Conclusions

53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015), Beijing, China
m.e.vanderwees@uva.nl

What’s in a Domain?

Analyzing Genre and Topic Differences in Statistical Machine Translation

Marlies van der Wees
Informatics Institute, University of Amsterdam

Arianna Bisazza

The Gen&Topic Evaluation Set

What’s in a Domain?

Impact of Genre and Topic Differences on SMT

The Gen&Topic Evaluation Set

Arabic-English
web-crawled manual translations

Two genres
newswire (NW) & comments (UG)

Five topics
culture, health, politics, economy & security

Controlled genre-topic distributions
similar-sized NW and UG fragments from all NW-UG document pairs that discuss the same article:

News (NW)

The 12 contestants competed during a May 3rd Prime before a panel of judges and millions of viewers across the Arab world.

Comment (UG)

Yemen is mulling the establishment of 13 industrial zones across its six planned administrative regions in a bid to stimulate development and create job opportunities.

Our contributions
- disentangling the concepts of genre and topic
- introducing the Gen&Topic evaluation set
- quantifying the impact of genre and topic on SMT
- analyzing OOV types for genres and topics

Translation model
balanced genres: 50% NW, 50% UG
wide range of topics

Language model
linear interpolation covers all genres & topics

Translation quality (BLEU)

\[
\Delta_{NW} = 3.9 \quad \Delta_{UG} = 1.1
\]

Average source-side phrase length

\[
\Delta_{NW} = 0.07 \quad \Delta_{UG} = 0.03
\]

Phrase pair model coverage

\[
\Delta_{NW} = 4.5 \quad \Delta_{UG} = 1.9
\]

Manual OOV Analysis

Five out-of-vocabulary (OOV) classes
rare (rare words), morph (morphological variants), dial (dialectal forms), spell (spelling errors), coll (colloquialisms)

Examples:

<table>
<thead>
<tr>
<th>class</th>
<th>Arabic OOV</th>
<th>English translation</th>
<th>explanation of OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>rare</td>
<td>دا عَشُرٍ</td>
<td>ISIL</td>
<td>new proper noun</td>
</tr>
<tr>
<td>dial</td>
<td>يَفْتَرِسُونَ (they) will forget</td>
<td>dialectal future tense</td>
<td></td>
</tr>
<tr>
<td>morph</td>
<td>يَفْتَرِسُونَ (they) revere</td>
<td>third person plural</td>
<td></td>
</tr>
<tr>
<td>spell</td>
<td>تَفْتَرِسُونَ (they)</td>
<td>creationofjobs</td>
<td></td>
</tr>
<tr>
<td>coll</td>
<td>(they)</td>
<td>volunteeers</td>
<td></td>
</tr>
</tbody>
</table>

Differences between genres
rare words dominate NW
spelling errors dominate UG

Differences between topics
more subtle than genre differences, dial & morph least uniformly distributed

Advice for topic adaptation: improve lexical selection
Advice for genre adaptation: increase model coverage

Conclusions

Genre differences have larger impact on SMT than topic differences