Extracting Bilingual Multi-word Terms from Comparable Corpora

Jun Liang Xiaodong Liu Kevin Duh Yuji Matsumoto
Nara Institute of Science and Technology
{liang.jun.lb5, xiaodong-l, kevinduh, matsu}@is.naist.jp

1 Introduction

Bilingual lexicon plays a very important role in many cross-lingual natural language processing tasks, such as cross-language information retrieval (CLIR) [1] and statistical machine translation (SMT) [2, 3]. Usually, bilingual lexicon is extracted from parallel corpora. But the resource of parallel corpora is only available for some language pairs or domains. Extracting bilingual lexicon from large-scaled comparable corpora becomes an attractive task in this field. Currently, there are two main Distributional Hypothesis [4] based methods, which are used to extract bilingual lexicon from comparable corpora. One is topic model based method (TMBM) [5, 6], the other is context based method (CBM) [7]. Both of them are mainly used to extract single-word (single-word term) pairs. While TMBM calculates word similarity based on the distribution of sharing the same topic information, CBM calculates word similarity based on the distribution of sharing the same context information.

However, we believe single-word term translation pairs are insufficient for many practical cross-lingual applications. Furthermore, constraining TMBM and CBM to operate on the word level risks losing important information because translations are not always word-for-word. For example, phrasal verbs and prepositions do not give their whole meaning without their related words.

Our proposed approach focus on multi-word term extraction using a hybrid method combining the TMBM approach of [6] with the C-value method [8], extracting multi-word term pairs using a word-alignment method. The key points of the combination are as follows:

1. C-value uses linguistic filters to extract the candidate multi-word terms from comparable corpora [8].
2. Multi-lingual Topic Model (MTM) [9] extracts semantic clusters of the multi-word terms sharing the same topic information [10] from the large-scale comparable corpora without using any seed lexicon and prior knowledge [5].

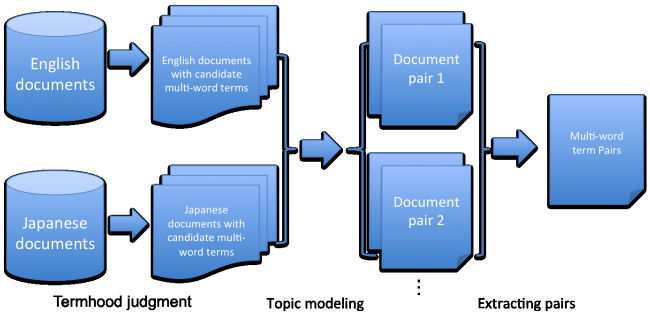

Figure 1: Process of bilingual lexicon extraction

3. Word-alignment is run on MTM results to extract the multi-word pairs sharing the same topic information.

Our approach does not require any prior knowledge; it improves the precision of multi-word term pairs and also proves the topic model can be used to deal with multi-word term problem. Experimental result shows that our proposed method on English-Japanese Kyoto Wikipedia corpora achieves good precision in extracting multi-word terms.

2 Proposed Method for Bilingual Lexicon Extraction

The process of our proposed approach is shown in Figure 1. Firstly, we use C-value to get the candidate multi-word terms from English corpus and Japanese corpus individually. Second, we run a MTM topic modeling method to get topic-aligned multi-word term-based corpora. Finally, we apply a word-alignment algorithm on the topic-aligned multi-word term-based corpora to extract the multi-word term pairs.

In the process of multi-word term pair extraction, we use word alignment algorithm of IBM Model 1.

2.1 C-value Method

In this section, we show how to use C-value method to extract multi-word terms. C-value combines two parts together, the linguistic part and statistical part; the linguistic part is based on the part-of-speech tagging and stop-word list for each language. The statistical part uses an algorithm to measure the termhood of every candidate multi-word term.

2.1.1 The Linguistic Part

In the linguistic part, firstly we do part-of-speech tagging. After tagging, we use linguistic filters to get the candidate multi-word terms. Finally, stop words are deleted based on an existing stop word list.

Depending on languages, we use different features and apply different linguistic filters [11, 12]. For English, we apply Stanford POS Tagger[1]; for Japanese, we apply KyTea[2].

