Using Word Signature Features for Terminology translation from Large Corpora

Pascale N. Fung

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

1997

Acknowledgments

I want to thank my parents, Luen-Wah Hui and Yuk-Sung Fung, for their unconditional love, sacrifice and support throughout my life. I want to thank them for their patience in answering my strangest questions when I was a child, thereby showing me that one should never stop asking questions and seeking the answers.

I want to thank my advisor, Kathleen McKeown, without whose guidance this thesis would not have been possible. I want to thank her for bringing me into the area of computational linguistics and taught me my first course in natural language processing. I want to thank her for always asking me to probe further whenever I was getting impatient with ... In addition, I want to thank her for her moral support during the personal crisis I have lived through as a graduate student.

I also owe a great deal to my other advisor, Ken Church. I want to thank him for introducing me to statistical natural language processing. I want to thank him...

I want to thank many of my mentors and teachers: Tatsuyo Kawahara for teaching me the basic knowledge of speech and pattern recognition, and for bringing me to the world of research; Richard Schwartz, for teaching me how to approach problems with ... ; Fred Juang and Chin-Hui Lee for ...., I want to thank my committee members, Al Aho for the most stimulating and interesting technical discussions, Shree Nayar for many interesting discussions and many great advices throughout my graduate school years, Mukesh Dalal for his support, both

Pascale N. Fung

COLUMBIA UNIVERSITY

May 1997

Using Word Signature Features

for Terminology translation from Large Corpora

Pascale N. Fung

Abstract

Translating technical and domain-specific terminology is one significant component of machine translation and machine-aided translation systems. Domain-specific terms are often not found in standard dictionaries. Human translators, not being experts in every technical or regional domain, need help for producing their translations more effectively. Automatic extraction of terminology translation is therefore highly desirable. A system which produces translation candidates can also be helpful to human translators.

This thesis describes a statistical approach, augmented with linguistic knowledge, to terminology translation. Most other statistical work on terminology translation use occurrence information as word features from clean parallel corpora. Given a sentence-aligned bilingual text, occurrence patterns of words are extracted and matched between pairs using similarity measures. However, parallel corpora

of clean, translated texts are hard to come by whereas there are more noisy, translated texts and many more monolingual texts in various domains. Word occurrence feature is susceptible to translation or OCR noises, and is not applicable to monolingual texts of the same domain.

In this thesis, we focus on extracting discriminatory statistical word signature features robust to noisy parallel texts and same-domain, non-parallel texts. We propose a translator-aid tool for term translation using noisy parallel corpora, and explore the new possibility of using same-domain monolingual texts of a pair of languages for translation extraction.

We obtain discriminatory statistical information which is related to the lexical properties of a term and its translation in any other language of the same domain. These lexical "signatures" are extracted from the corpora and are represented in vector form. In this thesis, we present two major types of such signature features: (1) the non-linear K-vec and dynamic K-vec (DK-vec) features from noisy parallel corpora, and (2) the Word Relation Matrix (WoRM) feature from non-parallel corpora. We will also present how we use matching techniques such as Dynamic Time Warping with similarity measures such as Euclidean distance and the Cosine Measre for finding associated bilingual word vector pairs. Compared to statisical word features employed in other work, our features are more robust to the syntactic structure, character sets, the language of the text, and to the domain.

Since most domain-specific terms are not single words in Indo-European languages, and since there are no space delimiters between words or terms in Chinese and Japanese, we need to extract technical terms from texts. We will also show methods for the extraction tasks in this thesis.

For noisy-parallel corpora using type (1) features, our evaluations have shown a 55.35% precision from a small AWK Japanese/English corpus and 89.93% precision from HKUST Chinese/English corpus in terms of word translation with best candidate, and a 47% improvement for human translators using our output, while translating domain-specific terms. For type (2) features, we have done some initial experiments in matching words with their translations in an English/English Wall Street Journal non-parallel corpus and an English-WSJ/Japanese-Nikkei non-parallel corpus. The evaluation results of using the Word Relation Matrix alone show a precision of about 30%. We have shown that this is nevertheless a useful result since human translation is improved almost two-fold by using this output as translation suggestions. Our results indicate a promising direction for exploring the vast quantity of online non-parallel material for terminology translation. We also discuss possible extensions to improve these results and potential applications of our features in other areas of natural language processing.

Contents

List of Tables

List of Figures

Chapter 1

Extracting technical terms from Japanese and Chinese texts

"The lexicon of a special subject language reflects the organizational characteristics of the discipline by tending to provide as many lexical units as there are concepts conventionally established in the subspace and by restricting the reference of each such lexical unit to a well-defined region. Besides containing a large number of items which are endowed with the property of special reference, the lexicon of a special language also contains items of general reference which do not usually seem to be specific to any discipline or disciplines and whose referential properties are uniformly vague or generalized. The items which are characterized by special reference within a discipline are the *terms* of that discipline, and collectively they form its *terminology*; those which function in general reference over a variety of sublanguages are simply called *words*, and their totality the *vocabulary*." [13].

