UiO : University of Oslo

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	**Statistical approaches**
+	Fine-grained control of conversation	Robust, data-driven models of dialogue
-	Limited account for uncertainties	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	Utility rules
What they encode:	*Conditional probability distributions* between state variables	*Utility functions* for system actions given state variables

General structure:

```
if ($condition_1$) then
    $P(effect_1) = \theta_1$,
    $P(effect_2) = \theta_2$,   ...
else if ($condition_2$) then
    $P(effect_3) = \theta_3$,  ...
...
```

```
if ($condition_1$) then
    $U(action_1) = \theta_1$,
    $U(action_2) = \theta_2$, ...
else if ($condition_2$) then
    $U(action_3) = \theta_3$,...
...
```

Demonstration of OpenDial

OpenDial

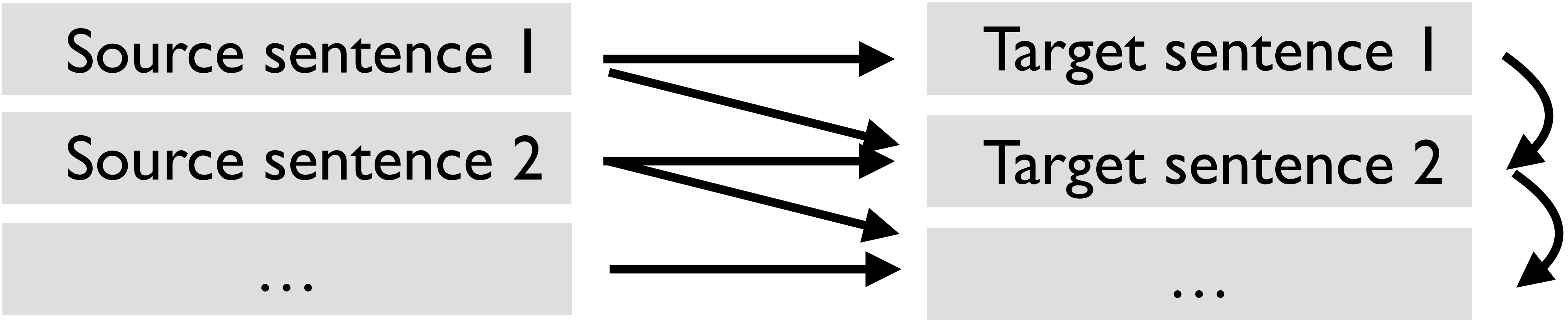
http://www.opendial-toolkit.net

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]



Source sentence 1

Source sentence 2

...

Target sentence 1

Target sentence 2

...

But so far little work on dialogue!

Example 1: Dialogue structure

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

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

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

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

All right, do your stuff!

好，一展身手吧！

- 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)

