Flexible Automatic Look-up of English Idiom Entries in Dictionaries

Koichi Takeuchi, Takashi Kanehila, Kazuki Hilao, Takeshi Abekawa and Kyo Kageura

Graduate School of Natural Science and Technology
Okayama University
Tsushimanaka 3-1-1, Okayamashi 700-8530, Japan
koichi@cl.it.okayama-u.ac.jp

Graduate School of Education
Tokyo University
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan
{abekawa,kyo}@p.u-tokyo.ac.jp

Abstract

Although many high-quality dictionaries contain a sufficient number of idioms for their intended users, the methods available for looking up entries in both paper and electronic dictionaries as well as in machine translation systems are not satisfactory. Providing an adequate automatic look-up function is complicated by the existence of idiom variants, which sometimes can be very creative. The problem is further complicated by the fact that the possible range of idiom variations has not been described in a computationally tractable way. Against this backdrop, we analysed the variation patterns of idioms using manually-created idiom variation data, and, on the basis of that, developed an idiom look-up system that automatically matches idiom variants in English texts with the canonical forms of idiom entries in dictionaries. The experimental results showed our system performs sufficiently well to be used in real-world settings, including as an aid for translators, which is our overall aim.

Introduction

We are currently developing a system that aids English-to-Japanese volunteer translators who translate online documents and publish translated documents online. Among the many reference functions and aspects of reference content that require enhancement, translators have identified improvement in idiom look-up functions as a key issue they would like to see addressed by translation aid systems (Kageura et. al., 2006).

Although idioms provided by many high-quality dictionaries (e.g. Sanseido, 2004; McCaleb and Iwasaki, 2003) are basically satisfactory for translators, the methods available for looking up idiom entries are far from satisfactory, not only in paper dictionaries but also in electronic dictionaries. This is partly because the user has to guess the core constituent words of an idiom in order to consult a dictionary.

Automatic look-up methods embodied in machine translation systems are not satisfactory, either. Take, for instance, the following examples:

(1) *He said that with his tongue in his cheek.*
(2) *He said that with his big fat tongue in his big fat cheek.*

Although many available machine translation systems successfully detect the idiomatic expression "with one's tongue in one's cheek" in (1), none among those we checked (e.g. Excite, 2005; Fujitsu, 2005; LogoVista, 2005; Sharp, 2004; Toshiba, 2005)[1] could properly translate (2). Most existing methods for looking up idioms cannot deal with the rich variations in idioms, that are abundant in ordinary texts[2].

In the field of natural language processing (NLP), much research has been carried out into the automatic extraction of collocations and idioms (e.g. Piao, 2006; Smadja, 1993; Widdows and Dorow, 2005), but not much work has been devoted to the automatic matching of idiom entries to their occurrences in running texts. As translators are basically satisfied with the idioms provided in existing high-quality dictionaries but frustrated with the poor look-up functions, to develop enhanced look-up functions for idioms is of utmost importance from the point of view of aiding translators. Although a few important related studies exist (Carl and Rascu, 2006; Jacquemin, 2001; Yoshihashi et. al., 2005), and many translation memory systems realise flexible approximate matching of similar sentence or phrasal constructions (Similis, 2006; Trados, 2006), the task of flexible automatic look-up of idioms is yet to be fully explored.

Against this backdrop, we are developing a mechanism that automatically matches English idiom occurrences in texts and their possible variations with idiom entries in dictionaries, as part of an overall project that aims at developing a system to aid English-to-Japanese online volunteer translators. In the following, we will first clarify translators' basic requirements. Then we will provide the basic patterns of idiom variations based on manually-constructed idiom variation data, and explain the automatic idiom look-up system that takes into account major syntagmatic idiom variations. Finally, we will provide an evaluation of the method and outline areas for further improvement.

Translators' basic requirements

In order to clarify requirements for translation-aid system, we consulted eight translators working online. We also sent a questionnaire to other online translators and obtained 12 replies. In relation to idiom look-up functions, two important features became clear.

Firstly, translators do not want the system to provide a single idiom entry that matches textual occurrences. This is more to do with the fact that checking multiple possibilities is an inherent and essential part of proper translation than with the fact that it is difficult to develop a satisfactorily high-performance automatic idiom look-up system. In other words, from the point of view of translators, for the system to provide multiple possibilities of matching idioms is not a defect but a necessity if the system is to be useful for them. After all, translators check many candidates that they do not eventually use in their translations. What is important is reducing translators' burden as well as the quality of candidates the system proposes.

Secondly, translators — as language practitioners — want the system to be able to deal with variations in a more flexible way than described by linguists. For instance, for language

[1] Sharp (2004) can detect some gapped idiom occurrences, but it fails to detect complex idiom variations.

