Semantic Classification of Automatically Acquired Nouns using Lexico-Syntactic Clues

Yugo Murawaki
Graduate School of Informatics
Kyoto University
murawaki@nlp.kuee.kyoto-u.ac.jp

Sadao Kurohashi
Graduate School of Informatics
Kyoto University
kuro@i.kyoto-u.ac.jp

Abstract

In this paper, we present a two-stage approach to acquire Japanese unknown morphemes from text with full POS tags assigned to them. We first acquire unknown morphemes only making a morphology-level distinction, and then apply semantic classification to acquired nouns. One advantage of this approach is that, at the second stage, we can exploit syntactic clues in addition to morphological ones because as a result of the first stage acquisition, we can rely on automatic parsing. Japanese semantic classification poses an interesting challenge: proper nouns need to be distinguished from common nouns. It is because Japanese has no orthographic distinction between common and proper nouns and no apparent morphosyntactic distinction between them. We explore lexico-syntactic clues that are extracted from automatically parsed text and investigate their effects.

1 Introduction

A dictionary plays an important role in Japanese morphological analysis, or the joint task of segmentation and part-of-speech (POS) tagging (Kurohashi et al., 1994; Asahara and Matsumoto, 2000; Kudo et al., 2004). Like Chinese and Thai, Japanese does not delimit words by white-space. This makes the first step of natural language processing more ambiguous than simple POS tagging. Accordingly, morphemes in a pre-defined dictionary compactly represent our knowledge about both segmentation and POS.

One obvious problem with the dictionary-based approach is caused by unknown morphemes, or morphemes not defined in the dictionary. Even though, historically, extensive human resources were used to build high-coverage dictionaries (Yokoi, 1995), texts other than newspaper articles, in particular web pages, contain a large number of unknown morphemes. These unknown morphemes often cause segmentation errors. For example, morphological analyzer JUMAN 6.0[1] wrongly segments the phrase "さっぽろ駅" (*saQporo eki*, "Sapporo Station"), where "さっぽろ" (*saQporo*) is an unknown morpheme, as follows:

"さ" (*sa*, noun-common, "difference"),
"っ" (*Q*, UNK), "ぽ" (*po*, UNK),
"ろ" (*ro*, noun-common, "sumac") and
"駅" (*eki*, noun-common, "station"),

where UNK refers to unknown morphemes automatically identified by the analyzer. Such an erroneous sequence has disastrous effects on applications of morphological analysis. For example, it can hardly be identified as a LOCATION in named entity recognition.

One solution to the unknown morpheme problem is unknown morpheme acquisition (Mori and Nagao, 1996; Murawaki and Kurohashi, 2008). It is the task of automatically augmenting the dictionary by acquiring unknown morphemes from text. In the above example, the goal is to acquire the morpheme "さっぽろ" (*saQporo*) with the POS tag "noun-location name." However, unknown morpheme acquisition usually adopts a coarser POS tagset that only represents the morphology level distinction among noun, verb and adjective. This means that "さっぽろ" (*saQporo*) is acquired as just a noun and that the semantic label "location name" remains to be assigned. The reason only the morphology level distinction is made is

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

Coling 2010: Poster Volume, pages 876–884,
Beijing, August 2010

that the semantic level distinction cannot easily be captured with morphological clues that are exploited in unknown morpheme acquisition.

In this paper, we investigate the remaining problem and introduce the new task of semantic classification that is to be applied to automatically acquired nouns. In this task, we can exploit syntactic clues in addition to morphological ones because, as a result of acquisition, we can now rely on automatic parsing. For example, since text containing "さっぽろ" (*saQporo*, noun-*unclassified*) is correctly segmented, we can extract not only the phrase "*saQporo* station," but the tree fragment "ϕ go to *saQporo*," and we can determine its semantic label.

Japanese semantic classification poses an interesting challenge: proper nouns need to be distinguished from common nouns. Like Chinese and Thai, Japanese has no orthographic distinction between common and proper nouns as there is no such thing as capitalization. In addition, there seems no morphosyntactic (i.e. grammatical) distinction between them.

In this paper, we explore lexico-syntactic clues that can be extracted from automatically parsed text. We train a classification model on manually registered nouns and apply it to automatically acquired nouns. We then investigate the effects of lexico-syntactic clues.

