Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner

Mary Hearne, John Tinsley, Ventsislav Zhechev & Andy Way

National Centre for Language Technology
School of Computing
Dublin City University
Parallel treebanks

A parallel treebank comprises:

- sentence pairs
- parsed
- word-aligned
- tree-aligned

(Volk & Samuelsson, 2004)

The role of alignments:

Santos (1996), paraphrasing Lab (1990):

*Having a linguistic description of two languages is not the same as having a linguistic description of the translation between them.*
Parallel treebanks

- Our work involves automatically obtaining a parallel treebank from a parallel corpus via parsing and tree alignment.
- Our overall objective is to use the parallel treebank for inducing a variety of syntax-aware and syntax-driven models of translation for use in data-driven MT.
- In this paper/presentation, the focus is on the capture of translational divergences through the application of a tree-aligner to gold-standard tree pairs.
Capturing translational divergences

We aim to:

▶ make explicit the syntactic divergences between source and target sentence pairs
▶ align to express as precisely as possible the translational equivalences between the tree pair
▶ constraining phrase-alignments in the data set is a consequence of aligning trees, but not an objective

We remain agnostic with regard to:

▶ which linguistic formalism is most appropriate for the expression of monolingual syntax
▶ how best to exploit parallel treebanks for syntax-aware data-driven MT
Outline

Tree Alignments

Translational Divergences

Automatic Tree-to-Tree Alignment

Evaluation

Conclusions & Future Work
Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner

Mary Hearne, John Tinsley, Ventsislav Zhechev & Andy Way
Tree-to-Tree Alignment

Links indicate *translational equivalence*:

- a link between root nodes indicates equivalence between the sentence pair
- a link between any given pair of source and target nodes indicates
  - equivalence between the substrings they dominate
  - equivalence between the substrings they do *not* dominate

```
John sees Mary
NP  VP  NP  VP
---  ---  ---  ---
sees Mary voit Mary
```

\[\text{◮ a link between root nodes indicates equivalence between the sentence pair}\]
\[\text{◮ a link between any given pair of source and target nodes indicates}\]
\[\text{◮ equivalence between the substrings they dominate}\]
\[\text{◮ equivalence between the substrings they do *not* dominate}\]
Tree-to-Tree Alignment

In the simplest case:

- the sentence lengths are identical
- the word order is identical
- the tree structures are isomorphic
Tree-to-Tree Alignment

Slightly more complex:

- not every node in each tree needs to be linked
- each node is linked at most once
- terminal nodes are not linked
**Tree-Alignment vs. Word-Alignment**

**Word-alignment:** unaligned words are problematic and to be avoided

**Tree-alignment:** unaligned nodes are informative

... Jacob’s ladder ... → ... l’échelle de Jacob ...

**Word alignment:** Tree alignment:

![Tree alignment diagram]

![Word alignment diagram]
Hierarchical alignments

On the relationship between ’s and de in

... Jacob’s ladder ... → ... l’échelle de Jacob ...

’s → de

X ’s Y → Y de X

NP₁ ’s NP₂ → NP₂ de NP₁

NP → NP₁ ’s NP₂ : NP → NP₂ de NP₁
Outline

Tree Alignments

Translational Divergences

Automatic Tree-to-Tree Alignment

Evaluation

Conclusions & Future Work
Nominalisation

removing the print head retraite P de la tête d’impression
Lexical Divergences
Context-Dependent Lexical Selection

1. S
   - PRON
     - you
   - V
     - need
   - PART
     - to
   - VPv
   - S
     - PRON
       - vous
     - V
       - devez

2. S
   - PRON
     - you
   - V
     - need
   - PART
     - to
   - VPv
   - S
     - PRON
       - il
     - V
       - faut

3. CONJPsub
   - CONJsub
     - if
   - S
     - PRON
       - you
     - V
       - need
   - P
     - pour
   - NPdet

Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner

Mary Hearne, John Tinsley, Ventsislav Zhechev & Andy Way

Tree Alignments
Translational Divergences
Automatic Tree-to-Tree Alignment
Evaluation
Conclusions & Future Work
Embedded Complexities

Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner

Mary Hearne, John Tinsley, Ventsislav Zhechev & Andy Way

Tree Alignments
Translational Divergences
Automatic Tree-to-Tree Alignment
Evaluation
Conclusions & Future Work
Structural Dissimilarity

Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner

Mary Hearne, John Tinsley, Ventsislav Zhechev & Andy Way

Tree Alignments
Translational Divergences
Automatic Tree-to-Tree Alignment
Evaluation
Conclusions & Future Work

Sadj.

CONJPsub

CONJsub

S

if

NPadj

VPaux

D

NP

NPadj

V

NP

A

N

AUX

V

the

N

PP

is

N

CONJ

N

unauthorised repair is performed remainder

P

NP

null and void

of

NPzero

the

N

N

warranty period

NPdet

VPv

toute

N

APvp

invaliderait

D

N

intervention

Amod

V

la garantie

non

autorisée

‘any unauthorised action would invalidate the guarantee’
Outline

Tree Alignments

Translational Divergences

Automatic Tree-to-Tree Alignment

Evaluation

Conclusions & Future Work
Tree-alignment algorithm

Alignment algorithm:

- hypothesise initial alignments: each source node can link to any target node and vice versa;
- assign a score to each hypothesised alignment;
- select a set of links meeting the well-formedness criteria according to a greedy search.

