Online Acquisition of Japanese Unknown Morphemes using Morphological Constraints

Yugo Murawaki **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
murawaki@nlp.kuee.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

Abstract

We propose a novel lexicon acquirer that works in concert with the morphological analyzer and has the ability to run in online mode. Every time a sentence is analyzed, it detects unknown morphemes, enumerates candidates and selects the best candidates by comparing multiple examples kept in the storage. When a morpheme is unambiguously selected, the lexicon acquirer updates the dictionary of the analyzer, and it will be used in subsequent analysis. We use the constraints of Japanese morphology and effectively reduce the number of examples required to acquire a morpheme. Experiments show that unknown morphemes were acquired with high accuracy and improved the quality of morphological analysis.

1 Introduction

Morphological analysis is the first step for most natural language processing applications. In Japanese morphological analysis, segmentation is processed simultaneously with the assignment of a part of speech (POS) tag to each morpheme. Segmentation is a nontrivial task in Japanese because it does not delimit words by white-space.

Japanese morphological analysis has successfully adopted dictionary-based approaches (Kurohashi et al., 1994; Asahara and Matsumoto, 2000; Kudo et al., 2004). In these approaches, a sentence is transformed into a lattice of morphemes by searching a pre-defined dictionary, and an optimal path in the lattice is selected.

This area of research may be considered almost completed, as previous studies reported the F-score of nearly 99% (Kudo et al., 2004). When applied to web texts, however, more errors are made due to unknown morphemes. In previous studies, experiments were performed on newspaper articles, but web texts include slang words, informal spelling alternates (Nishimura, 2003) and technical terms. For example, the verb “ググる” (*gugu-ru*, to google) is erroneously segmented into “ググ” (*gugu*) and “る” (*ru*).

One solution to this problem is to augment the lexicon of the morphological analyzer by extracting unknown morphemes from texts (Mori and Nagao, 1996). In the previous method, a morpheme extraction module worked independently of the morphological analyzer and ran in off-line (batch) mode. It is inefficient because almost all high-frequency morphemes have already been registered to the pre-defined dictionary. Moreover, it is inconvenient when applied to web texts because the web corpus is huge and diverse compared to newspaper corpora. It is not necessarily easy to build subcorpora before lexicon acquisition. Suppose that we want to analyze whaling-related documents. It is unnecessary and probably harmful to acquire morphemes that are irrelevant to the topic. A whaling-related subcorpus should be extracted from the whole corpus but it is not clear how large it must be.

We propose a novel lexicon acquirer that works in concert with the morphological analyzer and has the ability to run in online mode. As shown in Figure 1, every time a sentence is analyzed, the lexicon acquirer detects unknown morphemes, enumerates

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 429–437,
Honolulu, October 2008. ©2008 Association for Computational Linguistics

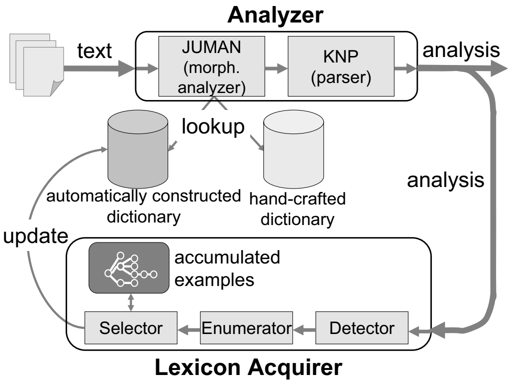


Figure 1: System architecture

candidates and selects the best candidates by comparing multiple examples kept in the storage. When a morpheme is unambiguously selected, the lexicon acquirer updates the automatically constructed dictionary, and it will be used in subsequent analysis. The proposed method is flexible and gives the system more control over the process. We do not have to limit the target corpus beforehand and the system can stop whenever appropriate.

We use the constraints of Japanese morphology that have already been coded in the morphological analyzer. These constraints effectively reduce the number of examples required to acquire an unknown morpheme. Experiments show that unknown morphemes were acquired with high accuracy and improved the quality of morphological analysis.

2 Japanese Morphology

In order to understand the task of lexicon acquisition, we briefly describe the Japanese morphological analyzer JUMAN.[1] We explain Japanese morphemes in Section 2.1, morphological constraints in Section 2.2, and unknown morpheme processing in Section 2.3.

