Parallel Treebanks in Machine Translation

John Tinsley

National Centre for Language Technology
Dublin City University
Ireland

Collaborators: Andy Way, Vensti Zhechev, Mary Hearne

CMU-LTI MT Lunch
Jan 13th 2009
Parallel Treebanks

Building Parallel Treebanks

Parallel Treebanks in PB-SMT

Parallel Treebanks in Syntax-Aware MT
Motivation and Hypothesis

- Incorporation of linguistics may help state-of-the-art MT
- Parallel treebanks are a linguistically rich resource
- Could MT be bettered by exploiting parallel treebanks as training data?
What is a parallel treebank?
What is a parallel treebank?

- Syntactically annotated (parsed) aligned sentences - also subsententially aligned
What is a parallel treebank?

- Syntactically annotated (parsed) aligned sentences - also subsententially aligned

```
S
  NP1  VP
  Anthony  V  NP2
  likes  Cleopatra
```
What is a parallel treebank?

- Syntactically annotated (parsed) aligned sentences - also subsententially aligned
- Links hold implication of translational equivalence
What is a parallel treebank?

- Syntactically annotated (parsed) aligned sentences - also subsententially aligned
- Links hold implication of translational equivalence

Recent concept - expensive to build, thus they are thin on the ground e.g. Cz-En, Sv-De-En
Parallel Treebanks

Building Parallel Treebanks

Parallel Treebanks in PB-SMT

Parallel Treebanks in Syntax-Aware MT
What do we need to build treebanks?

► Parallel corpus
► Monolingual parsers
► Sub-tree aligner
What do we need to build treebanks?

- Parallel corpus
- Monolingual parsers
- Sub-tree aligner

What is Sub-Tree Alignment
What do we need to build treebanks?

- Parallel corpus
- Monolingual parsers
- Sub-tree aligner

What is Sub-Tree Alignment

- Insertion of links between sub-trees
- Issues with previous approaches led us to develop our own aligner
What do we need to build treebanks?

- Parallel corpus
- Monolingual parsers
- Sub-tree aligner

What is Sub-Tree Alignment

- Insertion of links between sub-trees
- Issues with previous approaches led us to develop our own aligner

Nodes are aligned between trees using the following algorithm
What do we need to build treebanks?

- Parallel corpus
- Monolingual parsers
- Sub-tree aligner

What is Sub-Tree Alignment

- Insertion of links between sub-trees
- Issues with previous approaches led us to develop our own aligner

Nodes are aligned between trees using the following algorithm

- assume hypothetical links between all nodes in a given tree pair
What do we need to build treebanks?

- Parallel corpus
- Monolingual parsers
- Sub-tree aligner

What is Sub-Tree Alignment

- Insertion of links between sub-trees
- Issues with previous approaches led us to develop our own aligner

Nodes are aligned between trees using the following algorithm

- assume hypothetical links between all nodes in a given tree pair
- estimate scores for these hypothetical links
What do we need to build treebanks?

- Parallel corpus
- Monolingual parsers
- Sub-tree aligner

What is Sub-Tree Alignment

- Insertion of links between sub-trees
- Issues with previous approaches led us to develop our own aligner

Nodes are aligned between trees using the following algorithm

- Assume hypothetical links between all nodes in a given tree pair
- Estimate scores for these hypothetical links
- Use a greedy search to select the optimal set of alignments
Given tree pair $<S,T>$ and hypothesis $<s,t>$, we compute the following strings:

$$s_l = s_i...s_{i+1}$$
$$\overline{s_l} = S_1...s_{i-1}s_{i+1}...S_m$$
$$t_l = t_j...t_{j+1}$$
$$\overline{t_l} = T_1...t_{j-1}t_{j+1}...T_n$$

```
  a  -  w  -  z
  |  |  |  |  |
  b  c  x  y
```

$s_l = b\ c$
$t_l = x\ y$
$\overline{s_l} = a$
$\overline{t_l} = w\ z$
Hypothesis Scoring

The score for the given hypothesis \(<s,t>\) is computed according to:

\[
\gamma(\langle s, t \rangle) = \alpha(s_l|t_l) \alpha(t_l|s_l) \alpha(s_l|\overline{t_l}) \alpha(\overline{t_l}|s_l)
\] (1)

where:

\[
\alpha(x|y) = \prod_{i=1}^{y} \sum_{j=1}^{x} P(x_j|y_i)
\] (2)
Greedy Search Example