2 Semantic Classification Task

2.1 Two-Stage Approach to Unknown Morpheme Acquisition

Our goal is to identify unknown morphemes in unsegmented text and assign POS tags to them. In this section, we omit the details of boundary identification (segmentation) and review the Japanese POS tagset to see why we propose a two-stage approach to assign full POS tags.

The Japanese POS tagset derives from traditional grammar. It is a mixture of several linguistic levels: morphology, syntax and semantics. In other words, information encoded in a POS tag is more than how the morpheme behaves in a sequence of morphemes. In fact, POS tags given to pre-defined morphemes are useful for applications of morphological analysis, such as dependency parsing (Kudo and Matsumoto, 2002), named entity recognition (Asahara and Matsumoto, 2003; Sasano and Kurohashi, 2008) and anaphora resolution (Iida et al., 2009; Sasano and Kurohashi, 2009). In these applications, POS tags are incorporated as features for models.

On the other hand, the mixed nature of the POS tagset poses a challenge to unknown morpheme acquisition. Previous approaches (Mori and Nagao, 1996; Murawaki and Kurohashi, 2008) directly or indirectly reply on morphology, or our knowledge on how a morpheme behaves in a sequence of morphemes. This means that semantic level distinction is difficult to make in these approaches, and in fact, is left unresolved. To be specific, nouns are only distinguished from verbs and adjectives but they have subcategories in the original tagset. These are what we try to classify acquired nouns into in this paper.

2.2 Semantic Labels

The Japanese noun subcategories may require an explanation since they are different from the English ones (Marcus et al., 1993) in many respects. Singular and mass nouns are not distinguished from plural nouns because Japanese has no grammatical distinction between them. More importantly for this paper, proper nouns have subcategories such as person name, location name and organization name in addition to the distinction from common nouns. These subcategories provide important information to named entity recognition among other applications. For proper nouns, we adopt these subcategories as semantic labels in our task.

In contrast to proper nouns, common nouns have only one subcategory "common." However, we consider that subcategories of common nouns similar to those of proper nouns are useful for, for example, anaphora resolution (Sasano and Kurohashi, 2009). We adopt the "categories" of morphological analyzer JUMAN, with which common nouns in its dictionary are annotated. There are 22 "categories" including `PERSON`, `ORGANIZATION` and `CONCEPT`. We collapse these "categories" into coarser semantic labels that roughly correspond to those for proper nouns. To sum up, we define 9 semantic labels as shown

Table 1: List of semantic labels.

labels	P/C	sources[1]	manually registered nouns	automatically acquired nouns
PSN-P	proper	subPOS:person name	松井 (*matsui*, a surname) ジョージ (*jôji*, "George")	佐祐理 (*sayuri*, a given name) キョン (*kyoN*, a nickname)
LOC-P		subPOS:place name	京都 (*kyouto*, "Kyoto") ドイツ (*doitsu*, "Germany")	アキバ (*akiba*, "Akihabara") ワイキキ (*waikiki*, "Waikiki")
ORG-P		subPOS:organization name	日銀 (*nichigin*, a bank) NHK (a broadcaster)	マツダ (*matsuda*, "Mazda") ヤフー (*yahû*, "Yahoo")
OTH-P		subPOS:proper noun	平成 (*heisei*, an era name) スラブ (*surabu*, "Slav")	ジプシー (*jipushî*, "Gypsy")
PSN-C	common	category:PERSON	先生 (*seNsei*, "teacher") スタッフ (*sutaQfu*, "staff")	メル友 (*merutomo*, "keypal") ニート (*nîto*, "NEET")
LOC-C		category:PLACE-*[2]	職場 (*shokuba*, "office") カフェ (*kafe*, "cafe")	囲炉裏 (*irori*, "hearth") 圃場 (*hojou*, "farm field")
ORG-C		category:ORGANIZATION	政府 (*seifu*, "government") チーム (*chîmu*, "team")	メーカ (*mêka*, "manufacturer") 弊所 (*heisho*, "our office")
ANI-C		category:ANIMAL and category:ANIMAL-PART	犬 (*inu*, "dog") 顔 (*kao*, "face")	チワワ (*chiwawa*, "Chihuahua") マンタ (*maNta*, "manta")
OTH-C		other categories	主張 (*shuchou*, "argument") 枕 (*makura*, "pillow")	甚平 (*jiNbei*, a kind of clothing) 着メロ (*chakumero*, "ringtone")

[1] A subPOS refers to a subcategory of noun. For example, PSN-P corresponds to the POS tag "noun-person name".
[2] category:PLACE-INSTITUTION, category:PLACE-INSTITUION PART and others.

in Table 1.