2.1 Morpheme

In JUMAN, the POS tagset consists of four elements: class, subclass, conjugation type and conjugation form. The classes are noun, verb, adjective and others. Noun has subclasses such as common noun, *sa*-group noun, proper noun, organization, place, personal name. Verb and adjective have no subclasses.

Verbs and adjectives among others change their form according to the morphemes that occur after them, which is called conjugation. Conjugable morphemes are grouped by conjugation types such as vowel verb, *ra*-row verb, *i*-type adjective and *na*-type adjective. Each conjugable morpheme takes one of conjugation forms in texts. It has an invariant stem and an ending which changes according to conjugation type and conjugation form.

In this paper, the tuple of class, subclass and conjugation type is referred to as a POS tag. For simplicity, POS tags for nouns are called by their subclasses and those for verbs and adjectives by their conjugation types.

There are two types of morphemes: abstract dictionary entries, and examples or actual occurrences in texts. An entry consists of a stem and a POS tag while an example consists of a stem, a POS tag and a conjugation form. For example, the entry of the *ra*-row verb "走る" (*hashi-ru*, to run) can be represented as

(“走” (*hashi*), *ra*-row verb),

and their examples "走ら" (*hashi-ra*) and "走り" (*hashi-ri*) as

(“走” (*hashi*), *ra*-row verb, imperfective),

and

(“走” (*hashi*), *ra*-row verb, plain continuative)

respectively. As nouns do not conjugate, the entry of the *sa*-group noun "希望" (*kibou*, hope) can be represented as

(“希望” (*kibou*), *sa*-group noun)

and its sole example form is

(“希望” (*kibou*), *sa*-group noun, NIL).

2.2 Morphological Constraints

Japanese is an agglutinative language. Depending on its grammatical roles, a morpheme is followed by a sequence of grammatical suffixes, auxiliary verbs and particles, and the connectivity of these elements is bound by morphological constraints. For example, the particle "を" (*wo*, accusative case) can follow a verb with the conjugation form of plain continuative, as in "走りを" (*hashi-ri-wo*, running-ACC),

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

but it cannot follow an imperfective verb ("*走らを" (**hashi-ra-wo*)).

These constraints are used by JUMAN to reduce the ambiguity. They can be also used in lexicon acquisition.

2.3 Unknown Morpheme Processing

Given a sentence, JUMAN builds a lattice of morphemes by searching a pre-defined dictionary, and then selects an optimal path in the lattice. To handle morphemes that cannot be found in the dictionary, JUMAN enumerates unknown morpheme candidates using character type-based heuristics, and adds them to the morpheme lattice. Unknown morphemes are given the special POS tag "undefined," which is treated as noun.

Character type-based heuristics are based on the fact that Japanese is written with several different character types such as kanji, hiragana and katakana, and that the choice of character types gives some clues on morpheme boundaries. For example, a sequence of katakana characters are considered as an unknown morpheme candidate, as in "グーグル" (*gûguru*, Google) out of "グーグルが" (*gûguru-ga*, Google-NOM). Kanji characters are segmented per character, which is sometimes wrong but prevents error propagation.

These heuristics are simple and effective, but far from perfect. They cannot identify mixed-character morphemes, verbs and adjectives correctly. For example, the verb "ググる" (*gugu-ru*, to google) is wrongly divided into the katakana unknown morpheme " ググ" (*gugu*) and the hiragana suffix "る" (*ru*).

3 Lexicon Acquisition

3.1 Task

The task of lexicon acquisition is to generate dictionary entries inductively from their examples in texts. Since the morphological analyzer provides a basic lexicon, the morphemes to be acquired are limited to those unknown to the analyzer.

In order to generate an entry, its stem and POS tag need to be identified. Determining the stem of an example is to draw the front and rear boundaries in a character sequence in texts which corresponds to the stem. The POS tag is selected from the tagset given by the morphological analyzer.

3.2 System Architecture

Figure 1 shows the system architecture. Each sentence in texts is processed by the morphological analyzer JUMAN and the dependency parser KNP.[2] JUMAN consults a hand-crafted dictionary and an automatically constructed dictionary. KNP is used to form a phrasal unit called *bunsetsu* by chunking morphemes.

Every time a sentence is analyzed, the lexicon acquirer receives the analysis. It detects examples of unknown morphemes and keeps them in storage. When an entry is unambiguously selected, the lexicon acquirer updates the automatically constructed dictionary, and it will be used in subsequent analysis.

