Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Automatic turn segmentation for movie and TV subtitles

Pierre Lison, NR
plison@nr.no
(joint work with Raveesh Meena, KTH)

LTG seminar, UiO
29/11/2016

Introduction

- Movie & TV subtitles are a great resource for NLP:
  - **Linguistic perspective**: Broad spectrum of linguistic genres & speaker styles (including colloquial language), non-sentential utterances, complex conversational structures, etc.

  - **Data-driven perspective**: Huge amounts of available data (and meta-data), covering many languages (2.8M subtitles in 60 languages in OpenSubtitles 2016)

Introduction

- 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]

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

Introduction

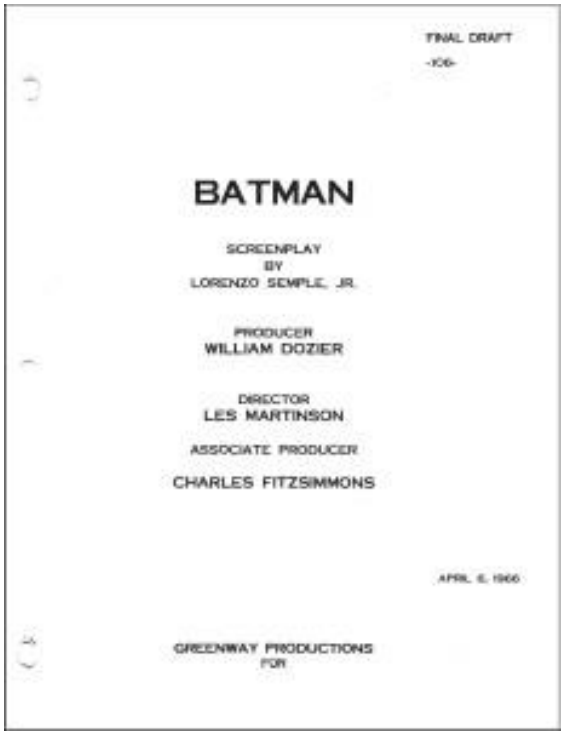
ID	Utterance	Start time	End time
1	If we wanted to kill you, Mr Holmes, we would have done it by now.	01:17:34.76	01:17:37.75
2	We just wanted to make you inquisitive.	01:17:37.80	01:17:40.59
3	Do you have it?	01:17:42.40	01:17:43.91
4	Do I have what?	01:17:43.91	01:17:45.43
5	The treasure.	01:17:45.48	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)

Key idea

- Subtitles do not contain speaker information…
- But movie and TV scripts (screenplays, transcripts, etc.) do!
- Outline of approach:
  1. Align the subtitles with movie and TV scripts
  2. Use alignments to *project* speaker information on the subtitles
  3. Use the subtitles augmented with speaker information to train a classifier that detects turn boundaries


FINAL DRAFT

-106-

BATMAN

SCREENPLAY
BY
LORENZO SEMPLE, JR.

