New Features for Statistical Machine Translation

David Chiang
Kevin Knight
Wei Wang
María no dio una bofetada a la bruja verde

This is a graph showing the syntactic structure of the sentence 'María no dio una bofetada a la bruja verde' in Spanish. The graph breaks down the sentence into its constituent parts, with each node representing a word and the edges showing the grammatical relationships between them.
Maria no dió una bofetada a la bruja verde
MOTIVATION

Maria no dió una bofetada a la bruja verde

...
Maria no dió una bofetada a la bruja verde

- Maria
- no
- dió
- una
- bofetada
- a
- la
- bruja
- verde
MOTIVATION

• Minimum error rate training (MERT) works for <30 features

• Margin infused relaxed algorithm (MIRA)
  • Online large-margin discriminative training
  • Scales better to large feature sets
  • Enables freer exploration of features
## RESULTS

**GALE 2008 Chinese-English data**

<table>
<thead>
<tr>
<th>System</th>
<th>Training</th>
<th>Features</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiero</td>
<td>MERT</td>
<td>11</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td>MIRA</td>
<td>10,990</td>
<td>37.6</td>
</tr>
<tr>
<td>Syntax</td>
<td>MERT</td>
<td>25</td>
<td>39.5</td>
</tr>
<tr>
<td></td>
<td>MIRA</td>
<td>283</td>
<td>40.6</td>
</tr>
</tbody>
</table>
OVERVIEW

• Training
• Features
• Experiments
Training
MIRA

- Crammer and Singer, 2003
- Applied to statistical MT by Watanabe et al., 2007
- Chiang, Marton, and Resnik, 2008:
  - use more of the forest
  - parallelize training
MERT

Model score

BLEU
MIRA

BLEU vs. Model score diagram with points and arrows indicating loss and margin.
FOREST-BASED TRAINING
PARALLEL TRAINING

• Run $n$ MIRA learners in parallel

• Share information among learners

<table>
<thead>
<tr>
<th>Hiero</th>
<th>$n = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>$n = 73$</td>
</tr>
</tbody>
</table>
Features
- Low counts are often overestimates

- Introduce a \textit{count=1} feature that fires on 1-count rules, etc.
TARGET SYNTAX FEATURES

UN inspectors VP

VP

expelled by NK

UN inspectors VP

VP

VBD

I

were

VP

expelled by NK

insert-were
TARGET SYNTAX FEATURES

S
  NP
  PP
  NP
  IN
  NP
  ""thinking of"
  NP
  the best-selling book
  VBN
  published
  NP
  ADVP
  NNP
  edo
  VP
  NP
  ""bad-rewrite
  NP
  in mind
  PP
  for the generation
  bad-rewrite
TARGET SYNTAX FEATURES

[Graphical representation of sentence structures with nodes labeled as S, NP, VP, and a highlighted node with value = ,]
TARGET SYNTAX FEATURES

第一个 站 出来

first stand come out

=IN

第一个 站 出来

first stand come out

=VP

the first to VP

stand up

IN

from

the first leg

PP

NP

IN

NP

IN

root=IN

root=VP
SOURCE CONTEXT FEATURES
Marton & Resnik 2008; Chiang et al 2008

- Use external parser to infer source-side syntax
- Rewards and penalties for matching/crossing brackets

新趋势在研究中的新动向.
new trends in the study cross-VP

这是一个值得的关注和研究的新动向.
this is a merit attention and study new trend
SOURCE CONTEXT FEATURES
Marton & Resnik 2008; Chiang et al 2008

VP

这是一个值得关注和研究的新动向。

meriting attention and study

match-VP

- Use external parser to infer source-side syntax
- Rewards and penalties for matching/crossing brackets
Norway restoring peace mediation in Sri Lanka.

Norway restore in Sri Lanka peace mediation.

to restore peace in Sri Lanka, the Norwegian mediation.
SOURCE CONTEXT FEATURES

• Word context features: similar to Watanabe et al. 2007 and work on WSD in MT (Chan et al. 2007, Carpuat & Wu 2007)

• Relate a word’s translation with its left or right neighbor on the source side (just the 100 most frequent types)

```
  f_{i-1}  f_i  e
  |       |
```

```
  f_i  f_{i+1}
  |       |
  e
```
he said because no voice he had to

since there is no voice, he said, he had to

he said that because of the lack of voice, he had to

he said, since there is no voice, he had to

he said that because of the lack of voice, he had to
Experiments
### TRAINING DATA

**GALE 2008 Chinese-English data**

<table>
<thead>
<tr>
<th></th>
<th>Hiero</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel data</td>
<td>260M</td>
<td>65M</td>
</tr>
<tr>
<td>Language model</td>
<td>2G</td>
<td>1G</td>
</tr>
<tr>
<td>MERT/MIRA</td>
<td>58k</td>
<td>58k</td>
</tr>
<tr>
<td>Test</td>
<td>57k</td>
<td>57k</td>
</tr>
</tbody>
</table>
## RESULTS (HIERO)

### Chinese-English

<table>
<thead>
<tr>
<th>Training</th>
<th>Features</th>
<th>#</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>baseline</td>
<td>11</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td>+source-side syntax +distortion</td>
<td>56</td>
<td>36.9</td>
</tr>
<tr>
<td>MIRA</td>
<td>+discount</td>
<td>61</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td>+word context</td>
<td>10,990</td>
<td>37.6</td>
</tr>
</tbody>
</table>
# RESULTS (SYNTAX)

## Chinese-English

<table>
<thead>
<tr>
<th>Training</th>
<th>Features</th>
<th>#</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>baseline</td>
<td>25</td>
<td>39.5</td>
</tr>
<tr>
<td></td>
<td>baseline</td>
<td>25</td>
<td>39.8</td>
</tr>
<tr>
<td></td>
<td>rule overlap</td>
<td>132</td>
<td>39.9</td>
</tr>
<tr>
<td>MIRA</td>
<td>node count</td>
<td>136</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td>+discount +bad rewrite +insertion</td>
<td>283</td>
<td>40.6</td>
</tr>
</tbody>
</table>
CONCLUSIONS

• Using underutilized information for new features:
  • Source context is computationally efficient
  • Target syntax provides a rich structure
• MIRA is working well on new features, systems, languages