3.3 Algorithm Overview

The process of lexicon acquisition has four phases: detection, candidate enumeration, aggregation and selection. First the analysis is scanned to detect examples of unknown morphemes. For each example, one or more candidates for dictionary entries are enumerated. It is added to the storage, and multiple examples in the storage that share the candidates are aggregated. They are compared and the best candidate is selected from it.

Take the *ra*-row verb "ググる" (*gugu-ru*) for example. Its example "ググってみた。" (*gugu-tte-mi-ta*, to have tried to google) can be interpreted in many ways as shown in Figure 2. Similarly, multiple candidates are enumerated for another example "ググるのは" (*gugu-ru-no-ha*, to google-TOPIC). If these examples are compared, we can see that the *ra*-row verb "ググる" (*gugu-ru*) can explain them.

3.4 Suffixes

Morphological constraints are used for candidate enumeration. Since they are coded in JUMAN, we first transform them into a set of strings called suffixes. A suffix is created by concatenating the ending of a morpheme (if any) and subsequent ancillary morphemes. Each POS tag is associated with a set of suffixes, as shown in Table 1. This means that a stem can be followed by one of the suffixes specified

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

Table 1: Examples of suffixes

POS tag	base form	stem	ending	conjugation form[1]	suffixes
ra-row verb	*hashi-ru*	*hashi*	*ra*	imperfective	*razu*, *ranaide*
			ri	plain continuative	*riwo*, *riwomo*
			ru	plain	*ru*, *rukawo*
vowel verb	*akogare-ru*	*akogare*	ϕ	imperfective	*zu*, *naide*
			ϕ	plain continuative	*wo*, *womo*
			ru	plain	*ru*, *rukawo*
sa-group noun	*kibou*	*kibou*	NIL	*wo*	*wo*, *womo*
			NIL	*suru*	*suru*, *shitara*

[1] The conjugation form of a noun is substituted with the base form of its immediate ancillary morpheme because nouns do not conjugate.

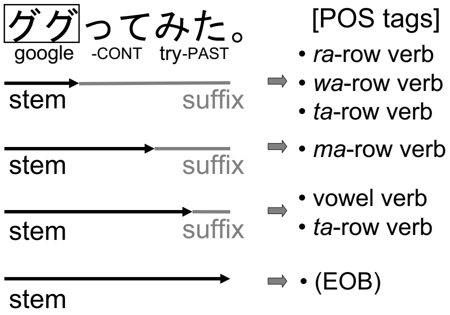


Figure 2: Candidate enumeration

by its POS tag and cannot be followed by any other suffix.

In preparation for lexicon acquisition, suffixes are acquired from a corpus. We used a web corpus that was compiled through the procedures proposed by Kawahara and Kurohashi (2006). Suffixes were extracted from examples of registered morphemes and were aggregated per POS tag.

We found that the number of suffixes did not converge even in this large-scale corpus. It was because ancillary morphemes included the wide variety of auxiliary verbs and formal nouns. Alternatively, we used the first five characters as a suffix. In the experiments, we obtained 500 thousand unique suffixes from 100 million pages. The number of POS tags that corresponded to a suffix was 1.33 on average.

3.5 Unknown Morpheme Detection

The first step of lexicon acquisition is unknown morpheme detection. Every time the analysis of a sentence was given, the sequence of morphemes are scanned, and suspicious points that probably represent unknown morphemes are detected.

Currently, we use the POS tag "undefined" to detect unknown morphemes. For example, the example "ググってみた。" is detected because "ググ" is given "undefined." This simple method cannot detect unknown morphemes if they are falsely segmented into combinations of registered morphemes. We leave the comprehensive detection of unknown morphemes to future work.

3.6 Candidate Enumeration

For each example, one or more candidates for the dictionary entry are enumerated. Each candidate is represented by a combination of a front boundary and the pair of a rear boundary and a POS tag.

The search range for enumeration is based on *bunsetsu* phrases, which is created by chunking morphemes. The range is at most the corresponding *bunsetsu* and the two immediately preceding and succeeding *bunsetsu*, which we found wide enough to contain correct candidates.

The candidates for the rear boundary and the POS tag are enumerated by string matching of suffixes as shown in Figure 2. If a suffix matches, the starting position of the suffix becomes a candidate for the rear boundary and the suffix is mapped to one or more corresponding POS tags.

