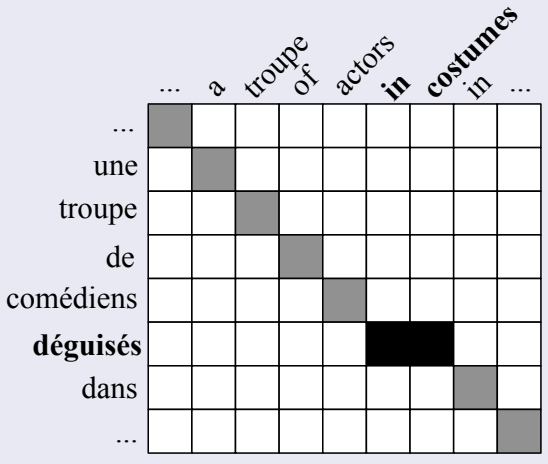
Improving Bilingual Sub-sentential Alignment by Sampling-based Transpotting

Li Gong, Aurélien Max, François Yvon

LIMSI-CNRS & Université Paris-Sud
Orsay, France

Context of this work

Building SMT systems, step 1 : align parallel corpus



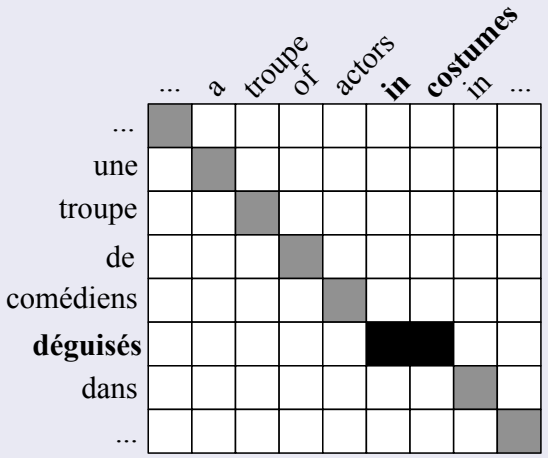
- parallel corpus can be huge
- we don't use/need everything
- we may regularly receive new data

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**)

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**)

Outline

Outline

Outline

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

	English	French
1	one coffee , please .	un café , s'il vous plaît .
2	the coffee is not bad .	ce café est correct .
3	yes , one tea .	oui , un thé .

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

	English	French
1	one coffee , please .	un café , s'il vous plaît .
2	the coffee is not bad .	ce café est correct .
3	yes , one tea .	oui , un thé .

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

	English	French
1	one coffee , please .	un café , s'il vous plaît .
2	the coffee is not bad .	ce café est correct .
3	yes , one tea .	oui , un thé .

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

	English	French
1	one coffee , please .	un café , s'il vous plaît .
2	the coffee is not bad .	ce café est correct .
3	yes , one tea .	oui , un thé .

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

	English	French
1	one coffee , please .	un café , s'il vous plaît .
2	the coffee is not bad .	ce café est correct .
3	yes , one tea .	oui , un thé .

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

	English	French
1	one coffee , please .	un café , s'il vous plaît .
2	the coffee is not bad .	ce café est correct .
3	yes , one tea .	oui , un thé .

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

	English	French
1	one coffee , please .	un café , s'il vous plaît .
2	the coffee is not bad .	ce café est correct .
3	yes , one tea .	oui , un thé .

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

words with same distribution profile			**profiles**
one ,	↔	un ,	[1, 0, 1]
diet coke	↔	coca zéro	[0, 0, 0]
please	↔	s'il vous plaît	[1, 0, 0]
.	↔	.	[1, 1, 1]

#(diet coke ↔ coca zéro) += 1
#(please ↔ s'il vous plaît) += 1
#(. ↔ .) += 1

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 N times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

words with same distribution profile			**profiles**
one ,	↔	un ,	[1, 0, 1]
diet coke	↔	coca zéro	[0, 0, 0]
please	↔	s'il vous plaît	[1, 0, 0]
.	↔	.	[1, 1, 1]

#(diet coke ↔ coca zéro) += 1
#(please ↔ s'il vous plaît) += 1
#(. ↔ .) += 1

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. Repeat steps 2 to 3 *N* times, so as to obtain an association table for the given sentence pair

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

one	[1, 0, 1]
diet	[0, 0, 0]
coke	[0, 0, 0]
,	[1, 0, 1]
please	[1, 0, 0]
.	[1, 1, 1]
un	[1, 0, 1]
coca	[0, 0, 0]
...	...

words with same distribution profile			**profiles**
one ,	↔	un ,	[1, 0, 1]
diet coke	↔	coca zéro	[0, 0, 0]
please	↔	s'il vous plaît	[1, 0, 0]
.	↔	.	[1, 1, 1]

#(diet coke ↔ coca zéro) += 1
#(please ↔ s'il vous plaît) += 1
#(. ↔ .) += 1

Sampling-based transpotting

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

   one diet coke , please . ↔ 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
4. 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	↔	un	830
coke	↔	coca	680
diet coke	↔	coca zéro	260
one diet coke	↔	un coca zéro	30
,	↔	,	900
please	↔	s'il vous plaît	160
.	↔	.	980

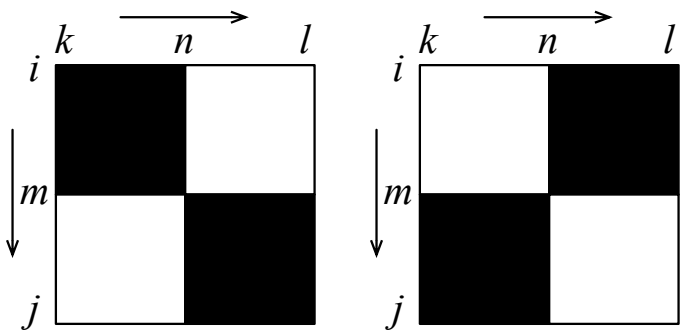
Outline

Sub-sentential alignment : algorithm illustration

	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 .

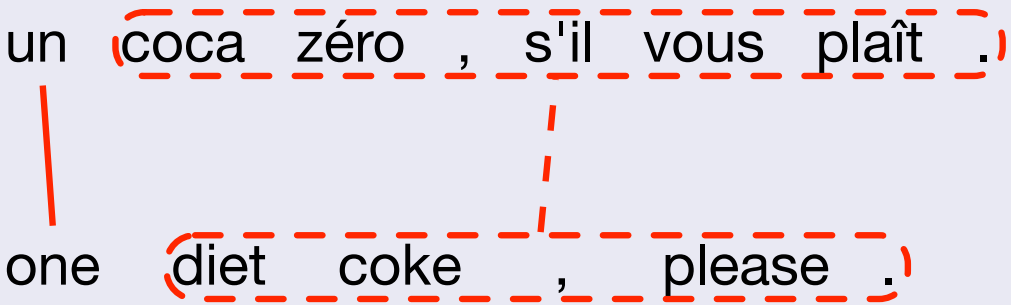
one diet coke , please .

