Multiword Expressions: Some Problems for Japanese NLP

Timothy Baldwin
CSLI, Stanford University
210 Panama St., Stanford, CA 94305, USA
tbaldwin@csli.stanford.edu

Francis Bond
NTT Communication Science Labs
2-4 Hikari-dai, Kyoto 619-0237, JAPAN
bond@cslab.kecl.ntt.co.jp

Abstract

Multiword expressions (MWEs) are notoriously difficult to handle in any language, due to syntactic and semantic idiosyncrasies. In this paper, we focus on Japanese in illustrating the types of difficulties MWEs present for NLP systems, in terms of both analysis and generation. We also outline a number of strategies which can be used to overcome such difficulties.

1 Introduction

The correct treatment of multiword expressions (MWEs) is increasingly being recognized as an important problem, both in linguistics and natural language processing (Sag et al., 2002). In English, Jackendoff (1997: 156) estimates that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words. In many on-line lexical resources almost half of the entries are multiword expressions. For example, in WordNet 1.7 (Fellbaum 1999), 41% of the entries are multiword.

The definition of a multiword expression in English is often given as a "word with spaces". This is not applicable to Japanese, which is typically written without spaces. Instead, we define multiword expressions very roughly as "idiosyncratic interpretations that cross word boundaries". That is, a multiword expression is one that is made up of several units that would normally be segmented into separate words, but whose meaning is not composed purely of the meanings of the individual words.

We divide MWEs into two classes: **lexicalized phrases** and **institutionalized phrases**. Lexicalized phrases have at least partially idiosyncratic syntax or semantics, or contain 'words' which do not occur in isolation; they can be further broken down into **fixed expressions**, **semi-fixed expressions** and **syntactically-flexible expressions**, in roughly decreasing order of lexical rigidity. Institutionalized phrases are syntactically and semantically compositional, but occur with markedly high frequency (in a given context).

In the next section, we give examples of the types of multiword expressions. This is followed by a discussion of some of the current approaches to handling them in Japanese NLP systems.

2 Types of MWEs

Fixed expressions are those that are totally frozen, and appear to act as a single word. An example is 前向き *mae muki* "positive (lit: facing forward)". The interpretation "positive" is not strictly compositional, in addition, there can be no variation in word order or internal modification. Also, in the normal compositional expression *mae-ni muku* "to face forward", there is a postposition, which is absent in the MWE. Another example is ちっとも *chitto mo* "(not) at all". Here, the final postposition is used with its normal meaning, but the first 'word' cannot be used alone. Because Japanese does not explicitly segment words, it is not clear that the above examples are genuinely composed of multiple words. The argument for treating them as MWEs is twofold. First, this analysis captures regularities: *chitto mo* behaves in many ways like a noun followed by *mo*. Second, it makes the task of the segmenter simpler: it can separate into small segments, and the grammar can build them into MWEs as necessary.

Semi-fixed expressions allow some variation. An example of this is the complex postposition: について *ni tsuite* "about (lit: placed in)", which can also have the more formal form: につきまして *ni tsukimashite*. This is similar to the English complex preposition *with regard to*. Although the individual parts can be used in other contexts, they have a fixed meaning in combination and allow no

internal modification, or variation in word order. There are similar expressions which allow slightly more variation, for example に 関 する *ni kansuru* "about (lit: in relation to)" and に 関 し て いる *ni kansuru* "about (lit: in relating to)". The grammaticalization of these words is discussed in detail in Matsumoto (1998).

Another class of semi-fixed expressions is proper names. There are many subregularities, of which we will discuss only one: baseball team names. Japanese baseball team names take the form `place/company_name` + `katakana_loan_word`: for example 阪神 タイガーズ *hanshin taigāzu* "the Hanshin Tigers" or the 広島 カープス *hiroshima kāpusu* "the Hiroshima Carps". Here, although perfectly good native Japanese words exist that mean "tiger" and "carp", the convention is to use a borrowed term.

Other semi-flexible MWEs are **non-decomposable** idioms. Nunberg et al. (1994) introduced the notion of 'semantic compositionality' in relation to idioms, as a means of describing how the overall sense of a given idiom is related to its parts. Idioms such as 腕 を 上げる *ude o ageru* "improve one's skill (lit: raise one's arm)", for example, can be analyzed as being made up of *ude* "arm" in a "skill" sense and *ageru* "raise" in an "improve" sense, resulting in the overall compositional reading of "improve one's skill". With an example such as お 目 に 掛かる *o me ni kakaru* "meet (lit: fix onto honorable eye)", on the other hand, no such analysis is possible: the idiom is non-decomposable.

