Bilingual Multi-Word Term Tokenization for Chinese–Japanese Patent Translation

Wei YANG and Yves LEPAGE
Graduate School of Information, Production and Systems, Waseda University
kevinyoogi@akane.waseda.jp; yves.lepage@waseda.jp

Abstract

We propose to re-tokenize data with aligned bilingual multi-word terms to improve statistical machine translation (SMT) in technical domains. For that, we independently extract multi-word terms from the monolingual parts of the training data. Promising bilingual multi-word terms are then identified using the sampling-based alignment method by setting some threshold on translation probabilities. We estimate that the bilingual multi-word terms extracted are correct in more than 70 % of the cases. We report a significant improvement of BLEU scores in experiments conducted on a Chinese–Japanese patent corpus.

1 Introduction

Making patents in one language available in other languages is crucial for sharing technology and increasing economical competitiveness in a global society. The crucial process of translation of patents should be helped by the use of statistical machine translation (SMT). The identification and translation of scientific or technical terms in patent corpora is thus a challenge for machine translation.

There exist previous work on extracting scientific or technical terms in different languages and different domains for applications like information retrieval, text categorization and also for machine translation. As an important milestone in terminology extraction, [1] describes a combination of linguistic filtering and statistical measure (C-value/NC-value) for the automatic extraction of multi-word terms from English scientific or technical texts. As an application for estimating the similarity of scientific papers, [5] shows how to extract English terms in the computer science and the medical domains using a C-value/NC-value extraction method. They make use of these terms to estimate the similarity of scientific papers in the vector space model. From these previous works, we can see that the C-value is commonly used as a domain-independent method for multi-word term extraction. As for language independence, it was shown in [6] that the C-/NC-value method is an efficient domain-independent multi-word term extraction technique not only in English but in Japanese as well.

In this paper, we adopt the C-value method to extract monolingual multi-word terms in Chinese and Japanese independently. We then apply an alignment technique, the sampling-based alignment method [4], on the re-tokenized Chinese–Japanese training corpus with monolingual multi-word terms to extract aligned candidate terms. We perform SMT experiments using the Chinese–Japanese experimental data re-tokenized again using the filtered bilingual multi-word terms. We obtain a better translation accuracy.

2 Extraction of Chinese–Japanese Bilingual Multi-Word Terms

2.1 Monolingual Multi-Word Term Extraction

The C-value is a widely used domain-independent approach for multi-word extraction. It combines a linguistic component and a statistical component. The advantage of the C-value is that it can compute multi-word terms made up of complex structures even when these structures have a low frequency. As for the linguistic component, in our experiments, for both Chinese and Japanese, we monolingually extract multi-word terms which contain a sequence of nouns or adjectives followed by a noun. This linguistic pattern can be written as follows using a regular expression[1]: (Adjective | Noun $)^+$ Noun.

The statistical component, the measure of termhood,

[1] Technically, for Chinese: (JJ | NN $)^+$ NN; for Japanese: (形容詞 | 名詞 $)^+$ 名詞. 'JJ' and '形容詞' are POS codes for adjectives, 'NN' and '名詞' are POS codes for nouns in the Chinese and the Japanese annotated corpora that we use.

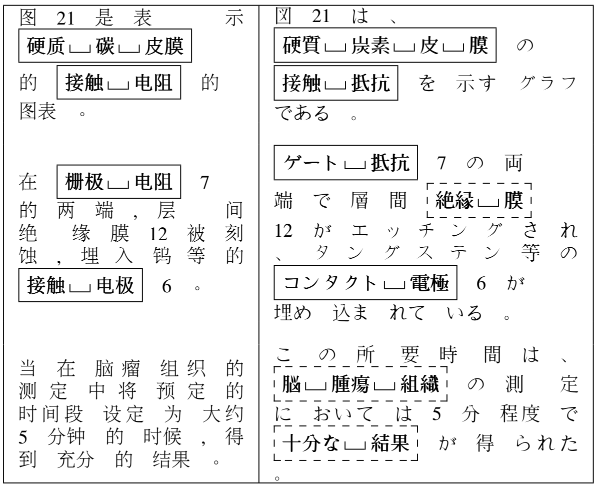

图 21 是 表 示 硬质␣碳␣皮膜 的 接触␣电阻 的 图表 。	図 21 は 、 硬質␣炭素␣皮␣膜 の 接触␣抵抗 を 示す グラフ である 。
在 栅极␣电阻 7 的 两 端 , 层 间 绝 缘 膜 12 被 刻 蚀 , 埋 入 钨 等 的 接触␣电极 6 。	ゲート␣抵抗 7 の 両 端 で 層 間 絶縁␣膜 12 が エ ッ チ ン グ さ れ 、 タ ン グ ス テ ン 等 の コンタクト␣電極 6 が 埋め 込ま れて いる 。
当 在 脑瘤 组织 的 测 定 中 将 预 定 的 时间段 设定 为 大约 5 分钟 的 时候 , 得 到 充分 的 结果 。	こ の 所 要 時 間 は 、 脳␣腫瘍␣組織 の 測 定 に おいて は 5 分 程度 で 十分な␣結果 が 得 られた 。