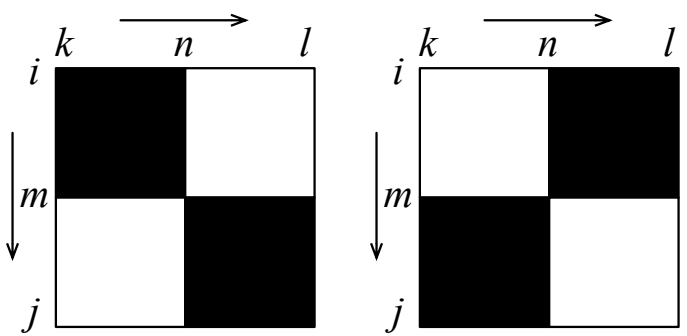


Straight rule Inversion rule

Sub-sentential alignment : algorithm illustration

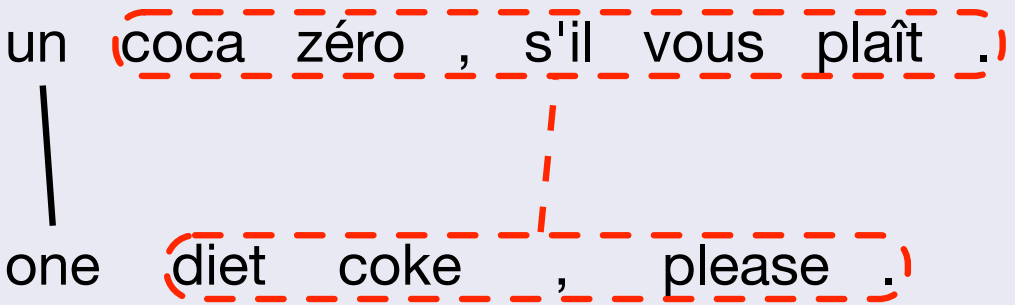
	un	coca	zéro	,	s'il	vous	plaît	.
one	0.846	ε	ε	ε	ε	ε	ε	ε
diet	ε	0.310	0.382	ε	ε	ε	ε	ε
coke	ε	0.738	0.132	ε	ε	ε	ε	ε
,	ε	ε	ε	0.624	ε	ε	ε	0.248
please	ε	ε	ε	ε	0.132	0.108	0.628	ε
.	ε	ε	ε	0.102	ε	ε	ε	0.873

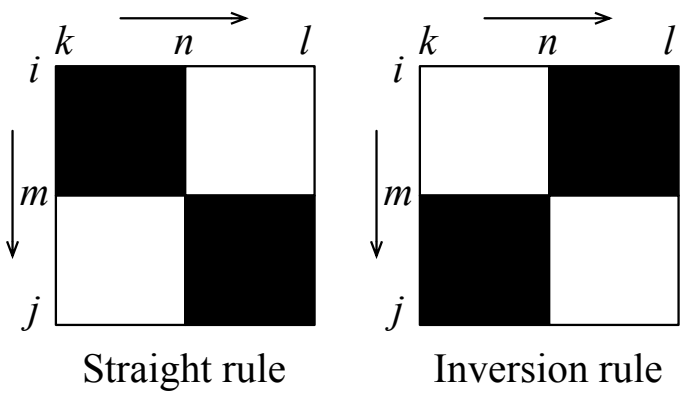




Sub-sentential alignment : algorithm illustration

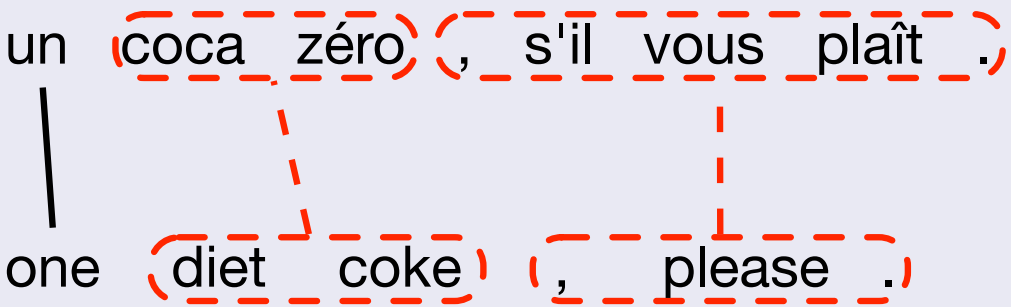
	un	coca	zéro	,	s'il	vous	plaît	.
one								
diet		0.310	0.382	ε	ε	ε	ε	ε
coke		0.738	0.132	ε	ε	ε	ε	ε
,		ε	ε	0.624	ε	ε	ε	0.248
please		ε	ε	ε	0.132	0.108	0.628	ε
.		ε	ε	0.102	ε	ε	ε	0.873

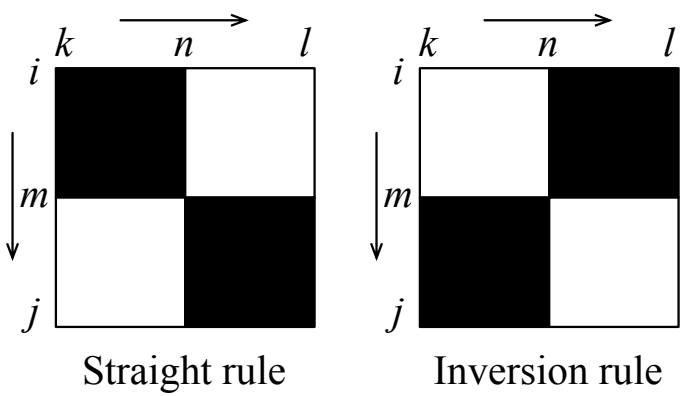




Sub-sentential alignment : algorithm illustration

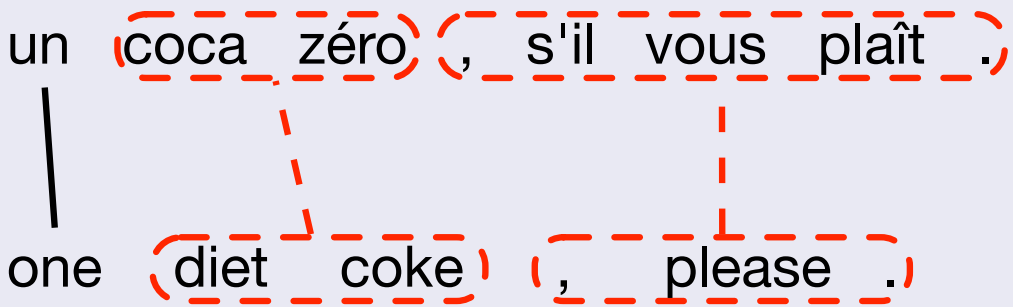
	un	coca	zéro	,	s'il	vous	plaît	.
one								
diet		0.310	0.382	ε	ε	ε	ε	ε
coke		0.738	0.132	ε	ε	ε	ε	ε
,		ε	ε	0.624	ε	ε	ε	0.248
please		ε	ε	ε	0.132	0.108	0.628	ε
.		ε	ε	0.102	ε	ε	ε	0.873

