Related Word-pairs Extraction without Dictionaries

Eiko Yamamoto 1  Kyoji Umemura 2

1 National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto, 619-0289 Japan
eiko@cri.go.jp
2 Toyohashi University of Technology
1-1 Tempaku, Toyohashi, Aichi, 441-8580 Japan
umemura@tutics.tut.ac.jp

Abstract

Although related pairs of words are useful lexical semantic resources, it is sometimes expensive to create and maintain the pairs. We propose a method that extracts pairs of related Japanese words from a text corpus, without the use of language knowledge, such as a dictionary, in any of the steps. This is difficult with a Japanese text because there are no spaces between words. The pairs are related words with similar usages and can be useful for understanding texts including unknown words. These related word pairs are extracted from a text corpus, our method outputs related word pairs using only the definitions of the words. In contrast, given a text and the pairs are used in a similar way. Knowing these kinds of words related to unknown words will help clarify the meanings of the unknown words. Using this concept, we extract special pairs of words for the text corpus including unknown words, and the pairs are used in a similar way. We assume that if two words appear with the same forward and backward strings, they are basically synonyms, and we judge them to be related words. In this paper, we define a set of related word-pairs Relevant as follows.

Related Word-Pair Set

\[ x \text{ and } y \text{ are strings. } a \text{ and } b \text{ are words. } (a,b) \text{ is a score function for } a \text{ and } b \text{ based on frequency information.} \]

\[ X \text{ is the number of documents in a text corpus.} \]

\[ \text{Relevants} = \{(a,b) | \text{score}(a,b) > \alpha \} \]

We define score function \( \text{score}(a,b) \) using \( cIDF \), which represents the characteristic of a word and is quantity-evaluated in terms of how often the word appears (Aizawa, 2000). Many of the measures used in existing search systems are based on this measure. We define the score function as follows.

Score Function

\[ cf(z) \text{ is the total frequency of string } z \text{ in a text corpus.} \]

\[ df^x(z) \text{ is the document frequency predicted by the Poisson distribution for string } z. \]

\[ N \text{ is the number of documents in a text corpus.} \]

\[ \text{score}(a,b) = \sum \text{score}(xay) \cdot \text{score}(xby) \]

where for each string, \( z \).

\[ \text{score}(z) = cf(z) \cdot IDF(z) \cdot \log(N), \]

\[ df^x(z) = N(1 - p(0; cf(z)/N)), IDP^x(z) = -\log(df^x(z)/N). \]
This score function sums up the products of $c_{IDF}$ for word $xay$ and $c_{IDF}$ for word $xby$, where $c_{f(xay)}>1$ and $c_{f(xby)}>1$. We use this estimation because document frequency is harder to calculate than corpus frequency.

**Evaluation**

**Corpora**

We tested our method experimentally by using a collection of summaries of conference papers (NTCIR, 2000) written in Japanese and articles published in the Mainichi Newspaper also written in Japanese. All the documents in each corpus have an ID number, title or heading, and summary or content.

A) NTCIR

We used three NTCIR corpora: NTCIR1, NTCIR2g, and NTCIR2k. NTCIR1 is NACSIS Test Collection 1, which contains documents selected from the Academic Conference Papers Database. The other two, NTCIR2g and NTCIR2k, are included in NII Test Collection 2. NTCIR2g comprises documents selected from the NACSIS Academic Conference Papers Database. NTCIR2k comprises documents selected from the NACSIS Grant-in-Aid Scientific Research Database; they are about three times as long as those in NTCIR2g. We thus divided NTCIR2k into the three corpora: NTCIR2k1, NTCIR2k2, and NTCIR2k3. Table 1 shows the number of documents in each corpus and its size.

| Corpus      | Number (Mbytes) |
|-------------|-----------------|
| NTCIR1      | 333,921 (125)   |
| NTCIR2g     | 116,177 (98)    |
| NTCIR2k1    | 100,000 (138)   |
| NTCIR2k2    | 100,000 (135)   |
| NTCIR2k3    | 87,071 (117)    |

Table 1: Japanese NTCIR Corpora

B) Mainichi Newspaper

The articles from the Mainichi Newspaper were published between 1991 and 1994 and are comprised of four corpora: MAI1991, MAI1992, MAI1993, and MAI1994. Table 2 shows the number of documents in each corpus and its size.

