Acquisition of Lexical Paraphrases from Texts

Kazuhide Yamamoto
ATR Spoken Language Translation Research Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
yamamoto@fw.ipsj.or.jp

Abstract

Automatic acquisition of paraphrase knowledge for content words is proposed. Using only a non-parallel text corpus, we compute the paraphrasability metrics between two words from their similarity in context. We then filter words such as proper nouns from external knowledge. Finally, we use a heuristic in further filtering to improve the accuracy of the automatic acquisition. In this paper, we report the results of acquisition experiments.

1 Introduction

Paraphrasing research has attracted increased attention, and the work in this field has become more active recently. Paraphrasing involves various types of transformations of expressions into the same language, and thus there is generally no all-purpose design and information resource. Among the many types of paraphrasing, a handwritten construction may be best for syntactic paraphrasing knowledge or knowledge of functional words because the number of resulting phenomena can be counted. On the other hand, we need to acquire lexical paraphrasing knowledge automatically or efficiently, since there is an enormous number of phenomena observed for an enormous number of content words.

Some works, such as Barzilay and McKeown (2001), have acquired paraphrasing knowledge automatically. All of those works found differences from a *paraphrase corpus*, where each expression is aligned to another expression (or more) with the same meaning and in the same language. Unfortunately, there is no paraphrase corpus widely available except for a few collections such as those prepared by Shirai et al. (2001) and Zhang et al. (2001). Most of those works collected paraphrase corpora by employing special situations, such as multiple news resources from the same events or multiple translations of the same (and well-known) story in other languages. However, since these situations seem to be really special, we believe that the collection of many paraphrase corpora in the near future is quite hopeless. Consequently, it is necessary to conduct a feasibility study on collecting various kinds of paraphrase knowledge from non-paraphrase corpora, particularly from raw text corpora. Although we have already reported extracting paraphrasing knowledge of Japanese noun modifiers from a raw corpus (Kataoka et al., 1999), we need to explore other types of expressions.

With this motivation, we have attempted to acquire paraphrasing knowledge on content words, mainly nouns and verbs. As a knowledge source, we use newspaper articles collected over one year in text format, which is regarded as the most generally used corpus. In this trial, we propose the following two principles of acquisition:

- Conditions for applying each type of paraphrasing knowledge should be obtained.
- Paraphrasing knowledge should have directionality.

In other words, all of the paraphrasing patterns obtained by the conventional methods seem to have been applied unconditionally, that is, conventional approaches tend to target only unconditional patterns. However, most paraphrasing phenomena depend on their context; paraphrasing can be possible only if a paraphrased expression fits the context.

Moreover, directionality of the rules is an important issue for paraphrasing, although no other works have discussed this. Despite the

existence of synonymy, even if expression E_1 can be replaced by expression E_2, it is unsure whether E_2 can be paraphrased by using E_1. We discuss this feature in the experiments.

2 Contextual Similarity vs. Synonymy

Paraphrasability is the degree of replacability for two expressions E_1 and E_2, which are regarded as different from each other in some sense. This definition implies the notion that E_1 should not be judged as the same (or similar) as E_2 in the sense of meaning. Of course, similarity of meaning and paraphrasability are very closely correlated with each other, and Kurohashi and Sakai (1999) utilize this feature to paraphrase Japanese expressions (in order to comprehend them more easily). They use a Japanese dictionary written for humans (or more precisely, children) to replace a part of the target expression with a different one by judging its local context computed by a thesaurus.

We propose that replacability (obtained by the corpus, for example) is a more important factor in judging the paraphrasability of expressions than their meaning as defined in a dictionary. For example, words used only in some special situations, such as for children or in ancient documents, should not be used in a paraphrase even though it has synonymy.

On the contrary, even if synonymy is not satisfied, we still focus on expressions that are replaceable. Hypernymy is one example of this. A hypernym is not a paraphrase in a strict sense due to the loss of information. However, this kind of paraphrasing is still useful from the engineering point of view. For instance, these *loose* (and therefore many) paraphrases are more effective in the case of reluctant processing, where we must necessarily change an expression for various reasons such as our requirement in paraphrase-based machine translation (Yamamoto, 2002). Moreover, this kind of paraphrase loses nothing when it is used as anaphora or when it is trivial and out of major interest in the context used. Not all hypernyms are always paraphrasable, so we cannot list this type of paraphrase from only a thesaurus.