PRODUCER
WILLIAM DOZIER

DIRECTOR
LES MARTINSON

ASSOCIATE PRODUCER
CHARLES FITZSIMMONS

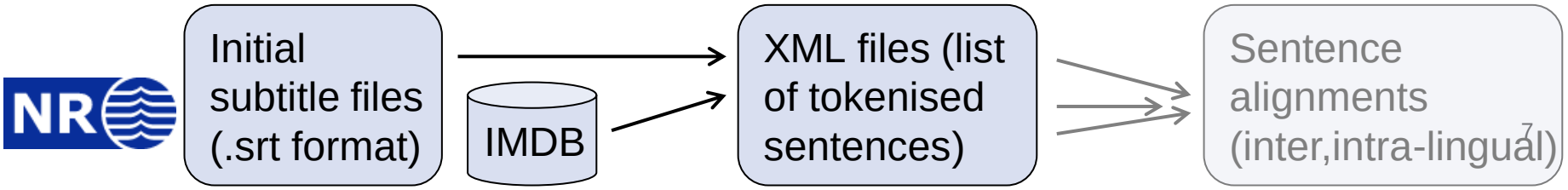
APRIL 6, 1966

GREENWAY PRODUCTIONS
FOR

Step 1:
Alignment with movie and TV scripts

OpenSubtitles 2016

- Earlier this year, Jörg Tiedemann and I released a new major version of the OpenSubtitles corpus
- What is it?
  - Collection of 2.8M subtitles from www.opensubtitles.org
  - 2.6 G sentences, 17.2 G tokens
  - 60 languages aligned at sentence-level (1689 bitexts)
- Preprocessing steps:
  1. Conversion,
  2. Sentence segmentation
  3. Tokenisation
  4. Correction of OCR and spelling errors
  5. Extraction of meta-data



Movie & TV scripts

- We crawled various websites hosting movie and TV scripts
  - Scrapped them to extract the sequence of dialogue turns
  - Result: total of **7,467** of dialogue transcripts
- NB: dialogues from screenplays can be very different from those found 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.

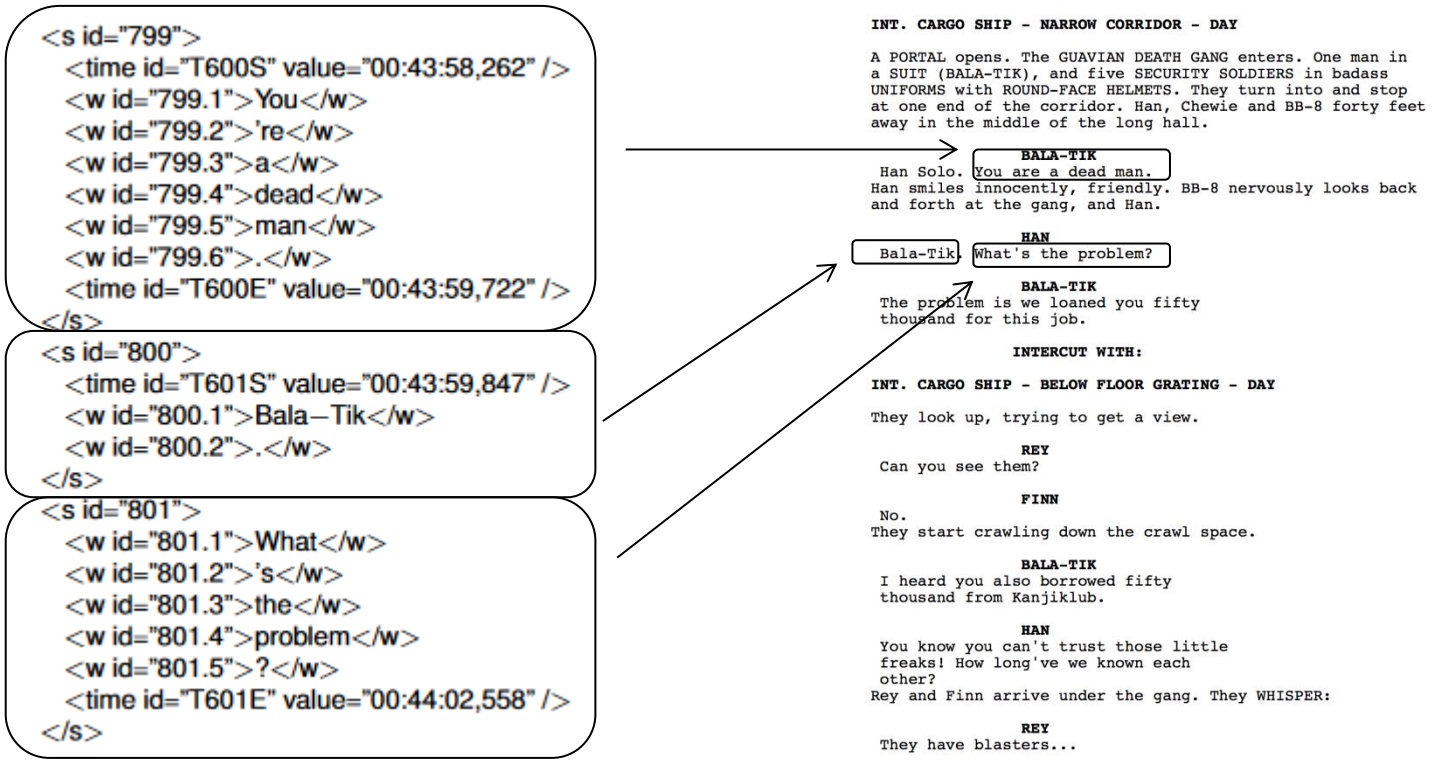
HAN
You know you can't trust those little freaks! How long've we known each other?

