A Japanese-English Technical Lexicon for Translation and Language Research

Fredric Gey[1], David Kirk Evans[2], Noriko Kando[2]
[1]University of California, Berkeley, CA, USA
[2]National Institute of Informatics, Tokyo, Japan
gey@berkeley.edu, devans@nii.ac.jp, kando@nii.ac.jp

Abstract

In this paper we present a Japanese-English Bilingual lexicon of technical terms. The lexicon was derived from the first and second NTCIR evaluation collections for research into cross-language information retrieval for Asian languages. While it can be utilized for translation between Japanese and English, the lexicon is also suitable for language research and language engineering. Since it is collection-derived, it contains instances of word variants and miss-spellings which make it eminently suitable for further research. For a subset of the lexicon we make available the collection statistics. In addition we make available a Katakana subset suitable for transliteration research.

1. NTCIR Cross-Language Retrieval

NTCIR[1] is a large evaluation initiative for Asian Language Search and Question Answering, currently in its Seventh evaluation. NTCIR is similar in scope to the TREC[2] series of evaluations for English and to CLEF, the Cross Language Evaluation Forum[3] a large European evaluation initiative dedicated to cross-language retrieval for European languages (Peters et al., 2007). NTCIR was developed to meet need for cross- and multi-lingual retrieval research specifically for East Asian languages (Chinese, Japanese and Korean). The first and second NTCIR Workshops utilized a collection of abstracts from the journal proceedings of 66 Japanese technical societies. As such, the NTCIR-1 and NTCIR-2 collections are the only evaluation resources available to test automatic retrieval of scientific and technical documents in Japanese. Further details about NTCIR-1 may be found in (Kando et al, 1999). Later NTCIR workshops utilized news collections from newspapers and newswire services and expanded the language scope to Japanese, Chinese and Korean. In this paper we are concerned with aspects of deriving a lexicon of technical terminology which can be utilized for both translation and language engineering for further research into finding technical content between the English and Japanese languages.

2. NTCIR Test Collections

Our lexicon is derived from the NTCIR-1 and NTCIR-2 workshop test collections. The collections consist of three disjoint sub-collections:

- NTCIR-1 J-E gakkai collection (339,483 documents) -- Author abstracts of articles from 65 Japanese scientific society hosted conferences for the period 1988-1992 (English and Japanese abstracts pre-joined, where English abstracts available).
- NTCIR-2 J-E gakkai collection – extension of the NTCIR-1 collection for years 1997-1999. 77,433 English abstracts, 116,177 Japanese abstracts, as independent files (not pre-joined)
- NTCIR-2 J-E kaken collection – abstracts of funded research final reports 1988-1997. 57,545 English abstracts, 287,071 Japanese abstracts, as independent files (not pre-joined)

2.1 NTCIR-1 J-E Collection

The NTCIR-1 J-E collection consists of 339,483 documents, of which 98.5% (334,515 documents have Japanese abstracts) and only 188,907 (55.6%) have equivalent English abstracts. The salient characteristic, however, of the collection is that 313,673 (92.3%) of the documents have author-assigned keywords in both Japanese and English. The following is an example of keywords assigned:

<KYWD TYPE="kanji">**画像センサ // コンピュテーショナルセンサ // 画像圧縮 // 画像符号化**</KYWD>
<KYWE TYPE="alpha">**Image Sensors // Computational Sensors // Image Compression // Image Coding**</KYWE>

Because only slightly more than half the documents have English abstracts, pairing keywords may be more useful than the more complicated task of pairing sentences in documents (the usual approach of statistical machine translation) to align term pairs.

2.2 NTCIR-2 J-E Gakkai Collection

The NTCIR-2 J-E Gakkai collection was basically an extension of the NTCIR-1 collection for the additional years 1997-1999. Because the collection only covered two years, it was a smaller collection than NTCIR-1, consisting of slightly more than 116,000 documents, of which only 77,000 documents (66.6%) had English abstracts and/or English keywords. Of these, 71,839 documents had both English and

[1] http://research.nii.ac.jp/ntcir/

[2] http://trec.nist.gov

[3] http://www.clef-campaign.org

Japanese keywords assigned by the authors. In order to extract a lexicon, the two independent files, Japanese abstracts and English abstracts had to be joined on a common document identification number (the NTCIR-1 collection had pre-joined abstracts from the two languages).

