call the resulting tool **CXtract**. Words found by CXtract are used to augment our dictionary.

In the following sections, we first describe the modifications in CXtract for finding Chinese words, and the corpus used for training. The resulting words and collocations are evaluated by human evaluators, and recall and precision are measured against the tokens in the training set. The significance of evaluated results will be discussed. Finally, we discuss a preliminary evaluation of the improvement in tokenization performance arising from the output of our tool.

2 A Collocation Extraction Tool

Xtract was originally developed by Smadja (1993) to extract collocations in an English text. It consists of a package of software tools used to find likely co-occurring word groups by statistical analysis.

In the first stage of Xtract, all frequent bigrams are found. These bigram words are permitted to occur within a window of 10 positions, specifically, at distance between -5 to 5 relative to each other. A threshold is set on the frequency, to discard unreliable bigrams. The remaining bigrams constitute part of the output from Xtract (along with the output from the second stage).

The second stage looks at a tagged corpus and to find collocations of involving more than two words—up to ten—using the bigram words found in the first stage as anchors. Again, a frequency threshold is set to discard unreliable collocations.

Xtract's output consists of two types of collocations. In the simpler case, a collocation is an adjacent word sequence such as "stock market" (extracted from the Wall Street Journal). More general collocations permit flexible distances between two word groups, as in "make a ... decision".

For our purpose, we were interested in looking for adjacent character groups without distances between the groups. We postulated that, just as "stock market" could be regarded as a compound word, we would discover that frequently appearing continuous character groups are likely to be words in Chinese. We were also interested in looking for multi-word collocations in Chinese since these would presumably give us many technical and regional terms.

Because Xtract was originally developed for English, many capabilities for handling non-alphabetic languages were lacking. We extended Xtract to process character-based Chinese texts without tags. Various stages of the software were also modified to deal with untagged texts.

Other parametric modifications arose from the difference between the distribution of characters that make up Chinese words, versus the words that make up English compounds. For example, the frequency threshold for finding reliable bigrams is different because CXtract returns far more Chinese character bigrams than English word bigrams returned by Xtract.

3 Experiment I: Dictionary Augmentation

Our experiments were aimed at determining whether our statistically-generated output contains legitimate words. We are using text from (the Chinese part of) the HKUST English-Chinese Parallel Bilingual Corpus (Wu 1994), specifically, transcriptions of the parliamentary proceedings of the Legislative Council. The transcribed Chinese is formalized literary Cantonese that is closer to Mandarin than conversational Cantonese. However, more vocabulary is preserved from classical literary Chinese than in Mandarin, which affects the ratio of bigrams to other words.

Evaluation of legitimate Chinese words is not trivial. It is straightforward to evaluate those outputs that can be found in a machine-readable dictionary such as the one used by the tokenizer. However, for unknown words, the only evaluation criterion is human judgement. We evaluated the output lexical items from CXtract by both methods.

3.1 Procedure

For Experiment I, we used a portion of the corpus containing about 585,000 untokenized Chinese characters (which turned out to hold about 400 thousand Chinese words after tokenization). The experiment was carried out as follows:

1: A dictionary of all unigrams of characters found in the text was composed. One example is the character 立 (*li*) which can mean "stand" or "establish" by itself.

2: From the unigram list, we found all the bigrams associated with each unigram and obtained a list of all bigrams found.

3: We kept only bigrams which occur significantly more than chance expectation, and which appear in a rigid way (Smadja 1993). This yields a list of possible bigrams and most frequent relative distance between the two characters. The distances are kept between -5 and 5 as in Xtract since this ultimately gives collocations of lengths up to 10, which we found sufficient for Chinese.

4: From this bigram list, we extracted only those bigrams in which the two characters occur adjacently. We assumed such bigrams to be Chinese words. For 立 (*li*), one output bigram was 立法 (*li fa*) which means "legislative", a legitimate word.

5: Using all bigrams (adjacent and non-adjacent) from (3), we extracted words and collocations of lengths greater than two. Outputs with frequency less than 8 were discarded.

6: We divided the output from (5) into lists of trigrams, 4-grams, 5-grams, 6-grams, and m-grams where $m > 6$. One of the trigrams, for example, is 立法局 (*li fa ju*) which means "Legislative Council" and is another legitimate word.

3.2 Results

A portion of the list of bigrams obtained from (4) is shown in Figure 1. We obtained 1695 such bigrams after thresholding.