| Corpus | Number (Mbytes) |
|--------|-----------------|
| MAI1991| 91,200 (85)     |
| MAI1992| 101,468 (85)    |
| MAI1993| 91,774 (85)     |
| MAI1994| 101,057 (115)   |

Table 2: Mainichi Newspaper Corpora

**Evaluation Method**

Because we deal with unknown words, it is not possible to make judgments automatically. Accordingly, we made our evaluations using the following method. First, we randomly chose 500 pairs from each related word list, under the condition that the minimum string length to be investigated is 2, and had five people judge these pairs using four criteria, deciding whether they felt each pair was valid. Next, we totaled the results of the five judges to obtain the precision for each text corpus as a tool for evaluation. The four decision criteria were as follows.

1. The pair is a word pair having a relationship such that the two words can be used in the same way.
2. The pair is a word pair in which there is some relationship between the two words.
3. The pair is a word pair in which there is no relationship between the two words.
4. The pair is not a word pair because one or both of the elements constituting the pair is not a word.

These criteria were given score values of 2, 1, -1, and -2, respectively. In addition, we defined two rules.

A) If the total score of the five judges is more than 4 points, then the pair is a related word pair.

B) If the total score of the five judges is less than -6 points, then the elements constituting the pair are not words.

The first rule means that if three of the five judges judged the pair to be a related word pair and the other two did not, then the overall judgment was that the pair was a related word pair. The second rule means that if one judge judged the pair to be a related word pair and the others did not, then the total judgment was that it is not a word pair. We defined the overall judgments so as to set a very high threshold for the decision “The pair is not a word pair.”

**Experimental Results**

Table 3 shows the number of related word pairs extracted from each text corpus for investigated string lengths of 2, 3, 4, 5, and 6.

| Corpus   | 2   | 3   | 4   | 5   | 6   |
|----------|-----|-----|-----|-----|-----|
| NTCIR1   | 1448| 547 | 230 | 75  | 38  |
| NTCIR2g  | 8442| 3564| 2061| 766 | 173 |
| NTCIR2k1 | 11399|3589|1458 | 590 | 270 |
| NTCIR2k2 | 10469|3183|1297 | 463 | 201 |
| NTCIR2k3 | 13204|3902|1396 | 473 | 214 |
| MAI1991  | 4112|1924|1154 |669 | 423 |
| MAI1992  | 1339| 604 |318 | 207 | 96  |
| MAI1993  | 1822|1053|510  |292 | 140 |
| MAI1994  | 12164|5429|2612|1271|681  |

Table 3: Number of Extracted Word Pairs

For lengths of 3 or less, many pairs were extracted from most of the corpora. For lengths of 5 or more, high precision was obtained, but not many pairs. Therefore, we determined that a length of 4 is suitable if both the number of extracted pairs and the precision are taken into consideration. If we give priority to a precision, a length of 5 is suitable because the precision may be much higher at this length than at the other lengths.

Table 4 shows for each case the rate of pairs judged to be valid and the rate of pairs judged to be a word pair. We represented these rates as a percentage. The former is precision, and the latter is the word pair rate. For example, for the NTCIR1 corpus with a string length of 3, the precision was 74.7% (127/170*100). This is because in this case the number of pairs obtained was 170 out of 500 judged pairs, and the number of pairs judged to be valid was 127 out of 170. In this case, the word pair rate was 92.9% (158/170*100) because the number of pairs judged to be word pairs was 158 out of 170.
We have described a method for generating a list of related word pairs that facilitates understanding of the words from a text corpus through a statistical word-extraction method, even if the words are new to the method. It efficiently creates word pairs and judges whether the two words are used in the same way. Furthermore, it extracts related word pairs without using any dictionaries. Using our proposed word definition method makes it possible to generate a list of related words that users have judged to be useful. Experimental results showed that a string length of 4 is suitable for a Japanese text corpus. Even though many of the parameters in the system depend on the corpus, and there remains a trade-off problem between the number of related word pairs and the obtainable precision, the system actually works. It is thus an important step in the development of a useful thesaurus.

**Discussion**

Since our method, including the steps for word segmentation and extraction, does not use dictionaries of any type, it can be used for any language whose word boundaries are not explicit, good examples of which are Chinese and Korean. Our method also makes it possible to extract pairs of words with similar usages in the English language.

