Consistent Improvement in Translation Quality of Chinese–Japanese Technical Texts by Adding Additional Quasi-parallel Training Data

Wei Yang and Yves Lepage

Graduate School of Information, Production and Systems
Waseda University

kevinyoogi@akane.waseda.jp ; yves.lepage@waseda.jp

Bilingual parallel corpora are an extremely important resource as they are typically used in data-driven machine translation. There already exist many freely available corpora for European languages, but almost none between Chinese and Japanese. The constitution of large bilingual corpora is a problem for less documented language pairs. We construct a quasi-parallel corpus automatically by using analogical associations based on certain number of parallel corpus and a small number of monolingual data. Furthermore, in SMT experiments, by adding this kind of Chinese–Japanese data into the baseline training corpus, on the same test set, the evaluation scores of the translation results we obtained were significantly or slightly improved over the baseline systems.

Building analogical clusters according to proportional analogies

- Proportional analogy establishes a general relationship between four objects A, B, C and D: "A is to B as C is to D". An efficient algorithm for the resolution of analogical equations has been proposed in (Lepage, 1998)[1].

$$A : B :: C : D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \ \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases}$$

- Sentential analogy:

早急に対応して下さい。	:	早急に対応して欲しい。	::	元に戻して下さい。	:	元に戻して欲しい。

- Analogical cluster: We can cluster sentential analogies as a sequence of lines, where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy.

早急に対応して下さい。：早急に対応して欲しい。
元に戻して下さい。：元に戻して欲しい。
やめて下さい。：やめて欲しい。

- We produced all possible analogical clusters from Chinese and Japanese unrelated unaligned monolingual data collected from the Web.

	Chinese	Japanese
# of different sentences	70,000	70,000
# of clusters	23,182	21,975

▲ Such clusters can be considered as rewriting models that can generate new sentences.

- Extracting corresponding clusters by computing similarity according to a classical Dice formula:

$$Sim = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \quad \Rightarrow \quad Sim_{C_{zh}-C_{ja}} = \frac{1}{2}(Sim_{left} + Sim_{right})$$

S_{zh} and S_{ja} denote the minimal sets of changes across the clusters (both on the left or right) in both languages (after translation and conversion).

Chinese cluster

left part	:	right part
经典游戏	:	游戏很不错
'*classic game*'	:	'*The game is very good.*'
喜欢经典	:	很不错喜欢
'*I like classic.*'	:	'*Very good, I like it.*'
经典啊	:	很不错啊
'*Classic!*'	:	'*Very good!*'

Japanese cluster

left part	:	right part
クラシック物語	:	この物語はとてもいい
'*classic narrative*'	:	'*The narrative is very good.*'
クラシック音楽	:	この音楽はとてもいい
'*classic music*'	:	'*The music is very good.*'

Generation of new sentences using analogical associations

- Generation of new sentences
  We use analogy as an operation by which, given two related forms (rewriting model) and only one form, the fourth missing form is coined[2]. Applied on sentences, this principle can be illustrated as follows:

早急に対応して下さい。 : 早急に対応して欲しい。 :: 正式版に戻して下さい。 : x
$\Rightarrow \quad x =$ 正式版に戻して欲しい。

- Experiments on new sentence generation and filtering by N-sequences
  We eliminate any sentence that contains an N-sequence of a given length unseen in our data. For valid sentences, we remember their corresponding seed sentences and the cluster identifiers they were generated from.

	Chinese		Japanese	
# of seed sentences	99,538		97,152	
# of clusters	23,182		21,975	
# of candidate sentences	105,038,200 Q= 29%		80,183,424 Q= 40%	
	unique	seed–new–#	unique	seed–new–#
# of filtered sentences	33,141	67,099	40,234	84,533
	Q= 96%		Q= 96%	

- Deducing and acquiring quasi-parallel sentences
  We deduce translation relations based on the initial parallel corpus and corresponding clusters between Chinese and Japanese.

Chinese	Japanese	Chinese–Japanese		
seed–new–#	seed–new–#	Initial parallel corpus	Corresponding clusters	Quasi-parallel corpus
67,099	84,533	103,629	15,710	35,817

A	:	B	::	C_{seed}	:	X_{new-zh}
经典游戏	:	游戏很不错				
喜欢经典	:	很不错喜欢	::	经典电影 '*classic film*'	⇒	电影很不错 '*The film is very good.*'
经典啊	:	很不错啊				很不错电影 '*That's very good, the film.*'

A	:	B	::	C_{seed}	:	X_{new-ja}
クラシック物語	:	この物語はとてもいい				
クラシック音楽	:	この音楽はとてもいい	::	クラシック映画 '*classic film*'	⇒	この映画はとてもいい '*The film is very good.*'

SMT experiments

- **Experimental protocol**: To assess the contribution of the generated quasi-parallel corpus, we compare two SMT systems. The first one is constructed using the initial given ASPEC-JC parallel corpus. This is the baseline. The second one adds the additional quasi-parallel corpus obtained using analogical associations and analogical clusters.

	Baseline	Chinese	Japanese
train	sentences	672,315	672,315
	words	18,847,514	23,480,703
	mean ± std.dev.	28.12 ± 15.20	35.05 ± 18.88
	+ Quasi-parallel	Chinese	Japanese
train	sentences	**708,132**	**708,132**
	words	19,212,187	24,512,079
	mean ± std.dev.	27.13 ± 14.19	34.23 ± 17.22
	Both experiments	Chinese	Japanese
tune	sentences	2,090	2,090
	words	60,458	73,177
	mean ± std.dev.	28.93 ± 15.86	35.01 ± 18.87
test	sentences	2,107	2,107
	words	59,594	72,027
	mean ± std.dev.	28.28 ± 14.55	34.18 ± 17.43

- **Experimental results** (using the different segmentation tools and moses version):
  - segmentation tools: urheen and mecab, moses 1.0: significant.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	29.10	7.5677	0.5352	0.5478	0.7801
	+ additional training data	**32.03**	**7.9741**	**0.5069**	**0.5172**	**0.7906**
ja-zh	baseline	22.98	7.0103	0.5481	0.5711	0.7893
	+ additional training data	**24.87**	**7.3208**	**0.5273**	**0.5482**	**0.8013**

  - segmentation tools: urheen and mecab, moses 2.1.1

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	33.41	8.1537	0.4967	0.5061	0.7956
	+ additional training data	**33.68**	**8.1820**	**0.4955**	**0.5039**	**0.7964**
ja-zh	baseline	25.53	7.3885	0.5227	0.5427	0.8053
	+ additional training data	**25.80**	**7.4571**	**0.5176**	**0.5378**	**0.8060**

  - segmentation tools: kytea, moses 1.0

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	28.35	7.3123	0.5667	0.5741	0.7610
	+ additional training data	**28.87**	**7.4637**	**0.5566**	**0.5615**	**0.7739**
ja-zh	baseline	22.83	6.9533	0.5633	0.5853	0.7807
	+ additional training data	**23.18**	**7.0402**	**0.5547**	**0.5778**	**0.7865**

[1] Yves Lepage. Solving analogies on words: An algorithm, *COLING-ACL'98, Volume I, pp. 728-735, Montréal, Aug. 1998.*
[2] Ferdinand de Saussure. Cours de linguistique générale, Payot, Lausanne et Paris, [1ère éd. 1916] edition, 1995.