A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation

Yasuharu Den*, Junpei Nakamura†, Toshinobu Ogiso‡, Hideki Ogura‡

*Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@cogsci.L.chiba-u.ac.jp

†Department of Computer and Information Sciences,
Tokyo University of Agriculture and Technology
2-24-16 Nakacho, Koganei, Tokyo 184-8588, Japan
naka-jun@fairy.ei.tuat.ac.jp

‡The National Institute for Japanese Language
10-2 Midoricho, Tachikawa, Tokyo 190-8561, Japan
{togiso,ogura}@kokken.go.jp

Abstract

In this paper, we discuss lemma identification in Japanese morphological analysis, which is crucial for a proper formulation of morphological analysis that benefits not only NLP researchers but also corpus linguists. Since Japanese words often have variation in orthography and the vocabulary of Japanese consists of words of several different origins, it sometimes happens that more than one writing form corresponds to the same lemma and that a single writing form corresponds to two or more lemmas with different readings and/or meanings. The mapping from a writing form onto a lemma is important in linguistic analysis of corpora. The current study focuses on disambiguation of *heteronyms*, words with the same writing form but with different word forms. To resolve heteronym ambiguity, we make use of *goshu* information, the classification of words based on their origin. Founded on the fact that words of some goshu classes are more likely to combine into compound words than words of other classes, we employ a statistical model based on CRFs using goshu information. Experimental results show that the use of goshu information considerably improves the performance of heteronym disambiguation and lemma identification, suggesting that goshu information solves the lemma identification task very effectively.

1. Introduction

Automatic morphological analysis is a widely-used technique for the development of NLP systems and linguistically-annotated corpora. Particularly in languages like Japanese, which lacks explicit indication of word boundaries in its writing system, morphological analyzers on computer are indispensable tools for both NLP researchers and corpus linguists. A great amount of studies have attempted to develop software for performing automatic morphological analysis with high accuracy, and several systems for Japanese morphological analysis have been freely available (Asahara and Matsumoto, 2000; Kudo et al., 2004). Although they are definitely useful not only for NLP researchers but also for linguists founded on corpus-based studies, they are insufficient when applied to linguistic analysis of corpora due to inappropriate formulation of the task to be solved.

In traditional morphological analysis on computer, the task is divided into two sub-tasks: i) segmentation of an input string into a sequence of units, and ii) assignment of a part of speech (POS) tag to each segmented unit. Linguistic analysis of corpora, however, requires more information than one provided by these sub-tasks (see e.g., Mizutani, 1983). For instance, Japanese words often have variation in orthography; verb *arawasu* (to express) is written either as "表わす," "表す," or "あらわす," and noun *sakura* (a cherry blossom) either as "桜," "サクラ," or "さくら." In examining the frequency of words occurring in a text, linguists usually want to collapse these variants. Japanese also has lots of *heteronyms*, two or more words that have the identical writing form but different word forms.[1] For instance, nouns *namamono* (raw food) and *seibutu* (a living thing) are both written as "生物." Japanese linguists would treat them as different words. These issues are crucial for linguistic analysis of corpora, but are not handled by traditional Japanese morphological analyzers on computer.

These tasks, though operating in opposite directions—one mapping two or more different writing forms onto a single word form and the other mapping one writing form onto more than one word form, can be regarded as the same problem; that is identification of *lemmas*, i.e., entry words in a dictionary. The task of identifying a lemma corresponding to each segmented unit in an input has been totally ignored in the study of automatic morphological analysis of Japanese, although a few studies on text-to-speech systems addressed this problem as a problem of pronunciation disambiguation (Nagano et al., 2005; Sumita and Sugaya, 2006).

