Lexicons or phrase tables? An investigation in sampling-based multilingual alignment

Adrien Lardilleux*  Jonathan Chevelu†*  Yves Lepage*
Ghislain Putois†  Julien Gosme*

*GREYC, Université de Caen Basse-Normandie
†Orange Labs, Lannion
France

12th November 2009
Outline

What is sampling-based multilingual alignment?

Some typical results on two typical tasks

What is in the phrase tables?

What is missing?
What is sampling-based multilingual alignment?

A sub-sentential alignment method

anymalign.py

Freely available, open source, easy to use, portable, pythonic

Extract lexical equivalences from sentence-aligned parallel corpora:

Multiword: extract translations of (dis)contiguous sequences of words
Multilingual: can process any number of languages at a time
“Anytime”: quality is not a matter of time.
Coverage is a matter of time.

Simple: very simple
An example-based sub-sentential alignment method

Alignments detection:
- based on strict distribution similarities of words on a multilingual parallel corpus

Alignments extraction:
- based on string differences

Alignments scoring:
- straightforward statistics
What is sampling-based multilingual alignment?

An example (1/3: strict distribution similarities)

Input: a subcorpus obtained by sampling the initial training corpus

1. One coffee, please. Un café, s’il vous plaît.
2. This coffee is excellent. Ce café n’est pas mauvais.
3. One strong tea. Un thé fort.
An example (1/3: strict distribution similarities)

Input: a subcorpus obtained by sampling the initial training corpus

1. One₁ coffee₁,₁ please₁,₁ Un₂ café₂,₂ s’il₂ vous₂ plait₂,₂
2. This₁ coffee₁ is₁ excellent₁,₁ Ce₂ café₂ n’est₂ pas₂ mauvais₂,₂
3. One₁ strong₁ tea₁,₁ Un₂ thé₂ fort₂,₂

⇓

|₁₁₁₁₁₀₁₀₁₁₁₀₀₀₁₁₁₁₁₁₀₀₁₀₀₀₁ |
|₀₁₀₁₁₁₁₁₁₀₁₁₁₀₀₀₀₀₀₀₀₀₀₀ |
|₀₁₀₁₀₀₁₀₀₁₀₁₁₀₀₀₀₀₁₁₁₀ |
An example (1/3: strict distribution similarities)

Input: a subcorpus obtained by sampling the initial training corpus

|   | One₁ coffee₁ ,₁ please₁ .₁ Un₂ café₂ ,₂ sʼil₂ vous₂ plaît₂ .₂ |
|   | This₁ coffee₁ is₁ excellent₁ .₁ Ce₂ café₂ nʼest₂ pas₂ mauvais₂ .₂ |
|   | One₁ strong₁ tea₁ .₁ Un₂ thé₂ fort₂ .₂ |

|₁₁₁₁₁₀₁₀₁₁₁₀₀₀₀₀₁₁₁₀₀₀₀₁ |
|₀₀₁₁₁₀₁₀₁₁₁₀₁₁₁₀₀₀₀₀₀₀₀₀ |
|₀₀₁₁₀₁₀₀₁₀₀₀₀₀₁₀₀₀₀₁₁₁₀ |

|₁₁₁₁₁₁₁₁₁₁₁₁₁₁₁¹² |
|₁₁₁₀₀₀₀₁₁₁₁₀₀₀₀₀₁¹² |
|₀₀₀₁₁₁₁₀₀₀₀₀₁₁₁¹² |

Adrien Lardilleux (GREYC) An investigation in sampling-based multilingual alignment
What is sampling-based multilingual alignment?

An example (2/3: string differences)

<table>
<thead>
<tr>
<th>The words:</th>
<th>appear on lines:</th>
<th>from which we extract:</th>
</tr>
</thead>
<tbody>
<tr>
<td>coffee₁ cafe₂</td>
<td>1</td>
<td>coffee₁ cafe₂</td>
</tr>
<tr>
<td></td>
<td></td>
<td>One₁,₁ please₁,₁ Un₂,₂ s’il₂ vous₂ plait₂,₂</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>coffee₁ cafe₂</td>
</tr>
<tr>
<td></td>
<td></td>
<td>This₁,₁ is₁ excellent₁,₁ Ce₂,₂ n’est₂ pas₂ mauvais₂,₂</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>
An example (2/3: string differences)

