Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD incorporating Idiom-Specific Features

Chikara Hashimoto
Graduate School of Science and Engineering
Yamagata University
Yonezawa, Yamagata, 992-8510, JAPAN
ch@yz.yamagata-u.ac.jp

Daisuke Kawahara
National Institute of Information and
Communications Technology
Sorakugun, Kyoto, 619-0289, JAPAN
dk@nict.go.jp

Abstract

Some phrases can be interpreted either idiomatically (figuratively) or literally in context, and the precise identification of idioms is indispensable for full-fledged natural language processing (NLP). To this end, we have constructed an idiom corpus for Japanese. This paper reports on the corpus and the results of an idiom identification experiment using the corpus. The corpus targets 146 ambiguous idioms, and consists of 102,846 sentences, each of which is annotated with a literal/idiom label. For idiom identification, we targeted 90 out of the 146 idioms and adopted a word sense disambiguation (WSD) method using both common WSD features and idiom-specific features. The corpus and the experiment are the largest of their kind, as far as we know. As a result, we found that a standard supervised WSD method works well for the idiom identification and achieved an accuracy of 89.25% and 88.86% with/without idiom-specific features and that the most effective idiom-specific feature is the one involving the adjacency of idiom constituents.

1 Introduction

Some phrases like *kick the bucket* are ambiguous with regard to whether they carry literal or idiomatic meaning in a certain context. This ambiguity needs to be resolved in the same manner as ambiguous words that have been dealt with in the WSD literature. We term the resolution of the literal/idiomatic ambiguity as idiom identification, hereafter.

Idiom identification is classified into two kinds; one is for idiom types and the other is for idiom tokens. With the former, phrases that *can* be interpreted as idioms are found in text corpora, typically for compiling idiom dictionaries. On the other hand, the latter helps identify a phrase in context as a true idiom or a phrase that should be interpreted literally (a literal phrase, henceforth). In this paper, we deal with the latter, i.e., idiom token identification.

Despite the recent enthusiasm for multiword expressions (MWEs) (Grégoire et al., 2007; Grégoire et al., 2008), the idiom token identification is in an early phase of its development. Given that many NLP tasks like machine translation or parsing have been developed as a result of the availability of language resources, idiom token identification should also be developed when adequate idiom resources are provided. To this end, we have constructed a Japanese idiom corpus. We have also conducted an idiom identification experiment using the corpus that we hope will be a good reference point for future studies on the task. We drew on a standard WSD framework with machine learning exploiting both features commonly used in the WSD studies and idiom-specific features. This paper reports in detail the corpus and the result of the experiment; herein, it must be noted that to the best of our knowledge, the corpus and the experiment are the largest ever of their kind.

We only deal with the ambiguity between literal and idiomatic interpretations. However, some phrases have two or more idiomatic meanings without context. For example, a Japanese idiom *te-o dasu* (hand-ACC stretch)[1] can be interpreted as ei-

[1] ACC is the accusative case marker. Likewise we use the following notation in this paper; NOM for the nominative case

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 992–1001,
Honolulu, October 2008. ©2008 Association for Computational Linguistics

ther "punch," "steal" or "make moves on." This kind of ambiguity should be placed on the agenda.

We do not tackle the problem of what constitutes the notion of "idiom." We simply regard phrases listed in Sato (2007) as idioms.

The reminder of this paper is organized as follows. In §2 we present related works. §3 shows the target idioms. After the idiom corpus is described in §4, we detail our idiom identification method and experiment in §5. Finally §6 concludes the paper.

2 Related Work

There have only been a few works on the construction of an idiom corpus. In this regard, Birke and Sarkar (2006) and Cook et al. (2008) are notable exceptions. Birke and Sarkar (2006) automatically constructed a corpus of English idiomatic expressions (words that can be used non-literally). They targeted 50 expressions and collected about 6,600 examples. They call the corpus TroFi Example Base, which is available on the Web.[2] Cook et al. (2008) compiled a corpus of English verb-noun combinations (VNCs) tokens. Their corpus deals with 53 VNC expressions and consists of about 3,000 example sentences. Like ours, they assigned each example with a label indicating whether an expression in the example is used literally or idiomatically. Our corpus can be regarded as the Japanese idiom counterpart of these works. However, note that our corpus targets 146 idioms and consists of as many as 102,846 example sentences. Another exception is Tsuchiya et al. (2006), who manually constructed an example database of Japanese compound functional expressions named MUST. They provide it on the Web.[3] Some of the compound functional expressions in Japanese are ambiguous like idioms are.[4]

The SAID dataset[5] provides data about the syntactic flexibility of English idioms. It does not concern itself with idiom token identification. However, as in Hashimoto et al. (2006b), Hashimoto et al. (2006a) and Cook et al. (2007) among others, the syntactic behavior of idioms is an important clue to idiom token identification.