2.3 Related Tasks

A line of research is dedicated to identify unknown morphemes with varying degrees of identification. Asahara and Matsumoto (2004) only focus on boundary identification (segmentation) of unknown morphemes. Mori and Nagao (1996), Nagata (1999) and Murawaki and Kurohashi (2008) assign POS tags at the morphology level. Uchimoto et al. (2001) assign full POS tags but unsurprisingly the accuracy is low. Nakagawa and Matsumoto (2006) also assign full POS tags. They address the fact that local information used in previous studies is inherently insufficient and present a method that uses global information, in other words, takes into consideration all occurrences of each unknown word in a document. They report an improvement in tagging proper nouns in Japanese.

A related task is named entity recognition (NER). It can handle a named entity longer than a single morpheme and is usually formalized as a chunking problem. Since Japanese does not delimit words by white-space, the unit of chunking can be a character (Asahara and Matsumoto, 2003; Kazama and Torisawa, 2008) or a morpheme (Sasano and Kurohashi, 2008). In either case, NER models encode the output of morphological analysis and therefore are affected by its errors. In fact, Saito et al. (2007) report that a majority of unknown named entities (those never appear in a training corpus) contain unknown morphemes as their constituents and that NER models perform poorly on them. A straightforward solution to this problem would be to acquire unknown morphemes and to assign semantic labels to them.

Another related task is supersense tagging (Ciaramita and Johnson, 2003; Curran, 2005; Ciaramita and Altun, 2006). A supersense corresponds to one of the 26 broad categories defined by WordNet (Fellbaum, 1998). Each noun synset is associated with a supersense. For example, "chair" has supersenses PERSON, ARTIFACT and ACT because it belongs to several synsets.

Since supersense tagging is studied in English, it differs from our task in several respects. In English, the distinction between common and proper nouns is clear. In fact, the tagging models can use POS features even for unknown nouns. In addition, the syntactic behavior of English nouns is different from that of Japanese nouns (Gil, 1987). Definiteness is not marked in Japanese as it lacks determiners (e.g. "the" and "a"), and Japanese has no obligatory plural marking. On the other hand, Japanese obligatorily uses numeral classifiers to indicate the count of nouns, as in

(1) *saN satsu no hoN*
three CL GEN book
three volumes of books, or three books,

where "*satsu*" is a numeral classifier for books. A number together with its numeral classifier forms a numeral quantifier. Numeral quantifiers would be informative about the semantic categories of nouns. Note that Japanese shares the above features with Chinese and Thai. Our findings in this paper may hold for these languages.

3 Proposed Method

3.1 Lexico-Syntactic Clues

In the task of semantic classification, we can exploit syntactic clues in addition to morphological ones. As a result of unknown morpheme acquisition, text containing acquired morphemes, or former unknown morphemes, is correctly segmented. Now we can treat automatic parsing as (at least partly) reliable with regard to acquired morphemes.

For noun X, we use the following sets of features for classification.

call: noun phrase Y that appears in a pattern like "Y called X" and "Y such as X," e.g. "`call`:*kuni*" from

X to iu kuni
X QT call country
a country called X.

cf: predicate with a case marker with which it takes X as an argument, e.g. "`cf`:*tooru*:*wo*" from

X wo tooru
X ACC pass
ϕ pass through X.

demo: demonstrative that modifies X, e.g. "`demo`:*kono*" from "*kono X*" (this X) and "`demo`:*doNna*" from "*doNna X*" (what kind of X).

ncf1: noun phrase which X modifies with the genitive case marker "*no*," e.g. "`ncf1`:*heya*" from

X no heya
X GEN room
X's room.

ncf2: noun phrase that modifies X with the genitive case marker "*no*," e.g. "`ncf2`:*subete*" from

subete no X
all GEN X
all X.

suf: suffix or suffix-like noun that follows X, e.g. "`suf`:*saN*" from "X *saN*" (Mr./Ms. X) and "`suf`:*eki*" from "X *eki*" (X station).