Rey and Finn arrive under the gang. They WHISPER:

REY
They have blasters...

Alignment

- We can then align the subtitles with the movie scripts
  - One alignment for each <subtitle,script> pair
  - We used both `hunalign` and `bleualign`



Alignment results

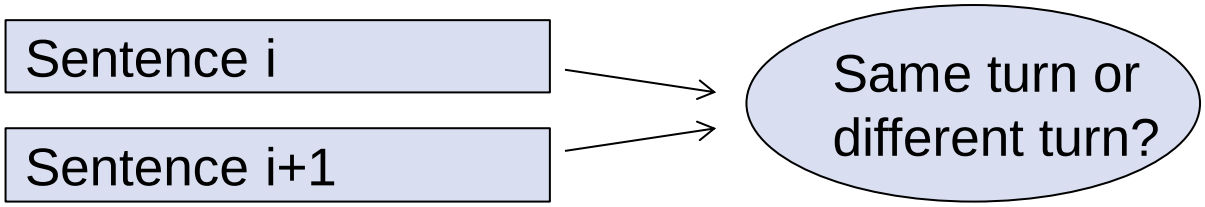
- 3,864,058 sentence pairs
  - 34% of the sentences for movies, 60% for TV episodes
- Quality of the alignments?
  - `hunalign` and `bleualign` were quite consistent (only 0.3% of conflicting alignments)
  - Comparison with a small, manually annotated corpus of TV series: 97.6% of the projected speaker labels matched the manually labelled ones
- We also projected the speaker information onto 6 other languages (using the bitexts from [Lison and Tiedemann, 2016])

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

Step 2:
Turn segmentation

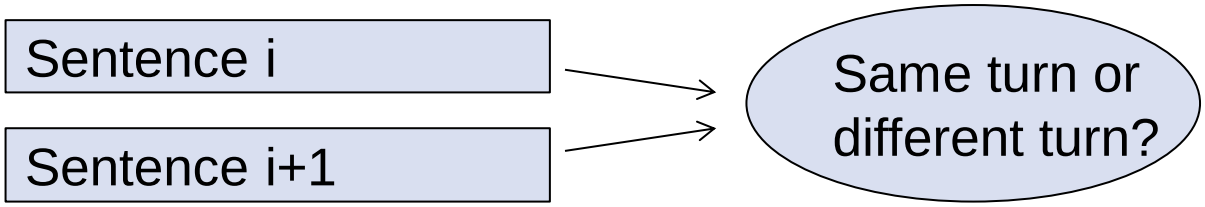
Taking stock

- **Where are we?**
  - Thanks to the alignments, we now have a subset of subtitles where a fraction of sentences are annotated with speaker information (speaker label + turn boundaries)
- **What do we want?**
  - A classifier that detects turn boundaries, using only textual and timing features from the subtitles themselves
- Binary classifier: given two consecutive sentences, predicts the *presence of a turn boundary* between them



Training data

- We extracted all consecutive sentence pairs in the subtitles that were annotated with speaker information
  - Total of 1,521,382 sentence pairs
  - Divided in training (60%), dev (20%) and test (20%) sets
- Binary scheme:
  - If the sentence i and i+1 have the same speaker and are part of the same turn in the script, mark it as "same turn"
  - Otherwise, mark it as "new turn"
- Quite balanced dataset: 52.3 % of "new turn" pairs