2.3 NTCIR-2 J-E Kaken collection

The NTCIR-2 J-E Kaken collection consists of abstracts of final reports for academic research funded by the Japanese government between the years 1988-1997. The two independent files were 287,071 Japanese abstracts and 57,545 English abstracts, which were again joined to create a bilingual abstract subset of 57,512 records with both Japanese keywords and English keywords. The Kakan collection exhibited considerably more diversity in subject matter as well as less direct correspondence between English and Japanese assigned keywords. Below are two examples of keyword assignments for this collection:

kaken-j-0965522600 |KYWE| **environmental issues | mass media | public opinion | social research | content analysis | effects of mass communication | global warming | social psychology** |KYWD| **環境問題 | マスメディア | 世論 | 社会調査 | 内容分析 | マスコミ効果論 | 地球温暖化 | 社会心理学**

kaken-j-0861763900 | KYWE| **Methylglyoxal | D-Lactate | HPLC | 2-Methylquinoxaline | o-Phenylenediamine | 4,5-Dichloro-o-phenylenediamine**|KYWD|**メチルグリオキサール | D-乳酸 | HPLC | 2-メチルキノキサリン | オルトフェニレンジアミン | 4.5-ジクロロオルトフェニレンジアミン | 6.7-ジクロロメチルキノキサリン**

3. Extracting the lexicon

For the NTCIR-1 and NTCIR-2 Gakkai collections of abstracts of conference papers by the 65 Japanese scientific societies, we observed that keyword sequences seemed to be ordered in both Japanese and English. This was confirmed by translating a number of the Japanese keywords using the GOOGLE Translate language tool for Japanese to English. Thus if we had a keyword sequences:
J_1 // J_2 // J_3 // J_4 // J_5 and
E_1 // E_2 // E_3 // E_4 // E_5

In almost all cases we would find that $J_i \equiv E_i$. For some records, we found that the count of Japanese keywords differed from the count of English keywords. This occurred in 5.2% (16,289 records) of NTCIR-1 documents with both English and Japanese keywords present and in 3.7% of NTCIR-2 gakkai documents with both language keywords present. For records with |KYWJ| = |KYWE| we simply extracted the corresponding keyword pairs and counted their number of occurrences in the collection. For records where |KYWJ| ≠ |KYWE| we chose to only process min(|KYWJ|, |KYWE|) in sequence.

3.1 NTCIR-1 Lexicon Creation Results

For the NTCIR-1 Gakkai collection of 313,673 records with both language keywords present, our keyword pairing strategy resulted in 598,439 unique Japanese-English term pairs with the following distribution shown in table 1:

Number of Occurrences	J-E Pair count
5 or more	34,044
4	11,698
3	23,063
2	64,726
1	464,908

Table 1. NTCIR-1 J-E Keyword Pair distribution

The lengthy tail of the distribution may include many erroneous pairs, including misspellings. The following table is a fragment (with occurrence count) from the lexicon for the Japanese term for “Information Retrieval”:

Frequency	Japanese	English
495	情報検索	information retrieval
8	情報検索	information retrival
4	情報検索	information retieval
3	情報検索	information retreival
3	情報検索	information retrilval
2	情報検索	information retrieving
1	情報検索	information retrilval

Table 2. Matches for “Information Retrieval”

We include these occurrences within the lexicon because they may be useful in studying cross-lingual search in the event of spelling errors.

3.2 NTCIR-2 Gakkai Creation Results

For the NTCIR-2 Gakkai collection of 71,839 records with both language keywords present, our keyword pairing strategy resulted in 172,400 unique Japanese-English term pairs with the following distribution shown in table 3:

Number of Occurrences	J-E Pair count
5 or more	8,032
4	3,210
3	6,644
2	19,380
1	135,134

Table 3. NTCIR-2 Gakkai Pair distribution

The following are the top 6 terms of the distribution:

528 シミュレーション | simulation

493 有限要素法 | finite element method

470 液状化 | liquefaction

466 インターネット | internet

412 遺伝的アルゴリズム | genetic algorithm

383 ニューラルネットワーク | neural network

3.3 NTCIR-2 Kaken Lexicon Creation

The Kaken collection proved to be qualitatively and quantitatively different from the other two sub-collections. As mentioned above there is considerably more diversity in subject of the documents in the collection; subjects are not grounded by 'domain' of a particular technical society as with the Gakkai collections. In addition their statistical characteristics differ, particularly with respect to equality of count of number of Japanese keywords matched to English keywords. Using the paired keyword approach of the Gakkai collections above, we generate 238,820 unique J-E pairs with the following distribution:

Number of Occurrences	J-E Pair count
5 or more	5,685
4	2,353
3	4,549
2	14,001
1	212,232

Table 4. NTCIR-2 Kaken Pair distribution

The following are the top 10 terms of the Kaken subcollection, together with collection counts:

282 ラット | rat

266 モノクローナル抗体 | monoclonal antibody

251 アポトーシス | apoptosis

243 サイトカイン | cytokine

236 遺伝子発現 | gene expression

233 免疫組織化学 | immunohistochemistry

188 シミュレーション | simulation

181 データベース | database

163 カルシウム | calcium

161 マウス | mouse

However, fully 27.1% (15,530 of 57,354 documents with both English and Japanese assigned keywords present) differed in count of keywords by language. This called for examination of keyword sequencing. We found by manual examination that the keywords which were translations often did not occur in the same linear sequential order. The most reasonable approach to generating this lexicon has been to take the maximalist approach of matching each Japanese keyword with all English keywords assigned to that document. Thus if a document has 5 English and 7 Japanese keywords:

E1 E2 E3 E4 E5
J1 J2 J3 J4 J5 J6 J7

we construct 35 keyword pairs:

(E1,J1)(E1,J2) … (E1,J7)
(E2,J1)(E2,J2) …

and proceed to accumulate collection statistics for each unique keyword pair according to the following 2-way contingency table:

Count	J_k	$\sim J_k$
E_i	**a**	**b**
$\sim E_i$	**c**	**d**

The lexicon is distributed in the following formatted list of term pairs with counts: J_k E_i **a b c d**
In this way the users of the Kakan lexicon can experiment with different measures of association such as Yates Chi Square (Yates, 1934) or Dunning's log likelihood ratio (Dunning, 1994) to choose the most likely equivalence. This maximalist approach has generated 2,219,878 J-E pairs, while the ordered sequence approach generates 238,820 pairs.

4. Lexicon Validation

Lexicon validation is a complex topic. Rather than doing a manual examination by professional Japanese translators, we chose to see what could be obtained from matching against other available dictionaries. The second author of the paper had assembled a selection of 21 freely available Japanese English dictionaries. The dictionaries had a total of 1,033,244 entries. Some of these are specialized lexicons that cover life sciences, computer terminology, and so on. A program was written to look up each Japanese term in all the dictionaries, and records when it finds a match. A match is an exact string match on the Japanese term. If the Japanese term is hiragana only, matches are also performed over the kana readings of dictionary terms. The summary below gives some information on how many terms were found and the weights of those terms. Nothing was done to automatically validate whether the matched translation from the lookup is valid (using, for example, a check whether the English is a complete substring of one of the translations – although that information is presented in Table 5) Such a check would be possible, but of course there are many translations that would be judged as incorrect despite being valid, or incorrectly marked valid when the English is only a small substring of a more complex translation. In addition to counting raw hits of an English term in the dictionaries, a term weighting by frequency scheme was also utilized to compensate for the frequency counts of the lexicon term. For the three lexicons, we had, respectively:

NTCIR-1:
unique terms: 598,424, total weights of terms: 1,314,161
number of dictionary hits:121, 032 (0.20225124)
number of misses: 477,392 (0.79774874)
weighted hits: 502,547 (0.382409)
weighted misses: 811,614 (0.617591)

NTCIR-2 gakkai:
unique terms: 172,400 weights of terms: 312,922
number of dictionary hits: 41,311 (0.23962297)
number of misses: 131,089 (0.76037705)
weighted hits: 119,763 (0.38272476)
weighted misses: 193,159 (0.61727524)

NTCIR-2 kaken:
Unique terms: 238,819 weights of terms: 331,900
number of hits: 59,370 (0.2485983)
number of misses: 179,449 (0.75140166)
weighted hits: 121,490 (0.36604398)
weighted misses: 210,410 (0.633956)

We also created three sample files of approximately 1000 J-E term pairs. The sample was stratified by frequency in order to obtain a larger sample for low-frequency term pairs. Thus we selected 100 pairs for frequency count >4, 100 pairs for frequencies = 3 or 4, 200 pairs for frequency 2 and 500 pairs for frequency 1.

For comparison, the same dictionary match was run for the NTCIR-1 sample with the following results:

NTCIR-1 sample:
uniqueTerms: 1000 weights of terms: 3212
number of hits: 229 (0.229)
number of misses: 771 (0.771)
weighted hits: 1062 (0.33063513)
weighted misses: 2150 (0.66936487)

If you compare this with the complete collection match statistics above, we find that the sample overestimates the raw number of hits and underestimates the number of weighted hits for this collection.

The CJK Dictionary Institute[4] volunteered to automatically match the sample terms against their technical dictionary. Their results were

163 (Japanese and English found in cjkiterm)
227 (Japanese found in cjkiterm with different English)
208 (English found in cjkiterm with different Japanese)
478 (Japanese, English not found)

We will be reviewing these results with CJKI.