**Conclusion**

We analyzed the pairs judged to be related word pairs to determine the relationship between them. Table 5 shows some of the related word pairs obtained from NTCIR1. First, even though Nos. 1-11 are clearly not pairs of words that represent the same thing, they are word pairs that are used in the same way. In particular, Nos. 5 and 11 are related word pairs typical of a specialized field. In our method, pairs of such related words are extracted more than pairs of related words that are simply synonyms. Nos. 12-17 are pairs of related words that are synonym pairs or pairs of a word and the word's abbreviation. Nos. 18-24 are pairs of words that have the same meaning but whose notations differ slightly, i.e., whether the Japanese hiragana syllabary or the katakana one is used to write the word and whether there are notations to which some characters are added. We call such differences “variation of notation.” Such related word pairs are known empirically. Nos. 25-28 are related word pairs whose character codes differ. These also fall in the “variation of notation” category. Nos. 29-33 are related word pairs that are antonym pairs or have the same upper word. Most of the related word pairs obtained from the NTCIR1 and NTCIR corpora can be classified under one of these five relationships. This is because the NTCIR corpus is comprised of documents of academic conference papers, and such documents tend to contain key words. And because the key words appear frequently, they can be extracted with relative ease. From this viewpoint, we consider our system to be a useful means of generating a special list of related words for a text corpus like NTCIR.

**Acknowledgements**

This research is the result of the IPA unexplored software project conducted in fiscal year 2001. We thank Akiko Sanada, Kenji Suzuki, Takashi Funatomi, and Chakma Junan for their cooperation in judging the related word pair lists. We used newspaper articles from The Mainichi Newspaper.

**References**

Aizawa A. (2000). The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 104-111).

Blondel, V.D. and Sennelart P. (2002). Automatic extraction of synonyms in a dictionary. In Proceedings of the SIAM Workshop on Text Mining.

Cho S. & Kakehi K. (2002). Extraction of Synonyms from Search Log by Using “Links.” In Proceedings of the 19th Annual Meeting of Japan Society for Software Science and Technology (pp. 1-5). (in Japanese).
Table 5: Partial list of related words from NTCIR1 corpus

| No. | Word1                        | Word2                        | No. | Word1                        | Word2                        |
|-----|------------------------------|------------------------------|-----|------------------------------|------------------------------|
| 1   | taikukan (体骨)              | kosuya (校令)                | 17  | seijiba (静脈)               | seijenzumonshinkogi (静脈前体静脈) |
| 2   | moi (文字) character         | tango (時語) (word)          | 18  | meikki (メキキ) (plating)     | meikki (めっき) (plating)      |
| 3   | akarusa (明るさ) (brightness)| kido (輝度) (luminance)      | 19  | reda (レーダ) (radar)         | reeda (レーダ) (radar)         |
| 4   | kigo (記号) (symbol)         | LISP                         | 20  | daibaishi (大二等) (flight)   | daibaishichi (大二等) (diversity) |
| 5   | hyooyoi (表情) (expression) | kaogao (顔表) (face image)   | 21  | tankai (戦災) (SiC)          | tankai (戦災) (SiC)           |
| 6   | ressy (列車) (train)         | tetsudo (鉄道) (railroad)    | 22  | tanpakushitsu (蛋白質) (protein)| tanpakushitsu (タンパク質) (protein) |
| 7   | hiko (飛行) (flight)         | koku (空) (flying)           | 23  | nendo (耐点) (clay)          | neseidou (耐性点) (clay)       |
| 8   | genomu (ゲノム) (genome)     | DNA                          | 24  | indenaranorigurizo (遺伝アルゴリズム) | identenaranorigurizo (遺伝アルゴリズム) |
| 9   | fukugoineshi (複合名詞)      | meishika (名詞) (名詞句)     | 25  | jinesi (軽性) (lightness)    | jinesi (軽性) (lightness)     |
| 10  | zogeshitsu (象牙質)          | enameshilushi               | 26  | fukuran (不撹乱) (unperturbed) | fukuran (不撹乱) (unperturbed) |
| 11  | renetsu (連接) (connection) | kyoki (共起) (co-occurrence) | 27  | hibaku (爆破) (atom bombed)  | hibaku (爆破) (atom bombed)   |
| 12  | Pararumadinesupurui (プラスマディスクプレイ) | (plasma display)            | 28  | keiba (顕微) (cervix)        | keiba (顕微) (cervix)         |
| 13  | juguo (ología) (lesson)      | kogi (講義) (lecture)       | 29  | Konbata (コンパートナー) (converter) | inbata (インバーター) (inverter) |
| 14  | koseishien (校正支援)        | suikoshien (校正支援)        | 30  | reibo (冷房) (air conditioner)| DRAM                         |
| 15  | halibebunsu (製物) (waste)  | gomi (ゴミ) (garbage)        | 31  | SIRAM                        |                              |
| 16  | zufukuri (周縁) (amplifier)  | anpu (アンプ) (amplifier)     | 32  | kobunkaiseki (構文解析)       | keitaisokakei (形態解析)      |