[2] For instance, our rough survey of online documents revealed that the idiom "hang on" in its basic form, "hang on" with insertion, and "hang on" with passivisation (with or without insertion) occur in roughly the same frequency. The same is true for "dumb down".

practitioners "shoot the breeze" could be passivised (or they could imagine a situation in which they face the passive form of this idiom or they passivise the idiom by themselves in the process of writing), while according to Numberg et. al. (1994), it is not possible. The idiom "go halves" can be used in the form "go exact halves", while Nicolas (1995) claimed that this is not possible. In a sense, even what descriptive (i.e. non-prescriptive) linguistics provides is too prescriptive for the reality of texts that translators face in their daily activities. This is especially the case for online documents (Aitchison and Lewis, 2003), which are dealt with by the translators our system targets.

From the point of view of system specifications, these two requirements mean that the system can — and should — provide overmatching results, with recall as close as 100%. They also mean that, in the evaluation of system performance, the concept of "precision" should be defined not in terms of the "correct" choice of candidate but in terms of its usefulness for translators. We will come back to this point later when we evaluate the performance of our system.

Idiom variation patterns

There are studies and reference books that describe (English) idiom variations at a variety of levels (Benson, 1985; Biber, 1999; Čermák, 1970; Fraser, 1970; Moon, 1998; Nicolas, 1995; Numberg et. al., 1994; Quirk et. al., 1985). On the basis of these, and taking into account the comments we obtained from translators, we first classified the idiom variation patterns as follows:

(1) Type variants or families: Types of idiom variations that are (or theoretically should be) registered in dictionary entries. An example is "run around [round] like a blue-arsed fly". From the practical point of view, this type of variant can be dealt with by the variation indications in idiom entries in dictionaries.

(2) Variations created by external factors: Passivisation ("the breeze was shot") and topicalisation ("It is these strings that he pulled") are typical examples of this type of variation. These variations are generally created by applying syntactic operations defined outside the idioms themselves, and could be dealt with uniformly by a few basic rules.

(3) Variations applied to parts or within the construction of idioms: many syntagmatic insertions ("go halves" → "go exact halves") and paradigmatic replacements ("head screwed on right" → "head screwed on wrong/left") fall under this category. This type of variation is expected to be neither straightforwardly clear nor completely unmanageable.

(4) Highly creative variations: "point of view" → "ball-point pen of view". This class of variation is expected to be unmanageable for the time being.

We focus on the third category (3) of idiom variations in this work, for reasons mentioned above.

In order to develop a look-up function for idiom entries, existing linguistic studies have three limitations: (i) as mentioned, they tend to be too restrictive from the point of view of the reality of texts that translators deal with; (ii) the description of variations is not given in a computationally tractable way; and (iii) the number of variations given in these studies is small[3].

Given the dearth of basic idiom variation data, we started by constructing idiom variation data manually. We took idiom entries from a widely-used English-Japanese idiom dictionary (McCaleb and Iwasaki, 2003), and asked three native English speakers (two of whom were professional editors) to create idiom variations with examples. We asked the informants to imagine they were writing or editing articles in the culture section of newspapers and be as creative as possible within that restriction. Table 1 shows the basic quantities of the idiom variation data. The quantity of data is rather small, and we are intending to augment it further in the same manner, asking informants to construct variations.

Informants	(a) ♯idioms	(b) ♯variants	(b)/(a)
h	475	469	0.99
j	661	890	1.35
s	777	822	1.06
Total	1913	2181	1.14

Table 1: Basic quantities of idiom variation data ('h', 'j' and 's' indicate the three informants)

Note, however, that using large corpora to collect the data is not our priority, for a few reasons: (i) it is difficult to extract idiom variations from large corpora, except for easily predictable regular types which could be covered by rules; (ii) it is difficult to identify the threshold of possible variations, which tend to occur in low frequencies (translators do not choose a text by the representativeness of the language in the text or by the overall frequencies in the corpora of idiom variations used in the text); (iii) translators are dealing with individual texts and not representative language expressions, so frequencies in large corpora do not automatically mean importance for translators; and (iv) the frequency of idiom variations detected in large corpora correlates with the frequency of individual idioms, and frequently occurring idioms tend to be the ones that translators are least interested in and therefore the variations of which are less important from the translators' point of view.

Our intention in referring to the data is not to observe common usage or dominant patterns but to define the tractable range of variation patterns, which would hopefully correspond to the range of practically possible variations; we have no interest in covering only frequently occurring patterns at the expense of less frequent but computationally tractable patterns. As such, it is expected that the use of large corpora would not cover up the shortcomings of the size of the manually constructed data. For instance, one of the informants provided the variation "take the wild plunge" of the idiom "take the plunge," which only brings up four hits in a Google search. From the point of view of translators, this and other rare variations, if technically possible, should be covered when they occur[4]. That our manually-constructed

[3] Some reference books (e.g. Oxford, 2001; Collins, 2002) give useful information, though not fully for variation patterns, for some class of idiomatic expressions, so we referred to them whenever was useful.