In this paper, we address a proper approach to Japanese

[1] An example of heteronym in English is "bow," which has two different meanings with different sounds, /bou/ and /bau/. In this paper, *writing forms* refer to representation in orthography, which corresponds to spelling in English. *Word forms*, on the other hand, are based on kana-reading and roughly correspond to sounds, although in a few cases, e.g., particles *wa* and *e*, there is dissociation between kana-reading and sound.

morphological analysis, taking lemma identification into account. We first propose an electronic dictionary for Japanese morphological analysis, UniDic, which employs hierarchical definition of word indexes to represent orthographic variants as well as allomorphs. In this hierarchical structure, heteronyms are represented as different nodes with the same index but with different super-nodes. We then propose a statistical model for resolving heteronym ambiguity, making use of *goshu* information, the classification of words based on their origin. Founded on the fact that words of some goshu classes are more likely to combine into compound words than words of other classes, we employ a statistical model based on CRFs using goshu information. We finally present the performance evaluation based on tripartite measures—accuracy of segmentation, POS assignment, and lemma identification. The results show that the use of goshu information considerably improves the performance of heteronym disambiguation and lemma identification, suggesting that goshu information solves the lemma identification task very effectively.

2. Dictionary

For a broad range of applications including not only NLP systems but also corpus-based linguistics and psycholinguistics, we developed an electronic dictionary, UniDic, with the aim of providing a proper tool for Japanese morphological analysis (Den et al., 2007).[2] The dictionary has the following features:

1. The unit for identifying a word is based on the *short unit word* (Maekawa, in press), which provides word segmentation in uniform size, without being harmed by too long words.
2. The indexes for words are defined at several levels, including *lemma*, *form*, and *orthography*, which enables us to represent orthographic variants as well as allomorphs.
3. An extensive amount of phonological information, such as lexical accent and sandhi, is also described and can be utilized in speech research.

For the current study, the second characteristic is relevant. In UniDic, as shown in Figure 1, word indexes are defined at three distinct levels: lemma, form, and orthography.[3] Lemmas roughly correspond to entry words in a traditional dictionary. Those words with the same meaning and grammatical function are represented by the same lemma. When there are more than one allomorphic variant for a single lemma, they are distinguished at the form level; *yahari* and *yappari* are two possible word forms for the same lemma meaning 'likewise.' When the same form under the same lemma has more than one writing form, they are distinguished at the orthography level; "矢張り" and "やはり" are two possible writing forms for the same word form *yahari*. Table 1 shows the number of entries at each of the three levels in our dictionary.

Lemma	Form	Orthography
⟨*yahari*, 矢張り, likewise⟩	*yahari*	矢張り
		やはり
	yappari	やっぱり
⟨*kyouzon*, 共存, coexistence⟩	*kyouzon*	共存
	kyouson	共存
⟨*kakeru*, 掛ける, to hang⟩	*kakeru*	掛ける
		かける
⟨*kakeru*, 欠ける, to chip off⟩	*kakeru*	欠ける
		かける
⟨*raito*, ライト, light⟩	*raito*	ライト
⟨*raito*, ライト, right⟩	*raito*	ライト
⟨*namamono*, 生物, raw food⟩	*namamono*	生物
		なま物
⟨*seibutu*, 生物, a living thing⟩	*seibutu*	生物
⟨*otto*, 夫, one's husband⟩	*otto*	夫
		良人
⟨*ryouzin*, 良人, a good person⟩	*ryouzin*	良人

Figure 1: Hierarchical definition of word indexes

Table 1: Number of entries in our dictionary

Level	# of entries
Lemma	107,623
Form	111,959
Orthography	153,564

By the hierarchical definition of word indexes, (part of) the task of collapsing orthographic variants is naturally resolved. For instance, when the two writing forms of word *yahari* (likewise) are both used in a text, we can assign the same lemma to them with reference to the structure in Figure 1. In contract, in order to assign the same lemma to orthographic variants "掛ける" and "かける," we first have to know whether a certain occurrence of "かける" in a text is a variant of "掛ける" (to hang) or "欠ける" (to chop off). This involves disambiguation of *homographs*, and cannot be solved by the dictionary design alone.

In Figure 1, there are five pairs of entries at the orthography level, each of which have the same writing form: "共存," "かける," "ライト," "生物," and "良人." The first instance "共存" appears on two adjacent lines on the rightmost column in Figure 1. Although they have distinct indexes at the form level, they belong to the same lemma. This is a spurious ambiguity, which linguists usually do not care.