In addition, the candidates for the front and rear boundaries are enumerated by scanning the sequence of morphemes. The boundary markers we use are

- punctuations,
- grammatical prefixes such as "御" (*go*-, honorific prefix), for front boundaries,

- grammatical suffixes such as “様” (*-sama*, honorific title), for rear boundaries, and
- *bunsetsu* boundaries given by KNP.

Each rear boundary candidate whose corresponding POS tag is not decided is given the special tag “EOB” (*end-of-bunsetsu*). This means that no suffix is attached to the candidate. Since nouns, vowel verbs and *na*-type adjectives can appear in isolation, it will be expanded to these POS tags when selecting the best POS tag.

3.7 Aggregation of Examples

Selection of the best candidate is done by comparing multiple examples. Each example is added to the storage, and then examples that possibly represent the same entry with it are extracted from the storage. Examples aggregated at this phase share the front boundary but may be unrelated to the example in question. They are pruned in the next phase.

In order to manage examples efficiently, we implement a trie. The example is added to the trie for each front boundary candidate. The key is the character sequence determined by the front boundary and the leftmost rear boundary. To retrieve examples that share the front boundary with it, we check every node in the path from the root to the node where it is stored, and collect examples stored in each node.

3.8 Selection

The best candidate is selected by identifying the front boundary, the rear boundary and the POS tag in this order. Starting from the rightmost front boundary candidate, multiple rear boundary candidates that share the front boundary are compared and some are dropped. Then starting from the leftmost surviving rear boundary candidate, the best POS tag is selected from the examples that share the stem. If the selected candidate satisfies simple termination conditions, it is added to the dictionary and the examples are removed from the storage.

For each front boundary candidate, some inappropriate rear boundary candidates are dropped by examining the inclusion relation between the examples of a pair of candidates. The assumption behind this is that an appropriate candidate can interpret more examples than incorrect ones. Let p and q be a pair of the candidates for the rear boundary, and R_p and R_q be the sets of examples for which p and q are enumerated. If p is a prefix of q and p is the correct stem, then R_q must be contained in R_p. In practice we loosen this condition, considering possible errors in candidate enumeration

For each stem candidate, the appropriate POS tag is identified. Similarly to rear boundary identification, POS identification is done by checking inclusion relation.

If the POS tag is successfully disambiguated, simple termination conditions is checked to prevent the accidental acquisition of erroneous candidates. The first condition is that the number of unique conjugation forms that appear in the examples should be 3 or more. If the candidate is a noun, it is substituted with the number of the unique base forms of their immediate ancillary morphemes. The second condition is that the front boundaries of some examples are decided by clear boundary markers such as punctuations and the beginning of sentence. This prevents oversegmentation. For example, the stem candidate “*撰組” (**sengumi*) is always enumerated for examples of “新撰組” (*Shingengumi*, a historical organization) since “新” (*shin-*, new) is a prefix. This candidate is not acquired because “*撰組” (**sengumi*) does not occur alone and is always accompanied by “新” (*shin-*). Thresholds are chosen empirically.

3.9 Decompositionality

Since a morpheme is extracted from a small number of examples, it is inherently possible that the acquired morpheme actually consists of two or more morphemes. For example, the noun phrase “顆粒タイプ” (*karyuu-taipu*, granular type) may be acquired as a morpheme before “顆粒” (*karyuu*, granule) is extracted. To handle this phenomenon, it is checked at the time of acquisition whether the new morpheme (*kairyuu*) can decompose registered morphemes (*kairyuu-taipu*). If found, a composite “morpheme” is removed from the dictionary.

Currently we leave the decompositionality check to the morphological analyzer. Possible compounds are enumerated by string matching and temporarily removed from the dictionary. Each candidate is analyzed by the morphological analyzer and it is checked whether the candidate is divided into a combination of registered morphemes. If not, the candidate is restored to the dictionary.

Table 2: Statistical information per query

query	number of sentences	number of affected sentences (ratio)	number of acquired morphs	number of correct morphs (precision)	number of examples[1]
捕鯨問題 (whaling issue)	135,379	2,444 (1.81%)	293	290 (99.0%)	4
赤ちゃんポスト (baby hatch)	74,572	775 (1.04%)	107	105 (98.1%)	4
ジャスラック (JASRAC)	195,928	6,259 (3.19%)	913	907 (99.3%)	4
ツンデレ (tsundere)	77,962	12,012 (15.4%)	243	238 (97.4%)	5
アガリクス (agaricus)	78,922	3,037 (3.85%)	114	107 (93.9%)	9

[1] The median number of examples used for acquisition.