[4] One of the translators we spoke with commented: "Linguists?

Type (tag)	Example	#
Paradigmatic replacement	the boiling point → the burning point	759
Syntagmatic augmentation	take off → take right off	1203
Deletion	not get to first base → got to first base	3
Dependent multiple replacement	more dead than alive → more alive than dead	39
Dependent multiple augmentation	can swing it → can swing it no problem	101
Replacement and augmentation	weak as a baby → strong as a baby ox	91
Deletion and replacement	go back to the basics → plunge into the basics	20
Deletion and augmentation	people will talk → people happily talk	8
Others	take it from me → rely on me	95

Table 2: Broad classification of idiom variations

dataset includes such rare cases, therefore, is not a demerit but a merit (on condition, of course, that basic patterns are covered). On the other hand, one might argue that this will result in the system potentially facing the problem of overmatching. But rare cases, if they do not occur, do not cause problems, because the basic task here is matching dictionary entries to textual occurrences, i.e. both ends are given.

This manually-constructed data was then analysed and variation patterns were identified (Kageura and Toyoshima, 2006). Table 2 shows the basic variation patterns. In the table, such types as "dependent multiple replacement" etc. indicate that more than one type of mutually dependent variations was observed.

As can be seen from Table 2, the major patterns are syntagmatic augmentations and paradigmatic replacement. Here we focus on variations by syntagmatic augmentation, and formalise the descriptions of syntagmatic augmentation patterns. In doing so, we assume the use of POS-taggers and morphological analysers. Though high-performance parsers exist, we did not use them for two main reasons: (i) idiomatic expressions often cross over the border of constituents given by parsers (in which case we would need to flatten the parse tree anyway), and (ii) to achieve recall as close to 100% as possible is most important, and for that aim the loose definition of patterns provides a better starting point than the rigid description of variations using structural information.

There are two different approaches for describing POS level patterns for variations of syntagmatic augmentations: (a) taking all constituents of the idiom into account, or (b) taking only binary constituents adjacent to words inserted into the idiom. For instance, assuming that the idiom expression "take the plunge" has as a variation "take the wild plunge", we can describe the POS level patterns as "Verb Det Noun" → "Verb Det Adj Noun" in approach (a), or "Det Noun" → "Det Adj Noun" in approach (b). In the current work we took approach (b) because we assume that: (i) inserted words are mostly bound by the adjacent words and their grammatical categories; (ii) to take into account the overall grammatical patterns would immediately lead us to taking into account the individual idioms with lexical substance; and (iii) the requirement of high recall is of utmost importance at the current stage.

Using the POS-information of adjacent elements, we formulated the basic patterns of idiom variations by syntagmatic augmentation as shown in Table 3 ("Prep", "Adj", "Adv", "PosPro" and "PerPro" denote preposition, adjective, adverb, possessive pronoun (e.g. "my", "her" etc.) and personal pronoun (e.g. "I", "he" etc.), respectively).

POS tags of constituents of idioms	POS of inserted word
(Noun, Noun)	Noun, Adj
(Noun, Prep)	Noun, Adj, Adv
(Noun, Adj)	Adj, Adv
(Noun, Verb)	Adv, Aux
(Adj, Noun)	Noun, Adj, Adv
(Adj, Prep)	Noun
(Adv, Adj)	Adv
(Adv, Adv)	Adv
(Adv, Prep)	Adv
(Adv, Verb)	Adv
(Conj, Verb)	Adv
(Conj, Prep)	Adv
(Conj, Noun)	Adj
(Verb, Adv)	Adv, Adj
(Verb, Noun)	Noun, Adj
(Verb, Adj)	Adv, Adj
(Verb, Prep)	Adv, Adj
(Prep, Noun)	Noun, Adj
(Prep, Adj)	Adv, Adj, Noun
(Det, Adj)	Adj, Adv, Noun
(Det, Noun)	Noun, Adj, Adv
(PosPro, Noun)	Adj, Noun
(PerPro, Noun)	Adj
(PerPro, Adv)	Adv
(PerPro, Prep)	Adj, Adv

Table 3: Idiom variation rules for insertion

Each variation rule in Table 3 indicates the POS sequence of constituents of an idiom and the POS tag of the inserted word. For instance, the POS pattern of (Det, Noun) in constituents of an idiom can take either a noun, adjective or adverb. This rule covers the variation from "take the plunge" to "take the *wild* plunge".

Note that the described range of variation patterns, when incorporated into automatic matching algorithms, can be overgenerative. We can, however, reasonably expect that the overmatching will be within the manageable range, because, as mentioned, the computational problem is defined here as a problem of matching when both ends are given, rather than

Ah, those who cannot read literary texts but still have the audacity to think themselves to be language specialists!" As computational linguists we should try to bridge this gap between linguists and translators.

a problem of generating acceptable variations.

