Identifying collocations using cross-lingual association measures

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

Abstract

We introduce a simple and effective cross-lingual approach to identifying collocations. This approach is based on the observation that true collocations, which cannot be translated word for word, will exhibit very different association scores before and after literal translation. Our experiments in Japanese demonstrate that our cross-lingual association measure can successfully exploit the combination of bilingual dictionary and large monolingual corpora, outperforming monolingual association measures.

1 Introduction

Collocations are part of the wide range of linguistic phenomena such as idioms (*kick the bucket*), compounds (*single-mind*) and fixed phrases (*by and large*) defined as Multiword Expressions (MWEs). MWEs, and collocations, in particular, are very pervasive not only in English, but in other languages as well. Although handling MWEs properly is crucial in many natural language processing (NLP) tasks, manually annotating them is a very costly and time consuming task.

The main goal of this work-in-progress is, therefore, to evaluate the effectiveness of a simple cross-lingual approach that allows us to automatically identify collocations in a corpus and subsequently distinguish them according to one of their intrinsic properties: the meaning of the expression cannot be predicted from the meaning of the parts, i.e. they are characterized by limited compositionality (Manning and Schütze, 1999). Given an expression, we predict whether the expression(s) resulted from the word by word translation is also commonly used in another language. If not, that might be evidence that the original expression is a collocation (or an idiom). This can be captured by the ratio of association scores, assigned by association measures, in the target vs. source language. The results indicate that our method improves the precision comparing with standard methods of MWE identification through monolingual association measures.

2 Related Work

Most previous works on MWEs and, more specifically, 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 are applied to rank the candidates based on association scores and consequently remove noise. 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. Our cross-lingual association measure ameliorates this problem by exploiting both corpora in two languages, one of which may be large.

A few studies have attempted to identify non-compositional MWE's using parallel corpora and dictionaries. Melamed (1997) investigates how non-compositional compounds can be detected from parallel corpus by identifying translation divergences in the component words. Pichotta and DeNero (2013) analyses the frequency statistics of an expression and its component words, using many bilingual corpora to identifying phrasal verbs in English. The disadvantage of such approach is that large-scale parallel corpora is available for only a few language pairs. On the other hand, monolingual data is largely and freely available for many languages. Our approach requires only a bilingual dictionary and non-parallel monolingual corpora in both languages.

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 109–113,
Gothenburg, Sweden, 26-27 April 2014.

Salehi and Cook (2013) predict the degree of compositionality using the string distance between the automatic translation into multiple languages of an expression and the individual translation of its components. They use an online database called Panlex (Baldwin et al., 2010), that can translate words and expressions from English into many languages. Tsvetkov and Wintner (2013) is probably the closest work to ours. They trained a Bayesian Network for identfying MWE's and one of the features used is a binary feature that assumes value is 1 if the literal translation of the MWE candidate occurs more than 5 times in a large English corpus.

3 Identifying Collocations

In this research, we predict whether the expression(s) resulted from the translation of the components of a Japanese collocation candidate is/are also commonly used in English. For instance, if we translate the Japanese collocation 面倒を見る *mendou-wo-miru* "to care for someone" (care-を-see)[1] into English word by word, we obtain "see care", which sounds awkward and may not appear in an English corpus very often. On the other hand, the word to word translation of the free combination 映画を見る *eiga-wo-miru* "to see a movie" (movie-を-see) is more prone to appear in an English corpus, since it corresponds to the translation of the expression as well. In our work, we focus on noun-verb expressions in Japanese. Our proposed method consists of three steps:

1) Candidate Extraction: We focus on noun-verb constructions in Japanese. We work with three construction types: object-verb, subject-verb and dative-verb constructions, represented respectively as "noun wo verb (noun-を-verb)", "noun ga verb (noun-が-verb)" and "noun ni verb (noun-に-verb)", respectively. The candidates are extracted from a Japanese corpus using a dependency parser (Kudo and Matsumoto, 2002) and ranked by frequency.

2) Translation of the component words: for each noun-verb candidate, we automatically obtain all the possible English literal translations of the noun and the verb using a Japanese/English dictionary. Using that information, all the possible verb-noun combinations in English are then generated. For instance, for the candidate 本を買う *hon-wo-kau* "to buy a book" (buy-を-book), we take the noun 本 *hon* and the verb 買う *kau* and check their translation given in the dictionary. 本 has translations like "book", "main" and "head" and 買う is translated as "buy". Based on that, possible combinations are "buy book" or "buy main" (we filter out determiners, pronouns, etc.).