Linguists have argued that the smallest translatable unit is often not a single word, as defined by a string of letters delimited by spaces, but a phrase (or a term) [12, 13]. A single word is often polysemous and is very likely mapped to multiple words in another language. On the other hand, a phrase or a term has specific meaning and often maps to only one term in the other language in the same domain. This notion applies both to English and to Asian languages with no space delimiters between words such as Chinese and Japanese.

Finding technical or domain-specific lexical entries for Chinese is hampered by a particularly obscure distinction between characters, morphemes, words, and compounds. Characters in Chinese (and in some other Asian languages such as Japanese) are not separated by spaces to delimit terms such as 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. Combinations of multiple characters form more terse lexical units, corresponding to either single word or multi-word terms in languages like English. Therefore, for lexical information acquisition, what is useful is not simply segmenting Chinese or Japanese into tokens corresponding to any definition of a word delimited by spaces, but rather extracting relevant lexical units from unsegmented texts.

In this chapter, we first describe a tool we have developed for extracting Chinese terms, CXtract, based on statistical information with morphological filters. Then, we show a technical term extractor we developed for Japanese texts, JBrat. This tool is rule-based, using morphosyntactic information about Japanese technical terms. The particular choice of different approaches for the two languages are based on their different linguistic properties. We advocate a predominantly lexical approach for Chinese, and a predominantly grammatical approach for Japanese. In the future, it is likely a combination of the two approaches with different weightings for Chinese and Japanese can have better performances.

1.1 Extracting terms from unsegmented Chinese texts

Despite a particularly difficult distinction between characters, morphemes, words, and compounds, some characteristics of Chinese lexical units 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* [3], 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 [9] showed that in general, 75% of Chinese words are bigrams, 14% trigrams, 6% n-grams with $n > 3$. In addition, just as in any other language, regional and domain-specific technical terms are likely to appear repeatedly in a large corpus.

Using a lexical approach, we extended and re-targeted a tool originally designed for extracting English compounds and collocations, Xtract[14], to find words in Chinese. We call the resulting tool CXtract. Terms found by CXtract are used to augment our dictionary. These terms either map to single words in European languages or multi-word terms.

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 are 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.

1.1.1 Overall algorithm

More precisely, the procedure and exact parameters we used were as follows:

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

2: From the unigram list, we found all the sentences containing each unigram. From these sentences, we extract all characters associated with each of the unigrams and obtained a list of all bigrams found. For example, we extract all words associated with 立. These bigrams each have a data structure containing their frequency information, distance between 立 and the other word, histogram of the frequency of appearances of the two words in given positions.

3: We kept only bigrams which occur significantly more than chance expectation, and which appear in a rigid way [14] as represented by the following formula:

$$
\begin{aligned}
\text{strength} &= \frac{\text{freq} - \overline{f}}{\sigma} \geq k_0 \\
\text{spread} &\geq U_0 \\
p_j^i &\geq \overline{p}_i + (k_1 x \sqrt{U_i} \\
\text{where} \quad \text{freq} &= \text{frequency of the bigram with character } li \\
\overline{f} &= \text{average frequency of all bigrams with } li \\
\sigma &= \text{standard deviation around } \overline{f} \\
k_0 &= \text{threshold for the strength of the bigram} \\
p_i^j &= \text{peak at position } i \\
U_i &= \frac{\sum_{j=1}^{10} p_i^j - \overline{p}_i^{\,2}}{10}
\end{aligned}
$$

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 terms 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.

The output of CXtract is in general domain specific and not found in common dictionaries. These are the terms for which automatic translation or machine-aided translation are most useful.

A portion of the list of bigrams obtained from (4) is shown in Figure 1.1.1. We obtained 1695 such bigrams after thresholding.

Part of the output from (5) is shown in Figure 1.2. 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 1.3, 1.4, 1.5, and 1.6.

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 1.1: Part of the bigram output, with glosses

Freq	Collocation
277	副 主 席 先 生 (Mr. Vice President) ,
337	政 府 (government)
50	利 益 (profit)
30	政 府 可 否 告 知 本 局 (could the government please tell me)
16	第 五 號 報 告 書 (the Number 5 report)
27	明 白 (understand)
20	私 事 (private affairs) ... 政 府 當 局 (government) .. 訂 究 證 文 (review of evidence document) . , . 私 (private) ...
16	謹 此 陳 辭 (hereby move that) , 支 持 動 議 。 (support the motion) ... 議 員 致 辭 (legislative member speaks)
12	工 商 界 (the industry)

Figure 1.2: Part of the CXtract output, with glosses in brackets.

不 公 平	unfair	不 知 交 和	
中 小 學	middle and elementary schools	中 英 雙 方	China and Britain
公 務 員	civil servants	中 國 政 府	Chinese government
及 他 方	and other aspects	公 共 援 助	public aid
及 其 他	and others	分 區 直 選	district direct election
尤 其 是	especially	支 持 動 議	support motion
大 多 數	majority	文 件 所 載	according to the document
大 家 都	everybody all	夾 心 階 層	sandwich class
大 部 份	most part	和 會 計 師	and accountant
工 市 政	work/city/politics	的 代 表 性	representation of
立 法 局	Legislative Council	的 是 人 一	
任 何 人	anyone	直 選 議 席	direct electorial seat
全 日 制	full-day system	社 會 人 士	people from the society
工 商 界	Industrial-commercial areas	社 會 福 利	social welfare