Part of the output from (5) is shown in Figure 2. The first and the last numbers on each line is the frequency for the occurrence of the n-grams.

Parts of the output from (6) are shown in Figures 3, 4, and 5.

立 法	legislative	人 士	personage	政 府	government
提 供	supply	發 展	develop(ment)	意 見	opinion
部 份	partial	包 括	include	教 育	educate(-tion)
要 求	demand	政 治	politics	而 且	moreover
制 度	system	研 究	research	一 般	in general
市 民	citizen	組 別	group/category	一 個	one *classifier*
法 律	law	條 例	regulation	報 告	report
缺 乏	lack/deprivation	關 係	relation	如 果	if/suppose
修 訂	revise(-sion)	表 示	express/indicate	關 注	attention
任 何	any	反 映	reaction	一 九	nineteen
利 益	benefit	部 門	department	增 加	increase
公 平	fair	本 身	itself	精 神	spirit
工 作	job	特 別	special	數 字	number

Figure 1: Part of the bigram output, with glosses

3.3 Human evaluation

For the first part of the precision evaluation, we relied on human native speakers of Mandarin and Cantonese. Many of the output words, especially domain-specific words and collocations, were not found in the tokenizer dictionary. Most importantly, we are interested in the percentage of output sequences that are legitimate words that can be used to augment the tokenizer.

Four evaluators were instructed to mark whether each entry of the bigram and trigram outputs was a word. The criterion they used was that a word must be able to stand by itself and does not need context to have a meaning. To judge whether 4-gram, 5-gram, 6-gram and m-gram outputs were words, the evaluators were told to consider an entry a word if it

Freq	Collocation
277	.... 副 主 席 先 生 , ...
337	.... 政 府 ....
50	.... 利 益 ....
30	.... 政 府 可 否 告 知 本 局 ....
16	.... 第 五 號 報 告 書 ....
27	.... 明 白 ....
20	.... 私 事 .... 政 府 當 局 .. 訂 究 證 文 . , . 私 ....
16	.... 謹 此 陳 辭 , 支 持 動 議 。 .... 議 員 致 辭 .....
12	.... 工 商 界 ....

Figure 2: Part of the CXtract output

不 公 平	不 知 交 和	供 某 人 政 a21
中 小 學	中 英 雙 方	兩 個 市 政 局
公 務 員	中 國 政 府	券 及 期 貨 事
及 他 方	公 共 援 助	委 員 會 提 出
及 其 他	分 區 直 選	的 生 活 方 式
尤 其 是	支 持 動 議	的 投 票 制 度
大 多 數	文 件 所 載	的 選 舉 制 度
大 故 是	人 權 法 案	的 選 舉 制 度
大 家 都	夾 心 階 層	建 議 修 訂 內
大 部 份	和 會 計 師	很 大 的 缺 失
工 市 政	的 代 表 性	施 政 報 告 中
立 法 局	的 是 人 一	施 政 報 告 內
任 何 人	直 選 議 席	副 主 席 先 生
全 日 制	社 會 人 士	商 務 委 員 會
工 商 界	社 會 福 利	專 責 委 員 會

Figure 3: Part of the trigram, 4-gram and 5-gram output

was a sequence of shorter words that taken together held a conventional meaning, and did not require any additional characters to complete its meaning.

Besides *correct*, the evaluators were given three other categories to place the n-grams. *Wrong* means the entry had no meaning or an incomplete meaning. *Unsure* means the evaluator was unsure. Note that the percentage in this category is not insignificant, indicating the difficulty of defining Chinese word boundaries even by native speakers. *Punctuation* means one or more of the characters was punctuation or ASCII markup.

Tables 1 and 2 show the results of the human evaluations. The *Precision* column gives

一九九七年後	after the year 1997
人權法案條例	Human Rights Bill
中英聯合聲明	Sino-British Joint Declaration
以及他們所提	and as they mentioned
加上九七年能	additionally in 1997 we can
本人謹此陳辭	I hereby move that
刑事罪行條例	Criminal Law Bill
至一九九七年	until the year 1997
我們必須審刻	we must examine
我們應該予支	we should allocate
我們應該糾支	we should correct
見是民上是內	*error*
兩個市政局的	two Urban Council's
券及期貨事務	*error* and Commodity Affairs
動議投贊成票	move to cast a supporting vote
第五號報告書	The Number 5 Report
港特別行政區	Hong Kong Special Administrative Region
機場核心工程	Airport Core Project
選舉委員會的	of the Elective Committee
總督施政報告	The Executive Report of the Governor
謹此提出動議	hereby beg to move

Figure 4: Part of the 6-gram output, with glosses

the percentage correct over total n-grams in that category.