3) Ranking of original and derived word to word translated expression: we compare the association score of the original expression in Japanese (calculated using a Japanese corpus) and its corresponding derived word to word translated expressions. If the original expression has a much higher score than its literal translations, it might be a good evidence that we are dealing with a collocation, instead of a free combination.

There is no defined criteria in choosing one particular association measure when applying it in a specific task, since different measures highlight different aspects of collocativity (Evert, 2008). A state-of-the-art, language independent framework that employs the standard methodology to identify MWEs is mwetoolkit (Ramisch, 2012). It ranks the extracted candidates using four different association measures: log-likelihood-ratio, Dice coefficient, pointwise mutual information and Student's *t*-score. We previously conducted experiments with these four measures for Japanese (results are ommited), and Dice coefficient performed best. Using Dice coefficient, we calculate the ratio between the score of the original expression and the average score of its literal translations. Finally, the candidates are ranked by the ratio value. Those that have a high value are expected to be collocations, while those with a low value are expected to be free combinations.

4 Experiment Setup

4.1 Data Set

The following resources were used in our experiments:

Japanese/English dictionary: we used Edict (Breen, 1995), a freely available Japanese/English Dictionary in machine-readable form, containing 110,424 entries. This dictionary was used to find all the possible translations of each Japanese word involved in the candidate (noun and verb). For our test set, all the words were covered by the dictionary. We obtained an average of 4.5 translations per word. All the translations that contains more

[1] In Japanese, を is a case marker that indicates the object-verb dependency relation.

than three words are filtered out. For the translations of the Japanese noun, we only consider the first noun appearing in each translation. For the translations of the Japanese verb, we only consider the first verb/phrasal verb appearing in each translation. For instance, in the Japanese collocation 恋に落ちる *koi-ni-ochiru* "to fall in love" (love-に-fall down)[2], the translations in the dictionary and the ones we consider (shown in bold type) of the noun 恋 *koi* "love" and the verb 落ちる *ochiru* "to fall down" are:

恋: **love** , tender **passion**
落ちる: to **fall down**, to **fail**, to **crash**, to **degenerate**, to **degrade**

Bilingual resource: we used Hiragana Times corpus, a Japanese-English bilingual corpus of magazine articles of Hiragana Times [3], 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 contains articles from 2003-2102, with a total of 117,492 sentence pairs. We used the Japanese data to extract the noun-verb collocation candidates using a dependency parser, Cabocha (Kudo and Matsumoto, 2002). For our work, we focus on the object-verb, subject-verb and dative-verb dependency relations. The corpus was also used to calculate the Dice score of each Japanese candidate, using the Japanese data.

Monolingual resource: we used 75,377 English Wikipedia articles, crawled in July 2013. It contains a total of 9.5 million sentences. The data was used to calculate the Dice score of each candidate's derived word to word translated expressions. The corpus was annotated with Part-of-Speech (POS) information, from where we defined POS patterns to extract all the verb-noun and noun-verb sequences, using the MWE toolkit (Ramisch, 2012), which is an integrated framework for MWE treatment, providing corpus preprocessing facilities.

Table 1 shows simple statistics on the Hiragana Times corpus and on the Wikipedia corpus.

	Hiragana Times	Wikipedia
# jp sentences	117,492	-
# en sentences	117,492	9,500,000
# jp tokens	3,949,616	-
# en tokens	2,107,613	247,355,886
# jp noun-verb	31,013	-
# en noun-verb	-	266,033
# en verb-noun	-	734,250

Table 1: Statistics on the Hiragana Times corpus and Wikipedia corpus, showing the number of sentences, number of words and number of noun-verb and verb-noun expressions in English and Japanese.

4.2 Test set

In order to evaluate our system, the top 100 frequent candidates extracted from Hiragana Times corpus were manually annotated by 4 Japanese native speakers. The judges were asked to make a ternary judgment for each of the candidates on whether the candidate is a collocation, idiom or free combination. For each category, a judge was shown the definition and some examples. We defined collocations as all those expressions where one of the component words preserves its literal meaning, while the other element assumes a slightly different meaning and its use is blocked (i.e. it cannot be substituted by a synonym). Idioms were defined as the semantically and syntactically fixed expressions where all the component words loose their original meaning. Free combinations were defined as all those expressions frequently used where the components preserve their literal meaning. The inter-annotator agreement is computed using Fleiss' Kappa statistic (Fleiss, 1971), since it involves more than 2 annotators. Since our method does not differentiate collocations from idioms (although we plan to work on that as future work), we group collocations and idioms as one class. We obtained a Kappa coefficient of 0.4354, which is considered as showing *moderate* agreement according to Fleiss (1971). Only the candidates identically annotated by the majority of judges (3 or more) were added to the test set, resulting in a number of 87 candidates (36 collocations and 51 free combinations). After that, we obtained a new Kappa coefficient of 0.5427, which is also considered as showing *moderate* agreement (Fleiss, 1971).