Previous studies have mostly focused on the idiom type identification (Lin, 1999; Krenn and Evert, 2001; Baldwin et al., 2003; Shudo et al., 2004; Fazly and Stevenson, 2006). However, there has been a growing interest in idiom token identification in recent times (Katz and Giesbrecht, 2006; Hashimoto et al., 2006b; Hashimoto et al., 2006a; Birke and Sarkar, 2006; Cook et al., 2007). Katz and Giesbrecht (2006) compared the word vector of an idiom in context and that of the constituent words of the idiom using LSA in order to determine if the expression is idiomatic. Hashimoto et al. (2006b) and Hashimoto et al. (2006a) (HSU henceforth) focused their attention on the differences in grammatical constraints imposed on idioms and their literal counterparts such as the possibility of passivization, and developed handcrafted rules for Japanese idiom identification. Although their task is exactly the same as ours and we draw on the grammatical knowledge provided by them, the scale of their experiment is very small, since only 108 sentences were used for idiom identification in their paper. Further, unlike HSU, we employ matured WSD technologies. Cook et al. (2007) (CFS henceforth) propose an unsupervised method for English on the basis of the observation that idioms tend to be expressed in a small number of fixed forms.

These studies used only the characteristics of idioms (or MWEs). On the other hand, we exploit a WSD method, for which there have been many studies and matured technologies, in addition to the characteristics of idioms. Birke and Sarkar (2006) also used WSD. However, they employed an unsupervised method, while ours is a completely supervised one.

Apart from idioms, Uchiyama et al. (2005) conducted the token classification of Japanese compound verbs exploiting supervised method.

marker, DAT for the dative case marker, and GEN for the genitive case marker. FROM and TO stand for the Japanese counterparts of *from* and *to*. NEG represents a verbal negation morpheme.

[2] http://www.cs.sfu.ca/~anoop/students/jbirke/

[3] http://nlp.iit.tsukuba.ac.jp/must/

[4] For example, *(something)-ni-atatte* ((something)-DAT-run.into) means either "run into (something)" or "on the occasion of (something)." The former is the literal interpretation and the latter is the idiomatic interpretation of the compound functional expression.

http://www.ldc.upenn.edu/Catalog/

[5] CatalogEntry.jsp?catalogId=LDC2003T10

3 Target Idioms

For this study, we selected 146 idioms through the following procedure. ① We extracted basic idioms from Sato (2007). Sato compiled about 3,600 basic idioms of Japanese from five books: two dictionaries for elementary school, two idiom dictionaries, and one linguistics book on idioms. We extracted those idioms that were described in more than two of these five books. The total number of such idioms added up to 926. ② From among these idioms, we chose ambiguous ones.[6] As a result, 146 idioms were selected.

As for ②, sometimes it is not trivial to determine if an idiom is ambiguous or not. Some idioms are rarely interpreted literally, while others, in all likelihood, take on the literal meaning. Is it meaningful to regard them as ambiguous and deal with them in this study? If not, how does one assuredly distinguish truly ambiguous idioms from those that are mostly interpreted either literally or figuratively? This can only be done if there is an accurate idiom identification system.

After all, we asked two native speakers of Japanese (Group A) to classify idioms into two classes: 1) truly ambiguous ones and 2) completely unambiguous or practically unambiguous ones. On the basis of the classification, one of the authors made final judgments.

To verify how stable this ambiguity endorsement was, we asked another two other native speakers of Japanese (Group B) to perform the same task and calculated the Kappa statistic between the two speakers. First, we sampled 101 idioms from the 926 chosen earlier. Then, the two members of Group B classified the sampled idioms into the two classes. The Kappa statistic was found to be 0.6576, which indicates middling stability.

Tables 2 and 3 list some of the target idioms.

4 Idiom Corpus

4.1 Corpus Specification

The corpus is designed for the idiom token identification task. That is, each example sentence in the corpus is annotated with a label that indicates whether the corresponding phrase in the example is used as an idiom or a literal phrase. We call the former the positive example and the latter the negative example. More specifically, the corpus consists of lines that each represent one example. A line consists of four fields as follows: 1 **Label** indicates whether the example is positive or negative. Label *i* is used for positive examples and *l* for negative ones. 2 **ID** denotes the idiom that is included in the example. In this study, each idiom has a unique number, which is based on Sato (2007). 3 **Lemma** also shows the idiom in the example. We assigned each idiom its canonical (or standard) form on the basis of Sato (2007). 4 **Example** is the example itself.

Given below is a sample of a negative example of *goma-o suru* (sesame-ACC crush) 'flatter'.

- *l*␣1417␣ごまをする␣すり鉢でごまをすり ···

The third field is the lemma of the idiom. The last one is the example that says 'crushing sesame in a mortar...'

Before working on the corpus construction, we prepared a reference by which human annotators could consistently distinguish between the literal and figurative meanings of idioms. To be more precise, this reference specified literal and idiomatic meanings for each idiom like dictionaries do. For example, the entry for *goma-o suru* in the reference is as follows.

Idiom: To flatter people.
Literal: To crush sesame.

As for the corpus size, we continued to annotate examples for each idiom, regardless of the proportion of idioms and literal phrases, until the total number of examples for each idiom reached 1,000.[7] In the case of a shortage of original data, we annotated as many examples as possible. The original data were sourced from the Japanese Web corpus (Kawahara and Kurohashi, 2006).

4.2 Corpus Construction

We constructed the corpus in the following manner: ❶ From the Web corpus, we collected example sentences that contained one of our target idioms whichever meaning (positive or negative) they

[6] Some idioms like *by and large* do not have a literal meaning. They are not dealt with in this paper.