We found some discrepancies between evaluators on the evaluation of *correct* and *unsure* categories. Most of these cases arose when an n-gram included the possessive 的 (*de*), or the copula 是 (*shi*). We also found some disagreement between evaluators from mainland China and those from Hong Kong, particular in recognizing literary idioms.

The average precision of the bigram output was 78.13%. The average trigram precision was 31.3%; 4-gram precision 36.75%; 5-gram precision 49.7%; 6-gram precision 55.2%; and the average m-gram precision was 54.09%.

3.4 Dictionary/text evaluation

The second part of the evaluation was to compare our output words with the words actually present in the text. This gives the recall and precision of our output with respect to the training corpus. Unfortunately, the training corpus is untokenized and too large to tokenize by hand. We therefore estimated the words in the training corpus by passing it through an automatic tokenizer based on the BDC dictionary (BDC 1992). Note that this dictionary's

一九九二年十月
一九九二年六月
一九九二至九三年度
一九九五年選舉
已是中英聯合聲明
支持麥理覺議員的
支持麥理覺議員的修訂
支持麥理覺議員的修訂動議
支持麥理覺議員的修訂動議
支持麥理覺議員的動議
由現在至一九九七年
在委員會審議階段
多及立法局改革委員會
委員會審議階段
持麥理覺議員的
政府可否告知本局
修訂動議經向委員會提出
選舉事宜專責委員會報告書
選舉委員會的成員
選舉委員會的組成
總督在施政報告
總督的施政報告
總督施政報告中
總督商務委員會
總督彭定康先生支
議員的修訂動議
議員致辭的譯文

Figure 5: Part of the m-gram output

entries were not derived from material related to our corpus. The tokens in the original tokenized text were again sorted into unique bigrams, trigrams, 4-grams, 5-grams, 6-grams, and m-grams with $m > 6$. Table 3 summarizes the precision, recall, and augmentation of our output compared to the words in the text as determined by the automatic tokenizer. *Precision* is the percentage of sequences found by CXtract that were actually words in the text. *Recall* is the percentage of words in the text that were actually found by CXtract. *Augmentation* is the percentage of new words found by CXtract that were judged to be correct by human evaluators but were not in the dictionary.

The recall is low because CXtract does not include n-grams with frequency lower than 8. However, we obtained 467 legitimate words or collocations to be added to the dictionary

Table 1: Human Evaluation of the Bigram Output Precision

Evaluator	wrong	unsure	punctuation	precision
A	339 20%	53 3.1%	111 6.5%	 75.2%
B	264 15.6%	31 1.8%	111 6.5%	 81.4%
C	269 15.87%	118 6.96%	111 6.5%	 75.6%
D	289 17%	23 1.4%	111 6.5%	 80.3%

Table 2: Human Evaluation of n-gram Output Precision

Evaluator	n	wrong	correct	unsure	punctuation	precision
A	3	205	81	33	25	23.5%
	4	98	89	5	20	44.1%
	5	33	48	1	6	54.5%
	6	9	32	5	1	68%
	m	14	32	3	0	65.3%
D	3	296	101	23	25	29.4%
	4	102	75	2	23	37.13%
	5	36	44	2	6	50%
	6	20	26	0	1	55.3%
	m	18	27	0	4	55.1%
E	3	168	134	16	26	39%
	4	89	81	10	22	40.1%
	5	29	44	5	10	50%
	6	10	0	11	1	53.2%
	m	12	0	7	4	53.06%
C	3	210	112	0	22	32.6%
	4	131	52	0	19	25.7%
	5	40	39	0	9	44.3%
	6	25	21	0	1	44.7%
	m	24	21	0	4	42.9%

and the total augmentation is 5.73%. The overall precision is 59.3%.

However, we believe the frequency threshold of 8 was too low and the 585K character size of the corpus was too small. Most of the "garbage" output had low frequencies. The precision rate can be improved by using a larger data base and raising the threshold as in

Table 3: Precision, Recall and Augmentation of CXtract Output

n	Token types	CXtract	Precision	Recall	Augmentation
2	6475	1201	852 (70.9%)	662 (10.2%)	190 (2.9%)
3	721	344	115 (33.4%)	10 (1.4%)	105 (14.6%)
4	911	202	75 (37.1%)	7 (0.008%)	68 (7.5%)
5	38	88	43 (48.9%)	0 (0%)	43 (113.2%)
6	7	47	29 (61.7%)	0 (0%)	29 (414.2%)
m	4	49	32 (65.3%)	0 (0%)	32 (800%)
Total	8156	1931	1146 (59.3%)	769 (14%)	467 (5.73%)

Experiment II.