4 Experiments

4.1 Experimental Design

We used the default dictionary of the morphological analyzer JUMAN as the initial lexicon. It contained 30 thousand basic morphemes. If spelling variants were expanded and proper nouns were counted, the total number of morphemes was 120 thousands.

We used domain-specific corpora as target texts because efficient acquisition was expected. If target texts shared a topic, relevant unknown morphemes were used frequently. In the experiments, we used search engine TSUBAKI (Shinzato et al., 2008) and casted the search results as domain-specific corpora. For each query, our system sequentially read pages from the top of the result and acquired morphemes. We terminated the acquisition at the 1000th page and analyzed the same 1000 pages with the augmented lexicon. The queries used were "捕鯨問題" (whaling issue), "赤ちゃんポスト" (baby hatch), "ジャスラック" (JASRAC, a copyright collective), "ツンデレ" (tsundere, a slang word) and "アガリクス" (agaricus).

4.2 Evaluation Measures

The proposed method is evaluated by measuring the accuracy of acquired morphemes and their contribution to the improvement of morphological analysis. A morpheme is considered accurate if both segmentation and the POS tag are correct. Note that segmentation is a nontrivial problem for evaluation. In fact, the disagreement over segmentation criteria was considered one of the main reasons for reported errors by Nagata (1999) and Uchimoto et al. (2001). It is difficult to judge whether a compound term should be divided because there is no definite standard for morpheme boundaries in Japanese. For example, "ミンク鯨" (*minku-kujira*, minke whale) can be extracted as a single morpheme or decomposed into "ミンク" and "鯨." While segmentation is an open question in Japanese morphological analysis, "correct" segmentation is not necessarily important for applications using morphological analysis. Even if a noun is split into two or more morphemes in morphological analysis, they are chunked to form a phrasal unit called *bunsetsu* in dependency parsing, and to extract a keyword (Nakagawa and Mori, 2002).

To avoid the decompositionality problem, we adopted manual evaluation. We analyzed the target texts with both the initial lexicon and the augmented lexicon. Then we checked differences between the two analyses and extracted sentences that were affected by the augmentation. Among these sentences, we evaluated randomly selected 50 sentences per query. We checked the accuracy of segmentation and POS tagging of each "diff" block, which is illustrated in Figure 3. The segmentation of a block was judged correct unless morpheme boundaries were clearly wrong.

In the evaluation of POS tagging, we did not distinguish subclasses of noun[3] such as common noun

[3] In the experiments, we regarded demonstrative pronouns as

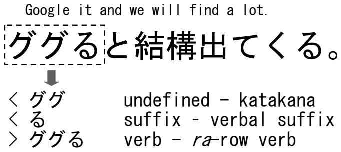
Table 3: Examples of acquired morphemes

query	examples
whaling issue	モラトリアム (moratorium), ツチクジラ (giant beaked whale), 混獲 (bycatch)
baby hatch	ダンナ (husband), 助産師 (midwife), 棄てる (to abandon), 訊く (to inquire)
JASRAC	ソフ倫 (an organization), シャ乱 Q (a pop-rock band), ヲタ (geek)
tsundere	アキバ (abbr. of Akihabara), 腐女子 (*fujoshi*, a slang word), モテる (to be popular)
agaricus	サプリ (abbr. of suppliment), アロマ (aroma), 食効 (enhanced nutritional function)

Table 4: Evaluation of “diff” blocks

	segmentation				POS tagging				
query	E → C	C → C	E → E	C → E	E → C	C → C	E → E	C → E	total
whaling issue	11	**45**	0	2	11	**45**	0	2	58
baby hatch	**37**	12	0	3	**37**	12	0	3	52
JASRAC	16	**23**	1	12	16	**23**	1	12	52
tsundere	17	**39**	0	1	17	**39**	0	1	57
agaricus	22	**31**	0	0	22	**31**	0	0	53

(Legend – C: correct; E: erroneous)



```
Google it and we will find a lot.
ググると結構出てくる。
< ググ      undefined - katakana
< る        suffix - verbal suffix
> ググる    verb - ra-row verb
```

Figure 3: A “diff” block in a sentence

and proper noun. The special POS tag “undefined” given by JUMAN was treated as noun.

