Improving Patent Translation using Bilingual Term Extraction and Re-tokenization for Chinese–Japanese

Wei YANG and Yves LEPAGE

Graduate School of Information, Production and Systems
Waseda University

We describe a method to improve Chinese–Japanese statistical machine translation (SMT) of patents by re-tokenizing the training corpus with aligned bilingual multi-word terms.

Chinese and Japanese tokenization on patent sentences

- Examples of terms in JPO[1] Chinese–Japanese patent sentences are tokenized at different levels of granularity. Segmentation tools used are Stanford[2] for Chinese and Juman[3] for Japanese.

Language	Sentence
Chinese	该/钽阳/极体/通常/是/烧结/的/。
Japanese	タンタル/陽極/ボディ/は/、/通常/、/焼結/さ/れて/いる/。
Chinese	贴片/52/-/58/也/通过/导线/连接/到/系统/控制器/30/。
Japanese	パッチ/52/～/58/は/、/また/、/電線/に/よって/システム/コント/ローラ/30/に/接続/さ/れる/。
Chinese	在/第一/热/处理/之后/，/氧化物/半导体层/变成/缺氧/的/氧化物/半导体/，/即/，/电阻率/变得/更低/。
Japanese	酸化/物/半導体/層/は/、/第/1/の/加熱/処理/後/に/酸素/欠乏/型/と/なり/、/低/抵抗/化/する/。

Monolingual multi-word term extraction using C-value

- The C-value is a commonly used automatic domain-independent method for multi-word term extraction. This method has two main parts: a linguistic part and a statistical part.
  - The linguistic pattern we use is the regular expression[4]: $(Adjective|Noun)^{+}\ Noun$
  - The statistical part, the measure of termhood, called the C-value, is given by the following formula:

$$\text{C–value}(a) = \begin{cases} \log_2 |a| \cdot f(a) & \text{if } a \text{ is not nested,} \\ \log_2 |a| \big(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \big) & \text{otherwise} \end{cases} \tag{1}$$

Chinese or Japanese sentences	Extracted monolingual terms
在♯P 该♯DT 方法♯NN 中♯LC ,♯PU 能够♯VV 得到♯VV 从♯P **心脏♯NN 周期♯NN** 内♯LC 的♯DEG **心收缩♯NN 期♯NN** 到♯VV 心♯NN 舒张♯VV 期♯NN 之间♯LC 的♯DEG 血液♯NN 移动♯VV 的♯DEC 1♯CD 个♯M 以上♯LC 的♯DEG 图像♯NN 。♯PU	心脏　周期 ‘cardiac cycle’ 心收缩　期 ‘systole’
この/指示詞 方法/名詞 に/助詞 おいて/動詞 は/助詞 、/特殊 **心臓**/名詞 **周期**/名詞 内/接尾辞 の/助詞 **心**/名詞 **収縮**/名詞 **期**/名詞 から/助詞 **心**/名詞 **拡張**/名詞 **期**/名詞 まで/助詞 の/助詞 間/名詞 の/助詞 **血液**/名詞 **移動**/名詞 の/助詞 1/名詞 枚/接尾辞 以上/接尾辞 の/助詞 画像/名詞 が/助詞 得/動詞 られる/接尾辞 。/特殊	心臓　周期 ‘cardiac cycle’ 心　収縮　期 ‘systole’ 心　拡張　期 ‘diastole’ 血液　移動 ‘blood moving’

- We re-tokenize such candidate terms in the corpus by enforcing the extracted monolingual multi-word terms to be considered as one token. Each candidate multi-word term is re-tokenized (aligned) with markers.

Bilingual multi-word term extraction

- We use the open source implementation of the sampling-based approach, `Anymalign` [A. Lardilleux and Y. Lepage, 2009].
  - consider multi-word to multi-word terms (green √)
    filtering by thresholds, ratio of lengths in words, and components of the bilingual multi-word terms
  - We use kanji-hanzi conversion method (Unihan Mapping Data, Langconv Traditional-Simplified Conversion data, Hanzi-kanji conversion) (blue √).
  - consider one side is multi-word term (red √)
    filtering by thresholds, ratio of lengths in words and components of the bilingual multi-word terms

Extract or not	Correct or not	Chinese	Japanese	$P(t\|s)$	$P(s\|t)$
○	√	接口_电子_线路	インタフェース_電子_回路	0.923077	0.928571
○	√	顶盖_主体	キャップ_本体	1.000000	0.833333
○	√	冷却_层	冷却_層	1.000000	0.951220
○	√	薄_膜片	薄膜_シート	1.000000	1.000000
○	√	肺气肿	肺_気腫	0.818182	0.900000
○	∗	激振_电极	主に_形成	0.861538	0.982456
○	∗	芯片_级_控制_手机_模块	チップ_レベル	1.000000	1.000000
×	√	废_热	廃_熱	0.844444	0.240506
×	√	变速_机	変速_機	1.000000	0.005988
×	√	壁部	壁_部	0.948247	0.677804
×	√	核酸	核_酸	0.974392	0.956030
×	√	极板	極_板	0.992000	1.000000
×	√	薄_膜	薄膜	0.197531	0.058252
×	√	贵_金属	貴金属	0.990548	0.984962
×	√	供油路	給油_路	1.000000	1.000000
×	√	输入_输出	入出力	0.952030	0.811321
×	√	制动液	ブレーキ_液	0.985437	0.902222
×	√	甲醛	ホルム_アルデヒド	0.997275	0.910448
×	√	存储器_控制器	メモリコントローラ	0.968831	0.917589
×	√	枢轴_板	ピボットプレート	0.977011	1.000000
×	∗	切换_步骤	Handover	1.000000	1.000000
×	∗	亭	キオスク_端末	1.000000	1.000000
×	∗	飞行物	前記_飛行_体	1.000000	1.000000
×	∗	总_能量_消耗量	総計	1.000000	1.000000

