Improving Bilingual Sub-sentential Alignment by Sampling-based Transpotting

Li Gong, Aurélien Max, François Yvon

LIMSI-CNRS & Université Paris-Sud Orsay, France







Experimental Results

Context of this work

Building SMT systems, step 1 : align parallel corpus



- parallel corpus can be huge
- we don't use/need everything
- we may regularly receive new data

Our method for parallel corpus alignment

- is very simple to describe and implement
- processes each sentence pair independently
- uses new data transparently (plug-and-play)

Experimental Results

Context of this work

Building SMT systems, step 1 : align parallel corpus



- parallel corpus can be huge
- we don't use/need everything
- we may regularly receive new data

Our method for parallel corpus alignment

- is very simple to describe and implement
- processes each sentence pair independently
- uses new data transparently (plug-and-play)

Method

Sampling-based transpotting Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task Incremental alignment task



Method

Sampling-based transpotting Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task Incremental alignment task

Outline

Method Sampling-based transpotting

Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task Incremental alignment task

one diet coke , please . \leftrightarrow un coca zéro , s'il vous plaît .

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- ④ Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

 Given a source-target sentence pair, extract an association table :

```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- A Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet coke	[1, 0, 1] [0, 0, 0] [0, 0, 0]		Frailab	French
, please un coca	[1, 0, 1] [1, 0, 0] [1, 1, 1] [1, 0, 1] [0, 0, 0]	1 2 3	one coffee , please . the coffee is not bad . yes , one tea .	un café , s'il vous plaît . ce café est correct . oui , un thé .

one diet coke , please . \leftrightarrow un coca zéro , s'il vous plaît .

```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- A Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair



```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- A Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet	one diet [1, 0, 1] [0, 0, 0]	coke	e , please . \leftrightarrow un coca ze	éro , s'il vous plaît .
	[0, 0, 0]		English	French
, please	[1, 0, 1] [1, 0, 0] [1, 1, 1]	1 2 3	one coffee , please . the coffee is not bad .	un café , s'il vous plaît . ce café est correct .
	[1, 0, 1] [0, 0, 0] 	0	yes, one tea .	

```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet	one diet [1, 0, 1] [0, 0, 0]	coke	\mathbf{e} , please . \leftrightarrow un coca ze	éro , s'il vous plaît .
соке , olease	[0, 0, 0] [1, 0, 1] [1, 0, 0]	1	English one coffee , please . the coffee is not bad	French un café , s'il vous plaît . ce café est correct .
	[1, 1, 1] [1, 0, 1] [0, 0, 0]	3	yes , one tea .	oui , un thé .
	[1, 1, 1] [1, 0, 1] [0, 0, 0]	3	yes , one tea .	oui , un thé

```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet	one diet [1, 0, 1] [0, 0, 0]	coke	e , please . \leftrightarrow un coca zé	iro , s'il vous plaît .
coke , please un coca	$\begin{bmatrix} 0, 0, 0 \\ [1, 0, 1] \\ [1, 0, 0] \\ [1, 1, 1] \\ [1, 0, 1] \\ [0, 0, 0] \end{bmatrix}$	1 2 3	English one coffee , please . the coffee is not bad . yes , one tea .	French un café , s'il vous plaît . ce café est correct . oui , un thé .

