

# Dialogue modelling: small data and big data

Pierre Lison, NR  
[plison@nr.no](mailto:plison@nr.no)

USC Institute for Creative Technologies  
13/12/2016



# Two parts

- ▶ **Part 1: Dialogue management**
  - Modelling approach
  - OpenDial toolkit
- ▶ **Part 2: Large-scale dialogue resources**
  - The OpenSubtitles collection
  - Detecting turn boundaries

# Part 1: Dialogue management

*Can we estimate probabilistic models of dialogue with small amounts of data?*

# Dialogue management

## Symbolic approaches

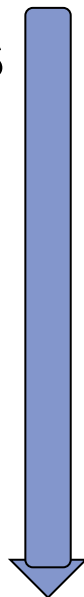
 Fine-grained control over interaction

 Limited account for uncertainties

## Statistical approaches

Robust, data-driven models of dialogue

Needs large amounts of training data



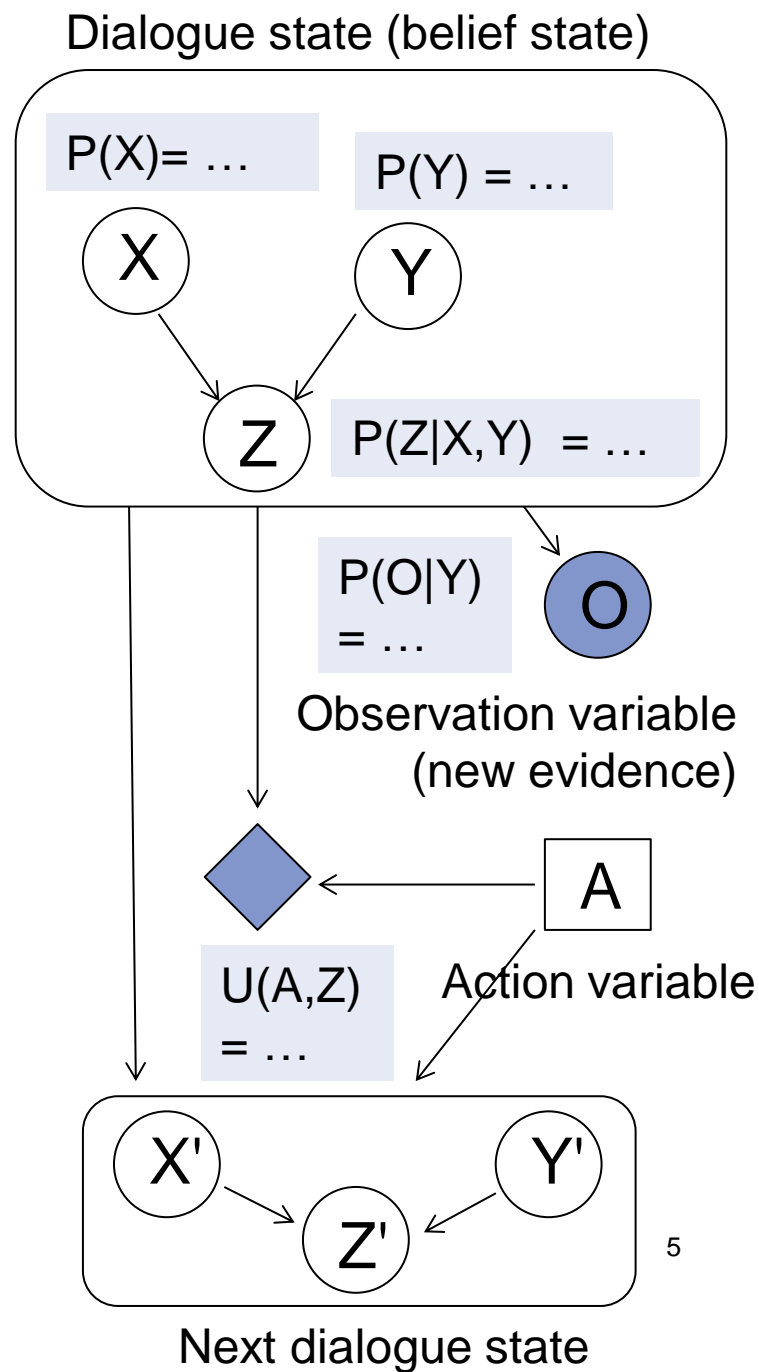
**Hybrid** approach that combines *probabilistic reasoning with expert knowledge* and *small amounts of data*