In the following sections, we discuss the significance of the evaluated results.

3.5 Bigrams are mostly words

Using human evaluation, we found that 78% of the bigrams extracted by our tool were legitimate words (as compared with $70.9\% + 2.9\% = 73.8\%$ by evaluation against the automatic tokenizer's output). Of all n-gram classes, the evaluators were least unsure of correctness for bigrams, although quite a few classical Chinese terms were difficult for some of the evaluators.

Since the corpus is an official transcript of formal debates, we find many terms from classical Chinese which are not in the machine-readable dictionary, such as 謹此 (*jin ci*, "I hereby").

Some of the bigrams are acronymic abbreviations of longer terms that are also domain specific and not generally found in a dictionary. For example, 中英 (*zhong ying*) is derived from 中國,英國 (*zhong guo, ying guo*), meaning Sino-British. This acronymic derivation process is highly productive in Chinese.

3.6 The whole is greater than the sum of parts

What is a *legitimate* word in Chinese? To the average Chinese reader, it has to do with the vocabulary and usage patterns s/he acquired. It is sometimes disputable whether 立法局 (*li fa ju*, "Legislative Council") constitutes one word or two. But for the purposes of a machine translation system, for example, the word 局 (*ju*) may be individually translated not only into "Council" but also "Station", as in 警察局 (*jing cha ju*, "Police Station"). So we might incorrectly get "Legislative Station". On the other hand, 立法局 (*li fa ju*) as a single lexical item always maps to “Legislative Council”

Another example is 大部份 (*da bu fen*) which means “the majority”. Our dictionary omits this and the resulting tokenization is 大 (*da*, “big”) and 部份 (*bu fen*, “part/partial”). It is clear that “majority” is a better translation than “big part”.

3.7 Domain specific compounds

Many of the n-grams for $n > 3$ found by CXtract are domain-specific compounds. For example, due to the topics of discussion in the proceedings, “the year 1997” appears very frequently.

Longer terms are frequently abbreviated into words of three or more characters. For example, 中英雙方 (*zhong ying shuang fang*) means “bilateral Sino-British”, and 中英聯合聲明 (*zhong ying lian he sheng ming*) means “Sino-British Joint Declaration”. Various titles, committee names, council names, projects, treaties, and joint-declarations are also found by our tool. Examples are shown in Figure 6.

Although many of the technical terms are a collocation of different words and sometimes acceptable word boundaries are found by the tokenizer, it is preferable that these terms be treated as single lexical items for purposes of machine translation, information retrieval, or spoken language processing.

3.8 Idioms and *cheng yu*

From n-gram output where $n > 3$, we find many idiomatic constructions that could be tokenized into series of shorter words. In Chinese especially, there are many four character words which form a special idiomatic class known as 成語 (*cheng yu*). There are dictionaries of *cheng yu* with all or nearly all entries being four character idioms (e.g., Chen & Chen 1983). In the training corpus we used, we discovered new *cheng yu* that were invented to describe a new concept. For example, 夾心階層 (*jia xin jie ceng*) means “sandwich class” and is a metaphorical term for families who are not well off but with income just barely too high to qualify for welfare assistance. Such invented terms are highly domain dependent, as are the usage frequencies of established *cheng yu*.

3.9 Names

Tokenizing Chinese names is a difficult task (Sproat *et al.* 1994) because Chinese names start with a unigram or bigram family name, and are followed by a given name freely composed of one or two characters. The given name usually holds some meaning, making it hard