Using automatically parsed text to extract syntactic features has an advantage. Since no manual annotation is necessary, we can utilize a huge raw corpus. On the other hand, parsing errors are inevitable. However, we can circumvent this problem by using the constraints of Japanese dependency structures: head-final and projective. The simplest example is the second last element of a sentence, which always depends on the last element. With these constraints, we can focus on syntactically unambiguous dependency pairs and extract syntactic features accurately. We follow Kawahara and Kurohashi (2001) to extract a pair of an argument noun and a predicate (**cf**), and Sasano et al. (2004) to extract a pair of nouns connected with the genitive case marker "*no*" (**ncf1** and **ncf2**).

Noun X can be part of a compound noun. We leave it for named entity recognition. Except for **suf**, we extract features only when X alone forms a word. Similarly, we extract **suf** features only when X and a suffix alone form a noun phrase.

For **call**, **ncf1**, and **ncf2**, we generalize numerals within noun phrases. For "*hoN*" (book) in example 1, we extract the feature "`ncf2`:<NUM>*satsu*."

3.2 Instances for Classification

Now that features are extracted for each noun, the question is how to combine them together to make an instance for classification. One factor we need to consider is polysemy: a noun can be a person name in one context and a location name in another. If we combine features extracted from the whole corpus, they may represent several semantic labels.

Modeling a mixture of semantic labels might be a solution, but we do not take this approach on the grounds that each occurrence of a noun corresponds to a single semantic label.

In our strategy, we perform classification multiple times for each noun and aggregate the results at the end. The features for each classification are extracted from a relatively small subset of a corpus where the noun is supposedly consistent in

terms of semantic labels. In the field of named entity recognition, it is known that label consistency holds strongly at the level of a document and less strongly across different documents (Krishnan and Manning, 2006). Thus we start with a document and gradually cluster related documents until a sufficient number of features are obtained. For the specific procedures we took in the experiments, see Section 4.1.

3.3 Training Data

Following unknown morpheme acquisition (Murawaki and Kurohashi, 2008), we create training data using manually registered nouns, for which we can obtain correct semantic labels. We perform the same procedure as above to make instances of registered nouns.

Some registered nouns are tagged with more than one semantic label, which we call "explicit polysemy." We drop them from the training data. The remaining problem is "implicit polysemy." Nouns are sometimes used with an uncovered sense. In preliminary experiments, we found that a typical case of implicit polysemy was that a proper noun derived from a basic noun. To alleviate this problem, we use an NE tagger for filtering. We run an NE tagger over a small portion of the corpus and extract common nouns that are frequently tagged as named entities. Then we remove these nouns from the training data.

We also drop nouns that appear extremely frequently such as "人" (*hito*, "person"), "事" (*koto*, "thing") and "私" (*watashi*, "I"[2]). Since acquired nouns to be classified are typically low frequency morphemes, they would not behave similarly to these basic nouns.

3.4 Classifier

To assign a semantic label to each instance, we use a multiclass discriminative classifier. The input it takes is an instance that is represented by a feature vector $x \in \mathbb{R}^d$. The output is one semantic label $y \in Y$, where Y is the set of semantic labels.

We use a linear classifier. It has a weight vector $w_y \in \mathbb{R}^d$ for each y and outputs y that maximizes the inner product of w_y and x.

$$y = \operatorname*{argmax}_y \langle w_y, x \rangle.$$

Several methods have been proposed to estimate weight vector w_y from training data. We use online algorithms because they are easy to implement and scale to huge instances. We try the Perceptron family of algorithms.

4 Experiments

4.1 Settings

We used JUMAN for morphological analysis and KNP[3] for dependency parsing. The dictionary of JUMAN was augmented with automatically acquired morphemes (Murawaki and Kurohashi, 2008). The number of manually registered morphemes was 120 thousands while there were 13,071 acquired morphemes, of which 12,615 morphemes were nouns.

We used a web corpus that was compiled through the procedures proposed by Kawahara and Kurohashi (2006). It consisted of 100 million pages.

We first extracted features from the web corpus. To keep the model size manageable, we used 447,082 features that appeared more than 100 times in the corpus.

We constructed training data from manually registered nouns and test data from automatically acquired nouns. For each noun, we combined text together until the number of features grew to more than 100. We started with a single web page, then merge pages that share a domain name and finally clustered texts across different domains. We split the web corpus into 40 subcorpora and applied this procedure in parallel. We used Bayon[4] for clustering domain texts. We sequentially read texts and applied the repeated bisections clustering every time some 5,000 pages were appended. The vectors for clustering were nouns, both registered and acquired, with their tf-idf scores. We obtained 4,843,085 instances for 10,613 registered nouns and 196,098 instances for 2,556 acquired nouns.