3 Approach and Implementation

This section describes our approach to acquiring paraphrase knowledge from a text corpus. We use Perl programming language to implement all of the following processes and experiments.

3.1 Collection of context from corpus

We first define the term *context* in this paper. The context of a certain content word is defined as direct dependency relations between the word and the words that surround it in texts. That is, the context of a content word c is, in our sense, defined as the collection of words upon which c directly depends as well as the collection of words that directly depend on c.

Under this definition, we first collected all of the dependency relations observed in the corpus. Each article is segmented and part-of-speech tagged by the morphological analyzer JUMAN[1] and then parsed by the KNP[2] parser. We then obtained a relation triplet (c_1, r, c_2) from each article, where a word c_1 depends on a word c_2 with the relation r. A complete list of r types and their examples is shown below:

- (Noun, r_1, Noun)
  e.g. '今度 (this time) の (of) 法律 (law)'
  'テロ (terrorism) 法案 (bill)'
- (Adjective, r_2, Noun)
  e.g. '新たな (new) 法律 (law)'
- (Noun, r_3, Verb)
  e.g. '衆議院 (Lower House) が (SUB) 可決する (approve)'
- (Verb, r_4, Noun)
  e.g. '空爆する (bombing) 米軍 (U.S. army)'

In this list, r_i can be a particle, such as a case particle (r_3) or an associative particle "の" (r_1). Another type of possible r_i in the list is a syntactic relation expressed without any particle or other functional marker. For instance, a verb or an adjective directly modifies a noun without using any functional words in Japanese. In this case, we introduce the notion of a constituent boundary proposed by Furuse and Iida

[1] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

[2] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

(1994), which is a virtual functional marker inserted between two consecutive content words, in order to more easily analyze a sentence. For instance, if there are two consecutive nouns, we assume that $\langle nn \rangle$ is inserted between the two nouns, and consequently the relation r_i is $\langle nn \rangle$.

3.2 Bigraph construction

We then transform the collection of triplets into a bigraph (2-partite graph). In the first step, each triplet in the collection is converted into two couplets consisting of a content word and an operator by the following definition: an operator consists of a content word c and a relation with directionality r. It is defined as either $r \rightarrow c$ (something depends on c by r) or $r \leftarrow c$ (c depends on something by r). For instance, suppose that a triplet is (c_1, r, c_2), then both a couplet for the first content word c_1, i.e., (c_1, $r \rightarrow c_2$) and a couplet for the second content word c_2, i.e., (c_2, $r \leftarrow c_1$) are extracted in this operation.

We perform this conversion for all of the triplets, and a list of couplets is obtained. From the viewpoint of graph theory, this couplet list is a bigraph, such as figure 1, which consists of two sets (content word set and operator set) and a list of edges, where each edge connects an element in one set to an element on the other side. This bigraph is a weighted graph, and each weight expresses the frequency of appearing in the corpus.

3.3 Paraphrasability computation

In the next step, we compute paraphrasability. In this work, the paraphrasability P for any two content words c_i and c_j is defined in the following formula:

$$P(c_i, c_j) = \frac{\sum_{m \in M(c_i) \cap M(c_j)} p(m, c_i)}{\sum_{m \in M(c_i)} p(m, c_i)} \quad (1)$$

$$M(c_i) = \{m | f(m, c_i) > 1\} \quad (2)$$

$$p(m, c_i) = \frac{f(m, c_i)}{\sum_c f(m, c)} \quad (3)$$

In this formula, let $f(m_0, c_0)$ be frequency of content word c_0 with operator m_0. In other words, $f(m_0, c_0)$ is a weight of edge (c_0, m_0) in the bigraph.