[7] For idioms that we sampled for preliminary annotation, we annotated more than 1,000 examples.

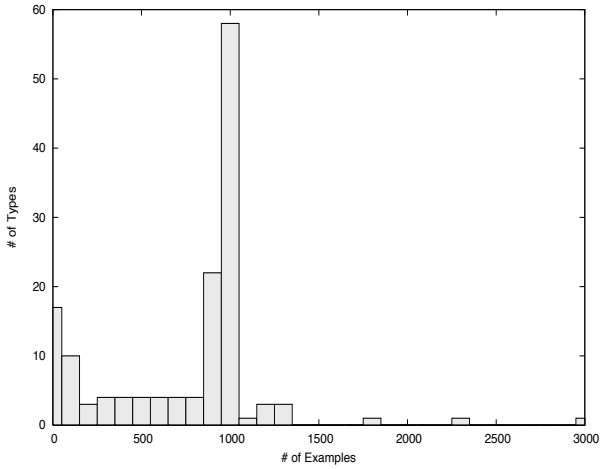


Figure 1: Distribution of the number of examples

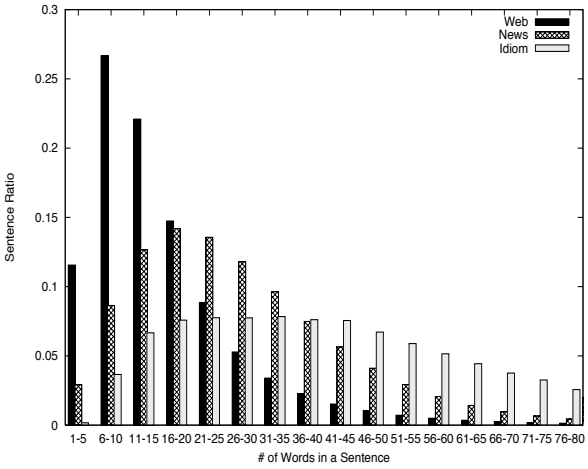


Figure 2: Distribution of sentence length

take on. Concretely speaking, we automatically collected sentences in which constituent words of one of our targets appeared in a canonical dependency relationship by using KNP[8], a Japanese dependency parser. ❷ We classified the collected examples as positive and negative. This was done by human annotators and was based on the reference to distinguish the two meanings. For annotation, longer examples were given higher priority than shorter examples. Note that we discarded examples that were collected by mistake due to dependency parsing errors and those that lacked a context that could help them be interpreted correctly.

This was done by the two members of Group A and took 230 hours.

4.3 Status of Corpus

The corpus consists of 102,846 examples.[9] Figure 1 shows the distribution of the number of examples. For 68 idioms, we annotated more than 1,000 examples. However, we annotated less than 100 examples for 17 idioms because of inadequate original data.

The average number of words in a sentence is 46. `Idiom` in Figure 2 shows the distribution of sentence length (the number of words) in the corpus. `Web` and `News` indicate the sentence length in the Web and a newspaper corpora, respectively. This is drawn from Kawahara and Kurohashi (2006). As you see, our corpus contains many more long sentences. This is because longer sentences were given priority for annotation, as stated in §4.2. Figure 3 shows the longest and shortest examples each for literal and idiomatic meanings of *goma-o suru* drawn from the corpus.

To determine how consistent the positive/negative annotation is across different human annotators, we sampled 1,421 examples from the corpus, asked the two members of Group B to do the same annotation, and calculated the Kappa statistic between the two. The value was 0.8519, which indicates very high agreement.

The corpus is available on the Web.[10] Currently we provide the list of the basic Japanese idioms we are dealing with, the idiom corpus, and the vector representation data used for the idiom identification experiment. The corpus is protected under the BSD license.

5 Idiom Identification Experiment

5.1 Method of Idiom Identification

We adopted a standard WSD method using machine learning. More specifically, we used SVM (Vapnik, 1995) with a quadratic kernel implemented in TinySVM.[11] The features we used are classified into either those that have been commonly used in WSD on the lines of Lee and Ng (2002) (LN hereafter),

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

[9] Note that the figures reported here are for the corpus of the 2008-06-25 version and will be slightly changed over time.

[10] `http://openmwe.sourceforge.jp/`

[11] `http://www.chasen.org/~taku/software/TinySVM/`

- ただし、１５６２年にグレシャムは１５６２年に、同一の額面価値で流通する素材価値を異にする２種類の貨幣が存在すると劣悪な貨幣が流通に残り、優良な貨幣は駆逐されるという「グレシャムの法則」を発表していることから、女衒がしたたかであったように、ＩＴ興行師もメーカーに**ごまをすり**、政府の省庁にこびを売り、知性が無くても生命力だけで、目新しい言葉のつまみ食いで、悪毒、図々しく、虚勢で生き続けることだろう。
(But I suspect that the show managers of IT ventures will remain sly and audacious, and survive by **flattering** manufacturers, bending over themselves to accede to the demands of governmental agencies, and talking glibly about buzz terms, without intelligence but with vitality, just like the brokers of prostitutes in the Edo period were, because Gresham's law of 1562 says that any circulating currency consisting of both good and bad money quickly becomes dominated by the bad money.)

- 上に**ごまをする**小役人タイプ。
(Just like a pretty official **flattering** his boss.)

