Automatic turn segmentation for movie & TV subtitles

Pierre Lison (plison@nr.no), Norwegian Computing Center (NR), Oslo
Raveesh Meena (raveesh@kth.se), KTH Royal Institute of Technology, Stockholm

Motivation

Subtitles are a very useful resource for dialogue processing tasks :

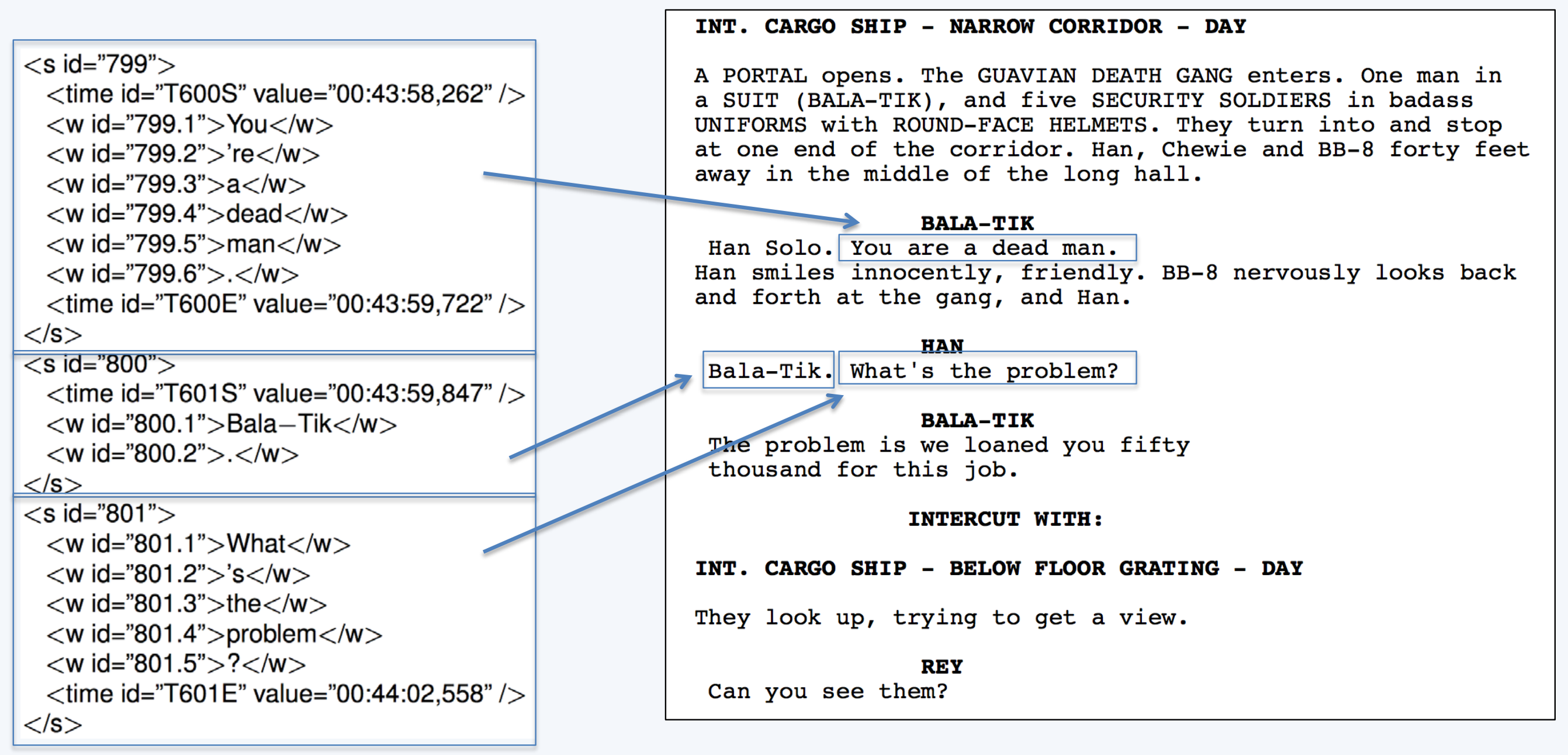
- Wide range of linguistic genres (incl. colloquial language), multiple speaker styles, complex conversational structures, etc.
- Large amounts of training data available online *(OpenSubtitles 2016*: 17.2 billion tokens covering no less than 60 languages!)
- **Use cases**: language modelling, machine translation, neural conversation modelling, dialogue systems, etc.

