Scene Linking Annotation and Automatic Scene Characterization in TV Series

Aman Berhe[1] aman.berhe@limsi.fr
Camille Guinaudeau[1] camille.guinaudeau@limsi.fr
Claude Barras[2] barras@vocapia.com

Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France[1]
Vocapia Research, 91400, Orsay, France[2]

Abstract

In the context of a large collection of multimedia documents, creating links between documents or scenes can help to organize the collection. For TV series, this organization can be achieved by means of narrative structure extraction through scene linking. Narrative characteristics such as speaking characters, entity mentions and theme can be used to characterize scenes. The linking of scenes can be between scenes inside an episode, between scenes in different episodes and/or in different seasons, since stories in TV series progress at different level of granularity. In this work, we have annotated the links between the scenes of the first two seasons of the TV series Game of Thrones, using predefined stories and sub stories. We have also automatically extracted the narrative characteristics of each scene. The dataset is composed by 444 scenes, involving 154 speaking character organized in 46 stories divided into 151 sub stories and 5 sub sub-stories.

1 Introduction

Organization of large multimedia collections can be done by means of scenes/documents linking. For example [$BVS^{+}17$] used cross modal approaches to link target videos to an anchor in a collection of archived multimedia documents and were able to extract common patterns between target video and the anchor. In the context of TV series, organizing a large and complicated TV series, Game of thrones for example, as a collection of scenes is a way of understanding its narratives.

The creation and diffusion of manual annotations for scene linking, even though difficult and time consuming, is necessary to evaluate automatic techniques and allow reproducible research. Some few works like [GLV14, Pro10, EF19] have been done on narrative structure annotations in folk tales, for example Propp's folktales [Pro10] and french folktales [GLV14]. However, to our knowledge, there is no publicly available annotated dataset of scene linking for narrative structure extraction of TV series or more generally for multimedia collections.

In this paper, we present a new dataset composed by the annotation for the two first seasons of Game of Thrones for scene linking with respect to narrative structure and the annotation of most reportable scene

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2Story'20 Workshop, Lisbon, Portugal, 14-April-2020, published at http://ceur-ws.org

(MRS), i.e. scenes that bring a drastic change on the lives of characters and a story change. They are key scenes to make connection of other scenes. The dataset (annotation + automatic characterization of scenes) is made publicly available in order to allow reproducible research.

Our paper is organized as follows. First, we present the background studies and data for narrative TV series and books in Section 2. Then, Section 3 describes scene linking annotation and its importance for narrative structure extraction. In Section 4, we introduce scene characterization with respect to narrative elements. Then, the data description is discussed in Section 5 and we conclude our paper in Section 6.

2 Background

Since 1970s, narratives and story telling has been investigated in validating scientific methods in the field of artificial intelligence for understanding and evaluating human cognition theories [Var17, AS90]. The field of computational narrative links the daily human activities (narratives) and the computing world (machines computations) by analyzing and modelling narratives, narrative understanding and machine readable representation of narratives with a purpose of enabling computers to tell a story.

Narrative elements annotation of a narrative document is a very time consuming task. [LCW^+17, EF19] have worked on the annotation of narrative elements of short stories in two different ways; [LCW^+17] produced a guide line for directly annotating the narrative structure based on Freytag's [Fre90] pyramid, and [EF19] provided a guideline for narrative characteristics annotation to collect human judgements on narrative characteristics. Garcia-Fernandez et al. [GLV14] proposed digitization and annotation of a tales corpus from narrative point of view (the only French tales corpus available) and classified it according to the Aarne & Thompson narrative classification.

Multimedia hyperlinking – a way to navigate information in videos by jumping from one video to another – has been studied by many people [BGJ^+17, BVS^+17, BDG18, CSR18] using different techniques. [BGP^+15, OAB15, KM16, AFM^+16, BGJ^+17] have designed some linking categories or typologies for multimedia hyperlinking and build graph for easily exploring news or links. Kim et al. [KM16] used narrative theory as a framework to identify the links between social media content. [OAB15] presented a video hyperlinking based on named entity identification.

In TV series, scenes can be linked using the concept of multimedia hyperlinking [KM16, ESB12, BGJ^+17, CSR18] which can be used to tie different videos together and recreate one whole narrative.

3 Scene linking annotation

A link, as in scene linking, is the relevance between two or more scenes according to the story and the narrative elements they share. Linking scenes from the same episode or from different episodes, and continuing this chain of links until the last episode of a TV series, can capture the narrative structure of the whole TV series.

