FBK's Machine Translation Systems for IWSLT 2012's TED Lectures

Nick Ruiz, Arianna Bisazza
Roldano Cattoni, Marcello Federico
Outline

- Common components
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion
Fill-Up
(Bisazza et al., 2011; Nakov, 2008)

à la chirurgie esthétique
da chirurgie esthétique
de la chirurgie esthétique
chirurgie esthétique
la chirurgie
la chirurgie plastique
derchirurgie esthétique
devraient subir une intervention chirurgicale
son ablation
de subir une intervention chirurgicale,
de subir une intervention chirurgicale,

cosmetic surgery
to undergo surgery

Phrase Table
- Europarl
- Giga French
- TED
Cross-Entropy LM Filtering
(Moore & Lewis, 2010)

- Cross-Entropy ranking of sentences in a out-of-domain corpus against TED
- Incrementally add sentences to minimize perplexity on a development set
- Also applicable to parallel corpora by filtering on target language
Cross-Entropy LM Filtering
(Moore & Lewis, 2010)

Cross-Entropy Filtering on English Corpora

Filtering tuned on TED dev2010 data
Outline

- Common features
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion
Arabic-English

- Early Distortion Cost
- Hybrid Language Modeling
- Phrase/Reordering Fill-Up (TED+MultiUN)
- Mixture LM (TED, Gigaword, WMT News)
Early Distortion Cost
(Moore & Quirk, 2007)

- Improved distortion penalty
- Anticipates gradual accumulation of total distortion cost
  - Incorporates an estimate of future jump's cost
  - Same distortion penalty as standard distortion cost over a complete hypothesis
- Benefits: Improves comparability of translation hypotheses with the same number of covered words
Early Distortion Cost
(Moore & Quirk, 2007)

\[ W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7 \]

\[ +1 + 6 + 0 + 6 + 6 + 0 + 5 + 0 \]

\[ \text{Tot(std)} = 12 \]
\[ \text{Tot(edc)} = 12 \]
Early Distortion Cost
(Moore & Quirk, 2007)

<table>
<thead>
<tr>
<th>DL</th>
<th>DC</th>
<th>tst2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>std</td>
<td>26.12/6.514</td>
</tr>
<tr>
<td>8</td>
<td>std</td>
<td>25.95/6.460</td>
</tr>
<tr>
<td>8</td>
<td>edc</td>
<td>26.31/6.551</td>
</tr>
</tbody>
</table>
Hybrid Language Modeling
(Bisazza & Federico, 2011)

- Replace bottom 25% of tokens with POS tags – corresponds to 2% of types

In-domain target data

Now you laugh, but that quote has kind of a sting to it, right. And I think the reason it has…

Now you VB, but that NN has kind of a NN to it, right. And I think the reason it has…

...a sting is because thousands of years of history don't reverse themselves without a lot of pain.

...a NN is because NNS of years of history don't VB PP without a lot of NN.

Hybridly mapped word/POS data

- Allows for the construction of 10-gram LMs
## Arabic-English results

<table>
<thead>
<tr>
<th></th>
<th>LM</th>
<th>DL</th>
<th>tst2011</th>
<th>tst2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>MixAll.4g + TED.Hybrid10g edc</td>
<td>8</td>
<td>25.46/6.232</td>
<td>27.86/6.881</td>
</tr>
<tr>
<td>C₁</td>
<td>MixAll.4g edc</td>
<td>8</td>
<td>25.19/6.205</td>
<td>27.74/6.903</td>
</tr>
<tr>
<td>C₂</td>
<td>MixFiltered.5g + TED.Hybrid10g</td>
<td>8</td>
<td>25.13/6.190</td>
<td>27.54/6.828</td>
</tr>
</tbody>
</table>
Outline

- Common features
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion
Turkish-English

- Morphological Segmentation
- Hierarchical phrase-based decoding
- Mixture LM
Morphological Splitting

• Rule-based vs. Unsupervised segmentation

<table>
<thead>
<tr>
<th>Distortion Limit</th>
<th>Distortion Calc</th>
<th>Seg</th>
<th>tst2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>std</td>
<td>MS6</td>
<td>13.61/5.280</td>
</tr>
<tr>
<td>15</td>
<td>std</td>
<td>MS15</td>
<td>14.38/5.273</td>
</tr>
<tr>
<td>15</td>
<td>std</td>
<td>Morfessor</td>
<td>13.45/5.080</td>
</tr>
</tbody>
</table>