But:

- They lack an important piece of information: the ***turn structure***!
- Can we automatically segment subtitles into dialogue turns? (without requiring access to the original audio material)

Alignment with movie & TV scripts

- We crawled various websites with movie and TV scripts and extracted 7,467 dialogue transcripts (1,069 movies and 6,398 TV episodes).
- We then applied two sentence aligners (`hunalign` and `bleualign`) on each pair <subtitle,script>:
- Based on these alignments, the speaker labels from the scripts were then *projected* onto the sentences from the subtitles.



```
<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?
```

- 5,413 English-language subtitles were labelled in this manner, covering on average 34% of the sentences in movies and 60% for TV episodes.
- We also used existing cross-lingual alignments to project speaker labels on 6 other languages (see Table 1).

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

Table 1. Number of subtitles and sentences per language automatically annotated with speaker labels.

Key idea

- Subtitles do not contain speaker information… but movie and TV scripts (screenplays, transcripts) do!
- **Approach**:
  1. Crawl the web for movie and TV scripts
  2. *Align* (at sentence-level) these scripts with the subtitles
  3. *Project* speaker labels on the subtitles based on the alignment
  4. Use the resulting data to create a dataset of *turn boundaries*
  5. *Learn a predictor* of turn boundaries from this training data
  6. Apply the estimated model to segment subtitles into turns!

Turn-taking example

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

Prediction model

- *Learning task*: given two consecutive sentences, predict whether a turn boundary exists between the two:

```
Sentence i   ──┐
Sentence i+1 ──┴─→ Same turn or new turn?
```

- *Dataset*: about 1.5M consecutive sentence pairs with projected speaker labels extracted from the subtitles.
- *Binary output*: "same turn" if the two sentences *i* and *i+1* were part of the same turn in the aligned script, else "new turn" (balanced dataset: 52.3 % of "new turn" pairs)
- Discriminative linear classifier with Vowpal Wabbit (with the features on the right + feature interactions)

Feature types:

Timing	*Time gaps and sentence durations*
Length	*Nb. of characters/tokens in each sentence*
Lexical	*BoW, bigrams, negation/question words, pronouns*
POS	*Part-of-speech tags & sequences, imperative mood*
Punctuation	*Marks at start/end of each sentence*
Edit distance	*Token-level distance between the two sentences*
Adjacency	*Specific patterns, such as likely polar answer, likely clarification request, pronoun inversion, etc.*
Global	*Character names, movie genre, sentence density, etc.*
Alignment	*Proportion of intra/inter-lingual alignments*
Visual	*Start/end of subtitle block*

Extensions

1. **Multilingual classifier**: if sentence pair is aligned to sentence pairs in other languages, combine the outputs of all per-language classifiers in weighted sum.
2. **with speaker diarization**: if audio is available, perform speaker diarization and add a new feature encoding whether the two sentences belong to the same cluster.

Experimental results

Baseline:

1. If sentence 2 starts with dash → new turn.
2. Else, if the 2 sentences belong to same "block" → same turn.
3. Else, → new turn (majority class in this context).

		Dev				Test				Tree Hill			
Approach	Turn	P	R	F_1	Acc	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	0.32	0.22	0.26	0.595
	New	0.81	0.98	0.89		0.80	0.98	0.88		0.75	1.00	0.85	
Classifier (basic)	Same	0.80	0.74	0.76	0.789	0.79	0.71	0.75	0.775	0.85	0.68	0.76	0.774
	New	0.78	0.84	0.81		0.77	0.83	0.80		0.72	0.87	0.79	
Classifier (multiling)	Same	0.80	0.74	0.77	**0.794***	0.79	0.72	0.75	**0.781***	/	/	/	/
	New	0.79	0.84	0.81		0.77	0.84	0.80		/	/	/	
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	**0.775***
	New	/	/	/		/	/	/		0.72	0.87	0.79	

Precision, recall, F1 and accuracy on the dev set (197K sentence pairs), test set (200K pairs), and the small Tree Hill data. The best results are in bold (p-values = 0.013 for Tree Hill and < 0.0001 for dev and test sets).

Results 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.

	Baseline	Classifier (basic)
Arabic	0.588	**0.716**
French	0.663	**0.743**
German	0.656	**0.741**
Czech	0.668	**0.756**
Turkish	0.662	**0.758**
Chinese	0.569	**0.670**

Compared accuracies for the baseline and classifier for the 6 languages other than English (on test set).