白 皮 書	White Paper
行 政 局	Executive Council
工 商 界	Industry and Trade
立 法 局	Legislative Council
保 安 司	Security Secretary
財 政 司	Financial Secretary
經 濟 司	Economics Secretary
聯 合 聲 明	Joint Declaration
一 九 九 七 年	the year 1997
立 法 局 選 舉	Election of the Urban Council
商 務 委 員 會	Commerce and Trading Committees
專 責 委 員 會	Select Committees
選 舉 委 員 會	Elective Committees
醫 院 管 理 局	Hospital Administration Committee
警 務 處 處 長	Police Chief
人 權 法 案 條 例	Human Rights Bill
中 英 聯 合 聲 明	Sino-British Joint Declaration
刑 事 罪 行 條 例	Criminal Law Bill
多 議 席 單 票 制	many-seats one-vote system
港 特 別 行 政 區	Hong Kong Special Administrative Region
機 場 核 心 工 程	Airport Core Project
計 由 現 在 至 一 九 九 七 年	counting from now until the year 1997
由 現 在 至 一 九 九 七 年	from now until the year 1997
委 員 會 審 議 階 段	examining period of the committee
教 育 統 籌 司 員 會	Education Commission
總 督 商 務 委 員 會	Trading Committee of the Governor

Figure 6: Some domain specific terms found by CXtract, with glosses

to distinguish names from other words. For names, we do not want to tokenize them into separate characters. In a large corpus, names are often frequently repeated. For example, in our data, the names of some parliamentary members are extracted by our tool as separate lexical items. Examples are shown in Figure 7. The last two characters of each example are the person's title.

4 Experiment II: Tokenization Improvement

Given the significant percentage of augmented words in Experiment I, we can see that many entries could be added to the dictionary used for automatic tokenization.

In the next stage of our work, we used a larger portion of the corpus to obtain more

Chinese words and collocations, and with higher reliability. These items were converted into dictionary format along with their frequency information.

To obtain a baseline performance, the tokenizer was tested with the original dictionary on two separate test sets. It was then tested with the statistically-augmented dictionary on the same test sets. Each of the tokenization outputs was evaluated by three human evaluators.

李柱銘議員
李華明議員
林鉅津議員
彭定康先生
馮檢基議員
黃宏發議員
詹培忠議員
劉慧卿議員
譚耀宗議員
涂謹申議員
周梁淑怡議員
麥理覺議員的

Figure 7: Some names and titles found by CXtract

4.1 Procedure

As training data we used about 2 million Chinese characters taken from the same HKUST corpus. This is about 4 times the size used in Experiment I. The tokenizer we used employs a maximal matching strategy with frequency preferences.

The original dictionary for the tokenizer holds 104,501 entries and lacks many of the domain-specific and regional words found in the corpus.

From the first stage of CXtract, we obtained 4,196 unique adjacent bigrams. From the second stage, we filtered out any CXtract output that occurred less than 11 times and obtained 7,121 lexical candidates. Additional filtering constraints on high-frequency characters were also imposed on all candidates.[1] After all automatic filtering, we were left with 5,554 new dictionary entries.

Since the original dictionary entries employed frequency categories of integer value from 1 to 5, we converted the frequency for each lexical item from the second stage output to

[1]A refined version of the linguistic filtering is discussed in Wu & Fung (1994).

this same range by scaling. The adjacent bigrams from the first stage were assigned the frequency number 1 (the lowest priority).

The converted CXtract outputs with frequency information were appended to the dictionary. Some of the appended items were already in the dictionary. In this case, the tokenization process uses the higher frequency between the original dictionary entry and the the CXtract-generated entry.

The total number of entries in the augmented dictionary is 110,055, an increase of 5.3% over the original dictionary size of 104,501.

4.2 Results

Two independent test sets of sentences were drawn from the corpus by random sampling with replacement. TESTSET I contained 300 sentences, and TESTSET II contained 200 sentences. Both sets contain unretouched sentences with occasional noise and a large proportion of unknown words, i.e., words not present in the original dictionary. (Sentences in the corpus are heuristically determined.)

Each test set was tokenized twice. *Baseline* is the tokenization produced using the original dictionary only. *Augmented* is the tokenization produced using the dictionary augmented by CXtract.

Three human evaluators evaluated each of the test sets on both baseline and augmented tokenizations. Two types of errors were counted: false joins and false breaks. A false join occurs where there should have been a boundary between the characters, and a false break occurs where the characters should have been linked. A conservative evaluation method was used, where the evaluators were told to not to mark errors when they felt that multiple tokenization alternatives were acceptable.

The results are shown in Tables 4, 5, and 6. Baseline error is computed as the ratio of the number of errors in the baseline tokenization to the total number of tokens found. Augmented error is the ratio of the total number of errors in the augmented tokenization to the total number of tokens found.