Dialogue management models represented by **probabilistic rules**

(*abstraction layer* on top of classical probabilistic models)

# Key idea

- ▶ Dialogue state encoded as a Bayesian Network
  - Each variable captures a relevant aspect of the interaction
  - Regularly updated with new observations (dialogue acts etc.)
  - Used to select system actions
  - (and predict future states)
- ▶ But the probabilistic models are expressed using a *high-level representation*
  - Making use of *logical abstractions*



# Rule structure

```
if (condition1 on X) then
  P(Y=val1) =  $\theta_1$ 
  P(Y=val2) =  $\theta_2$ 
  ...
else if (condition2 on X) then
  P(Y=val3) =  $\theta_3$ 
  P(Y=val4) =  $\theta_4$ 
  ...
else
  ...
```

- ▶ If-then-else skeleton, which maps between conditions and (probabilistic) effects
- ▶ Specifies a probability distribution  $P(\mathbf{Y}|\mathbf{X})$  where  $\mathbf{X}$  and  $\mathbf{Y}$  are state variables
- ▶ *Conditions*: logical formulae on some state variables  $\mathbf{X}$
- ▶ *Effects*: assignment of values to some state variables  $\mathbf{Y}$
- ▶ Can include logical operators, universal quantifiers, etc.

# Rule structure (utility functions)

```
if (condition1 on X) then
  U(A=val1) =  $\theta_1$ 
  U(A=val2) =  $\theta_2$ 
  ...
else if (condition2 on X) then
  U(A=val3) =  $\theta_3$ 
  U(A=val4) =  $\theta_4$ 
  ...
else
  ...
```

- ▶ If-then-else skeleton, which maps between conditions and utility functions
- ▶ Specifies a utility function  $U(\mathbf{X}, \mathbf{A})$  where  $\mathbf{X}$  are state variables,  $\mathbf{A}$  action variables
- ▶ *Conditions*: logical formulae on some state variables  $\mathbf{X}$
- ▶ *Effects*: assignment of values to the action variables  $\mathbf{A}$

# Two examples

$\forall x,$

if (*last-user-act* =  $x \wedge$  *system-action* = AskRepeat) then  
 $P(\textit{next-user-act} = x) = 0.9$

“If the system asks the user to repeat his last dialogue act  $x$ , the user is predicted to comply and repeat  $x$  with probability 0.9”

---

$\forall x,$

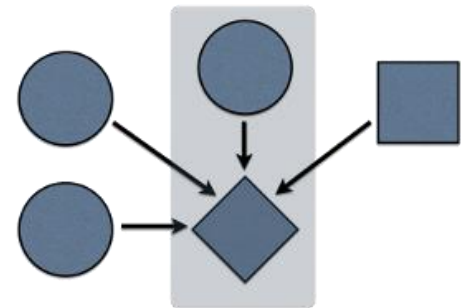
if (*last-user-act* = Request( $x$ )  $\wedge$   $x \in$  *perceived-objects*) then  
 $U(\textit{system-action} = \textit{PickUp}(x)) = +5$

“If the user asks the system to pick up a given object  $x$  and  $x$  is perceived by the system, then the utility of picking up  $x$  is 5”



# Rule instantiation

- ▶ At runtime, the rules are “executed” by instantiating them as distinct nodes in the dialogue state
  - Probabilistic rules are thus “high-level templates” for traditional probabilistic (graphical) models
- ▶ State update & action selection are then performed using well-known inference algorithms (e.g. sampling)
- ▶ (Demonstration in OpenDial)



