OpenDial: Hybrid dialogue management

OpenSubtitles: Dialogue modelling for MT

Pierre Lison
University of Oslo

Svensk Dialogverkstad 2016, Göteborg
Outline of the talk

• Part 1: Dialogue management
  • A hybrid logical/probabilistic approach
  • The OpenDial toolkit

• Part 2: Dialogue modelling for SMT
  • General motivation
  • Dialogues from OpenSubtitles 2016
Part 1: dialogue management
Dialogue management (DM)

### Logical approaches
- Fine-grained control of conversation
- Limited account for uncertainties

### Statistical approaches
- Robust, data-driven models of dialogue
- Need large quantities of training data

- PhD thesis on hybrid approaches to DM
- Development of new representation for DM models: probabilistic rules
Two types of rules

**Probability rules**

- Conditional probability distributions between state variables

**General structure:**

```
if (condition₁) then
  P(effect₁) = θ₁,
  P(effect₂) = θ₂, ...
else if (condition₂) then
  P(effect₃) = θ₃, ...
...
```

**Utility rules**

- Utility functions for system actions given state variables

**General structure:**

```
if (condition₁) then
  U(action₁) = θ₁,
  U(action₂) = θ₂, ...
else if (condition₂) then
  U(action₃) = θ₃, ...
...
```
Demonstration of OpenDial

http://www.opendial-toolkit.net
Part 2: dialogue modelling for MT
MT and the role of context

- MT systems translate sentences in isolation
  - Source text viewed as unstructured "bag of sentences"
  - No use of linguistic information expressed at cross-sentential level

- Recent interest in *discourse* aspects of MT
  - Lexical cohesion, word-sense disambiguation, discourse connectives, verb tenses, pronominal anaphora, etc.

[see e.g. Hardmeier (2012) for a survey]

But so far little work on dialogue!
Example 1: Dialogue structure

A: Which way goes into town?
B: Right.

A: So, those two don’t work for Miletto. They work for Crenshaw.
B: Right.

A: Hvilken vei fører til byen?
B: Høyre.

A: Så de to arbeider ikke for Miletto. De arbeider for Crenshaw.
B: Riktig.

[Source: OpenSubtitles parallel corpus]
Example 2: fragments

A: Mother… what was it like for you?
B: For me?

A: Mor… hvordan var det for deg?
B: For meg?

A: You made this?
B: For me?

A: Har du bygget den?
B: Til meg?

[Source: OpenSubtitles parallel corpus]
Example 3: Entrainment

A: Please, don’t make the mistake of not taking me seriously, Roschmann.
B: I do take you seriously.

A: Ikke gjør den feilen å ikke ta meg på alvor, Roschmann.
B: Jeg tar Dem på alvor.

Reuse of expression “take X seriously”
The OpenSubtitles collection

• Collaboration with Jörg Tiedemann on a new major release of OpenSubtitles
  • Collection of bitexts extracted from movie & TV subtitles
  • 2.6 billion sentences in 60 languages!
  • Largest multilingual corpus currently available?
### Some statistics (20 biggest languages)

