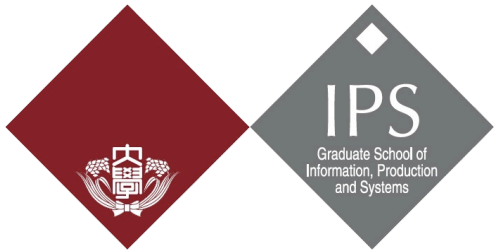
Inflating a Training Corpus for SMT by Using Unrelated Unaligned Monolingual Data



Wei Yang and Yves Lepage

EBMT/NLP Lab of IPS, Waseda University

PolTAL, September 18, 2014

Contents

1. Motivation and Related Research
2. Data Preparation for the Experiments
3. Construction of Analogical Clusters
4. Generation of New Sentences Using Analogical Associationss
5. SMT Experiments
6. Synthesis of Results and Conclusion

Contents

1 Motivation and Related Research

2 Data Preparation for the Experiments

3 Construction of Analogical Clusters

4 Generation of New Sentences Using Analogical Associationss

5 SMT Experiments

6 Synthesis of Results and Conclusion

Motivation and Related Research

- Sentence–level aligned parallel corpora are necessary resource as training data in statistical machine translation (SMT).
- There exist numerous freely available bilingual or multilingual parallel corpora for language pairs that involve English, such as the Europarl parallel corpus (Koehn, 2005).
- But the linguistic resources between languages like: Chinese, Japanese, Thai, Hindi or Bahasa Indonesian are relatively scarce.

Motivation and Related Research

- We propose a method to construct a quasi-parallel corpus based on the notion of proportional analogy.
- We report SMT experiments and compare a baseline system using a relatively small amount of parallel data and a system built on the baseline by adding the quasi-parallel corpus as additional training data.
- We report improvements in translation quality.

Contents

1 Motivation and Related Research

2 Data Preparation for the Experiments

3 Construction of Analogical Clusters

4 Generation of New Sentences Using Analogical Associationss

5 SMT Experiments

6 Synthesis of Results and Conclusion

Data Preparation for the Experiments

- Collection of parallel Chinese–Japanese data

	Language	# of different sentences	size of sentences in characters mean ± std.dev.	# of characters	# of words
Subtitle Corpus (106,310)	Chinese	99,251	8.68 ± 3.59	861,723	589,757
+ JEC (3,804)	Japanese	90,406	11.99 ± 4.36	1,084,287	647,285

Table: Statistics on the Chinese–Japanese subtitle data combined with a part of JEC sentences. This constitutes our initial parallel corpus (110,114 sentence pairs in total: 106,310 + 3,804).