The other four instances involve either homograph or heteronym. The writing form "かける" has two distinct super-nodes at the lemma level, i.e., ⟨*kakeru*, 掛ける, to hang⟩ and ⟨*kakeru*, 欠ける, to chip off⟩, which are different in the standardized Kanji notation as well as meaning—a case of homograph. "ライト" also has two distinct lemma-level indexes, which are distinguished by their meanings, i.e., light vs. right—another case of homograph. The remaining two instances both involve heteronyms. The writing form "生物" has two distinct super-nodes at the lemma level, i.e., ⟨*namamono*, 生物, raw food⟩ and

[2] Freely available at `http://download.unidic.org/`.

[3] Throughout the paper, writing forms, i.e., indexes at the orthography level, are written in Japanese characters, and word forms, i.e., indexes at the form level, are written in Romaji. A lemma is expressed by a triple consisting of a standardized word form, a standardized writing form, and a meaning articulated in English.

Table 2: Distribution of goshu classes in our dictionary and corpora

	Dictionary	Corpora
Class	**# of lemmas**	**# of lemmas**
Wago	24,126 (22.4%)	870,355 (54.9%)
Kango	34,737 (32.3%)	440,674 (27.8%)
Gairaigo	10,119 (9.4%)	35,891 (2.3%)
Konshugo	3,316 (3.1%)	12,564 (0.8%)
Proper name	32,437 (30.1%)	51,140 (3.2%)
Symbol	2,581 (2.4%)	174,049 (11.0%)
Unknown	307 (0.3%)	1,654 (0.1%)
Total	107,623 (100.0%)	1,586,327 (100.0%)

⟨*seibutu*, 生物, a living thing⟩, which have different word forms. "良人" also has two distinct lemma-level indexes with different word forms (*otto* vs. *ryouzin*) and standardized Kanji notations ("夫" vs. "良人").

Resolving these lemma ambiguities is an important issue in our proper approach to Japanese morphological analysis. Since a throughout solution to this problem is hardly achieved, in this paper we focus on the resolution of heteronyms, i.e., the last two cases in the above instances.

3. Model

3.1. The Basic Idea

Our idea for resolving heteronyms in Japanese is to make use of *goshu* information. Goshu is the classification of words based on their origin. In addition to native Japanese words, which have been used in Japan since ancient times, Japanese has imported lots of foreign words from Chinese and some European languages including English. In Japanese linguistics, they are classified into four major classes: i) *wago*, native Japanese words, ii) *kango*, words of Chinese origin, iii) *gairaigo*, words of foreign origin other than Chinese, and iv) *konshugo*, words made of components belonging to different classes. Table 2 shows the distribution of these four goshu classes, as well as other three miscellaneous classes, i.e., proper names, symbols, and unclassified words, in our dictionary and corpora.

Words of different classes may occur in different configurations in a sentence. It is known that short unit words belonging to the kango class often combine into compound words whereas wago words are likely to be used solely. For instance, in a Japanese dictionary *Daijirin* (version 2), there are 30 entries of compound words that contain a kango noun *seibutu* (a living thing), such as *seibutu-heiki* (a biological weapon), *huyuu-seibutu* (plankton), etc. For wago noun *namamono* (raw food), on the other hand, no compound words are found in the dictionary. Hence, if we know that the goshu class of the words adjacent to "生物," which is ambiguous between *namamono* and *seibutu*, is kango, then we can know that this instance of "生物" is likely to be a kango word *seibutu*, not a wago word *namamono*.

In this way, some portion of heteronym ambiguity can be reduced to goshu ambiguity. That is, resolving goshu ambiguity may give us the resolution of heteronyms. To see the applicability of this idea, we next turn to a corpus-based analysis for estimating how many heteronyms can be resolved by using goshu information.