The idiom look-up system

The system consists of three processing modules: (1) the pre-processing module in which the input text is processed to facilitate automatic matching, including POS-tagging and normalisation of expressions, (2) the surface matching module in which all the possible idiom entries are detected by using AND matching of constituent elements of idioms with words occurring in texts, and (3) the filtering module in which the undesired candidates detected in the surface matching module are filtered out by using the POS-based variation restriction rules constructed based on the patterns given in Table 3. The input of the target system is an English text and the output is idiom candidates occurring in the text with their Japanese translations provided in the dictionary. Figure 1 shows the overall flow of the system. We will elaborate each of these modules below and illustrate the system interface.

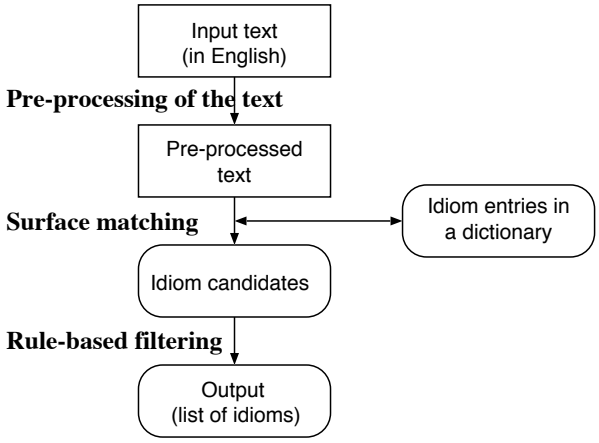


Figure 1: The overall flow of the system

Pre-processing

In the pre-processing module, we first assign POS information to the input text by using Tree-tagger (TreeTagger, 2004). After that, we apply the standardisation rules to the surface word forms that occur in texts. This is because, in many cases, the word form in the text is different from the word form in dictionary entries. For example, "took his seat" can be found in the text, but what is registered in the dictionary is "take one's seat". We adjusted the word form in the text so that it could be matched to the dictionary entries. Four types of formal standardisations are applied at this stage:

(1) Inflected forms of verbs are transformed into basic forms. In addition, we added "do" or "doing" to absorb the matching of idioms whose entries are registered as something like "cannot help *doing*" in the dictionary.

(2) Plural forms of nouns are transformed into singular forms.

(3) Articles are paradigmatically expanded so that the occurrence "a", for instance, can be matched with an entry with "the". This may often lead to false matching as some idioms require the strict use of either definite or indefinite articles, but we found we can gain more than we lose by applying this processing, from the point of view of system requirements.

(4) Personal pronouns are standardised into basic forms.

Table 4 shows the basic standardisation patterns of word forms. Note that in the actual matching, we also retain the original forms. In addition to these, we applied a small amount of pre-processing such as splitting hyphenated words, etc.

Input word	Pattern
(1) verb	surface, basic form, do, doing
(2) plural noun	surface, singular noun
(3) particle	a, an, the
(4) my, his, etc.	surface, one's
(4) myself, herself, etc.	surface, oneself

Table 4: Standardisation of words

Surface matching

In the surface matching module, we carry out an extensive retrieval in which all the possible idiom candidates can be detected. We retrieve idiom entries whose constituents all match the textual sequences of words in order. In the dictionary entries, such examples as "make A of B" or "have ... in" exist. In the surface matching module, we deal with these "position fillers" as wild cards. Figure 2 shows an example of surface matching. In Figure 2, the dictionary entry "have one's eye on" is detected as an idiom candidate, because its constituent words all match the input words in the text.

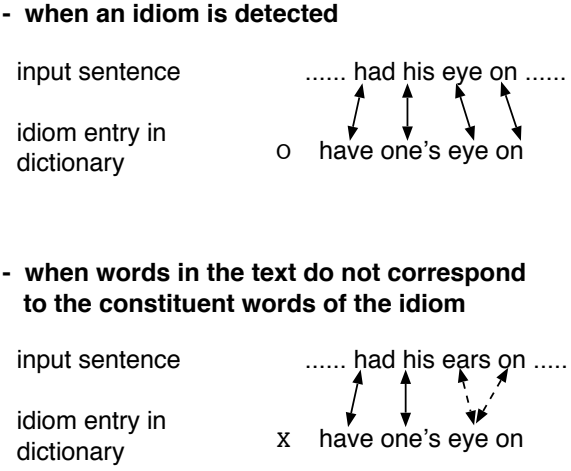


Figure 2: Surface matching

Filtering with POS patterns

As we emphasise exhaustive retrieval in the surface matching module, many of the idiom candidates detected in the surface matching module are expected to be non-relevant idioms. In the filtering module, we filter out many irrelevant idioms by using the rules constructed on the basis of POS-patterns given in Table 3.

