English-Hindi Translation in 21 Days

Ondřej Bojar, Pavel Straňák, Daniel Zeman

ÚFAL MFF, Univerzita Karlova, Praha
Data

• Parallel (en-hi)
  – TIDES (50k training sentences, 1.2M hi words)
  – EILMT (7k training sentences, 181k hi words)
  – EMILLE (200k en words)
  – Daniel Pipes (322 texts)
  – Agriculture (17k en ~ 13k hi words)

• Monolingual (hi)
  – Hindi news web sites (18M sentences, 309M words)
Impact of additional data

• Larger parallel data helps
  – Test data: EILMT
  – Training & dev data:
    • EILMT              18.88 ± 2.05
    • EILMT+TIDES       19.27 ± 2.22
    • EILMT+TIDES+20k web sents 20.07 ± 2.21
Impact of additional data

• Larger Hindi LM data does not help
  – Test data: EILMT
  – Parallel training data: EILMT + TIDES + 20k web sentences
  – LM training data:
    • EILMT + web (>300M words): $18.82 \pm 2.13$
    • EILMT (181k words): $20.07 \pm 2.21$
  – Out of domain
  – Incompatible tokenization?
Moses setup

• Alignment heuristics: grow-diag-final-and (GDFA)
  – 4 times more extracted phrases than GDF
  – BLEU + 5 points (*table*)
## Alignment heuristics

<table>
<thead>
<tr>
<th></th>
<th>EILMT</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>grow-diag-final</td>
<td>13.82 ± 1.46</td>
<td>14.67 ± 1.46</td>
</tr>
<tr>
<td>grow-diag-final-and</td>
<td>18.88 ± 2.05</td>
<td>20.07 ± 2.21</td>
</tr>
</tbody>
</table>
## Alignment heuristics: CS-EN

<table>
<thead>
<tr>
<th></th>
<th>CS to EN</th>
<th>EN to CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>grow-diag-final</td>
<td>17.37 ± 0.46</td>
<td>14.40 ± 0.88</td>
</tr>
<tr>
<td>grow-diag-final-and</td>
<td>17.67 ± 0.44</td>
<td>14.50 ± 0.87</td>
</tr>
</tbody>
</table>
Moses settings

• Alignment using first four characters ("light stemming")
  – helps with GDF (not significantly)
  – does not help with GDFA (not significantly)

• MERT tuning of feature weights
  – (not included in official baseline)
Rule-based reordering

• Move finite verb forms to the end of the sentence (not crossing punctuation, “that”, WH-words).
• Transform prepositions to postpositions
• TectoMT, Morče tagger (perceptron), McDonald’s MST parser
Reordering example

Technology **is** the most obvious part: the telecommunications revolution **is** far more pervasive and spreading more rapidly than the telegraph or telephone **did in** their time.

Technology the most obvious part **is**: the telecommunications revolution far more pervasive **is** and spreading more rapidly than the telegraph or telephone **their time in did**.
Unsupervised stem-suffix segmentation

• Factors in Moses
  – Lemma + tag: but we do not have a tagger
  – Stem + suffix: unsupervised learning is language independent

  – A tool by Dan Zeman (Morpho Challenge 2007, 2008)
Core Idea

• Assumption: 2 morphemes: stem+suffix
  – Suffix can be empty
• All splits of all words
  – (into a stem and a suffix)
• Set of suffixes seen with the same stem is a paradigm
  – In a wider sense, paradigm = set of suffixes + set of stems seen with the suffixes
Paradigms get filtered

• Remove the paradigm if:
  – There are more suffixes than stems
  – All suffixes begin with the same letter
  – There is only one suffix

• Merge paradigms A and B if:
  – B is subset of A
  – A is the only superset of B
Paradigm Examples (en)

• Suffixes: e, ed, es, ing, ion, ions, or
• Stems: calibrat, decimat, equivocat, ...

• Suffixes: e, ed, es, ing, ion, or, ors
• Stems: aerat, authenticat, disseminat, ...

• Suffixes: 0, d, r, r’s, rs, s
• Stems: analyze, chain-smoke, collide, ...
Paradigm Examples (hi)

• Suffixes: 0, ा, े, ों
• Stems: अहात, खांच, घुटन, चढाव, …

• Suffixes: 0, ं, ंगे, गा
• Stems: कराए, दर्शाए, पेंके, बदले, …

• Suffixes: 0,ि,ियां, ियों
• Stems: अनुभूत, अभिव्यक्त, …
Learning Phase Outcomes

- List of paradigms
- List of known stems
- List of known suffixes
- List of stem-suffix pairs seen together

- How can we use that to segment a word?
Morphemic Segmentation

• Consider all possible splits of the word
  1. Stem & suffix known and allowed together
  2. Stem & suffix known but not together
  3. Stem is known
  4. Suffix is known
  5. Both unknown

• We use 4 (longest known suffix)
Impact of our preprocessing

<table>
<thead>
<tr>
<th></th>
<th>EILMT</th>
<th>TIDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Moses, Distance Reordering</td>
<td>18.88±2.05</td>
<td>10.06±0.76</td>
</tr>
<tr>
<td>Baseline Moses, Reordering Using en+hi Forms</td>
<td>19.77±2.03</td>
<td>10.95±0.75</td>
</tr>
<tr>
<td>Suffix LM+Reord</td>
<td>20.09±2.18</td>
<td>10.18±0.74</td>
</tr>
<tr>
<td>Rule-based Reordering + Suffix LM+Reord</td>
<td>21.01±2.18</td>
<td>10.29±0.69</td>
</tr>
</tbody>
</table>