- 煮た大豆をつぶすには、ミンチみたいな器具があればいいのですが、ない　場合は、**ごまをする**もので潰すか、もっと簡単な方法としては、ビニール　袋に大豆を入れ、封をしてタオルをかけてその上から瓶でこするようにす　るといいでしょう。
(In order to mash boiled soybeans, it is the best to use a meat chopper, but if you don't have one, use the thing to **crush sesame**, or put them into a plastic bag, cover it with a towel, then mash it with a glass bottle, which is easier.)

- **ごまをすり**調味料とあえる
(**Crushing sesame**, then adding seasonings to it.)

Figure 3: The longest and shortest examples for both literal and idiomatic meanings of *goma-o suru*

or those that have been designed for Japanese idiom identification proposed by HSU.[12]

- Common WSD Features
  - **f1:** POS of three words on the left side of idiom and three words on the right side
  - **f2:** Local collocations
  - **f3:** Single words in the surrounding context
  - **f4a:** Lemma of the rightmost word among those words that are the dependents of the leftmost constituent word of idiom[13]
  - **f4b:** POS of the rightmost word among those words that are the dependents of the leftmost constituent word of idiom
  - **f5a:** Lemma of the word which the rightmost constituent word of idiom is the dependent of
  - **f5b:** POS of the word which the rightmost constituent word of idiom is the dependent of
  - **f6:** Hypernyms of words in the surrounding context
  - **f7:** Domains of words (Hashimoto and Kurohashi, 2007; Hashimoto and Kurohashi, 2008) in the surrounding context
- Idiom-Specific Features
  - **f8:** Adnominal modification flag
  - **f9:** Topic case marking flag
  - **f10:** Voice alternation flag
  - **f11:** Negation flag
  - **f12:** Volitional modality flag
  - **f13:** Adjacency flag

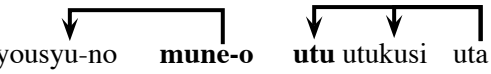
We used JUMAN,[14] a morphological analyzer of Japanese, and KNP to extract these features.

[12] Remember that HSU implemented them in handcrafted rules. We adapted them to a machine learning framework.

[13] Note that Japanese is a head final language.

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

f2 and **f3** are the same as those described in LN. But **f1** is slightly different in that we did not use the P_0 of LN. **f4** and **f5** roughly correspond to the syntactic relations of LN. We adapted it to Japanese idioms along with some simplifications. In the case of the example of *mune-o utu* (chest-ACC hit) 'impress' below,[15] **f4** is the POS and lemma of *tyousyu* and **f5** corresponds to those of *uta*.[16]



- tyousyu-no **mune-o** **utu** utukusi uta
  audience-GEN **chest-ACC** **hit** beautiful song
  'A beautiful song that **impresses** the audience'

f6 and **f7** are available from JUMAN's output. For example, the hypernym of *tyousyu* (audience) is `human` and its domain is `culture/media`. Those of *uta* (song) are `abstract-thing` and `culture/recreation`. They are not used in LN, but they are known to be useful for WSD (Tanaka et al., 2007; Magnini et al., 2002).

f8 indicates whether a nominal constituent of an idiom, if any, undergoes adnominal modification. **f9** indicates whether one of Japanese topic case markers is attached to a nominal constituent of an idiom, if any. **f10** is turned on when a passive or causative suffix is attached to a verbal constituent of an idiom, if any.[17] **f11** and **f12** are similar to **f10**. The former is used for negated forms and the latter for volitional modality suffixes of a predicate part of an idiom, if any.[18] Volitional modality includes expressions like order, request, permission, prohibition, and volition. Finally, **f13** indicates whether the constituents of an idiom is adjacent to each other.

As discussed in HSU, the idiom-specific features are effective to distinguish idioms from literal phrases. For example, the idiom *goma-o suru* does not allow adnominal modification, while its literal counterpart does. Similarly, the idiom *mune-o utu* cannot take volitional modality unlike its literal counterpart.

[15]The arrows indicate dependency relations.

[16]Functional words attaching to either the **f4** word or the **f5** word are ignored. In the example, *no* (GEN) is ignored.

[17]Passivization is indicated by the suffix *(r)are* in Japanese. But the same suffix is also used for honorification, potentials and spontaneous potentials. Since it is beyond the current technology, we gave up distinguishing them.

[18]Note that **f10**, **f11** and **f12** are applied to only those idioms that can be used as predicates.

5.2 Experimental Condition

In the experiment, we dealt with 90 idioms for which more than 50 examples for both idiomatic and literal usages were available.[19] We conducted experiments for each idiom.

The performance measure is the accuracy.

$$\text{Accuracy} = \frac{\text{\# of examples correctly identified}}{\text{\# of all example}}$$

The baseline system uniformly regards all examples as either positive or negative depending on which is more dominant in the idiom corpus. Naturally, this is prepared for each idiom.

$$\text{Baseline} = \frac{\max(\text{\# of positive}, \text{\# of negative})}{\text{\# of all example}}$$

The accuracy and the baseline accuracy for each idiom are calculated in a 10-fold cross validation style; we split examples of an idiom into 10 pieces in advance of the experiment.