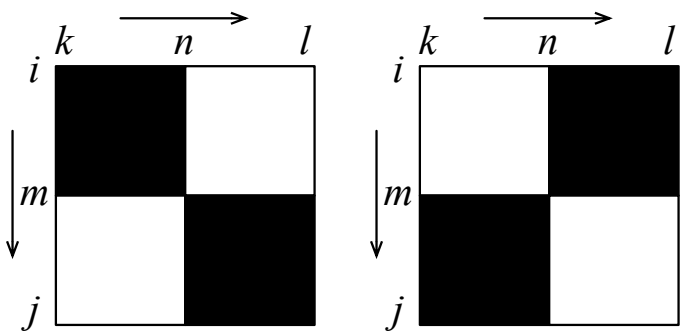




Sub-sentential alignment : algorithm illustration

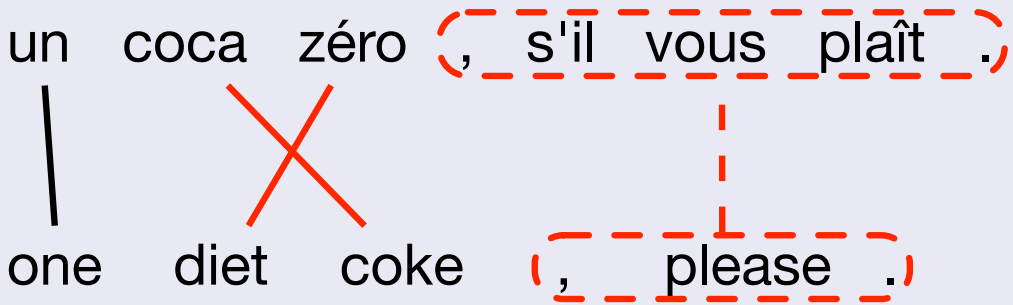
	un	coca	zéro	,	s'il	vous	plaît	.
one								
diet		0.310	0.382					
coke		0.738	0.132					
,				0.624	ε	ε	ε	0.248
please				ε	0.132	0.108	0.628	ε
.				0.102	ε	ε	ε	0.873

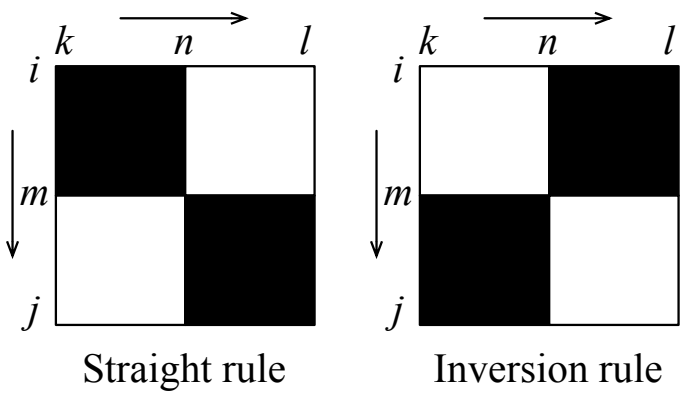




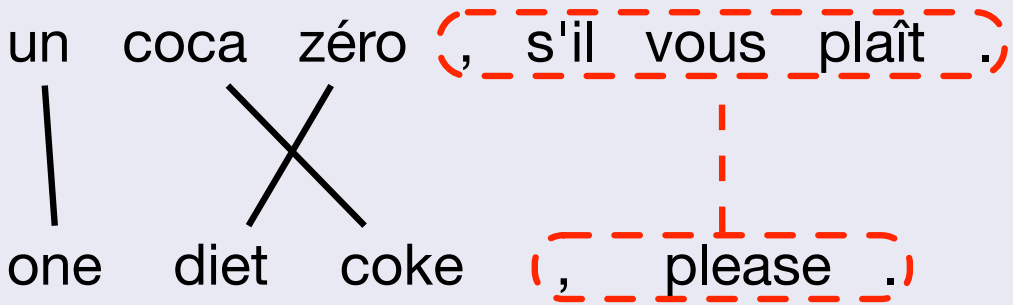
Sub-sentential alignment : algorithm illustration

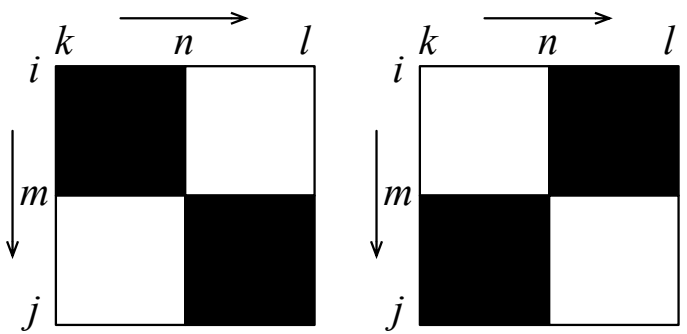
	un	coca	zéro	,	s'il	vous	plaît	.
one								
diet		0.310	0.382					
coke		0.738	0.132					
,				0.624	ε	ε	ε	0.248
please				ε	0.132	0.108	0.628	ε
.				0.102	ε	ε	ε	0.873





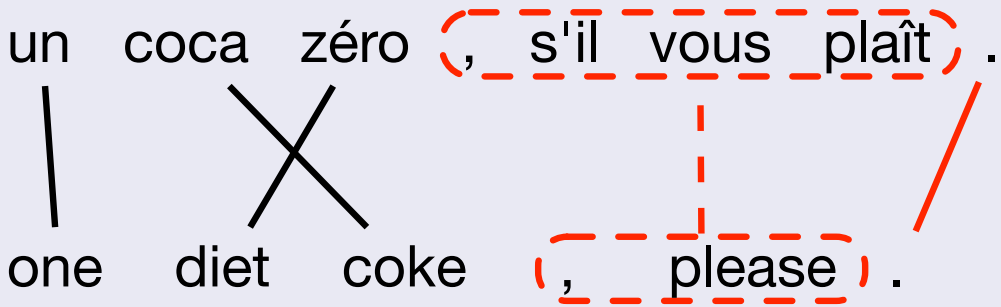
Sub-sentential alignment : algorithm illustration