```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet	one diet [1, 0, 1] [0, 0, 0]	coke	$e , {\sf please} . \leftrightarrow {\sf un} {\sf coca} {\sf ze}$	ero , s'il vous plaît .
coke , please un coca	[0, 0, 0] [1, 0, 1] [1, 0, 0] [1, 1, 1] [1, 0, 1] [0, 0, 0]	1 2 3	English one coffee , please . the coffee is not bad . yes , one tea .	French un café , s'il vous plaît . ce café est correct . oui , un thé .

```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet [1, 0, 1] [0, 0, 0]	coke	e , please . \leftrightarrow un coca zé	éro , s'il vous plaît .
[0, 0, 0]		English	French
[1, 0, 1] [1, 0, 0] [1, 1, 1] [1, 0, 1] [0, 0, 0]	1 2 3	one coffee , please . the coffee is not bad . yes , one tea .	un café , s'il vous plaît . ce café est correct . oui , un thé .
	one diet [1, 0, 1] [0, 0, 0] [0, 0, 0] [1, 0, 1] [1, 0, 0] [1, 1, 1] [1, 0, 1] [0, 0, 0]	one diet coke [1, 0, 1] [0, 0, 0] [1, 0, 1] 1 [1, 0, 0] 2 [1, 1, 1] 3 [1, 0, 1] [0, 0, 0]	one diet coke , please . \leftrightarrow un coca zé [1, 0, 1] [0, 0, 0] [0, 0, 0] [1, 0, 1] [1, 0, 0] [1, 1, 1] [1, 0, 1] [1, 0, 1] [1, 0, 1] [1, 0, 1] [1, 0, 0] 2 1 1 one coffee , please . the coffee is not bad . [1, 0, 1] [0, 0, 0] [1, 0, 0] [1, 0, 0] [2, 0] [1, 0, 1] [1, 0, 0] [2, 0] [3, 0] [4, 0] [4, 0] [5, 0]

 Given a source-target sentence pair, extract an association table :

```
one diet coke, please. \leftrightarrow un coca zéro, s'il vous plaît.
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word

	one diet d	ske, please	. ↔ ui	r coca zero, s il vous pi	an.
one	[1, 0, 1]	words wit	h sam	e distribution profile	profiles
ulet	[0, 0, 0]	one,	\leftrightarrow	un ,	[1, 0, 1]
соке	[0, 0, 0]	diet coke	\leftrightarrow	coca zéro	[0, 0, 0]
. ,	[1, 0, 1]	please	\leftrightarrow	s'il vous plaît	1.0.0
please	[1, 0, 0]	P	\leftrightarrow		[1 1 1]
	[1, 1, 1]	· · ·	L/ ell e t e		[', ', ']
un	[1, 0, 1]			$oke \leftrightarrow coca zero) += 1$	
c003	້ທີ່ທີ່ດ້			$e \leftrightarrow s'il vous plaît) += 1$	
coca	[0, 0, 0]			+= 1	

 Given a source-target sentence pair, extract an association table :

one diet coke , please . \leftrightarrow un coca zéro , s'il vous plaît .

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- 3 Increment the count for each contiguous phrase pairs
- A Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet coke, please \leftrightarrow un coca zero, sin vous plant.					
one	[1, 0, 1]	words wit	h same	distribution profile	profiles
ulet	[0, 0, 0]	one.	\leftrightarrow	un.	[1, 0, 1]
coke	[0. 0. 0]			, , , , , , , , , , , , , , , , ,	
	[4, 0, 4]	diet coke	\leftrightarrow	coca zero	[0, 0, 0]
, nlagog	[1, 0, 1]	please	\leftrightarrow	s'il vous plaît	[1, 0, 0]
please	[1, 0, 0]		\sim		[1 1 1]
	[1, 1, 1]	· · ·	~ /	·	[[', ', ']
		#	t(diet co	$ke \leftrightarrow coca zéro) += 1$	
un	[1, 0, 1]		(
0003		#	(piease	$e \leftrightarrow s$ ii vous plait) += 1	
coca	[0, 0, 0]	Ħ	(\rightarrow)	⊥– 1	
		"	(• \ / •)	· = ·	

