FrNewsLink : a corpus linking TV Broadcast News Segments and Press Articles

Nathalie Camelin[1], Géraldine Damnati[2], Abdessalam Bouchekif[1,3], Anais Landeau[1], Delphine Charlet[2], Yannick Estève[1]

(1) LIUM, Maine University, France – {firstname.surname}@univ-lemans.fr

(2) Orange Labs, Lannion, France – {firstname.surname}@orange.com

(3) LSE, EPITA, France – {firstname.surname}@epita.fr

Abstract

In this article, we describe *FrNewsLink*, a corpus allowing to address several applicative tasks that we make publicly available. It gathers several resources from TV Broadcast News (TVBN) shows and press articles such as automatic transcription of TVBN shows, text extracted from on-line press articles, manual annotations for topic segmentation of TVBN shows and linking information between topic segments and press articles. The *FrNewsLink* corpus is based on 112 (TVBN) shows recorded during two periods in 2014 and 2015. Concomitantly, a set of 24,7k press articles has been gathered. Beyond topic segmentation, this corpus allows to study semantic similarity and multimedia News linking.

Keywords: content linking, semantic textual similarity, topic segmentation

1. Introduction

Several corpora are dedicated to evaluation in the domain of topic segmention, semantic similarity or resource linking (textual or multi-modal). An overview of such corpora is proposed in Section 2. contextualizing our proposed corpus. We describe in this article the corpus *FrNewsLink* that can be downloaded on the LIUM website[1] in its first version. During two different periods of time: 7 consecutive days in February 2014 and 2 consecutive days in January 2015, several TVBN shows from 8 different French channels have been recorded and web press articles appearing on the Google News homepage have been downloaded. As a consequence, *FrNewsLink* resources (whose elements are described in Section 3. and detailed statistics are given in Section 5.) deal with various fields such as politics, sports, cinema, *etc.* and some events that are very dynamic during a given day. Then, a manual annotation process (described in Section 4.) has been performed in order to obtain: 1- topic segmentation annotations and 2- linking annotations between topic segments and press articles. Therefore, this corpus is very useful for many tasks such as topic segmentation, topic titling, video linking, semantic similarity, events follow-up and topic modeling (grouping documents addressing similar topics). Section 6. presents several tasks that can be addressed with our corpus, along with evaluations of theses tasks performed on it.

2. Related work

2.1. Corpora for topic segmentation

The C99 corpus designed by (Choi, 2000) has been widely used for evaluating topic segmentation of written text. It is an artificial corpus composed of 7000 segments randomly selected from articles of the Brown corpus. These segments are grouped into 700 documents where each document is the concatenation of 10 segments. A segment is composed of the first n sentences of the original article.

The *ICSI* corpus (Shriberg et al., 2004) (Shriberg et al., 2000) contains 75 documents transcribed automatically from meeting records (approximately one hour each). For each conversation turn the speaker, start time, end time, and word content are marked. This corpus has been exploited in several works as (Eisenstein and Barzilay, 2008) and (Galley et al., 2003).

The *TDT* (Topic Detection and Tracking) corpus has become a standard for topic segmentation. This corpus contains English, Arabic and Chinese (Mandarin) documents. The corpus with its different versions (from TDT1 to TDT5) is used to evaluate many topic segmentation systems as (Rosenberg and Hirschberg, 2006) and (Xie et al., 2010). It is important to note that several works dedicated to topic segmentation use their own corpus (Malioutov and Barzilay, 2006), (Eisenstein and Barzilay, 2008). In (Malioutov and Barzilay, 2006), the authors have created their corpus from physics course recordings. In (Eisenstein and Barzilay, 2008), the authors have made available a medical book in which each section is considered as a new topic segment. In the domain of French TVBN topic segmentation (Guinaudeau, 2011) created a corpus of 57 news programs broadcasted in February and March 2007 on the French television channel France 2. It contained 1180 topic boundaries. As will be developed in section 5.1. the particularity of our corpus lies in the diversity of TVBN shows sources and formats.