Also, we calculated the overall accuracy and baseline accuracy from the individual results. We summed up all accuracy scores of all the 90 idioms and then divided it by 90, which is called the macro-average. We did this for the baseline accuracy, too.

Another performance measure is the relative error reduction (RER).[20]

$$\text{RER} = \frac{\text{ER of baseline} - \text{ER of system}}{\text{ER of baseline}}$$

The overall RER is calculated from the overall accuracy and baseline by the above formula.

5.3 Experimental Result

Table 1 shows the overall performance. The first column is the baseline accuracy (%). The second column is the accuracy (%) and relative error reduction (%) of the system without the idiom-specific features. The third column is those of the system with the idiom features. Tables 2 and 3 show the individual results of the 90 idioms. The first column shows

[19]Some examples were unavailable due to the feature extraction failure. Thus, examples used for the experiment are fewer in number than those included in the corpus.

[20]ER stands for Error Rate in the formula.

Table 2: Individual Results (1/2)

Type	Base	(Pos ; Neg)	w/o I (RER)	w/ I (RER)
青筋を立てる (blue.vein-ACC emerge) 'burst a blood vessel'	83.38	(286 ; 57)	86.32 (17.68)	**86.61 (19.45)**
あぐらをかく (sit cross-legged) 'rest on one's laurels'	62.45	(587 ; 353)	92.66 (80.45)	**92.87 (81.02)**
足が付く (leg-NOM attach) 'find a clue to solving a case'	72.21	(184 ; 478)	77.20 (17.96)	**79.62 (26.68)**
足が出る (leg-NOM go.out) 'run over the budget'	77.59	(188 ; 651)	92.61 (67.01)	**93.08 (69.13)**
足元を見る (one's feet-ACC look.down) 'see someone coming'	57.53	(420 ; 310)	**85.89 (66.77)**	85.75 (66.45)
足を洗う (leg-ACC wash) 'wash one's hands of ...'	68.47	(632 ; 291)	92.65 (76.68)	92.65 (76.69)
足を伸ばす (leg-ACC stretch) 'go a little further'	80.24	(727 ; 179)	95.26 (76.03)	**95.38 (76.59)**
頭が痛い (head-NOM ache) 'harass oneself about ...'	57.87	(158 ; 217)	83.94 (61.89)	83.94 (61.89)
頭を抱える (head-ACC fold) 'tear one's hair out'	87.28	(796 ; 116)	91.35 (31.99)	91.35 (31.99)
頭をもたげる (head-ACC lift) 'rear its head'	83.14	(804 ; 163)	93.40 (60.83)	**93.50 (61.45)**
脂が乗る (fat-NOM put.on) 'warm up to one's work'	83.69	(196 ; 1006)	92.94 (56.69)	92.94 (56.69)
油を売る (oil-ACC sell) 'shoot the breeze'	86.67	(507 ; 78)	92.63 (44.70)	92.63 (44.70)
油を絞る (oil-ACC squeeze) 'rake someone over the coals'	66.83	(69 ; 139)	84.64 (53.71)	**86.14 (58.23)**
網を張る (net-ACC spread) 'wait expectantly'	70.10	(366 ; 858)	**81.28 (37.41)**	80.96 (36.31)
息が詰まる (breath-NOM choke.up) 'stifling'	71.61	(681 ; 270)	**79.82 (28.91)**	79.50 (27.80)
一から十まで (one-FROM ten-TO) 'all without exception'	92.00	(770 ; 67)	93.48 (18.51)	93.48 (18.51)
色を失う (color-ACC lose) 'turn pale'	73.32	(262 ; 720)	84.23 (40.91)	84.23 (40.91)
腕が上がる (arm-NOM go.up) 'develop one's skill'	57.06	(481 ; 362)	84.47 (63.85)	**88.75 (73.80)**
尾を引く (tail-ACC pull) 'have a lasting effect'	87.72	(843 ; 118)	93.14 (44.15)	**93.35 (45.84)**
顔を出す (face-ACC present) 'show up'	84.48	(697 ; 128)	88.60 (26.49)	**88.82 (27.93)**
肩を並べる (shoulder-ACC juxtapose) 'on a par'	89.38	(842 ; 100)	**93.20 (35.97)**	93.10 (34.97)
角が取れる (corner-NOM remove) 'become mature'	57.45	(370 ; 274)	**78.35 (49.13)**	78.04 (48.39)
唇をかむ (lip-ACC bite) 'bite one's lip'	70.89	(587 ; 241)	78.40 (25.78)	**79.36 (29.10)**
口を切る (mouth-ACC cut) 'break the ice'	51.50	(210 ; 223)	**84.83 (68.73)**	83.69 (66.36)
口をとがらせる (mouth-ACC sharpen) 'pout'	86.33	(663 ; 105)	**87.61 (9.40)**	87.35 (7.47)