Because of their opaque semantics, non-decomposable idioms have very limited syntactic variability, e.g. in the form of internal modification or passivization. The main type of lexical variation observable in non-decomposable idioms is inflection.

Japanese also has a wide range of complex predicates, which are semi-flexible. Matsumoto (1996) argues that syntactic V-V compounds (such as 読み たい *yomi tai* "want to read"), lexical V-V compounds (such as 読み 上げる *yomi hajimeru* "read out loud") and complex motion predicates (such as 読み に 行く *yomi ni iku* "go to read") are all two words at the surface level, but have a single interpretation. Syntactic V-V compounds and complex motion predicates are basically compositional, therefore we do not include them as MWEs. Lexical V-V compounds on the other hand, often have very idiomatic interpretations, and thus need to be considered as MWEs.

Syntactically-flexible expressions include **decomposable idioms** such as *ude o ageru* "improve one's skill (lit: raise one's arm)" and light verb constructions such as 電話 を 掛ける/する *denwa o kakeru/suru* "to make/do a phonecall". Both for decomposable idioms and light verbs, the elements can be modified, and there is considerable variation possible in word order. However, the combinations of lexical items which make up the MWE are fixed: *te o ageru* "raise one's hand" does not have the meaning "improve one's skill", and *denwa* "telephone" can only be used with *suru* and *kakeru*, not with other light verbs such as *toru* "take".

Institutionalized phrases include compounds such as 機械 翻訳 *kikai hon'yaku* "machine translation" and phrases like 悪口 を 言う *waruguchi o iu* "badmouth someone (lit: say bad things)". These are semantically and syntactically compositional, but statistically idiosyncratic. There is no particular reason why one could not say *#konpyūtā honyaku* "computer translation", or *#waruguchi o shaberu* "talk bad things" but people don't. The idiosyncrasy is statistical rather than linguistic, in that it is observed with much higher relative frequency than any alternative lexicalization of the same concept. We refer to potential lexical variants of a given institutionalized phrase which are observed with zero or markedly low frequency as **anti-collocations**.

3 Current NLP Approaches

The approach to multiword expressions depends on the nature of the NLP task. For tasks that use minimal semantics, such as bag-of-word based indexing for document retrieval, a lot can be done with just simple segmentation. For machine translation, where the interpretation of meaning is the key problem, treatment of MWEs needs to be quite sophisticated.

For almost all NLP systems, fixed expressions are entered whole into the lexicon. Capturing the regularities has no effect on the performance of the system, and does not affect the maintainability to any great extent.

Overall, there are two basic strategies for segmenter/taggers, such as Chasen (Matsumoto

et al., 1999) or JUMAN. One is to split the input into the smallest meaningful units and leave any aggregation to the next stage. In this case something like *ni kan suru* would end up as just that: *ni kan suru*, possibly with *kan* and *suru* grouped together. The other strategy is to try and get the largest meaningful unit: in this case *ni kan suru* would be a single lexical entry: *nikansuru*. The problem with the first approach is that the unified nature of the expression is lost. The problem with the second is that all possible forms of the MWE have to be entered into the lexicon, and possible compositional uses will not be identified.

Shirai et al. (1993) augment the first approach by doing an initial dependency parse after segmentation, and then using rewrite rules (equivalent to grammatical constructions) to chunk together the units. The advantage of this method is that the rewrite rules can use information about syntactic dependency, not just lexical information. For *ni kan suru*, if the postposition *ni* is a dependent of the verbal noun plus verb group *kan suru*, the whole element is rewritten into a pseudo-postposition *nikansuru*. Eliminating the verb phrase *kan suru* early reduces spurious ambiguity in the parse, as well as removing potential zero anaphora.

Typically, many proper names are entered fully into the lexicon. However, as there are a vast number of proper names it is common to also have specialized rules for creating new proper names: for example `proper_name` + `school_type` ⇒ `school_name`: for example 奈良 *Nara* + 大学 *daigaku* "university" ⇒ 奈良大学 *Nara daigaku* "University of Nara". The **ALT-J/E** Japanese-to-English machine translation system (Ikehara et al., 1991), has 130 proper name classes designed specifically to deal with the analysis of names.

Most work on processing idioms, both decomposable and non-decomposable, has been done for machine translation systems. The emphasis has been on analysis, because it is common to translate from a more complex representation to a simpler one, therefore the analysis processing tends to be more sophisticated than the generation. If analysis is the main concern, the question of whether or not modification and word order variation is possible is less urgent, or at the very least, the problem becomes one of disambiguation between idiomatic and non-idiomatic expressions based on selectional restrictions or similar.