OpenDial

# Parameter estimation

- ▶ The probabilistic rules may include parameters  $\theta$  to learn (unknown probabilities or utilities)
- ▶ Bayesian approach:
  - Start with a prior distribution  $P(\theta)$
  - Collect some data  $\mathcal{D}$  and estimate the posterior  $P(\theta|\mathcal{D})$ :
  - Can be applied to both supervised (e.g. on Wizard-of-Oz data) or reinforcement learning

$$P(\theta | \mathcal{D}) = \eta P(\mathcal{D}; \theta) P(\theta)$$

Posterior distribution      Normalisation factor      Likelihood of the data      Prior distribution

# OpenDial

- ▶ A software toolkit to develop (spoken) dialogue systems
  - Blackboard architecture centered on the dialogue state
  - External modules for ASR, NLU, TTS, etc.
  - The toolkit itself is fully domain-independent; domain-specific information is provided through XML files
- ▶ Deployed in several application domains:
  - Human-robot interaction
  - Multi-modal navigation assistants
  - Intelligent tutoring systems

Lison, P. (2015). A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232-255

Lison, P. and Kennington, C. (2016). *OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules*. ACL.

# **Part 2: Building large-scale NLP resources with movie & TV subtitles**

*Joint work with Jörg Tiedemann (U. Helsinki)  
and Raveesh Meena (KTH, Sweden)*

# Movie & TV subtitles

- ▶ Subtitles are a very interesting resource for NLP:
  1. Broad spectrum of linguistic genres, multiple speaker styles, complex conversational structure, etc.
  2. Huge amounts of training data available (more than 3 million subtitles in [www.opensubtitles.org](http://www.opensubtitles.org) )
  3. Tight coupling between the subtitles and their source material → easier to align across languages
  4. Can be augmented with meta-data, audio-visual sources, etc.



# OpenSubtitles 2016



opensubtitles  
.org

- ▶ We compiled a new major release of the **OpenSubtitles** corpus collection:
  - 2.8 million subtitles (**17.2** billion tokens) covering no less than **60** languages!
  - The subtitles are first preprocessed (format conversion, sentence segmentation, tokenisation, spellchecking, inclusion of meta-data, etc.) and stored in XML format
  - The subtitles are then aligned at document- and sentence-level across all language pairs

The complete collection is freely available in the OPUS corpus repository!

# The initial dataset

- ▶ The administrators of [www.opensubtitles.org](http://www.opensubtitles.org) kindly provided us with a full dump of their database
  - 3.36 million subtitle files
  - with meta-data for each subtitle (language code, format, IMDB identifier, subtitle rating, etc.)
- ▶ The subtitles are structured in **blocks**, short text segments associated with a start and end time.

5

00:01:15,200 --> 00:01:20,764

Nehmt die Halme, schlagt sie oben ab,  
entfernt die Blätter

6

00:01:21,120 --> 00:01:24,090

und werft alles auf einen Haufen  
für den Pflanztrupp.

7

00:01:24,880 --> 00:01:30,489

Das Zuckerrohr beißt euch nicht.  
Nicht so zaghaft! Na los, Burschen, los!

# Some statistics (20 biggest languages)

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



# Preprocessing

## 1. Conversion to Unicode

5

00:01:15,200 --> 00:01:20,764

Nehmt die Halme, schlägt sie oben ab,  
entfernt die Blätter

6

00:01:21,120 --> 00:01:24,090

und werft alles auf einen Haufen  
für den Pflanztrupp.

7

00:01:24,880 --> 00:01:30,489

Das Zuckerrohr beißt euch nicht.  
Nicht so zaghaft! Na los, Burschen, los!

# Preprocessing

1. Conversion to Unicode
2. Sentence segmentation

5

00:01:15,200 --> 00:01:20,764

Nehmt die Halme, schlägt sie oben ab,  
entfernt die Blätter

6

00:01:21,120 --> 00:01:24,090

und werft alles auf einen Haufen  
für den Pflanztrupp.

