Collocation or Free Combination? Applying Machine Translation Techniques to Identify Collocations in Japanese

Lis Pereira[1], Elga Strafella[2], Yuji Matsumoto[1]
[1]Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{lis-k,matsu}@is.naist.jp
[2]National Institute for Japanese Language and Linguistics, 10-2 Midoricho, Tachikawa, Tokyo 190-8561, Japan
strafelga@gmail.com

Abstract

This work presents an initial investigation on how to distinguish collocations from free combinations. The assumption is that, while free combinations can be literally translated, the overall meaning of collocations is different from the sum of the translation of its parts. Based on that, we verify whether a machine translation system can help us perform such distinction. Results show that it improves the precision compared with standard methods of collocation identification through statistical association measures.

Keywords: collocation, free combination, statistical machine translation.

1. Introduction

This work stems from a basic observation on collocational behaviour of words: the overall meaning of a collocation is not the sum of the individual meanings of its components, rather they have a slightly figurative connotation, which is difficult to guess without a previous specific knowledge of the observed phrase. For example, the meaning of the Japanese expression お茶を入れる [ocha-o-ireru] is not ‘to put in tea’ as someone who already knows the meaning of the noun お茶 [ocha] ‘tea’ and of the verb 入れる [ireru] ‘to put in’ might think. It rather means ‘to make tea’. This example shows an important feature of collocations, that is they cannot be translated literally (word by word). The same semantic feature (commonly referred to as ‘non-compositionality of the meaning’) is shown by another kind of multiword expressions, so-called idioms, the meaning of which is even more abstract and difficult to guess by a non-native speaker. On the other hand, natural languages are also made up of some frequent regular combinations (free combinations) of words whose meaning is nothing more than the combination of each word’s meaning, as for example, the expression お茶を飲む [ocha-o-nomu] ‘to drink tea’, which is the combination of the noun お茶 [ocha] ‘tea’ and the verb 飲む [nomu] ‘to drink’.

The main goal of this work-in-progress is to conduct an initial investigation on whether translation information can help us to distinguish collocations from free combinations. Given an expression, we use a machine translation system to predict whether its meaning can be derived from the translation of its parts. If not, that might be evidence that the original expression is a collocation (or an idiom). The results indicate that it improves the precision compared with standard methods of collocation identification through statistical association measures.

2. Related Work

Previous works on collocation identification (Evert, 2008; Seretan, 2011; Pecina, 2010; Ramisch, 2012) employ a standard methodology consisting of two steps: 1) candidate extraction, where candidates are extracted based on n-grams or morphosyntactic patterns and 2) candidate filtering, where association measures (AMs) are applied to rank the candidates based on association scores and consequently remove noise. Association measures assign an association score to each candidate pair. High association score indicates strong association, and can be used to identify the collocations among the recurrent word pairs found in a corpus (Stefan, 2008). One drawback of such method is that association measures might not be able to perform a clear-cut distinction between collocation and non-collocations, since they only assign scores based on statistical evidence, such as co-occurrence frequency in the corpus.

3. Collocation Identification

In our work, we focus on classifying Japanese noun-verb expressions into: free combination and collocation. Our approach consists of four steps:

1) For each candidate pair, we find all the possible translations of each Japanese word involved in the candidate (noun and verb), using a Japanese-English dictionary;

2) We then look for all the entries in the phrase-table (generated after training a Statistical Machine Translation (SMT) system using a Japanese-English bilingual corpus) that contain the candidate pair string and check if at least one of the possible literal translations appear as its corresponding translation. For instance, for the candidate pair 本を買う [hon-o-kau] ‘to buy a book’, we take the noun 本 and the verb 買う and check their translation given in the dictionary. 本 has translations such as ‘book’, ‘main’ and ‘head’, and 買う is translated as ‘to buy’. Based on that, one possible combination is ‘buy book’, or ‘buy main’. Therefore, we check in the phrase table whether 本を買う was translated as’ buy book’ or’ buy main’.

3) For the matched entries, we compute the average of the sum of the candidate’s direct and inverse phrase translation probability scores.

4) Finally, the candidates are ranked by the average score described in step 3.

4. Data Set

The following resources were used in our experiments:

1) **Bilingual Dictionary**: EDICT (Breen, 1999), a freely available Japanese/English Dictionary in machine-readable form, containing 110,424 entries, was used to find all the possible meanings of each Japanese word involved in the candidate (noun and verb). For our test set, all words were covered by the dictionary.