4.3 Results

Table 2 summarizes statistical information per query. The number of sentences affected by the augmentation varied considerably (1.04%–15.4%). The initial lexicon of the morphological analyzer lacked morphemes that appeared frequently in some corpora because morphological analysis had been tested mainly with newspaper articles.

The precision of acquired morphemes was high (97.4%–99.3%), and the number of examples used for acquisition was as little as 4–9. These results are astonishing considering that Mori and Nagao (1996) ignored candidates that appeared less than 10 times (because they were unreliable).

Table 3 shows some acquired morphemes. As expected, the overwhelming majority were nouns (93.0%–100%) and katakana morphemes (80.7%–91.6%). Some were mixed-character morphemes (“ソフ倫” and “シャ乱 Q”), which cannot be recognized by character-type based heuristics, and slang words (“腐女子,” “ヲタ,” etc.) which did not appear in newspaper articles. Some morphemes were spelling variants of those in the pre-defined dictionary. Uncommon kanji characters were used in basic words (“棄てる” for “捨てる” and “訊く” for “聞く”) and katakana was used to change nuances (“モテる” for “もてる” and “ダンナ” for “旦那”).

Table 4 shows the results of manual evaluation of “diff” blocks. The overwhelming majority of blocks were correctly analyzed with the augmented lexicon (E → C and C → C). On the other hand, adverse effects were observed only in a few blocks (C → E). In conclusion, acquired morphemes improve the quality of morphological analysis.

4.4 Error Analysis

Some short katakana morphemes oversegmented other katakana nouns. For example, “サーバー” (*sâbâ*, server) was wrongly segmented by newly-acquired “サー” (sâ, sir) and preregistered “バー” (*bâ*, bar). Neither the morphological analyzer and the lexicon acquirer could detect this semantic mismatch. Curiously, one example of “サー” (sâ) was actuallly part of “サーバー” (*sâbâ*), which was erro-

nouns because their morphological behaviors were the same as those of nouns. Although demonstrative nouns are closed class morphemes, their katakana forms such as “コレ” (this) were acquired as nouns. The morphological analyzer assumed that demonstrative pronouns were written in hiragana, e.g., “これ,” as they always are in a newspaper.

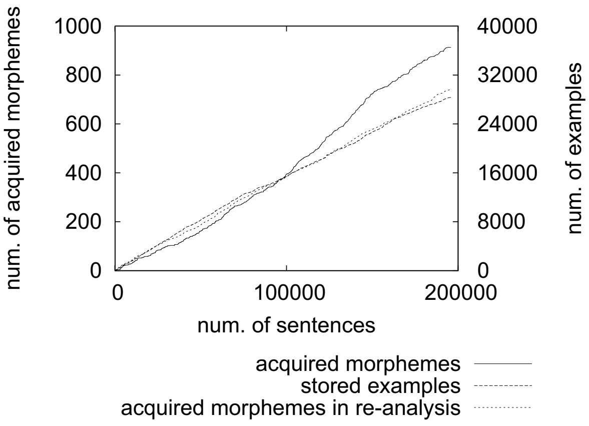


Figure 4: Process of online acquisition

neously segmented when extracting sentences from HTML.

The katakana adjective "イイ" (*i-i*, good), a spelling variant of the basic morpheme "いい," was falsely identified as a noun because its ending "イ" was written in katakana. The morphological analyzer, and hence the lexicon acquirer, assume that the ending of a verb or adjective is written in hiragana. This assumption is reasonable for standard Japanese, but does not always hold when we analyze web texts. In order to recognize unconventional spellings that are widely used in web texts (Nishimura, 2003), more flexible analysis is needed.

4.5 Discussion

It is too costly or impractical to calculate the recall of acquisition, or the ratio of the number of acquired morphemes against the total number of unknown morphemes because it requires human judges to find *undetected* unknown morphemes from a large amount of raw texts.

Alternatively, we examined the ratio against the number of *detected* unknown morphemes. Figure 4 shows the process of online acquisition for the query "JASRAC." The monotonic increase of the numbers of acquired morphemes and stored examples suggests that the vocabulary size did not converge. The number of occurrences of acquired morphemes in re-analysis was approximately the same with the number of examples kept in the storage during acquisition. This means that, in terms of frequency of occurrence, about half of unknown morphemes were acquired. Most unknown morphemes belong to the "long tail" and the proposed method seems to have seized a "head" of the long tail.