Data Preparation for the Experiments

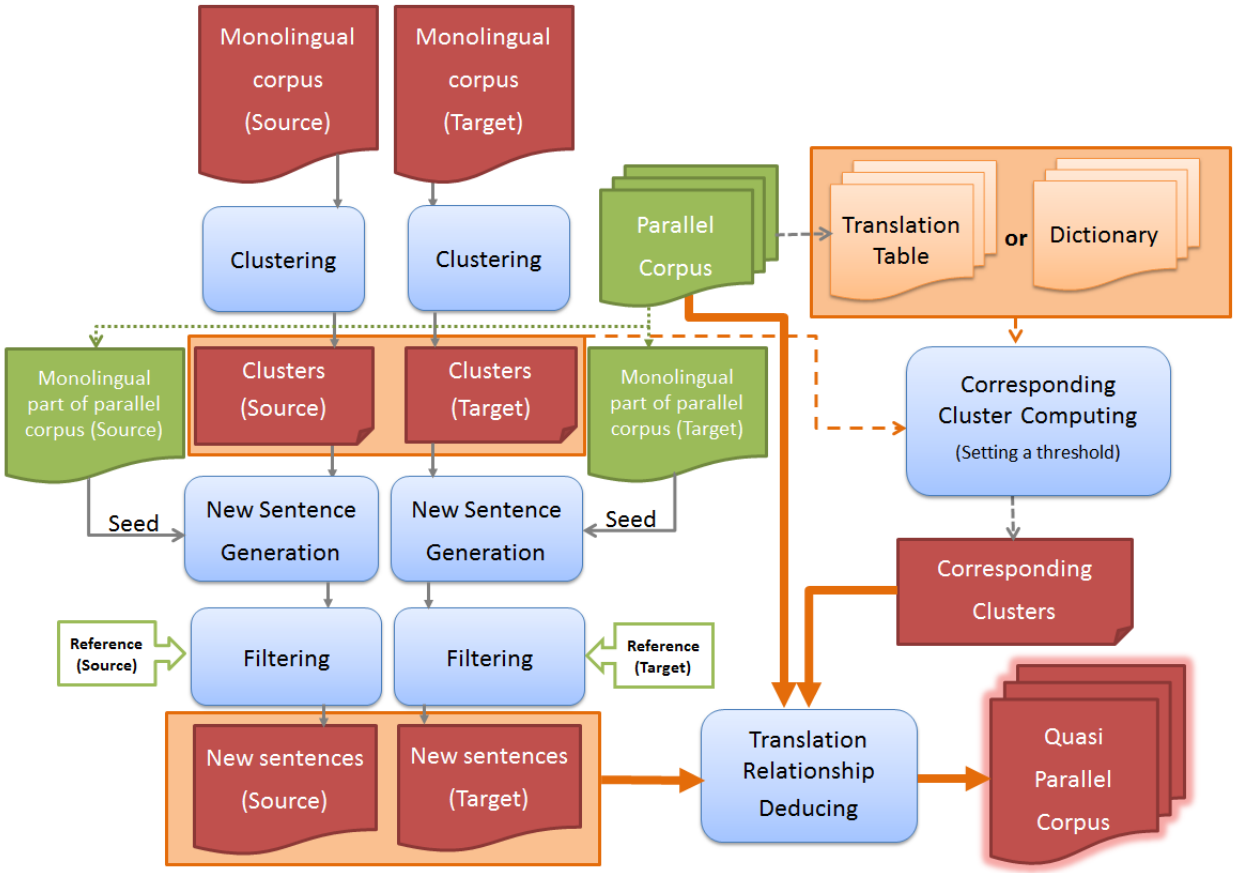
- Collection of monolingual resources

	# of different sentences (cleaned)	size of sentences in characters mean ± std.dev.	# of characters	# of words
Chinese	70,000	10.29 ± 6.21	775,530	525,462
Japanese	70,000	15.06 ± 6.34	1,139,588	765,085

Table: Statistics on the cleaned Chinese and Japanese monolingual short sentences used in clustering experiment.

Data Preparation for the Experiments

- The flow chart of our proposed method



Contents

1 Motivation and Related Research

2 Data Preparation for the Experiments

3 Construction of Analogical Clusters

4 Generation of New Sentences Using Analogical Associationss

5 SMT Experiments

6 Synthesis of Results and Conclusion

Construction of Analogical Clusters

- Proportional analogies

$$A:B::C:D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases} \quad (1)$$

Construction of Analogical Clusters

- Sentential analogies

紅茶が 飲みたい。 : あなたは紅茶が好きですか。 :: ビールが飲みたい。 : あなたはビールが好きですか。

I'd like a black tea. : *Do you like black tea?* :: *I'd like a beer.* : *Do you like beer?*

  - $d(A, B) = d(C, D) = 13$ and $d(A, C) = d(B, D) = 5$
  - 茶 (tea): 1 (in A) - 1 (in B) = 0 (in C) - 0 (in D)

Construction of Analogical Clusters

- Analogical cluster

紅茶が飲みたい。	:	あなたは紅茶が好きですか。
'I'd like a cup of black tea.'		*'Do you like black tea?'*
ビールが飲みたい。	:	あなたはビールが好きですか。
'I'd like a beer.'		*'Do you like beer?'*
ジュースが飲みたい。	:	あなたはジュースが好きですか。
'I'd like some juice.'		*'Do you like juice?'*
冷たいお水が飲みたい。	:	あなたは冷たいお水が好きですか。
'I'd like some cold water.'		*'Do you like cold water?'*

Construction of Analogical Clusters

- Cluster construction

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

Table: Statistics on the Chinese and Japanese clusters constructed from our unrelated monolingual data independently in each language.

Construction of Analogical Clusters

- Computing the correspondence between clusters
  - We calculate the similarity between two Chinese and Japanese word sets according to a classical Dice formula:

$$Sim = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \tag{2}$$

  - We take the arithmetic mean to compute the similarity between two Chinese and Japanese clusters:

$$Sim_{C_{zh}-C_{ja}} = \frac{1}{2}(Sim_{left} + Sim_{right}) \tag{3}$$

  - About 14,578 corresponding clusters were extracted ($Sim_{C_{zh}-C_{ja}} \geq 0.300$) by the above steps.

Construction of Analogical Clusters

- Computing the correspondence between clusters

I like beer.	:	**Do you** like beer?
I like juice.	:	**Do you** like juice?
language 1		
I study maths.	:	**Do you** study maths?
I watch movies.	:	**Do you** watch movies?
I read books.	:	**Do you** read books?
language 2		

Table: Two corresponding clusters. They do not have the same sizes and the sentences contained are not translations. But the change between the left and the right columns (‘I . . . .’ $\rightarrow$ ‘Do you . . . ?’) is the same.