Table 3: Distribution of heteronym categories in our dictionary and corpora

	Dictionary	Corpora
Category	**# of heteronyms**	**# of heteronyms**
Proper name	1,519 (23.9%)	1,990 (1.6%)
Same	2,332 (36.7%)	27,132 (21.4%)
Different	2,114 (33.3%)	75,076 (59.3%)
Partly diff.	383 (6.0%)	22,488 (17.8%)
Total	6,348 (100.0%)	126,686 (100.0%)

Table 4: Size of corpora

Corpus	# of tokens	Training	Test
RWCP	899,347	802,954	96,393
CSJ	458,760	413,168	45,592
GWP	228,220	113,689	114,531
Total	1,586,327	1,329,811	256,516

3.2. A Corpus-based Analysis

To estimate how many heteronyms can be dealt with by goshu disambiguation, we first collected heteronyms in our dictionary, UniDic, and check whether they are distinguishable by goshu classes. Among ca 154 thousands entries at the orthography level, 6,348 entries, or 4.1%, were heteronyms, which had another entry with the same writing form but with a different word form/lemma. We categorized them, according to their lemmas' goshu classes, in the following way:

Proper name: All goshu classes are proper names.

Same: All goshu classes are the same and not proper names.

Different: The goshu classes are different from each other.

Partly different: Some lemmas belong to the same goshu class and others do not.

For instance, the writing form "生物" has two word forms, *namamono* and *seibutu*, the former belonging to the wago class and the latter to the kango class; they are categorized as 'Different.' The two word forms of "対," *tai* (versus) and *tui* (a pair), on the other hand, both fall into the kango class, and, hence, they are categorized as 'Same.' When there are three or more word forms corresponding to the same writing form, they may be divided into two goshu classes, either of which contains more than one word form. "人気" is an example; its word forms *ninki* (popularity) and *zinki* (the traits of a certain area or region) both belong to the kango class, whereas another word form *hitoke* (a sign of life) to the wago class. Thus, they are categorized as 'Partly different.'

On the middle column of Table 3 is the distribution of heteronym categories in our dictionary. It is shown that about one third of the heteronyms in our dictionary can be distinguished by their goshu classes.

We next examined the distribution of heteronym categories in our corpora. Table 4 shows the size and the components

of the corpora. Each number on the second column indicates the number of tokens in a component corpus. (The numbers on the third and the fourth columns will be explained later.) The corpora contains the data from three different sources. The *RWCP Text Corpus* is a corpus of written Japanese collected from newspaper articles. It is widely used in NLP research in Japan, and, in particular, major statistical morphological analyzers, such as ChaSen (Asahara and Matsumoto, 2000) and Mecab (Kudo et al., 2004), are trained on this corpus. The *Corpus of Spontaneous Japanese* (CSJ) (Maekawa, in press) is a corpus of spoken Japanese mainly collected from academic presentations and simulated public speech. Our set of CSJ contains only the data with hand-corrected morphological and clause unit annotations. The government white paper (GWP) data is taken from a written Japanese corpus under development, the *Balanced Corpus of Contemporary Written Japanese* (Maekawa, 2008). All the data are manually segmented into short unit words and annotated with POS tags and lemma-level indexes, and, thus, goshu information can be imported from our dictionary.

In the corpora, consisting of ca 1.6M tokens, 126,686 tokens, or 8.0%, were heteronyms. Categorizing them into the above mentioned four categories, we obtained the distribution shown on the rightmost column of Table 3. It is evident that as much as 60% of the heteronyms found in the corpora can be distinguished by their goshu classes. If we include the cases that can be partly differentiated by goshu information, the percentage reaches 77%. This result suggests that the use of goshu information in heteronym disambiguation is a promising way.

3.3. A Statistical Morphological Analyzer

The corpus-based analysis in the previous section showed that about three quarters of heteronyms may be resolved by using goshu information. To realize our idea, we then constructed a statistical morphological analyzer using goshu information.

The published version of UniDic runs with the morphological analyzer ChaSen (Asahara and Matsumoto, 2000), which employs (an extension of) a hidden Markov model (HMM) to determine the optimal segmentation and POS assignment but can only bring a poor modeling for lemma identification, i.e., the uni-gram probability of lemmas given a writing form. Incorporating statistical information of goshu classes into an HMM-based analyzer, however, is problematic since in HMMs the only way to utilize goshu information is to introduce new tags that consist of combinations of POS tags and goshu classes and this will easily lead us to the data sparseness.

