AUTOMATIC TURN SEGMENTATION FOR MOVIE & TV SUBTITLES

Pierre Lison

Norwegian Computing Centre
Oslo, Norway

Raveesh Meena

KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

Movie and TV subtitles contain large amounts of conversational material, but lack an explicit turn structure. This paper present a data-driven approach to the segmentation of subtitles into dialogue turns. Training data is first extracted by aligning subtitles with transcripts in order to obtain speaker labels. This data is then used to build a classifier whose task is to determine whether two consecutive sentences are part of the same dialogue turn. The approach relies on linguistic, visual and timing features extracted from the subtitles themselves and does not require access to the audiovisual material – although speaker diarization can be exploited when audio data is available. The approach also exploits alignments with related subtitles in other languages to further improve the classification performance. The classifier achieves an accuracy of 78 % on a held-out test set. A follow-up annotation experiment demonstrates that this task is also difficult for human annotators.

1. INTRODUCTION

Movie and TV subtitles constitute interesting resources for dialogue modelling. Although they transcribe scripted interactions, subtitles do cover a large variety of dialogue phenomena, including e.g. the widespread use of colloquial language, multiple speaker styles, and the presence of complex conversational structures. Furthermore, large corpora of subtitles are now available online: the latest release of the OpenSubtitles corpus [1] contains more than 2.8 million subtitles in over 60 languages, complete with detailed timing information and meta-data about the corresponding source material, typically a movie or TV episode. These resources can be useful in multiple areas of spoken language technology, such as statistical language modelling, conversational modelling [2], machine translation of dialogues [3, 4], and even spoken dialogue systems [5].

However, one important information is missing from these subtitles from a dialogue modelling perspective: the turn structure. Subtitles are of course meant to be displayed together with its associated video and not read in isolation. As a consequence, they do not provide any indication about who is speaking at a given time, since this information is superfluous for the viewers. Previous work has mostly focused on exploiting the audiovisual material to identify the speakers associated with each sentence of the subtitle [6, 7, 8, 9, 10]. However, this approach is difficult to scale, as it necessitates access to a large number of copyrighted material.

This paper presents a novel approach to turn segmentation based on a combination of linguistic, visual and timing features extracted from the subtitles themselves. The approach is decomposed in two processing steps. The subtitles are first aligned with a collection of movie and TV scripts to associate with speaker labels. This dataset is then used to train a classifier that, given a pair of two consecutive sentences, determines whether they are part of the same turn or not.

The next section describes the alignment of subtitles with movie & TV scripts. Section 3 describes the data, features and training regime of the classifier. Section 4 then presents the evaluation results, and Section 5 provides a short discussion of the approach.

2. ALIGNMENT PROCEDURE

2.1. Data

The source material for our approach is the collection of movie & TV subtitles from the OpenSubtitles2016 release [1]. The subtitles in this collection are segmented into sentences, which are themselves segmented into tokens (one per line). The sentences are also annotated

with start and times which are derived from the timestamps of the original subtitle. The result is encoded in a simple XML format, as illustrated in Listing 1.

```
<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>
```

Listing 1. Excerpt of three tokenised sentences from the OpenSubtitles2016 corpus. The last two sentences were extracted from the same subtitle block, shown between 00:43:59,847 and 00:44:02,558.

As previously noted, subtitles do not provide any information about who is speaking at a given time. But scripts (also called screenplays) do provide such details, through rich annotations including scene descriptions, filming instructions and dialogue transcripts. However, they typically blend the actual dialogue lines with scene-related instructions, as illustrated in Figure 1. In addition, movie scripts available online are often first drafts (written before any actual filming has taken place). As a consequence, their transcripts are often somewhat different from the dialogues in the final movie.

