Detecting Japanese Idioms with a Linguistically Rich Dictionary

Chikara Hashimoto
Graduate School of Informatics, Kyoto University

Satoshi Sato
Graduate School of Engineering, Nagoya University

Takehito Utsuro
Graduate School of Systems and Information Engineering, University of Tsukuba

Abstract. Detecting idioms in a sentence is important to sentence understanding. This paper discusses the linguistic knowledge for idiom detection. The challenges are that idioms can be ambiguous between literal and idiomatic meanings, and that they can be "transformed" when expressed in a sentence. However, there has been little research on Japanese idiom detection with its ambiguity and transformations taken into account. We propose a set of linguistic knowledge for idiom detection that is implemented in an idiom dictionary. We evaluated the linguistic knowledge by measuring the performance of an idiom detector that exploits the dictionary. As a result, more than 90% of the idioms are detected with 90% accuracy.

Keywords: Word Sense Disambiguation, Idiom Detection, Linguistic Knowledge

1. Introduction

Detecting idioms in a sentence is important to sentence understanding. Failure of detecting idioms leads to, for example, mistranslation. In the case of the translation service of Excite (www.excite.co.jp/world/), it sometimes mistranslates sentences that contain idioms such as (1a), due to the detection failure.

(1) a. Kare-wa mondai-no kaiketu-ni *hone-o* *o*-tta.
(he-TOP problem-GEN solving-DAT *bone*-ACC *break*-PAST)
"He *made an effort* to solve the problem."
b. "He broke his bone to the resolution of a question."

(1a) has an idiom, *hone-o oru* (bone-ACC break) "make an effort." (1b) is the mistranslation of (1a), where the idiom is interpreted literally.

In this paper, we discuss the linguistic knowledge for idiom detection. The knowledge is implemented in an idiom dictionary that is used by an idiom detector we implemented. Note that the idiom detection we define involves distinguishing literal and idiomatic meanings.[1] Though there has been a growing interest in MWEs (Sag et al., 2002), few proposals on idiom detection take into account ambiguity and transformations. Note also that we tentatively define an idiom as a phrase that is semantically non-compositional. A precise characterization of the notion “idiom” is beyond the scope of the paper.[2]

Two factors make idiom detection difficult: **ambiguity** between literal and idiomatic meanings and the **transformations** that idioms could undergo. In fact, the mistranslation in (1) is caused by the failure to disambiguate between the two meanings. “Transformation” also causes mistranslation. Sentences in (2a) and (2b) contain an idiom, *yaku-ni tatu* (part-DAT stand) “serve the purpose.”

(2) a. Kare-wa *yaku-ni tatu.* (he-TOP *part*-DAT *stand*)
“He *serves the purpose.*”

b. Kare-wa *yaku-ni* sugoku *tatu.* (he-TOP *part*-DAT very *stand*)
“He really *serves the purpose.*”

c. “He stands enormously in part.”

Google’s translation system (www.google.co.jp/language_tools) mistranslates (2b) as in (2c), which does not make sense,[3] though it successfully translates (2a). The only difference between (2a) and (2b) is that bunsetu[4] constituents of the idiom are detached from each other.

Section 2 discusses the classification of Japanese idioms, the requisite lexical knowledge, and implementation of an idiom detector. Section 3 evaluates the detector that exploits the knowledge. After the overview of related works in Section 4, we conclude the paper in Section 5.

2. Linguistic Knowledge for Idiom Detection

2.1. Classification of Japanese Idioms

Requisite linguistic knowledge to detect an idiom depends on how difficult it is to detect it. Thus, we first classify idioms based on **detection difficulty**. The detection difficulty is determined by the two factors: ambiguity and transformability. Consequently, we identify three classes. **Class A** is neither transformable nor ambiguous. **Class B** is transformable but not ambiguous.[5] **Class C** is transformable and ambiguous. Class A amounts to unambiguous single words, which are easy to detect, while Class C is the most difficult. Only Class C needs lexical knowledge for disambiguation (**disambiguation knowledge**). As disambiguation knowledge, we exploit grammatical differences between literal and idiomatic usages. For instance, the phrase, *hone-o oru*, does not allow passivization when used as an idiom, though it does when used literally. Thus, (3), in which the phrase is passivized, cannot be an idiom.

(3) *hone*-ga *o*-rareru (*bone*-NOM *break*-PASS) "A bone is broken."