Table 6: Partial list of related words from Mainichi Newspaper corpus

| Word1                        | Word2                        | Word1                        | Word2                        |
|------------------------------|------------------------------|------------------------------|------------------------------|
| Nishikashi (西口) (Mr. Nishioka) | Koizumishi (小泉類) (Mr. Koizumi) | chihoo (痴ろう) (dementia) | chihoo (痴ろう) (dementia) |
| Obuchi (小淵) (Mr. Obuchi) | Hashimoto (橋本) (Mr. Hashimoto) | eizu (エイズ) (AIDS) | HIV |
| Jiko (ジコ) (Ziko) | Arashindo (アルシンド) (AirShindo) | daiyaraqutasu (ダイアツルQ2) | daiyaraqutasu (ダイアツルQ2) |
| Kurochajinin (クロチャ者が) (Croatian)| Serubijani (セルビア人) (Serbian) | rakusatsu (落札) (successful bid) | nyusatsu (札) (bid) |
| Chugoku (中国) (China) | Taiwan (台灣) (Taipei) | jokoku (上告) (final appeal) | kiso (起訴) (indictment) |
| Nagano (長野県) | Shizuokaken (静岡県) | hanketsu (判決) (judgment) | sooyo (訴訟) (lawsuit) |
| Nagano Prefecture (長野県) | Shizuoka Prefecture (静岡県) | boto (暴投) (wild pitch) | shikyu (投射) (flying) |
| Minamiafurika (南アフリカ) (South Africa)| Minamia (南ア) (South Africa) | senjumin (先住民) | Aina (アイナ) (Aina) |
| Sekisuiahuu (水水ハウス) (Sekisui House) | Unichika (ユニチカ) (Unitika) | EAECD | APEC |
| Toyota (トヨタ) | Yanase (ヤナセ) (YANASE) | Wakahamada (若花田) | Wakahamada (若花田) |
| Nihonbikuta (日本ビクタ) (JVC) | Yusassangyo (ユアサ産業) (YUSA) | yorikiri (落札) | oshiдачи (押し出し) |

Church, K.W. (2000). Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to p/2 than p2. In Proceedings of the 18th International Conference on Computational Linguistics (pp. 180–186).
Curran, J. (2002). Ensemble Methods for Automatic Theaurus Extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 59–66).
EDR Electronic Dictionary (1995). http://www2.crl.go.jp/kk/e416/EDR/index.html
Fellbaum, C. (1998). WordNet: an electronic lexical database. The MIT Press.
Fujita, A. & Inui, K. (2001). Paraphrase of Common Nouns to Its Synonyms by using Definition Statements. In Proceedings of the Seventh Annual Meeting of the Association for Natural Language Processing (pp. 331-334) (in Japanese).
Grefenstette, G. (1994). Explorations in Automatic Theaurus Discovery. Kluwer Academic Publishers.
Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Imai, O., and Imamura, T. (1997). Japanese Morphological Analysis System ChaSen. Nara Advanced Institute of Science Technology Technical Report.
NTCIR Projet (2000). http://research.nii.ac.jp/ntcir/.
Okamoto, H., Sato, K., and Saito, H. (2003). Preferential Presentation of Japanese Near-Synonyms Using Definition Statements. In Proceedings of the Second International Workshop on Paraphrasing (pp. 17–24).
Rapp, R. (2002). The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In Proceedings of the 19th International Conference on Computational Linguistics (pp. 821–827).
Takeda, Y. & Umemura, K. (2002). Selecting Indexing Strings using Adaptation. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 427–428).
Wu, H. & Zhou, M. (2003). Optimizing Synonym Extraction Using Monolingual and Bilingual Resources. In Proceedings of the Second International Workshop on Paraphrasing (pp. 72–79).