Semantic Type Disambiguation for Japanese Verbs

Shohei Okada and Kazuhide Yamamoto
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka, Japan
{*okada, yamamoto*}*@jnlp.org*

***Abstract*—The interest has been increasing in recent years in extracting and analyzing evaluations and opinions of service or products from large bodies of text. It is important to classify predicates according to sense because whether or not a statement includes the speaker's opinion depends strongly on its predicate. It is generally assumed that Japanese part-of-speech (POS) for predicates is classified according to sense; however, the POS classifications differ from their semantic classification. On this subject, semantic types, which aim to classify predicates, have been proposed. In this paper, we describe semantic types and present our construction of a disambiguator for Japanese verbs. Specifically, we constructed this disambiguator using a support vector machine by building feature vectors. We used semantic categories of noun and results of morphological analysis for the feature vectors. We then achieved 69.9% accuracy of disambiguation for newspaper articles using 10-fold cross-validation.**

***Keywords*-Word Sense Disambiguation (WSD)**

I. Introduction

The interest in extracting and analyzing evaluations and opinions of services or products from large bodies of text has been increasing in recent years. Whether or not a statement includes the speaker's opinion depends strongly on its predicate. Yang and Cardie [1] said "the majority of opinion expressions involve verb phrases". Therefore, it is important to classify predicates according to sense, and Japanese predicates consist of adjectives and verbs. It is generally assumed that adjectives signify properties or states of people or objects, and while verbs signify observable actions or changes. However, there are verbs that signify properties, such as "優れる (to excel)" and emotions such as "むかつく (to get frustrated)". Therefor, *part-of-Speech* (POS) differs from semantic classification for Japanese predicates.

In their 2011 study, Nakayama and Yamamoto [2] proposed that predicates should be classified according to sense and, accordingly, they defined four *semantic types*: *action*, *change*, *emotion*, and *modification*. According to Nakayama and Yamamoto, all adjectives are of the type *modification*, while verbs are of multiple types. They then annotated all verbs in a dictionary for type, and, in order to retain information, they annotated all possible types that could relate to a given verb. Verbs that have ambiguity of sense can be annotated multiple types. In this case, one of the annotated types can be determined from the context. Okada and Yamamoto [3] investigated the relationship between semantic types and case frame information. They found that the combination of surface case information with the semantic categorization of a noun in a statement is an effective means to determine semantic type.

In this study, therefore, we constructed a semantic type disambiguator, that determines verb type from context based on Nakayama and Yamamoto's classification. We supposed that the disambiguator only utilizes information from the results of morphological analysis and existing language resources. Thus, we were able to directly add type information to the results of morphological analysis like POS tags. Morphological analysis and POS tagging are fundamental operations for *Natural Language Processing* (NLP). By enabling us to consider information that more accurately reflects verb sense using semantic type instead of POS, this study can contribute to Japanese NLP which uses POS as semantic categories. For an example of opinion mining, when applying the method proposed by Scholz and Conrad [4] to Japanese, improvement of precision can be expected using semantic type categorization instead of POS.

In this paper, we first introduce Nakayama and Yamamoto's definition of semantic types and the concept of a *Semantic Type Dictionary*. In the third section, we describe the construction of the semantic type disambiguator. We will evaluate the disambiguator in section IV, before concluding in the final section.

II. Semantic Types

A. Definition of Semantic Type

In their study, Nakayama and Yamamoto proposed classifying predicates according to sense and discussed the associated problems. For example, a verb "優れる (to excel)" behaves like an adjective insofar as it signifies the property of a person or an object. A verb "走る (to run)", which usually behaves as a verb, forms a phrase "虫唾が走る", which means "to give someone the creeps" in English and behaves like an adjective. Nakayama and Yamamoto found that aggregation of verb senses was necessary to be preprocessed for the classification due to the diversity in sense of verbs. They then proposed that verbs have four semantic types: *action*, *change*, *emotion* and *modification*. Verbs do not conform uniquely to one type, but can have multiple types. The four types are defined as follows:

action: Expressions that signify objectively observable motion and in which the state does not change before and after the motion.

