BUILDING A FREE, GENERAL-DOMAIN PARAPHRASE DATABASE FOR JAPANESE

Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura

Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan
masahiro-mi@is.naist.jp

ABSTRACT

Previous works have used parallel corpora and alignment techniques from phrase-based statistical machine translation to extract and generate paraphrases. In Japanese, paraphrases for a number of paraphrase categories or domains have been extracted by this method. However, most of these resources focus on a particular phenomenon in Japanese, and there are still no Japanese paraphrase resources that cover all varieties of phrases from several domains, and are freely available. In addition, because Japanese and English vary in grammar and word ordering, we perform syntax-based preprocessing to reduce this mismatch and extract paraphrases similar in quality to those extracted using more similar language pairs. The data used in creating the Japanese paraphrases is either in the public domain, or available under the Creative Commons license, and spans a variety of genres for wide coverage.

Index Terms— Paraphrasing, Free Data, General-Domain

1. INTRODUCTION

Paraphrases are alternative ways of conveying the same meaning, and are useful in a number of NLP applications such as machine translation and question answering [1, 2]. In this paper, we concern ourselves with building paraphrase resources for Japanese. In Japanese, methods and resources have been proposed for paraphrasing for a number of categories of paraphrases or domains [3, 4, 5, 6]. However, most of these resources focus on a particular phenomenon in Japanese, and there are still no broad-coverage and freely available resources.

In previous research on paraphrasing, methods that use bilingual corpora have proven successful [7, 8]. In these methods, paraphrases for one language (e.g., English) are acquired by treating another language (e.g., French) as an intermediate meaning representation, as described more completely in Section 3. In this paper, we describe a paraphrasing resource that we constructed for Japanese using a similar method, with Japanese as our target language and English serving as the intermediate meaning representation. In contrast, most previous work has focused on using bilingual corpora for language pairs such as English-French, English-Spanish and other Germanic and Romance languages in which the word order and grammar are similar, as shown in the example in Figure 1.

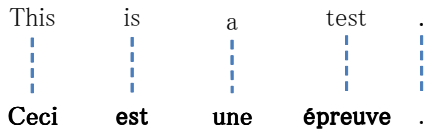

Fig. 1. Example of alignment for a language pair with similar word order and grammar (e.g., English-French).

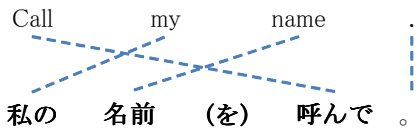

Fig. 2. Example of an alignment in English-Japanese.

In contrast, there is a large divergence in both the word order, and the grammatical structure between Japanese and English, as shown in Figure 2. We describe in Section 4 how we use a syntactic preprocessing method, Head Finalization [9], to help compensate for this difference.

As a target for this method, we collect Japanese-English bilingual data that is either in the public domain, or available under the Creative Commons license, as described in Section 5 and use it to create a broad-coverage and freely available Japanese paraphrase database.[1] Finally, we perform an analysis of the database in Section 6, and find that it achieves a standard of accuracy similar to that of previously reported paraphrasing resources.

2. EXTRACTING PARAPHRASES

We extract paraphrases using Bannard and Callison-Burch's bilingual pivoting method [7]. This method is a general-

[1] http://ahclab.naist.jp/resource/jppdb/

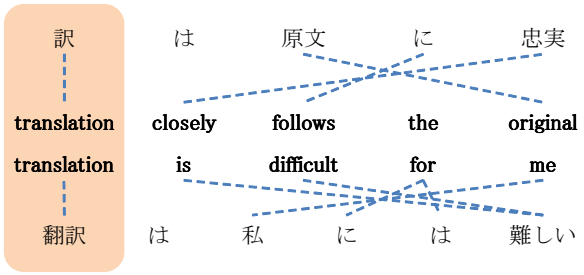

Fig. 3. Phrase-based paraphrases are extracted via bilingual pivoting.

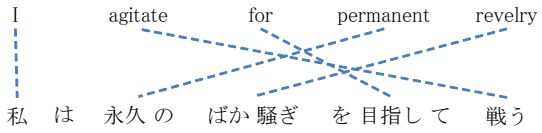

Fig. 4. Example of alignment in standard English-Japanese.