Well-formedness criteria:

- each node can only be linked once;
- descendants of a source linked node may only link to descendents of its target linked counterpart;
- ancestors of a source linked node may only link to ancestors of its target linked counterpart.
Tree-alignment algorithm
Tree-alignment algorithm

\[ s_l = b \ c \quad \overline{s_l} = a \]
\[ t_l = x \ y \quad \overline{t_l} = w \ z \]

Computing hypothesis scores:
Assume tree pair \(<S,T>\), hypothesis \(<s,t>\), the following strings and GIZA++ / Moses word-alignment probabilities.

\[ s_l = s_i...s_{ix} \quad \overline{s_l} = S_1...s_{i-1}s_{ix+1}...S_m \]
\[ t_l = t_j...t_{jx} \quad \overline{t_l} = T_1...t_{j-1}t_{jx+1}...T_n \]

Hypothesis score: \( \gamma(<s,t>) = \alpha(s_l|t_l) \alpha(t_l|s_l) \alpha(\overline{s_l}|\overline{t_l}) \alpha(\overline{t_l}|\overline{s_l}) \)

String correspondence score: \( \alpha(x|y) = \prod_{j=1}^{\left| y \right|} \frac{\sum_{i=1}^{\left| y \right|} P(x_j|y_i)}{|y|} \)
Outline

Tree Alignments

Translational Divergences

Automatic Tree-to-Tree Alignment

Evaluation

Conclusions & Future Work
Methodology

- **dataset**: HomeCentre English-French corpus, parsed and aligned, 810 sentence pairs
- **Alignment evaluation**:
  - precision and recall of automatic alignments vs. manual alignments
- **Translation evaluation**:
  - split the data into training and test, 6 splits, averaged results
  - MT system used: DOT (Hearne & Way, EAMT-06)
  - train the system on manual vs. automatic alignments
- **Manual analysis of translational divergences**
Alignment Evaluation vs. Gold Standard

<table>
<thead>
<tr>
<th>Configs</th>
<th>all links</th>
<th></th>
<th>lexical links</th>
<th></th>
<th>non-lexical links</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>scr1</td>
<td>0.6162</td>
<td>0.7783</td>
<td>0.5057</td>
<td>0.7441</td>
<td><strong>0.8394</strong></td>
<td>0.7486</td>
</tr>
<tr>
<td>scr2</td>
<td>0.6215</td>
<td>0.7876</td>
<td>0.5131</td>
<td>0.7431</td>
<td>0.8107</td>
<td>0.7756</td>
</tr>
<tr>
<td>scr1_sp1</td>
<td><strong>0.6256</strong></td>
<td><strong>0.8100</strong></td>
<td><strong>0.5163</strong></td>
<td><strong>0.7626</strong></td>
<td>0.8139</td>
<td><strong>0.8002</strong></td>
</tr>
<tr>
<td>scr2_sp1</td>
<td>0.6245</td>
<td>0.7962</td>
<td><strong>0.5184</strong></td>
<td>0.7517</td>
<td>0.8031</td>
<td>0.7871</td>
</tr>
</tbody>
</table>
Translation Evaluation vs. Gold Standard

<table>
<thead>
<tr>
<th>Configs</th>
<th>Bleu</th>
<th>NIST</th>
<th>Meteor</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>manual</td>
<td>0.5222</td>
<td>6.8931</td>
<td>71.8531</td>
<td>68.5417</td>
</tr>
<tr>
<td>scr1</td>
<td>0.5091</td>
<td>6.9145</td>
<td>71.7764</td>
<td>71.8750</td>
</tr>
<tr>
<td>scr2</td>
<td>0.5333</td>
<td>6.8855</td>
<td>72.9614</td>
<td>72.5000</td>
</tr>
<tr>
<td>scr1_sp1</td>
<td>0.5273</td>
<td>6.9384</td>
<td>72.7157</td>
<td>72.5000</td>
</tr>
<tr>
<td>scr2_sp1</td>
<td>0.5290</td>
<td>6.8762</td>
<td>72.8765</td>
<td>72.5000</td>
</tr>
</tbody>
</table>
Capturing Translational Divergence

Simple, isomorphic alignments:

```
  NP  --  NP
   D   N  D   N
 the  scanner  le  scanner
```

```
  PP  --  PP
  P   NP  P   NP
  to  D   N  à  D   N
 the  Home Centre  le  Home Centre
```
Capturing Translational Divergence

Nominalisation

- VP
  - V
  - NP
- NP
  - N
  - PP
  - retraite P NP
  - de

removing
Capturing Translational Divergence

Lexical Selection

S

PRON
you

V
need
PART
to

VPv

S

PRON
vous

V
devez

VPinf

S

PRON
you

V
need
PART
to

VPv

S

PRON
il

V
faut

VPverb

VPinf

VPverb
Outline

Tree Alignments

Translational Divergences

Automatic Tree-to-Tree Alignment

Evaluation

Conclusions & Future Work
Conclusions

- aligner performance is better at the phrase level than the lexical level
- imbalance between precision and recall at the lexical level
  - aligner uses GIZA++ word-alignment probabilities
  - GIZA++ prioritises broad coverage over high precision
  - in terms of capturing translational divergences between tree pairs, the preference is for the opposite
- it is appropriate for tree-alignment to prioritise precision over recall
- MT systems should use high-precision tree alignments in conjunction with broad-coverage models to preserve robustness
Future Work

- investigate alternative word-alignment methods to further improve the accuracy of the tree-alignment algorithm
- investigate the impact of imperfect parse quality on tree-alignment
- investigate the extraction of translation models from automatically-annotated parallel treebanks
The End.