Current State of JMWEL: a Comprehensive Japanese MWE Lexicon and its Applications

Masahito Takahashi[1], Toshifumi Tanabe[2], Kosho Shudo [3], and Jack Halpern[4]

[1]Kurume Institute of Technology, JAPAN
[2]Fukuoka University, JAPAN
[3]Fukuoka University, emeritus, JAPAN
[4]CJK Dictionary Institute, Inc., JAPAN
[1]taka@kurume-it.ac.jp
[2]tanabe@fukuoka-u.ac.jp
[3]viggo_ksf@jcom.home.ne.jp
[4]jack@cjki.org

Abstract.

JMWEL is a comprehensive lexicon of Japanese Multiword Expressions (MWEs) with a rich set of grammatical attributes fine-tuned for phrase-based NLP applications such as machine translation and information retrieval, as well as morpho-syntactic analysis of a wide-range of Japanese documents. It currently contains about 150,000 lemmas covering almost every kind of linguistically idiosyncratic, but commonly used Japanese phrases, e.g., idioms, quasi-idioms, collocations, quasi-collocations, clichés, quasi-clichés, proverbs, and old sayings.

JMWEL consists of eight basic sub-lexicons reflecting their distinctive grammatical functions: sub-lexicon of nominal MWEs; sub-lexicon of verbal MWEs; sub-lexicon of adjectival MWEs; sub-lexicon of adjective-verbal MWEs; sub-lexicon of adverbial MWEs; sub-lexicon of adnominal MWEs; sub-lexicon of connective MWEs; and sub-lexicon of functional MWEs. Cross-sectionally, JMWEL also has topic-based sub-lexicons: sub-lexicons of standard-idioms, onomatopoeic collocations, proverbs/sayings/clichés, syntactically ill-formed-phrases, four-kanji-idioms, and cranberry expressions.

The fields for basic information which are common to all sub-lexicons of JMWEL are:

i. lemma written in hiragana without space-marker;
ii. segmentation by space-markers into a string of morphemes;
iii. most kanji or katakana notations for each morpheme;
iv. syntactic function as a whole;
v. morpho-syntactic structure with markers for possible internal modification;
vi. left context condition;
vii. right context condition.

Extensiveness of collected MWEs and information written in iii., v., vi., and vii. are notable features of JMWEL. In particular, information in v. enables JMWEL to cope with a wide-range of the syntactic rigidity of MWE.

The most fundamental application of JMWEL is the annotation of MWEs on the morpho-syntactic composition of sentences in corpora. Fixing each MWE as a chunk

or a loose chunk with internal modifiers in morpho-syntactic analysis is an approach that simulates the process of human speech understanding. The framework of morpho-syntactic analysis accompanied by MWE-annotation will provide a firm basis for a variety of forthcoming NLP applications which function on the basis of the correct meanings of linguistically idiosyncratic word-strings.

Keywords: MWE, MWU, Phrase-based NLP, RBMT, PBMT, SMT, PBSMT, NMT, Lexical Bundles, Formulaic Language, Construction Grammar

References

1. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. : Multiword expressions: a pain in the neck for NLP. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, pp. 1–15. Mexico City (2002).
2. Shudo, K., Tanabe, T., Yoshimura, K.: MWEs as non-propositional content indicators. In: Proceedings of the ACL, Workshop on Multiword Expressions: Integrating Processing, pp. 31-39. Barcelona (2004).
3. Tanabe, T., Takahashi, M., Shudo, K.: A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. In: Computer Speech and Language, vol. 28, pp. 1317-1339. Elsevier (2014).
4. http://jefi.info