2.2. Corpora for semantic textual similarity

Semantic textual similarity measures the meaning similarity of texts, beyond lexical similarity. It has been treated in several evaluation campaigns in SemEval in the recent years (Cer et al., 2017). Available corpora are mainly in English, even though some have been introduced in Spanish and Arabic. In the semantic textual similarity task of SemEval, similarity is measured between short sentences, according to a scale of 5 levels, ranging from "no relation at all" to "paraphrase". Another task of the SemEval campaign deals with semantic similarity in the context of Community Questions Answering (Nakov et al., 2017). In particular: one subtask addresses "question-question" similarity, where the questions are questions posted on an english-

[1] https://lium.univ-lemans.fr/frnewslink/

speaking forum (Qatar Living Forum) which deals with any aspect of daily life for foreign people in Qatar. In the corpus, for each original question, a set of 10 questions retrieved by a classical information retrieval system are to be ranked, according to their semantic similarity to the original question. Manual reference for semantic similarity is on a 3-level scale (Perfect Match, Relevant, Irrelevant).
To the best of our knowledge, there isn't any freely available corpus in French, annotated in semantic similarity.

2.3. Multi-modal linking

Multi-modal linking is a domain where semantic similarities are searched among multi-modal documents and/or across modalities. Video hyperlinking is a task in multimedia evaluation campains such as MediaEval (Eskevich et al., 2014) or TRECVid (see *e.g.* (Bois et al., 2017b)). The objective here is to be able to link an anchor (a piece of a BBC program which has been selected by experts as a segment of interest) to other segments defined as targets, that can be extracted from 2,700 hours of programs. This task, similarly to textual semantic similarity tasks, refers to homogeneous data: the objective is to link a fragment to another fragment from the same source. Some other works attempt to link heterogeneous sources but from an alignment perspective (*e.g.* books and movies (Zhu et al., 2015) or video lectures and scientific papers (Mougard et al., 2015)).
In the News domain there has been several studies about linking press articles with other information sources. (Aker et al., 2015) explore linking press articles and comments on the AT corpus (Das et al., 2014) which has been built from article of The Guardian. Linking press articles and Tweets have also been studied (Guo et al., 2013). Closer to our purpose is the work of (Bois et al., 2017a) who attempt to build graph representations for News browsing. The authors have collected over a 3 week period (May 20−Jun 8, 2015) a corpus of documents in French including press articles, videos (*e.g.* daily news from France 2, political news), and radio podcasts (*e.g.* news programs from France Inter, political interviews from RMC). This corpus is not distributed so far. The FrNewsLink corpus allows addressing several multi-modal linking tasks, with heterogeneous data from various sources and of various length.

3. The FrNewsLink corpus

The *FrNewsLink* corpus is composed of several resources extracted from TVBN shows and web press articles collected during the same period.

3.1. TV Broadcast News data

The *FrNewsLink* corpus contains 86 TVBN shows recorded during one week from the 10^{th} to the 16^{th} of February 2014 and 26 TVBN shows recorded on 26^{th} and 27^{th} January 2015. The first epoch of the corpus is referred to as TV_W07_14 (7^{th} week of 2014) and the second as TV_W05_15. Contrarily to many other TVBN corpora, the particularity of this corpus lies in the large variety of formats : 14 different shows from 8 different channels (Arte, D8, Euronews, France2, France3, M6, NT1, TF1) at various times of the day. Our original objective was to develop automatic structuration approaches that are robust to format variations. Hence the corpus contains both traditional shows (with a succession of reports introduced by one anchorman) and more modern setups (with two anchormen or no anchorman at all, with debates on stage, with short pieces of news, *etc.*). Section 5. provides detailed statistics on the TVBN part of the corpus.
Due to legal issues, we cannot distribute the videos corresponding to these shows, but we propose to make available the automatic transcription as well as the automatic speaker diarization results. The automatic transcription is performed with a variant of the LIUM ASR system described in (Tomashenko et al., 2016) in its chain-TDNN version with discriminative learning (Peddinti et al., 2015). Speaker diarization results are provided by the LIUM_SpkDiarization system (Meignier and Merlin, 2010). On similar corpora, such systems observed a performance between 9% and 13% of WER and between 11% and 14% of DER.

