Second International Conference in New Methods for Language Processing, {NEMLP} Bilkent, Turkey: Sep. 1996. pp. 148-159

Extracting Japanese Domain and Technical Terms is Relatively Easy

Pascale Fung
Min-yen Kan
Yurie Horita
Computer Science Department
Columbia University, New York, NY 10027
{pascale,min}@cs.columbia.edu

Abstract

We argue that the important task of extracting domain and technical terms is much easier in Japanese than is commonly believed, and that technical term extraction should and can be more widely used. We present a method which shows that the implementation of such a tool for Japanese can be remarkably simple due to the regularity and rigidity of morphosyntactic characteristics of Japanese technical terms. Our learning algorithm bootstraps from the tagged output of a Japanese tokenizer/tagger to learn syntactic and morphological properties of technical terms. We show that very simple linguistic patterns are reliable enough to yield high precision technical terms from tagged texts, even for terms which occur only once and are usually overlooked by English technical term extractors. In addition, by incorporating a model for unknown words, we are able to extract correct technical terms from words not properly tagged.

1 Introduction

Technical terms provide essential knowledge for various NLP systems. They are important for information retrieval, machine translation, translator-aid and technical manual generation tasks. Professional technical writers, translators and machine translations systems often have difficulties generating or translating domain and technical terms. There is no formal definition of technical terms[12]. However, they can be intuitively characterized as being highly specific to a domain. They are lexically rigid, i.e. the same term is always used to refer to a domain-specific concept. And they are often new. Published technical dictionaries do not provide enough coverage of these terms [10]. One reason for the inadequacy of coverage is that there are many sub-areas in a domain with specialized terminology. Another reason is that in areas such as computer science, new terms are being created constantly and technical dictionaries often fail to be up-to-date. In addition, efforts at providing a standard up-to-date dictionary of new terms each year require a lot of time and manual labor. It can be very helpful to have an automatic tool which can extract domain and technical terms from large corpora as an aid to information retrieval systems, dictionary editors, translators, technical writers, and machine translation systems. It is also desirable that this tool be easy to implement and robust for different domains.

Japanese is generally regarded as a more difficult language than English or other European languages for natural language processing. There are many morphosyntactic and stylistic singularities in Japanese, such as no space delimiter between words, three different character sets, large number

of homonyms, same word written in different forms, relatively free word order, few function words, etc. Perhaps due to such perceived difficulties, much research on Japanese terms has been concentrated on the analysis of some sub-classes such as four character compound nouns [13], rather than a full study of Japanese domain and technical terms. On the other hand, tokenization and tagging of Japanese texts has been greatly facilitated by tools such as JUMAN, which is freely available. We show in this paper that the implementation of a Japanese domain and technical term extraction tool is remarkably simple given a Japanese tokenizer/tagger.

For many other languages, much research has been done on tools for automatic extraction of technical terms from large corpora. These languages include English [12, 4], Greek [2], French [3], and Chinese [8]. There are two kinds of main approaches—*lexical* and *morphosyntactic*. Lexical approaches use statistical correlation scores between single words (or characters) [8, 4]. Morphosyntactic approaches use syntactic and morphological information provided by taggers [12, 2, 3]. The advantages and disadvantages of these two approaches are listed in Figure 1.

Lexical approach	**Morphosyntactic approach**
More adaptive—the same statistical measure can be used over different domain texts.	Less robust–prior knowledge differ for different domains.
Statistical measure can reliably recognize unknown words.	Reliance on prior knowledge cannot solve unknown word problem.
Precision suffers because many statistically correlated words are not terms (e.g. "I hereby claim that").	High precision, if prior knowledge matches text.
Cannot process rare terms.	Able to robustly process less frequent terms.

Figure 1: Advantages and disadvantages of lexical vs. morphosyntatic approach

The design of a technical term extractor involves two trade-offs: (1) the trade-off between efficiency and robustness—how to tune an algorithm to suit a particular sub-domain for a high yield of technical terms versus how easy the algorithm can adapt to another domain; (2) the trade-off between precision and recall.

We choose an approach of using morphosyntactic information, augmented with an unknown word model, for Japanese technical term extraction. Our approach is motivated by the following reasons:

- 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. Statistically significant non-technical collocations are not as useful. On the other hand, technical terms which occur infrequently should also be included.
- 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.
- We can make use of a freely available Japanese POS tokenizer/tagger, JUMAN.
- 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 paper 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 to 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.