[2] Japanese personal pronouns are treated as common nouns because they show no special morphosyntactic behavior.

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

[4] http://code.google.com/p/bayon/

Table 2: Results of semantic classification.

learning algorithms	acquired nouns	registered nouns
Averaged Perceptron	86.40% (432 / 500)	88.59% (123,113 / 138,971)
Passive-Aggressive	**87.00**% (435 / 500)	**91.68**% (127,407 / 138,971)
Confidence-Weighted	85.20% (426 / 500)	89.66% (124,604 / 138,971)
baseline[1]	69.60% (348 / 500)	79.14% (109,980 / 138,971)

[1] assign OTH−C to all instances.

Table 3: Examples of aggregated instances.

acquired nouns	instances	labels
ヒカル (*hikaru*, a person name)	84	PSN−P:58.33%, PSN−C:41.67%
チワワ (*chiwawa*, "Chihuahua")	128	ANI−C:54.69%, OTH−C:45.31%
かみさん (*kamisaN*, colloq. "wife")	131	PSN−C:100%
ラスベガス (*rasubegasu*, "Las Vegas")	136	LOC−P:97.06%, LOC−C:2.94%
アップル (*aQpuru*, "Apple/apple")	187	ORG−P:63.10%, PSN−C:34.76%, OTH−C:2.14%
メルマガ (*merumaga*, abbr. of "mail magazine")	1,622	OTH−C:99.32%, LOC−C:0.55%, PSN−C:0.06%

In order to handle polysemy, we evaluated semantic classification on an instance-by-instance basis. We randomly selected 500 instances from the test data and manually assigned the correct labels to them. For comparison purposes, we also classified registered nouns. We split the training data: 829 nouns or 138,971 instances for testing and the rest for training.

We trained the model with three online learning algorithms, (1) the averaged version (Collins, 2002) of Perceptron (Crammer and Singer, 2003), (2) the Passive-Aggressive algorithm (Crammer et al., 2006), and (3) the Confidence-Weighted algorithm (Crammer et al., 2009). For Passive-Aggressive algorithm, we used *PA-I* and set parameter C to 1. For Confidence-Weighted, we used the single-constraint updates. All algorithms iterated five times through the training data.

4.2 Results

Table 2 shows the results of semantic classification. All algorithms significantly improved over the baseline. As suggested by the gap in accuracy between acquired and registered nouns in the baseline method, the label distribution of the training data differed from that of the test data, but the decrease in accuracy was smaller than expected.

The Passive-Aggressive algorithm performed best on both acquired and registered nouns. For the rest of this paper, we report the results of the Passive-Aggressive algorithm.

Table 3 shows aggregated instances of some acquired nouns. Although classification sometimes failed, correct labels took the majority. However, it is noticeable that PSN−P was frequently misidentified as PSN−C while PSN−C was correctly identified. This phenomenon is clearly seen in the confusion matrix (Table 4). Half of PSN−P instances were misidentified as PSN−C but the percentage of errors in the opposite direction was just above 9%. We will investigate this in the next section.

4.3 Discussion

Our interest is in determining what kinds of features are effective in semantic classification. We first performed standard ablation experiments. We trained a series of models on the training data after removing each feature set. The training and test data were the same with those in Section 4.1.

Table 5 shows the results of ablation experiments. Significant decreases in accuracy are observed in the **cf** dataset. This is easily explained by the fact that more than half of features belonged to **cf**. The ratio of **ncf1** was much the same with that of **ncf2**, but the removal of **ncf1** resulted in a worse performance in classifying registered nouns than that of **ncf2**. This means that a modifiee of a noun explains more about the noun than its modifier.

The ablation experiments cannot capture interesting properties of features because each feature set has a great diversity within it. Next, we directly examine features instead. Since we use a simple linear classifier, a feature has $|Y|$ corresponding weights, each of which represents how likely a noun belongs to label y. For example, features whose weights for PSN−C are the largest

Table 4: Confusion matrix of acquired nouns.