one diet coke , please . \leftrightarrow un coca zéro , s'il vous plaît .

```
one diet coke, please. \leftrightarrow un coca zéro, s'il vous plaît.
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one diet coke , please . \leftrightarrow un coca zéro , s'il vous plaît .				
words with sam	ne distribution profile	profiles		
one, \leftrightarrow	un,	[1, 0, 1]		
diet coke \leftrightarrow	coca zéro	[0, 0, 0]		
please \leftrightarrow	s'il vous plaît	[1, 0, 0]		
」· ↔		[1, 1, 1]		
#(diet of	$coke \leftrightarrow coca zéro) += 1$			
#(pleas	se \leftrightarrow s'il vous plaît) += 1			
#(. ↔ .) += 1			
() 	et coke , please . \leftrightarrow un] words with sam] one , \leftrightarrow] diet coke \leftrightarrow] diet coke \leftrightarrow] diet coke \leftrightarrow] \leftrightarrow] \leftrightarrow] \leftrightarrow] $(diet of a a a a a a a a a a a a a a a a a a $	et coke , please . \leftrightarrow un coca zéro , s'il vous pl words with same distribution profile one , \leftrightarrow un , diet coke \leftrightarrow coca zéro please \leftrightarrow s'il vous plaît . \leftrightarrow #(diet coke \leftrightarrow coca zéro) += 1 #(please \leftrightarrow s'il vous plaît) += 1 #(. \leftrightarrow .) += 1		