Figure 3 shows an example of filtering. In this case, the surface matching module detected the three idiom candidates "make a habit of doing", "wake up", "make after" for the input text "I make a habit of stretching after I wake up." The basic adjacent patterns given in Table 3 require that what can be inserted between a verb and a preposition is either

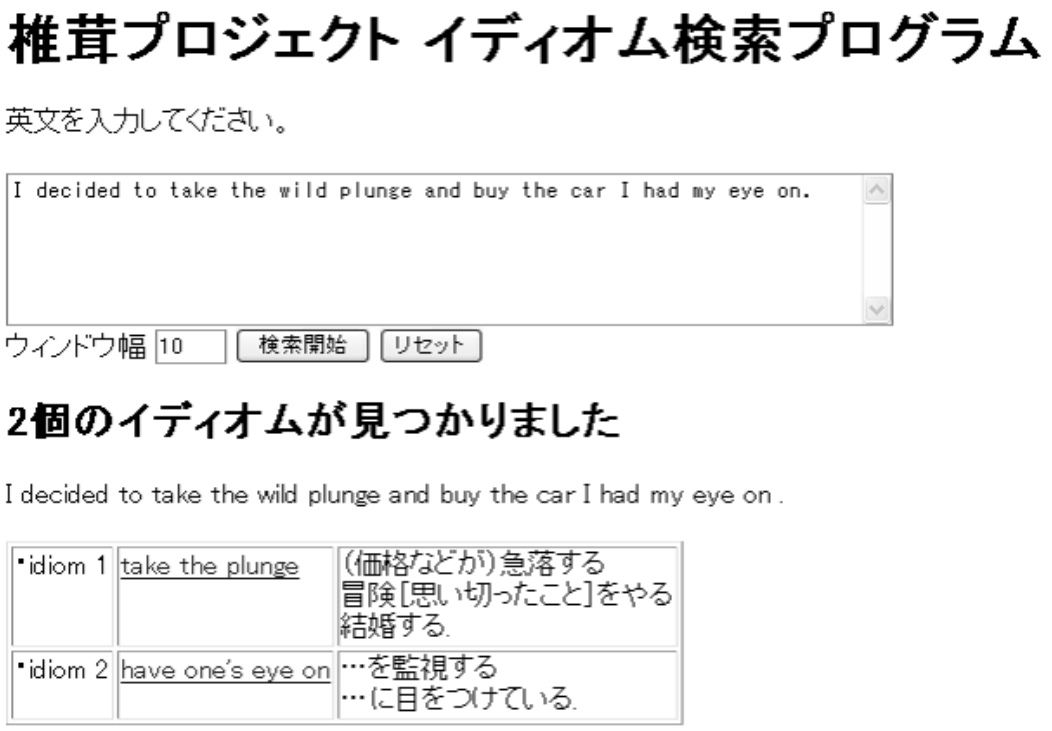


Figure 4: System interface

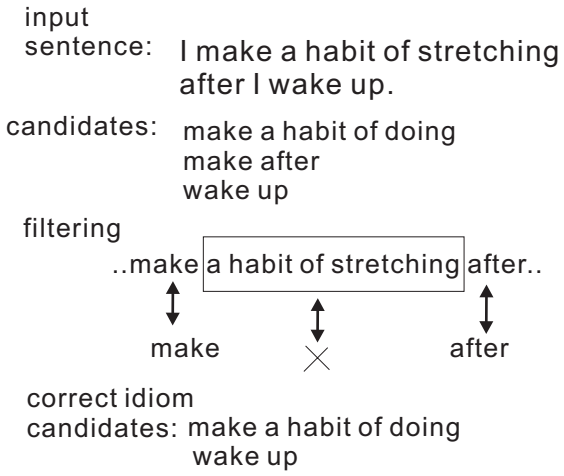


Figure 3: Filtering with POS patterns

an adverb or an adjective. As the construction "a habit of stretching" is neither an adverb nor an adjective, the candidate "make after" is filtered out and excluded from the final output.

Interface of the independent idiom look-up system
Figure 4 shows the system interface. The interface consists of an input area, parameter specification area in which the user can specify a window size of a certain number of words within which idiom candidates are to be searched, and an output area. When the user inputs an English text, and clicks the "search" button, the system outputs the idiom candidates with their meaning (in Japanese). By moving the mouse over an output idiom, the matched parts of the input text in the input area are emphasised. Figure 4 shows that "take the plunge" and "have one's eye on" were detected for the input: "I decided to take the wild plunge and buy the car I had my eye on."

Integration to the translation editor environment
In addition to the independent system interface, we incorporated the idiom look-up system into the integrated translation editor environment QRedit (Abekawa and Kageura, 2007). Figure 5 shows the interface in which automatic idiom look-up functions within the integrated environment. The screen shot shows that the idiom entry "(with) one's tongue in one's cheek" matches the sentence "He said that with his big fat tongue in his big fat cheek."