		Actual								
		PSN-P	LOC-P	ORG-P	OTH-P	PSN-C	LOC-C	ORG-C	ANI-C	OTH-C
Predicted	PSN-P	16		1		4				1
	LOC-P									1
	ORG-P			4						
	OTH-P									
	PSN-C	16				39			1	2
	LOC-C	2	2	1			10			4
	ORG-C									2
	ANI-C								28	
	OTH-C	3	1	1		1	13		9	338

Table 5: Results of ablation experiments.

feature set	ratio[1]	acquired nouns	registered nouns
-call	0.23%	87.60% (438 / 500)	91.58% (127,276 / 138,971)
-cf	54.84%	84.80% (424 / 500)	88.96% (123,630 / 138,971)
-demo	2.40%	**88.00**% (440 / 500)	91.38% (126,996 / 138,971)
-ncf1	19.03%	87.20% (436 / 500)	89.23% (124,008 / 138,971)
-ncf2	18.40%	85.60% (428 / 500)	91.54% (127,220 / 138,971)
-suf	5.10%	87.40% (437 / 500)	91.30% (126,889 / 138,971)
all		87.00% (435 / 500)	**91.68**% (127,407 / 138,971)

[1] The proportion of each feature set that appears in the instances of the test data.

include:

- cf:*nakusu*:*wo* ("ϕ lose X to the disease"),
- cf:*oshieru*:*ni* ("ϕ_1 teach X ϕ_2"),
- ncf2:*ooku* ("many/much X"), and
- ncf2:<NUM>*niN* (X is modified by <NUM> plus a numeral classifier for persons).

As briefly mentioned in Section 2.3, Japanese numeral quantifiers received scholarly attention in the fields of linguistic philosophy and linguistics in relation to the count/mass distinction (Quine, 1969; Gil, 1987). In our feature sets, numeral quantifiers typically appear as **ncf2**, e.g. "ncf2:<NUM>*niN*." The weights given to them demonstrate their effectiveness in semantic classification. They discriminate common nouns from proper nouns as the weights given to common nouns are larger with wide margins. It is not surprising because, say, the phrase "two Johns" is semantically acceptable but extremely rare in reality. They are also informative about the distinction among PSN, LOC and others. For example, the classifier "*niN*" for persons suggest the noun in question is a person while "*keN*" for houses would modify a location-like noun. However, we found quite a few "noises" about these features in data. The modifiee of a numeral expression is not always the noun to be counted, as demonstrated by the following example:

(2) *saN niN no moNdai*
three CL GEN problem
matters among the three persons.

From the above, the feature "ncf2:<NUM>*niN*" is extracted although "*moNdai*" is OTH-C. Theis "noise" is attributed to the genitive case marker "*no*" because it can denote a wide range of relations between two nouns. We might be able to avoid this problem if we focus on "floating" numeral quantifiers. A floating numeral quantifier has no direct dependency relation to the noun to be counted, as in

(3) *seito ga saN niN keQseki shita*
student NOM three CL absence do
three students were absent,

where the numeral quantifier modifies the verb phrase instead of the noun. Further work is needed to anchor floating numeral quantifiers since they bring a different kind of ambiguity themselves (Bond et al., 1998).

Closely related to numeral quantifiers are quantificational nouns that appear as "ncf2:*ooku*" ("many/much"), "ncf2:*subete*" ("all") and others. They distinguish common nouns from proper

nouns but does not make a further classification. The same is true of other numeral expressions such as "cf:*hueru*:*ga*" ("X increase in number") and "cf:*nai*:*ga*" ("there is no X" or "X do not exist"). We found that, other than numeral expressions, some features distinguished common nouns from proper nouns because they indicated the noun denoted an attribute. Such features include "cf:*naru*:*ni*" ("ϕ become X") and "cf:*kaneru*:*wo*" ("ϕ double as X").

We expected that demonstratives (**demo**) served similar functions to quantificational expressions, but it turned out to be more complex. The distal demonstrative "*ano*" ("that") often modifies proper nouns to give emphasis. In fact, the model gave larger weights to proper nouns. On the other hand, interrogative demonstratives such as "*dono*" ("which") and "*doNna*" ("what kind of") are rarely used with proper nouns although semantically acceptable.