Figure 1: Examples of Chinese *(left)* and Japanese *(right)* sentences re-tokenized using the extracted monolingual multi-word terms. The boxes in plain line show the monolingual multi-word terms which correspond across languages. The boxes in dashed line show the monolingual multi-word terms which are re-tokenized only in their language.

called the C-value, is given by the following formula:

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

where a is the candidate string, $f(a)$ is its frequency of occurrence in the corpus, T_a is the set of extracted candidate terms that contain a, $|T_a|$ is the number of these candidate terms. In our experiments, following the C-value computation method and the linguistic pattern, we extract multi-word terms from a Chinese and a Japanese training corpus respectively. Then, we re-tokenize the training corpus with these extracted monolingual multi-word terms by enforcing these terms to be considered as one token. Technically, we just replace each space inside a multi-word term by a non-space word separator, so that each multi-word term is considered as one token.

The segmenter and part-of-speech tagger that we use are the Stanford parser[2] for Chinese and Juman[3] for Japanese. Figure 1 shows examples of re-tokenized Chinese–Japanese sentences with monolingual multi-word terms in Chinese and Japanese respectively.

2.2 Bilingual Multi-Word Term Extraction

Bilingual multi-word terms are multi-word term to multi-word term alignments, i.e., we only want to extract corresponding terms which are multi-word terms at the same time in both languages. We extract them by performing word-to-word, or, better said, token-to-token alignment on the Chinese–Japanese training corpus re-tokenized as described in the previous section. For that, we use the open source implementation of the sampling-based alignment method, Anymalign [4][4]. We further select the best bilingual multi-word terms by setting some threshold P on each of the direct and inverse translation probabilities ($0<P\leq 1$). Table 1 shows examples from the results of bilingual multi-word term extraction.

2.3 Using Bilingual Multi-Word Terms in SMT

We propose two protocols to use these extracted bilingual multi-word terms in SMT experiments. We will compare these two protocols with a standard baseline system.

The first protocol (System re-tok-all) is as follows. We train the translation model based on the training corpus re-tokenized with the best bilingual multi-word terms. The language model is learnt on the target part of the re-tokenized training corpus. The tuning and test sets used are also re-tokenized with the same bilingual multi-word terms as used for re-tokenizing the training data. We remove the non-space word separators after decoding and before the evaluation process.

The second protocol (System re-tok-train-only) is as follows. We train the translation model only based on the training corpus re-tokenized with the best bilingual multi-word terms. But the language model is learnt based on the original, un-re-tokenized, target language part of the training corpus. For consistency, we remove the non-space word separators from the phrase tables before performing tuning and decoding.

3 Experiments

3.1 Chinese and Japanese Data Used

The Chinese–Japanese parallel sentences used in our experiments are randomly extracted from the Chinese–Japanese

[2] http://nlp.stanford.edu/software/segmenter.shtml

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[4] Anymalign is a phrase-to-phrase alignment tool, but the use of option -N 1 limits its functionality to word-to-word alignment. Technically, we identify the multi-word term to multi-word term alignments by spotting the non-space word separators inserted inside multi-word terms in place of spaces.

Chinese	Japanese	Meaning	$P(t\|s)$	$P(s\|t)$	Kept	Good match
硬质␣碳␣皮膜	硬質␣炭素␣皮␣膜	‘diamond-like carbon’	1.000	1.000	yes	yes
接触␣电极	コンタクト␣電極	‘contact electrode’	0.946	0.972	yes	yes
极板	極␣板	‘electrode plate’	0.992	1.000	no	yes
废␣热	廃␣熱	‘waste heat’	0.844	0.241	no	yes
变速␣机	変速␣機	‘variable-speed motor’	1.000	0.006	no	yes
芯片␣级␣控制␣手机␣模块	チップ␣レベル	–	1.000	1.000	yes	no
激振␣电极	主に␣形成	–	0.862	0.982	yes	no

Table 1: Examples of bilingual multi-word terms extracted and then filtered by our method: firstly, pairs with only one word on any side are rejected, then pairs of multi-word terms where one of the translation probabilities is below the threshold (0.6 here) are rejected. The last column shows which extracted multi-word term pairs were considered correct or not by manual inspection.

JPO Patent Corpus (JPC)[5]. For our experiments, we randomly extract 100,000 parallel sentences from the training part, 500 parallel sentences from the tuning part, and 1,000 from the test part.

3.2 Monolingual and Bilingual Multi-Word Term Extraction

We apply the method described in Section 2.1 to independently extract monolingual multi-word terms from the 100,000 sentences of the training data of our Chinese–Japanese parallel corpus. We independently obtain 81,618 multi-word terms in Chinese and 93,105 in Japanese. The extracted monolingual multi-word terms were ranked by decreasing order of C-values. We re-tokenize the training corpus with the same number of Chinese and Japanese monolingual multi-word terms respectively. These terms are the first 80,000 monolingual multi-word terms with the highest C-values in each language.