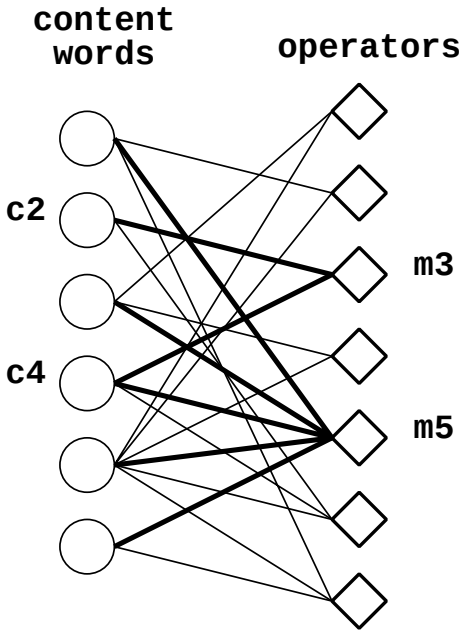

Figure 1: Example of a bigraph (2-partite graph)

This formulation can be explained as follows. Paraphrasability between two content words c_i and c_j increases if these words behave similarly in terms of their dependency relations. That is, this metrics compares the similarity of the contextual situations of the two input words. The definition states that paraphrasability computes the number of operators that c_j links among the operators that c_i links.

However, we believe that the importance of each operator m is not equivalent to that of the others. For example, in figure 1, the operator m_3 is linked by only two words, c_2 and c_4, while the operator m_5 is linked by almost all of the words. In this situation, it is not reasonable to handle the two operators equally, since m_3 may confirm that the two words are similar or paraphrasable, whereas m_5 may be a general operator widely used in various situations. In other words, when we compute paraphrasability from c_2 to c_4, the edge (c_4, m_5) is judged as less important than the edge (c_4, m_3) or (c_2, m_3). Consequently, each operator is weighted by the definition of formula (3). Moreover, instances of low frequency are regarded as accidental and insignificant, so we filter out links where an in-

stance appears only once.

It is obvious in the definition of (1) that $0 \leq P(c_i, c_j) \leq 1$, and a higher score expresses a higher possibility of paraphrasing. More importantly, the definition indicates the relation $P(c_i, c_j) \neq P(c_j, c_i)$, i.e., there is a directionality that gives larger differences than any similarity metrics. Even if an expression E_1 has a large paraphrasability for an expression E_2, it is completely uncertain whether the paraphrasability of E_2 into E_1 is high or low.

3.4 Paraphrase knowledge filtering

By only taking the discussion of the last subsection into account, we can compute paraphrasability between any two content words. However, this measure is not the final judgment of paraphrasability: some pairs score very high even though they are not paraphrasable. For example, the pair *three* and *four* may have a very high score but are of course not paraphrasable. In our observation, the following kinds are found to be misjudged as paraphrasable by our definitions.

1. number, e.g., '三'(three) →'四'(four)
2. proper noun,
   e.g., '北京'(Beijing) →'台北'(Taipei)
3. antonym, e.g., '右'(right) →'左'(left)

Obviously, these errors occurred due to one of the limitations of our approach; since the formula only has an interest in the context of the words found in the corpus, not in the sense of the words found in a dictionary.

However, we can filter out these kinds of word pairs by introducing language resources external to the corpus. First, we can judge whether the word is a number by applying some simple rules. Second, we can now easily obtain extensive lists of both major proper nouns and antonyms. We obtain the proper noun list from GoiTaikei[3], one of the largest Japanese electronic thesauri, in which 169,682 proper noun entries are extracted. We obtain the antonym list from both Gakken Kokugo Daijiten (a Japanese word dictionary) and Kadokawa Ruigo Shinjiten (a Japanese thesaurus), which have 11,981 antonym pairs in total.