7

00:01:24,880 --> 00:01:30,489

Das Zuckerrohr beißt euch nicht.  
Nicht so zaghaft! Na los, Burschen, los!

Detection of sentence boundaries is language-specific (depends on writing system, punctuation marks, etc.)

# Preprocessing

1. Conversion to Unicode
2. Sentence segmentation
3. Tokenisation

5

00:01:15,200 --> 00:01:20,764

Nehmt die Halme , schlägt sie oben ab ,  
entfernt die Blätter

6

00:01:21,120 --> 00:01:24,090

und werft alles auf einen Haufen  
für den Pflanztrupp .

7

00:01:24,880 --> 00:01:30,489

Das Zuckerrohr beißt euch nicht .  
Nicht so zaghaft ! Na los , Burschen , los !

- tokenizer.perl script from Moses
- jieba and kytea libraries for Chinese and Japanese word segmentation

# Preprocessing

1. Conversion to Unicode
2. Sentence segmentation
3. Tokenisation
4. OCR error correction

4

00:01:34,000 --> 00:01:38,471

<i>"Along with the simplification  
I sought in my first films,</i>

5

00:01:39,080 --> 00:01:40,991

<i>"I wanted to be revolutionary,</i>

- Noisy-channel approach with Google N-grams (11 European languages)
- Total of 9.04 million corrected tokens
- > 99% precision, but lower recall

# Preprocessing

1. Conversion to Unicode
2. Sentence segmentation
3. Tokenisation
4. OCR error correction
5. Language identification

Lui & Baldwin (2012) `langid.py`: An Off-the-shelf language identification tool, ACL 2012

# Preprocessing

1. Conversion to Unicode
2. Sentence segmentation
3. Tokenisation
4. OCR error correction
5. Language identification
6. Extraction of meta-data

- **Infos on source media:**  
release year, language, duration, genre,...
- **Infos on the subtitle:**  
upload date, rating, duration,...
- **Infos on conversion:**  
file encoding, number of sentences, tokens,...

# Preprocessing

1. Conversion to Unicode
2. Sentence segmentation
3. Tokenisation
4. OCR error correction
5. Language identification
6. Extraction of meta-data
7. Generation of XML file

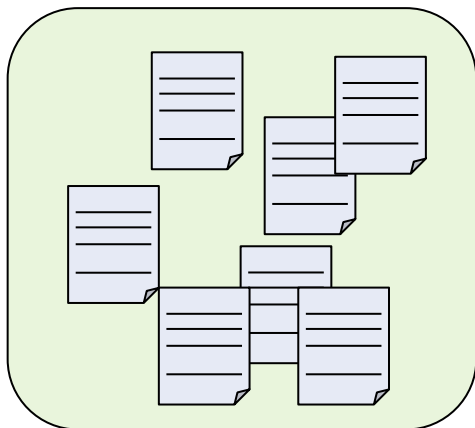
```
...  
<s id="801">  
  <w id="801.1">What</w>  
  <w id="801.2">'s</w>  
  <w id="801.3">the</w>  
  <w id="801.4">problem</w>  
  <w id="801.5">?</w>  
  <time id="T601E" value="00:44:02,558" />  
</s>  
...
```

- 2.8 M subtitles
- 153K movies & TV episodes
- 60 languages
- 2.6G sentences
- 17.2G tokens

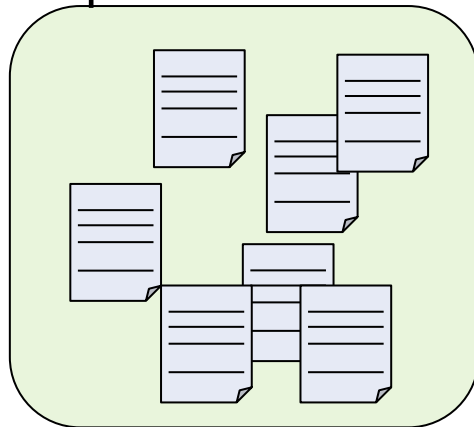
# Alignment

- ▶ The processed subtitles are then aligned with one another to create a collection of parallel corpora