```
one diet coke , please . 
 \leftrightarrow un coca zéro , s'il vous plaît .
```

- 2 Draw a random sub-corpus from the parallel corpus and compute profiles for each word
- Increment the count for each contiguous phrase pairs
- A Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

source phrase		target phrase	count
one	\leftrightarrow	un	830
coke	\leftrightarrow	coca	680
diet coke	\leftrightarrow	coca zéro	260
one diet coke	\leftrightarrow	un coca zéro	30
,	\leftrightarrow	,	900
please	\leftrightarrow	s'il vous plaît	160
	\leftrightarrow		980



Method Sampling-based transpotting Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task Incremental alignment task

Method	

	un	coca	zéro	,	s'il	vous	plaît	
one	0.846	ε	ε	ε	ε	ε	ε	ε
diet	ε	0.310	0.382	ε	ε	ε	ε	ε
coke	ε	0.738	0.132	ε	ε	ε	ε	ε
,	ε	ε	ε	0.624	ε	ε	ε	0.248
please	ε	ε	ε	ε	0.132	0.108	0.628	ε
	ε	ε	ε	0.102	ε	ε	ε	0.873

un coca zéro , s'il vous plaît .

one diet coke , please .









Method Experimental Results

Sub-sentential alignment : algorithm illustration



Straight rule

Inversion rule





Inversion rule



m

Inversion rule

m

Straight rule





Inversion rule

Straight rule



i m jStraight rule Invo



Inversion rule

 Method
 Experimental Results
 Conclusion and

 Soco
 Soco



Inversion rule

Straight rule

8/26





 Method
 Experimental Results
 Conclusion and future

 Sub-sentential alignment : algorithm illustration





m

1 Association score w(s, t) between source and target words :

$$w(s,t) = p(s|t) * p(t|s)$$

2 Segmentation criterion :

$$\operatorname{cut}(X,Y) = \operatorname{cut}(\bar{X},\bar{Y}) = W(X,\bar{Y}) + W(\bar{X},Y)$$



Outline

1 Method

Sampling-based transpotting Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task Incremental alignment task

Outline

1 Method

Sampling-based transpotting Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task

Incremental alignment task

Method

Experimental Results

Conclusion and future work

Basic alignment task : systems



Baseline : giza++ with default setting Our method : sba, drawing 1,000 sub-corpora per sentence pa

Method

Basic alignment task : systems



Baseline : giza++ with default setting Our method : sba, drawing 1,000 sub-corpora per sentence pair

Basic alignment task : systems



Baseline : giza++ with default setting Our method : sba, drawing 1,000 sub-corpora per sentence pair

Basic alignment task : data

Language pairs

- English to French (1 reference translation)
- French to English (7 reference translations)

Experimental Results

- Chinese to English (7 reference translations)
- Development and test set (from BTEC)

Corpus	#lines	Avg(#token _{en})	#token _{fr}	#token _{zh}
devel03	506	4,098 (16 refs)	4,220	3,435
test09	469	3,928 (7 refs)	4,023	3,031

Training Data

Corpus	# lines	#token _{en}	# token _{fr}	# token _{zh}
BTEC	20K	182K	207K	-
HIT	62K	600K	690K	590K

• English \rightarrow French (1 reference) :

	BTEC (in-domain)			HIT (out-of-domain)				
	BLEU		TER	# entries	BLEU		TER	# entries
giza++	45.68	76.26	37.03	360K	39.65	68.20	44.50	1,217K
sba	47.81		36.60	315K	39.70	68.45	43.56	921K

• French \rightarrow English (7 references) :

	BTEC (in-domain)			HIT (out-of-domain)				
				# entries				# entries
giza++			24.59		45.52			1,224K
			24.22	315K	45.34			

	HIT (out-of-domain)									
				# entries						
giza++										
	27.85									

• English \rightarrow French (1 reference) :

	BTEC (in-domain)			HIT (out-of-domain)				
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
giza++	45.68	76.26	37.03	360K	39.65	68.20	44.50	1,217K
sba	47.81	77.78	36.60	315K	39.70	68.45	43.56	921K

• French \rightarrow English (7 references) :

	BTEC (in-domain)				HIT (out-of-domain)			
				# entries				# entries
giza++			24.59		45.52			1,224K
			24.22	315K	45.34			

	HIT (out-of-domain)								
				# entries					
giza++									
	27.85								

• English \rightarrow French (1 reference) :

		BTEC (in-domain)				HIT (out-of-domain)			
		BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
giza	++	45.68	76.26	37.03	360K	39.65	68.20	44.50	1,217K
sba	L	47.81	77.78	36.60	315K	39.70	68.45	43.56	921K

• French \rightarrow English (7 references) :

