Translation Using JAPIO Patent Corpora: JAPIO at WAT2016

Satoshi Kinoshita, Tadaaki Oshio,

Tomoharu Mitsuhashi, Terumasa Ehara

一般財団法人
日本特許情報機構
Japan Patent Information Organization

Questions

1. Can PB-SMT with large corpora generate better translations than systems with an advanced framework such as NMT?

2. Are large patent corpora useful for translation of scientific domain?

JAPIO Patent Corpora

- Parallel corpora of EJ/CJ/KJ, which are automatically extracted from patent families using an alignment tool by NICT
- Corpus size:

  EJ: 250 million

  CJ: 100 million

  KJ: 5 million
- Purpose
  - Train SMT
  - Search by human translators
  - Build translation dictionaries

Systems

- NICT PB-SMT toolkit for EJ & CJ (Preordering included)
  Moses for KJ (character-based tokenizer)

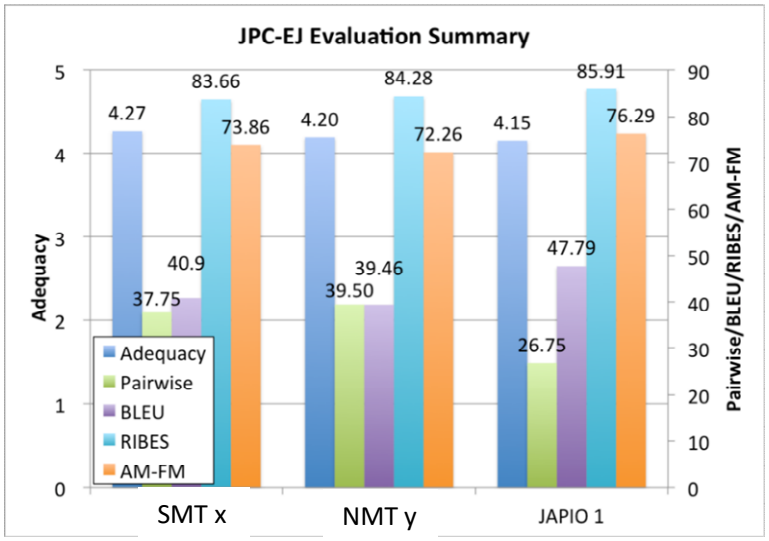
- Additional post-editing functions:
  ASPEC-EJ/CJ: modify punctuations and patent specific expressions
  EJ: Recapitalization of OOVs
  KJ: Resolving unbalanced parentheses

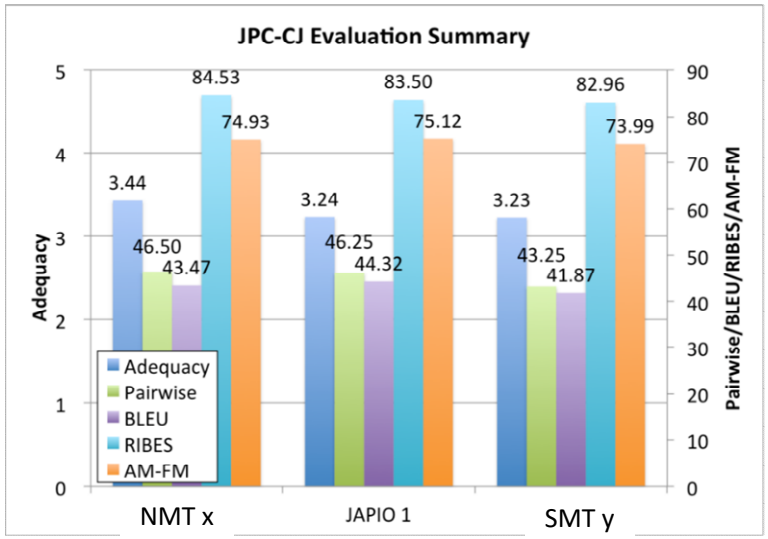
Results (#1)