Before, we start scene linking annotation, stories and sub stories of the TV series are defined based on the main characters' stories and the story of the overall Game of Thrones TV series. A scene can start by a story or a sub-story, for example a scene of Jon Snow's (character in Game of Thrones) story starts with "Jon Snow going to the wall" which is a sub-story of "Jon Snow as Lord Commander". A scene is assigned to one or more stories and can have many sub stories inside the assigned stories depending on the theme of the story that the scene focuses on.

Our scene linking annotation is performed in two steps. First, current scene is linked to the most related scenes that come before it. One scene might be linked to more than 1 scene. For example, a scene (S3) may starts with an event or story that is linked to a scene (S2) and it may also focus on an event or story that is in another scene (S1). Therefore S3 is linked to both S2 and S1 but there may not be a link between S2 and S1. Second, a story title is given for each scene, based on the predefined stories and sub-stories. As for the stories, manually annotated stories can have up to three levels of granularity. The given stories are used as linking category. A linking category is a title given to link two or more scenes according to a story assigned. When we

think of automatic assignment of story title to a scene, a linking category may be seen as a cluster name that group related scenes together.

Furthermore, in each story, we have also annotated the most reportable scenes (MRS), i.e. the scenes that bring a major change to a story or a situation inside a scene.

4 Semi-automatic scene characterization

The process of scene linking annotation is time consuming and ambiguous in a sense of assigning stories to a scene. Therefore, an automatic tool that can characterize a scene to make the links between them is vital. Annotating a scene with narrative characteristics helps to create a link between scenes with respect to narrative structure specially in case of complicated TV series. Thus we have designed a method to automatically characterize a scene. The whole process is discussed as follows.

Before scene characterization, data of each episode need to be prepared. Therefore, video episodes are extracted from DVDs of the TV series with their respected audio and subtitles in many different languages (e.g. French, English, German, Czech and Spanish, Polish and Hungarian). Manual transcripts of episodes are scraped from different websites and fan pages of TV series with the speaking character names for each line of the transcripts. Speaking character names are normalized to the characters' list found in IMDB[1] by adding "_" instead of space and change all letters to lower case. Finally, forced-alignment of transcripts is performed using LIMSI text-to-audio alignment tool [GLA02] with the audio files extracted automatically. At this step we have the timing of each word in an episode.

Shot segmentation is performed based on shot boundary detection (SBD) algorithm implemented in the open-source Pyannote-Video toolkit [Bre15], which is based on displaced frame differences (DFF) and uses landmark features of the frames. Scene segmentation is performed based on grouping of adjacent shots and relying on a combination of multimodal neural features using our previous work [BBG19]. In this work, we define a scene as a set of contiguous shots which are connected by a central concept or theme. The method detects shot boundaries and compute visual features of each shot using VGG16 pretrained model provided by [SZ14] to extract deep visual features for each frame and embed each speech (text) inside a shot boundary using word2vec model built on subtitles and books of Game of Thrones and the temporal information of each shot is augmented. After the features are computed a shot clustering is performed and each shot is grouped into a cluster and then a sequence grouping algorithm is used to regroup adjacent shots together to form a scene segment.

Finally, each scene is represented by its narrative characteristics or elements. The narrative elements are characters (speaking characters and characters mentioned in the conversation inside a scene), entity mentions (locations and organizations) and the theme of the scene. Figure 1 depicts the characterization of a scene and the way it is computed is described in Section 4.1.

4.1 Automatic scene characterization

In order to automatically represent the semantic content of a collection of short documents (in this case, scenes), vector representation of words and documents (word2vec and doc2vec respectively), term frequencyinverse document frequency (TF-IDF) and latent Dirichlet allocation (LDA) are among the most famous and effective methods. Considering the narrative elements, the automatic scene characterization is done as follows:

First, we extracted the speaking characters (from manually annotated transcripts) and the entities that are involved in a scene. Transcripts with their speaking character names are scraped and the names are normalized to our standard character naming, since the transcripts come from different websites that can use different naming of characters. E.g. a character nick named "little finger" is normalized to "petyr_baelish". There are 154 speaking characters in the first two season of Game of Thrones.

Since any situation or events evolve around characters or entities, identifying entities will serve as a connection-link between scenes that have the same events or situations based on the common mentioned entities.