<table>
<thead>
<tr>
<th>The words:</th>
<th>appear on lines:</th>
<th>from which we extract:</th>
</tr>
</thead>
<tbody>
<tr>
<td>coffee₁ café₂</td>
<td>One₁ ,₁ please₁ .₁ Un₂ ,₂ s’il₂ vous₂ plaît₂ .₂</td>
<td></td>
</tr>
<tr>
<td>this₁ is₁ excellent₁ .₁ Ce₂ n’est₂ pas₂ mauvais₂ .₂</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>coffee</td>
<td>↔</td>
<td>café</td>
</tr>
<tr>
<td>One , , please .</td>
<td>↔</td>
<td>Un , , s’il vous plaît .</td>
</tr>
<tr>
<td>This , is excellent .</td>
<td>↔</td>
<td>Ce , n’est pas mauvais .</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>
What is sampling-based multilingual alignment?

An example (3/3: score alignments)

- The same process is repeated for numerous random subcorpora.
- All alignments from all subcorpora are collected.
- Translation probabilities are computed based on alignments’ counts.

Result:
A full-fledged translation table.
An example (3/3: score alignments)

- The same process is repeated for numerous random subcorpora.
- All alignments from all subcorpora are collected.
- Translation probabilities are computed based on alignments’ counts.

Result:
A full-fledged translation table.

... or not?
What is sampling-based multilingual alignment?

Some typical results on two typical tasks

What is in the phrase tables?

What is missing?
Two typical tasks

1. A machine translation task
2. A bilingual lexicon induction task

We compare the outputs of two word aligners:

1. Anymalign
2. MGIZA++, augmented by Moses for symmetric alignment and phrase extraction and scoring

We use two bilingual parallel corpora of different natures:

1. 40,000 pairs of Japanese-English sentences from the BTEC (average sentence length: 10 words)
2. 200,000 pairs of French-English sentences from the Europarl corpus (average sentence length: 31 words)
Some typical results on two typical tasks

Evaluation 1: a machine translation task

Using the Moses phrase-based SMT decoder

**BTEC: short Japanese-English sentences**

<table>
<thead>
<tr>
<th>Phrase table origin</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anymalign</td>
<td>0.39</td>
<td>0.45</td>
</tr>
<tr>
<td>MGIZA++/Moses</td>
<td>0.38</td>
<td>0.45</td>
</tr>
</tbody>
</table>

**Europarl: long French-English sentences**

<table>
<thead>
<tr>
<th>Phrase table origin</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anymalign</td>
<td>0.25</td>
<td>0.60</td>
</tr>
<tr>
<td>MGIZA++/Moses</td>
<td>0.29</td>
<td>0.56</td>
</tr>
</tbody>
</table>
Evaluation 2: a bilingual lexicon induction task (1/3)

- We compare the phrase tables to a reference bilingual lexicon.
- The reference bilingual lexicon is filtered so that it contains only translation pairs that can actually be obtained from the training parallel corpus.
- We compute precision, recall, and f-measure. Translation pairs from the phrase tables are weighted according to their source-to-target translation probabilities.
Some typical results on two typical tasks

Evaluation 2: a bilingual lexicon induction task (2/3)

BTEC: ja-en

Anymalign: precision
Anymalign: f-measure
Anymalign: recall
MGIZA++/Moses: precision
MGIZA++/Moses: f-measure
MGIZA++/Moses: recall

Adrien Lardilleux (GREYC) An investigation in sampling-based multilingual alignment EBMT3 12 / 24
Some typical results on two typical tasks

Evaluation 2: a bilingual lexicon induction task (3/3)

Europarl: fr-en

Anymalign: precision
Anymalign: f-measure
Anymalign: recall
MGIZA++/Moses: precision
MGIZA++/Moses: f-measure
MGIZA++/Moses: recall
Some typical results on two typical tasks

Conclusion of the two experiments

Anymalign typically yields equal or worse results on phrase-based machine translation tasks

+ Anymalign typically yields equal or better results on bilingual lexicon induction tasks, involving mainly unigrams
Conclusion of the two experiments

Anymalign typically yields equal or worse results on phrase-based machine translation tasks

+ 

Anymalign typically yields equal or better results on bilingual lexicon induction tasks, involving mainly unigrams

= 

Aren’t we just aligning unigrams, and missing longer n-grams?
What is sampling-based multilingual alignment?

Some typical results on two typical tasks

What is in the phrase tables?

What is missing?
What is in the phrase tables?