A more recent morphological analyzer Mecab (Kudo et al., 2004) is based on a novel statistical method, conditional random fields (CRFs) (Lafferty et al., 2001), which overcome several problems of HMMs including label bias, length bias, and difficulty in using features that can co-occur at the same position such as the POS tag and the goshu class. With the lexicon contained in ChaSen's standard dictionary and the RWCP Text Corpus as the training data, Mecab is shown to outperform ChaSen in the segmentation and POS assignment tasks.

Table 5: Features concerning goshu information. p, t, f, l, w, and g indicate the part of speech, conjugation type, conjugation form, lemma index, writing form, and goshu class, respectively.

Uni-gram features
$\langle p, g\rangle$
$\langle p, t, f, g\rangle$
$\langle l, g\rangle$
$\langle w, g\rangle$
$\langle p, t, f, l, w, g\rangle$

Bi-gram features (left + right)
$\langle g\rangle$ + $\langle g\rangle$
$\langle p\rangle$ + $\langle g\rangle$
$\langle g\rangle$ + $\langle p\rangle$
$\langle p, g\rangle$ + $\langle p, g\rangle$
$\langle p, t, f, g\rangle$ + $\langle p, t, f, g\rangle$
$\langle w, g\rangle$ + $\langle w, g\rangle$
$\langle p, w, g\rangle$ + $\langle p, w, g\rangle$
$\langle p, t, f, w, g\rangle$ + $\langle p, t, f, w, g\rangle$

In utilizing statistical information of goshu classes imported from UniDic into our corpora, we employed Mecab using goshu-related features shown in Table 5. In addition to the bi-gram of goshu classes, other features that examine only the goshu class of either position, as well as more complex features that introduce combinations of goshu and other morphological information, were also used. Note that for bi-gram features, only the writing forms of words of limited types, e.g., particles, auxiliary verbs, and affixes, were actually used. Ordinary uni-gram and bi-gram features involving no goshu classes were implemented as well.

4. Evaluation

4.1. Experimental Settings

To validate the efficacy of the proposed model, we conducted an experiment on performance evaluation. In traditional studies on morphological analysis, only the performance of the segmentation and POS assignment tasks is evaluated. This, however, is not sufficient as we discussed in Section 1. The performance of the lemma identification task is comparably important in our proper approach to Japanese morphological analysis. Therefore, we evaluated the performance based on tripartite measures—accuracy of segmentation, POS assignment, and lemma identification.

For comparison, the performance of ChaSen- and Mecab-based systems that use no goshu information was also measured. In the ChaSen-based system, only the uni-gram probability of lemmas given a writing form may contribute to the lemma identification task. In the Mecab-based system without goshu information, all the features used in the proposed model, except for the goshu-related ones listed in Table 5, were implemented, and, thus, bi-gram features involving lemma indexes may enhance the performance of lemma identification.

All the three systems were trained on the same set of sentences randomly selected from the corpora described in Section 3.2, and the remaining part of the corpora were held out for the test data. The third and the fourth columns of Table 4 show the sizes of the training and the test data. A total of 1.3M tokens were committed to training.

In order for the comparison not to be affected by the factor of unknown word handling, the vocabulary for the systems was obtained from all entries of UniDic and contained words from not only the training but also the test data.

Table 6: Performance of the three systems on the segmentation, POS assignment, and lemma identification tasks for the three test corpora. In each system's performance, the top line indicates the recall, the center line the precision, and the bottom line the F-score, respectively.

	RWCP			CSJ			GWP		
	segment	POS	lemma	segment	POS	lemma	segment	POS	lemma
ChaSen (w/o goshu)	99.20	97.90	97.13	99.25	97.31	96.48	99.62	98.92	98.66
	99.35	98.05	97.27	99.30	97.36	96.53	99.67	98.97	98.72
	99.27	97.98	97.20	99.27	97.33	96.50	99.65	98.95	98.69
Mecab (w/o goshu)	99.65	98.79	98.00	99.57	98.22	97.37	99.86	99.34	99.01
	99.66	98.80	98.01	99.70	98.35	97.50	99.88	99.35	99.02
	99.66	98.80	98.01	99.63	98.29	97.44	99.87	99.35	99.02
Mecab (w goshu)	99.62	98.79	98.47	99.61	98.27	97.76	99.86	99.36	99.22
	99.66	98.83	98.52	99.71	98.37	97.87	99.86	99.36	99.21
	99.64	98.81	98.49	99.66	98.32	97.82	99.86	99.36	99.21