The first step was to crawl various websites hosting movie and TV scripts. Each script was parsed to extract its sequence of dialogue turns, based on simple layout heuristics (to distinguish e.g. speaker names from scene descriptions). This results in a set of 7,467 dialogue transcripts (one per script), among which 1,069 movies and 6,398 TV episodes. The movie scripts had an average of 756 turns (1383 sentences) per transcript, while the TV scripts had an average of 275 turns (482 sentences). The definition of "turn" employed in this paper is directly derived from the script: two consecutive sentences are assumed to be from the same turn if they are part of the same visual "block" in the transcript (for instance, "Bala-Tik." and "What's the problem?" in Figure 1 are part of the same turn). In the vast majority of cases, we found that this operational definition follows quite well the traditional linguistic criteria for turn boundaries [11].

```
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...
```

Fig. 1. Excerpt from a movie script.

2.2. Sentence alignment

The dialogue transcripts are aligned to their corresponding English subtitles from OpenSubtitles using standard sentence alignment techniques. Two state-of-the-art sentence aligners were employed: hunalign [12] and bleualign [13]. Based on the alignments generated by these tools, we can then *project* the speaker names from the dialogue transcripts onto their corresponding subtitles. A total of 5,413 English-language subtitles were automatically annotated in this manner, amounting to 3,864,058 sentences with speaker information. This corresponds to 34 % of the total number of sentences for the movies and 60 % for the TV episodes. This relatively low annotation ratio is mainly due to the structural

differences between the transcripts and the subtitles, as explained in the previous section. In addition, we relied on strict thresholds (i.e. favouring precision over recall) for the sentence aligners to ensure that the speaker labels projected onto the subtitles were of sufficiently high quality. The two sentence aligners were quite consistent, with only 0.3 % of conflicting alignments.

A small-scale evaluation was conducted to assess the quality of the projected speaker labels compared to gold standard annotations. We extracted 5 episodes from a manually annotated corpus of TV series (see [14]) and compared their labels to the ones derived from the alignments with the transcripts. 97.6 % of the projected speaker labels matched the manually annotated ones on this dataset, thus confirming the high accuracy of the alignment procedure.

2.3. Projection to other languages

Contrary to movie and TV transcripts (which are only available in the original language of the audiovisual material, usually English), subtitles are available in many languages. Furthermore, parallel corpora such as OpenSubtitles provide sentence alignments to all language pairs for which subtitles exist. This allows us to project the speaker annotations extracted using the procedure outlined above to other, non-English subtitles.

We performed this projection of speaker labels onto six other languages, namely Arabic, Chinese, Czech, French, German and Turkish. Table 1 details the number of speaker-annotated subtitles and individual sentences resulting from this projection.

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.

3. TURN SEGMENTATION

3.1. Training data

Based on this annotated dataset of subtitles, we train a classifier that determines the likelihood of two consecutive sentences being part of the same turn. We extracted from the subtitles all consecutive sentences pairs where both sentences were annotated with a speaker name (1,521,382 pairs), and divided this collection into training (60%), development (20%) and test (20%) sets.

The following two-class scheme was adopted: If the speaker names are identical for the two consecutive sentences *and* those sentences were part of the same dialogue turn in their corresponding transcript, the pair is marked as "same turn". Otherwise, the sentence pair is marked as "new turn". The resulting dataset is quite balanced, with 52.3 % of "new turn" pairs.

3.2. Features

Various linguistic markers can contribute to the detection of turn boundaries. For instance, adjacency pairs such as question/answer or statement/clarification request are indicative of a turn change. On the other hand, sentences that share an identical pronoun as subject often denote a continuation from the same speaker.

The following features (reflecting the semantic/pragmatic content of the two sentences as well as their relation with one another) are used by the classifier:

Timing features : Time gaps (continuous and discretised) between the end of the first sentence and start of the second; duration of each sentence; time gaps for the preceding and next sentence pairs.

Punctuation features : Occurrences of punctuation marks at the start and end of each sentence (most importantly, sentence-initial dashes).

Lexical features : Bag-of-words and bigram features for the two sentences ; first and final tokens, first and final bigrams ; occurrence of a negation word, a first or second-person pronoun, or a question word.

Part-of-speech features : First and final POS tag of each sentence (using the NLTK perceptron tagger with a pretrained model), together with a feature to capture imperative sentences (`VB` tag occurring before any `NN` or `PRP` tag, and not ending with a question mark).