Disambiguation knowledge depends on its POS and internal structure. As for **POS**, disambiguation of verbal idioms can be performed by the knowledge of passivizability, while that of adjectival idioms cannot. Regarding **internal structure**, detachability should be annotated on every boundary of bunsetus. Consequently, the number of annotations of detachability depends on the number of bunsetus of an idiom.

Thus, Class C needs further classification according to its POS and internal structure, while there is no need for further classification of Class A and B. Thus, Japanese idioms are classified as in Figure 1. The whole picture of the subclasses of Class C remains to be seen.[6]

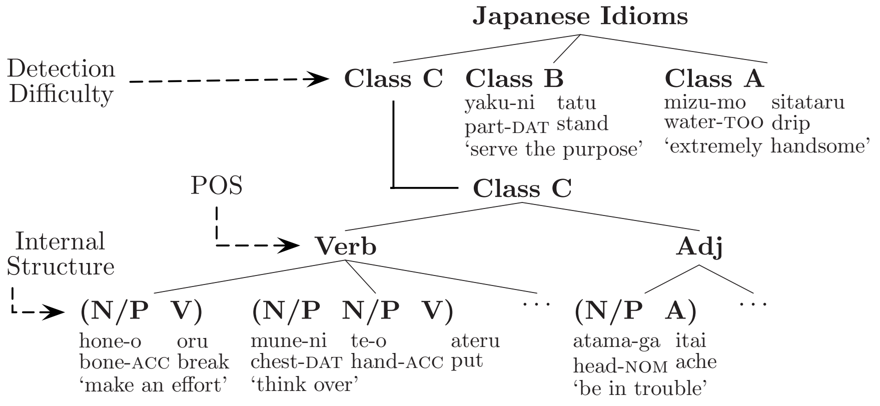

Figure 1. Classification of Japanese Idioms for the Detection Task

2.2. Knowledge for Each Class

Class A needs only string information; idioms of this class amount to unambiguous single words.

Class B requires not only a string but also knowledge about the transformations idioms could undergo, such as passivization. We identify three types of idiom transformations: **1)** Detachment of Bunsetu Constituents, **2)** Predicate's Change, and **3)** Particle's Change. Predicate's change includes inflection, attachment of a negative morpheme, a passive morpheme, and so on. Particle's change represents attachment of topic or restrictive particles.

To normalize the transformations, we utilize a dependency structure, and we call it the **dependency knowledge**.

Class C requires the disambiguation knowledge, as well as all the knowledge for Class B.

A comprehensive idiom detector calls for all the disambiguation knowledge for all the subclasses of Class C, but we have not figured all of them so far. Then, we decided to discover the disambiguation knowledge of the most commonly used idioms as a first step.

2.3. Disambiguation Knowledge for the Verbal (N/P V)

The **verbal (N/P V)** type like *hone-o oru* (bone-ACC break) is the most abundant in terms of both type and token. 1,834 out of 4,581 idioms ($\simeq$40%) in Kindaichi and Ikeda (1989), which is a Japanese dictionary with more than 100,000 words, belong to this type. Also, 167,268 out of 220,684 idiom tokens in Mainichi newspaper of 10 years ('91–'00) ($\simeq$76%) are this type.

To discover the disambiguation knowledge of this type, we first examined the linguistic literature (Miyaji, 1982; Ishida, 2000) on Japanese idioms. Then, among the characteristics, we picked those that could help with the disambiguation of this type and summarized them in (4).

(4) a. Adnominal Modification Constraints
{Relative Clause/Genitive Phrase/Adnominal Word} Prohibition
b. Topic/Restrictive Particle Constraints
c. Voice Constraints
{Passivization/Causativization} Prohibition
d. Modality Constraints
{Negation/Volitional Modality[7]} Prohibition
e. Detachment Constraint
f. Selectional Restriction

For example, the idiom, *hone-o oru*, does not allow adnominal modification by a genitive phrase. Thus, (5) can be interpreted only literally.

(5) **kare-no** *hone-o oru*. (**he-GEN** *bone*-ACC *break*)
"(Someone) breaks **his** bone."

That is, Genitive Phrase Prohibition is in effect for the idiom.

Note that the constraints in (4) are not always in effect for an idiom. For instance, the Causativization Prohibition is invalid for the idiom, *hone-o oru*. In fact, it can be interpreted both literally and idiomatically even when it is causativized.