Construction of Analogical Clusters

经典游戏	:	游戏很不错			
‘*classic game*’		‘*The game is not bad.*’			
喜欢经典	:	很不错喜欢	クラシック物語	:	この物語はとてもいい
‘*I like classic.*’		‘*Not bad, I like it.*’	‘*classic narrative*’		‘*The narrative is not bad.*’
经典啊	:	很不错啊	クラシック音楽	:	この音楽はとてもいい
‘*Classic!*’		‘*Not bad!*’	‘*classic music*’		‘*The music is not bad.*’
$S_{zh_{left}}$ = { 经典 }		$S_{zh_{right}}$ = { 很, 不错 }	$S_{ja_{left}}$ = { クラシック }		$S_{ja_{right}}$ = { この, は, とても, いい }

Table: An example of the extracted corresponding clusters which constructed based on unrelated unaligned monolingual.

Contents

1 Motivation and Related Research

2 Data Preparation for the Experiments

3 Construction of Analogical Clusters

4 Generation of New Sentences Using Analogical Associationss

5 SMT Experiments

6 Synthesis of Results and Conclusion

Generation of New Sentences Using Analogical Associations

- Generation of new sentences using analogical associations

Analogy ($A : B :: C : D$) is not only a structural relationship. It is also a process (Itkonen, 2005) by which, given two related forms and only one form, the fourth missing form is coined (Saussure, 1916).

紅茶が飲みたい。 : あなたは紅茶が好きですか。 :: ビールが飲みたい。 : x $\Rightarrow$ $x =$ あなたはビールが好きですか。

'I'd like a black tea.' : *'Do you like black tea?'* :: *'I'd like a beer.'* : x $\Rightarrow$ $x =$ *'Do you like beer?'*

Generation of New Sentences Using Analogical Associations

- Experiments on new sentence generation and filtering

		Chinese	Japanese
Initial data	# of seed sentences # of clusters	99,251 23,182	90,406 21,975
New sentence generation	# of candidate sentences	192,121,764 Q= 20%	50,418,891 Q= 50%
Quality assessment (filtered)	# of new valid sentences	unique | seed–new–# 34,230 | 105,537 Q= 99%	unique | seed–new–# 142,820 | 191,409 Q= 99%

Table: Statistics on new sentence generation in Chinese and Japanese. Q is the quality of the new candidate sentences or new valid sentences after filtering.

Generation of New Sentences Using Analogical Associations

- Deduction of translation relations between new generated sentences

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 not bad.*' ⋆ 很不错电影 '*That's not bad, the film.*'
经典啊	:	很不错啊				

A	:	B	::	C_{seed}	:	X_{new-ja}
クラシック物語	:	この物語はとてもいい	::	クラシック映画 '*classic film*'	⇒	⋆ この映画はとてもいい '*The film is not bad.*'
クラシック音楽	:	この音楽はとてもいい				

Table: The result of new sentence generation in Chinese and Japanese based on a pair of parallel seed sentences according to the clusters were given.

Generation of New Sentences Using Analogical Associations

- Deduction of translation relations between new generated sentences

Chinese	Japanese	Chinese–Japanese		
seed–new–#	seed–new–#	Initial parallel corpus	Corresponding clusters	Quasi-parallel corpus
105,537	191,409	110,114	14,578	76,151

Table: Statistics on the quasi-parallel corpus deducing.

Contents

1 Motivation and Related Research

2 Data Preparation for the Experiments

3 Construction of Analogical Clusters

4 Generation of New Sentences Using Analogical Associationss

5 SMT Experiments

6 Synthesis of Results and Conclusion

SMT Experiments

- Experimental protocol

	Baseline	Chinese	Japanese
train	sentences	110,114	110,114
	words	637,036	721,850
	mean ± std.dev.	5.94 ± 2.60	6.69 ± 2.94

	+ Quasi-parallel	Chinese	Japanese
	sentences	**186,265**	**186,265**
	words	1,147,098	1,318,747
	mean ± std.dev.	6.06 ± 2.61	7.16 ± 3.08

	Both experiments	Chinese	Japanese
tune	sentences	500	500
	words	3,582	5,042
	mean ± std.dev.	7.15 ± 2.86	10.12 ± 3.39
test	sentences	1,000	1,000
	words	7,285	10,126
	mean ± std.dev.	7.28 ± 2.87	10.15 ± 3.30

Table: Statistics on the Chinese–Japanese corpus used for the training, tuning, and test sets in baseline (left) and baseline + quasi-parallel data (right). The tuning and testing sets are the same in both experiments.

SMT Experiments

- Experimental results of SMT

		BLEU	NIST	WER	TER
zh-ja	baseline	13.10	4.1732	0.7229	0.7344
	+ extra training data	**19.27**	**4.7013**	**0.6880**	**0.6933**
ja-zh	baseline	10.94	4.4028	0.7545	0.7621
	+ extra training data	**17.66**	**4.7989**	**0.7140**	**0.7214**

Table: Evaluation results for Chinese–Japanese translation across two SMT systems (baseline and baseline + additional quasi-parallel data).