Classifier

- **Goal**: train a binary classifier that, given a pair of two consecutive sentences, outputs the probability of a turn boundary between them
- We used a linear discriminative classifier for this task
  - Using *Vowpal Wabbit*, a high-performance linear classifier
- Which features to use?
  - Various linguistic markers can be useful
  - For instance, adjacency pairs (such as question/answer) often denote a turn change
  - Another example: reuse of same pronoun as subject in the two sentences often denote a turn continuation

Features

Timing features:
Time gaps and sentence durations

Edit distance features
Token-level dist. between the two sentences

Length
Nb. of characters/tokens in each sentence

Adjacency features
Occurrence of specific patterns, such as
- *Likely polar answer*
- *Likely clarification request*
- *Pronoun inversion*

Lexical features:
BoW, bigrams, occurrence of negation/question words, pronouns

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

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

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

Punctuation features:
Marks at start/end of each sentence

"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.)

Extension 1: multilingual classifier

- We also have speaker annotations for non-English subtitles
  - Can we use them to further improve the classification?
  - Useful markers of turn change might be absent in a particular language but present in another one.
- We combine all classifiers in a weighted sum:

$$P_{multiling}(turn|s_{i-1}, s_i) = \alpha \left[P_L(turn|s_{i-1}, s_i) + \sum_{L'} w_{L'} P_{L'}(turn|s_{j-1}, s_j) \right]$$

Probability of turn boundary for English sentence pair

Probability of turn boundary for sentence pair in language L

Extension 2: speaker diarization

- When the corresponding audiovisual material is available, we can also exploit it to improve the segmentation
- More precisely, we can apply *speaker diarization* techniques to segment the audio stream into clusters
- Again, we integrate the result in a weighted sum:

$$P_{Classif+Dia}(turn = same | s_{i-1}, s_i) = \alpha \left[P(turn = same | s_{i-1}, s_i) + w_{Dia} \mathbb{1}(C(s_{i-1}) = C(s_i)) \right]$$

$$P_{Classif+Dia}(turn = new | s_{i-1}, s_i) = \alpha \left[P(turn = new | s_{i-1}, s_i) + w_{Dia} \mathbb{1}(C(s_{i-1}) \neq C(s_i)) \right]$$

Indicator function (1 if the s_{i-1} and s_i are part of the same diarization cluster, else 0)

Step 3:
Experimental results

Experiments

- **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)
- And 3 alternative approaches:
  - Basic discriminative classifier
  - Ensemble of multilingual classifiers (extension 1)
  - Classifier with speaker diarization (extension 2)

For the speaker diarization, we extracted the audio of one season (21 episodes of ~ 40 minutes each) of the “One Tree Hill” TV series, and applied the LIUM diarization toolkit on the data.

Results

		Dev				Test			
Approach	Turn	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	**0.794***	0.79	0.72	0.75	**0.781***
	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

		Tree Hill			
Approach	Turn	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	**0.775***
	New	0.72	0.87	0.79	

Accuracy, precision, recall and F1 scores on the small "Tree Hill" dataset.
The best result is statistical significant with p-value = 0.013

Results

	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**

Table 3. Compared accuracies for the baseline and classifier for 6 non-English languages (test set).

NB: Some features (e.g. adjacency features) were not present for these languages.

Discussion

- The 3 approaches outperform the baseline, but the results are far from perfect
  - Is this the result of a bad classification model
  - … or of the inherent difficulty of the task?
- Small-scale annotation experiment with *3 annotators*
  - The annotators were shown 100 sentence pairs, together with their associated start and end times.
  - Fleiss' kappa of 0.35 ("fair" agreement) among the three annotators and the "gold standard" from the script
  - Classification accuracy not better than the baseline (68%, 72%, 65% respectively)

↘

But they ignored the timing information, which is often crucial to detect turn boundaries

Step 4: Conclusion

Conclusion

- **Two contributions:**
  - An extension of the OpenSubtitles dataset with speaker information extracted from movie & TV scripts
  - A data-driven approach to the segmentation of subtitles into dialogue turns, based on linguistic and timing features
- Although the approach focused on subtitles, it can easily be adapted to other types of dialogue transcripts.
- Future work:
  - More advanced segmentation approach? Neural architectures, structured prediction, etc.
  - Use of the resulting turn structure for downstream tasks

NR