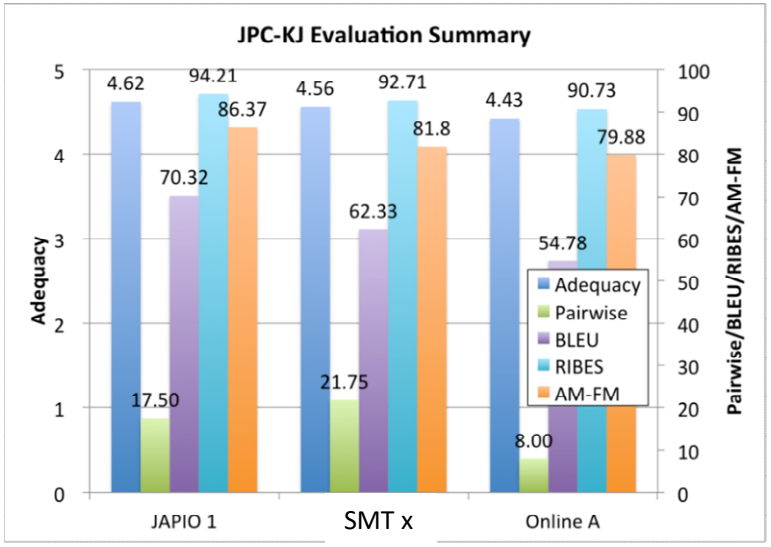
#	Subtask	System	Corpus	Size (million)	BLEU	RIEBS	AMFM	HUMAN
1	JPC-EJ	JAPIO-a	JAPIO-test	5	45.57	0.851376	0.747910	17.750
2		JAPIO-b	JAPIO-test+JPC	6	47.79	0.859139	0.762850	26.750
3		JAPIO-c	JAPIO	5	50.28	0.859957	0.768690	—
4		JAPIO-d	JPC	1	38.59	0.839141	0.733020	—
5	JPC-CJ	JAPIO-a	JAPIO-test	3	43.87	0.833586	0.748330	43.500
6		JAPIO-b	JAPIO-test	4	44.32	0.834959	0.751200	46.250
7		JAPIO-c	JAPIO	49	58.66	0.868027	0.808090	—
8		JAPIO-d	JPC	1	39.29	0.820339	0.733300	—
9	JPC-KJ	JAPIO-a	JAPIO	5	68.62	0.938474	0.858190	-9.000
10		JAPIO-b	JAPIO+JPC	6	70.32	0.942137	0.863660	17.500
11		JAPIO-c	JPC	1	69.10	0.940367	0.859790	—

- SMTs with larger corpora are competitive but not superior than state-of-the-art systems like NMTs trained with 1M JPC corpora.

Results (#2)







*Figures on this page are provided by Organizer.

Results (#3)

#	Subtask	System	Corpus	Size (million)	BLEU	RIEBS	AMFM	HUMAN
12	ASPEC-EJ	JAPIO-a	JAPIO	10	20.52	0.723467	0.660790	4.250
13		Online x	–	–	18.28	0.706639	0.677020	49.750
14		RBMT x	–	–	13.18	0.671958	–	–
15	ASPEC-CJ	JAPIO-a	JAPIO	49	26.24	0.790553	0.696770	16.500
16		Online x	–	–	11.56	0.589802	0.659540	-51.250
17		RBMT x	–	–	19.24	0.741665	–	–

- SMTs with Patent Corpora are better than RBMT and Online (except AMFM and HUMAN in EJ)
- SMTs with Patent Corpora could be used practically for translation in scientific-paper domain

Error Analysis

- Error analysis of patent subtask (#sent=200)

Error Type	EJ	CJ	KJ
Insertion	0	0	6
Deletion	4	9	1
OOV	6	9	2
Mistranslation(content word)	44	41	30
Mistranslation(functional word)	21	51	0
Pre-ordering	33	45	0
Other	6	7	2
Total	114	162	41

# of mistranslation of functional words and errors related to pre-ordering suggests improvement of pre-ordering is essential.

Conclusion

1. Can PB-SMT with large corpora generate better translations than systems with an advanced framework such as NMT?

   → No

   Improvement of pre-ordering seems more effective than increasing corpus

2. Are large patent corpora useful for translation of scientific domain?

   → Yes

   Could be used without domain adaptation

Suggestion

Suggestions to Workshop Organizer:

To make results of WAT more useful,

- Make a new test set without using patent families

  If impossible, remove training examples which are extracted from the same documents as test set examples are extracted.

- Balance sentence length of test sets of JPC-EJ/CJ
  - Sentences of EJ test set are much shorter than CJ
  - Current results may mislead readers