Although some previous studies emphasized correct identification of low frequency terms (Nagata, 1999; Asahara and Matsumoto, 2004), it is no longer necessary because very large scale web texts are available today. If a small set of texts needs to be analyzed with high accuracy, we can incorporate similar texts retrieved from the web, to increase the number of examples of unknown morphemes. The proposed method can be modified to check if unknown morphemes detected in the initial set are acquired and to terminate whenever sufficient acquisition coverage is achieved.

5 Related Work

Since most languages delimit words by white-space, morphological analysis in these languages is to segment words into morphemes. For example, Morpho Challenge 2007 (Kurimo et al., 2007) was evaluations of unsupervised segmentation for English, Finnish, German and Turkish.

While Japanese is an agglutinative language, other non-segmented languages such as Chinese and Thai are analytic languages. Among them, Chinese has been a subject of intensive research. Peng et al. (2004) integrated new word detection into word segmentation. They detected new words by computing segment confidence and re-analyzed the inputs with detected words as features.

The Japanese language is unique in that it is written with several different character types. Heuristics widely used in unknown morpheme processing are based on character types. They were also used as important clues in statistical methods. Nagata (1999) integrated a probabilistic unknown word models into the word segmentation model. Uchimoto et al. (2001) incorporated them as feature functions of a Maximum Entropy-based morphological analyzer. Asahara and Matsumoto (2004) used them as a feature of character-based chunking of unknown words using Support Vector Machines.

Mori (1996) extracted words from texts and estimated their POSs using distributional analysis. The appropriateness of a word candidate was measured by the distance between probability distributions of the candidate and a model. In this method, morphological constraints were indirectly represented by distributions.

Nakagawa and Matsumoto (2006) presented a method for guessing POS tags of pre-segmented unknown words that took into consideration all the occurrences of each unknown word in a document. This setting is impractical in Japanese because POS tagging is inseparable from segmentation.

6 Conclusion

We propose a novel method that augments the lexicon of a Japanese morphological analyzer by acquiring unknown morphemes from texts in online mode. Unknown morphemes are acquired with high accuracy and improve the quality of morphological analysis.

Unknown morphemes are one of the main sources of error in morphological analysis when we analyze web texts. The proposed method has the potential to overcome the unknown morpheme problem, but it cannot be achieved without recognizing or being robust over various phenomena such as unconventional spellings and typos. These phenomena are not observed in newspaper articles but cannot be ignored in web texts. In the future, we will work on these phenomena.

Morphological analysis is now very mature. It is widely applied as preprocessing for NLP applications such as parsing and information retrieval. Hence in the future, we aim to use the proposed method to improve the quality of these applications.

References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech tagger. In *Procs. of COLING 2000*, pages 21–27.

Masayuki Asahara and Yuji Matsumoto. 2004. Japanese unknown word identification by character-based chunking. In *Procs. of COLING 2004*, pages 459–465.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *Procs. of LREC-06*, pages 1344–1347.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Procs. of EMNLP 2004*, pages 230–237.

Mikko Kurimo, Mathias Creutz, and Ville Turunen. 2007. Overview of Morpho Challenge in CLEF 2007. In *Working Notes of the CLEF 2007 Workshop*, pages 19–21.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Procs. of The International Workshop on Sharable Natural Language Resources*, pages 22–38.

Shinsuke Mori and Makoto Nagao. 1996. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Procs. of COLING 1996*, pages 1119–1122.

Masaaki Nagata. 1999. A part of speech estimation method for Japanese unknown words using a statistical model of morphology and context. In *Procs. of ACL 1999*, pages 277–284.

Tetsuji Nakagawa and Yuji Matsumoto. 2006. Guessing parts-of-speech of unknown words using global information. In *Procs. of COLING-ACL 2006*, pages 705–712.

Hiroshi Nakagawa and Tatsunori Mori. 2002. A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002*, pages 29–35.

Yukiko Nishimura. 2003. Linguistic innovations and interactional features of casual online communication in Japanese. *Journal of Computer-Mediated Communication*, 9(1).

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Procs. of COLING '04*, pages 562–568.

Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. 2008. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Procs. of IJCNLP-08*, pages 189–196.

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 2001. The unknown word problem: a morphological analysis of Japanese using maximum entropy aided by a dictionary. In *Procs. of EMNLP 2001*, pages 91–99.