978-1-4799-5330-1/14/$31.00 ©2014 IEEE

Table I
Annotated semantic types and number of verbs.

action	change	emotion	modification	# of verbs
Y	N	N	N	6637
N	Y	N	N	1531
N	N	Y	N	1441
N	N	N	Y	358
Y	Y	N	N	469
Y	N	Y	N	1127
Y	N	N	Y	190
N	Y	Y	N	396
N	Y	N	Y	113
N	N	Y	Y	72
Y	Y	Y	N	143
Y	Y	N	Y	61
Y	N	Y	Y	35
N	Y	Y	Y	42
Y	Y	Y	Y	33
total				12648

Y: annotated, N:not-annotated

Table II
Features for SVM.

feature	# of features	value
target verb	2681	0 or 1
inflected form	19	0 or 1
semantic types	4	0 or 1
particles	118	0 or 1
semantic categories	21	real value

e.g.) 泳ぐ (to swim), 食べる (to eat)

change: Expressions that signify a state as a result of a motion and in which the state after the motion differs from the state before the motion.

e.g.) 乾く (to dry), 死ぬ (to die)

emotion: Expressions that signify the operation of a sense organ, such as eyes, ears, and skin, or mental actions. According to Nakayama and Yamamoto, verbs that express intentional operations, such as “見る (to look)” and “考える (to think)”, have the type *action* and those that express unintentional operations, such as “見える (to see)” and “ 感じる (to feel)”, have the type *emotion*.

modification: Expressions that signify properties, shapes, beings, or relations. According to Nakayama and Yamamoto, all of adjectives are of this type.

e.g.) 優れる (to excel), 異なる (to differ)

All *action* or *change* verbs behave like verbs and *modification* verbs behave like adjectives.

B. Semantic Type Dictionary

Nakayama and Yamamoto manually annotated the semantic types for all 12,648 verbs in the *IPADIC for Japanese*(1). They annotated all types that can be related to a verb without losing the verb's information. The breakdown of annotated types and number of verbs are shown in Table I. 2,681 verbs (*i.e.*, approximately 20% of all verbs) are annotated as multiple types.

III. Construction of the Semantic Type Disambiguator

Some verbs are annotated as multiple types in the dictionary due to ambiguity of verb sense. We assumed that such verb types were determined from context information. For example, the verb “満たす “ has both the sense of “to fill” and “to satisfy” in English. If the verb appears in the sentence like “コップに水を満たす (to fill a glass with water)“ the type is *action* and “条件を満たす (to satisfy a condition)“ *modification*. In this study, we aim to disambiguate the types of verb that have ambiguity and multiple type using context information.

In their study, Okada and Yamamoto [3] investigated the relationship between semantic types and case frame information. They found that using surface frame information is insufficient on its own, and combining this information with semantic categorization of nouns in the statement is an effective way to determine semantic type. Therefore, we achieve disambiguation as a multi-class classification problem using a *support vector machine* (SVM) by building feature vectors. In the following it explains what feature has been used for the feature vector.

A. Feature Vector

First, an input sentence, which includes a target verb, was analyzed by the morphological analyzer *MeCab*(2). Morphological features were then generated from the results of the analysis. These features are shown in Table II. Description for each feature is as follows:

1) target verb: We assumed that the verb type was dependent on the verb itself. Hence, we used the target verb as a feature. The target verb takes value 1 and the other verbs take 0 for all 2,681 verbs which are annotated as multiple types in the dictionary.

2) inflected form of target verb: The inflected form of the verb relates to its role in the sentence. Hence the inflected form relates to verb type. Therefore, we used the inflected form of the target verb as a feature. The inflected form takes value 1 and the other forms take 0 for all 19 inflected forms in IPADIC.

3) semantic types: We used the types annotated to the verb in the dictionary as features. The types annotated take value 1 and the other types take 0.

4) particles: The case is indicated by particular particles in Japanese. We used the particles that decide the case, specifically case particle and binding particle, as features. The particles that appear in the sentence take value 1 and other particles take 0 for 118 of the particles in IPADIC.