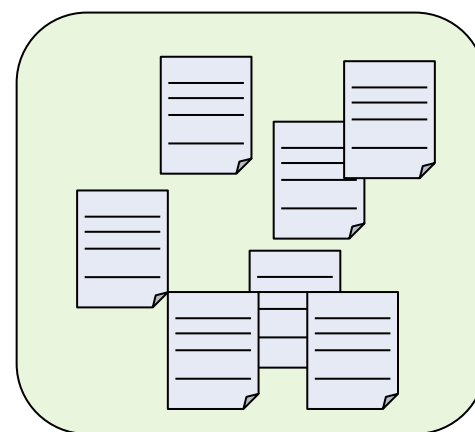
English subtitles



Japanese subtitles




Turkish subtitles

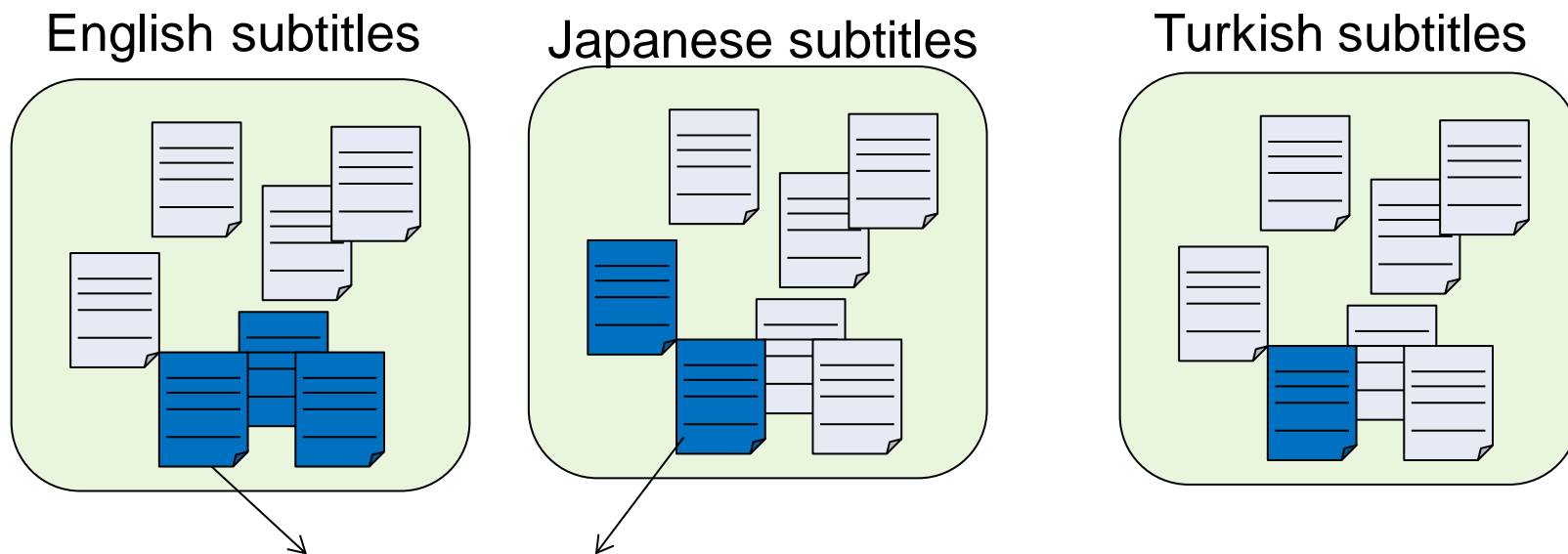




# Alignment


 = Subtitles for  
"Love actually" (2003),  
(using IMDB identifier)

- ▶ The processed subtitles are then aligned with one another to create a collection of parallel corpora