3.2. Press articles

During the same periods of time, we have extracted articles from various press websites, using the Google News homepage for selection. Google News articles are clustered and presented by news topic. For each topic cluster, one article is highlighted as the main article and other ones are shown as related. We have chosen to download the Google News homepage every hour. Each referenced article becomes an entry of our database which is composed of an id number, the article's title, the date and time of the record, the link to the referenced article and the id number of the main topic article. This extraction has been performed during seven consecutive days from the 10^{th} to the 16^{th} of February 2014, leading to the Press_W07_14 sub-corpus, and on the 26^{th} and 27^{th} of January 2015, leading to the Press_W05_15 sub-corpus).
Only the article content itself is relevant while the remaining (*e.g.* navigation menu, reader comments, pictures, *etc.*) is uninformative and has to be discarded. To that purpose, we have used the Boilerpipe (Kohlschütter et al., 2010) library that provides quite accurate algorithms to detect and remove all the surplus of a web page.

4. Annotation process

4.1. Reference annotation for topic segments

The task consists in placing boundaries between topic segments. The definition of a topic can be subject to various interpretations. In our work, we consider a topic as a precise information, which takes place at a given moment and place. For instance in the case of several consecutive reports on sport results, there will be as many topic segments as addressed sport disciplines. Within the context of the crisis in Greece, if several consecutive reports concern several consequences of the crisis, they will be split into consecutive topic segments. For greater clarity, and in order to facilitate the other annotation tasks, a small textual description has been inserted describing the content of each segment. Note that the first and last topic segments of a show have been annotated with a particular tag when they correspond to the titles presentation or the summary. Actually for topic

segmentation evaluation they are usually discarded. Moreover, in some BN, the anchor gives a small description in the middle of the show of the topics that will be dealt with later. These segments are also particularly tagged during the annotation phase. One human annotator has proceed to the whole annotation : segmentations, summaries and discarded segments. Time spent on this first annotation phase is about 8 months (about 1.1k hours).

4.2. Reference annotation for linking

Originally our objective was to define a titling strategy where a given topic segment from a TVBN show could be assigned a title from a set of candidate press article titles (Bouchekif et al., 2016). Hence the protocol for manually annotating the ground truth for this task was defined as follows. Given a reference topic segment, and the set of article titles collected during the same day, the annotation process consists in specifying for each candidate title if *(i)* a title is suitable for the segment, if *(ii)* a title could be suitable for the segment but appears to be too specific or too generic or if *(iii)* it does not reflect the segment. Indirectly, this three scale annotation allows to assess to which extent a press article and a TVBN topic segment can be linked. The 3-level scale is analogous to the scale used for the question-question textual similarity task described in section 2.2..

In order to facilitate the annotation process, we have defined a strategy to reduce the amount of potential titles to evaluate: only the titles of articles of the day having at least two content words in common with the news segment are proposed to the annotator. Hence, on average the annotator had to evaluate 127 titles by TVBN topic segment. If the annotation has been performed with a scale of three possible values, we consider in the rest of the article that a TVBN segment and a press article can be linked if the title satisfies condition *(i)* (the title is suitable for the segment).

The same human annotator as the one of the first annotation phase has linked press articles and segments according to the 3 values. Time spent on this second annotation phase is about less than 2 months (about 0.25k hours).

5. Corpus statistics

5.1. TV Broadcast News data

Table 1 describes our two sub-corpora in terms of number of topic boundaries, number of segments and average segment duration. In several studies we have observed that segment duration can have a significant impact on downstream modules performances. When analyzing the duration distribution, it appears that they can be easily divided into two sets, where the threshold between *short* and *long* segments is set to $30s$. Table 2 describes the two sub-corpora in terms of segment duration and type. Note that the longest segment in the TV_W07_14 is an exception and corresponds to a long report on a particular society subject inserted in the Sunday midday show.

Corpus	TV_W07_14	TV_W05_15
Duration	23.3h	9.9h
# boundaries	895	271
# segments	997	297
# *long* segments	761	227
# *short* segments	236	70
av. segment duration	105.1	120.5
(min, max)	(5.2, 1145.4)	(5.1, 655.5)

Table 1: TVBN corpus description

5.2. Press articles

The data collection process described in section 3.2. results into a database of 22141 entries after suppressing duplicates and erroneous exports (dead urls or irrelevant contents): 17324 articles for PRESS_W07_14 and 4821 for PRESS_W05_15. On average around 2460 press articles are available for each day. When reducing to the first article of each Goggle News cluster, this figure goes down to 590 candidates for linking on average each day. The full corpus can be used to train vectorial representations as in (Bouchekif et al., 2015) where we show that using semantic relations derived from word embeddings could help for the topic segmentation task. To this purpose table 3 provides figures in terms of number of words in the overall press articles sub-corpora.