5) semantic categories of noun: We used occurrence frequency of semantic categories of noun that appear in the sentence as features. For the categories of noun, we used semantic categories of noun and proper noun in *Goitaikei — A Japanese Lexicon*(3). GoiTaikei is a Japanese thesaurus that consists of 300,000 words and 3,000 semantic categories. The categories are used with generalization in order to avoid sparseness problem. Specifically, we limited the depth of the thesaurus of the categories from the root category. We set the depth limitation to 3 for the results of the preliminary experiment. The number of categories were reduced to 21 by the generalization. There

Table III
FREQUENCIES OF SEMANTIC TYPE IN THE DATA SET.

action	change	emotion	modification	total
451	614	255	680	2000

are certain nouns that have multiple categories. Therefore, we formulated occurrence frequency $freq(c)$ for semantic categories c as follows:

$$freq(c) = \sum_{n \in N_c} \frac{freq(n)}{|C_n|}, \qquad (1)$$

where N_c is set of nouns that have the category c, $freq(n)$ is occurrence frequency of noun n in the sentence, and C_n is set of categories that the noun n has.

Conventionally, syntactic structure information is utilized in a task of *Word Sense Disambiguation* (WSD) or *Semantic Role Labeling* (e.g. Johansson and Nugues [5]), which the semantic type disambiguation is similar to. However, we use only morphological features because dependency parsers may not be robust for casual text such as blog text.

We apply the Libsvm(4) for classification. We used a linear kernel for the kernel function due to sparseness of features and adjusted the parameters to maximize accuracy.

IV. EVALUATION

We evaluated our disambiguator using newspaper articles and compared with two baselines. In the following it explains about the setup of evaluation and the results.

A. Setup

We used newspaper articles(5) published in 2004 as the data set. First, sentences that included a verb that have multiple types were extracted from the data set. These verbs were then used as the classification target. Next, 2,000 of these sentences were randomly sampled as the data set. One of the authors manually annotated the most appropriate type to each verb. The frequency of each type, as a result of annotation, is shown in Table III.

We evaluated the accuracy of the disambiguator for the data set using 10-fold cross-validation. In order to achieve this, we set following two baselines:

1) random: One of the types that the verb has was selected at random.

2) frequency: The most frequently type in the types that the verb has was selected. The priority follows the order of *modification* $>$ *change* $>$ *action* $>$ *emotion* for Table III.

In this study, the disambiguator is not compared with other systems. It is because there is no other studies about disambiguation for the semantic types.

B. Results and Discussion

The evaluation result is shown in Table IV. We achieved 69.9% accuracy and marked better than baseline results. The breakdown of the disambiguation results is shown in Table V. According to the table, both of recall and precision for *modification* mark the worst in the four types. We considered this is because the number of training data for *modification* is the least; therefore, increasing data set for *modification* or weighting based on number of the data set for each type during training can improve the disambiguation accuracy.

Table IV
EVALUATIVE RESULT.

method	accuracy [%]
SVM	69.9
baseline1 (random)	24.1
baseline2 (frequency)	30.1

Table V
BREAKDOWN OF THE DISAMBIGUATION RESULTS.

	predicted					recall
actual	act	cha	emo	mod	total	[%]
action	276	64	61	50	451	61.2
change	31	435	10	138	614	70.8
emotion	63	28	147	17	255	57.6
modification	23	99	18	540	680	79.4
total	393	626	236	745	2000	-
precision [%]	70.2	69.5	62.3	72.5	-	69.9

The achieved accuracy seems low, however, we think it will not be serious problem in applications. As we can see for Table I, approximately 80% of all verbs have single semantic type. We roughly estimate more than 90% of verbs in sentences can be labeled with a correct semantic type by the disambiguator.

Furthermore, we examined the change of accuracy by size of data set in order to evaluate whether the size was sufficient. The results are shown in Figure 1. According to the figure, because the accuracy is on an upward trend in the range, we expected higher accuracy with a larger data set.

V. RELATED WORK

We can assume that our study is a WSD task in which the word sense is limited to four types. WSD for verbs is a fundamental topic of NLP. In English, the classification based on two verb categorization schemes, *VerbNet* [6] and *FrameNet* [7], has been addressed. Croce et al. [8] proposed a verb classification model according to VerbNet and FrameNet. They used the similarity of two verbs based on syntactic dependencies and kernel function.

In Japanese, Takeuchi et al. [9] constructed a system that estimates verb sense in a sentence. This system conducts argument structure analysis using manually constructed language resource. In this analysis, the sense of a predicate is estimated and its semantic role is added to the argument, in contrast with our study, which only utilized shallow information.