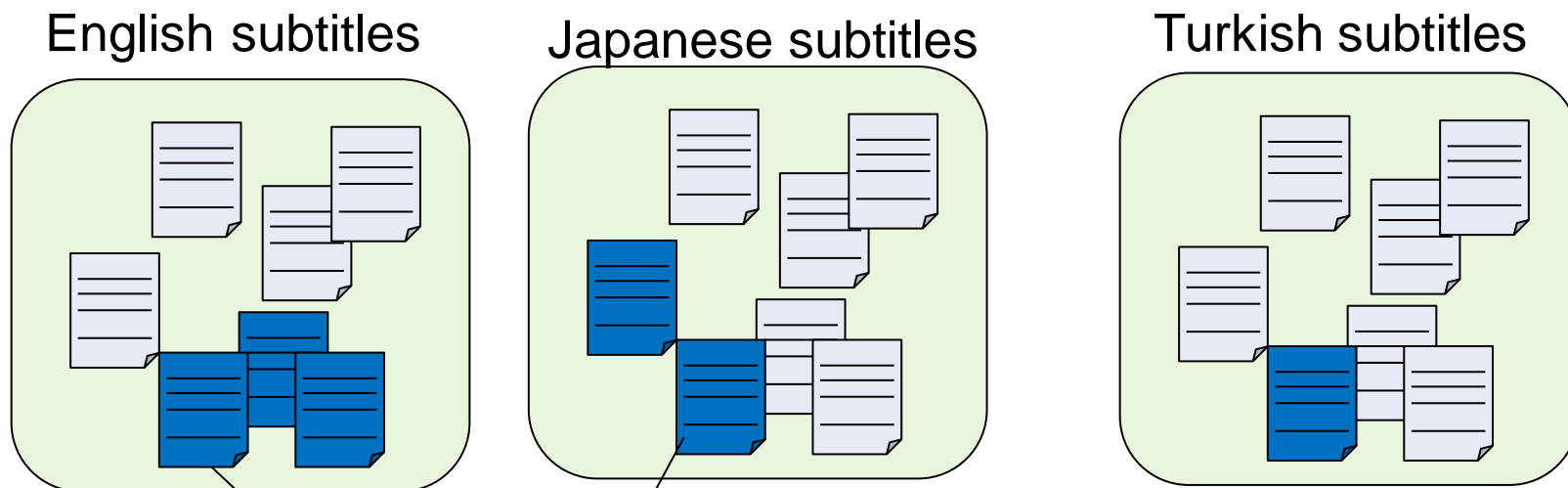


Handcrafted scoring function to determine the best subtitle pairs  
(based on subtitle quality measures + time overlap between the two)

# Alignment

 = Subtitles for  
"Love actually" (2003),  
(using IMDB identifier)

- ▶ The processed subtitles are then aligned with one another to create a collection of parallel corpora



Generation of sentence alignments  
based on **timing information**

- Unknown start/end times are interpolated
- Speed ratio and offset are adjusted using anchor points (e.g. cognates)

# Alignment results

- ▶ 1689 bitexts across 60 languages

Language pair	Aligned docs	Sentence pairs	Tokens
English - Spanish	62.2K	50.1M	620.0M
English - Portuguese (BR)	61.1K	47.6M	587.4M
Spanish - Portuguese (BR)	56.3K	43.0M	521.1M
English - Romanian	48.8K	39.5M	484.5M
English - Turkish	47.4K	37.3M	404.9M
Spanish - Romanian	45.5K	35.1M	431.1M
Portuguese (BR) - Romanian	45.5K	34.7M	422.2M
English - Serbian	44.1K	34.6M	411.6M
English - Hungarian	44.7K	33.6M	381.4M
English - French	42.6K	33.5M	432.1M

- ▶ + *intra-lingual alignments* between alternative subtitles

# Downstream tasks?

- ▶ Resources from movie and TV subtitles are already used for various NLP tasks:
  - Language modelling
  - Machine translation
  - Multilingual and cross-lingual NLP
  - Conversation modelling & dialogue systems

[e.g. Vinyals and Q. V. Le (2015), "A Neural conversational model", ICML Deep Learning Workshop]



- ▶ However, they lack a crucial piece of information: the **turn structure**
  - Who is speaking at a given time?