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Greedy Search Example

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Greedy Search Example

![Diagram of a syntactic tree with parallel elements and a table showing marking pairs and cost matrix.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Greedy Search Example

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
```

John Tinsley - CMU-LTI MT Lunch 13/01/09
Greedy Search Example

Parallel Treebanks
Building Parallel Treebanks
Parallel Treebanks in PB-SMT
Parallel Treebanks in Syntax-Aware MT

John Tinsley - CMU-LTI MT Lunch 13/01/09
Greedy Search Example

Parallel Treebanks
Building Parallel Treebanks
Parallel Treebanks in PB-SMT
Parallel Treebanks in Syntax-Aware MT

Greedy Search Example

Parallel Treebanks in Machine Translation

John Tinsley - CMU-LTI MT Lunch 13/01/09
Greedy Search Example

```
HEADER-1 PP-1
PP-2 COLON-9 P-2 NP-7
P-3 NP-4 : P-3 D-5 P-6 D-8 NP-9
from D-5 NP-6 ` a partir de une N-10 N-11
N-7 N-8 application Windows
```

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
```
Greedy Search Example

Parallel Treebanks
Building Parallel Treebanks
Parallel Treebanks in PB-SMT
Parallel Treebanks in Syntax-Aware MT

HEADER-1 PP-1
PP-2 COLON-9 P-2 NP-7
P-3 NP-4 : P-3 D-5 P-6 D-8 NP-9
from D-5 NP-6 ` a partir de une N-10 N-11
NP-6 N-7 N-8
Windows Application

