A Simple Platform for Defining Idiom Variation Matching Rules

Koichi Takeuchi[#], Ulrich Apel[§], Ray Miyata[¶], Ryo Murayama[¶],
Ryoko Adachi[¶], Wolfgang Fanderl[§], Iris Vogel[§], Kyo Kageura[¶]
[#]Okayama University, Japan,
[§]Tübingen Eberhard Karls University, Germany,
[¶]University of Tokyo, Japan
koichi@cl.cs.okayama-u.ac.jp, ulrich.apel@uni-tuebingen.de, ray.miyata@gmail.com,
utatanenohibi@gmail.com, littlemarmalade@gmail.com, fanderl@japanologie.uni-tuebingen.de,
iris@fruchtfledermaus.de, kyo@p.u-tokyo.ac.jp

Abstract

In this demonstration, we present a system which enables people who are learning languages to define idiom variation matching rules, with special reference to variations created by insertion. The system is fully operational, currently providing Japanese idiom entries taken from Japanese-English and Japanese-German dictionaries, and is used by both graduate and undergraduate university students who are studying Japanese and Japanese native speakers.
Keywords: Idiom variations; Language learning; Japanese-German dictionary

1 Introduction

Matching idiom occurrences in texts to dictionary entry forms is critical for developing a satisfactory automatic dictionary lookup system. There are important studies of idioms in linguistics (Čermàk, 2001; Fraser, 1970; Moon, 1998; Nicolas, 1995; Numberg et al., 1994) but they do not provide tractable variation rules in different languages. In the field of computational linguistics, some important contributions have been made in idiom variation matching and related issues since the mid 1990s (Breidt et al., 1996; Breidt and Feldweb, 1997; Carl and Rascu, 2006; Michiels, 2000; Proszeky and Kis, 2002; Takeuchi et al., 2007). Nevertheless, most available dictionary lookup systems and MT systems do not incorporate flexible idiom matching functions. Given this situation, we developed a system which enables language learners and practitioners to define flexible idiom matching rules.

2 Idiom variations

Major idiom variations can be categorised into three types (Kageura and Toyoshima, 2006), namely (i) insertion (e.g. “make unholy allowance for” as a variation of “make allowance for”); (ii) change of or-

der (e.g. "the bucket is kicked" as a variation of "kick the bucket"); and (iii) paradigmatic replacement (e.g. "head screwed on wrong" as a variation of "head screwed on right"). We focus on variation by insertion in our platform, because (a) this is the most frequently observed variation, (b) simple change of order can mostly be dealt with straightforwardly and such complex variations as combinations of change of order with insertion are relatively rare, and (c) dealing with paradigmatic replacement is a problem to be solved not by defining syntagmatic rules but by lexical resources such as thesauri. We may extend the target classes to include change of order variations in the future.

Note that while linguists are likely to argue that "you cannot passivise 'kick the bucket' and say 'the bucket is kicked'," these kinds of variations do occur, albeit rarely, in real-world texts, and as such it is important for language practitioners and learners to be able to retrieve the underlying idiom from the variation.

3 System for Defining Idiom Variation Matching Rules

3.1 Access

The system can be accessed at http://edu.ecom.trans-aid.jp.

3.2 System Concepts

The basic policies we adopted are as follows:

(a) We took a restrictive approach rather than a generative approach; we assume that gapped matching of constituent elements is carried out by the base lookup algorithm, and that the rules defined in the platform are to be used to filter out false positives. This has two practical merits. First, the idiom variation rules can be added to dictionary lookup systems as a separate module. Second, if we assume that the matching rules will be used in a translation-aid environment, overmatching (as long as it is not excessive) is less harmful than misses.

(b) We only assume morphological analysers and/or POS-taggers for preprocessing; we do not use parsers. This is because (i) morphological analysers and POS-taggers are available for a wider range of languages than parsers, (ii) we found that there is no difference in performance all in all in a test in English, and (iii) as overmatching is less critical than misses, the merits of using parsers are less important in the application we assume.

(c) We assume that the rules will be defined not only by trained linguists but by ordinary speakers, practitioners and learners of that language. To facilitate this, we restricted the range of variations that can be specified in one rule, by prohibiting the combinations of AND and OR choices. For instance, one cannot write: N (adj|adv|N)+(postp) V in a single rule.

3.3 System Constitution

The system requires dictionary entries consisting of ordinary words. In addition, it requires a separate list of idioms. They should be registered to the system in advance. Currently, a Japanese-English dictionary and a Japanese-German dictionary are registered, through which Japanese idiom variation rules can be defined and validated. The entries are morphologically analysed, and indexes are made for the entry forms as well as the constituent elements of entries. The base lookup module consists of (a) matching individual entries and (b) gapped matching, with up to eight intervening elements, of the constituent elements of idioms. Variation matching rules are defined by users of the system through the Web interface. Currently, we assume that the target idioms for which variation rules are defined are determined by users. To develop variation rules systematically, it would be better for the system to provide users with idioms. This is to be incorporated at the next developmental stage. The rules are used as filters for the gapped matching of idioms. They can be downloaded as a separate file, which can be used as an add-on providing filtering rules for lookup systems.

Interface and Usage

The initial screen consists of a search box into which text (a sentence) that contains an idiom can be input. Figure 1 shows the system output when a user inputs the sentence “このままでは足がすぐ出る。” (kono mama deha ashi ga sugu deru = we will run short of money soon). The system outputs an idiom entry matched to this input, together with word-level matching information. Note that the idiom entry “足が出る” is retrieved through gapped matching.