For English, there are three different filters:

1. $Noun^+\ Noun$

   $Ex. kyoto + protocol$

2. $(Adj^+Noun)^+Noun$

   $Ex. actual + time$,
   $northern + imperial + court$,
   $short + preparation + time$

3. $((Adj \mid Noun)^+ \mid ((Adj \mid Noun)^*$
   $(NounPre)?\ (Adj \mid Noun)^*)\ Noun$

   $Ex. south + korean + presidential + committee$,
   $kyoto + international + university + tanabe +$
   $central + hospital$

For Japanese, we also use three different linguistic filters:

1. $Noun\{2,\}$

   $Ex.$ 奈良 + 時代 ($nara\ period$)

2. $(Prefix \mid Adv)(Noun \mid Adj \mid Suffix)^+\ Noun^+$

   $Ex.$ 回転 + 寿司 ($conveyor\ belt\ sushi$),
   第 +16+ 師団 ($16th\ division$)

3. $Prefix\ Noun^+\ Suffix$

   $Ex.$ 再 + 初期 + 化 ($re - initialize$),
   未 + 定義 + 型 ($undefined\ type$)

2.1.2 The Statistical Part

The algorithm of termhood, called C-value, is given as follows:

$$C\text{-}value(a)=\begin{cases}\log_2|a|\cdot f(a) & a\ is\ not\ nested,\\ \log_2|a|\left(f(a)-\frac{1}{Z(T_a)}\sum_{b\in T_a}f(b)\right) & otherwise\end{cases} \quad (1)$$

[1]http://nlp.stanford.edu/software/tagger.shtml
[2]http://www.phontron.com/kytea/

Where
a: candidate string.
$|a|$: length of the candidate string a.
$f(.)$: its frequency of occurrence in the corpus.
T_a: all the candidate terms that contain a.
$Z(T_a)$: the number of all candidate terms.

C-value is based on the frequency of the occurrence of the candidate string. Logarithm on the length of the candidate string is used to moderate the influence of it. The independence of a candidate string depends on the frequency of this candidate string being obtained by other longer candidate terms. The greater $f(b)$ is, the bigger a's independency is (and vice versa).

2.2 Multi-lingual Topic Model (MTM)

In our approach, we use multilingual topic model proposed by [9], which is based on monolingual Latent Dirichlet Allocation model [13]. After applying the C-value method, we run MTM on the comparable corpora to extract topic distributions for each single-word term or multi-word term. This part is essentially the same as the method of [6], with the exception that bag-of-words input is augmented with multi-word terms. Because MTM is often used to deal with single word term, here we use "_" to combine more than one word together in a multi-word term.

We have three system settings:

In "Combined" system, we combined the words (single-word terms) extracted from word segmentation with multi-word terms extracted from C-value method together.

In "Modified C-value" system, we extracted the defined single-word terms and multi-word terms using the modified C-value algorithm. We changed the $\log_2|a|$ part of original C-value algorithm into $\log_2|a+1|$ as modified C-value algorithm [14] to get the defined candidate single-word terms to do comparison with original C-value method. The other parts are the same as equation (1).

In "C-value" system, we extracted all the defined multi-word terms in English and Japanese without any single-word term.

2.3 Word Alignment

After MTM, we get a topic-aligned corpus (e, f). Here $e = [w_1^e, w_2^e, ..., w_e^e]$ is a set of $|e|$ English terms in a topic. And $f = [w_1^f, w_2^j, ..., w_f^f]$ is a set of $|f|$ Japanese terms in the same topic. We give English multi-word term index by i, give Japanese multi-word term index by j to notate the multi-word terms. We can get two multi-word terms and

Table 1: 100 Term (single-word term and multi-word term) Pairs Precision on "Combined" and "Modified C-value" Systems

Method	K(Topic)	Iteration	Precision
Combined	100	500	12%
Modified C-value	100	500	12%
Combined	400	500	34%
Modified C-value	400	500	**37%**
Combined	400	1,000	**50%**
Modified C-value	400	1,000	33%

the probabilities through the word-alignment function $a : i \rightarrow j$ mapping terms in e to terms in f (and vice versa).

3 Experiment

We evaluate our proposed method on comparable Japanese-English Kyoto Wiki corpora of [6], which is prepared by crawling the English documents according to the Japanese documents in Kyoto Wiki Corpus[3], where all the English documents are guaranteed to have the corresponding Japanese documents. We assign the document IDs on the same topic via the interlanguage links and choose 4090 Japanese-English document pairs as our training data.

3.1 Parameter Setting

To get higher probability candidate multi-word terms, the threshold of C-value is assigned to 2.0.

We set the hyper parameters of Dirchlet distribution $\alpha = 50/K$, and $\beta = 0.01$ following [5], where K denotes topic number. Here, we trained the multilingual topic model using Gibbs sampling with 500 and 1000 iteratiions.