Fig. 5. Example of alignment in head-finalized English-Japanese.

purpose paraphrase extraction method, with the intuition that two English strings e_1 and e_2 that translate to the same foreign string f can often be assumed to have the same meaning.

In this work, instead of English, we extract paraphrases for Japanese, over Japanese-English bilingual parallel corpora. In Japanese-English, we can thus pivot over e and extract $\langle j_1, j_2 \rangle$ as a pair of paraphrases, as illustrated in Figure 3. We estimate the conditional paraphrase probability $P(j_2|j_1)$ by marginalizing over all shared English translations e:

$$P(j_2|j_1) = \sum_e P(j_2|e)P(e|j_1) \quad (1)$$

To calculate these pivoted pairs and probabilities, we need to calculate the conditional probabilities $P(j_2|e)$ and $P(e|j_1)$. This is done by first extracting phrase pairs from a Japanese-English corpus according to the standard phrase extraction algorithm of phrase-based statistical machine translation [10]. For example, translation probability $P(e|j)$ is calculated according to maximum likelihood estimation based on the counts of each phrase pair $c(e, j)$ and its constituent phrases:

$$P(e|j) = \frac{c(e, j)}{\sum_e c(e, j)} \quad (2)$$

This method has been shown to accurately extract a diverse set of paraphrases in past research[7].

3. SYNTACTIC PREPROCESSING

In order to use the previously described method, it is necessary to acquire phrase alignments in parallel corpus as pivots between English phrases. In general, these automatic alignments are produced in an unsupervised manner with the GIZA++ toolkit [11]. However, for languages with greatly different syntax and word order, standard alignment with GIZA++ has worse performance in comparison to languages with more similar syntax and word order.

In this paper, we help ameliorate this problem using recent syntactic preprocessing approaches to statistical machine translation. Specifically, we use the Head Finalization (HF; [9]) syntactic preprocessing method to change the English sentence to a syntactic structure similar to Japanese before running alignment and phrase extraction.

In English-Japanese translation, one of the most serious problems is the difference of grammar and word ordering between the two languages. Isozaki et al. proposed the HF preprocessing method, which helps reduce this syntactic divergence and increase the accuracy of statistical machine translation results [9].

The main element of the method lies in reordering English into a similar order to Japanese. This is done by using a syntactic parser to parse the English sentence, then moving the head of all phrases to the end of the phrase, transforming English into head-final order, similar to Japanese. In addition, the HF algorithm considers the fact that Japanese contains explicit case markers after the subject and object which are not present in English, and inserts pseudo-words that correspond to these particles into the head-finalized English. As these will be aligned into Japanese particles, it can be expected that performing this processing will also improve the accuracy of acquiring paraphrases for these particles. We show examples of alignment in standard English-Japanese and head-finalized English-Japanese in Figure 4 and 5.

4. PPDB : JAPANESE

Based on this data, we extracted Japanese paraphrases according to the proposed method.We extract alignments from a 1.9M sentence Japanese-English parallel corpus, the details of which are shown in Tables 1 and 2. For alignment, we use GIZA++, with the English side being pre-processed with HF as mentioned in the previous section. For tokenization, we use the Stanford Parser[2] [17] for English and KyTea[3] [18] for Japanese.

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] http://www.phontron.com/kytea/

Table 1. The details of corpus

	Sentences	Words (en)	Words (ja)	Licenses
Tanaka [12]	150k	1.4M	2.1M	Public Domain
AOZORA [13]	108k	1.6M	2.5M	Public Domain
Common Crawl [14]	821k	13.8M	22.0M	Public Domain
WWWJDIC [15]	373k	866k	373k	CC BY-SA 3.0[4]
Kyoto Wiki [16]	440k	11.5M	11.8M	CC BY-SA 3.0[4]
Total	1.9M	29.2M	38.3M	CC BY-SA 3.0[4]