- We re-tokenize parallel training corpus with extracted bilingual multi-word terms. Each multi-word term is re-tokenized (aligned) with markers.

Experiments and results

- Baseline (zh→ja) JPO corpus (lines)
  training: 100,000, tuning: 500, test: 1,000 and 2,000
- Monolingual multi-word term are extracted from training data:
  Chinese: 81,618 and Japanese: 93,105
- SMT experiments
  - Baseline system (no re-tokenization)
  - Several systems based on re-tokenized training data using different number of bilingual multi-word terms.

	Filtering by thresholds (a)			Filtering by thresholds (a) + the ratio of lengths + the components (b) + kanji-hanzi conversion (c)			
Thresholds	♯ of bilingual multi-word terms (a)	BLEU	p-value	♯ of bilingual multi-word terms (a + b)	♯ of bilingual multi-word terms (a + b + c)	BLEU	p-value
≥ 0.0	52,785 (35%)	32.44	> 0.05	48,239 (63%)	49,474 (70%)	**33.19**	< 0.05
≥ 0.1	31,795 (52%)	32.23	> 0.05	29,050 (68%)	30,516 (78%)	**33.09**	< 0.05
≥ 0.2	27,916 (58%)	32.00	> 0.05	25,562 (75%)	27,146 (83%)	**33.12**	< 0.05
Baseline (1,000)	-	32.35	-	-	-	32.35	-
≥ 0.3	25,404 (63%)	**33.08**	< 0.01	23,321 (78%)	25,006 (83%)	**33.25**	< 0.01
≥ 0.4	23,515 (72%)	**32.77**	< 0.05	21,644 (80%)	23,424 (84%)	**33.31**	< 0.01
≥ 0.5	21,846 (76%)	**33.02**	< 0.01	20,134 (85%)	22,000 (88%)	**33.23**	< 0.01
≥ 0.6	20,248 (78%)	**33.32**	< 0.01	18,691 (88%)	20,679 (89%)	**33.75**	< 0.01
≥ 0.7	18,759 (79%)	**32.85**	< 0.01	17,340 (88%)	19,460 (90%)	**33.41**	< 0.01
≥ 0.8	17,311 (79%)	**33.25**	< 0.01	16,001 (89%)	18,265 (90%)	**33.38**	< 0.01
≥ 0.9	15,464 (80%)	**33.20**	< 0.01	14,284 (92%)	16,814 (93%)	**33.43**	< 0.01

	Considering one side multi-word terms + filtering by constraints (d) + (a + b + c)				
Thresholds	♯ of one side multi-word terms	♯ of filtered one side multi-word terms (d)	♯ of combination of multi-word terms (a + b + c + d)	BLEU	p-value
≥ 0.0	72,428 (2%)	27,116 (40%)	75,425 (64%)	32.55	> 0.05
≥ 0.1	18,395 (7%)	7,570 (55%)	37,059 (78%)	**33.36**	< 0.01
≥ 0.2	14,179 (12%)	6,031 (62%)	32,224 (85%)	**33.20**	< 0.01
≥ 0.3	11,849 (15%)	5,161 (70%)	29,280 (90%)	**33.41**	< 0.01
≥ 0.4	10,259 (17%)	4,537 (76%)	27,125 (90%)	**33.37**	< 0.01
≥ 0.5	9,069 (17%)	4,050 (76%)	25,270 (90%)	**33.63**	< 0.01
≥ 0.6	7,875 (30%)	3,575 (76%)	23,522 (93%)	**34.27**	< 0.01
≥ 0.7	6,900 (30%)	3,088 (80%)	21,874 (93%)	**33.90**	< 0.01
≥ 0.8	6,026 (30%)	2,726 (80%)	20,318 (93%)	**33.85**	< 0.01
≥ 0.9	5,062 (30%)	2,275 (82%)	18,484 (95%)	**33.75**	< 0.01

Test is 2,000 sentences (zh)	Evaluation result
Baseline	32.29
Re-tokenization	**33.61**
	(p-value < 0.01)

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html
[2] http://nlp.stanford.edu/software/segmenter.shtml
[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[4] Pattern for Chinese: $(JJ|NN)^{+}\ NN$, pattern for Japanese: (形容詞 | 名詞)$^{+}$ 名詞. ‘JJ’ and ‘形容詞’ are codes for adjectives, ‘NN’ and ‘名詞’ are codes for nouns in the Chinese and the Japanese taggers that we use.

The 3rd Workshop on Asian Translation (WAT2016), Osaka, Japan, 2016

The 3rd Workshop on Asian Translation (WAT2016), Osaka, Japan, 2016