Semi-automatic Refinement of
the JMdict/EDICT Japanese-English Dictionary

Francis Bond* and Jim Breen**
* NICT Computational Linguistics Group
** Monash University
<bond@ieee.org, jim.breen@infotech.monash.edu.au>

Abstract

The JMdict/EDICT Japanese-English Dictionary is a freely-available dictionary distributed in XML (JMdict) and text (EDICT) formats. It is widely used as a source of lexical material in dictionary systems and text-processing projects. We propose two refinements to make the dictionary more computationally tractable: marking entries where the English is not a translation equivalent and expanding contracted entries. We then propose and apply semi-automatic methods to refine existing entries. The resulting dictionary is shown to be more suitable for the construction of machine translation rules.

1 Introduction

Resources built for one task can often be useful in others. WordNet, for example, started off as an test-bed for a particular model of lexical organization (Fellbaum, 1998, p4) and is now widely used in natural language processing (NLP) applications. In this paper we look at the Japanese/English lexicon JMdict/EDICT (Breen, 2004), which started out as a voluntary project to produce a freely available Japanese/English Dictionary in machine-readable form. In addition to being useful for people as a bilingual dictionary, it is also widely used in NLP applications. For example, it has been the base to make compound noun lexicons (Tanaka and Matsuo, 1999; Ohmori and Higashida, 1999), new bilingual lexicons (Paik et al., 2001; Apel, 2002; Sjöbergh, 2005; Zhang et al., 2005; Fujita and Bond, 2006; Bond and Ogura, 2007) and machine translation transfer rules Bond et al. (2005).

We look at two ways of making the dictionary even more tractable for NLP tasks: (1) marking entries where the English is not a translation equivalent and (2) expanding contracted entries. Finally, we discuss some planned future enhancements.

2 JMdict/EDICT

The JMdict/EDICT project now has approximately 110,000 Japanese/English entries recorded, with the number increasing at about 1,000 per month. A WWW-based system for submitting amendment and new entry suggestions is yielding about 100 submissions each day, which close to the limit that can be handled by the sole editor (Breen).

The project dictionary is distributed in three formats:

1. the full JMdict (XML) format, both in Japanese-English and Japanese-English/German/French/etc. versions.

2. the original EDICT format, which only allows for one kanji word and one reading per entry. Thus JMdict entries which have alternative kanji or okurigana forms, or which have alternative readings will result in multiple entries in the EDICT file.

3. the EDICT2 format (shown below) which allows for multiple kanji words and readings in an entry, and is in effect a human-readable equivalent of the JMdict entries.
The dictionary files are generated daily, and are available via ftp and rsync, thus allowing WWW servers that use the files to stay up-to-date.\(^1\)

At present all editing is taking place at Monash University, with semi-automated creation of new entries, and manual amendment of existing entries. A new WWW-based maintenance system in nearing completion which will enable distributed editing with a pool of editors. The new system has a more flexible database which will allow additional information to be included in the entries, and greater access to the data by project members.

3 Enhancements to the Dictionary Structure

A typical entry (for jiten “dictionary”) in the EDICT2 format is shown below:

(1) 辞典(P);辞典(oK) /[じてん] /n dictionary/(P)/

The various marks indicate that it is a common word (P), there is an orthographic variant with old Kanji (oK), and that the Japanese part of speech is a noun (n). In this entry, as in most entries, the English gloss is a translation equivalent and the entry is effectively reversible: 辞典 → dictionary). This allows the use of JMdict as an English → Japanese lexicon, even though it basically Japanese → English.

However there are some entries where the reversibility does not hold. For example consider the simplified entry for ten “piece” (2):

(2) てん[てん] (1) /(n,n-suf) spot/mark/
(2) point/dot/
(3) (n-suf) counter for goods or items/(P)/

In this case the third gloss is not a translation equivalent, but rather an explanation: ten “piece” is used as a suffix for numbers when counting goods or items. We would not expect to want to look up this directly, and could not directly use the gloss to create translation rules.