SMT Experiments

- Analysis of the results

		Target language							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source language	1-grams	23,789	35,494	25,069	13,670	6,568	2,982	1,257	108,829
	2-grams	34,865	52,596	38,612	22,429	12,300	6,413	3,113	170,328
	3-grams	18,904	33,116	39,633	29,465	19,262	11,881	6,617	158,878
	4-grams	8,097	15,948	24,779	28,160	23,629	17,628	11,495	129,736
	5-grams	3,235	7,020	12,656	18,532	21,166	19,277	15,072	96,958
	6-grams	1,195	2,860	6,027	10,405	14,537	16,470	15,245	66,739
	7-grams	466	1,223	2,615	5,196	8,395	11,003	12,239	41,137
	total	90,551	148,257	149,391	127,857	105,857	85,654	65,038	772,605

Table: Distribution of phrase pairs in the phrase translation table of GIZA++ (baseline zh→ja).

SMT Experiments

- Analysis of the results

		Target language							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source language	1-grams	23,752	**38,758**	**29,920**	**18,025**	9,671	4,777	2,211	127,114
	2-grams	38,997	**56,814**	**44,647**	**28,149**	16,733	9,539	4,985	199,864
	3-grams	23,240	38,360	**45,596**	**35,724**	25,148	16,653	9,977	194,698
	4-grams	10,954	19,398	29,343	**33,124**	28,991	22,801	16,078	160,689
	5-grams	4,779	9,143	15,540	21,864	25,515	24,242	20,052	121,135
	6-grams	1,858	3,787	7,475	12,683	17,799	20,577	19,996	84,175
	7-grams	765	1,577	3,357	6,531	10,335	13,969	15,787	52,321
	total	104,345	167,837	175,878	156,100	134,192	112,558	89,086	**939,996**

Table: Distribution of phrase pairs in the phrase translation table (baseline + extra training data: zh→ja). Compare with Table 13, increase in entry numbers in boldface.

SMT Experiments

- Analysis of the results

Baseline + quasi-parallel corpus	Baseline	只能这样了	|||	これ で 行く しか ない	|||
		只能这样了	|||	それ しか ない ん だ よ	|||
		只能这样了	|||	やる しか ない だろ	|||
	Increased	只能这样了	|||	これ で 行く しか なかっ た	|||
		只能这样了	|||	これ しか ない	|||
		只能这样了	|||	それ しか ない ん だ	|||
		只能这样了	|||	やる しか ない	|||
		只能这样了	|||	やる しか なかっ た	|||

Table: Samples of phrase alignments in zh→ja phrase table. The same Chinese phrase and corresponding Japanese phrases.

Contents

1 Motivation and Related Research

2 Data Preparation for the Experiments

3 Construction of Analogical Clusters

4 Generation of New Sentences Using Analogical Associationss

5 SMT Experiments

6 Synthesis of Results and Conclusion

Synthesis of Results and Conclusion

- We followed an original path for improving performance of PB-SMT by adding new training data (usually extracted from comparable corpora).
- We chose to add (110K to 186K) more data that may be not so well aligned and not so correct (quasi parallel corpus).
- On the same test set, the translation quality increased very significantly over the baseline systems (more than 6 BLEU points).

Thank you for listening.

ご清聴ありがとうございました。

谢谢大家。

References

- Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, 2005.
- Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, 2006.
- Yujie Zhang, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. Building an annotated Japanese-Chinese parallel corpus–a part of nict multilingual corpora. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 71–78, 2005.
- Yves Lepage and Etienne Denoual. Purest ever example-based machine translation: detailed presentation and assessment. *Machine Translation*, 19:251–282, 2005.
- Jean-François Lavallée and Philippe Langlais Morphological acquisition by formal analogy. In *Morpho Challenge 2009*, Corfu, Greece, oct 2009.

References

- Peter D. Turney and Michael L. Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278, 2005.
- Yves Lepage. Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Conference of the Association Proceedings of the 36th Annual Conference of the Association for Computational Linguistics (COLING-ACL'98)*, pages 728–735, August 1998.
- Lloyd Allison and Trevor I. Dix. A bit string longest common subsequence algorithm. *Information Processing Letter*, 23:305–310, 1986.
- Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Lausanne et Paris, [1ère éd. 1916] edition, 1995.
- Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL-2003)*, pages 71–78, 2003.

References

- George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference (HLT2002)*, pages 128– 132, San Diego, CA, USA, 2002. Morgan Kaufmann.
- Yves Lepage and Etienne Denoual. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *the 3rd International Workshop on Paraphrasing (IWP2005)*, pages 57–64, 2005.
- Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21:168–173, 1974.