首が回らない (neck-NOM turn-NEG) 'up to one's neck'	66.63	(619 ; 310)	**86.41 (59.28)**	86.22 (58.71)
首を切る (neck-ACC cut) 'give the axe'	53.90	(449 ; 384)	**89.93 (78.15)**	89.80 (77.88)
首をひねる (neck-ACC twist) 'think hard'	93.16	(885 ; 65)	**94.11 (13.85)**	93.79 (9.23)
事によると (thing-DAT depend) 'perhaps'	67.15	(231 ; 113)	96.50 (89.35)	**97.35 (91.94)**
ごまをする (sesame-ACC crush) 'flatter'	50.29	(87 ; 88)	**92.75 (85.42)**	90.99 (81.88)
背を向ける (back-ACC train) 'turn one's back'	66.70	(597 ; 298)	89.06 (67.14)	89.06 (67.14)
血が通う (blood-NOM flow) 'humane'	50.18	(422 ; 419)	82.41 (64.70)	**83.24 (66.37)**
宙に浮く (midair-DAT float) ''	58.07	(382 ; 529)	88.03 (71.46)	**88.69 (73.03)**
土が付く (dirt-NOM attach) 'be defeated in sumo wrestling'	72.66	(70 ; 186)	**79.48 (24.97)**	78.76 (22.33)
手が届く (hand-NOM reach) 'afford' 'reach an age' 'attentive'	80.76	(470 ; 112)	87.66 (35.85)	87.66 (35.85)
手がない (hand-NOM there.isn't) 'have no remedy'	86.94	(799 ; 120)	92.61 (43.38)	**92.83 (45.06)**
手が離れる (hand-NOM get.away) 'get one's work done'	53.49	(360 ; 414)	**92.37 (83.59)**	92.36 (83.57)
手に乗る (hand-DAT ride) 'fall into someone's trap'	61.05	(372 ; 583)	92.86 (81.68)	**93.49 (83.30)**
手を入れる (hand-DAT insert) 'obtain'	53.21	(373 ; 328)	93.44 (85.99)	**93.59 (86.29)**
手を掛ける (hand-ACC hang) 'give a lot of care'	70.57	(241 ; 578)	91.19 (70.04)	**91.31 (70.46)**
手を切る (hand-ACC cut) 'break away'	57.85	(468 ; 341)	91.08 (78.83)	91.08 (78.83)
手を取る (hand-ACC take) 'give every possible help (to learn)'	88.89	(91 ; 728)	**92.74 (34.67)**	92.62 (33.56)
手を握る (hand-ACC grasp) 'conclude an alliance'	90.51	(73 ; 696)	**95.44 (51.93)**	95.17 (49.16)
手を延ばす (hand-ACC stretch) 'extend one's business'	89.55	(95 ; 814)	94.01 (42.69)	**94.22 (44.72)**
手を広げる (hand-ACC open.up) 'extend one's business'	70.52	(579 ; 242)	89.17 (63.26)	**90.15 (66.57)**
手を回す (hand-ACC turn) 'take measures'	68.86	(246 ; 544)	93.04 (77.64)	**93.92 (80.49)**
峠を越す (mountain.pass-ACC go.over) 'get over the hump'	72.18	(685 ; 264)	89.28 (61.46)	**89.49 (62.23)**
泥を塗る (mud-ACC daub) 'drag someone through mud'	74.38	(543 ; 187)	91.64 (67.38)	**91.92 (68.45)**
波に乗る (wave-DAT ride) 'catch a wave'	86.23	(783 ; 125)	**93.05 (49.55)**	92.94 (48.74)
熱が冷める (heat-NOM get.cool) 'fever goes down'	89.90	(890 ; 100)	92.02 (21.00)	**92.22 (23.00)**
熱を上げる (heat-ACC raise) 'go ape'	92.52	(903 ; 73)	94.50 (26.45)	**94.71 (29.21)**
熱を入れる (heat-ACC feed.in) 'enthuse'	85.06	(723 ; 127)	90.71 (37.80)	**91.76 (44.88)**
根を下ろす (root-ACC take.down) 'take root'	85.83	(824 ; 136)	93.23 (52.21)	93.23 (52.21)
根を張る (root-ACC spread) 'take root'	60.00	(564 ; 376)	87.66 (69.15)	87.66 (69.15)
バスに乗り遅れる (bus-DAT miss) 'miss the boat'	76.97	(199 ; 665)	90.50 (58.74)	**92.36 (66.81)**
バトンを渡す (baton-ACC give) 'have someone succeed his position'	65.33	(471 ; 250)	81.70 (47.23)	**82.25 (48.81)**
鼻息が荒い (nasal.breathing-NOM heavy) 'full of big talk'	52.77	(286 ; 256)	75.33 (47.77)	**76.62 (50.50)**
鼻が高い (nose-NOM high) 'proud'	50.27	(659 ; 652)	81.01 (61.81)	**82.30 (64.42)**
鼻を折る (nose-ACC break) 'humble (someone)'	56.60	(69 ; 90)	69.58 (29.91)	**74.92 (42.20)**
鼻を鳴らす (nose-ACC make.a.sound) 'make light of ...'	55.72	(536 ; 426)	80.79 (56.63)	**81.21 (57.57)**
腹を割る (belly-ACC cut) 'have a heart-to-heart talk'	95.62	(1265 ; 58)	96.68 (24.16)	96.68 (24.16)
歯を食い縛る (teeth-ACC clench) 'grit one's teeth'	65.54	(194 ; 102)	**71.97 (18.66)**	71.63 (17.66)
人を食う (human-ACC eat) 'look down on someone'	74.95	(727 ; 243)	87.01 (48.15)	87.01 (48.15)
火花を散らす (spark-ACC spread) 'fight heatedly'	75.99	(728 ; 230)	89.57 (56.56)	**89.68 (57.00)**