2.4. Implementation

A rough sketch of the detection algorithm is as follows. **1)** Analyze the morphology and dependency structures of an input sentence. **2)** Look up **dependency patterns** in the idiom dictionary that match a part of the dependency of the input sentence. The dependency pattern of an idiom, which is equipped with all the requisite knowledge to detect it, tells the idiom detector how it can be realized in a sentence. **3)** Mark constituents of the idiom in the sentence if any. We use ChaSen (Matsumoto et al., 2000) as a morphological analyzer and CaboCha (Kudo and Matsumoto, 2002) as a dependency analyzer. Dependency matching is performed by TGrep2 (Rohde, 2005).

The only difference in treatments of Class B and C lies in their dependency patterns. The dependency pattern of Class B consists of only its dependency knowledge, while that of Class C consists of not only its dependency knowledge but also its disambiguation knowledge.

The idiom dictionary consists of 100 idioms, which are all verbal (N/P V) and belong to either Class B or C.[8] Among the knowledge in (4), Selectional Restriction has not been implemented yet. The 100 idioms are those that are listed in either Kindaichi and Ikeda (1989) or Miyaji (1982) and that are used most frequently in 10 years of the Mainichi newspaper. As a result, 66 out of the 100 idioms were Class B, and the other 34 idioms were Class C.[9]

For the detailed of the idiom detector, see Hashimoto et al. (2006).

3. Evaluation

3.1. Experiment Condition

As an **evaluation corpus**, we collected 309 example sentences of the 100 idioms from the Mainichi newspaper of '95. Table I shows the breakdown of the data. "Positive" indicates sentences including a true idiom, while "Negative" indicates those including a literal-usage "idiom."

Table I. The Evaluation Corpus

	Class B	Class C	Total
Positive	200	66	266
Negative	1	42	43

A **baseline system** was prepared to see the effect of the disambiguation knowledge. The baseline system was the same as the detector except that it exploited no disambiguation knowledge.

3.2. Result

The result is shown in Table II. The differences between the performance of the two systems are marked with **bold**. Recall(R), Precision(P), and F-Measure(F) are calculated using the following equations.

Table II. Performance of the Detector (left) and the Baseline (right)

	Class B	Class C	All	Class B	Class C	All
Recall	0.975	0.939	0.966	0.975	0.939	0.966
Precision	1.000	**0.697**	**0.905**	1.000	0.602	0.862
F-Measure	0.987	**0.800**	**0.935**	0.987	0.734	0.911

$$R = \frac{|Correct\ Outputs|}{|Positive|} \quad P = \frac{|Correct\ Outputs|}{|All\ Outputs|} \quad F = \frac{2 \times P \times R}{P + R}$$

As a result, more than 90% of the idioms can be detected with 90% accuracy. Note that the detector made fewer errors due to the employment of the disambiguation knowledge.

The result shows good performance. However, there is still a long way to go to solve the most difficult problem of idiom detection: drawing a line between literal and idiomatic meanings. In fact, the precision of detecting idioms of Class C remains less than 70% as in Table II. Besides, the detector successfully rejected only 15 out of 42 negative sentences (35.71%).

3.3. Discussion of the Disambiguation Knowledge

Disambiguation amounts to **i)** rejecting negative sentences with observable evidence, **ii)** rejecting negative ones without observable evidence, or **iii)** accepting positive ones. i) is relatively easy since evidence in a sentence tells us that it is NOT an idiom, while ii) and iii) are difficult. Our method performs only i), and thus has an obvious limitation.

Next, we look at cases of success or failure of rejecting negative sentences. There were 15 cases where rejection succeeded, which correspond to i). The breakdown is as follows[10]: Genitive Phrase Prohibition rejected 6 cases; Relative Clause Prohibition rejected 5 cases; Detachment Constraint rejected 2 cases; Negation Prohibition rejected 1 case. Thus, Adnominal Modification Constraints are the most effective.

27 cases where rejection failed are classified into two types; those that could have been rejected by the Selectional Restriction (5 cases), and those that might be beyond the current technology (22 cases).

Thus, Selectional Restriction would have been effective. A part of a sentence that Selectional Restriction could have rejected is below.

(6) basu-ga *tyuu-ni ui*-ta. (bas-NOM *midair*-DAT *float*-PAST)
"The bus floated in midair."

An idiom, *tyuu-ni uku* (midair-DAT float) "remain to be decided," takes as its argument something that can be decided, i.e., ⟨`1000:abstract`⟩ rather than ⟨`2:concrete`⟩ in the sense of the *Goi-Taikei* ontology (Ikehara et al., 1997). Thus, (6) has no idiomatic sense. An example of a case that might be beyond the current technology is illustrated in (7).