<table>
<thead>
<tr>
<th>Language</th>
<th>Number of files</th>
<th>Number of blocks</th>
<th>Covered IMDBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>70.1K</td>
<td>53.2M</td>
<td>34.1K</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>95.8K</td>
<td>68.1M</td>
<td>49.3K</td>
</tr>
<tr>
<td>Czech</td>
<td>134K</td>
<td>93.4M</td>
<td>51.3K</td>
</tr>
<tr>
<td>Greek</td>
<td>118K</td>
<td>216M</td>
<td>49.9K</td>
</tr>
<tr>
<td>English</td>
<td><strong>344K</strong></td>
<td><strong>347M</strong></td>
<td><strong>106K</strong></td>
</tr>
<tr>
<td>Spanish</td>
<td><strong>205K</strong></td>
<td><strong>167M</strong></td>
<td><strong>76.1K</strong></td>
</tr>
<tr>
<td>Finnish</td>
<td>46.9K</td>
<td>27.9M</td>
<td>31.8K</td>
</tr>
<tr>
<td>French</td>
<td>110K</td>
<td>200M</td>
<td>56.4K</td>
</tr>
<tr>
<td>Hebrew</td>
<td>85.0K</td>
<td>60.6M</td>
<td>35.6K</td>
</tr>
<tr>
<td>Croatian</td>
<td>106K</td>
<td>64.8M</td>
<td>41.3K</td>
</tr>
<tr>
<td>Hungarian</td>
<td>103K</td>
<td>78.6M</td>
<td>52.7K</td>
</tr>
<tr>
<td>Italian</td>
<td>98.9K</td>
<td>70.5M</td>
<td>41.9K</td>
</tr>
<tr>
<td>Dutch</td>
<td>104K</td>
<td>68.7M</td>
<td>46.6K</td>
</tr>
<tr>
<td>Polish</td>
<td>169K</td>
<td>122M</td>
<td>44.0K</td>
</tr>
<tr>
<td>Portuguese</td>
<td>102K</td>
<td>94.9M</td>
<td>36.2K</td>
</tr>
<tr>
<td>Portuguese (BR)</td>
<td><strong>228K</strong></td>
<td><strong>188M</strong></td>
<td><strong>77.0K</strong></td>
</tr>
<tr>
<td>Romanian</td>
<td>170K</td>
<td>134M</td>
<td>58.1K</td>
</tr>
<tr>
<td>Slovenian</td>
<td>58.6K</td>
<td>37.8M</td>
<td>22.8K</td>
</tr>
<tr>
<td>Serbian</td>
<td>164K</td>
<td>226M</td>
<td>56.3K</td>
</tr>
<tr>
<td>Turkish</td>
<td><strong>181K</strong></td>
<td><strong>115M</strong></td>
<td><strong>55.0K</strong></td>
</tr>
</tbody>
</table>
Turn segmentation

• Subtitles lack an important information for dialogue modelling: the *turn structure*!

• Ideally, we could use the audiovisual data
  • But requires access to large amounts of copyrighted material!

• Joint work with Raveesh on automatic turn segmentation from subtitles

  • Step 1: create "annotated" data using movie scripts
  • Step 2: train a classifier on this data
Alignment from movie scripts

INT. CARGO SHIP - NARROW CORRIDOR - DAY

A PORTAL opens. The GUAVIAN DEATH GANG enters. One man in a SUIT (BALA-TIK), and five SECURITY SOLDIERS in badass UNIFORMS with ROUND-FACE HELMETS. They turn into and stop at one end of the corridor. Han, Chewie and BB-8 forty feet away in the middle of the long hall.

BALA-TIK

Han Solo: You are a dead man.

Bala-Tik. You are a dead man. Han smiles innocently, friendly. BB-8 nervously looks back and forth at the gang, and Han.

HAN

Bala-Tik. What's the problem?

BALA-TIK

The problem is we loaned you fifty thousand for this job.

INTERCUT WITH:

INT. CARGO SHIP - BELOW FLOOR GRATING - DAY

They look up, trying to get a view.

REY

Can you see them?

786,195 sentences annotated with speaker notation
Turn segmentation

• Classification on consecutive sentence pairs, with two outputs: same or new turn
• Combination of various linguistic, contextual and temporal features
• Modest accuracy: 0.78 on test data
  • But human also find the task difficult: Fleiss' $\kappa$ of 0.35 with three annotators on 100 sentence pairs
Conclusion

- **OpenDial**: an open-source toolkit for developing spoken dialogue systems
  - Well-suited for domains that combine a complex state-action space and little to no training data

- **Project on dialogue modelling for MT**
  - Released a large (2.6G sentences!) collection of corpora extracted from movie & TV subtitles
  - *Current work*: extract useful, dialogue-related features from this data