Visual features : Binary features indicating whether the sentence is starting/completing a subtitle "block" in the original subtitle (a subtitle block may indeed contain more than one sentence). This feature is determined through the presence/absence of timestamps in the XML entity for the sentence (see Listing 1).

Length features : Number of characters and tokens in each of the two sentences.

Adjacency features : Occurrence of specific patterns between the two sentences, such as a likely polar answer (first sentence starting with an auxiliary followed by a yes/no response), clarification request (second sentence contained in the first and followed by a question mark) or a person inversion (sentence with second-person pronouns followed by a sentence with first-person pronouns, or vice versa).

Edit distance features : Features capturing the (token-level) edit distance between the two sentences. If the edit distance is equal to 1, we also add a lexical feature with the token being inserted/deleted/replaced.

Global features : Features capturing whether the sentences contains a likely character name (uppercase token repeated many times in the subtitle); the genre of the movie, its sentence/token density (total number of sentences/tokens divided by movie duration), and the position of the sentence in the subtitle.

Alignment features : Finally, we exploit the sentence alignments from [1], and add features capturing the proportion of inter- and intra-lingual alignments[1] where the two sentences are mapped into one single sentence in another subtitle. Alignments of type 2:1 are indeed much more likely to occur if the two sentences are from the same speaker.

The above list of features were extracted from the training set and fed into the Vowpal Wabbit machine learning library [15]. Vowpal Wabbit is an efficient online learner that can be applied to various regression and classification tasks. The classification relied on a discriminative linear model. To account for interacting features, we also added the cross-product of the most important features to the model. Stochastic gradient descent is used for the parameter optimisation. Section 4 details the empirical evaluation of this classifier.

[1]Interlingual alignments are sentence alignments between subtitles in different languages, whereas intra-lingual alignments connect alternative subtitles in the same language.

3.3. Multilingual classification

The alignments from Section 2 enabled the projection of speaker labels not only on English-language subtitles, but also on corresponding subtitles in other languages. As these subtitles are aligned with one another at the sentence level in the OpenSubtitles collection, we can exploit these alignments to further improve the segmentation performance. Indeed, useful (linguistic or non-linguistic) markers of turn change might be absent in a particular language but present in another one.

To this end, we trained separate classifiers for each language in Table 1. The results of all classifiers are then combined in a weighted sum. Formally, let $(s_{i-1}, s_i)_L$ denote a pair of consecutive sentences for language L, and let $P_L(turn|s_{i\text{-}1}, s_i)$ denote the posterior distribution for $turn = \{\textit{new}, \textit{same}\}$ according to the classifier for L. Now, if the pair $(s_{i\text{-}1}, s_i)$ is aligned with some sentence pairs $\{(s_{j\text{-}1}, s_j)_{L'}\}$ in some other languages L', we can define the multilingual classifier as:

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

where:

- $(s_{j\text{-}1}, s_j)$ denotes the sentence pair in language L' which is aligned with the $(s_{i\text{-}1}, s_i)$ pair.
- α is a normalisation factor to ensure the probabilities for the two values of $turn$ sum up to 1.
- $w_{L'}$ is the (tunable) weight of the classifier for L'.

3.4. Speaker diarization

The classifier outlined so far is designed to operate on the basis of the subtitles themselves, without requiring access to the audiovisual material. However, in case audiovisual data is available, the turn segmentation can be augmented in order to take advantage of speaker diarization [16]. To this end, we integrated the LIUM speaker diarization toolkit [17] into the processing pipeline of the classifier. The toolkit performs speech activity detection, BIC segmentation and hierarchical clustering. A Gaussian Mixture Model is then estimated for each resulting cluster via Expectation-Maximisation, after which the signal is re-segmented through Viterbi decoding. Finally, the final clustering is produced based on the

		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	

Table 2. Accuracy, precision, recall and F_1 scores based on the development set, test set, and on the small Tree Hill dataset. Diarization results are not available for the DEV and TEST sets due to their lack of audio data. The best results are written in bold and are all statistical significant using a bootstrap test with a confidence level $\alpha = 0.05$ (p-values are 0.013 for TREE HILL and < 0.0001 for DEV and TEST).