(7) ase-o nagasi-te huku-o kiru-yorimo, hadaka-ga gouriteki-da.
(sweat-ACC shed-and clothes-ACC wear-rather.than, nudity-NOM rational-DECL)
"It makes more sense to be naked than wearing clothes in a sweat."

The phrase *ase-o nagasu* (sweat-ACC shed) could have been an idiom meaning "work hard." It is contextual knowledge that prevented it from being the idiom. Our technique is unable to handle such a case, since no observable evidence is available.

Finally, the 42 negative sentences consist of 15 sentences, which we could disambiguate, 5 sentences, which Selectional Restriction could have disambiguated, and 22, which are beyond the current technique. Thus, the real challenge lies in 7% ($\frac{22}{309}$) of all idiom occurrences.

4. Related Work

Few attempts have been made to detect idioms in a sentence with ambiguity and transformations taken into account. In fact, most of them only create catalogs of idiom (Tanaka, 1997; Shudo et al., 2004).

A notable exception is Oku (1990); his idiom detector takes the ambiguity and transformations into account. However, he only uses Genitive Phrase Prohibition, Detachment Constraint, and Selectional Restriction, which would be too few to disambiguate idioms. Although Oku (1990) seems to think little of constraints on what forms an idiom itself is allowed to appear in, linguistic knowledge about idiom forms plays an important role in detecting idioms in a language, such as Japanese, where syntactic arguments are easily dropped and hence Selectional Restriction often cannot help.

Our technique has the limitation that we cannot reject literal-usage phrases without observable evidence. In that case, the technique discussed in Katz and Giesbrecht (2006), who tried to disambiguate German MWEs by means of Latent Semantic Analysis, would be helpful.

However, using only statistical techniques would not give a satisfactory solution, since each idiom shows various kinds of peculiarities of its own and thus poses a serious sparseness problem. Rather, combining statistical techniques with categorical linguistic knowledge such as those discussed in this paper will provide a far better result.

Fazly and Stevenson (2006) proposes a statistical method to see in which syntactic forms a given idiom can appear. Though we relied on native speakers' intuition to construct the disambiguation knowledge, it would be helpful to make use of their method for the disambiguation knowledge construction.

As for the classification of Japanese idioms, Oku (1990) classifies idioms according to only the transformability and does not take the ambiguity into account. On the other hand, Shudo et al. (2004) make a very fine distinction between Japanese idioms. Basically, they assign fine-grained linguistic knowledge that corresponds to our disambiguation knowledge to all idioms whether they are ambiguous or not. But, from the viewpoint of the idiom detection, this is too much; only ambiguous idioms need detailed linguistic information. Related to this is that while they take the compositionality into account, they do not care about the ambiguity, which is indispensable for the idiom detection.[11]

Our classification of idioms correlates loosely with that of MWEs by Sag et al. (2002). Japanese idioms that we define correspond to *lexicalized phrases*. Among lexicalized phrases, *fixed expressions* are equal to Class A. Class B and C roughly correspond to *semi-fixed* or *syntactically-flexible expressions*. Note that, though the three subtypes of lexicalized phrases are distinguished based on what we call **transformability**, no distinction is made based on the **ambiguity**.[12]

5. Conclusion

Aiming at Japanese idiom detection with ambiguity and transformations taken into accout, we proposed a set of linguistic knowledge for idioms and implemented a linguistically rich idiom dictionary and an idiom detector that exploits the dictionary. We maintain that requisite knowledge depends on its transformability and ambiguity; transformable idioms require the dependency knowledge, while ambiguous ones require the disambiguation knowledge as well as the dependency knowledge. As the disambiguation knowledge, we proposed a set of constraints applicable to a phrase when it is used as an idiom. The experiment showed that more than 90% idioms could be detected with 90% accuracy but the success rate of rejecting negative sentences remained 35.71%. The experiment also revealed that, among the disambiguation knowledge, Adnominal Modification Constraints and Selectional Restriction are the most effective.

For future work, we will reveal all the subclasses of Class C and all the disambiguation knowledge, and apply a machine learning technique to disambiguating those cases that the current technique is unable to handle, i.e., cases without observable evidence.

Acknowledgements

A special thank goes to Gakushu Kenkyu-sha, also known as Gakken, who permitted us to use Gakken's Dictionary for our research.

Notes

[1] Some idioms represent two or three idiomatic meanings. But we only check whether a phrase is used as an idiom or not.

[2] For a detailed discussion of what constitutes the notion of (Japanese) idiom, see Miyaji (1982), which details usages of commonly used Japanese idioms.

[3] In fact, the idiom has no literal interpretation.