5.3. Linking statistics

As a result of the linking annotation process, TVBN topic segments can be separated into two sets. MATCH (M) contains segments that can be linked to at least one press article. NOMATCH ($\overline{M}$) contains segments that have no corresponding article in the candidate list. Table 4 provides detailed description of the cross-modal linking annotation for the TV_W07_14 sub-corpus.

Interestingly, short TVBN topic segments are more likely to be linked with a press article. Actually they usually correspond to a brief recall of news such as sport results, international events, *etc.* that are also referred to in press articles. Longer topic segments on TVBN are more likely to correspond to a particular seasonal report or a general society question illustrated in a particular place, that is not necessarily relevant for a press article or not necessarily put forward by the Google News algorithms.

Beyond $segment \leftrightarrow article$ linking, other linking tasks can be derived. Namely it is possible to derive $segment \leftrightarrow segment$ linking by identifying segments that share similar articles in their list of linked articles. Conversely we can derive $article \leftrightarrow article$ linking by looking for articles that are linked to a same TVBN topic segment. In our TV_W07_14 sub-corpus, out the 658 segments of the MATCH set, 489 segments (74%) can be linked to at least one other segment and among them 275 can be linked to more than 5 other articles. Conversely for the same period, out of the 4129 candidate articles for linking, 1605 articles (39%) have been linked to at least one TVBN topic segment. Among them, 469 articles can be linked to more than five other articles. These figures indicate that our corpus would be well suited to study graph-linking of a set of news fragments.

6. A variety of addressable tasks

This corpus enables to develop and evaluate several tasks such as thematic segmentation and semantic-based content

show	average segment duration (min; max)		
	TV_W07_14	TV_W05_15	Type
Arte_LeJournal	123.6 (19.7 ; 341.4)	158.9 (14.0; 452.1)	T
D8_LeJournal	170.0 (39.9 ; 258.2)	-	T
Euronews_LeJournal	68.8 (15.5 ; 159.2)	67.6 (19.0 ; 314.4)	M
France2_7Heures	62.7 (9.5 ; 290.0)	59.8 (13.7 ; 152.4)	M
France2_8Heures	56.5 (9.6 ; 137.4)	61.6 (12.8 ; 151.6)	T
France2_13Heures	147.4 (8.7 ; 447.3)	190.3 (10.8 ; 560.7)	T
France2_20Heures	126.2 (9.9 ; 447.5)	176.5 (11.7 ; 655.6)	T
France3_12/13Heures	84.9 (11.4 ; 274.2)	118.4 (15.2; 320.9)	T
France3 _19/20Heures	101.2 (13.8 ; 258.2)	132.7 (18.0; 315.2)	T
M6_12h45	81.7 (8.8 ; 232.9)	105.9 (17.3 ; 300.0)	T
M6_19h45	92.0 (20.7 ; 373.7)	117.6 (18.8 ; 360.0)	T
TF1_13Heures	126.5 (5 .2 ; 1145.4)	113.1 (5.1 ; 265.2)	T
TF1_20Heures	121.8 (15.3 ; 451.7)	131.3 (14.1 ; 348.9)	T
NT1_LeJournal	54.6 (15.9 ; 107.0)	68.2 (21.0 ; 83.2)	M

Table 2: Description of TVBN shows in terms of segment duration and type (**T** for Traditional and **M** for Modern)

Corpus	PRESS_W07_14	PRESS_W05_15
# articles	17324	4821
# words	378.6K	109.8K

Table 3: Press articles corpus description

Corpus	MATCH (M)	NOMATCH ($\overline{M}$)
# segments	658	339
# *long* segments	467	294
# *short* segments	191	45

Table 4: TV_W07_14 sub-corpus description in terms of linking

linking. In this section we provide a few results on different tasks that can be considered as baseline for further developments.

6.1. Topic Segmentation

The *FrNewsLink* corpus has been extensively used to develop and evaluate topic segmentation (Bouchekif et al., 2015). The approach was inspired by TextTiling (Hearst, 1997), a sliding window based algorithm, with several major improvements, mainly in the representation of segments and in the similarity measure between 2 consecutive windows. The similarity measure is computed between vectors encoding not only the words in the window but also the speakers who uttered the words. What's more, the similarity score embeds a semantic relation matrix in order to take into account words which are semantically related. The semantic relation matrix is based on word similarity, where words are represented by their embeddings, estimated on the set of the news articles of the same day, through word2vec tools (Mikolov et al., 2013).