Table 3. Examples of paraphrases with their rough English gloss

Seed	Paraphrases
メンバー	メンバ, 一族, 一員, 員, 会員, 加盟, 会員の
member	*member, family, a member (of), member, membership, member (join), member's*
魏志 倭人 伝 に	魏志 倭人 伝 の 記述 に, 魏志 倭人 に, 魏志 倭人 伝 に は
in the Gishi-wajin-den	*According to a description in the Gishi-wajin-den, in the Gishi-wajin, in Gishi-wajin-den*
論争	の 論争, 争議, 紛争, 争い, 討論, 議論
argue	*'s dispute, dispute, conflict, controversy, discussion, argue*
突如	とつぜん, 急 に, 不意 に, 突然
suddenly	*all of a sudden, hastily, abruptly, sudden*
で	に より, に よって, に よる, に, の
by	*because of, depending on, according to, to, 's*

Table 2. The details of the phrase table

Phrases	67.1M
Alignment	GIZA++ [11]
Tokenization (en)	Stanford Parser [17]
Tokenization (ja)	Kytea [18]
Max phrase length	7 words

In paraphrasing, we chose paraphrases where the conditional probability of the target is less than 1% to reduce the number of extracted paraphrases with low probability. The total number of extracted Japanese paraphrases were 10.5M pairs. We analyze and evaluate the paraphrases, the detailed results of which are below.

4.1. Analysis

In the analysis, we evaluated mean phrase length of paraphrases, number of paraphrases provided for single words (as opposed to phrases), number of paraphrases of particles, and a histogram of the phrase lengths of paraphrases.

The mean phrase length of paraphrases is 3.42 words. The number of paraphrases provided for words is 60.8k, and the number of paraphrases for phrases of length two or more is 986k. We were able to acquire 134 paraphrases of Japanese particles in this method. We show examples of the acquired paraphrases in Table 3.

We show a histogram of every phrase length of paraphrases in Figure 6.

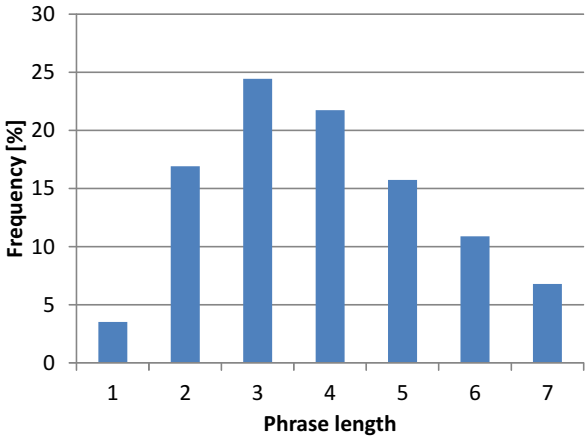

Fig. 6. Histogram of every phrase length in the acquired paraphrases.

4.2. Evaluation

We evaluate the paraphrases under the same conditions as previous work [7], with the details as follows. We substituted candidate paraphrases into 24 sentences which contained the original phrase, creating a total of 85 unique sentences through substitution. We had 3 native Japanese speakers produce judgments as to whether the new sentences preserved the meaning of the original phrase and as to whether they remained grammatical. Paraphrases that were judged to preserve both meaning and grammaticality were considered to be

Table 4. Evaluation of the acquired paraphrases

	Correct rate [%]
Meaning	84.7
Grammar	55.3
Both	45.8

correct, and examples which failed on either judgment were considered to be incorrect. We show results of the judgement in Table 4.

The accuracy of paraphrases is 45.8%, almost same with previous work [7]. The inter-annotator agreement for these judgements was measured at $\kappa = 0.60$ [19], which is conventionally interpreted as "moderate" agreement [20].

Ignoring the constraint that the new sentences remain grammatically correct, these paraphrases were judged to have the correct meaning 84.7% of the time. This indicates that the paraphrases are semantically correct, but may vary in their syntactic categories or contexts. These tendencies are similar to Bannard and Callison-Burch [7].

5. RELATED WORKS

With regards to related resources for paraphrasing created using bilingual data, there are English paraphrase data extracted by the same method [8]. In their work, they use English-French, English-Spanish and other European language bilingual parallel corpora. Finally, 16.7M paraphrases were extracted from 1G sentences of parallel data. Fujita et al. [21] also extract 28M Japanese paraphrases from 3.2M sentence pairs of Japanese-English patent translation data, although these paraphrases are limited to the patent domain and not publicly available as a resource.