[4] A bunsetu is a syntactic unit in Japanese, consisting of one independent word and more than zero ancillary words.

[5] One can devise a context that makes the literal interpretation of those Classes possible. However, virtually no phrase of Class A or B is interpreted literally in real texts, and we think our generalization safely captures the reality of idioms.

[6] There were many more variations in the internal structure of idiom than we had expected. To make clear what internal structures there are in Japanese idioms, careful investigation is required, which we could not carry out in this study.

[7] "Volitional Modality" represents those verbal expressions of order, request, permission, prohibition, and volition.

[8] It might seem unfeasible to compile a large-scale idiom dictionary that is equipped with the lexical knowledge described so far. In fact, only Class C requires detailed linguistic information (the disambiguation knowledge), which must be described by relying on native speakers' intuition, while the lexical knowledge of Class A and B (two-thirds of all idioms) is compiled automatically. Related to this, the disambiguation knowledge for Class C has been compiled by the authors' intuition in this study. And we found that there were far fewer disagreements about the judgments than we had expected.

[9] The most frequently used 100 idioms in Kindaichi and Ikeda (1989) cover 53.49% of all tokens in the Mainichi newspaper of 10 years. Thus, our dictionary accounts for approximately half of all idiom tokens in a corpus.

[10] One rejection was done by the dependency analysis error.

[11] Semantic compositionality does not play an important role in the idiom detection, although most papers concerning MWEs are obsessed with it.

[12] The notion of *decomposability* of Sag et al. (2002) and Nunberg et al. (1994) is independent of **ambiguity**. In fact, ambiguous idioms are either decompos-

able (*hara-ga kuroi* (belly-NOM black) "black-hearted") or non-decomposable (*hiza-o utu* (knee-ACC hit) "have a brainwave"). Also, unambiguous idioms are either decomposable (*hara-o yomu* (belly-ACC read) "fathom someone's thinking") or non-decomposable (*saba-o yomu* (chub.mackerel-ACC read) "cheat in counting").

References

Fazly, A. and S. 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)*. pp. 337–344.

Hashimoto, C., S. Sato, and T. Utsuro: 2006, 'Japanese Idiom Recognition: Drawing a Line between Literal and idiomatic Meanings'. In: *COLING/ACL 2006*. Sydney, pp. 353–360.

Ikehara, S., M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi: 1997, *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten.

Ishida, P.: 2000, 'Doushi Kanyouku-ni taisuru Tougoteki Sousa-no Kaisou Kankei (On the Hierarchy of Syntactic Operations Applicable to Verb Idioms)'. *Nihongo Kagaku (Japanese Linguistics)* **7**, 24–43.

Katz, G. and E. 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*. pp. 12–19.

Kindaichi, H. and Y. Ikeda (eds.): 1989, *Gakken Kokugo Daijiten (Gakken's Dictionary)*. Gakushu Kenkyu-sha.

Kudo, T. and Y. Matsumoto: 2002, 'Japanese Dependency Analysis using Cascaded Chunking'. In: *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*. pp. 63–69.

Matsumoto, Y., A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara: 2000, 'Morphological Analysis System ChaSen version 2.2.1 Manual'. Nara Institute of Science and Technology.

Miyaji, Y.: 1982, *Kanyouku-no Imi-to Youhou (Usage and Semantics of Idioms)*. Meiji Shoin.

Nunberg, G., I. A. Sag, and T. Wasow: 1994, 'Idioms'. *Language* **70**, 491–538.

Oku, M.: 1990, 'Nihongo-bun Kaiseki-ni-okeru Jutsugo Soutou-no Kanyouteki Hyougen-no Atsukai (Treatments of Predicative Idiomatic Expressions in Parsing Japanese)'. *Journal of Information Processing Society of Japan* **31**(12), 1727–1734.

Rohde, D. L. T.: 2005, 'TGrep2 User Manual version 1.15'. Massachusetts Institute of Technology. http://tedlab.mit.edu/~dr/Tgrep2/.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger: 2002, 'Multiword expressions: A pain in the neck for NLP'. In: *Computational Linguistics and Intelligent Text Processing: Third International Conference*. pp. 1–15.

Shudo, K., T. Tanabe, M. Takahashi, and K. Yoshimura: 2004, 'MWEs as Non-propositional Content Indicators'. In: *the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*. pp. 32–39.

Tanaka, Y.: 1997, 'Collecting Idioms and Their Equivalents'. In: *IPSJ SIGNL 1997-NL-121*.