	BTEC (in-domain)					HIT (out-of-domain)			
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries	
giza++	59.50	77.23	24.59	360K	45.52	68.58	33.99	1,224K	
sba	59.92	77.50	24.22	315K	45.34	69.59	33.79	937K	

	HIT (out-of-domain)								
				# entries					
giza++									
	27.85								

• English \rightarrow French (1 reference) :

	BTEC (in-domain)			HIT (out-of-domain)				
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
giza++	45.68	76.26	37.03	360K	39.65	68.20	44.50	1,217K
sba	47.81	77.78	36.60	315K	39.70	68.45	43.56	921K

• French \rightarrow English (7 references) :

	BTEC (in-domain)				HIT (out-of-domain)			
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
giza++	59.50	77.23	24.59	360K	45.52	68.58	33.99	1,224K
sba	59.92	77.50	24.22	315K	45.34	69.59	33.79	937K

	HIT (out-of-domain)					
	BLEU	oracle-BLEU TER # entrie				
giza++	27.88	51.69	50.76	1,139K		
sba	27.85	53.05	50.93	655K		

• English \rightarrow French (1 reference) :

		BTEC (in-domain)			HIT (out-of-domain)				
		BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
giza	++	45.68	76.26	37.03	360K	39.65	68.20	44.50	1,217K
sba	L	47.81	77.78	36.60	315K	39.70	68.45	43.56	921K

• French \rightarrow English (7 references) :

		BTEC (in-c	lomain)		HIT (out-of-domain)			
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
giza++	59.50	77.23	24.59	360K	45.52	68.58	33.99	1,224K
sba	59.92	77.50	24.22	315K	45.34	69.59	33.79	937K

	HIT (out-of-domain)					
	BLEU	oracle-BLEU TER # entries				
giza++	27.88	51.69	50.76	1,139K		
sba	27.85	53.05	50.93	655K		

Outline

1 Method

Sampling-based transpotting Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task Incremental alignment task

Method

Incremental alignment task : system & data



Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system. Supp PT : only contains entries of OOV words Baseline system : giza++/Moses on HIT corpus (French → English with 7 references)

Corpus	# lines	#token _{en}	# token _{fr}
HIT	62K	600K	690K
WMT	11,745K	317,688K	383,076K
	3.3K	111K	121K

Method

Incremental alignment task : system & data



Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

Supp PT : only contains entries of OOV words Baseline system : giza++/Moses on HIT corpus (French \rightarrow English with 7 references)

Corpus	# lines	#token _{en}	# token _{fr}
HIT	62K	600K	690K
WMT	11,745K	317,688K	383,076K
	3.3K	111K	121K

Incremental alignment task : system & data



Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

Supp PT : only contains entries of OOV words Baseline system : giza++/Moses on HIT corpus (French \rightarrow English with 7 references)

Corpus	# lines	#token _{en}	# token _{fr}
HIT	62K	600K	690K
WMT	11,745K	317,688K	383,076K
	3.3K	111K	121K

Incremental alignment task : system & data



Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

Supp PT : only contains entries of OOV words

Baseline system : giza++/Moses on HIT corpus (French \rightarrow English with 7 references

Corpus	# lines	#token _{en}	# token _{fr}
HIT	62K	600K	690K
WMT	11,745K	317,688K	383,076K
	3.3K	111K	121K

Incremental alignment task : system & data



Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

Supp PT : only contains entries of OOV words

Baseline system : giza++/Moses on HIT corpus

(French \rightarrow English with 7 references)

Corpus	# lines	#token _{en}	# token _{fr}
HIT	62K	600K	690K
WMT	11,745K	317,688K	383,076K
supp	3.3K	111K	121K

	Phrase ta		HIT			
main	sup	olementa	ry			
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	Δ -BLEU	TER
giza++	none	-	-	45.52	0	33.99
	forced	59	1,993	47.94	+2.42	34.62
	concat	60	1,190	48.69		33.09
	sba	64	681	49.83		30.61
Í	concat++	62	1,218	50.23		29.81
sba	none	-	-	45.34	-0.18	33.79
	sba	64	681	50.45	+4.93	29.94

none: baseline system

- forced: forced alignment, trained on HIT
- concat: giza++ alignment learnt on the concatenation of HIT and supp

sba: our sampling-based alignment method

	Phrase ta	ables			HIT	
main	sup	plementa	ry			
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	Δ -BLEU	TER
giza++	none	-	-	45.52	0	33.99
	forced	59	1,993	47.94	+2.42	34.62
	concat	60	1,190	48.69		33.09
Ì	sba	64	681	49.83		30.61
Í	concat++	62	1,218	50.23		29.81
sba	none	-	-	45.34	-0.18	33.79