# Finding turn boundaries

ID	Utterance	Start time	End time
1	If we wanted to kill you, Mr Holmes, we would have done it by now.	01:17:34.76	01:17:37.75
2	We just wanted to make you inquisitive.	01:17:37.80	01:17:40.59
3	Do you have it?	01:17:42.40	01:17:43.91
4	Do I have what?	01:17:43.91	01:17:45.43
5	The treasure.	01:17:45.48	01:17:46.43
6	I don't know what you're talking about.	01:17:46.43	01:17:48.91
7	I would prefer to make certain.	01:17:48.96	01:17:52.03
8	Everything in the West has its price.	01:17:57.00	01:17:59.63
9	And the price for her life - information.	01:17:59.68	01:18:04.55



**Question:** can we automatically segment this dialogue into turns?

(without having access to the audiovisual material)

# Exploiting movie & TV scripts

- ▶ Subtitles do not contain speaker information...
- ▶ But movie and TV scripts (screenplays, transcripts, etc.) do!
- ▶ We crawled various websites hosting movie and TV scripts
  - Scrapped 7.5K transcripts
  - The dialogues can be markedly different from those in the subtitles!



**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?

**FINN**

No.  
They start crawling down the crawl space.

**BALA-TIK**

I heard you also borrowed fifty thousand from Kanjiklub.

**HAN**

You know you can't trust those little freaks! How long've we known each other?  
Rey and Finn arrive under the gang. They WHISPER:

**REY**

They have blasters...