As seen above, there is an abundant variety of features that distinguish common nouns from proper nouns. Also, it is not difficult to make a distinction among PSN, LOC and others although the far largest cluster OTH−C sometimes absorbs other instances. The remaining question is how to distinguish proper nouns from common nouns, or specifically PSN−P from PSN−C. We examined features that gave larger weights to PSN−P than to PSN−C. They generally had smaller margins in weights than those which distinguish PSN−C from PSN−P. Among them, features such as "cf:*utau*:*ga*" ("X sing") and "cf:*hanasu*:*ni*" ("ϕ talk to X") have no problem with being used for common nouns in terms of both semantics and pragmatics. They seem to have resulted from over-training. There were seemingly appropriate features such as "suf:*saNchi*" ("X's house") and "suf:*seNshu*" (honorific suffix for players), but they were not ubiquitous in the corpus. PSN−P instances suffered from lack of distinctive features.

One solution to this problem is to combine additional knowledge about person names. For example, a Japanese family name is followed by a given name, and most Chinese names consist of three Chinese characters. However, quite a few person names in the web corpus do not follow the usual patterns of person names because they are handles (or nicknames) and names for fictional characters. Thus it would be desirable to be able to classify person names without additional knowledge.

5 Conclusion

In this paper, we presented the new task of semantic classification of Japanese nouns and applied it to nouns automatically acquired from text. Unlike in unknown morpheme identification in previous studies, we can exploit automatically parsed text. We explored lexico-syntactic clues and investigated their effects. We found plenty of features that distinguished common nouns from proper nouns, but few features worked in the opposite direction. Further work is needed to overcome this bias.

References

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

Asahara, Masayuki and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proc. of HLT/NAACL 2003*, pages 8–15.

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

Bond, Francis, Daniela Kurz, and Satoshi Shirai. 1998. Anchoring floating quantifiers in Japanese-to-English machine translation. In *Proc. of COLING 1998*, pages 152–159.

Ciaramita, Massimiliano and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP 2006*, pages 594–602.

Ciaramita, Massimiliano and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In *Proc. of EMNLP 2003*, pages 168–175.

Collins, Michael. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proc. of EMNLP 2002*, pages 1–8.

Crammer, Koby and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Crammer, Koby, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Crammer, Koby, Mark Dredze, and Alex Kulesza. 2009. Multi-class confidence weighted algorithms. In *Proc. of EMNLP 2009*, pages 496–504.

Curran, James R. 2005. Supersense tagging of unknown nouns using semantic similarity. In *Proc. of ACL 2005*, pages 26–33.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Gil, David. 1987. Definiteness, NP configurationality and the count-mass distinction. In Reuland, Eric J. and Alice G. B. ter Meulen, editors, *The Representation of (In)definiteness*, pages 254–269. MIT Press.

Iida, Ryu, Kentaro Inui, and Yuji Matsumoto. 2009. Capturing salience with a trainable cache model for zero-anaphora resolution. In *Proc. of ACL/IJCNLP 2009*, pages 647–655.

Kawahara, Daisuke and Sadao Kurohashi. 2001. Japanese case frame construction by coupling the verb and its closest case component. In *Proc. of HLT 2001*, pages 204–210.

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

Kazama, Jun'ichi and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proc. of ACL 2008*, pages 407–415, June.

Krishnan, Vijay and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proc. of COLING-ACL 2006*, pages 1121–1128.

Kudo, Taku and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proc. of CONLL 2002*, pages 1–7.

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

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

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Mori, Shinsuke and Makoto Nagao. 1996. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Proc. of COLING 1996*, volume 2, pages 1119–1122.

Murawaki, Yugo and Sadao Kurohashi. 2008. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proc. of EMNLP 2008*, pages 429–437.

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

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

Quine, Willard Van. 1969. *Ontological Relativity and Other Essays*. Columbia University Press.

Saito, Kuniko, Jun Suzuki, and Kenji Imamura. 2007. Extraction of named entities from blogs using CRF. In *Proc. of The 13th Annual Meeting of The Association for Natural Language Processing*, pages 107–110. (in Japanese).

Sasano, Ryohei and Sadao Kurohashi. 2008. Japanese named entity recognition using structural natural language processing. In *Proc. of IJCNLP 2008*, pages 607–612.

Sasano, Ryohei and Sadao Kurohashi. 2009. A probabilistic model for associative anaphora resolution. In *Proc. of EMNLP 2009*, pages 1455–1464.

Sasano, Ryohei, Daisuke Kawahara, and Sadao Kurohashi. 2004. Automatic construction of nominal case frames and its application to indirect anaphora resolution. In *Proc. of COLING 2004*, pages 1201–1207.

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

Yokoi, Toshio. 1995. The EDR electronic dictionary. *Communications of the ACM*, 38(11):42–44.