To classify verbs according to sense contributes to Japanese language education. According to Taniguchi [10], understanding for types of verb is important to correct use of verbs. Classification by aspect proposed by Kindaichi [11] is one of the sense types. The second verb (continuing verb) in the classification relates to the type *action*, the third verb (instant verb) relates to the type

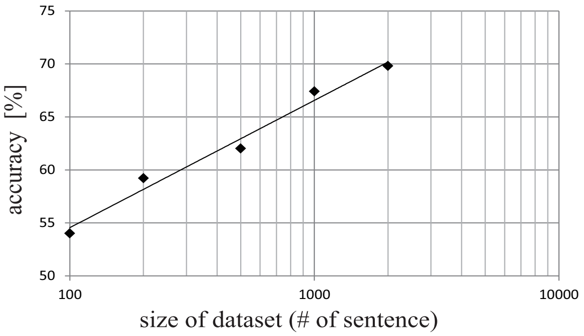

Figure 1. Relationship between size of data set and accuracy.

change and the first (state verb) and fourth (verb always entails “テイル”) relate to the type *modification*. Hence, the semantic type adds information for Japanese language learners.

VI. Conclusion

We constructed a disambiguator for semantic types, as proposed by Nakayama and Yamamoto. This disambiguator was achieved using an SVM that utilizes information from existing resources and the results of the morphological analyzer. We then achieved a 69.9% accuracy rate for newspaper articles using 10-fold cross-validation.

In future work, we will evaluate the effectiveness of these semantic types using the semantic type dictionary and the disambiguator by applying them to a particular task such as opinion mining.

The disambiguator will be available to the public.

Tool and resources

(1) Information-technology Promotion Agency DICtionary (IPADIC) for Japanese. Version 2.7.0. http://mecab.sourceforge.net/src/

(2) Morphological analyzer MeCab. Version 0.993. http://mecab.sourceforge.net/

(3) Goitaikei — A Japanese Lexicon. http://www.kecl.ntt.co.jp/icl/lirg/resources/GoiTakikei/index-en.html

(4) Libsvm. Version 3.18. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

(5) Nihon Keizai Shinbun (literally, Japanese Economics Newspaper) all article database CD-ROM, 2004.

References

[1] B. Yang and C. Cardie, “Extracting Opinion Expressions with semi-Markov Conditional Random Fields”, In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pp. 1335-1345, 2012.

[2] T. Nakayama and K. Yamamoto, “New Semantic Types for Predicates”, In *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, pp. 560-563, 2011. (in Japanese).

[3] S. Okada and K. Yamamoto, “Investigation of Relation with Case Frame Information for Semantic Type Disambiguation” (original Japanese), In *Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing*, pp. 524-527, 2014. (in Japanese).

[4] T. Scholz and S. Conrad, “Opinion Mining in Newspaper Articles by Entropy-based Word Connections”, In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp. 1828-1839, 2013.

[5] R. Johansson and P. Nugues, “Dependency-based Semantic Role Labeling of PropBank”, In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 69-78, 2008.

[6] K. K. Schuler, “VerbNet: A broad-coverage, comprehensive verb lexicon”, *Dissertations available from ProQuest*. Paper AAI3179808. University of Pennsylvania, 2005.

[7] C. F. Baker, C. J. Fillmore, and J. B. Lowe, “The Berkeley FrameNet Project”, In *Proceedings of the 1998 Joint Conference on 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING)*, pp. 86-90, 1998.

[8] D. Croce, R. Basili, A. Moschitti, and M. Palmer , “Verb Classification using Distributional Similarity in Syntactic and Semantic Structures”, In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 263-272, 2012.

[9] K. Takeuchi, S. Tsuchiyama, M. Moriya, and Y. Moriyasu, “Construction of Argument Structure Analyzer Toward Searching Same Situations and Actions”, *the Institute of Electronics, Information and Communication Engineers (IEICE) Technical Report. Natural Language Understanding and Models of Communication (NLC) 109(390)*. pp. 1-6. 2010. (in Japanese).

[10] S. Taniguchi, “Verb Classification for Japanese Language Education”, *Oita University International Student Center bulletin No. 2*, pp. 53-63, 2005. (in Japanese).

[11] H. Kindaichi, “Classification for Japanese Verbs”, *Aspect of Japanese Verb*, pp. 5-26, 1976. (in Japanese).