Figure 1.3: Part of the trigram and 4-gram output, with glosses

供 某 人 政 a21	
兩 個 市 政 局	two Urban Councils
券 及 期 貨 事	stocks and bonds
委 員 會 提 出	committe proposes
的 生 活 方 式	life style of
的 投 票 制 度	voting system of
的 選 舉 制 度	election system of
建 議 修 訂 内	within proposal amendment
很 大 的 缺 失	a huge lack of
施 政 報 告 中	in government report
施 政 報 告 内	inside government report
副 主 席 先 生	Mr. Vice President
商 務 委 員 會	Trade Committee
專 責 委 員 會	special duty committee

Figure 1.4: Part of the 5-gram, with glosses

1.1.2 Evaluation of the statistical output

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

Tables 1.1 and 1.2 show the results of the human evaluations. The *Precision* column gives the percentage correct over total n-grams in that category.

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 1.5: Part of the 6-gram output, with glosses

Table 1.1: 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%

一 九 九 二 年 十 月	October 1992
一 九 九 二 年 六 月	June 1992
一 九 九 二 至 九 三 年 度	from 1992 to 93
一 九 九 五 年 選 舉	election in 1995
已 是 中 英 聯 合 聲 明	already is Sino-British Joint Decleration
支 持 麥 理 覺 議 員 的	support McGregore's amendment
支 持 麥 理 覺 議 員 的 動 議	support McGregore's motion
由 現 在 至 一 九 九 七 年	from now until 1997
在 委 員 會 審 議 階 段	during committee review period
多 及 立 法 局 改 革 委 員 會	thanks to Legislative Council reform committee
委 員 會 審 議 階 段	commitee review period
政 府 可 否 告 知 本 局	could the government tell me
修 訂 動 議 經 向 委 員 會 提 出	amendment motion has been aleady proposed to the committee
選 舉 事 宜 專 責 委 員 會 報 告 書	report of the special task committee on election matters
選 舉 委 員 會 的 成 員	members of the election committee
選 舉 委 員 會 的 組 成	establishing the election committee
總 督 的 施 政 報 告	Governor's executive report
總 督 施 政 報 告 中	in the Governor's executive report
總 督 商 務 委 員 會	Governor trading committee
總 督 彭 定 康 先 生 支	Govrnor Chris Patten
議 員 的 修 訂 動 議	member's amendment motion
議 員 致 辭 的 譯 文	translation of the member's speech

Figure 1.6: Part of the m-gram output, with glosses

Table 1.2: 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%

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 1.7.

1.1.3 Evaluation of CXtract output for the tokenization task

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 [2]. Note that this dictionary's 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 1.3 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.

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 1.7: Some domain specific terms found by CXtract, with glosses

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.

Table 1.3: Precision, Recall and Augmentation of CXtract Output

n-gram	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%)

Given the significant percentage of augmented words in the above experiment, we can see that many entries could be added to the dictionary used for automatic tokenization. Indeed, in [4], we reported using about 2 million Chinese characters taken from the same HKUST corpus to augment a dictionary with 5,554 new dictionary entries, mostly technical and domain-specific terms. We reported that the output of CXtract helped improve a Chinese tokenizer performance from 76% to 84%, with an error reduction rate of 33%.

1.1.4 Morphological filters for CXtract

CXtract initially produces excellent recall at the expense of low precision. Ideally, the false candidates should be rejected by some automatic means, without eliminating valid lexical entries. To this end, we investigated a set of simple filters based on linguistic principles. The morphological and syntactic filters are shown in Figure 1.8 and Figure 1.9.

	Type (see text)
Number markers	
1. 們 . . . (*mén*, [plural])	B
Numeral markers	
2. . . . 第 (*dì*, [ordinal])	B
3. 次 . . . (*cì*, [instance])	B
Case markers	
4. . . . 的 . . . (*de*, [genitive/associative])	A/B
Aspect markers	
5. 了 . . . (*le*, [perfective])	B
6. . . . 在 (*zài*, [durative])	B

Figure 1.8: Morphological filters for CXtract

The filtering process first segments the candidate using the original dictionary, to identify the component words. It then applies a number of morphological and syntactic constraints to eliminate (A) and sequences that should remain multiple segments and (B) ill-formed sequences.

We reject a hypothesized lexical entry if it matches any filtering pattern. We treat the particular characters in these filters as morphological affixes. We can then reject any sequence with the affix on the wrong end (the special case of the genitive 的 (*de*) is considered below). Because morphemes such as the plural marker 們 (*mén*) or the instance marker 次 (*cì*) are suffixes, we can eliminate candidate sequences that begin with them. Similarly, we can reject sequences that end with the ordinal prefix 第 (*dì*) or the preverbial durative 在 (*zài*).

Applying these morphological filters to CXtract output, we improved tokenization results significantly. A summary of all the precision rates is shown for comparison in Figure 1.10. We used eight human evaluators for a nk-Blind evaluation as described in [16]. The precision numbers under k where k is from one to eight are obtained from counting the agreement between the judges. When $k = 1$, it means a machine segmentation is correct if at least one human judge segments it in the same way. When $k = 1$, it means all eight judges have to agree with the machine output. Comparing the precisions over the baseline segmentation evaluation without CXtract output, with CXtract output, and with additional morphological filters on CXtract output, we see that tokenization using the (unfiltered) augmented dictionary is consistently improved over the baseline by a substantial margin. With filtering, performance improves again significantly as shown in Figure 1.10.