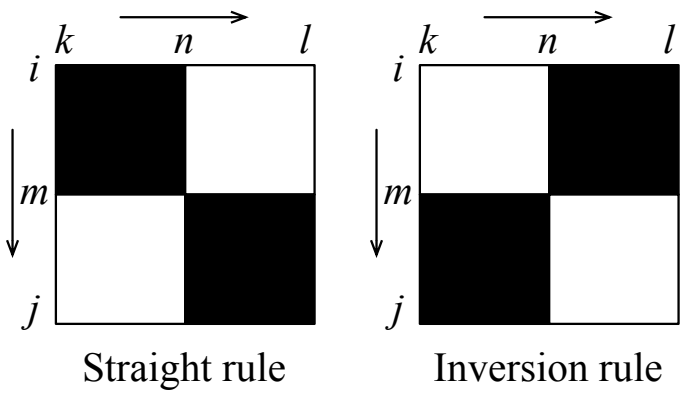




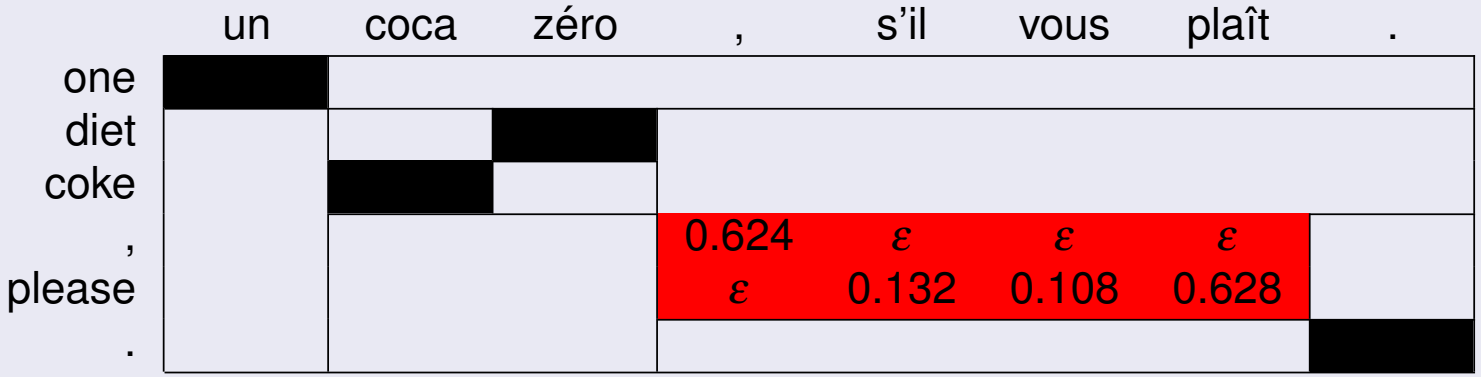
Sub-sentential alignment : algorithm illustration

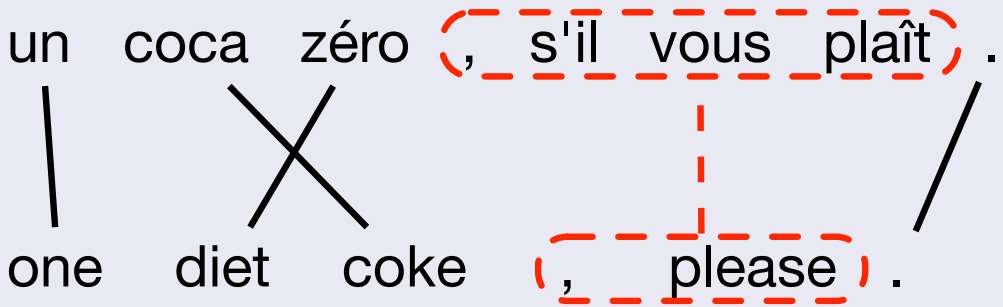
	un	coca	zéro	,	s'il	vous	plaît	.
one	■							
diet			■					
coke		■						
,				0.624	ε	ε	ε	0.248
please				ε	0.132	0.108	0.628	ε
.				0.102	ε	ε	ε	0.873

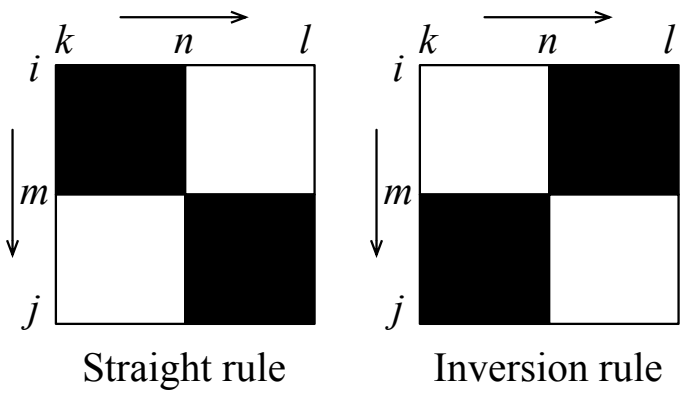




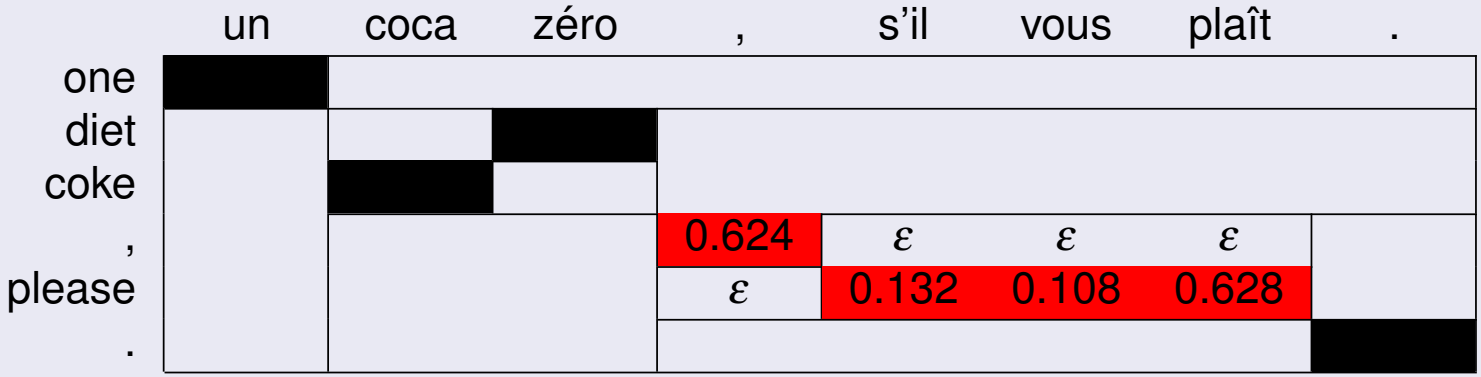
Sub-sentential alignment : algorithm illustration