We then extract bilingual multi-word terms from the Chinese–Japanese training corpus re-tokenized using these 80,000 monolingual multi-word terms, by following the method described in Section 2.2. We measured the number of bilingual multi-word terms extracted from the re-tokenized training corpus of 100,000 sentence pairs by the sampling-based alignment method which meet the constraint of having both translation probabilities above a given threshold. The second column in Table 2 shows this number when the threshold varies. In addition, we manually checked the correspondence between these bilingual multi-word terms. The percentage of good matches was roughly estimated to be over 70 % when the threshold becomes greater than 0.4.

3.3 SMT systems

We train a standard baseline system (no re-tokenization) using the GIZA++/MOSES pipeline [3]. We train the Chinese-to-Japanese translation model with 100,000 training parallel corpus. The monolingual part in the target language (Japanese) is used to learn a language model using KenLM [2] in word-based 5-grams. The development data with 500 parallel sentences is used for tuning by minimum error rate training [7]. For decoding, we use the default options of Moses, the distortion limit is set to 20.

Different from the baseline SMT system, here we make use of bilingual multi-word terms following the two experimental protocols described in Section 2.3.

Table 2 (column 3) shows the evaluation results for the first protocol (System re-tok-all) in BLEU scores [8]. We did not obtain significant difference in BLEU in comparison with the baseline system, except for BLEU scores which are significant lower than those of the baseline system when the thresholds are $P \geq 0.1$ and $P \geq 0.4$.

Because re-tokenization of all of the data did not lead to improvement, we decide to only re-tokenize the Chinese–Japanese training parallel corpus (System re-tok-train-only).

Table 2 (column 4) shows the evaluation of the results for the second protocol (System re-tok-train-only). Compared with the baseline system and the System re-tok-all, we obtained significantly better results in BLEU scores for thresholds equal to or greater than 0.3, while the scores for lower thresholds are similar to and not significantly different from the score of the baseline system. This shows that this protocol at least does not hurt and may be beneficial when applied with any value for the threshold.

[5] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html

Thresholds	♯ of bilingual multi-word terms	BLEU	BLEU
	(filtered by thresholds)	(System: re-tok-all)	(System: re-tok-train-only)
$\geq$ 0.0	52,785 (35 %)	32.08±1.07	32.44±1.07
$\geq$ 0.1	31,795 (52 %)	31.88±1.10	32.23±1.18
$\geq$ 0.2	27,916 (58 %)	32.42±1.14	32.00±1.16
Baseline	-	32.35 ±1.15	32.35±1.15
$\geq$ 0.3	25,404 (63 %)	31.85±1.08	33.08±1.12*
$\geq$ 0.4	23,515 (72 %)	31.45±1.13	32.77±1.15*
$\geq$ 0.5	21,846 (76 %)	32.11±1.12	33.02±1.14*
$\geq$ **0.6**	20,248 (78 %)	32.68±1.13	**33.32±1.15***
$\geq$ 0.7	18,759 (79 %)	32.61±1.12	32.85±1.19*
$\geq$ 0.8	17,311 (79 %)	32.34±1.15	33.25±1.06*
$\geq$ 0.9	15,464 (80 %)	32.16±1.11	33.20±1.15*

Table 2: Results of bilingual multi-word extraction and evaluation results for Chinese-to-Japanese translation with the two proposed protocols (Systems re-tok-all and re-tok-train-only) for different thresholds on the translation probabilities. The score of the baseline is given on line 4. The best BLEU score obtained (33.32) is for the System re-tok-train-only with a threshold of 0.6 (boldfaced score). BLEU scores marked with * are significantly better than the score of the Baseline system at $p < 0.01$, except for threshold ≥ 0.4 at $p < 0.05$.

4 Conclusion

We proposed an approach to improve translation accuracy in statistical machine translation of Chinese–Japanese patents by re-tokenizing the parallel training corpus with extracted bilingual multi-word terms using our proposed methods. We did not use any other additional corpus or terminological lexicon. An investigation of the results of our experiments indicate that the bilingual multi-word terms extracted have over 70 % precision (good match) for threshold values over 0.4.

We proposed two experimental protocols for using the extracted bilingual multi-word terms in SMT experiments. The first protocol re-tokenized all of the data with the bilingual multi-word terms. The second protocol only re-tokenized the training data to produce the phrase tables of the SMT system. The first protocol did not lead to improvements in translation accuracy compared with the baseline system. The second protocol led to statistically significant improvements for thresholds equal to or greater than 0.3.

References

[1] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[2] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 187–197. Association for Computational Linguistics, 2011.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, and Richard Zens. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180. Association for Computational Linguistics, 2007.

[4] Adrien Lardilleux and Yves Lepage. Sampling-based multilingual alignment. In *Proceedings of the 7th Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, 2009.

[5] E Milios, Y Zhang, B He, and L Dong. Automatic term extraction and document similarity in special text corpora. In *Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics (PACLIC 2003)*, pages 275–284, 2003.

[6] 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.

[7] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, volume 1, pages 160–167. Association for Computational Linguistics, 2003.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 4Oth Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318, 2002.