When evaluating topic segmentation as a boundary detection task, with a tolerance margin of 10s around the actual boundary, this system yields a F-measure of 76.3% decomposing into 73.6% recall and 79.1% precision on the TV_W07_14 sub-corpus. More recently (Bouchekif et al., 2017) introduced a new metric in order to evaluate topic segmentation as a segment retrieval task and showed that this model has a retrieval score of 66.3% in terms of number of segments and a retrieval score of 75.1% in terms of segment duration.

6.2. Content linking by textual semantic similarity

For each topic segment S from the speech-data, a list $Accept(S)$ (possibly empty) of news texts, which are semantically similar, is provided.

Let M be the set of speech segments which have non-empty $Accept(S)$ and $\bar{M}$ be the set of speech segments without any corresponding article.

Note that, if one restricts to the news text ranked in first position, and the semantic similarity score between this text and the topic segment is above a certain threshold, the title of this news text can be considered as an acceptable title for the topic segment. This was the paradigm of the topic titling task proposed in (Bouchekif et al., 2016)

6.2.1. Ranking by semantic textual similarity

Content linking can be seen as a ranking task: for a given topic segment, rank all the news texts of the same day, in order to retrieve the news texts that are semantically similar to the topic segment. The semantically similar news texts ($Accept(S)$) must be ranked before the other ones. Such ranking can be evaluated in the usual way in Information Retrieval with Mean Average Precision (MAP) or Mean Reciprocal Rank (MRR). We perform experiments on textual similarity scoring. Speech segments and news articles are represented as weighted bags of selected lemmas. The selected lemmas are names, adjectives and non-auxiliary verbs, and the weights are $Okapi - BM25$ weights (Jones et al., 2000). For speech segment, Okapi computation is performed considering each topic segment in a given TVBN show as a document, and the whole TVBN show as the collection. For news article, Okapi computation is

performed considering all the articles of the same day as the collection. Ranking of news articles is done, based on the cosine similarity score between weighted bags of selected lemmas of the given speech segment and the article. For each speech segment, the similarity score is computed with all the articles of the same day (on average 590 articles). We compute a MAP@10, *i.e.* a MAP evaluation restricted to the list of the 10 first ranked news articles. In this task, the focus is made on ranking the articles according to their similarity score, and not on the thresholding of this score. Thus, the MAP evaluation is restricted to the set of 658 segments in M (subset for which $Accept(S)$ is non-empty). The obtained MAP@10 in such configuration is 83.7%.

6.2.2. Titling

In (Bouchekif et al., 2016), it is proposed to assign to a speech segment the title of the news article of the same day, which is the most similar to the speech segment and whose similarity is above a given threshold. As all speech segments in the corpus don't have a corresponding article, 3 kinds of errors can occur. A substitution (Sub) occurs when a segment S of M is assigned a title that doesn't belong to $Accept(S)$, a false rejection (FR) occurs when a segment S of M is not assigned any title and a false alarm (FA) occurs when a segment S from $\bar{M}$ is assigned a title. The Titling Error Rate (TER) is defined as follows :

$$TER = \frac{\#Sub + \#FR + \#FA}{|M| + |\bar{M}|} \quad (1)$$

One can also adopt the evaluation framework proposed for answer triggering in (Yang et al., 2015): the goal is to detect whether a speech segment has a corresponding article in the news corpus, and return the best corresponding article if there exists such one. Thus, the task can be evaluated as the detection of correctly titled speech-segments, with conventional metrics for detection evaluation: precision, recall and f-measure.

In (Bouchekif et al., 2016), on the set of 997 speech segments of TV_W07_14, the best results were obtained when computing a cosine similarity score between weighted bags of selected lemmas of the speech segment and of the news article (as explained in section 6.2.1.). The optimal threshold on this score gives, for reference topic segments, a TER of 11.8%, with $Sub = 3.9\%$, $FA = 3.8\%$ and $FR = 1.1\%$, which translates, for the detection of correctly titled segments, in $Re = 87.8\%$, and $Pr = 88.2\%$. Thus, for manually segmented topic segments, the results, with a simple textual similarity performs pretty well, but there is still room for improvement. (Bouchekif et al., 2016) extends the evaluation to the case of automatically segmented segments, with a global segmentation and titling error rate, where the titling error is computed only if the segmentation reaches a sufficient level of quality.

