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@crl.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 based on judgments of whether two words are used in a similar way. We report the precisions of pair lists extracted from various kinds of corpora and analyze the tendencies of each list.

Introduction

A thesaurus is a commonly used lexical semantic resource; examples include WordNet (Fellbaum, 1998) and the EDR Electronic Dictionary (EDR, 1995). We have taken on the challenge of building a thesaurus automatically from text corpora with the goal of using it to help understand texts including unknown words.

A good way to get an understanding of an unknown word is to find a synonym or near-synonym of the word. We extract related word pairs based on the assumption that synonyms may always not occur together but simply be two words with similar usages. This approach is similar to that used by Rapp (Rapp, 2002).

The most resources used for synonym extraction are text corpora (Grefenstette, 1994; Cho & Kakehi, 2002; Curran, 2002; Wu & Zhou, 2003). These works use a certain dictionary for word segmentation, word selection, and/or addition of part-of-speech information. Other resources, such as sets of definition statements extracted from dictionaries, are also used for synonym extraction (Fujita & Inui, 2001; Blondel & Sennelart, 2002; Okamoto et al., 2003). The synonyms extracted by these methods depend on the definitions of the words. In contrast, given a text corpus, our method outputs related word pairs using only information about the frequency of strings in the text for each step, from word segmentation to related word pair judgment. Therefore, our system can consider not only known words but also unknown words.

In this paper, we describe the extraction from various corpora lists of related word pairs as a special thesaurus for the corpus. We report the precisions and analyze the tendencies for each list of pairs.

Method

The first step of our method is word segmentation. Words in a Japanese text are usually segmented using a Japanese morphological-analysis system as "ChaSen" (Matsumoto et al., 1997), which uses many dictionaries. In contrast, we use a keyword extraction system that does not need a dictionary; it segments words merely by using information about the frequency of strings in the text (Takeda & Umemura, 2002). This system mainly uses Adaptation (Church, 2000). Using this system, we can get a list of keywords from texts.

The second step is related word pair judgment. To extract related word pairs from candidate pairs, the method judges whether two words are used in a similar way. In particular, the method investigates whether the two words have the same forward and backward strings. For example, the Japanese sentences "Watashi wa nengajo wo insatsu sita" and "Boku wa nengajo wo purinto sita" both mean "I printed New Year's cards." In this case, the verb "insatsu sita (printed)" can be replaced with "purinto sita (printed)," because "insatu" and "purinto" are synonyms in Japanese. Antonyms and hyponyms can also be 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 string, they are basically synonyms, and we judge them to be related words. In this paper, we define a set of related word-pairs *Relevants* as follows.

Related Word-Pair Set

x and *y* are strings. *a* and *b* are words. *xay* and *xby* are strings that join *x* before *a* or *b* and *y* after *a* or *b*. *score(a,b)* is a score function for *a* and *b* based on frequency information.

$$\text{Relevants} = \{(a,b) \mid score(a,b) > \alpha\}$$

We define score function *score(a,b)* using *cfIDF*, 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) is the total frequency of string *z* in a text corpus. *df"(z)* is the document frequency predicted by the Poisson distribution for string *z*. *N* is the number of documents in a text corpus. *score(z)* is the score for string *z*.

$$score\ (a,b) = \sum_{x,y} score\ (xay) \cdot score\ (xby)$$

where for each string,

$$score\ (z) = cf\ (z) \cdot IDF"(z) / \log(N),$$

$$df"(z) = N(1 - p(0; cf(z)/N)), IDF"(z) = -\log(df"(z)/N).$$

This score function sums up the products of *cfIDF"* for word *xay* and *cfIDF"* for word *xby*, where *cf(xay)>1* and *cf(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.

Corpus	Rate [%]	2	3	4	5	6
NTCIR1	Precision	66.6	74.7	77.0	80.0	70.0
	Word	88.6	92.9	91.8	86.7	80.0
NTCIR2g	Precision	52.4	60.3	65.3	73.0	40.0
	Word	84.8	87.0	89.8	89.2	80.0
NTCIR2k1	Precision	52.4	56.8	60.4	77.8	66.7
	Word	85.6	89.9	94.3	100.0	100.0
NTCIR2k2	Precision	46.4	52.7	46.2	61.5	57.1
	Word	84.4	84.0	75.0	76.9	57.1
NTCIR2k3	Precision	39.2	49.0	42.9	50.0	33.3
	Word	86.8	86.7	78.6	75.0	66.7
MAI1991	Precision	35.4	22.6	24.0	24.0	23.9
	Word	79.4	75.8	78.8	64.0	63.0
MAI1992	Precision	20.1	20.5	25.4	18.8	16.7
	Word	67.6	56.8	63.4	62.5	55.6
MAI1993	Precision	16.0	8.5	11.1	6.1	7.1
	Word	68.0	64.8	67.5	63.3	64.3
MAI1994	Precision	19.4	27.3	28.7	40.0	30.8
	Word	54.8	58.8	61.1	66.0	61.5

Table 4: Precision and Word Pair Rate by String Length

The rate of pairs judged to be word pairs for a length of 4 was very high (75.0-94.3% in the NTCIR corpora and 61.1-78.8% in the Mainichi corpora). The corresponding precisions were 42.9-77.0% and 11.1-28.7%. Therefore, the rates of pairs judged to be word pairs and related word pairs were 54.6-83.9% in the NTCIR corpora and 16.4-47.0% in the Mainichi corpora. We can thus see that the precision for the Mainichi corpora was lower than that for the other corpora. To clarify the reason for this, we analyzed the obtained related word lists.