1.2 Japanese technical term extractor–JBrat

We have seen that tools based on lexical approaches such as CXtract are very useful for extracting technical or domain-specific terms, especially from unsegmented texts. We have also seen that even minimal morphosyntactic information helped increase the performance of a tokenizer. CXtract is especially useful for Chinese because there is really little morphosyntactic information for technical or domain terms.

When we work with Japanese texts, we realize that another approach for extracting technical terms from Japanese texts might be appropriate. We choose the approach of using morphosyntactic information, augmented with an unknown word model, for Japanese technical term extraction. Our approach is motivated by

		Type (see text)
Prepositions/coverbs		
1.	對 . . . (*dùi*, towards)	A
2.	在 . . . (*zài*, at)	A
3.	. . . 在 (*zài*, at)	B
4.	到 . . . (*dào*, to)	A
5.	. . . 到 (*dào*, to)	B
6.	. . . 內 (*nèi*, inside)	A
7.	. . . 後 (*hòu*, after)	A
Conjunctions		
8.	和 . . . (*hé*, and)	A
9.	. . . 和 (*hé*, and)	A
10.	與 . . . (*yú*, and)	A
11.	. . . 與 (*yú*, and)	A
12.	但 . . . (*dàn*, but)	A
13.	. . . 但 (*dàn*, but)	A
Modals		
14.	應 . . . (*yīng*, should)	A
Adverbs		
15.	再 . . . (*zài*, again)	A
16.	將 . . . (*jiāng*, will [future])	A
17.	. . . 已 (*yǐ*, already)	B
18.	. . . 曾 (*céng*, once [used to be])	B
19.	. . . 不 (*bù*, not])	B
Promiscuous verbs		
20.	是 . . . (*shì*, be [copula])	A
21.	. . . 是 (*shì*, be [copula])	A
22.	有 . . . (*yǒu*, have)	A
23.	使 . . . (*shǐ*, make/cause/enable)	A
Verbs taking sentential complements		
24.	想 . . . (*xiǎng*, think)	A
25.	. . . 問 (*wèn*, ask)	A
Pronouns		
26.	. . . 我 . . . (*wǒ*, I)	A
27.	. . . 你 . . . (*nǐ*, you)	A
28.	. . . 您 . . . (*nín*, you [formal])	A

Figure 1.9: Closed-class syntactic filters for CXtract

Experiment	k							
	1	2	3	4	5	6	7	8
baseline	85.02	77.69	72.97	68.09	62.15	55.69	47.64	28.54
CXtract	89.01	84.85	81.27	75.49	69.42	63.21	53.99	32.97
filtered CXtract	91.71	88.72	85.32	80.24	73.54	65.84	55.78	32.76

Figure 1.10: Comparison of nk-Blind Precision Percentages for CXtract Output

the following reasons:

- We can make use of a freely available Japanese POS tokenizer/tagger, JUMAN. Unlike Chinese, where the text was not tokenized nor tagged before applying CXtract, we have a tokenized and tagged text as input to the term extractor. This makes the term extraction task in Japanese more similar to that in English, than that in Chinese.

- We want to use the output of the term extractor for compiling bilingual term dictionaries. Thus precision is relatively important. Minimal post-filtering is desirable. Since we focus on term extraction rather than tokenization, statistically significant non-technical collocations are not as useful. On the other hand, technical terms which occur infrequently should also be included if possible.

- We believe that syntactic and morphological patterns for Japanese technical terms are quite regular and can be learned easily from a small sample set. Japanese technical terms can be distinguished from non-technical terms with simple syntactic and morphological filtering. This is the distinctive advantage for Japanese techincal term extraction over that of English or Chinese.

- By incorporating an unknown word model, we can relax the dependency on prior knowledge of the tool. This will allow us to take advantage of both lexical and syntactic approaches.

This section describes some findings of morphological and syntactic information for Japanese technical terms from both a bilingual corpus and published technical dictionaries. We describe how the extractor learns such knowledge. Note that we chose to learn from a bilingual corpus because we plan to use the Japanese technical term extractor together with an English extractor for a term translation project. Learning from a bilingual corpus can give us more insight into the various characteristics of technical terms in both languages. The findings we present in this paper are applicable to monolingual tasks as well and can also be learned from a monolingual Japanese corpus. We also discuss the output of the term extractor and its applications. It is counter-intuitive that although Japanese is generally more difficult than English for NLP systems, the particular task of technical term extraction turns out to be relatively easy in Japanese.