Evaluation

We carried out evaluation experiments in order to observe the overall performance of the system, as well as the following three aspects: (1) the effect of standardisation of words; (2) the effect of the POS-based filtering; (3) the overall performance of the system.

Experimental setup
We observed how many correct idiom candidates our system was able to locate with each set of data. The data set used for evaluation were: (a) 100 sentences containing idiom variations, randomly extracted from the idiom variation data mentioned in section 2, and (b) data consisting of 20 newspaper and journal articles (five articles taken from the BBC online news site, five articles from *The Nation* website, five articles from the *The Independent* website, and five articles from the *New York Times* website). We manually identified and tagged the idioms for these articles. The window size is set to a sentence. For both the data (a) and (b), we compared the method of surface matching only with the method of surface matching and filtering with POS-based information.

Correct outputs were defined as follows:

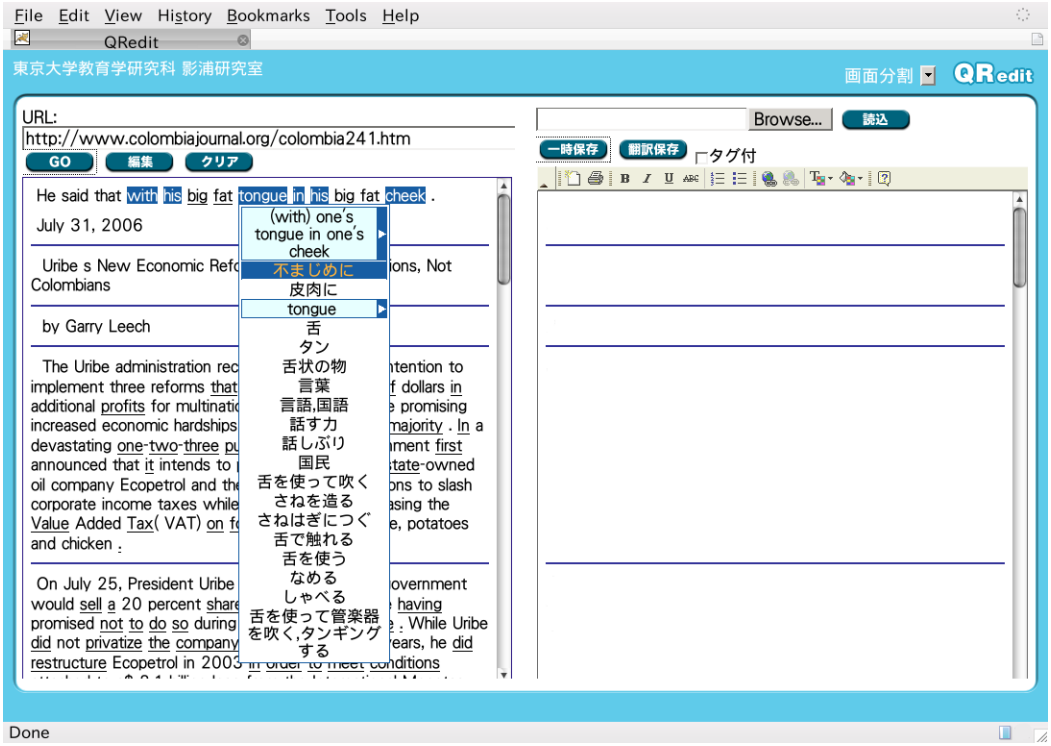


Figure 5: Idiom look-up functions integrated into the translation editor

(1) Idioms that are actually used in the input text;

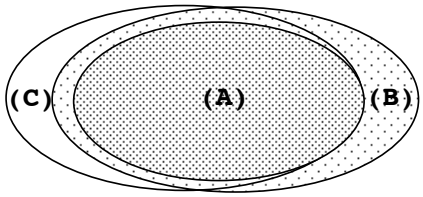
(2) Variations of (1) created by replacement of articles. For instance, when "come to the point" is the idiom actually used in the text, "come to a point" is evaluated as correct output.

(3) Variations of (1) created by the singular/plural forms of nouns. When a constituent noun in the actual idiom is plural, such as "in spirits", "in spirit" is evaluated as correct output.

(4) Embedded idioms. When the actual idiom is "come to the point", we evaluate "to the point" as correct output.

These criteria were set on the basis of the consultation with eight online volunteer translators. Translators, when they come across expressions they cannot readily translate analytically, check several possible idiom candidates to reach the final correct translation. As mentioned, checking multiple candidates is not an optional, extraneous process that translators would like to omit if they can, but rather an essential process by which they make sure that their final decision is correct. As such, given that no automatic processing can substitute for human decision making, translators want the automatic idiom look-up system to show multiple candidates that are close to the set of candidates that translators actually check. The criteria given here are an approximation to this. Figure 6 shows the range of the correct output defined here[5].