4.3 Baseline

We compare our proposed method with two baselines: an association measure based system and a Phrase-Based Statistical Machine Translation

[2] に is the dative case marker in Japanese.
[3] http://www.hiraganatimes.com

(SMT) based system.

Monolingual Association Measure: The system ranks the candidates in the test set according to their Dice score calculated using the Hiragana Times Japanese data.

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:

$$\Phi(\overline{e}|\overline{f}) = \frac{count(\overline{f},\overline{e})}{\sum_{\overline{f}} count(\overline{f},\overline{e})} \quad (1)$$

$$\Phi(\overline{f}|\overline{e}) = \frac{count(\overline{f},\overline{e})}{\sum_{\overline{e}} count(\overline{f},\overline{e})} \quad (2)$$

Where $\overline{f}$ and $\overline{e}$ indicate a foreign phrase and a source phrase, independently.

The candidates are ranked according to the average score as described previously.

5 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.

Table 2 presents the MAP values for our proposed method and for the two baselines. Our cross-lingual method performs best in terms of MAP values against the two baselines. We found out that it performs statistically better only compared to the Monolingual Association Measure baseline[4]. The Monolingual Association Measure baseline performed worst, since free combinations were assigned high scores as well, and the system was not able to perform a clear separation into collocations and non-collocations. The Phrase-Based SMT system obtained a higher MAP value than Monolingual Association measure, but the score may be optimistic since we are testing in-domain. One concern is that there are only a very few bilingual/parallel corpora for the Japanese/English language pair, in case we want to test with a different domain and larger test set. The fact that our proposed method outperforms SMT implies that using such readily-available monolingual data (English Wikipedia) is a better way to exploit cross-lingual information.

Method	MAP value
Monolingual Association Measure	0.54
Phrase-Based SMT	0.67
Proposed Method	**0.71**

Table 2: Mean average precision of proposed method and baselines.

Some cases where the system could not perform well include those where a collocation can also have a literal translation. For instance, in Japanese, there is the collocation 心を開く *kokoro-wo-hiraku* "to open your heart" (heart-を-open), where the literal translation of the noun 心 *kokoro* "heart" and the verb 開く *hiraku* "open" correspond to the translation of the expression as well.

Another case is when the candidate expression has both literal and non-literal meaning. For instance, the collocation 人を見る *hito-wo-miru* (person-を-see) can mean "to see a person", which is the literal meaning, but when used together with the noun 目 *me* "eye", for instance, it can also can mean "to judge human character". When annotating the data, the judges classified as idioms some of those expressions, for instance, because the non-literal meaning is mostly used compared

[4]Statistical significance was calculated using a two-tailed t-test for a confidence interval of 95%.

with the literal meaning. However, our system found that the literal translated expressions are also commonly used in English, which caused the performance decrease.

6 Conclusion and Future Work

In this report of work in progress, we propose a method to distinguish free combinations and collocations (and idioms) by computing the ratio of association measures in source and target languages. We demonstrated that our method, which can exploit existing monolingual association measures on large monolingual corpora, performed better than techniques previously applied in MWE identification.

In the future work, we are interested in increasing the size of the corpus and test set used (for instance, include mid to low frequent MWE's), as well as applying our method to other collocational patterns like Noun-Adjective, Adjective-Noun, Adverb-Verb, in order to verify our approach. We also believe that our approach can be used for other languages as well. We intend to conduct a further investigation on how we can differentiate collocations from idioms. Another step of our research will be towards the integration of the acquired data into a web interface for language learning and learning materials for foreign learners as well.

Acknowledgments

We would like to thank Masashi Shimbo, Xiaodong Liu, Mai Omura, Yu Sawai and Yoriko Nishijima for their valuable help and anonymous reviewers for the helpful comments and advice.

References

Timothy Baldwin, Jonathan Pool, and Susan M Colowick. 2010. Panlex and lextract: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 37–40. Association for Computational Linguistics.

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

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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.

Christopher D Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*, volume 999. MIT Press.

Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. EMNLP.

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.

Karl Pichotta and John DeNero. 2013. Identifying phrasal verbs using many bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 636–646.

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.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages.

Violeta Seretan. 2011. *Syntax-based collocation extraction*, volume 44. Springer.

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

Yulia Tsvetkov and Shuly Wintner. 2013. Identification of multi-word expressions by combining multiple linguistic information sources.