1.2.1 Overall algorithm

The algorithm is as follows:

Training:

1. Tag both texts of a small English/Japanese bilingual corpus
2. Extract English NPs from the English side of the corpus
3. Select English technical terms from English NPs

4. Align Japanese translations to English technical terms manually

5. Sort frequency distribution of tag sequences of these Japanese technical terms

6. Obtain regular expressions which cover tag sequence patterns of Japanese technical terms

Application:

1. Tokenize and tag a Japanese test corpus

2. Use a program with learned regular expressions to extract technical terms from the test corpus

3. Filter the output with any of the frequency, morphology or string length constraints as needed in the application

1.2.2 Using a Japanese tokenizer and tagger

As a preprocess for learning the syntactic and morphological rules of Japanese technical terms, we use JUMAN, a tool developed by Matsumoto and Nagao [10], to perform two tasks: (1) to insert delimiters between Japanese words, and then (2) to label these words with POS tags.

We tokenize and tag both the Japanese translation of the AWK[1] manual [1], and the Nihon Kezai Shimbun (the NIKKEI corpus) [11]. Part of the tagged AWK text is used for training of syntactic regular expression, and the NIKKEI corpus is used for open testing of our term extractor.

[1]Access to the translation of the AWK manual is provided by Bell Laboratories.

Since JUMAN uses a dictionary-based algorithm, it does not recognize any word or term not contained in its dictionaries. When such an unknown word is encountered, it is tokenized into single character sequences, and labelled with the tag *unknown*. Unknown word recognition is one of the biggest limitations of JUMAN. Since many technical terms, especially in domains such as computer science, are new and not likely covered by JUMAN dictionaries, a technical term extractor cannot rely solely on syntactic information from JUMAN.

1.2.3 Some morphosyntactic properties of Japanese

Japanese is often regarded as a more difficult language to process than Indo-European languages such as English. It has many morphological, syntactic and stylistic singularities. Some of the morphosyntactic properties of Japanese pertinent to technical terms are as follows:

- Japanese uses three character sets, each with different morphosyntactic properties:
  - Chinese character set **Kanji**. It is used for words and terms borrowed from China, or composed by Japanese scholars in the Chinese manner. These are always content words, such as nouns, adjectives, verbs, etc.
  - Phonetic set **Hiragana**. It is used for function words, grammatical markers, inflectional endings and some Japanese words not written in *Kanji*.
  - Phonetic set **Katakana**. This is used for words borrowed from foreign languages, except from Chinese. These words are often nouns, sometimes

verbs, but rarely in other syntactic categories.

Some words, such as verbs and adjectives, can be a mixture of character sets. e.g. *Kanji* with *Hiragana* inflections, *Katakana* mixed with *Kanji*. Japanese also uses Roman letters to represent technical terms such as *NAFTA*, without translation[2]

- There are more homonyms than in English. More importantly, there is a significant number of polymorphic words, i.e. the same word written in different characters or character sets[3].

However, we note that most of the syntactic and stylistic complexities of Japanese do not affect technical terms. Character set information in Japanese is a specific type of morphological feature which exists in many Asian languages with ideograms. Indeed, this type of morphological feature in Japanese helps to classify Japanese words into different syntactic categories. Technical terms belong to only a few of these syntactic categories. Consequently, we can take advantage of morphological and syntactic properties of Japanese to render the task of technical term extraction easier. As an example, seeing a multi-character string of *Kanji*s is analogous to seeing, say, the suffix *-tion* of a word in English.

[2]Japanese has a more extensive usage of Roman letters than Chinese where translation or transliteration in Chinese characters are almost exclusively used.

[3]A remote English analogy is *center* written as *Center*, *center*, or *centre*.

1.2.4 Learning morphological patterns from technical dictionaries

In order to verify our knowledge of Japanese technical terms, we consulted several Japanese technical dictionaries [8, 6, 15]. We drew random samples from the dictionaries by taking the first entry on each page of the index, and computed the distribution of character sets shown in Figure 1.11. Each column represents the percentage of terms containing that character set or the nominalizer. A term containing multiple character sets is counted multiple times.

dictionary	Kanji	Katakana	Hiragana	*no*/of	Roman	Total
Financial	85.0%	9.8%	4.2%	6.9%	1.0%	287
Statistics	71.4%	17.1%	4.3%	3.6%	3.6%	140
Computer Science	50.3%	33.2%	4.2%	3.9%	8.4%	310

Figure 1.11: Percentage of terms containing different character sets

Even more than we expected, these findings from technical dictionaries support our intuition about morphosyntactic patterns in the Japanese language.

Dictionary entries are mostly in *Kanji*. This is because most content words, and thus technical words are in *Kanji*. The usage of *Kanji* is the highest in the financial domain dictionary because it is the most traditional and established field among the three.

The second most frequent character set is *Katakana*. It is used for transliteration of imported foreign terms such as *pu-ro-gu-ra-mu/program*. Scientific dictionaries have many more *Katakana* terms than financial domain dictionaries because the former correspond to newer areas. Computer science, being one of the newest technical field, employs the largest number of *Katakana* characters.

The usage pattern of Roman letters is similar to that of *Katakana*. Roman letters are used where European words, especially English words, are employed directly in the Japanese text. Again, the newer the field, the more Roman letters are used. In computer science, words with Roman letters are mostly computer language commands.