2) **Parallel Corpus**: we used Hiragana Times corpus, a Japanese-English bilingual corpus of magazine articles of Hiragana Times [1], a bilingual magazine written in Japanese and English to introduce Japan to non-Japanese, covering a wide range of topics (culture, society, history, politics, etc.). The corpus contain articles from 2003-2102, with a total of 117,492 sentence pairs. The Japanese data contains 3,949,616 words and the English data contains 2,107,613 words, as shown in Table 1.

	English	Japanese
#sentences	117,492	117,492
#tokens	2,107,613	3,949,616

Table 1: Statistics on the Hiragana Times corpus

Hiragana Times corpus was used to:

a) Extract noun-verb collocation candidate pairs using a dependency parser, Cabocha (Kudo and Matsumoto, 2002), applying different co-occurrence threshold values: $3 \leq f < 5$, $5 \leq f < 10$, $10 \leq f < 20$, $20 \leq f < 30$ and $f \geq 30$. In theory, any pair of words that co-occur at least twice in a corpus is a potential collocation. However, in order to reduce the enormous amounts of data that have to be processed, it is common to apply frequency thresholds (Stefan, 2008). The purpose of applying different threshold values was to verify from which threshold value a noun-verb pair should appear in the corpus so the MT system would be able to generate its translation. A total number of 8480 candidates were extracted. For evaluation, we selected, from the candidates extracted, a number of free combinations and true collocations from the candidates extracted, shown in Table 2.

b) Train a Japanese-English phrase-based SMT system, described in details in the next section.

Frequency Threshold	**Free Combinations**	**Collocations**
$3 \leq f < 5$	5	5
$5 \leq f < 10$	7	14
$10 \leq f < 20$	4	7
$20 \leq f < 30$	5	8
$f \geq 30$	7	9
Total	**28**	**43**

Table 2: Number of free combinations and true collocations from the candidates extracted from Hiragana Times corpus applying different co-occurrence threshold values.

5. Phrase-based SMT System

A standard non-factored phrase-based SMT system was built using the open source Moses toolkit (Koehn et al., 2007) with parameters set similar to those of Neubig (2011), who provides a baseline system previously applied to a Japanese-English corpus built from Wikipedia articles. For training, we used Hiragana Times bilingual corpus. The Japanese sentences were word-segmented and the English sentences were tokenized and lowercased. All sentences with size greater than 60 tokens were previously eliminated. The whole English corpus was used as training data for a 5-gram language model built with the SRILM toolkit (Stolcke, 2002). Similar to what we did for our proposed method, for each candidate in the test set, we find all the possible literally translated expressions (as described in Section 3). In the phrase-table generated after the training step, we look for all the entries that contain the original candidate string and check if at least one of the possible literal translations appear as their corresponding translation. For the entries found, we compute the average of the sum of the candidate's direct and inverse phrase translation probability scores. The direct phrase translation probability and the inverse phrase translation probability (Koehn et al., 2003) are respectively defined as:

$$\varphi(\bar{e}, \bar{f}) = \frac{\mathrm{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \mathrm{count}(\bar{f},\bar{e})} \qquad (1)$$

$$\varphi(\bar{f}, \bar{e}) = \frac{\mathrm{count}(\bar{f},\bar{e})}{\sum_{\bar{e}} \mathrm{count}(\bar{f},\bar{e})} \qquad (2)$$

Where $\bar{f}$ and $\bar{e}$ indicate a foreign phrase and a source phrase, independently. The candidates are ranked according to the average score as described previously.

6. Baseline System

For our baseline system, all selected candidate pairs were ranked according to three different association measures:

1) **Pointwise Mutual Information (PMI)** (Church and Hanks, 1990): it is an association measure related to Information Theory concepts and is perhaps the most widely used association measure in collocation extraction. It measures how often two words co-occur, compared with what we would expect if they occur independently.

2) **Log-likelihood ratio** (Dunning, 1993): this measure is argued to be appropriate when the data are sparse (Dunning 1993). It compares two hypotheses to determine which hypothesis is more likely to occur than the other. The first hypothesis proposes that two terms occur independently from each other, while the second hypothesis proposes that the occurrence of one of the terms is dependent on the occurrence of the other term.

3) **Weighted Dice** (Kitamura and Matsumoto, 1997): Weighted Dice coefficient has been reported to improve the performance of the Dice coefficient (Smadja et al. 1996), another commonly applied statistical measure. While in Dice coefficient the maximum value is 1, when

[1] http://www.hiraganatimes.com/

	Proposed Method	**Weighted Dice**	**PMI**	**Log-likelihood**	**AM's combined**
$3 \leq f < 5$	0.31	0.73	**0.91**	0.53	0.88
$5 \leq f < 10$	0.43	0.81	0.76	0.69	**0.82**
$10 \leq f < 20$	**1**	0.71	0.34	0.69	0.73
$20 \leq f < 30$	**1**	0.89	0.74	0.87	0.82
$f \geq 30$	0.75	0.64	**0.78**	0.76	0.61

Table 3 MAP values considering different co-occurrence frequency threshold values.

the pair always co-occurs, regardless of the frequency of the occurrence, Weighted Dice takes the absolute value of the co-occurrences into consideration.