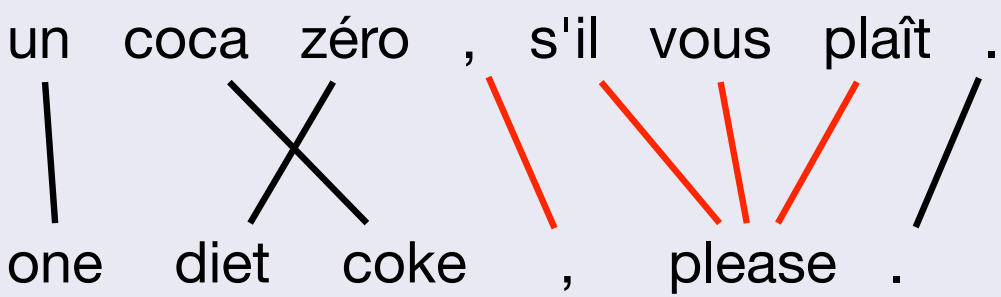


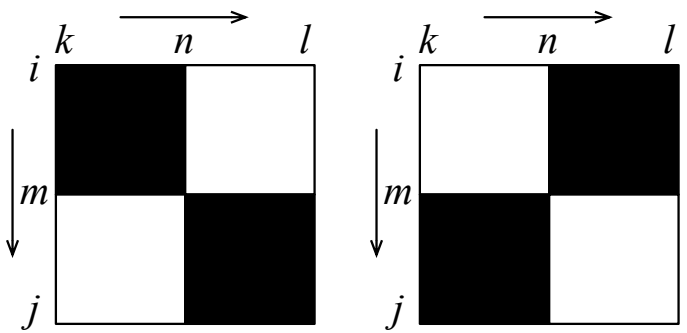




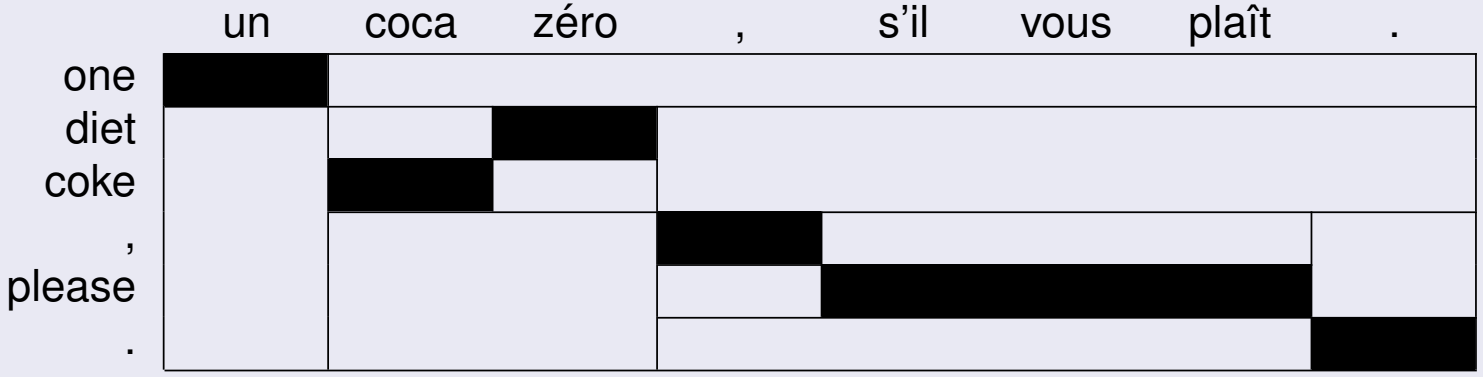
Sub-sentential alignment : algorithm illustration

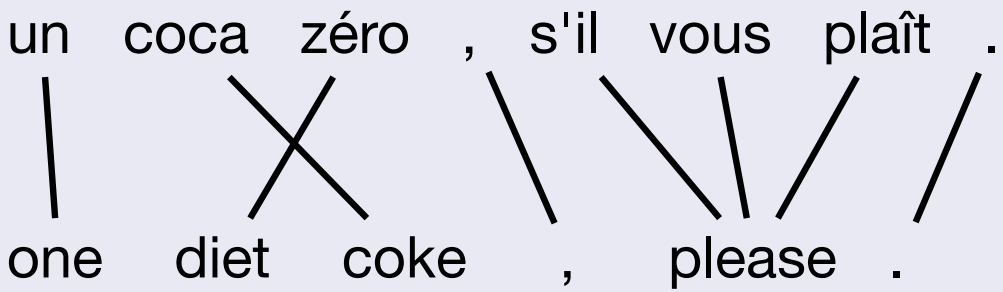


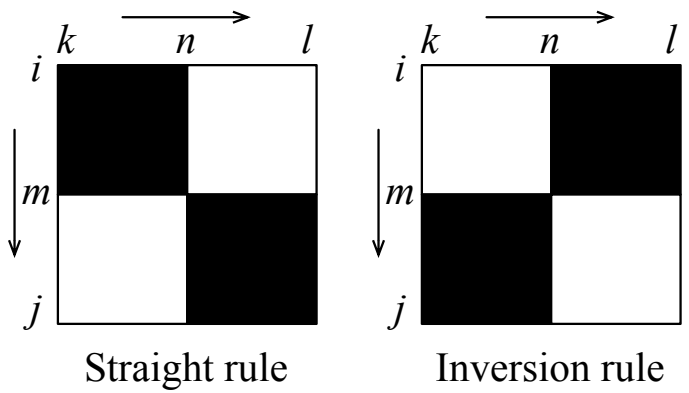




Sub-sentential alignment : algorithm illustration







Sub-sentential alignment : details

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

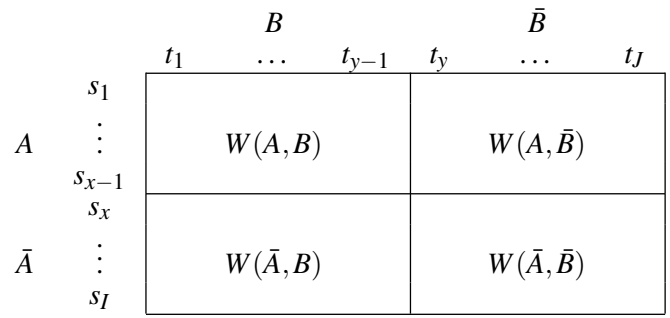
$$w(s,t) = p(s|t) * p(t|s)$$

2. Segmentation criterion :

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

where :