# Aligning scripts and subtitles

- ▶ We can then align the subtitles with the movie scripts, using two standard alignment toolkits:
  - One alignment for each <subtitle,script> pair

```
<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 GUVAVIAN 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?

**FINN**  
No.  
They start crawling down the crawl space.

**BALA-TIK**  
I heard you also borrowed fifty thousand from Kanjiklub.

**HAN**  
You know you can't trust those little freaks! How long've we known each other?  
Rey and Finn arrive under the gang. They WHISPER:

**REY**  
They have blasters...

# Alignment results

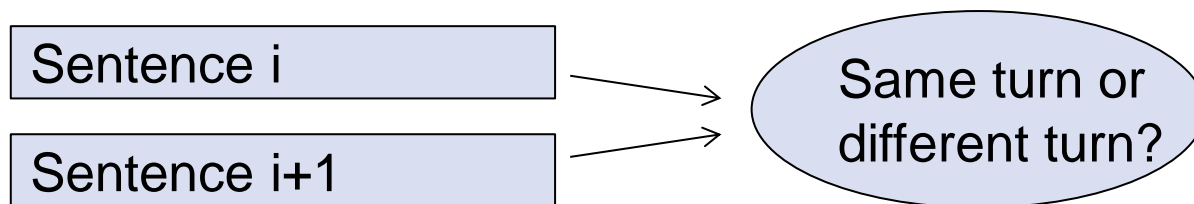
- ▶ Speaker labels from the scripts projected on 3.87 M sentences (~1% of total) in the English subtitles
  - Two aligners quite consistent (0.3% of conflicting labels)
  - Quality of the alignments? Comparison with a small, manually annotated corpus of TV series: 97.6% of the projected labels matched the manually labelled ones
- ▶ Using the cross-lingual alignments, we also projected the speaker labels onto 6 other languages

Language	Nb. of subtitles	Nb. of sentences
Arabic	1,340	1,413,326
Chinese	591	805,191
Czech	1,874	1,835,896
English	5,413	3,864,058
French	1,872	1,894,925
German	766	911,609
Turkish	1,863	1,953,208



# Predicting turn boundaries

- ▶ We can use the subtitles enriched with speaker labels to train a classifier to predict turn boundaries
- ▶ Using only *textual* and *timing* features from the subtitles



- ▶ Training data:
  - Consecutive sentence pairs in the subtitles that were annotated with speaker information (about 1.5M pairs)
  - Binary output: "*same turn*" if sentences  $i$  and  $i+1$  were part of the same turn in the aligned script, "*new turn*" otherwise
  - Balanced dataset: 52.3 % of "new turn" pairs

# Features

## Timing features:

*Time gaps and sentence durations*

## Length

*Nb. of characters/tokens in each sentence*

## Lexical features:

*BoW, bigrams, occurrence of negation/question words, pronouns*

## POS features

*POS tags and sequences, likely imperative mood (VB before NN or PRP and no question mark)*

## Punctuation features:

*Marks at start/end of each sentence*

## Edit distance features

*Token-level dist. between the two sentences*

## Adjacency features

*Occurrence of specific patterns, such as*

- *Likely polar answer*
- *Likely clarification request*
- *Pronoun inversion*

## Global features

*Occurrence of character names, movie genre, sentence/token density, sentence number*

## Alignment features

*Proportion of inter- and intra-lingual alignments in the OpenSubtitles bitexts.*

## "Visual" features

*Start/end of subtitle block*



(Alignments of type 2:1 are much more likely to occur if the two sentences are from the same speaker.)

# Experimental setup

## ► Baseline

- If second sentence starts with a “-” dash → new turn
- Otherwise, if the time gap is exactly zero → same turn
- Else, → new turn (majority class in this context)

## ► Alternative approaches:

- **Basic classifier**, trained with the high-performance classifier *Vowpal Wabbit* with all features + feature interactions
- **Extension 1 (multilingual classifier)**: if a sentence pair is aligned to sentence pairs in other languages, combine the output of all per-language classifiers in a weighted sum
- **Extension 2 (w/ speaker diarization)**: if the audio is available, perform speaker diarization and add a new feature encoding whether the two sentences belong to the same cluster

# Results

Approach	Turn	DEV				TEST			
		P	R	$F_1$	ACC	P	R	$F_1$	ACC
Baseline	Same	0.48	0.36	0.41	0.694	0.43	0.32	0.37	0.669
	New	0.81	0.98	0.89		0.80	0.98	0.88	
Classifier (basic)	Same	0.80	0.74	0.76	0.789	0.79	0.71	0.75	0.775
	New	0.78	0.84	0.81		0.77	0.83	0.80	
Classifier (multiling)	Same	0.80	0.74	0.77	<b>0.794*</b>	0.79	0.72	0.75	<b>0.781*</b>
	New	0.79	0.84	0.81		0.77	0.84	0.80	

Accuracy, precision, recall and F1 scores based on the development set (197K sentence pairs) and test set (200K sentence pairs). The best results are written in bold and are all statistical significant using a bootstrap test (p-values < 0.0001)

# Results (with diarization)

Approach	Turn	TREE HILL			
		P	R	$F_1$	ACC
Baseline	Same	0.32	0.22	0.26	0.595
	New	0.75	1.00	0.85	
Classifier (basic)	Same	0.85	0.68	0.76	0.774
	New	0.72	0.87	0.79	
Classifier (multiling)	Same	/	/	/	/
	New	/	/	/	
Diarization only	Same	0.75	0.39	0.51	0.617
	New	0.57	0.86	0.69	
Classifier+Diarization	Same	0.85	0.68	0.76	<b>0.775*</b>
	New	0.72	0.87	0.79	

Accuracy, precision, recall and F1 scores on a small dataset with one season (21 episodes of ~ 40 minutes each) of the “One Tree Hill” TV series, using the LIUM toolkit for speaker diarization.

The best result is statistical significant with p-value = 0.013

# Conclusion (Part 2)

- ▶ **OpenSubtitles 2016**: the world's largest collection of parallel corpora currently available
  - 17.2 billion tokens in 60 languages
  - Widely used for MT and multilingual NLP
  - But also very interesting for dialogue modelling
  - Can be aligned with movie & TV scripts to extract speaker labels and train a detector of turn boundaries
- ▶ Feel free to get in touch if you would like to collaborate on exploiting this corpus for downstream tasks
  - E.g. building chatbots with small-talk capabilities?
  - Or computer-assisted language learning?