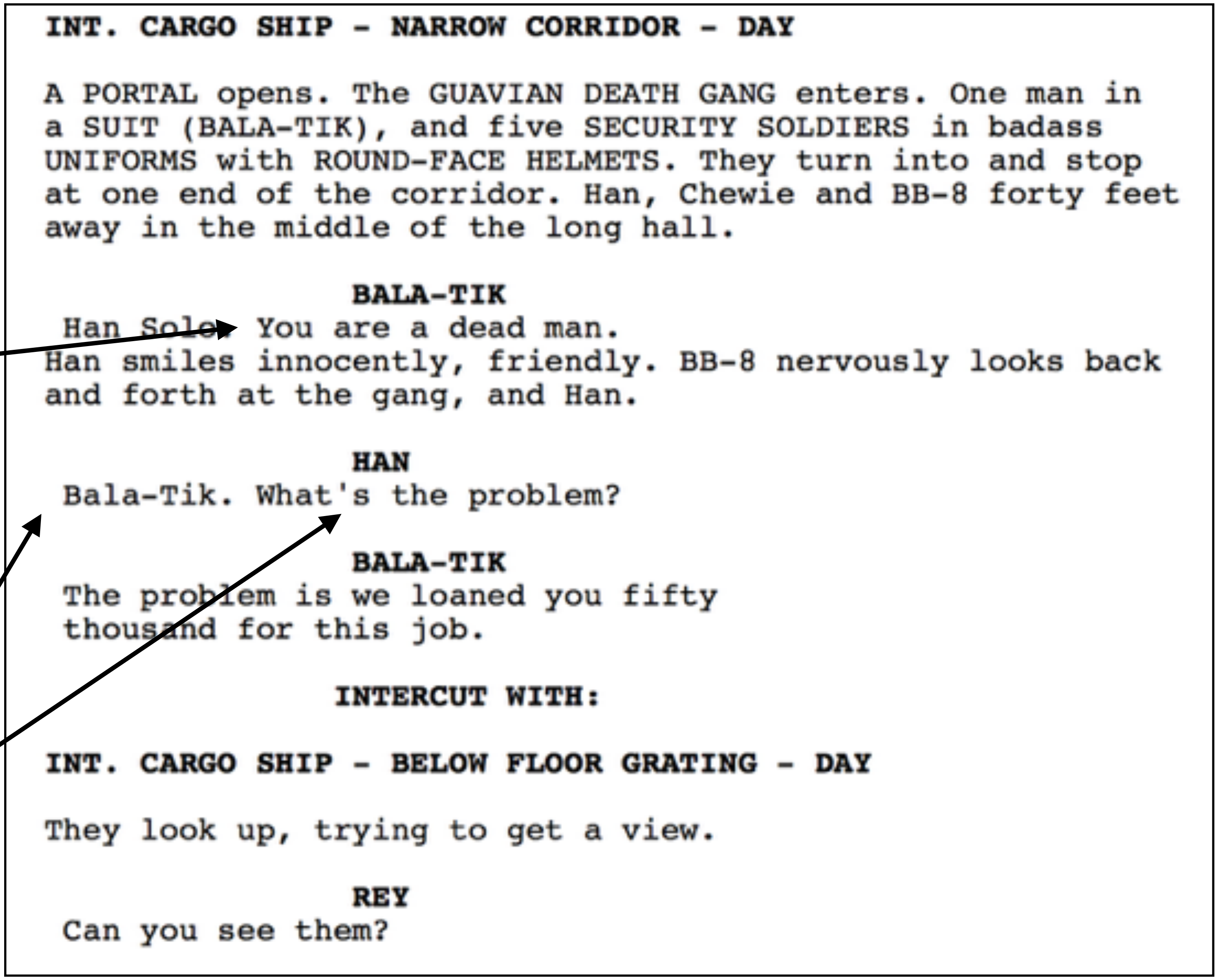
Language	Number of files	Number of blocks	Covered IMDBs
Arabic	70.1K	53.2M	34.1K
Bulgarian	95.8K	68.1M	49.3K
Czech	134K	93.4M	51.3K
Greek	118K	216M	49.9K
English	**344K**	**347M**	**106K**
Spanish	**205K**	**167M**	**76.1K**
Finnish	46.9K	27.9M	31.8K
French	110K	200M	56.4K
Hebrew	85.0K	60.6M	35.6K
Croatian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

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

```
<s id="799">
  <time id="T600S" value="00:43:58,262" />
  <w id="799.1">You</w>
  <w id="799.2">'re</w>
  <w id="799.3">a</w>
  <w id="799.4">dead</w>
  <w id="799.5">man</w>
  <w id="799.6">.</w>
  <time id="T600E" value="00:43:59,722" />
</s>
<s id="800">
  <time id="T601S" value="00:43:59,847" />
  <w id="800.1">Bala–Tik</w>
  <w id="800.2">.</w>
</s>
<s id="801">
  <w id="801.1">What</w>
  <w id="801.2">'s</w>
  <w id="801.3">the</w>
  <w id="801.4">problem</w>
  <w id="801.5">?</w>
  <time id="T601E" value="00:44:02,558" />
</s>
```



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.

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' κ 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