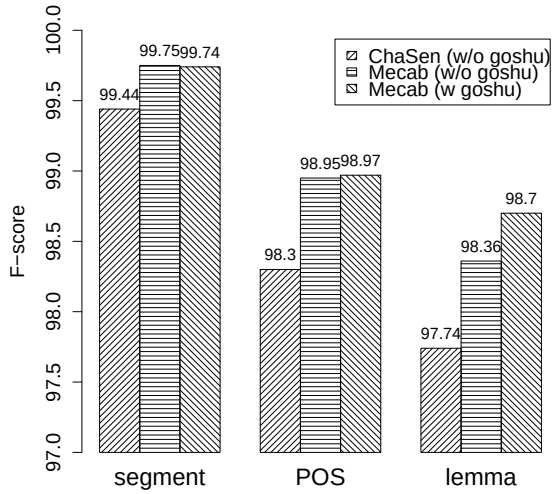

Figure 2: F-scores for the three test corpora in total

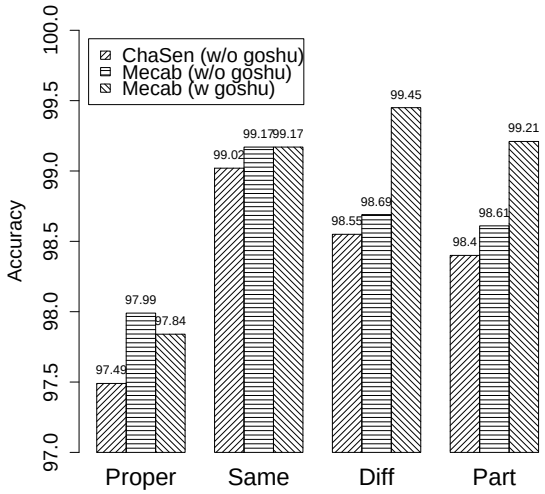

Figure 3: Accuracy of heteronym disambiguation for the three test corpora in total

4.2. Results

Table 6 shows the performance of the three systems on the segmentation, POS assignment, and lemma identification tasks for the three test corpora. In each system's performance, the top line indicates the recall, the center line the precision, and the bottom line the F-score, i.e., the unweighted harmonic mean of the recall and the precision, respectively. Figure 2 also shows the F-scores of the three systems for the three test corpora in total.[4]

Clearly, the Mecab-based systems outperformed the ChaSen-based system in any task. More significantly, the use of goshu information considerably enhanced the performance in the lemma identification task. Although the two versions of a Mecab-based system were comparable in the segmentation and POS assignment tasks, the system with goshu information performed better in the lemma identification task than the one without goshu information; the improvement in the F-score was about 0.34% in total—0.19% to 0.48% depending on the test corpus.

[4]When we removed from the vocabulary the words only appearing in the test data, the F-score of the proposed system fell to 98.76%, 97.82%, and 97.50% in each task, respectively. Note, however, that we did not use specific features that may be effective for unknown words, such as the one used in (Kudo et al., 2004).

To see more closely how our heteronym disambiguation benefited the lemma identification task, we calculated the accuracy of heteronym disambiguation for each of the four heteronym categories defined in Section 3.2. Table 7 shows the accuracy of the three systems for the three test corpora. For each system, the accuracy for the four heteronym categories, i.e., 'Proper name,' 'Same,' 'Different,' and 'Partly different,' are shown. Figure 3 also shows the accuracy of the three systems for the three test corpora in total.

Although for the first two categories the performance of the proposed system was not necessarily superior to the other systems, the accuracies for the 'Different' and 'Partly different' categories were better in the system that uses goshu information than in the systems that do not. When heteronym ambiguity can be totally reduced to goshu ambiguity, the improvement in the accuracy, compared with the Mecab-based system without goshu information, was 0.76% in total—0.53% to 1.22% depending on the test corpus; an improvement of 0.6% in total—0.28% to 1.36% depending on the test corpus—was also achieved, when heteronyms can be partly differentiated by goshu classes.