Analysis

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.

Table 6 shows some of the related word pairs obtained from the Mainichi Newspaper. From this corpus, we obtained many related word pairs like personal names, company names, place names, and group names. We also obtained a number of related word pairs, e.g., general words. Moreover, there were many pairs that were judged to be word pairs although the words were not related. Most of such pairs were word pairs containing numerals. For example, the pair of "61.2 kiro (61.2 kg)" and "50.8 kiro (50.8 kg)" were judged to be a related word pair by two judges, but judged to be not a word pair by the other three judges. Overall, therefore, such pairs were judged to be unrelated pairs on a point basis. We found that such pairs resulted in low precision in these corpora.

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 airs of words with similar usages in the English language.

Conclusion

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.

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).

No.	Word1	Word2	No.	Word1	Word2
1	taiikukan (体育館) (gym)	kousya (校舎) (schoolhouse)	17	seijiha (静磁波) (magnetostatic wave)	seijizenshintaisekiha (静磁前進体積波) (magnetostatic forward volume wave)
2	moji (文字) (character)	tango (単語) (word)	18	mekki (メッキ) (plating)	mekki (めっき) (plating)
3	akarusa (明るさ) (brightness)	kido (輝度) (luminance)	19	reda (レーダー) (radar)	reeda (レーダ) (radar)
4	kigo (記号) (symbol)	LISP	20	daibaashitexi (ダイバーシティ) (diversity)	daibaashichi (ダイバーシチ) (diversity)
5	hyojyo (表情) (expression)	kaogazo (顔画像) (face image)	21	tankakeiso (炭化珪素) (SiC)	tankakeiso (炭化ケイ素) (SiC)
6	ressya (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 (航空) (flying)	23	nendo (粘土) (clay)	nenseido (粘性土) (clay)
8	genomu (ゲノム) (genome)	DNA	24	indenarugorizumu (遺伝アルゴリズム) (genetic algorithm)	identekiarugorizumu (遺伝的アルゴリズム) (genetic algorithm)
9	fukugoumeishi (複合名詞) (compound noun)	meishiku (名詞句) (noun phrase)	25	jinsei (靭性) (toughness)	jinsei (靱性) (toughness)
10	zogeshitsu (象牙質) (dentin)	enamerushitsu (エナメル質) (enamel)	26	fukakuran (不攪乱) (undisturbed)	fukakuran (不撹乱) (undisturbed)
11	rensetsu (連接) (connection)	kyoki (共起) (co-occurrence)	27	hibaku (被爆) (atom bombed)	hibaku (被曝) (atom bombed)
12	Purazumadexisupurei (プラズマディスプレイ) (plasma display)	PDP (plasma display)	28	keibu (頸部) (cervix)	keibu (頚部) (cervix)
13	jugyo (授業) (lesson)	kogi (講義) (lecture)	29	Konbata (コンバーター) (converter)	inbata (インバーター) (inverter)
14	koseishien (校正支援) (proofreading support)	suikoshien (推敲支援) (elaboration support)	30	reibo (冷房) (air conditioner)	danbo (暖房) (heater)
15	haikibutsu (廃棄物) (waste)	gomi (ごみ) (garbage)	31	SRAM	DRAM
16	zufukuki (増幅器) (amplifier)	anpu (アンプ) (amplifier)	32	kobunkaiseki (構文解析) (syntax analysis)	keitaisokaiseki (形態素解析) (morphological analysis)

Table 5: Partial list of related words from NTCIR1 corpus

Word1	Word2	Word1	Word2
Nishiokashi (西岡氏) (Mr. Nishioka)	Koizumishi (小泉氏) (Mr. Koizumi)	chiho (痴呆) (dementia)	chiho (痴ほう) (dementia)
Obuchi (小渕) (Mr. Obuchi)	Hashimoto (橋本) (Mr. Hashimoto)	eizu (エイズ) (AIDS)	HIV
Jiko (ジーコ) (Ziko)	Arushindo (アルシンド) (Alcindo)	daiyaruqutsu (ダイアル Q2) (Dial Q2)	daiyarutsuqu (ダイアル 2Q) (Dial Q2)
Kuroachiajin (クロアチア人) (Croatian)	Serubiajin (セルビア人) (Serbian)	rakusatsu (落札) (successful bid)	nyusatsu (入札) (bid)
Chugoku (中国) (China)	Taiwan (台湾) (Taipei)	jokoku (上告) (final appeal)	kiso (起訴) (indictment)
Naganoken (長野県) (Nagano Prefecture)	Shizuokaken (静岡県) (Shizuoka Prefecture)	hanketsu (判決) (judgment)	sosyo (訴訟) (lawsuit)
		boto (暴投) (wild pitch)	shikyu (四球) (walk)
Minamiafurika (南アフリカ) (South Africa)	Minamia (南ア) (South Africa)	senjumin (先住民) (aborigine)	Ainu (アイヌ) (Ainu)
Sekisuihausu (積水ハウス) (Sekisui House)	Unichika (ユニチカ) (Unitika)	EAEC	APEC
Toyota (トヨタ) (Toyota)	Yanase (ヤナセ) (YANASE)	Wakahanada (若花田) (Wakahanada)	Wakanohana (若乃花) (Wakanohana)
NihonBikuta (日本ビクター) (JVC)	Yuasasangyo (ユアサ産業) (YUASA)	yorikiri (寄り切り) (yorikiri)	oshidashi (押し出し) (oshidashi)

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

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 Thesaurus 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 Thesaurus 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).