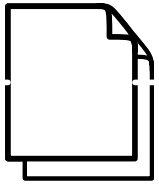
- $(X,Y) \in \{A,\bar{A}\} \times \{B,\bar{B}\}$
- $W(X,Y) = \sum_{s \in X, t \in Y} w(s,t)$



Outline

Outline

Basic alignment task : systems

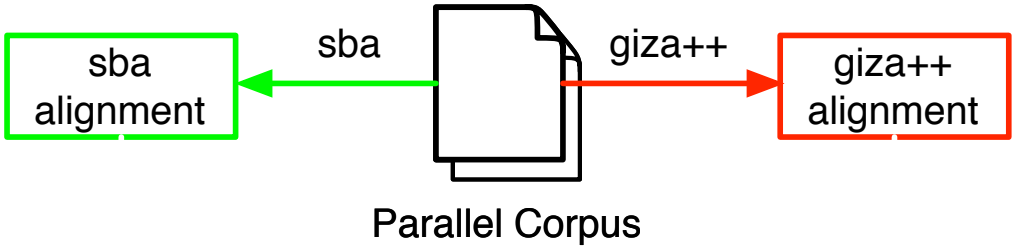


Parallel Corpus

Baseline : `giza++` with default setting

Our method : `sba`, drawing 1,000 sub-corpora per sentence pair

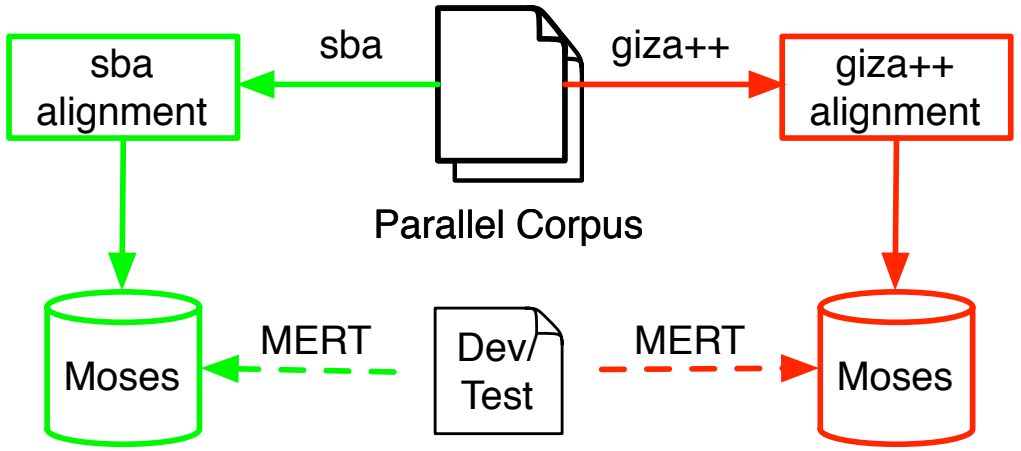
Basic alignment task : systems



Baseline : `giza++` with default setting

Our method : `sba`, drawing 1,000 sub-corpora per sentence pair

Basic alignment task : 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)
  - 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

Basic alignment task : results

- English → 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 → 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

- Chinese → English (7 references) :

	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

Basic alignment task : results

- English → 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 → 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

- Chinese → English (7 references) :

	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

Basic alignment task : results

- English → 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 → 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

- Chinese → English (7 references) :

	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

Basic alignment task : results

- English → 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 → 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

- Chinese → English (7 references) :

	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

Basic alignment task : results

- English → 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 → English (7 references) :

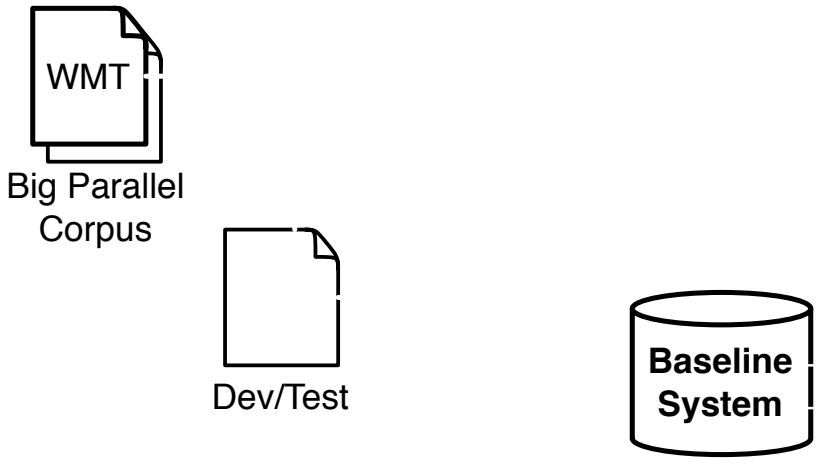
	BTEC (in-domain)				**HIT** (out-of-domain)			
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
giza++	59.50	77.23	24.59	360K	45.52	68.58	33.99	1,224K
sba	59.92	77.50	24.22	315K	45.34	69.59	33.79	937K

- Chinese → English (7 references) :

	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

Incremental alignment task : system & data



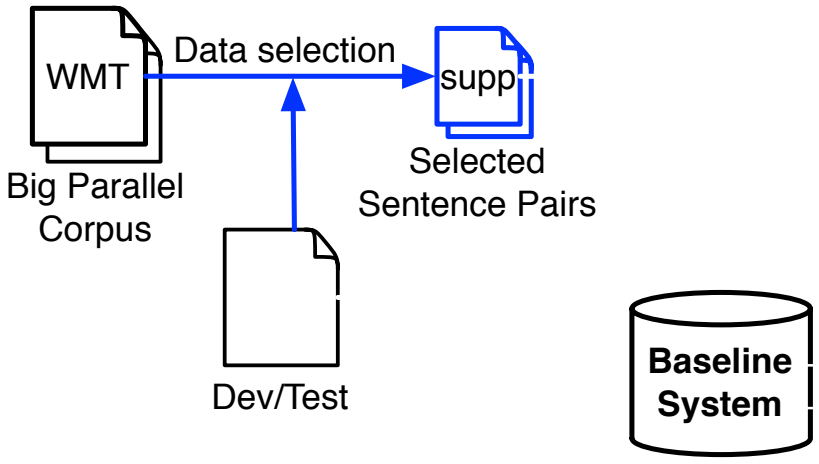
Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