7. Conclusion

This article has presented a new resource that enables to study topic segmentation and semantic similarity through linking tasks (including cross-modality linking). Along several consecutive days, it gathers TVBN shows and press articles. Topic segments from the TVBN shows are linked with press articles of the same day whenever associations are possible. Further annotations could be added to precisely specify the links between segments and articles. Furthermore linking was restricted to articles and segments produced during the same day but studying the evolution of a topic across days could also be annotated and studied on the basis of the *FrNewsLink* corpus.

8. Bibliographical References

Aker, A., Kurtic, E., Hepple, M., Gaizauskas, R., and Di Fabbrizio, G. (2015). Comment-to-article linking in the online news domain. In *Proceedings of the SIGDIAL 2015 Conference*, pages 245–249. ACL.

Bois, R., Gravier, G., Jamet, É., Morin, E., Robert, M., and Sébillot, P. (2017a). Linking multimedia content for efficient news browsing. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, pages 301–307.

Bois, R., Vukotić, V., Simon, A.-R., Sicre, R., Raymond, C., Sébillot, P., and Gravier, G., (2017b). *Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity*, pages 185–197. Springer International Publishing, Cham.

Bouchekif, A., Damnati, G., Estève, Y., Charlet, D., and Camelin, N. (2015). Diachronic semantic cohesion for topic segmentation of TV broadcast news. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2932–2936.

Bouchekif, A., Damnati, G., Charlet, D., Camelin, N., and Estève, Y. (2016). Title assignment for automatic topic segments in TV broadcast news. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 6100–6104.

Bouchekif, A., Charlet, D., Damnati, G., Camelin, N., and Estève, Y. (2017). Evaluating automatic topic segmentation as a segment retrieval task. In *Interspeech 2017*, pages 2924–2928.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 26–33.

Das, M. K., Bansal, T., and Bhattacharyya, C. (2014). Going beyond corr-lda for detecting specific comments on news & blogs. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 483–492. ACM.

Eisenstein, J. and Barzilay, R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, October.

Eskevich, M., Aly, R., Racca, D., Ordelman, R., Chen, S.,

and Jones, G. J. (2014). The search and hyperlinking task at mediaeval 2014.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 562–569.

Guinaudeau, C. (2011). *Structuration automatique de flux télévisuels*. Ph.D. thesis, Institut National des Sciences Appliquées de Rennes, France, December.

Guo, W., Li, H., Ji, H., and Diab, M. T. (2013). Linking tweets to news: A framework to enrich short text data in social media. In *ACL (1)*, pages 239–249.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information processing & management*, 36(6):779–808.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, USA*, pages 441–450.

Malioutov, I. and Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Sydney, Australia, July.

Meignier, S. and Merlin, T. (2010). Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Mougard, H., Riou, M., de la Higuera, C., Quiniou, S., and Aubert, O. (2015). The paper or the video: Why choose? In *Proceedings of the 24th International Conference on World Wide Web*, pages 1019–1022. ACM.

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K. (2017). Semeval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH*, pages 3214–3218.

Rosenberg, A. and Hirschberg, J. (2006). Story segmentation of broadcast news in english, mandarin and arabic. In *Proceedings of the Human Language Technology Conference of the NAAC*, pages 125–128.

Shriberg, E., Stolcke, A., Hakkani-Tür, D. Z., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The icsi meeting recorder dialog act $mrda$ corpus. Technical report.

Tomashenko, N., Vythelingum, K., Rousseau, A., and Estève, Y. (2016). Lium asr systems for the 2016 multi-genre broadcast arabic challenge. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 285–291. IEEE.

Xie, L., Yang, Y., Liu, Z.-Q., Feng, W., and Liu, Z. (2010). Integrating acoustic and lexical features in topic segmentation of chinese broadcast news using maximum entropy approach. In *Audio Language and Image Processing (ICALIP)*, pages 407–413, Nov.

Yang, Y., Yih, W.-t., and Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In *EMNLP*, pages 2013–2018.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.