	sba	64	681	50.45	+4.93	29.94

none: baseline system

forced: forced alignment, trained on HIT

sba: our sampling-based alignment method

	Phrase ta	ables			HIT	
main	sup	olementa	ry			
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	Δ -BLEU	TER
giza++	none	-	-	45.52	0	33.99
	forced	59	1,993	47.94	+2.42	34.62
l i	concat	60	1,190	48.69	+3.17	33.09
	sba	64	681	49.83		30.61
Í	concat++	62	1,218	50.23		29.81
sba	none	-	-	45.34	-0.18	33.79
	sba	64	681	50.45	+4.93	29.94

none: baseline system

forced: forced alignment, trained on HIT

sba: our sampling-based alignment method

	Phrase ta	ables			HIT	
main	sup	plementa	ry			
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	Δ -BLEU	TER
giza++	none	-	-	45.52	0	33.99
	forced	59	1,993	47.94	+2.42	34.62
	concat	60	1,190	48.69	+3.17	33.09
l ì	sba	64	681	49.83	+4.31	30.61
Í	concat++	62	1,218	50.23		29.81
sba	none	-	-	45.34	-0.18	33.79
	sba	64	681	50.45	+4.93	29.94

none: baseline system

forced: forced alignment, trained on HIT

- - sba: our sampling-based alignment method

	Phrase ta	ables			HIT	
main	sup	plementa	ry			
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	Δ -BLEU	TER
giza++	none	-	-	45.52	0	33.99
	forced	59	1,993	47.94	+2.42	34.62
	concat	60	1,190	48.69	+3.17	33.09
l ì	sba	64	681	49.83	+4.31	30.61
	concat++	62	1,218	50.23	+4.71	29.81
sba	none	-	-	45.34	-0.18	33.79
	sba	64	681	50.45	+4.93	29.94

none: baseline system

forced: forced alignment, trained on HIT

sba: our sampling-based alignment method

	Phrase ta	ables			HIT	
main	sup	plementa	ry			
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	Δ -BLEU	TER
giza++	none	-	-	45.52	0	33.99
	forced	59	1,993	47.94	+2.42	34.62
	concat	60	1,190	48.69	+3.17	33.09
l ì	sba	64	681	49.83	+4.31	30.61
	concat++	62	1,218	50.23	+4.71	29.81
sba	none	-	-	45.34	-0.18	33.79
	sba	64	681	50.45	+4.93	29.94

none: baseline system

forced: forced alignment, trained on HIT

- - sba: our sampling-based alignment method
- concat++: giza++ alignment learnt on the WMT corpus

Method 00

Incremental alignment task : results

main	Phrase tables main supplementary					
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	Δ -BLEU	TER
giza++	none	-	-	45.52	0	33.99
	forced	59	1,993	47.94	+2.42	34.62
	concat	60	1,190	48.69	+3.17	33.09
	sba	64	681	49.83	+4.31	30.61
	concat++	62	1,218	50.23	+4.71	29.81
sba	none	-	-	45.34	-0.18	33.79
	sba	64	681	50.45	+4.93	29.94

none: baseline system

- forced: forced alignment, trained on HIT
- - sba: our sampling-based alignment method
- concat++: giza++ alignment learnt on the WMT corpus

Outline

1 Method

Sampling-based transpotting Sub-sentential alignment extraction

2 Experimental Results

Basic alignment task Incremental alignment task

Conclusion

Conclusion

- alignment time can be controlled
- only useful sentence pairs need to be aligned
- integrating new data is plug-and-play

Previous work

- sampling-based transpotting (Anymalign) (Lardilleux and Lepage, 2008)
- Inversion Transduction Grammars (Wu, 1997)
- binary recursive bi-sentence segmentation (Lardilleux, Yvon and Lepage, 2012)

Experimental Results

Conclusion and future work

Hypothesis : sba performs better on rare words



With this framework, we can :

- perform the alignment process and phrase table construction on a **per-need basis**
- work at the level of tera-scale translation using huge quantities of unaligned parallel corpora
- perform domain adaptation by careful example selection

Thank you !

Sub-sentential alignment : details

1 Association score w(s, t) between source and target words :

$$w(s,t) = p(s|t) * p(t|s)$$

2 Segmentation criterion :

$$\operatorname{cut}(X,Y) = \operatorname{cut}(\bar{X},\bar{Y}) = W(X,\bar{Y}) + W(\bar{X},Y)$$



To avoid unbalanced segmentation, we use instead a normalized variant :

$$\mathsf{Ncut}(X,Y) = \frac{\mathsf{cut}(X,Y)}{\mathsf{cut}(X,Y)+2 \times W(X,Y)} + \frac{\mathsf{cut}(\bar{X},\bar{Y})}{\mathsf{cut}(\bar{X},\bar{Y})+2 \times W(\bar{X},\bar{Y})}$$

Sub-sentential alignment

- The greedy strategy is used to find the best segmentation point and the direction (direct or swap).
- 2 The recursive procedure ends when the source or target segment contains only one word.