2 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

3 Using a Japanese tokenizer and tagger

As a preprocess for learning the syntactic and morphological rules of Japanese technical terms, we use JUMAN to perform two tasks: (1) to insert delimiters between Japanese words, and then (2) to label these words with POS tags. This tool, developed by Matsumoto and Nagao [15], uses a set of Japanese Word Construction Grammar and syntactic grammar rules for transition rules, a morpheme dictionary and various grammar dictionaries to tokenize a text into word-segmented, POS tagged format with inflection and phonetic information. Two levels of POS classification are used in the POS dictionary. There are about 6000 grammar rules, 20 grammar and morpheme dictionaries. There are seven dictionaries for nouns, corresponding to seven noun classes:

- 普通名詞 /**Nn** Nominal noun
- 副詞的名詞 /**Na** Adverbial noun, a noun which can be an adverb
- 形式名詞 /**Nk** Keishiki noun, nouns containing particles
- 固有名詞 /**Np** Proper noun
- サ変名詞 /**Nv** Verbal noun, a noun which becomes a verb with the suffix する /*suru, do*
- 数詞 /**Nm** Numerals

word-token	phonetics	root	POS tag	POS sub-class	usage
細川	(ほそかわ)	細川	固有名詞		
首相	(しゅしょう)	首相	普通名詞		
は	(は)	は	副助詞		
一日	(いちにち)	一日	時相名詞		
午後	(ごご)	午後	時相名詞		
の	(の)	の	名詞接続助詞		
衆院	(しゅういん)	衆院	普通名詞		
予算委員会	(よさんいいん	予算委員会	普通名詞		
で	(で)	だ	判定詞	判定詞	ダ列タ系連用
、	(、)	、	読点		

Figure 2: JUMAN output format

- 地名 **/Nt** Place names

A typical JUMAN output is as follows in Figure 2 (taken from Nihon Kezai Shimbun [16]):

Phonetics are useful in applications such as speech synthesis. The morphological *root* word differs from the word-token only in cases of verb conjugation, verbalization of nouns, etc. This information is useful when morphological normalization is needed. Japanese technical terms, being mostly noun phrases, have fairly rigid forms. Thus we do not need to use morphological normalization[1]. *POS sub-class* represents a finer division of certain POS tags. *Usage* indicates how certain function words are used. Again, POS sub-class and usage are not directly relevant to extracting Japanese technical terms. Therefore, we use only the tokenized word output together with their POS tags. The input to our tool consists of the first and fourth column of Figure 2.

We tokenize and tag both the Japanese translation of the AWK[2] manual [1], and the Nihon Kezai Shimbun (the NIKKEI corpus) [16]. 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.

There have been various reports on the tokenization and tagging accuracy of JUMAN, ranging from 95% to 100%. However, the evaluation sets for these tests varied, and it is not clear from the reports whether the evaluation was on closed test sets or open sets. From using JUMAN on our corpora, we find that the actual error rate in tokenization and tagging is higher than 5%. 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 未定義語 (Unk)/unknown. Unknown word recognition is one of the biggest limitation 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.

4 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:

[1] The morphological information we use in this paper refers to the properties of character sets in Asian languages. The particular character set of a word is a type of morphological feature.

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

- 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[3]

- There is no space delimiter between words in Japanese.
- 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[4].

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 syntactical 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.

5 Learning morphological patterns from technical dictionaries

In order to verify our knowledge of Japanese technical terms, we consulted several Japanese technical dictionaries [14, 11, 17]. We drew random samples from the dictionaries and computed the distribution of character sets shown in Figure 3. 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	の /of	Roman
Financial	85.0%	9.8%	4.2%	6.9%	1.0%
Statistics	71.4%	17.1%	4.3%	3.6%	3.6%
Computer Science	50.3%	33.2%	4.2%	3.9%	8.4%

Figure 3: Percentage of terms containing different character sets

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

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

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

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 プログラム /*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 close set. Common nouns consisting of only *Hiragana* are not among the technical terms. *Hiragana* in technical terms are mostly the suffix する /*suru, do* or the nominalizer の /, 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 あげる /*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[5]). Simple verbs can have new usage 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 する /*suru, do*. This limitation of new verbs to verbalized nouns seems to be a common phenomenon to many other languages, including English and French [9].

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 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**

6 Learning syntactic patterns from a bilingual corpus

The next step is to 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

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

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 [12], and (2) their translations in Japanese are technical terms as well, regardless of their POS tags, we bootstrap the learning process from the 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 4.

%	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 4: Top 20 frequent POS patterns of Japanese technical term POS patterns

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.

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** $(Adj*(Np*Nn*Unk+(NnNm)*Nv*)+)$

The tool was tested without rule 6 on part of the NIKKEI corpus. Some examples extracted by our tool are as shown in Figure 5. Precision yield of these tests are shown in Figure 6.

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 5: Some technical terms extracted from Nikkei

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

Figure 6: Precision of extracted technical terms from Nikkei

7 Empirical finding: Unknown words are technical terms

We argue that there is a larger amount of unknown or new words/terms in Japanese texts than in English texts due to its 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 is 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 ミサイル /*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 3 indicates that the newer a domain is, the higher the percentage of *Katakana* and English terms in the Japanese text. Figure 4 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 frequent abbreviated structure of Chinese words, which are the roots of these *Kanji* terms. For example, the term 政教分離 / *separation of state & church* is an abbreviated term with 政 being the first character of *politics, state*, 教 the first character of *church* and the second character of *religion*, and 分離 is the noun/verb *separate, separation*. We previously noted in [8] 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 equally high degree of abbreviation as Chinese.