[3] http://www.kecl.ntt.co.jp/icl/mtg/resources/GoiTaikei/

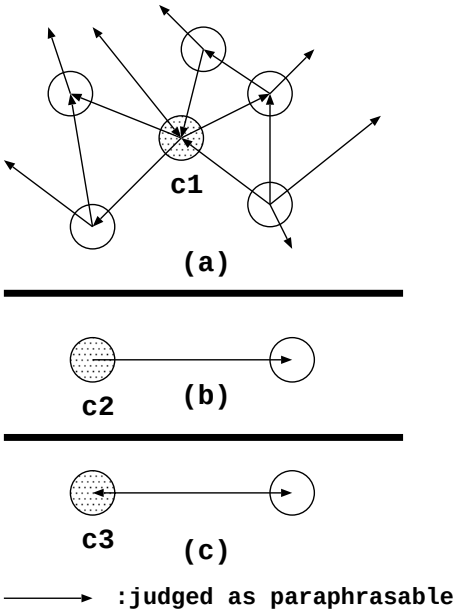

Figure 2: Heuristic by number of links

In fact, further filtering is necessary in order to reduce errors. For example, in English, *guitar*, *piano*, and *flute* have very similar contexts, such as "*to play the* ____," "*an electric* ____," "*a violin and a* ____," and so on, although they are naturally not paraphrasable. We predict that in order to use lexical paraphrase collection for filtering, future research will need to concentrate on how to collect word pairs that are not paraphrasable but have the same context.

3.5 Further filtering by heuristic method

In the final process, we filter the pairs further by using our proposed heuristic to improve the acquisition accuracy.

From our observations of the results obtained by the above operations, we found a clear tendency in words that have a very high frequency or very broad sense: these words tend to be judged as having a high paraphrasability from many words or to many words, even if they are not actually paraphrasable. For example, in figure 2 (a), a content word such as c_1 tends to be misjudged as paraphrasable if c_1 links to many words and/or if c_1 is linked by many words. In other words, case (b) of the figure, where a word c_2 connects to only one word, would more likely have its paraphrasing judged as proper. We also build a hypothesis that case (c), where

Table 1: Evaluation for Content Words

	Case 1	Case 2	Total
Extracted	668	1149	1684
Paraphrasable	422	780	1117
Accuracy	63.2%	67.9%	66.3%

Case 1: a word paraphrases to one word
Case 2: a word is paraphrased from one word

two words are exchangeable, has more accuracy than the other two cases, which are evaluated in the next section.

We assume these errors occurred because such words can have dependency relations with many words, i.e., such words are general and frequently appearing. Consequently, such cases are unexpectedly judged as being highly paraphrasable from or to many words. As these words are used many times in many contexts, the possibility of inserted noises also increases. Therefore, distinguishing noises from real paraphrases becomes difficult.

These spurious paraphrases should not remain in the final results, so we conduct another filtering according to the above analysis. The actual process is conducted as follows. For each c_i, we count the number of c_j that satisfies the relation $P(c_i, c_j) > P_{const}$. If there is only one word c_j that satisfies this relation, we finally determine that c_i can paraphrase to c_j. Similarly, for each c_j, we count the number of c_i that satisfies the relation $P(c_i, c_j) > P_{const}$. If there is only one word c_i that satisfies the relation, we finally determine that c_i can paraphrase to c_j. In the experiment below, we set $P_{const} = 0.1$.

In this heuristic filtering, some word pairs that are actually paraphrasable may, unfortunately, also be lost. The problem of saving them remains for our future work.

4 Knowledge Acquisition Experiment

4.1 Experiment on content word paraphrasing

We have conducted an experiment of paraphrasing knowledge acquisition in the following conditions. The corpus we used was all articles of The Mainichi Shimbun, which is one of the national daily newspapers of Japan, published in the year 1995. The size of the corpus is 87.3 MB, consisting of 1.33 million sentences.

Table 1 illustrates evaluation results of knowledge acquisition. The results show that our proposed process can choose approximately 1,700 paraphrase pairs that have 66% accuracy. Although this accuracy is not satisfactory for an automatic process, it is already helpful from the engineering point of view; accordingly, we can obtain a large amount of high-quality paraphrase pairs with a minimum human check in significantly less time than one day.

We also show the acquired paraphrase pairs with the highest paraphrasabilities in Table 2. Note that P (←) in the table denotes the paraphrasability of the inverted paraphrases, from right to left direction, and the symbol * indicates that this direction is also judged as paraphrasable, i.e., these two words are determined to be paraphrasable with each other. We found that most of the entries in the list are correctly judged to be paraphrasable, even though some of them cannot be paraphrasable, such as "したてなげ (underarm throwing)" into "うわてなげ (overarm throwing)"[5].