```
 \begin{array}{l} \textbf{procedure } align(S,T):\\ \textbf{if} length(S) = 1 \textbf{ or } length(T) = 1:\\ link each word of S to each word of T\\ \textbf{stop procedure}\\ minNcut = 2\\ (X,Y) = (S,T)\\ \textbf{for each} (i,j) \in \{2...I\} \times \{2...J\}:\\ \textbf{if Neut}(A,B) < minNcut :\\ minNcut = Neut(A,B)\\ (X,Y) = (A,B)\\ \textbf{if Neut}(A,\bar{B} < minNcut :\\ minNcut = Neut(A,\bar{B})\\ (X,Y) = (A,\bar{B})\\ \textbf{if Neut}(A,\bar{B} < minNcut :\\ (X,Y) = (A,\bar{B})\\ align(X,Y)\\ align(X,Y) \end{array}
```

Sub-sentential alignment

- The greedy strategy is used to find the best segmentation point and the direction (direct or swap).
- Provide the source or target segment contains only one word.

```
 \begin{aligned} & \textbf{procedure } \text{align}(S,T): \\ & \textbf{if} | \text{length}(S) = 1 \textbf{ or } | \text{length}(T) = 1: \\ & \text{link each word of } S \text{ to each word of } T \\ & \textbf{stop procedure} \\ & minNcut = 2 \\ & (X,Y) = (S,T) \\ \hline & \textbf{for each}(i,j) \in \{2...,I\} \times \{2...,I\}: \\ & \textbf{if} \text{Ncut}(A,B) < minNcut : \\ & minNcut = \text{Ncut}(A,B) \\ & (X,Y) = (A,B) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & minNcut = \text{Ncut}(A,\bar{B}) \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut : \\ & \textbf{if} \text{Ncut}(A,\bar{B} < minNcut :
```

Sub-sentential alignment

- The greedy strategy is used to find the best segmentation point and the direction (direct or swap).
- 2 The recursive procedure ends when the source or target segment contains only one word.