Figure 1: Initial system output for “このままでは足がすぐ出る”.

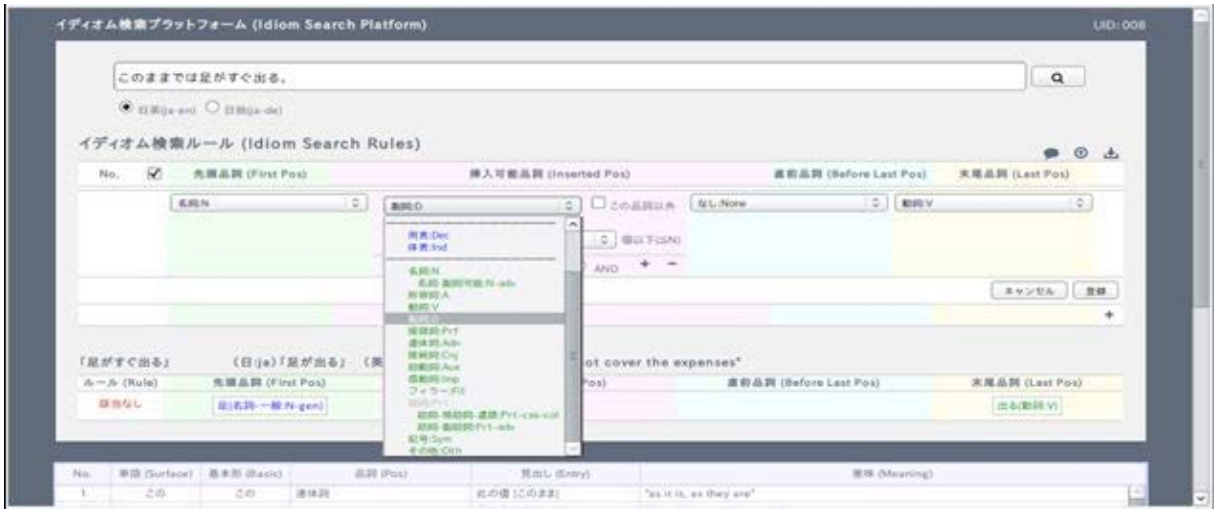

Figure 2: Defining an idiom variation matching rule.

Figure 3: Validating the rule.

This example shows that an adverb can be inserted between “足が” (ashi ga) and “出る” (deru), which enables us to define a rule allowing for the insertion of (an) adverb(s) in this position. The definition of a rule can be done by choosing POS patterns from the pull-down menu for the two constituent elements of the idioms and possible POSes which can be inserted in between (Figure 2). Once the rule is registered, the matching is carried out by applying the rule, which is shown by the rule number given in the result in Figure 3. Rules are defined as POS-based filtering patterns, so rules can be applied to other idioms which were not used for defining the pattern. The rule set can be downloaded as a text file.

4 Prospects

The system is deliberately simple for two reasons: (a) so that non-computationally oriented language practitioners and learners can use it and (b) so that the resultant variation rules can be exported to a variety of dictionary lookup systems. The variation matching rules defined in this platform can be straightforwardly incorporated into the dictionary lookup routine of the translation aid system we maintain (Utiyama, et al. 2009) (and in other systems if only mapping of POS sets is made). Currently, the system assumes that individual users define rules independently. While adding a collective mode is technically not difficult, whether that would be effective in defining rules is not yet clear. We are currently discussing this issue with a limited number of users of the system, while testing the system by providing users with a set of idioms and variation examples, and asking them to individually construct rules. The rules thus created will be unified and adjusted. After this cycle, we will be able to determine whether adopting collective coordination from the start would be more useful or not.

5 References

Breidt, E., Segond, F., & Valetto, G. (1996). Formal description of multi-word lexemes with the finite-state formalism IDAREX. In *Proceedings of the 16th International Conference on Computational Linguistics.* Copenhagen, Denmark, pp. 1036-1040.

Breidt, E., & Feldweg, H. (1997). Accessing foreign languages with COMPASS. In *Machine Translation*, 12, pp. 153-174.

Carl, M., & Rascu, E. (2006). A dictionary lookup strategy for translating discontinuous Phrases. In *Proceedings of the European Association for Machine Translation 2006,* Oslo, Norway, pp. 49-58.

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

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

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

Michiels, A. (2000). New developments in the DEFI matcher. In *International Journal of Lexicography,* 13(3), pp. 151-167.

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

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

Numberg, G., Sag, A, I., & Wasow, T. (1994). Idioms. In *Language,* 70(3), pp. 491-538.

Proszeky, G., & Kis, B. (2002). Context-sensitive electronic dictionaries. In *Proceedings of the 19th International Conference on Computational Linguistics,* Taipei, Taiwan, pp. 1-5.

Takeuchi, K., Kanehila, T., Hilao, K., Abekawa, T., & Kageura, K. (2007). Flexible automatic look-up of English idiom entries in dictionaries. In *Proceedings of the Machine Translation Summit XI,* Copenhagen, Denmark, pp. 451-458.

Utiyama, et al. (2009) Minna no Hon'yaku: A website for hosting, archiving, and promoting translations. In *Translating and the Computer* 31, London.

Acknowledgements

This work is supported by the JSPS-DAAD bilateral collaboration project “Flexible matching of Japanese collocations in a translation environment with Japanese as the source language”.