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

Incremental alignment task : system & data



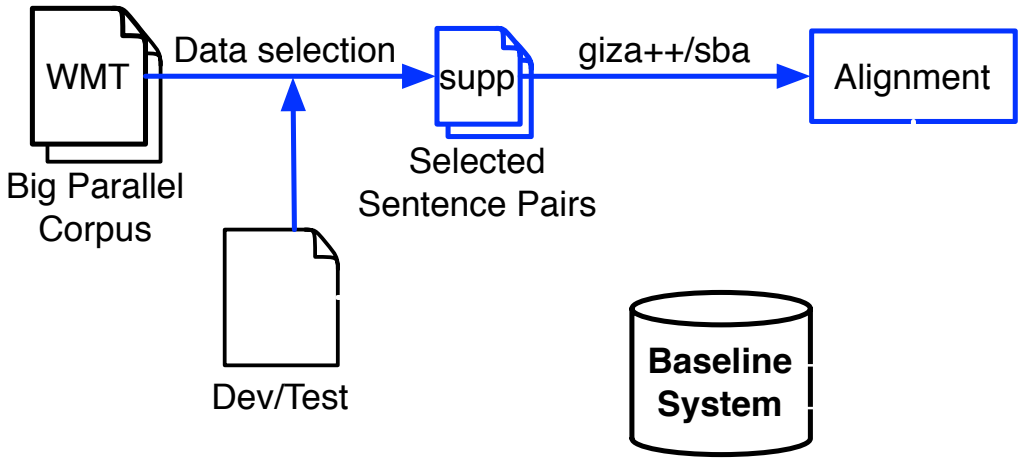
Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

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

Incremental alignment task : system & data



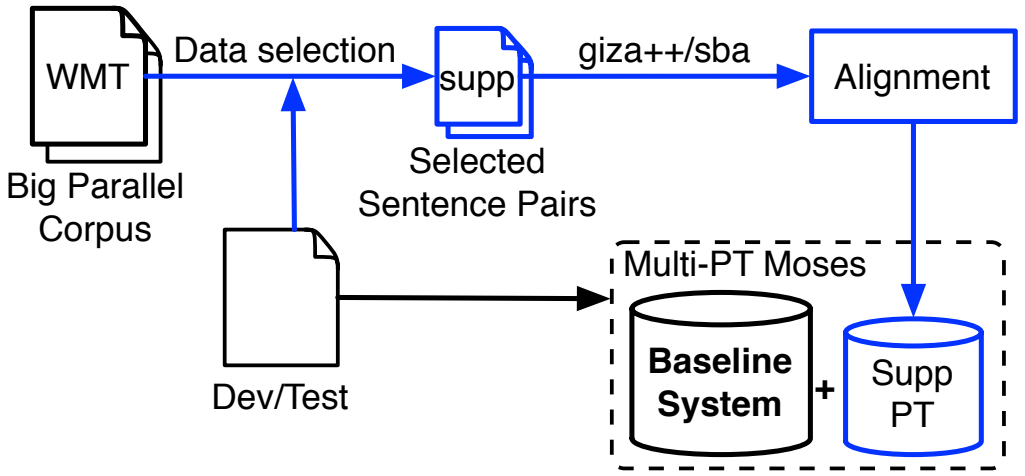
Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

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

Incremental alignment task : system & data



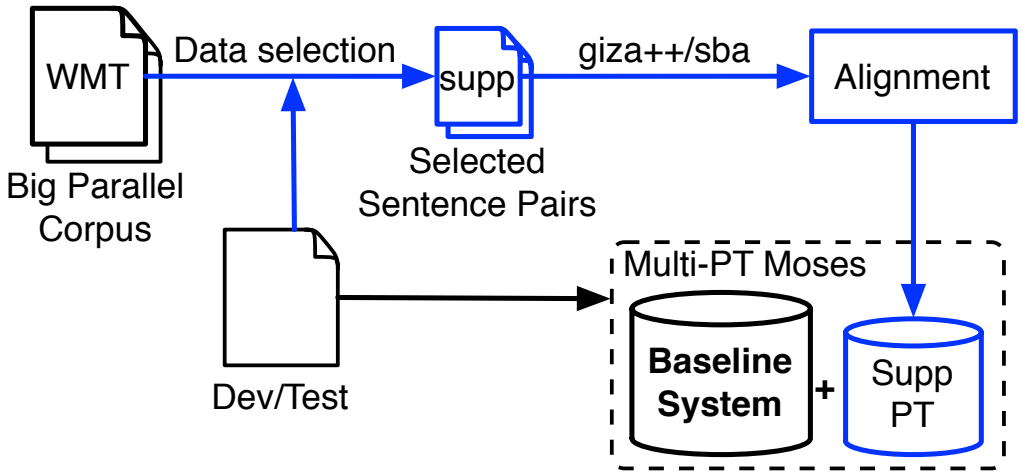
Data selection : select sentence pairs which contain at least one occurrence of a word in the input text and is out-of-vocabulary (OOV) in the baseline system.

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

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

Incremental alignment task : results

Phrase tables				HIT		
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

`concat`: `giza++` alignment learnt on the concatenation of HIT and supp

`sba`: our sampling-based alignment method

`concat++`: `giza++` alignment learnt on the `WMT` corpus

Incremental alignment task : results

Phrase tables				HIT		
main (62K HIT)	**supplementary** (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

`concat`: `giza++` alignment learnt on the concatenation of HIT and supp

`sba`: our sampling-based alignment method

`concat++`: `giza++` alignment learnt on the `WMT` corpus

Incremental alignment task : results

Phrase tables				HIT		
main (62K HIT)	supplementary (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

`concat`: `giza++` alignment learnt on the concatenation of HIT and supp

`sba`: our sampling-based alignment method

`concat++`: `giza++` alignment learnt on the `WMT` corpus

Incremental alignment task : results

Phrase tables				HIT		
main (62K HIT)	supplementary (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

`concat`: `giza++` alignment learnt on the concatenation of HIT and supp

`sba`: our sampling-based alignment method

`concat++`: `giza++` alignment learnt on the `WMT` corpus

Incremental alignment task : results

Phrase tables				HIT		
main (62K HIT)	supplementary (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

`concat`: `giza++` alignment learnt on the concatenation of HIT and supp

`sba`: our sampling-based alignment method

`concat++`: `giza++` alignment learnt on the `WMT` corpus

Incremental alignment task : results

Phrase tables				HIT		
main (62K HIT)	supplementary (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

`concat`: `giza++` alignment learnt on the concatenation of HIT and supp

`sba`: our sampling-based alignment method

`concat++`: `giza++` alignment learnt on the `WMT` corpus

Incremental alignment task : results

Phrase tables				HIT		
main (62K HIT)	supplementary (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