Table 5 shows the number of Japanese terms that were found in the dictionary lookup and the number of instances where the English term was found as a substring of the translation. As expected, the percentage of English terms in the lexicon that are found as substrings in the dictionaries increases with the frequency of the observed JA-EN term pair. A matched English substring in the dictionary translation is a likely indication of a good translation.

Number of Occurrences	Term Pairs	Matched Japanese Terms (pct)	English Substring Matches (pct)
10 or more	**14,146**	**52.1**	**39.3**
9	**1,964**	**39.3**	**24.3**
8	**2,576**	**38.2**	**24.2**
7	**3,379**	**37.1**	**22.3**
6	**4,775**	**34.2**	**19.4**
5	**7,204**	**30.8**	**16.1**
4	**11,698**	**28.7**	**14.1**
3	**23,063**	**26.3**	**11.1**
2	**64,725**	**21.9**	**7.1**
1	**464,894**	**17.9**	**2.5**

Table 5. NTCIR-1 J-E Dictionary Matches by Frequency

[4] http://www.cjk.org/

Another approach which we have not yet explored is to use the GOOGLE language API to translate the Japanese sample terms and compare to the lexicon English derivation for that term. We know anecdotally that many high-frequency pairs in the lexicon have the same English as the Google translation of the Japanese. Low frequency pairs have a different pattern. For example, for one lexicon pair:

1 判別木 | classification tree

The Japanese was translated by Google as "taxonomic wood."

The problem with our approach to the lexicon validation is illustrated by Table 2 above, where the multiple pairs in English for the Japanese 情報検索 (information retrieval) are spelling variants. How do you validate a lexicon for these low-frequency mis-spellings?

5. Lexicon Research Uses

Besides providing another public-use technical lexicon for use as a translation aid, the characteristics of this lexicon can be exploited for other research purposes.

5.1 Spelling Correction for Lexicons

As Table 2 above shows, Japanese researchers have sometimes made spelling mistakes when assigning English keyword equivalents to Japanese terms. It would be an interesting research question as to the degree to which application of spell correction algorithms could improve the accuracy of the lexicon by increasing frequency counts between Japanese terms and the class of English terms represented by the corrected spelling.

5.2 Transliteration of Katakana

Consider again the top 6 terms of the NTCIR-2 gakkai sub-collection distribution – three of these terms:

528 シミュレーション | simulation

466 インターネット | internet

383 ニューラルネットワーク | neural network

are in the Katakana alphabet. The Japanese Katakana alphabet is a phonetic alphabet usually used for the introduction of words borrowed from other languages. Use of phonetic models for transliteration and matching between Japanese Katakana and English was studied by Knight (Knight & Graehl 1997). We will make available a subset of the lexicon restricted to those Japanese terms which are only written with the Katakana alphabet. The Katakana subset will include a Hepburn Romanization[5] of all Japanese terms, according to the Hepburn module in the PERL CPAN archive.[6] Thus researchers will be able to use the lexicon for discovery of improved phonetic transliteration algorithms, as well as to compare machine learning transliteration to the accuracy of rule-based Japanese Romanization algorithms.

6. Lexicon Availability

We will make the lexicon files described in the paper under a GNU style public licensing. The lexicon and its description will be found at the following url:

http://metadata.berkeley.edu/japanese

Files will be in gzip format.

7. Acknowledgment

Most of this work was performed while the first author was a visiting researcher at the National Institute of Informatics, Tokyo, Japan, during the summer of 2007, under a travel and expense grant from that institution.

8. References

T Dunning (1994), Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics, v19, no 1, 1994, pp 61-74.

Kando, N, et al (1999): Overview of IR Tasks at the First NTCIR Workshop. In Proceedings of the First NTCIR Workshop on Research In Japanese Text Retrieval and Term Recognition,NTCIR-1. ISBN 4-924600-77-6. Tokyo Japan, August 30-Sept 1, 1999, pp 11-44.

K Knight & J Graehl (1997), Machine Transliteration; Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics.

Peters, Carol et al. (Eds., 2007) Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer Lecture Notes in Computer Science, LNCS 4730, July 2007.

Yates, F (1934). Contingency tables involving small numbers and the χ^2 test, Journal of the Royal Statistical Society (Supplement) **1**: 217-235.

[5] http://en.wikipedia.org/wiki/Hepburn_romanization

[6] http://www.cpan.org/modules/by-module/LWP/MIYAGAWA/Lingua-JA-Hepburn-Passport-0.01.readme