Table 3: Individual Results (2/2)

Type	Base	(Pos ; Neg)	w/o I (RER)	w/ I (RER)
筆を入れる (painting.brush-ACC add) 'correct (writings or paintings)'	75.80	(213 ; 68)	83.99 (33.84)	**84.70 (36.79)**
船をこぐ (ship-ACC row) 'nod'	50.76	(167 ; 162)	75.82 (50.88)	**76.37 (52.01)**
骨が折れる (bone-NOM break) 'have difficulty'	62.30	(575 ; 348)	94.14 (84.46)	94.14 (84.47)
骨を埋める (bone-ACC bury) 'make it one's final home'	82.82	(757 ; 157)	89.84 (40.85)	**90.60 (45.31)**
骨を折る (bone-ACC break) 'make efforts'	60.89	(350 ; 545)	92.74 (81.43)	**92.96 (82.01)**
幕が開く (curtain-NOM open) 'start'	55.64	(533 ; 425)	**86.32 (69.17)**	86.22 (68.94)
右から左 (right-FROM left) 'passing through without staying'	73.88	(794 ; 2246)	**89.90 (61.34)**	89.87 (61.21)
水と油 (water-AND oil) 'oil and water'	55.66	(1053 ; 839)	83.19 (62.10)	**85.84 (68.07)**
水に流す (water-DAT flush) 'forgive and forget'	67.08	(652 ; 320)	85.91 (57.19)	**89.40 (67.81)**
身に付ける (body-DAT put.on) 'learn'	90.29	(725 ; 78)	**96.51 (64.11)**	96.39 (62.82)
耳が痛い (ear-NOM ache) 'make one's ears burn'	59.49	(333 ; 489)	88.69 (72.08)	**89.54 (74.19)**
耳に入れる (ear-DAT insert) 'get word of ...'	74.89	(501 ; 168)	89.50 (58.20)	**90.38 (61.67)**
実を結ぶ (fruit-ACC bear) 'bear fruit'	89.39	(826 ; 98)	**95.79 (60.33)**	95.68 (59.31)
胸が痛む (chest-NOM ache) 'suffer heartache'	93.59	(876 ; 60)	95.82 (34.78)	**95.93 (36.46)**
胸が膨らむ (chest-NOM expand) 'feel one's heart leap'	55.58	(338 ; 423)	94.08 (86.68)	**94.48 (87.57)**
胸を打つ (chest-ACC hit) 'impress'	92.39	(801 ; 66)	96.45 (53.34)	**96.68 (56.39)**
芽が出る (germ-NOM come.out) 'close to making the top'	56.57	(377 ; 491)	91.33 (80.03)	**91.55 (80.55)**
目がない (eye-NOMthere.isn't) 'have a passion for ...'	91.81	(829 ; 74)	**95.70 (47.47)**	95.25 (42.05)
メスを入れる (scalpel-ACC insert) 'take drastic measures'	88.96	(741 ; 92)	96.28 (66.30)	96.28 (66.30)
目に入る (eye-DAT enter) 'catch sight of ...'	84.76	(623 ; 112)	90.22 (35.79)	**91.16 (41.97)**
目を覆う (eye-ACC cover) 'be in a shambles'	87.24	(725 ; 106)	91.45 (32.99)	**92.06 (37.72)**
目を覚ます (eye-ACC awake) 'snap out of ..'	83.26	(118 ; 587)	87.92 (27.85)	**88.64 (32.12)**
目をつぶる (eye-ACC close) 'turn a blind eye'	70.13	(533 ; 227)	90.26 (67.40)	90.26 (67.40)
目を細くする (eye-ACC thin) 'one's eyes light up'	53.44	(115 ; 132)	**75.20 (46.74)**	75.11 (46.54)
指をくわえる (finger-ACC suck) 'look enviously'	92.50	(876 ; 71)	**95.68 (42.41)**	95.58 (41.09)
弓を引く (bow-ACC draw) 'defy'	88.06	(138 ; 1018)	**95.51 (62.41)**	95.43 (61.68)

Table 1: Overall Result

Base	w/o I (RER)	w/ I (RER)
72.92	88.86 (58.87)	89.25 (60.30)

the target idioms. The second column shows baseline accuracy (%) and the numbers of positive and negative examples for each idiom. The accuracy (%) and relative error reduction (%) of the system without the idiom-specific features are described in the third column. The fourth column is those of the system with the idiom features. Bold face indicates a better performance.

All in all, we see relatively high baseline performances. Nevertheless, both systems outperformed the baseline. Especially, the system without the idiom-specific features has a noticeable lead over the baseline, showing that WSD technologies are effective in the idiom identification. Incorporating the idiom features into the system improved the overall performance, which is statistically significant (McNemar test, $p<0.01$). But performances of some idioms slightly degraded by the incorporation of the idiom features.