[5]In an evaluation of a system that provides translation information for human translators, Sharoff et. al. (2006) introduces five grades: 5 = the suggestion is an appropriate translation as it is; 4 = the suggestion can be used with some minor amendment; 3 = the suggestion is useful *as a hint for another, appropriate translation*; 2 = the suggestion is not useful, even though it is still in the same domain; 1 = the suggestion is totally irrelevant. The grades "5", "4" and "3" can be interpreted as corresponding roughly to (C) in Figure 6, which includes, but is not limited to, the correct idioms.



(A) Correct idioms in running text
(B) Correct candidates of idioms defined here
(C) Ideally desirable idiom candidates for translators

Figure 6: The position of correct idiom candidates

Evaluation measures

From the viewpoint of translation support, it is important for our system to detect all the idioms that appear in the text. Therefore, recall is the most important factor at this stage. Improvements in precision should be elaborated without negatively affecting recall. F-measure is irrelevant.

$$precision = \frac{\#CorrectSystemOutputs}{\#SystemOutputs}$$

$$recall = \frac{\#CorrectSystemOutputs}{\#AllCorrectIdioms}$$

Result of the experiments

Tables 5 and 6 show the results of the experiments. In both sets of data, the recall is as high as 0.97, although the precision varies between the data sets (a) and (b). The high recall is very promising, as the essential problem that prompted us to develop this system was the difficulty experienced by translators in looking up idioms, the improvement of which requires high performance in recall. The precision of 0.5 to

0.7 is in a practically useful range. This figure means that translators are provided with twice as many candidates as they need to check. For most texts this will not be an excessive number. Experimental use by a translator in an integrated editor environment has shown that the problem is more to do with the quality of unnecessary candidates rather than the quantity (because they reduce the translators' expectations of the system).

The result of filtering with POS-based patterns for the data set (a) and (b) shows that this filtering improved precision greatly, without negatively affecting recall, which proves the usefulness of POS-based patterns for our aim. The result of experiments with filtering on data set (b) shows that the recall for data set (b) is higher than that for data set (a), but the precision for data set (b) is lower than that for data set (a). The cause of the low precision can be summarised as follows: (1) it is easy for our system to detect incorrect idioms for data set (b), because real-world English texts tend to have longer sentences; (2) there is room for improvement in the filtering rules. On the other hand, the higher recall for data set (b) can be explained by the fact that the manually constructed basic data include some highly creative examples, which are rather difficult to detect, while the real-world data contains less of these extremely creative variations.

	precision	recall
surface matching	0.418	0.991
	(218/521)	(218/220)
surface matching	0.734	0.968
+ filtering	(213/290)	(213/220)

Table 5: The result of experiments on data set (a)

	precision	recall
surface matching	0.147	0.996
	(456/3100)	(458/460)
surface matching	0.528	0.978
+ filtering	(450/853)	(450/460)

Table 6: The result of experiments on data set (b)

Diagnosis

Upon analysing the results, certain patterns of errors and misses were identified.

(1) Errors resulting from insufficiency of lemmatisation by the POS-tagger. For instance, "She horribly damned him with faint praise" is based on the idiom "horribly damn with faint praise". However, our system could not detect this idiom because "damned" was recognised as an adverb rather than the verb "damn". This could be avoided by the improvement of the POS-tagging performance. Another related pattern is errors resulting from the errors of POS-taggers and/or lack of parsing. For instance, the system wrongly output "from high" to the input text "I graduated from high school". This was because, on the one hand, as we do not give structures to input texts, information about the proper construction of "from (high school)" is not provided, while on the other hand the dictionary entry "from high" was wrongly tagged as "from:prep high:adj" instead of "from:prep high:nn". Although theoretically it is preferable to use a high-performance parser, many of these errors can practically be avoided by an improvement in the POS-tagging performance.

(2) Errors resulting from the lack of restrictions on the side of such idiom entries as "make A of B" or "have ... in". The dictionary we used does not give detailed information for the slot "A", "B" and "...". As a result, the system output several irrelevant idioms. This problem can be solved by imposing restrictions on each idiom entry with place holders.

(3) Misses resulting from input text variations in which long phrases are inserted into the idiom constructions. For instance, the dictionary entry "take apart" was not detected for the input text: "She takes (her daughter-in-law) apart with stinging criticism." In order to deal with this, we need to further our understanding of possible idiom variations.

In summary, most of the errors can be avoided (a) if we impose further restrictions on the variation patterns and place holders of idiom entries and (b) if the POS-tagging performance is improved. On the other hand, if we systematically try to deal with the misses, further understanding of the potential range of idiom variations is needed. The experimental results show that the rules we have established cover most of the variations that can be described formally as POS-based patterns. The remaining variations may well be ones that are more context dependent, creative, and/or related to constructions larger than those that can be conveniently described by POS-based patterns. We are currently dealing with misses through experimental use of the system and modifications on the basis of user feedback.