One of the most interesting findings is the consistent and low usage of *Hiragana* in all dictionaries. It shows that words with *Hiragana* are in a closed set. Common nouns consisting of only *Hiragana* are not among the technical terms. *Hiragana* in technical terms are mostly the suffix *su-ru/do* or the nominalizer *no*, appended to a *Kanji* or *Katakana* noun. There are very few pure verbs, as opposed to verbal nouns, as entries in the technical dictionaries. The very few verbs, such as *akeru/open*, are highly specific in context when used in the financial domain (it refers to a type of bank transaction bookkeeping rather than the common verb[4]). Simple verbs can have new usages, such as in the previous example, but there is no new invention of simple verbs. New verbal terms are exclusively *Kanji* or *Katakana* verbal nouns with *suru/do*. This limitation of new verbs to verbalized nouns seems to be a common phenomenon in many other languages, including English and French [5].

Technical terms are usually noun phrases, just like in English. Nouns and noun phrases are usually composed of only one character set, where most common Japanese nouns are in *Hiragana*, and more specific nouns in *Kanji* or *Katakana*.

We also found from these dictionaries that most terms are composed of more

[4]For translation purpose, it might be useful to include these terms with their context, instead of as single words.

than two characters and up to thirteen characters. Four character word terms are common and usually consist of two bi-character *Kanji* words. However, they are by no means the most common technical terms, despite the emphasis on analysis of such terms in the literature. From this, we decided that our tool should not be restricted to looking only at two bi-character *Kanji* terms.

From the above observations, we deduce the following properties of the morphosyntactic form of Japanese technical terms:

rule 1	**they contain more than two characters**
rule 2	**they are composed of either all Kanji characters, or all Katakana, or Kanji mixed with Katakana.**
rule 3	**the only Hiragana contained in technical terms are the nominalizer** *no* **and the suffix** *suru/do*
rule 4	**they are not likely to be verbs since most verbs are common words**

1.2.5 Learning syntactic patterns from a bilingual corpus

We then learn the full syntactic pattern of Japanese technical terms. One possible way is to tokenize and tag a monolingual Japanese text, then manually select technical terms from the text, and learn the part-of-speech pattern of these terms. For our eventual application of term translation, we use some bilingual training examples for learning because we are interested not only in the Japanese technical term properties, but the comparative properties of Japanese and English technical terms as well. Since (1) it is reported that most technical terms in English are noun phrases [7], and (2) their translations in Japanese are technical terms as well, regardless of their POS tags, we bootstrap the learning process from the

%	POS tags
15.7%	普通名詞 (Nn) 普通名詞 (Nn)
11.6%	普通名詞 (Nn)
9.6%	普通名詞 (Nn) 未定義語 (Unk)
9.1%	未定義語 (Unk) 未定義語 (Unk)
8.8%	未定義語 (Unk) 普通名詞 (Nn)
5.8%	サ変名詞 (Nv) 普通名詞 (Nn)
3.8%	普通名詞 (Nn) サ変名詞 (Nv)
2.8%	サ変名詞 (Nv) 未定義語 (Unk)
2.3%	未定義語 (Unk) 普通名詞 (Nn) 普通名詞 (Nn)
2.3%	未定義語 (Unk)
1.8%	未定義語 (Unk) サ変名詞 (Nv)
1.0%	名詞接頭辞 普通名詞 (Nn)
1.0%	未定義語 (Unk) 未定義語 (Unk) 未定義語 (Unk)
0.8%	普通名詞 (Nn) 普通名詞 (Nn) 未定義語 (Unk)
0.8%	動詞 (V) 未定義語 (Unk)
0.8%	サ変名詞 (Nv)
0.5%	副詞的名詞 (Na) 普通名詞 (Nn)
0.5%	普通名詞 (Nn) 普通名詞 (Nn) 普通名詞 (Nn)
0.5%	普通名詞 (Nn) 動詞 (V)
0.5%	普通名詞 (Nn) 固有名詞 (Np)

Figure 1.12: Top 20 frequent POS patterns of Japanese technical term POS patterns

Japanese translations of English noun phrases. For this purpose, a simple English NP finder is used to extract noun phrases from a small part of the English AWK manual. These noun phrases are manually aligned to the matching terms in the Japanese AWK translation The bilingual corpus helps to disambiguate certain terms in Japanese, filter out non-technical terms in English NPs, and ensure that many of the *Katakana* terms are indeed technical. Altogether, 396 technical terms are found. The most frequent patterns are shown in Figure 1.12.

The high occurrence of unknown words are due to the fact that AWK, being a computer language manual, contains a large number of *Katakana* terms imported from English. This leads us to believe that in order for our tool to be robust, we need to include unknown word patterns.

Additional syntactic patterns of Japanese technical terms are in rule 5 and rule 6.