`concat`: `giza++` alignment learnt on the concatenation of HIT and supp

`sba`: our sampling-based alignment method

`concat++`: `giza++` alignment learnt on the `WMT` corpus

Outline

Conclusion

Conclusion

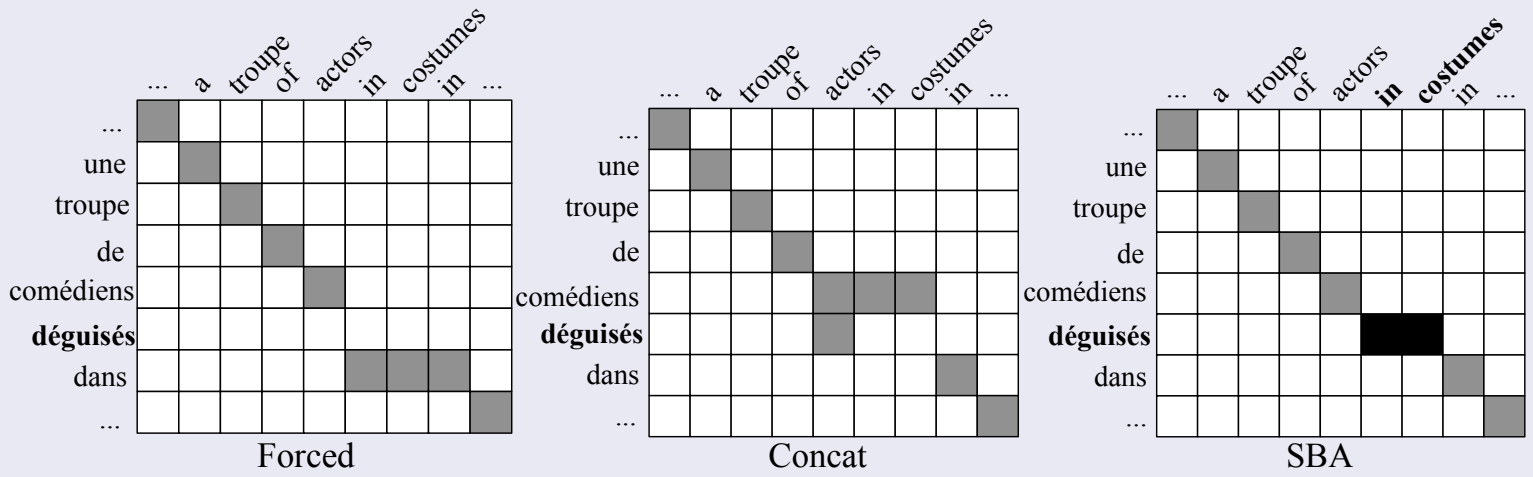
- alignment time can be controlled
- only useful sentence pairs need to be aligned
- integrating new data is plug-and-play

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)*

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 :

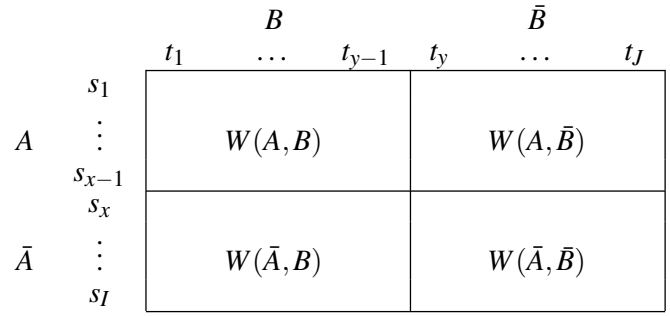
$$w(s,t) = p(s|t) * p(t|s)$$

2. Segmentation criterion :

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

where :

- $(X,Y) \in \{A,\bar{A}\} \times \{B,\bar{B}\}$
- $W(X,Y) = \sum_{s \in X, t \in Y} w(s,t)$



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

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

Sub-sentential alignment

1. 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.

procedure align(S,T) :
  if length(S) = 1 **or** length(T) = 1 :
    link each word of S to each word of T
    stop procedure
  $minNcut = 2$
  $(X,Y) = (S,T)$
  for each $(i,j) \in \{2 \ldots I\} \times \{2 \ldots J\}$:
    if Ncut$(A,B) < minNcut$:
      $minNcut$ = Ncut(A,B)
      $(X,Y) = (A,B)$
    if Ncut$(A,\bar{B} < minNcut$:
      $minNcut$ = Ncut$(A,\bar{B})$
      $(X,Y) = (A,\bar{B})$
  align(X,Y)
  align$(\bar{X},\bar{Y})$

Sub-sentential alignment

1. 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.

procedure align(S,T) :
  if length(S) = 1 **or** length(T) = 1 :
    link each word of S to each word of T
    stop procedure
  $minNcut = 2$
  $(X,Y) = (S,T)$
  for each $(i,j) \in \{2 \ldots I\} \times \{2 \ldots J\}$:
    if Ncut$(A,B) < minNcut$:
      $minNcut$ = Ncut(A,B)
      $(X,Y) = (A,B)$
    if Ncut$(A,\bar{B} < minNcut$:
      $minNcut$ = Ncut$(A,\bar{B})$
      $(X,Y) = (A,\bar{B})$
  align(X,Y)
  align$(\bar{X},\bar{Y})$

Sub-sentential alignment

1. 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.

procedure align(S,T) :
  if length$(S) = 1$ **or** length$(T) = 1$:
    link each word of S to each word of T
    stop procedure
  $minNcut = 2$
  $(X,Y) = (S,T)$
  for each $(i,j) \in \{2 \ldots I\} \times \{2 \ldots J\}$:
    if Ncut$(A,B) < minNcut$:
      $minNcut =$ Ncut(A,B)
      $(X,Y) = (A,B)$
    if Ncut$(A,\bar{B} < minNcut$:
      $minNcut =$ Ncut$(A,\bar{B})$
      $(X,Y) = (A,\bar{B})$
  align(X,Y)
  align$(\bar{X},\bar{Y})$

All results

	BTEC				HIT				BTEC+HIT			
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
English→French (1 reference)												
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
French→English (7 references)												
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
Chinese→English (7 references)												
giza++	-	-	-	-	27.88	51.69	50.76	1,139K	-	-	-	-
sba	-	-	-	-	27.85	53.05	50.93	655K	-	-	-	-

Incremental alignment task

Phrase tables				HIT					
main (62K HIT)	**supplementary** (3.3K supp)	# words	# entries	BLEU	1g	2g	3g	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

`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`

Incremental alignment task

Phrase tables				HIT					
main	supplementary								
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	3g	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

`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`

Incremental alignment task

Phrase tables				HIT					
main	**supplementary**								
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	3g	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

`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`

Incremental alignment task

Phrase tables				HIT					
main (62K HIT)	**supplementary** (3.3K supp)	# words	# entries	BLEU	1g	2g	3g	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

`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`

Incremental alignment task

Phrase tables				HIT					
main	supplementary								
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	3g	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

`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`

Incremental alignment task

Phrase tables				HIT					
main	**supplementary**								
(62K HIT)	(3.3K supp)	# words	# entries	BLEU	1g	2g	3g	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

`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`

Hypothesis

SBA performs better on rare words.