We also compare with the score generated when we combine the ranks given by each Association Measure. The combined score is calculated by the Mean Reciprocal Rank (MRR), which is calculated as follows:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank(i)} \tag{3}$$

Where N is the number of Association Measures used.

7. Evaluation

In our evaluation, we average the precision considering all true collocations and idioms as threshold points, obtaining the mean average precision (MAP). Differently from the traditional approach used to evaluate an association measure, using MAP we do not need to set a hard threshold.

8. Results

Table 3 shows the MAP values for the baseline and proposed method. It shows that, in general, association measures perform quite well for all threshold values, while our proposed method performs poorly for low frequency candidates (frequency less than 10). Due to the low number of occurrences in the corpus, the SMT system was not able to produce acceptable translations for such candidates. On the other hand, for candidates that occurred in the corpus more than 10 times, the system was able to classify the candidates into free combination or collocation with high accuracy, outperforming the baseline systems for candidates that occurred more than 10 times and less than 30 times in the corpus. Although the baseline systems assigned most true collocations with high scores, some free combinations were assigned high scores as well, and it was not able to perform a clear separation into collocations and non-collocations. For candidates that occurred more than 30 times in the corpus, the baseline systems obtained a slightly better performance.

9. Discussion

Cases where our proposed method could not handle correctly were due to three main reasons:

1) **Data sparseness**: the MT system was not able to assign correct translations for many cases, especially low frequent candidates. Moreover, the coverage of the corpus was not sufficient for finding the translations generated using the bilingual dictionary, for many test instances. A larger parallel corpus might help, although it is a quite scarce resource for the Japanese/English pair.
2) **Collocations that can be literally translated:** For instance, in Japanese, there is the collocation 責任を負う [sekinin-o-ou] 'to bear the responsibility', where the literal translation of the noun 責任 [sekinin] 'responsibility' and the verb 負 [ou] 'to bear' correspond to the translation of the expression as well.
3) **Expressions that have both literal and non-literal meaning**: For instance, the collocation 目に入る [me-ni-hairu] can mean 'to enter the eye', which is the literal meaning (目 means 'eye' and 入る means 'to enter'), but it can also can mean 'to come into view'.

10. Conclusion

In this paper, we conducted an initial investigation on how to distinguish collocations from free combinations in Japanese language. The assumption is that, while free combinations can be translated literally, in true collocations the overall meaning slightly differs from the meaning of the parts. Based on that, we built an SMT system that can help classify noun-verb pairs into free combination or true collocation. Our system could classify the candidates into free combination or collocation with high accuracy for candidates that appeared 10 times or more in the data. Although the method was tested on a small test set, we believe that this can be a first step in order to speed up the work of lexicographers in developing resources, for instance.

In order to verify our approach, our next steps will be to evaluate on a larger test data and to check if the literal translation of a candidate pair exists directly using a large scale corpus of English, due to the lack of large parallel corpus for the Japanese/English pair. In addition, although our method made no distinction between collocations and idioms, it may be possible to make such distinction, since idioms are to be translated completely in non-literal expressions while collocations can be translated partially into literal expressions, and we plan to conduct experiments to verify it.

11. Acknowledgements

We would like to thank Dr. Kevin Duh for his valuable help and advice.

12. References

Jim Breen. 1995. Building an electronic Japanese-english dictionary. In Japanese Studies Association of Australia Conference. Citeseer.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. Computational linguistics 16, 1 (1990), 22–29.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. Computational linguistics 19, 1 (1993), 61–74.

Stefan Evert. 2008. Corpora and collocations. Corpus Linguistics. An International Handbook, 2.

Mihoko Kitamura and Yuji Matsumoto. 1997. Automatic extraction of translation patterns in parallel corpora. Transactions of IPSJ 38, 4 (1997), 727–735.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. 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, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180. Association for Computational Linguistics.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In proceedings of the 6th conference on Natural language learning-Volume 20, pages 1–7. Association for Computational Linguistics.

Graham Neubig. 2011. The kyoto free translation task. Available on line at http://www. phontron. com/kftt.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. Language resources and evaluation, 44(1-2):137–158.

Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: from acquisition to applications. In Proceedings of ACL 2012 Student Research Workshop, pages 61–66. Association for Computational Linguistics.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. Computational linguistics 22, 1 (1996), 1–38.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In INTERSPEECH.