Cross-Likelihood Ratio (CLR) metric. The diarization approach employed for the evaluation (see next Section) remains relatively simple and can be further improved through the estimation of speaker models or the use of the video stream in addition to the audio signal [10].

The diarization results are also integrated into the segmentation using a weighted sum. Since the subtitles include precise start and end timestamps, we associate every sentence s in a given subtitle to its corresponding cluster label $C(s)$ in the diarization output. The diarization will thus indicate a turn change between $s_{i\text{-}1}$ and s_i when $C(s_{i\text{-}1}) \neq C(s_i)$. We define the posterior probability given a sentence pair $(s_{i\text{-}1}, s_i)$ as:

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

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

where $\mathbb{1}$ is the indicator function, α is again the normalisation factor and w_{Dia} the weight of the diarization.

4. EVALUATION

4.1. Experimental setup

The development and test sets are respectively composed of 197K and 200K sentence pairs from the English-language subtitles. To reduce the risk of incorrect annotations, only sentence pairs for which the two aligners (hunalign and bleualign) agreed on the speaker labels were included in the development and test sets. The weights for the multilingual classifiers and the diarization were manually set to a value of 0.5.

The baseline for the evaluation is defined through the following procedure:

1. If the second sentence starts with a "-" dash, classify the pair as "new turn".
2. Otherwise, if the two sentences are part of the same subtitle block, classify them as "same turn".
3. Otherwise, classify the pair as "new turn" (which is the majority class in this context).

To evaluate the use of speaker diarization, we also extracted the audio data of one season (21 episodes of about 40 minutes each) of the "One Tree Hill" TV series, and applied the LIUM diarization toolkit on this data.

4.2. Results

The accuracy, precision, recall and F_1 scores are given in Table 2. We can observe that the classifier improves the classification performance by reaching an accuracy of 78 % on the test set compared to 67 % for the baseline. Table 3 also compares the accuracies of the baseline to the classifier for 6 other languages.

	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).

The most informative features for this segmentation task are the alignment features, the time gaps between the two sentences, the occurrence of a starting dash, the final punctuation of the first sentence, and whether the two sentences were part of the same subtitle block. Some lexical features such as the occurrence of "well,..." or "you mean" are useful as well.

We can observe the somewhat disappointing results of the diarization. This seems to be partly caused to a poor synchronization between the subtitle and the audio data. Furthermore, speaker diarization of movies and TV series is known to be an inherently difficult task, notably due to the presence of many short speech segments with few pauses between them [10, 18].

5. DISCUSSION AND RELATED WORK

Although the evaluation results show a clear improvement relative to the baseline, they remain prone to various segmentation errors. But is this relatively low accuracy the result of a bad classification model, or of the inherent complexity of the task? In order to shed some light on this question, we performed a small-scale experiment with 3 human annotators. The annotators were asked to label sentence pairs as to whether the second sentence is from the same speaker as the first sentence, or from a different speaker. The annotators were shown both the two sentences as well as their associated start and end times. A sample of 100 sentence pairs were randomly selected from the training set. The Fleiss' kappa among the three annotators and the gold standard annotations was 0.35, which is typically interpreted as expressing only a fair agreement [19]. This suggests that given the provided context, the task is perceived as difficult for human annotators. This is further evidenced by the Fleisse's kappa of 0.33 among the three annotators. We also analysed the accuracy of the annotators against the gold standard. The accuracy scores for the three annotators were 68%, 72% and 65%, respectively. Surprisingly, these scores are lower than the baseline model performance. A closer look at the samples indicated that the human annotators made little use of the timing information, which is often an important element in the segmentation (in particular, the absence of a time gap means that the two sentences are part of the same visual block, which often indicates a continued turn).

To the best of our knowledge, this paper is the first to investigate the problem of segmenting movie and TV subtitles in dialogue turns without access to audiovisual data. The idea of aligning movie scripts with subtitles has, however, been investigated in previous work [7]. Other related works include the dialogue corpus extracted from movie scripts of [20], and the approach of [21] on using movie scripts to learn character models for generating expressive dialogue. [10] presents a recent approach on speaker diarization of TV series exploiting both audio and video data.