Conclusions

This paper has reported a method for automatically looking up idiom entries in dictionaries vis-à-vis idiom occurrences in texts that may include variations. We started by defining translators' requirements, and then observed the range of idiom variations and formalised the variation patterns of syntagmatic augmentations as POS-based patterns. The result of the experiment showed that the system performance is very promising. The precision for real-world texts is slightly above 0.5, which is in a practically useful range, as users' satisfaction depends more on an improvement of what is currently provided than on "ideal" performance.

As for the technical aspect, we used a morphological analyser but not a parser. The experimental evaluation and the error analyses suggested that most errors and misses can be dealt with without delving into the structural level information given by parsers, although this needs further analysis and examination.

We are currently working in three mutually related directions:

(1) Making the system available for experimental use by translators and obtaining feedback from them, including levels of satisfaction and detailed patterns of errors/misses. We have obtained feedback from two translators and two more translators will take part in the user-based evaluations;

(2) Developing a mechanism that deals with paradigmatic replacements. A basic mechanism has already been developed, and we are currently carrying out evaluation experiments; and

(3) Refining the algorithms for dealing with variations by syntagmatic augmentation.

Acknowledgements

This research is partly supported by grant-in-aid (A) 17200018 "Construction of online multilingual reference tools for aiding translators" by the Japan Society for the Promotion of Sciences (JSPS). We would like to thank Sanseido Publishing Company for allowing us to use the *Grand Concise English-Japanese Dictionary* for our experiments.

References

Abekawa, T. and Kageura, K. (2007). A translation-aid system with a stratified lookup interface. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (to appear).

Aitchison, J. and Lewis, D. M. (Eds.) (2003). *New Media Language*. London: Routledge.

Benson, M. (1985). Collocations and idioms. In Ilson, R. (Ed.) *Dictionaries, Lexicography and Language Learning* (pp. 61–68). Oxford: Pergamon Press.

Biber, D. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Carl, M. and Rascu, E. (2006). A dictionary lookup strategy for translating discontinuous phrases. In *The 11th Annual Conference of the European Association for Machine Translation*.

Čermák, F. (2001). Substance of idioms: Perennial problems, lack of data, or theory? *International Journal of Lexicography* 14(1) (pp. 1–20).

Collins. (2002). *Collins COBUILD Phrasal Verbs Dictionary.* Glasgow: HarperCollins.

Excite. (2005). http://www.excite.co.jp/world/.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language* 6 (pp. 22–42).

Fujitsu. (2005). Atlas Personal Translation.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Mass: MIT Press.

Kageura, K. et. al. (2006). Strategies for enhancing language reference tools to help translators. In *Proceedings of the 12th Annual Meeting of the Japan Society of Natural Language Processing* (pp. 707–710).

Kageura, K. and Toyoshima, M. (2006). Analysis of idiom variations in English for the enhanced automatic look-up of idiom entries in dictionaries. *Proceedings of the 12th Euralex International Congress* (pp. 989–995).

LogoVista. (2005). LogoVista X PRO 2005.

McCaleb, J. G. and Iwasaki, M. (2003). *All-Purpose Dictionary of English Idioms and Expressions*. Tokyo: Asahi Publishing.

Moon, R. (1998). *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

Nicolas, T. (1995). Semantics of idiom modification. In Everaert, M. et. al. (Eds.) *Idioms: Structural and Psychological Perspectives* (pp. 233–252). Hillsdale: Lawrence Erlbaum Associates.

Oxford. (2001). *Oxford Phrasal Verbs: Dictionary for Learners of English*. Oxford: Oxford University Press.

Piao, S. S. L. et. al. (2006). Extracting multiword expressions with a semantic tagger. In *Proceedings of ACL2003 Workshop on Multiword Expressions* (pp. 143–177).

Quirk, R. et. al. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Numberg, G. et. al. (1994). Idioms. In *Language* 70(3) (pp. 491–538).

Sanseido. (2004). *Grand Concise English-Japanese Dictionary*. Tokyo: Sanseido.

Sharoff, S. et. al. (2006). Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the COLING/ACL 2006* (pp. 739–746).

Sharp. (2004). Hon'yaku Kore Ippom.

Similis. http://similis.fr/similis.html.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1) (pp. 143–177).

Toshiba Solutions Corporation. (2005). Translation Professional, Version 10.

Trados. (2006). http://www.trados.com/.

TreeTagger. (2004). http://www.ims.uni-stuttgart.de/projekte/corplex/treetagger/.

Widdows, D. and Dorow, B. (2005). Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of ACL2005 SIGLEX Workshop on Deep Lexical Acquisitions* (pp. 48–56).

Yoshihashi, K. et. al. (2005). Retrieving complex elements from Japanese texts using XPath. In *Proceedings of the 11th Annual Meeting of the Japan Society of Natural Language Processing* (pp. 257–260).