```
\begin{array}{l} \textbf{procedure } align(S,T):\\ \textbf{if} length(S) = 1 \textbf{ or } length(T) = 1:\\ link each word of S to each word of T\\ \textbf{stop procedure}\\ \hline minNcut = 2\\ (X,Y) = (S,T)\\ \textbf{for each } (i,j) \in \{2...I\} \times \{2...J\}:\\ \textbf{if Ncut}(A,B) < minNcut :\\ minNcut = Ncut(A,B)\\ (X,Y) = (A,B)\\ \textbf{if Ncut}(A,\bar{B} \ minNcut :\\ minNcut = Ncut(A,\bar{B})\\ (X,Y) = (A,\bar{B})\\ align(X,Y)\\ align(X,Y) \end{array}
```

All results

		BTE	С			HIT				BTEC+	HIT	
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
					Eng	glish→French	(1 refer	rence)				
giza++	45.68	76.26	37.03	360K	39.65	68.20	44.50	1,217K	47.97	83.62	35.45	1,546K
sba	47.81	77.78	36.60	315K	39.70	68.45	43.56	921K	47.55	84.40	37.22	1,241K
					Frei	nch→English	(7 refer	ences)				
giza++	59.50	77.23	24.59	360K	45.52	68.58	33.99	1,224K	63.69	84.00	21.95	1,551K
sba	59.92	77.50	24.22	315K	45.34	69.59	33.79	937K	64.44	83.57	22.31	1,241K
					Chin	ese→English	(7 refe	rences)				
giza++	-	-	-	-	27.88	51.69	50.76	1,139K	-	-	-	
sba	-	-	-	-	27.85	53.05	50.93	655K	-	-	-	-

	Phrase ta	ables				Н	IT		
main	sup	supplementary							
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	Зg	4g	TER
giza++	none	-	-	45.52	76.5	52.2	37.8	27.1	33.99
	forced	59	1,993	47.94	76.8	55.4	41.0	29.2	34.62
	concat	60	1,190	48.69	78.4	56.1	41.4	29.8	33.09
	sba	64	681	49.83	80.9	57.3	42.0		30.61
İ	concat++	62	1,218	50.23	81.5	57.8	42.6	31.1	29.81
sba	none	-	-	45.34	77.0	52.1	37.4	26.9	33.79
	sba	64	681	50.45	81.8		42.5	30.9	29.94

none: baseline system

- forced: forced alignment, trained on HIT
- concat: giza++ alignment learnt on the concatenation of HIT and supp
 - sba: our sampling-based alignment method
- concat++: giza++ alignment learnt on the corpus WMT

	Phrase ta	ables				Н	IT		
main	sup	supplementary							
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	Зg	4g	TER
giza++	none	-	-	45.52	76.5	52.2	37.8	27.1	33.99
	forced	59	1,993	47.94	76.8	55.4	41.0	29.2	34.62
	concat	60	1,190	48.69	78.4	56.1	41.4	29.8	33.09
	sba	64	681	49.83	80.9	57.3	42.0		30.61
İ	concat++	62	1,218	50.23	81.5	57.8	42.6	31.1	29.81
sba	none	-	-	45.34	77.0	52.1	37.4	26.9	33.79
	sba	64	681	50.45	81.8		42.5	30.9	29.94

none: baseline system

forced: forced alignment, trained on HIT

- concat: giza++ alignment learnt on the concatenation of HIT and supp
 - sba: our sampling-based alignment method
- concat++: giza++ alignment learnt on the corpus WMT

	Phrase ta	ables				Н	IT		
main	sup	supplementary							
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	Зg	4g	TER
giza++	none	-	-	45.52	76.5	52.2	37.8	27.1	33.99
	forced	59	1,993	47.94	76.8	55.4	41.0	29.2	34.62
l i	concat	60	1,190	48.69	78.4	56.1	41.4	29.8	33.09
	sba	64	681	49.83	80.9	57.3	42.0		30.61
İ	concat++	62	1,218	50.23	81.5	57.8	42.6	31.1	29.81
sba	none	-	-	45.34	77.0	52.1	37.4	26.9	33.79
	sba	64	681	50.45	81.8		42.5	30.9	29.94

- none: baseline system
- forced: forced alignment, trained on HIT

sba: our sampling-based alignment method

	Phrase ta	ables				Н	IT		
main	sup	supplementary							
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	Зg	4g	TER
giza++	none	-	-	45.52	76.5	52.2	37.8	27.1	33.99
	forced	59	1,993	47.94	76.8	55.4	41.0	29.2	34.62
l i	concat	60	1,190	48.69	78.4	56.1	41.4	29.8	33.09
l i	sba	64	681	49.83	80.9	57.3	42.0	30.5	30.61
İ	concat++	62	1,218	50.23	81.5	57.8	42.6	31.1	29.81
sba	none	-	-	45.34	77.0	52.1	37.4	26.9	33.79
	sba	64	681	50.45	81.8		42.5	30.9	29.94

- none: baseline system
- forced: forced alignment, trained on HIT
- - sba: our sampling-based alignment method

	Phrase ta	ables		HIT					
main	sup	olementa	ry						
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	Зg	4g	TER
giza++	none	-	-	45.52	76.5	52.2	37.8	27.1	33.99
	forced	59	1,993	47.94	76.8	55.4	41.0	29.2	34.62
	concat	60	1,190	48.69	78.4	56.1	41.4	29.8	33.09
	sba	64	681	49.83	80.9	57.3	42.0	30.5	30.61
İ	concat++	62	1,218	50.23	81.5	57.8	42.6	31.1	29.81
sba	none	-	-	45.34	77.0	52.1	37.4	26.9	33.79
	sba	64	681	50.45	81.8	58.3	42.5	30.9	29.94

- none: baseline system
- forced: forced alignment, trained on HIT
- - sba: our sampling-based alignment method
- concat++: giza++ alignment learnt on the corpus WMT

Phrase tables					HIT					
main	supplementary									
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	Зg	4g	TER	
giza++	none	-	-	45.52	76.5	52.2	37.8	27.1	33.99	
	forced	59	1,993	47.94	76.8	55.4	41.0	29.2	34.62	
	concat	60	1,190	48.69	78.4	56.1	41.4	29.8	33.09	
	sba	64	681	49.83	80.9	57.3	42.0	30.5	30.61	
	concat++	62	1,218	50.23	81.5	57.8	42.6	31.1	29.81	
sba	none	-	-	45.34	77.0	52.1	37.4	26.9	33.79	
	sba	64	681	50.45	81.8	58.3	42.5	30.9	29.94	

- none: baseline system
- forced: forced alignment, trained on HIT
- - sba: our sampling-based alignment method
- concat++: giza++ alignment learnt on the corpus WMT

Hypothesis

SBA performs better on rare words.