[1] https://www.imdb.com/list/ls068919538/

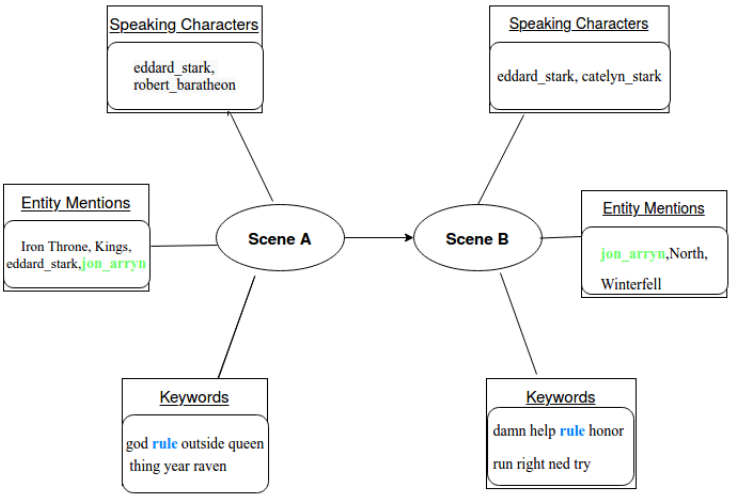

Figure 1: Scene Characterization

State of the art neural based named entity recognition (NER) called Flair [ABV18] is used to extract name mentions in the dialogues of each scene. The Flair NER technique performs better than Stanford CoreNLP in detecting entities with longer names, for example titles like 'Robert of the House Baratheon' are detected by Flair but not by CoreNLP.

Then, the main keywords that can represent the theme inside a scene are extracted. TF-IDF is used to extract the 10 most representative keywords from scene text, since it is the simplest and efficient method for extracting keywords and capture the importance of a word from a short text. The first five seasons of Game of Thrones is used as the documents collection, using scene as documents. The document collection is then composed by 1018 scenes and 92970 words. If a scene has less than 10 words then all the words are assumed as the keywords of a scene.

Nevertheless, TF-IDF keyword extraction method may not captures words which are synonyms, therefore we added a topic model to assign a topic, which may capture the main theme, for each scene. To this end, the topic of each scene is extracted so that we can relate the topic of a scene with another. We used Latent Dirichlet Algorithm (LDA), to assign each scene to one topic (the one associated with the highest membership or probability value) by using all the TV series as the corpus and scenes as documents.

Finally, a document to vector (Doc2Vec) embedding is used to represent the scenes' transcript. Each scene is treated as a document and is represented by a vector using Doc2Vec [LM14]. The embeddings of each scene have a vectors size of 100 values. This Doc2Vec representation of scenes' transcript can then be used to compute the semantic similarity of scenes, considering that scenes that talk about the same stories have a high content similarity.

5 Description of the annotation

Our data focused on the Game of Throne TV series which is one of the most complicated and popular TV series. It is complex due to the number of stories, and the number of characters and their intertwined stories.

We have annotated[2] the first 2 seasons of Game of Thrones by assigning stories from predefined stories for the purpose of linking scenes, though almost all stories try to converge into one story. The annotation is performed by one annotator and it took around one hour and thirty minutes per episode. The annotated dataset have 444 scenes[3] with 46 main stories and 151 sub-stories. The largest story is composed of 76 scenes and the smallest only of 1 scene. The main stories have a maximum of 11 sub stories and 1 minimum sub-story. The sub-stories

[2]https://github.com/aman-berhe/Game-of-Thrones-Dataset
[3]Scenes that do not contain speech are ignored in the characterization step (87 scenes).

increase as we continue to the seasons of Game of Thrones and some new stories are also created. Figure 2 illustrates the average number of stories (resp. sub-stories) per scene, while Figure 3 shows the average number of speaking characters per scene. Most of the scenes contain only 2 speaking characters. The maximum number of speaking characters in a scene is 8 and the minimum is 1[4].

