Manually Parsed, Manually Aligned Bilingual Data for MT

Kevin Knight, USC/ISI
Member of BAC
Parsed, Aligned, Bilingual Data

• Understanding weak points of automatic parsers and aligners

• Building better parsers and aligners

• Designing radically new MT models
Understanding Weak Points

GOLD parses  
GOLD alignments  
GOLD translation rules  

AUTOMATIC parses  
AUTOMATIC alignments  
AUTOMATIC translation rules  

Chinese/English

<table>
<thead>
<tr>
<th>Alignments</th>
<th>Parses</th>
<th>Rule Quality (f-score vs. GOLD rules)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>GOLD</td>
<td>100</td>
</tr>
<tr>
<td>GOLD</td>
<td>Automatic (Collins)</td>
<td>75</td>
</tr>
<tr>
<td>Automatic</td>
<td>GOLD</td>
<td>57</td>
</tr>
<tr>
<td>Automatic</td>
<td>Automatic</td>
<td>47</td>
</tr>
</tbody>
</table>

HOW BAD ARE THESE, AND WHY?
Building Better Aligners and Parsers

Problem: GIZA only aligns string pairs.
Idea: Use parse trees to improve alignments.

Impact: +0.7 Bleu on Chinese/English

[Fossum, Knight, Abney 08]
Designing radically new MT models

• There are lots of ways to envision syntactic/semantic translation patterns
  – Mathematical considerations
  – Linguistic considerations

• Which models best fit the data?

• Best way:
  – See how new models physically work on clean data
  – Clean = manually parsed, manually aligned
Look at many examples.

Ask: “What is going on here?”
but not the same world here too the world is

lkn hnA AyDA lm yEd AIEAlm *Ath
but the same world here too is not the same world.
not is the same world the too here but too the world .

Select region to translate.

Select rule:

\[ S(CC(but) \ x0:S\text{-BAR}) \rightarrow lkn \ , \ x0 \]

Apply rule.
Ikn, here too the world is not the same.
Lkn, 

here too the world is not the same.
The world is the same.
This rule takes care of SVO→VSO re-ordering.

Problems
- Period is part of rule: lots of SVO→VSO rules are needed
- Even if period were higher, there might be PPs
- Verb still unknown, may control re-ordering?
not is the same world
Problem
- Translation of “is” by this rule requires an RB to the right
- So, there are zillions of rules that translate “is” to the same Arabic word
not the same world
Ikn, hnA, AyDA, not the same world.
the same world
the same world
lkn, hnA, AyDA, lm, yEd, AlEAIm, *Ath.
but here too the world is not the same.
is yEd
is

STAG Explanation
not is
is the same
not is the same world
not the same world is the same world
not is the same world.
not is the same world.
not is the same world too here the world.

VP

VBZ not the same is

VP

RB DT JJ

VP

NP

NP

DT NN

S

S

ADVP

RB ADVP

here RB NP

too DT NN

S

.
ADVP
  RB here
  RB too
  DT the
  NN world

VP
  VBZ not
  DT the
  JJ same
  RB is
but the same world here too the world is not the same is
but the world too is the same and here too is the world.
Parsed, Aligned, Bilingual Data

• Understanding weak points of automatic parsers and aligners

• Building better parsers and aligners

• Designing radically new MT models