The tool was tested without rule 6 on part of the NIKKEI corpus. Some

rule 5	**their POS tag sequence matching the regular expression** $(Adj * (Np * Nn * (NnNm) * Nv*)+)$
rule 6	**for terms containing unknown words their POS tag sequence matches** $(Adj * (Np * Nn * Unk + (NnNm) * Nv*)+)$

自民党	Liberal Democratic Party
社会党	Socialist Party
不動産	real estate
政治改革	political reform
細川首相	Prime Minister Hosokawa
土地税制	The Land Ownership Tax
政治改革法案	Political Reform Bill
衆院予算委員会	Parliamentary Budget Committee
諮問機関	Advisory Committee
社会資本	social capital
成長率	Growth rate
自衛隊	Self Defense Force
経済改革研究会	Economic Reform Study Group
日経平均株価	Nikkei Average Stock Index
難民センター	Refugee Center

Figure 1.13: Some technical terms extracted from Nikkei

examples extracted by our tool are as shown in Figure 1.13. Precision yield of these tests are shown in Figure 1.14.

corpus	size	freq	no. terms-Juman error	no. correct	precision
Nikkei	18511	> 1	250-13=237	237-23=214	90.3%

Figure 1.14: Precision of extracted technical terms from Nikkei

1.2.6 Unknown word model

We argue that there is a larger amount of unknown or new words/terms in Japanese texts than in English texts due to their linguistic characteristics. On the other hand, these linguistic characteristics help us to identify unknown words as mostly being part of a technical term. Our method takes into account this property and incorporate an unknown word model in the regular expression for extracting technical terms.

In order to deal with unknown words, we incorporated unknown word insertion into the technical term POS patterns. This is because most new technical terms are composed of words not found in the dictionaries of JUMAN, especially in highly specialized domains such as computer science. These terms are nevertheless very important for various applications. Machine translation systems can greatly benefit from a bilingual lexicon of such terms and their translations. Moreover, technical dictionaries should also include these terms as lexical entries.

The unknown word problem exists to some extent in other languages also where there are previously unknown technical terms. However, this situation is aggravated in languages such as Japanese and Chinese. In English, it is rare that the individual lexical components of a term are also unknown. (e.g. *trade agreement* might not be found in a dictionary, but both *trade* and *agreement* are known words). In Japanese, the very definition of a "word" is based on the dictionaries a segmenter such as JUMAN uses. If the word is not listed in the dictionary, JUMAN simply treats each individual character as an unknown word. For example, the *Katakana* word *mi-se-e-ru/missile* is labeled as four individual unknown words. This is roughly akin to having a dictionary-based English tagger label *emacs* as five separate letters, each taggged as an unknown lexical item.

Figure 1.11 indicates that the newer a domain is, the higher the percentage of *Katakana* and English terms in the Japanese text. Figure 1.12 shows that there is a high percentage of unknown words in the Japanese AWK manual, corresponding to the *Katakana* and English terms.

In addition to the *Katakana* terms, there are many *Kanji* terms which are tagged as unknown. This is due to the frequently abbreviated structure of Chinese words, which are the roots of these *Kanji* terms. For example, the term *separation of state & church* is an abbreviated term with the first character of *politics, state*, the first character of *church* and the second character of *religion*, and the noun/verb *separate, separation*. We previously noted in [4] that this type of abbreviation increases the amount of unknown words in Chinese to a higher level than that of European languages. In fact, Japanese *Kanji* terms have an equally high degree of abbreviation as Chinese.

Some of the terms with unknown words extracted by our tool are shown in Figure 1.15. Test results are shown in Figure 1.16. Precision ranges from 68% to 86% when tokenization errors are discounted. It is remarkable that such a simple lexical model can capture many new terms, even when they occur very few times.

Incorporating unknown word lexical patterns into the syntactic regular expression finder is a first attempt in taking advantage of both lexical information and morphosyntactic information. Eventually we plan to incorporate statistical scores into the unknown word model to improve the performance of our tool.

ビジョン (vision)	ガソリン (Gasoline)
金利スワップ (interest rate swape)	リストラ (restructure)
ＰＥＭＥＸ	ＮＡＦＴＡ
クリントン (Clinton)	利ザヤ (profit margin)
富士重 (Fuji Heavy Industry Co.)	不動産 (real estate)
特別委 (Special Committee)	渡辺元副 (ex-Vice, Watanabe)
繊維工 (textile worker)	衆院予算委 (Lower House Budget Comittee)
取引事 (transaction time)	リスク (RISC)
リース (lease)	ヨーカ (Yokada Inc)
政教分離 (separation of state & church)	空母ミンスク (aircraft carrier Minsk)
空対空ミサイル (air-to-air missile)	ボスニア (Bosnia)
東京原油スポット市 (Tokyo Crude Oil Spot Market)	上海市 (Shanghai City)
ＲＩＳＣ	民主フォーラム (democracy forum)

Figure 1.15: Part of NIKKEI output containing unknown words, with some glosses

corpus	size	no. terms	tok. error	no. correct	precision tok. error	precision
Nikkei	18511	163	34	163-34-18=111	68.1 %	86.0 %

Figure 1.16: Precision of partial output containing unknown words

1.2.7 Rare terms can also be found

Justeson and Katz [7] report that the repetition of noun phrases provide discourse information for technical terms. This can be regarded as a most naive statistical significance measure. We previously used frequency threshold to extract statistically significant Chinese words and terms [4]. However, we argue that since technical terms in Japanese have more rigid morphological patterns than their counterparts in other languages, they are more easily distinguishable from non-technical terms. Our tool relies on this fact to extract Japanese technical terms which occur only once in a text. Some of the single frequency output is shown in Figure 1.17. Precision of partial output from the NIKKEI and AWK corpora are shown in Figure 1.18. The 87.8% precision rate for NIKKEI shows that our method is indeed effective in finding many technical terms with occurrence frequency one.