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

John Tinsley - CMU-LTI MT Lunch 13/01/09
Greedy Search Example

```
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

John Tinsley - CMU-LTI MT Lunch 13/01/09
Evaluation

- Xerox HomeCentre Corpus: 810 English–French treepairs
- Manually parsed and aligned gold standard
- 2 DOT systems trained on automatically built treebank and manually built treebank
Evaluation

- Xerox HomeCentre Corpus: 810 English–French treepairs
- Manually parsed and aligned gold standard
- 2 DOT systems trained on automatically built treebank and manually built treebank

<table>
<thead>
<tr>
<th>Training Treebank</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>0.7185</td>
<td>68.5417%</td>
</tr>
<tr>
<td>automatic</td>
<td>0.5334</td>
<td>6.9210</td>
<td>0.7297</td>
<td>71.8750%</td>
</tr>
</tbody>
</table>

parallel treebanks in machine translation
Evaluation

- Xerox HomeCentre Corpus: 810 English–French treepairs
- Manually parsed and aligned gold standard
- 2 DOT systems trained on automatically built treebank and manually built treebank

<table>
<thead>
<tr>
<th>Training Treebank</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>0.7185</td>
<td>68.5417%</td>
</tr>
<tr>
<td>automatic</td>
<td>0.5334</td>
<td>6.9210</td>
<td>0.7297</td>
<td>71.8750%</td>
</tr>
</tbody>
</table>

Observations

- Scores competitive with manually built treebank
- DOT coverage increases significantly
- We have developed a viable alternative
- Described fully in Tinsley et al. (2007a)
- Further intrinsic and manual evaluation - described in Hearne et al. (2007)
Parallel Treebanks

Building Parallel Treebanks

Parallel Treebanks in PB-SMT

Parallel Treebanks in Syntax-Aware MT
Question

- Will machine translation quality improve if we incorporate phrase pairs extracted from a parallel treebank into the translation model of a phrase-based SMT system?

Why would it?

- More reliable phrases?
- Increased coverage
- Boosted probabilities
Figure 2: Example of phrase extraction for the given sentence pair depicting: (a) the aligned parallel tree pair; (b) the word alignment matrix (the rectangular areas represent extracted phrase pairs); (c) the combined set of extracted phrase pairs where: ◊ = only extracted from (a); † = only extracted from (b); * = extracted from both (a) and (b).
Experiments

Small scale

- English ↔ German 10,000 sentence pairs
- English ↔ Spanish 5,000 sentence pairs

<table>
<thead>
<tr>
<th>Config.</th>
<th>en-es</th>
<th>es-en</th>
<th>en-de</th>
<th>de-en</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>0.1765</td>
<td>0.1754</td>
<td>0.1186</td>
<td>0.1622</td>
</tr>
<tr>
<td>+Tree</td>
<td>0.1867</td>
<td>0.1880</td>
<td>0.1259</td>
<td>0.1687</td>
</tr>
</tbody>
</table>
Experiments

Small scale

- English ↔ German 10,000 sentence pairs
- English ↔ Spanish 5,000 sentence pairs

<table>
<thead>
<tr>
<th>Config.</th>
<th>en-es</th>
<th>es-en</th>
<th>en-de</th>
<th>de-en</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>0.1765</td>
<td>0.1754</td>
<td>0.1186</td>
<td>0.1622</td>
</tr>
<tr>
<td>+Tree</td>
<td>0.1867</td>
<td>0.1880</td>
<td>0.1259</td>
<td>0.1687</td>
</tr>
</tbody>
</table>

Observations

- Adding treebank data improved over baseline system
- Treebank data of higher quality - especially word alignments
- Not enough treebank data to suffice alone
- Described fully in Tinsley et al. (2007b)
Experiments

Large scale

- Full English $\rightarrow$ Spanish Europarl v2 ($\sim$730,000 sentence pairs)
- 1,000 devset, 2,000 testset

<table>
<thead>
<tr>
<th>Config.</th>
<th>Bleu</th>
<th>NIST</th>
<th>%METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>0.3341</td>
<td>7.0765</td>
<td>57.39</td>
</tr>
<tr>
<td>+Tree</td>
<td>0.3397</td>
<td>7.0891</td>
<td>57.82</td>
</tr>
<tr>
<td>Tree only</td>
<td>0.3153</td>
<td>6.8187</td>
<td>55.98</td>
</tr>
</tbody>
</table>
Experiments

Large scale

- Full English → Spanish Europarl v2 (≈730,000 sentence pairs)
- 1,000 devset, 2,000 testset

<table>
<thead>
<tr>
<th>Config.</th>
<th>Bleu</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>0.3341</td>
<td>7.0765</td>
<td>57.39</td>
</tr>
<tr>
<td>+Tree</td>
<td>0.3397</td>
<td>7.0891</td>
<td>57.82</td>
</tr>
<tr>
<td>Tree only</td>
<td>0.3153</td>
<td>6.8187</td>
<td>55.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource</th>
<th>#Tokens</th>
<th>#Types</th>
<th>intersection</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>72,940,465</td>
<td>24,708,527</td>
<td>1,447,505</td>
</tr>
<tr>
<td>Treebank</td>
<td>21,123,732</td>
<td>6,432,771</td>
<td>-</td>
</tr>
</tbody>
</table>
More experiments

**What?**: only use treebank *phrase* pairs  
**Why?**: large proportion of high-risk word alignments  
**Outcome**: Further slight improvement
More experiments

What?: only use treebank phrase pairs
Why?: large proportion of high-risk word alignments
Outcome: Further slight improvement

<table>
<thead>
<tr>
<th>Config.</th>
<th>BLEU</th>
<th>NIST</th>
<th>%METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.3341</td>
<td>7.0765</td>
<td>57.39</td>
</tr>
<tr>
<td>+Tree</td>
<td>0.3397</td>
<td>7.0891</td>
<td>57.82</td>
</tr>
<tr>
<td>Strict phrases</td>
<td>0.3414</td>
<td>7.1283</td>
<td>57.98</td>
</tr>
</tbody>
</table>
More experiments

- **What?**: Increase counts of treebank phrase pairs
- **Why?**: Give them more weight in the model
- **Outcome**: No improvement
More experiments

- **What?**: Increase counts of treebank phrase pairs
- **Why?**: Give them more weight in the model
- **Outcome**: No improvement

- **What?**: Use separate phrase tables
- **Why?**: Allow model to chose which to use
- **Outcome**: No improvement
More experiments

- **What?**: Increase counts of treebank phrase pairs
- **Why?**: Give them more weight in the model
- **Outcome**: No improvement

- **What?**: Use separate phrase tables
- **Why?**: Allow model to chose which to use
- **Outcome**: No improvement

- **What?**: Use treebank alignments to seed SMT phrase extraction
- **Why?**: create more robust word alignment
- **Outcome**: No translation improvement but large reduction in model size when using union of SMT and treebank alignments

All experiments detailed in **Tinsley et al. (2009)**
Parallel Treebanks

Building Parallel Treebanks

Parallel Treebanks in PB-SMT

Parallel Treebanks in Syntax-Aware MT
So why am I here?

- Stat-XFER essentially uses parallel treebanks as training data
- Similar experiments involving combinations of data
- Compare our respective methodologies within this framework
References


Other Relevant Literature


Thank you
http://computing.dcu.ie/~jtinsley
http://nclt.dcu.ie/mt