We can also confirm that the directionality of the proposed measure works quite well. For example, we can paraphrase the term "逸話 (anecdote)" with the more general term "話 (story)," but it is impossible to replace the latter with the former except in some restricted context. The outputs seen in this table illustrate such an intuition.

Table 2: Highest Paraphrasable Pairs

Paraphrase pair	P	P (←)
逸話 (anecdote) → 話 (story)	1	.0015
したてなげ → うわてなげ	1	.2539*
かぎり (only) → 限り (only)	1	.1877*
図式 (scheme) → 構図 (form)	.9982	.2671*
パニック (panic) → 混乱 (confusion)	.9978	.0496
勝ち (win) → 勝ち (win)[4]	.9802	.5176*
ホッケー (hockey) → 野球 (baseball)	.9752	.0286
結党 (formation) → 結成 (formation)	.9672	.0449
違和感 (incongruity) → 痛み (pain)	.9667	.0352
激変 (drastic change) → 変化 (change)	.9582	.0177

[4]These two words have the same string but different part-of-speech, so our tagger judges these two as different.

[5]Both are names of techniques in *sumo* wrestling.

Table 3: Paraphrasability of Operators

Paraphrase pair	P
依頼 (request) が → 注文 (order) が	1
理事 (director) を → 教授 (professor) を	.9940
支部 (branch) で → 地裁 (district court) で	.9334
高裁 (high court) で → 地裁 で	.8734
⟨*nn*⟩ 短期 (short term) →⟨*nn*⟩ 短大 (college)	.8723
⟨*nn*⟩ ヤ (Swallows) →⟨*nn*⟩ 巨 (Giants)	.8553
毎週 (every week) ⟨*nn*⟩→⟨*nn*⟩ 夜 (night)	.8123
市議 (city councillor) ⟨*nn*⟩→ 県議 ⟨*nn*⟩	.8063
十数 (several) ⟨*nn*⟩→ 数 (several) ⟨*nn*⟩	.7961
県議 (pref. assemblyman) ⟨*nn*⟩→ 市議 ⟨*nn*⟩	.7859

If the process judges that the two words can paraphrase each other, these words are considered to be a paraphrase in a narrow sense. In this experiment, we can extract 114 pairs that satisfy this relation, and 75 of these pairs are evaluated as being correct, for an accuracy of 65.8%.

4.2 Experiment for acquisition of operator paraphrase

So far in this paper, we have been using an operator set to compute any of two words in the content word set in the bigraph. We found that we can also do this in the reverse way: computing any of two operators by using the content word set. This is possible because even if we turn a bigraph upside-down, it is still a bigraph. In this subsection, we report an experiment on computing the paraphrasability of operators by the same procedure as above.

After multiple filtering, 432 pairs were judged as paraphrasable. From these we found that the number of correct pairs was 312 (72.2% accuracy). Table 3 illustrates the final paraphrasable pairs with the highest paraphrasability.

Unfortunately, these pairs include errors, so their performance in an automatic process should be improved. However, this performance is still promising for a human-assisted tool.

We investigated the pairs and found that there were various kinds of paraphrasing knowledge obtained in this process. Not only paraphrases of content words but also paraphrase knowledge of the following types were obtained in this experiment.

- insertion and deletion of the particle "の" in noun-noun sequences
- paraphrasing for case particles; in Japanese, it may be possible to change a particle under a certain context.
- voice conversion
- different description of the same word, e.g., from a Chinese-origin word to a native Japanese word

5 Related Works

Lexical paraphrasing is very useful in information retrieval, since it is necessary to expand terms for improving retrieval coverage. Jacquemin et al. (1997) have proposed acquiring syntactic and morpho-syntactic variations of the multi-word terms using a corpus-based approach. They have searched for variation, i.e., similar expressions using (a part of) the input words, such as *technique for measurement* against *measurement technique*, while our target is the paraphrase of a single content word.

The goal of our work is to obtain lexical knowledge for paraphrasing. For this purpose we use contextual similarity, which is also used in the sense similarity computation task in the fields of natural language processing, artificial intelligence, and cognitive science. Moreover, the idea of corpus-based context extraction is basically the same and also used in the task of automatic construction of thesauri or sense determination of unknown words.