Considering work on Japanese paraphrasing not limited to those extracted from parallel data, there is Japanese honorifics paraphrase data [3], action word paraphrases [4] and others. The Japanese honorifics paraphrase data offers 130k paraphrases gathered from 50k sentences of monolingual data covering honorifics. In the action word paraphrase data, 1.1k paraphrases covering action words such as verbs are included.

In these related works, various types of paraphrases are suggested, but there are few resources freely available. In addition, there are many paraphrase resources that have some kind of theme, but there are few large-scale and general paraphrase resources. In comparison, our paraphrase data is large, general, and freely available.

6. CONCLUSION

In the end, we were able to acquire 1.2M paraphrases from 1.9M sentences of our bilingual parallel corpus. Our paraphrase data is larger than some previous works that created Japanese paraphrases [3, 4]. And our proposed method was able to acquire paraphrase data that is large, covers several domains, and is high quality. We hope that our paraphrase data will be able to contribute to future studies that require paraphrases in the Japanese language.

[4]CC BY-SA 3.0 : Creative Commons Attribution-ShareAlike 3.0 Unported License

7. REFERENCES

[1] Chris Callison-Burch, Philipp Koehn, and Miles Osborne, "Improved statistical machine translation using paraphrases," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, New York City, USA, June 2006, pp. 17–24.

[2] Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu, "Natural language based reformulation resource and wide exploitation for question answering.," in *Proceedings of the 11th Text Retrieval Conference*, 2002.

[3] Kiyonori Ohtake and Kazuhide Yamamoto, "Paraphrasing honorifics," in *Automatic Paraphrasing: Theories and Applications (NLPRS2001 Post-Conference Workshop)*, 2001, pp. 13–20.

[4] Hiroshi Nakagawa and Hidetaka Masuda, "Extracting paraphrases of Japanese action word of sentence ending part from web and mobile news articles," in *Information Retrieval Technology*, vol. 3411 of *Lecture Notes in Computer Science*, pp. 94–105. Springer Berlin Heidelberg, 2005.

[5] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata, "Large scale relation acquisition using class dependent patterns," in *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 2009, pp. 764–769.

[6] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi, "Extracting paraphrases from definition sentences on the web," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, June 2011, pp. 1087–1097.

[7] Colin Bannard and Chris Callison-Burch, "Paraphrasing with bilingual parallel corpora," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 597–604.

[8] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch, "PPDB: The paraphrase database," in *Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, June 2013, pp. 758–764.

[9] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh, "Head finalization: A simple reordering rule for SOV languages," in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden, July 2010, pp. 244–251.

[10] Philipp Koehn, Franz Josef Och, and Daniel Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, Stroudsburg, PA, USA, 2003, The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 48–54.

[11] Franz Josef Och and Hermann Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[12] Yasuhito Tanaka, "Compilation of a multilingual parallel corpus," *Proceedings of the 2001 Conference of the Pacific Association for Computational Linguistics*, pp. 265–268, 2001.

[13] Masao Utiyama and Mayumi Takahashi, "English-Japanese translation alignment data," `http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/align/`.

[14] Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez, "Dirt cheap web-scale parallel text from the common crawl," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 1374–1383.

[15] Jim Breen, "WWWJDIC Japanese dictionary server user guide," `http://www.csse.monash.edu.au/~jwb/wwwjdicinf.html`.

[16] Masao Utiyama and Mayumi Takahashi, "Japanese-English bilingual corpus of Wikipedia's Kyoto articles," `http://alaginrc.nict.go.jp/WikiCorpus/index_E.html`.

[17] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng, "Parsing with compositional vector grammars," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2013.

[18] Graham Neubig, Yosuke Nakata, and Shinsuke Mori, "Pointwise prediction for robust, adaptable Japanese morphological analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, Stroudsburg, PA, USA, 2011, HLT '11, pp. 529–533.

[19] Joseph.L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[20] J. Richard. Landis and G. Gary Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[21] Atsushi Fujita, Pierre Isabelle, and Roland Kuhn, "Enlarging paraphrase collections through generalization and instantiation," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Stroudsburg, PA, USA, 2012, EMNLP-CoNLL '12, pp. 631–642.