Investigating the contents of alignments: settings

We now resort to 1,000,000 pairs of French-English sentences from the Europarl corpus.

- We obtained the worst results on this corpus in the previous experiments.
- A large training corpus will highlight differences between the two phrase tables.
Investigating phrase table coverage

![Bar chart showing Europarl coverage (%) vs N-gram length for MGIZA++/Moses and Anymalign.

- **X-axis**: N-gram length (1 to 7)
- **Y-axis**: Europarl coverage (%)

Legend:
- Blue line: MGIZA++/Moses
- Red line: Anymalign

The chart illustrates the coverage of Europarl phrases for different N-gram lengths, comparing MGIZA++/Moses and Anymalign.]
What is in the phrase tables?

Less data is worse data

Anymalign’s phrase table is 42 times smaller than MGIZA++/Moses’!

- Anymalign is much better at unigram extraction.
- Anymalign is much much much much worse at n-gram extraction ($n \geq 2$).

⇒ Quantity, not quality!
What is in the phrase tables?

Failing at aligning n-grams?

Manual inspection of the content of phrase tables suggests that Anymalign would not align sequences of words with different frequencies.

⇒ We plot the distribution of bigrams according to the frequency of the words they are made of.
What is in the phrase tables?

Investigating bigrams distribution

Europarl

MGIZA++/Moses

Anymalign
Investigating bigrams distribution

Europarl

MGIZA++/Moses

Anymalign
But why?

Basics of the method:
Extract sequences of words that share exactly the same distribution in a subcorpus.

Words with very different frequencies never share the same distribution, whatever the subcorpus!
What is in the phrase tables?

But why?

Basics of the method:
Extract sequences of words that share exactly the same distribution in a subcorpus.

Words with very different frequencies never share the same distribution, whatever the subcorpus!

From this corpus:

1  a b ?  α β ;
2  a c ?  α γ ;
3  d ?  δ ;
What is in the phrase tables?

But why?

Basics of the method:
Extract sequences of words that share exactly the same distribution in a subcorpus.

Words with very different frequencies never share the same distribution, whatever the subcorpus!

From this corpus:

<table>
<thead>
<tr>
<th></th>
<th>a b ?</th>
<th>α β ;</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>a c ?</td>
<td>α γ ;</td>
</tr>
<tr>
<td>3</td>
<td>d ?</td>
<td>δ ;</td>
</tr>
</tbody>
</table>

we can extract:

b ↔ β

? ↔ ;

;
What is in the phrase tables?

But why?

Basics of the method:
Extract sequences of words that share exactly the same distribution in a subcorpus.

Words with very different frequencies never share the same distribution, whatever the subcorpus!

From this corpus:

<table>
<thead>
<tr>
<th></th>
<th>a b ?</th>
<th>α β ;</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>a c ?</td>
<td>α γ ;</td>
</tr>
<tr>
<td>3</td>
<td>d ?</td>
<td>δ ;</td>
</tr>
</tbody>
</table>

we can extract:

\[ b \leftrightarrow \beta \]
\[ ? \leftrightarrow ; \]

but we cannot extract:

\[ b \ ? \leftrightarrow \beta ; \]
What is missing?

What is sampling-based multilingual alignment?

Some typical results on two typical tasks

What is in the phrase tables?

What is missing?
What is missing?

What remains to be done

Recombine alignments together in order to produce longer alignments:

Alignments known

\[
b \leftrightarrow \beta
\]
\[
? \leftrightarrow ;
\]
\[
a \ b \ ? \leftrightarrow \alpha \ \beta \ ;
\]

⇒

New alignment

\[
b \ ? \leftrightarrow \beta \ ;
\]

≃ extract phrase alignments consistent with word alignments
≃ phrase extraction for phrase-based SMT
Conclusion

- An example-based sub-sentential alignment method
- Better results on lexicon induction tasks than on MT tasks
  \[ \rightarrow \text{better at unigram extraction} \]
- Does not align together words with different frequencies
- We would just need to recombine word alignments together
  in order to produce longer alignments

anymalign.py

http://users.info.unicaen.fr/~alardill/anymalign/
Conclusion

- An example-based sub-sentential alignment method
- Better results on lexicon induction tasks than on MT tasks
  ⇒ better at unigram extraction
- Does not align together words with different frequencies
- We would just need to recombine word alignments together in order to produce longer alignments

anymalign.py

http://users.info.unicaen.fr/~alardill/anymalign/

Thank you!