Although this is the first work to use context for paraphrase knowledge extraction, many previously reported works have used context for similarity calculation. Paraphrasability and word sense similarity may seem like similar metrics, but there are critical differences between the two tasks. First, similarity satisfies the symmetrical property while paraphrasability does not (explained in 3.3). Second, similarity is a relative measure while paraphrasability is an absolute measure; in many cases, we can answer "*Can* E_1 *paraphrase to* E_2*?*", but it is hard to answer "*Is* E_1 *similar to* E_2*?*". In other words, it is important to collect paraphrases while it may be pointless to collect *similar words*, since the border for the former is clearer than that of the latter.

The kind of information used for defining context is important. For this question, Nagamatsu

and Tanaka (1996) used a deep case (seen in a semantically tagged corpus), and Kanzaki et al. (2000) only extracted relations of nominal modification. The most closely related work in terms of similarity source is the work of Grefenstette (1994), where they obtained subject-verb, verb-object, adjective-noun, and noun-noun relations from a corpus. In contrast, as discussed in subsection 3.1, we propose extracting all of the dependency relations around content words, i.e., nouns, verbs, and adjectives. This is the first attempt to introduce these features into a context definition, and it is obvious that coverage of extracted pairs becomes wider by using various features. However, we have not conducted enough experiments to prove that these factors are effective. This remains for our future work.

6 Conclusions

We propose a process to acquire paraphrasing pairs of content words from a non-parallel raw corpus. We utilize contextual similarity, obtained from the corpus, to compute paraphrasability between any two content words. Some of the word pairs that unexpectedly have high paraphrasability are filtered out by using external linguistic knowledge such as proper nouns and antonyms. Moreover, our proposed heuristic, obtained through observation, can increase acquisition accuracy. These processes in combination are able to obtain more than 1,700 paraphrase pairs with approximately 66% accuracy.

Our interest in this research is not to pursue higher accuracy in automatic processing but to obtain any kind of paraphrasing knowledge as fast as possible. From this point of view, the coverage of the acquisition process is a more serious problem for us than accuracy. Our preliminary experiment showed that a drastic drop in accuracy is observed even if we increase coverage gradually. We need to find another filtering criterion to avoid this problem.

Acknowledgment

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

References

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. of ACL-2001*, pages 50–57.

Osamu Furuse and Hitoshi Iida. 1994. Constituent boundary parsing for example-based machine translation. In *Proc. of Coling'94*, pages 105–111.

Gregory Grefenstette. 1994. Corpus-derived first, second, third-order word affinities. In *Proc. of EURALEX'94*.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proc. of ACL-EACL'97*, pages 24–31.

Kyoto Kanzaki, Qing Ma, and Hitoshi Isahara. 2000. Similarities and differences among semantic behaviors of Japanese adnominal constituents. In *Proc. of ANLP/NAACL 2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing System*, pages 59–68.

Akira Kataoka, Shigeru Masuyama, and Kazuhide Yamamoto. 1999. Summarization by shortening a Japanese noun modifier into expression "A no B". In *Proc. of NLPRS'99*, pages 409–414.

Sadao Kurohashi and Yasuyuki Sakai. 1999. Semantic analysis of Japanese noun phrases: A new approach to dictionary-based understanding. In *Proc. of ACL'99*, pages 481–488.

Kenji Nagamatsu and Hidehiko Tanaka. 1996. Estimating point-of-view-based similarity using POV reinforcement and similarity propagation. In *Proc. of Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 373–382.

Satoshi Shirai, Kazuhide Yamamoto, and Francis Bond. 2001. Japanese-English paraphrase corpus. In *Proc. of NLPRS2001 Workshop on Language Resources in Asia*, pages 23–30.

Kazuhide Yamamoto. 2002. Machine translation by interaction between paraphraser and transfer. In *Proc. of COLING2002*.

Yujie Zhang, Kazuhide Yamamoto, and Masashi Sakamoto. 2001. Paraphrasing utterances by reordering words using semi-automatically acquired patterns. In *Proc. of NLPRS2001*, pages 195–202.