湾岸戦争 (Gulf War)	両院議長 (Chair of both Houses)
湾岸諸国国際会議 (gulf countries international conf.)	和平交渉再開 (peace talk reopen)
老人施設 (Home for the aged)	予備契約書 (preliminary contract)
労働問題研究委員会報告 (labor problem study report)	予定配当率 (planned dividend)
労働党政治局員 (Labor Party Rep.)	流動資産 (liquid assets)
連邦仲裁裁判所 (federal supreme court)	予算年内編成論 (year-ed budget compilation)
連邦赤字 (Commonwealth deficit)	予算措置 (budget facility)
冷戦終結 (end of the Cold War)	輸入販売管理 (import sales management)
冷戦時代 (Cold War era)	輸入契約分 (partial import contract)
臨時議員総会 (special Diet meeting)	輸入禁止品目 (import ban item)
臨時株主総会 (special shareholder meeting)	輸入禁止対象 (import ban target)

Figure 1.17: Partial list of terms occurring only once

corpus	size	no. terms	no. correct	precision
Nikkei	18511	950	950-116=834	87.8%

Figure 1.18: Precision of terms occurring only once

1.2.8 Conclusion

We have presented two systems for technical and domain-specific lexical unit extraction—CXtract, for Chinese, is based on statistical lexical information with morphological filters; and JBrat, for Japanese, uses morphosyntactic rules of Japanese domain and technical terms. CXtract is algorithmic while JBrat is more heuristic-based.

The motivation for developing a statistics-oriented CXtract is due to a lack of general morphological and syntactic information for Chinese. CXtract is adapted from Xtract, an English collocation finder, so it is in theory independent of the corpus language characteristics. We have used some morphological filters for Chinese to improve the statistical output. We evaluated the output of CXtract by using it in a tokenization task. Using a best case evaluation, where a segmentation is correct when at least one human evaluator agrees, the precision of the baseline tokenizer is

85.02%. The precision when CXtract output is used is 89.01%. It is 91.71% when additional morphological filters are applied to the statistical output. In the average case, where half of the human evaluators agree, the baseline precision is 68.09%. When CXtract statistical and then morphologically filtered statistical output are used, the precisions are 75.49% and 80.24% respectively.

JBrat, developed for Japanese technical term extraction, is based on a different approach. Japanese differs from Chinese in that it has three different character sets whose usage follow regular syntactic patterns. JBrat uses simple morphological and syntactic information learned from a small annotated sample, in combination with findings from published technical dictionaries. It has 90.3% precision in finding technical terms in a open test set, when the terms contain no unknown words, and when they occur more than once. This tool has 87.8% precision in finding technical terms which occur only once. By incorporating unknown word lexical model into the syntactic model, JBrat achieved 86% precision in finding technical terms containing unknown words. Even when tokenization error is not discounted, it still has 68.1% precision in finding unknown word technical terms.

What is remarkable about CXtract and JBrat are their powerful performances despite simple origins—CXtract adapts simple statistical equations from Xtract; JBrat is based on a few heuristic rules. In fact, the very simplicity of the morphosyntactic rules for JBract seems to have caused some controversies. After all, Chinese and Japanese are *supposed* to be much more difficult than European languages. Our response to such doubts is that Chinese and Japanese are difficult only if one seeks to fit them into an Indo-European model of linguistic theories. Using a combination of language-independent techniques, such as CXtract's statistical model, and langauge-dependent linguistic knowledge, such as JBrat's morphological and syntactic rules, we should be able to further the state-of-the-art in Asian language processing.

Bibliography

[1] A. Aho, B. Kernighan, and P. Weinberger. *The AWK Programming Language.* Addison-Wesley, Reading, Massachusetts, USA., 1980.

[2] BDC. *The BDC Chinese-English Electronic Dictionary (version 2.0).* Behavior Design Corporation, 1992.

[3] FDMC. *Xiandai Hanyu Pinlu Cidian (Frequency Dictionary of Modern Chinese).* Beijing Language Institute Press, 1986.

[4] Pascale Fung and Dekai Wu. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, pages 69–85, Kyoto, Japan, June 1994.

[5] Louis Guilbert. *La créativité lexicale.* Larousse, Paris, 1975.

[6] IWNM. *Iwa Nami Information Science Dictionary.* Iwa Nami Shoten, 1990. In Japanese.

[7] John S. Justeson and Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

[8] KYJBJD. *Financial Management Dictionary.* Financial Study Group Inc., 1975. In Japanese.

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

[10] Yuji Matsumoto and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 22–28, August 1994.

[11] NIKKEI. *Nihon Kezai Shimbun.* Nihon Kezai Shimbun, Inc., 1994.

[12] Isadore Pinchuck. *Scientific and technical translation.* Andre Deutsch, 1977.

[13] Juan C. Sager. *A Practical Course in Terminology Processing.* John Benjamins B.V., 1990.

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

[15] TKSHS. *Statistics Dictionary.* Toyo Kezai Shin Ho Sha, 1989. In Japanese.

[16] Dekai Wu and Pascale Fung. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 180–181, Stuttgart, Germany, October 1994.