3.2 Evaluation Criterion

We use Combined method, Modified C-value method, and C-value method to extract all the candidate term (single-word and multi-word) pairs.

After word alignment, we randomly chose 100 terms (include single-word terms and multi-word terms) combining both English-Japanese translation result and Japanese-English translation result.

We get the final translation probability, which is larger than 0.5, by calculating the average value of term pair's translation probabilites from English to Japanese and Japanese to English.

3.3 Result

Table 1 shows the manually-judged precision of randomly extracted 100 term (single-word and multi-word term) pairs of "Combined" system and "Modified C-value" system and Table 2 shows the precision of randomly extracted 100 whole multi-word term pairs per system.

[3]http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

Table 2: 100 Multi-word Term Pairs-only Precision on Three Different Systems

Method	K(Topic)	Iteration	Precision
Combined	100	500	1%
Modified C-value	100	500	10%
C-value	100	500	**14%**
Combined	400	500	9%
Modified C-value	400	500	28%
C-value	400	500	**31%**
Combined	400	1,000	11%
Modified C-value	400	1,000	27%
C-value	400	1,000	**36%**

In "Combined" system, by checking the results of Table 1 and Table 2, which shows that single-word term works better than multi-word term because single-word term's occurrence is higher than multi-word term's, and singl-word term influences multi-word term.

In "Modified C-value" system, some single-word terms in Japanese can be translated by multi-word terms in English (and vice versa). E.g. "豆乳" in Japanese is corresponding to "soy_milk" in English. Also "b.subtilis" in English is corresponding to "納豆_菌" in Japanese. In Table 1, under the condition of K=400, iteration=500, the precision is the highest.

In "C-value" system, we extracted all the defined multi-word terms in English and Japanese without any single-word term. The result in Table 2 shows the precision under "C-value" system is the highest.

Table 3: Part of Extracted Multi-word Terms from C-value

Japanese	English
大日如来	wisdom buddhas
アルコール度数	alcohol content
勧善懲悪	poetic justice
イセエビ	japanese spiny lobster
水墨画	ink wash painting

Finally, Table 3 gives examples of the extracted multi-word terms in "C-value" system.

4 Conclusion

We propose a method to extract multi-word term translations from comparable corpora. The idea is to combine the C-value method for termhood likelihood with the Multilingual Topic Model (MTM) approach to lexicon extraction. We compare three different

settings for our hybrid approach and show 36% precision for multi-word term pairs and 50% precision for mixed multi- and single-word term pairs.

For future work we plan to use this approach on scientific-oriented medical domain corpora, where multi-word terms play more important role and have a more specific meaning. We also plan to extend this approach to different languages, such as Chinese, Spanish, and French.

To augment both termhood and unithood of the extracted candidate multi-word terms[15], we are now working on combining context information and topic information together.

5 Acknowledgments

We thank Masashi Shimbo, Hiroyuki Shindo, Erlyn Manguilimotan, Philip Arthur, Yutaro Shigeto, Mai Omura, Yuki Tawara and the anonymous reviewers for valuable discussions and comments.

References

[1] Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4:209–230, 2001.

[2] Hal Daume III and Jagadeesh Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412. Association for Computational Linguistics, 2011.

[3] Hua Wu, Haifeng Wang, and Chengqing Zong. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 993–1000. Coling 2008 Organizing Committee, 2008.

[4] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[5] Ivan Vulic, Wim De Smet, Marie-Francine Moens, and KU Leuven. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 479–484, 2011.

[6] Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus.

[7] Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, College Park, Maryland, USA, June 1999. Association for Computational Linguistics.

[8] Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii. The c-value/nc-value method of automatic recognition for multi-word terms. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, pages 585–604, London, UK, UK, 1998. Springer-Verlag.

[9] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[10] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[11] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms: the c-value/nc-value method, 2000.

[12] Hideki Mima and Sophia Ananiadou. An application and evaluation of the c/nc-value approach for the automatic term recognition of multi-word units in japanese. *Terminology*, 6(2):175–194, 2001.

[13] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[14] Alberto Barrón-Cedeño, Gerardo Sierra, Patrick Drouin, and Sophia Ananiadou. An improved automatic term recognition method for spanish. In Alexander F. Gelbukh, editor, *CICLing*, volume 5449 of *Lecture Notes in Computer Science*, pages 125–136. Springer, 2009.

[15] Thuy Vu, Ai Ti Aw, and Min Zhang. Term extraction through unithood and termhood unification. In *In Proc. of Int 'l Joint Conf on Natural Language Proc*, 2008.