Our baseline rates demonstrate how sensitive tokenization performance is to dictionary coverage. The accuracy rate of 76% is extremely low compared with other reported percentages which generally fall around the 90's (Chiang *et al.* 1992; Lin *et al.* 1992; Chang & Chen 1993; Lin *et al.* 1993). We believe that this reflects the tailoring of dictionaries to the particular domains and genres on which tokenization accuracies are reported. Our experiment, on the other hand, reflects a more realistic situation where the dictionary and

Table 4: Result of TESTSET I - 300 sentences

Eval-uator	# tokens	Baseline # errors	Error rate	Accu-racy	# tokens	Augmented # errors	Error rate	Accu-racy
A	4194	1128	27%	73%	3893	731	19%	81%
F	4194	1145	27%	73%	3893	713	18%	82%
G	4194	1202	29%	71%	3893	702	18%	82%

Table 5: Result of TESTSET II - 200 sentences

Eval-uator	# tokens	Baseline # errors	Error rate	Accu-racy	# tokens	Augmented # errors	Error rate	Accu-racy
A	3083	737	24%	76%	2890	375	13%	87%
H	3083	489	16%	84%	2890	322	11%	89%
I	3083	545	18%	82%	2890	339	12%	88%

Table 6: Average accuracy and error rate over all evaluators and test sets

Experiment	Total # tokens	Average error	Error rate	Accuracy
Baseline	7277	1749	24%	76%
Augmented	6783	1061	16%	84%

text are derived from completely independent sources, leading to a very high proportion of missing words. Under these realistic conditions, CXtract has shown enormous utility. The error reduction rate of 33% was far beyond our initial expectations.

5 Conclusion

We have presented a statistical tool, CXtract, that identifies words without supervision on untagged Chinese text. Many domain-specific and regional words, names, titles, compounds, and idioms that were not found in our machine-readable dictionary were automatically extracted by our tool. These lexical items were used to augment the dictionary and to improve tokenization.

The output was evaluated both by human evaluators and by comparison against dictionary entries. We have also shown that the output of our tool helped improve a Chinese tokenizer performance from 76% to 84%, with an error reduction rate of 33%.

6 Acknowledgement

We would like to thank Kathleen McKeown for her support and advice, and Frank Smadja and Chilin Shih for helpful pointers. We would also like to thank our evaluators, Philip Chan, Eva Fong, Duanyang Guo, Zhe Li, Cindy Ng, Derek Ngok, Xuanyin Xia, and Michelle Zhou. The machine-readable dictionary (BDC 1992) was provided by Behavior Design Corporation.

References

BDC. 1992. *The BDC Chinese-English electronic dictionary (version 2.0)*. Behavior Design Corporation.

Chang, Chao-Huang & Cheng-Der Chen. 1993. HMM-based part-of-speech tagging for Chinese corpora. In *Proceedings of the Workshop on Very Large Corpora*, 40–47, Columbus, Ohio.

Chen, Yong-zhen & Spring Chen. 1983. *Chinese idioms and their English equivalents.* Hong Kong: Shang Wu Yin Shu Ju.

Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin, & Keh-Yih Su. 1992. Statistical models for word segmentation and unknown resolution. In *Proceedings of ROCLING-92*, 121–146.

FDMC. 1986. *Xiandai hanyu pinlu cidian (Frequency dictionary of modern Chinese).* Beijing Language Institute Press.

Lin, Ming-Yu, Tung-Hui Chiang, & Keh-Yih Su. 1993. A preliminary study on unknown word problem in Chinese word segmentation. In *Proceedings of ROCLING-93*, 119–141.

Lin, Yi-Chung, Tung-Hui Chiang, & Keh-Yih Su. 1992. Discrimination oriented probabilistic tagging. In *Proceedings of ROCLING-92*, 85–96.

Liu, Y. 1987. New advances in computers and natural language processing in China. *Information Science*, 8:64–70. In Chinese.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Sproat, Richard, Chilin Shih, William Gale, & N. Chang. 1994. A stochastic word segmentation algorithm for a Mandarin text-to-speech system. In *Proceedings of the*

32nd Annual Conference of the Association for Computational Linguistics, Las Cruces, New Mexico. To appear.

WU, DEKAI. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, Las Cruces, New Mexico. To appear.

WU, DEKAI & PASCALE FUNG, 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. Submitted.

WU, ZIMIN & GWYNETH TSENG. 1993. Chinese text segmentation for text retrieval: Achievements and problems. *Journal of The American Society for Information Science*, 44(9):532–542.