These results suggest that our model for heteronym disambiguation using goshu information is an effective way to solve the lemma identification task.

Table 7: Accuracy of heteronym disambiguation of the three systems for the three test corpora relative to the four heteronym categories, i.e., 'Proper name,' 'Same,' 'Different,' and 'Partly different'

	RWCP				CSJ				GWP			
	Proper	Same	Diff	Part	Proper	Same	Diff	Part	Proper	Same	Diff	Part
ChaSen (w/o goshu)	97.65	99.43	98.85	98.93	98.21	98.34	98.51	97.81	93.62	98.58	97.54	96.36
Mecab (w/o goshu)	98.15	99.57	98.95	99.13	98.21	98.37	98.57	97.98	94.68	99.13	97.88	96.75
Mecab (w goshu)	98.04	99.56	99.61	99.41	98.21	98.35	99.10	99.34	93.62	99.19	99.10	97.63

4.3. Discussion

We have already achieved very high performance in heteronym disambiguation and lemma identification. However, there is still some room for improvement.

One crucial deficit of our approach is that it is not applicable to the cases where heteronyms cannot be resolved by goshu classes. The 'Same' and 'Proper name' categories are such cases. Although the accuracy for the 'Same' category shown in Figure 3 was quite high (over 99%), this would certainly be because of biased distribution of heteronyms in this category since the ChaSen-based system, which uses only the uni-gram probability of lemmas, performed comparably well.

To overcome this weakness, it may be effective to look at broader context beyond bi-gram. Yarowsky (1996) proposed such method in his attempt to disambiguate homographs in English. Sumita and Sugaya (2006), particularly focusing on the disambiguation of the pronunciation of proper names, applied a similar method using the Web as a training source. The applicability of these methods should be investigated in the future.

For full treatment of lemma identification, other kinds of lemma ambiguity should also be addressed. Homographs, discussed in Section 2, are a typical example. Our method proposed in this paper, however, may not be applicable to homograph disambiguation since in most cases homographs fall into the same goshu class. Investigation of this problem is left for the future study.

5. Conclusion

In this paper, we discussed the problem of lemma identification in order to provide a proper tool for Japanese morphological analysis that benefits not only NLP researchers but also corpus linguists. To tackle the problem, we proposed a heteronym disambiguation method using goshu information, the classification of words based on their origin. Utilizing an electronic dictionary for morphological analysis that contains goshu information, we constructed a statistical morphological analyzer based on CRFs for resolving heteronym ambiguity. The experimental results showed that the use of goshu information considerably improves the performance of heteronym disambiguation and lemma identification. It is suggested that goshu information can solve the lemma identification task very effectively.

6. Acknowledgment

This work is supported by a Grant-in-Aid for Scientific Research on Priority-Area Research, *Japanese Corpus*, led by Kikuo Maekawa, supported by the Ministry of Education, Culture, Sports, Science and Technology.

7. References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech tagger. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 21–27, Saarbrücken, Germany.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22:101–123.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, MA.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*, pages 101–102, Hyderabad, India.

Kikuo Maekawa. in press. Analysis of language variation using a large-scale corpus of spontaneous speech. In Shu-Chuan Tseng, editor, *Linguistic patterns in spontaneous speech*, Language and Linguistics Monograph Series. Institute of Linguistics, Academia Sinica, Taipei.

Shizuo Mizutani. 1983. *Vocabulary (in Japanese)*, volume 2 of *Asakura New Series on Japanese*. Asakura-Shoten, Tokyo.

Tohru Nagano, Shinsuke Mori, and Masafumi Nishimura. 2005. A stochastic approach to phoneme and accent estimation. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005)*, pages 3293–3296, Lisbon, Portugal.

Eiichiro Sumita and Fumiaki Sugaya. 2006. Word pronunciation disambiguation using the Web. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 165–168, New York City.

David Yarowsky. 1996. Homograph disambiguation in text-to-speech synthesis. In Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, pages 159–174. Springer-Verlag, New York.