ALT-J/E deals with predicate-based idioms by allowing predicates to have fixed fillers, as shown in Figure 1. In the first example the fixed filler (corresponding to N1) is actually a leaf node in the semantic ontology. This allows for some variation in the form of the idiom. It will in fact also match with patterns that do not normally occur (for example the very formal 輿望 *yobō* "popularity"), however this is not a problem for analysis. In the second example a Japanese idiom is mapped onto an English idiom. Because idiomatic MWEs can potentially also have straight compositional interpretations, it is important to ensure that the semantic restrictions used to select the patterns are sufficiently constrained.

Light verb constructions have proved to be problematic. Creating separate patterns for all or most light verb constructions has been suggested (Matsuo et al., 1997), but it has proved difficult in practice as there are tens of thousands of possible light verb combinations. The same is true of V-V lexical constructions and institutionalized phrases such as noun-noun compounds: it is in theory possible to list them, but in practice there are too many.

One approach to solving this problem, which is gaining in popularity due to the increasing availability of large corpora, is to attempt to automatically extract such expressions (and possibly their translations) from corpora. Initial work focussed on single word units extracted from aligned corpora (e.g., Fung (1995)), but recent work has started to also investigate the use of non-parallel corpora (Tanaka and Iwasaki, 1996) and the extraction of multiword expressions (Tanaka and Matsuo, 1999).

4 Conclusion

We have shown that MWEs are both diverse and interesting. Like the issue of disambiguation, MWEs constitute a key problem that must be resolved in order for linguistically precise NLP to succeed. In this paper we mainly illustrated the diversity of the problem, but we have also examined known NLP techniques for dealing with these problems. Although these techniques take us further than one might think, there is much descriptive and analytic work on MWEs that has yet to be done.

It is our hope that increasingly sophisticated

```
Pattern ID: -500889-00-
+ N1 (popularity) (ga) (OBL)
+ N3 (*) (ni)
+ atsumaru ''gather''

Pattern ID: -201257-00-
+ N1 (agent) (ga)
+ chie (o)
+ N3 (abstract) (ni)
+ shiboru ''squeeze''
```

```
U-SENT (active)
+ NP:Subj N3
+ PRED - VERB become
+ ADJ-P popular

U-SENT (active no-passive)
+ NP:Subj N1
+ PRED - VERB rack
+ NP:Obj N1's brains
+ PP: over N3
```

Figure 1: Predicate based Idiom Frames

linguistic analyses: hierarchically organized lexicons, restricted combinatoric rules, lexical selection, idiomatic constructions, circumscribed constructions and simple statistical affinity can begin to make the problems more understandable, and lead to more principled solutions.

Acknowledgements

This research was supported in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University.

References

Christine Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

Pascale Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *33th Annual Conference of the Association for Computational Linguistics*, pages 236–243, 1995.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Third Machine Translation Summit: MT Summit III*, pages 101–106, Washington DC, 1991. (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

Ray Jackendoff. *The Architecture of the Language Faculty.* MIT Press, Cambridge, MA, 1997.

Yō Matsumoto. *Complex Predicates in Japanese.* Kuroshio Shuppan & CSLI, Tokyo & Stanford, 1996.

Yō Matsumoto. Semantic change in the grammaticalization of verbs into postpositions in Japanese. In Toshio Ohori, editor, *Studies in Japanese Grammaticalization.* Kuroshio Shuppan & CSLI, Tokyo, 1998.

Yuuji Matsumoto, Akira Kitauchi, Tatsu Yamashita, and Yoshitake Hirano. *Japanese Morphological Analysis System ChaSen Version 2.0 Manual.* Technical Report NAIST-IS-TR99009, NAIST, 1999.

Yoshihiro Matsuo, Satoshi Shirai, Akio Yokoo, and Satoru Ikehara. Direct parse tree translation in cooperation with the transfer method. In Daniel Joneas and Harold Somers, editors, *New Methods in Language Processing*, pages 229–238. UCL Press, London, 1997.

Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. Idioms. *Language*, 70:491–538, 1994.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002.* Springer-Verlag, Heidelberg/Berlin, 2002.

Satoshi Shirai, Satoru Ikehara, and Tsukasa Kawaoka. Effects of automatic rewriting of source language within a Japanese to English MT system. In *Fifth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-93*, pages 226–239, Kyoto, 1993.

Kumiko Tanaka and Hideya Iwasaki. Extraction of lexical translations from non-aligned corpora. In *16th International Conference on Computational Linguistics: COLING-96*, pages 580–585, Copenhagen, 1996.

Takaaki Tanaka and Yoshihiro Matsuo. Extraction of translation equivalents from non-parallel corpora. In *Eighth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99*, pages 109–119, Chester, UK, 1999.