Table 4: Overall Results without Using One of the Idiom Features

Feature Type	Acc
All	89.25
−f8 (w/o Adnominal modification flag)	89.24
−f9 (w/o Topic case marking flag)	89.22
−f10 (w/o Voice alternation flag)	89.15
−f11 (w/o Negation flag)	89.17
−f12 (w/o Volitional modality flag)	89.19
−f13 (w/o Adjacency flag)	89.09

Table 4 shows overall results without using one of the idiom features.[21] As you see, the adjacency flag (f13) contributes to idiom identification accuracy the most.[22] On the other hand, the adnominal modification flag (f8) contributes to the task only slightly.[23]

[21]The first row shows the result with all idiom features used, just for ease of reference.

[22]Note that greater performance drop indicates greater contribution.

[23]This result is inconsistent with the result obtained in HSU, where they reported that grammatical constraints involving adnominal modification was most effective. This inconsistency might be attributed to the differences of datasets being used for idiom identification experiment. HSU used only 108 sentences

Table 5: Results reported in CFS

	Accu	RER
Baseline	61.9	—
Unsupervised	72.4	27.6
Supervised	76.2	37.5

Table 5 shows the results reported in CFS. Their baseline system regards all instances as idioms. The performance of the supervised one is obtained by the method of Katz and Giesbrecht (2006). Though we cannot simply compare this with our results due to the difference in experimental conditions, this implies that our WSD-based method was equally good or possibly better than their methods that are tailored to MWEs.

6 Conclusion

In this paper, we reported on the idiom corpus we have constructed and the idiom identification experiment using the corpus.

As mentioned in §4.3, some idioms are short of examples in the current idiom corpus. We plan to collect more examples by using different characters. In the Japanese language, there are basically three character systems: Hiragana, Katakana, and Chinese characters. Thus, you can write an idiom in different characters. For example, *mune-o utu* (chest-ACC hit) 'impress' can be either 胸を打つ or 胸をうつ.

In spite of its imperfection, we are sure that we can learn a lot about the idiom identification from the corpus, since, as far as we know, it is the largest-ever one, and so is the idiom identification experiment reported in §5.

Also, we showed that a standard supervised WSD method works well for the idiom identification. Our system achieved the accuracy of 89.25% and 88.86% with/without idiom-specific features.

Though we dealt with as many as 90 idioms, practical NLP systems are required to deal with many more idioms. Toward a scalable idiom identification, we have to develop an unsupervised or semi-supervised method. The unsupervised method of Birke and Sarkar (2006) requires WordNet. Fortunately, the Japanese WordNet is now available (Isahara et al., 2008), thus we can try their method. Also, CFS propose a language-independent unsupervised method. These could be of help.

At any rate, our idiom corpus will play an important role in the development of unsupervised or semi-supervised methods, and the experimental results obtained in this study will be a good reference point to evaluate those methods.

for the experiment, while 75,011 sentences were used for our experiment. Also, the dataset of HSU came from newspaper articles, while our dataset came from the web.

Acknowledgments

This work was conducted as a part of the collaborative research project of Kyoto University and NTT Communication Science Laboratories.

The work was supported from NTT Communication Science Laboratories and JSPS Grants-in-Aid for Young Scientists (B) 19700141.

We would like to thank the members of the collaborative research group of Kyoto University and NTT Communication Science Laboratories and Francis Bond for their stimulating discussion. Our thanks go as well to Prof. Sato Satoshi, who kindly gave us the list of basic idioms of Japanese.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 89–96.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recoginition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 329–336.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE2008)*, pages 19–22.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference*

of the European Chapter of the Association for Computational Linguistics (EACL-2006), pages 337–344.

Nicole Grégoire, Stefan Evert, and Su Nam Kim, editors. 2007. *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, Prague.

Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors. 2008. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*. ACL Special Interest Group on the Lexicon (SIGLEX), Marrakech.

Chikara Hashimoto and Sadao Kurohashi. 2007. Construction of Domain Dictionary for Fundamental Vocabulary. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07) Poster*, pages 137–140.

Chikara Hashimoto and Sadao Kurohashi. 2008. Blog Categorization Exploiting Domain Dictionary and Dynamically Estimated Domains of Unknown Words. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08) Short paper, Poster*, pages 69–72.

Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006a. Detecting Japanese idioms with a linguistically rich dictionary. *Language Resources and Evaluation*, 40(3–4):243–252.

Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006b. Japanese Idiom Recognition: Drawing a Line between Literal and Idiomatic Meanings. In *The Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006) Poster*, pages 353–360, Sydney, July.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *The sixth international conference on Language Resources and Evaluation (LREC2008)*.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop, COLING/ACL 2006, Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, July.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 1344–1347.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? a case study on extracting pp-verb collocations. In *Proceedings of the ACL-01 Workshop on Collocations*, pages 39–46.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 41–48.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceeding of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering, special issue on Word Sense Disambiguation*, 8(3):359–373.

Satoshi Sato. 2007. Compilation of a comparative list of basic Japanese idioms from five sources. In *IPSJ 2007-NL-178*, pages 1–6. (in Japanese).

Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi, and Kenji Yoshimura. 2004. MWEs as Non-propositional Content Indicators. In *the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 32–39.

Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. 2007. Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 477–485.

Masatoshi Tsuchiya, Takehito Utsuro, Suguru Matsuyoshi, Satoshi Sato, and Seiichi Nakagawa. 2006. Development and analysis of an example database of Japanese compound functional expressions. *Transactions of Information Processing Society of Japan*, 47(6):1728–1741. (in Japanese).

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):497–512.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.