6. CONCLUSION

The contributions of this paper are twofold. We first presented a new multilingual dataset (freely available under an open-source license) of subtitles annotated with speaker information. These annotations are automatically extracted from alignments with movie and TV transcripts. Second, we reported on a data-driven approach to the segmentation of movie & TV subtitles into dialogue turns. The classifier relies on a wide spectrum of features (including cross-lingual alignments) extracted from the subtitles. The approach does not require access to the actual movie or TV episode, although speaker diarization can be employed when audiovisual data is available. The evaluation results highlight the importance of linguistic and timing features for the detection of turn boundaries. Although the present paper focused on subtitles, the approach can be adapted to other types of dialogue transcripts.

In future work, we will consider the use of more advanced speaker diarization methods, such as the estimation of a-priori speaker models or the simultaneous processing of video and audio streams [10]. We also wish to investigate whether the subtitles can be used to provide an initial segmentation to the diarization tool.

7. REFERENCES

[1] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[2] O. Vinyals and Q. V. Le, "A neural conversational model," *CoRR*, vol. abs/1506.05869, 2015.

[3] V. Petukhova, R. Agerri, M. Fishel, S. Penkale, A. del Pozo, M. S. Maucec, A. Way, P. Georgakopoulou, and M. Volk, "SUMAT: Data collection and parallel corpus compilation for machine translation of subtitles," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012.

[4] L. Wang, X. Zhang, Z. Tu, A. Way, and Q. Liu, "Automatic construction of discourse corpora for dialogue translation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, may 2016.

[5] D. Ameixa, L. Coheur, P. Fialho, and P. Quaresma, "Luke, i am your father: dealing with out-of-domain requests by using movies subtitles," in *International Conference on Intelligent Virtual Agents*. Springer, 2014, pp. 13–21.

[6] Y. Li, S. S. Narayanan, and C.-C.J. Kuo, "Adaptive speaker identification with audiovisual cues for movie content analysis," *Pattern Recognition Letters*, vol. 25, no. 7, pp. 777 – 791, 2004.

[7] R. Turetsky and N. Dimitrova, "Screenplay alignment for closed-system speaker identification and analysis of feature films," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*. IEEE, 2004, vol. 3, pp. 1659–1662.

[8] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar, "Movie/script: Alignment and parsing of video and text transcription," in *European Conference on Computer Vision*. Springer, 2008, pp. 158–171.

[9] A. Roy, C. Guinaudeau, H. Bredin, and C. Barras, "TVD: a reproducible and multiply aligned TV series dataset," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, 2014.

[10] X. Bost, G. Linarès, and S. Gueye, "Audiovisual speaker diarization of TV series," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4799–4803.

[11] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[12] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, "Parallel corpora for medium density languages," in *Recent Advances in Natural Language Processing (RANLP 2005)*, 2005, pp. 590–596.

[13] R. Sennrich and M. Volk, "Iterative, MT-based sentence alignment of parallel texts," in *NODALIDA 2011, Nordic Conference of Computational Linguistics*, 2011.

[14] X. Bost and G. Linarès, "TV Series Corpus," 7 2016, URL: http://figshare.com/articles/TV_Series_Corpus/3471839.

[15] A. Agarwal, O. Chapelle, M. Dudík, and J. Langford, "A reliable effective terascale linear learning system," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1111–1133, 2014.

[16] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb 2012.

[17] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proceedings of Interspeech*, Lyon (France), 25-29 Aug. 2013.

[18] I. Kapsouras, A. Tefas, N. Nikolaidis, G. Peeters, L. Benaroya, and I. Pitas, "Multimodal speaker clustering in full length movies," *Multimedia Tools and Applications*, pp. 1–20, 2016.

[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[20] R. E Banchs, "Movie-DiC: a movie dialogue corpus for research and development," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 203–207.

[21] M. Walker, G. Lin, and J. Sawyer, "An annotated corpus of film dialogue for learning and characterizing character style," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, May 2012, pp. 1373–1378.