Some of the terms with unknown words extracted by our tool are shown in Figure 7. Test results are shown in Figure 8. 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.

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 7: 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 8: Precision of partial output containing unknown words

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.

8 Empirical finding: Rare terms can also found

Justeson and Katz [12] 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 [8]. 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 are shown in Figure 9. Precision of partial output from NIKKEI and AWK corpora are shown in Figure 10. 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 9: Partial list of terms occurring only once

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

Figure 10: Precision of terms occurring only once

9 How to use the term extractor

Based on our findings, we believe the tool can be applied to the following applications:

- **Machine translation systems** translate technical terms without any human intervention. It should only use terms with frequency greater than one. This gives a higher hit rate for technical terms. For terms containing unknown words, those in *Katakana* would have a higher hit rate.
- **Human translators** need as much help as available and manual filtering is relatively easy. So this tool can be used with an unknown word model, without frequency counts. Recall is more important for humans than precision is.
- **Technical dictionary update** needs a list of new terms with high recall. We suggest that this tool be used without frequency counts, including an unknown word model. Terms in *Hiragana* should not be filtered out since they could be common verbs used in a specific domain-dependent way. Manual post-filtering can be done to select a final list for the dictionary.
- **Information retrieval** systems can use this tool with frequency counts, including an unknown word model, terms with *Hiragana*, and terms with all character lengths. Since recipients of such information are human, high recall rate is desired and noise-filtering is easy. Moreover, since statistically significant collocations and key words have high information content, we suggest that this tool be combined with a statistical model which can extract collocations and key words other than technical terms.

10 Future work

We obtained the precision of our tool by human evaluation of the output. However, there is no satisfying way for us to carry out a recall evaluation due to the subjective nature of defining technical terms. In [12], recall was computed for a short paper by asking the author of the paper to manually select what he believes to be technical terms from his own paper, and then evaluating the tool output against his selection. Unfortunately, it is not feasible for us to ask the authors of AWK manual or various articles in NIKKEI to manually select a technical term list. We also ruled out the AWK manual index as a recall baseline because (1) many of the index entries do not correspond to any term in the text. e.g. *leap year computation* refers to the page where a program segment for computing leap year is presented and described, and the term *leap year computation* does not actually appear in the text; (2) many of the index entries are not technical terms per se, such as *records with headers*; and (3) some others could be error in indexing, such as the index entry *happiness 30*. Under such circumstances, any recall evaluation would be overly subjective and possibly meaningless.

We plan to further evaluate the performance of our term extractor in conjunction with a terminology translation tool we are currently developing [6, 5, 7]. This translation tool will match these extracted Japanese technical terms to their English correspondence in a corpus. The output of our Japanese technical term extractor can then be evaluated against that of an English term extractor, and vice versa.

11 Conclusion

We have shown a method for developing an automatic tool for extracting Japanese technical terms. This tool is very easy to implement and as shown in our open set tests, robust to domain variations. It uses simple morphological and syntactic information learned from a small annotated sample, in combination with findings from published technical dictionaries. This tool filters out all candidate terms having character size smaller than 3, and candidate terms which start in *Hiragana*. 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, our tool 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. Our empirical findings show that Japanese technical terms have a very regular morphosyntactic pattern and are therefore easily extractable from texts.

12 Acknowledgment

This research was partially supported by a joint grant from the Office of Naval Research and the Defense Advanced Research Projects Agency under contract # N000 14-95-1-0745 and by National Science Foundation Grant GER-90-2406.

References

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

[2] Sophia Ananiadou. A methodology for automatic term recognition. In *Proceedings of COLING-94*, pages 1034–1038, 1994.

[3] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of COLING 92*, pages 977–981, 1992.

[4] Ido Dagan and Kenneth W. Church. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 34–40, Stuttgart, Germany, October 1994.

[5] Pascale Fung. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Annual Workshop on Very Large Corpora*, pages 173–183, Boston, Massachusettes, June 1995.

[6] Pascale Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, pages 236–233, Boston, Massachusettes, June 1995.

[7] Pascale Fung. Space-frequency analysis for domain word translation. In *Proceedings of ICASSP 96*, Atlanta, Georgia, May 1996. To appear.

[8] 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.

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

[10] Stephanie W. Haas. Covering the vocabulary of technical abstracts using standard and specialized dictionaries. *Journal of Information Science*, 18:363–373, 1992.

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

[12] 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.

[13] Yosiyuki Kobayasi, Takenobu Tokunaga, and Hozumi Tanaka. Analysis of syntactic structures of japanese compound noun. In *Proceedings of NLPRS'95*, pages 326–331, 1995.

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

[15] 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.

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

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