• MS6: Nominal suffixes (case + possessive) only
• MS15: Nominal and verbal suffixes
  – e.g. person-subject, negation, passive, etc.
• Morfessor:
  – Concatenates non-initial “morphs” into word endings
  – Could perhaps be trained with better configurations
Morphological Splitting

<table>
<thead>
<tr>
<th>Original</th>
<th>Analyzed</th>
<th>MS15</th>
<th>Morfessor</th>
<th>Trans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendisine Don diyelim .</td>
<td>kendī+Pron+Reflex+A3sg+P3sg+Dat</td>
<td>kendi+Pron+Reflex+A3sg</td>
<td>Kendi</td>
<td>Let's call him Don .</td>
</tr>
<tr>
<td></td>
<td>don+Noun+A3sg+Pnon+Nom</td>
<td>+Dat</td>
<td>+sine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>de+Verb+Pos+Opt+A1pl</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>de+Verb+Opt+A1pl</td>
<td>+A1pl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Phrase-Based Decoding

- Better able to handle mismatches in predicate-argument structure between languages
- Robust with respect to long-distance reordering

<table>
<thead>
<tr>
<th>Turkish (source)</th>
<th>English (target)</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>[X] söyle+Verb+Fut</td>
<td>will say [X]</td>
<td>SOV → SVO</td>
</tr>
<tr>
<td>[X] +Dat bak</td>
<td>look at [X]</td>
<td>S Comp V → S V Comp</td>
</tr>
<tr>
<td>[X] +Dat baktı</td>
<td>looked at [X]</td>
<td>S Comp V → S V Comp</td>
</tr>
</tbody>
</table>
## Turkish-English results

<table>
<thead>
<tr>
<th>System</th>
<th>Seg</th>
<th>tst2011</th>
<th>tst2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Hierarchical</td>
<td>MS15</td>
<td>17.24/5.560</td>
</tr>
<tr>
<td>C₁</td>
<td>Phrase-based (dl=15, edc)</td>
<td>MS15</td>
<td>15.45/5.289</td>
</tr>
</tbody>
</table>
Outline

- Common features
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion
Dutch-English

- Language properties
  - Similar to German
    - SVO for main clauses, SOV for subordinates
    - Noun casing, but less than German
  - Only “gendered” and “neutered” nouns/determiners
  - Compound nouns and verbs
Dutch-English

- Compound Splitting
- Phrase/Reordering Fill-Up (TED+Europarl)
- Mixture LM
Compound Splitting  
(Koehn & Knight, 2003)

- Preliminary experiments on German, carried over to Dutch
- Moses Compound Splitting tool
  - Split candidate words into tokens already existing in a corpus' vocabulary
  - Default (normal) setting: min 4 characters per split
  - Aggressive setting: reduce minimum to 2 chars
    - e.g. “aanvragen”, “afvallen”
He said he didn 't know . He would ask around .

Hij zei dat hij het niet wist . Hij zou rondvragen

(Normal/Aggressive splitting)

And he said that he did not know . He would ask around .

rond vragen
Compound Splitting

Not by the latest combine and tractor invention

niet door de laatste combine- en tractoruitvinding

(Normal splitting)

(Aggressive splitting)
## Dutch-English results

<table>
<thead>
<tr>
<th></th>
<th>Splitter</th>
<th>tst2011</th>
<th>tst2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Normal</td>
<td>36.11/7.921</td>
<td>32.68/7.743</td>
</tr>
<tr>
<td>C₁</td>
<td>Normal</td>
<td>36.23/7.946</td>
<td>32.48/7.722</td>
</tr>
<tr>
<td>C₂</td>
<td>Aggressive</td>
<td>35.82/7.881</td>
<td>32.68/7.725</td>
</tr>
</tbody>
</table>

- **P**: 4-gram Mix LM
- **C₁**: 5-gram Mix LM
- **C₂**: 6-gram Mix LM
Dutch-English results

<table>
<thead>
<tr>
<th>Splitter</th>
<th>tst2011</th>
<th>tst2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Normal</td>
<td>36.11/7.921</td>
</tr>
<tr>
<td>C₁</td>
<td>Normal</td>
<td>36.23/7.946</td>
</tr>
<tr>
<td>C₂</td>
<td>Aggressive</td>
<td>35.82/7.881</td>
</tr>
</tbody>
</table>

- P: 4-gram Mix LM
- C₁: 5-gram Mix LM
- C₂: 6-gram Mix LM
Conclusion

- We present several ideas for Arabic-, Turkish-, and Dutch-English machine translation

- Contributions:
  - Early distortion limit (Arabic, attempted w/ Turkish)
  - Morphological Segmentation (Turkish)
  - Compound Splitting (Dutch)
  - Corpora Filtering