Another example where the reverse look up fails is disjunctive entries such as (2):

In this case, two translations have been collapsed into the second gloss: rice field and rice paddy. This contraction of entries is important in paper dictionaries, where space is precious, but causes problems for electronic access: the translation equivalent rice field will not be an exact match and the translation rice paddy is not even a contiguous string.

The solution to the first problem is to explicitly mark the type of each gloss. The default type is equ (translation equivalent) whereas explanations are marked as exp. Simplified examples of this marked up in xml are shown in (4):

(4) てん[てん] ...
<gloss g_type="equ">spot</gloss>
<gloss g_type="exp">counter for goods or items</gloss>

The solution to the second is even simpler: split the entry with “or” into two separate entries:

(5) てん[てん]; てんし] /(n) farmland/ rice field/ rice paddy/

4 Expanding Disjunctive Entries

There were 2843 entries containing ” or ” in JMdict (1.3% of the Japanese-English entries contained a disjunctive gloss). Four word entries were the most common, with the longest entry consisting of 35 words.

An initial survey of the glosses found three major types (G) good translation equivalents (D) disjunctive glosses and (E) explanations.

\(^1\)http://ftp.monash.edu.au/pub/nihongo/00INDEX.html
The vast majority were of short entries (5 words or less) were of type D, while the longer entries were mainly of type E.

The algorithm for rewriting was simple:

1. Remove any articles from the gloss
2. if the final word is *two*, *other* or *another* ⇒ G
3. elsif the Japanese entry ends in *か* ⇒ G
4. elsif the gloss appear more than 3 times (*e.g.* *yes* or *no*) ⇒ G
5. elsif there 6 or more words ⇒ E
6. elsif *or* is the second word (*w_2*) D split into *w_1, w_4, w_5, \ldots* and *w_3, w_4, w_5, \ldots*
7. elsif *or* is the second last word (*w_{n-2}*), D split into *\ldots w_{n-4}, w_{n-3}, w_{n-2} and \ldots w_{n-4}, w_{n-3}, w_n*
8. else leave to be hand-checked

As a result of this 72 entries were judged to be good (G), 1,500 to be disjunctive (D) and the remaining 1,271 to be explanations (E) or requiring further checking.

In the examples given above, the disjunctive entries are rewritten as follows:

(6) 国際語[*こくさいご*] / (n) international language/universal language/
(7) 蕎落[*はくろう*] / (n, vs) poor reception/ inhospitable reception/

5 Evaluation

The effectiveness of splitting was tested by manual evaluation. For those judged G, there were only three errors, all caused by errors in the original entries.

For those split (D), only 2% were erroneous. The main source of errors was splitting good entries, such as (8).

(8) 虚実[*きおうち*] / (n) truth or falsehood/
4. Extension of cross-referencing including indication of synonyms and antonyms. At present there is a relatively low level of cross-referencing. Some experimentation using bag-of-words techniques with the English glosses has shown that this may be a fruitful approach for identifying synonyms.

5. Marking of domains. At present there is limited domain marking in a number of entries, and it is highly desirable that this be extended. An issue is determining an appropriate set of word domains to use. A possibility being explored is the application of Wordnet synsets.

6. Adding verb translations for verbal nouns (サ変名詞) verbs. At present EDICT generally records only the noun translation: for example: 検査[けんさ] / (n, vs) inspection/. We would like to expand the entry to: 検査[けんさ] / (n) inspection/ (vs) inspect. This can be done semi-automatically and hand checked, as we did for the disjunctive entries. In this case the verbal form can be deduced from the nominal one using Nomlex (Macleod et al., 1998) or WordNet 2.0. However, there are over 10,000 verbal nouns, so semi-automatic checking becomes even more important.

7 Conclusion

In this paper we showed a semi-automatic approach (automatic generation followed by manual checking) to improve disjunctive entries in the JMdict/EDICT lexicon, and outlined some of the future plans. Because the lexicon is freely available, any improvements will be multiplied by the number of projects that use the lexicon, making even small improvements valuable.

References