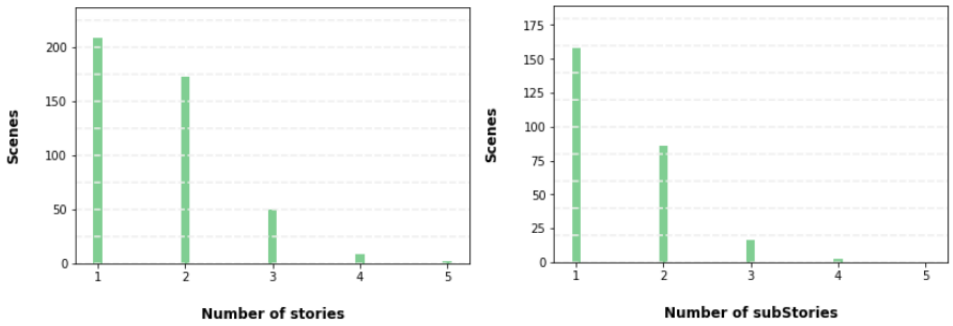

Figure 2: The average number of stories and sub stories per scene

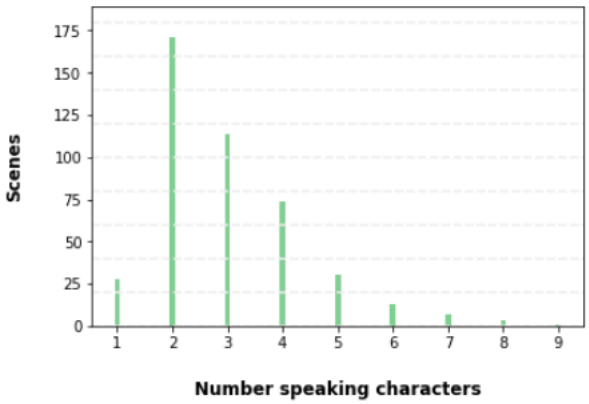

Figure 3: The average number of speaking characters per scene

For the two seasons of Game of Thrones, the dataset has 444 scenes with an average scene length of 123.23 seconds, a total of 92970 words and 70 most reportable scenes (MRS). Story wise, the dataset has 197 predefined stories (main stories and sub stories). In average, there are 3 MRS per story.

6 Conclusion

Scene linking as a way to construct the narrative structure is a good way of representing a very complicated and long connected stories, in TV series or other similar documents. The complication in TV series is that all scene stories feed to the main story of the TV series.

As the manual annotation is quite difficult and time consuming, a more complicated automatic annotation can be achieved for better scene characterization and linking besides what we have proposed.

Clustering techniques can be used on the dataset for creating links between scenes. Thus, soft or fuzzy clustering can group an element of the document in more than one cluster. And in case of scenes, fuzzy clustering can group a scene in multiple clusters which in turn can capture the membership of a scene to multiple stories. Additionally, most repeatable scene (MRS) can be used as a connection center for scene that come before and after the MRS.

[4]Unknown characters, for example "soldier#1" are removed, as they have no value for scene linking.

In the near future we plan to work on automatic MRS detection and fuzzy clustering of scenes. Different ways of document to vector representations can also be used to represent the scenes' text, rather than just Doc2Vec model.

Acknowledgements

This research was partly funded by the French National Research Agency (ANR) through the PLUMCOT (ANR-16-CE92-0025) project, and the Digiteo StoryArcs project.

References

[Luc68] Lucas, Donald W. Aristotle Poetics. *University of Michigan Press*, 1968.

[TW69] Todorov, Tzvetan and Weinstein, Arnold. Structural Analysis of Narrative. *Novel: a Forum on Fiction*, 70–76, 1969.

[Fre90] Freytag, Gustav. Die Technik des Dramas. *S. Hirzel*, 1890.

[Pro10] Propp, Vladimir. Morphology of the Folktale. *University of Texas Press*, 2010.

[Lea70] Leach, Edmund. Claude Levi Struass (Modern Master Series edited by Frank Kermode). *University of Chicago Press*, 1970.

[KM16] Kim, Joy and Monroy-Hernandez, Andres. Storia: Summarizing SMContent Based on Narrative Theory Using Crowdsourcing. *ACM on Computer-Supported Cooperative Work & Social Computing*, 1018–1027, 2016.

[EBS^+11] Ercolessi, Philippe and Bredin, Hervé and Sénac, Christine and Joly, Philippe. Segmenting TV Series into Scenes Using Speaker Diarization. *Workshop on Image Analysis for Multimedia Interactive Services*, 13–15, 2011.

[BGJ^+17] Bois, Rémi and Gravier, Guillaume and Jamet, Eric and Morin, Emmanuel and Robert, Maxime and Sébillot, Pascale. Linking Multimedia Content for Efficient News Browsing. *ACM on Multimedia Retrieval*, 301–307, 2017.

[CSR18] Chaturvedi, Snigdha and Srivastava, Shashank and Roth, Dan. Where Have I Heard This Story Before? Identifying Narrative Similarity in Movie Remakes. *NAACL on Human Language Technologies*, 673–678, 2018.

[Var17] Vargas, Josep Valls. Narrative Information Extraction with Non-Linear Natural Language Processing Pipelines. *Drexel University*, 2017.

[AS90] Andersen, Sandy and Slator, Brian M. Requiem for a theory: the story grammarstory. *Journal of Experimental & Theoretical Artificial Intelligence*, 253–275, 1990.

[BBG19] Berhe, Aman and Barras, Claude and Guinaudeau, Camille. Video Scene Segmentation of TV Series Using Multimodal Neural Features. *Series - International Journal of TV Serial Narratives*, 59–68, 2019.

[ABV18] Akbik, Alan and Blythe, Duncan and Vollgraf, Roland. Contextual String Embeddings for Sequence Labeling. *COLING on Computational Linguistics*, 1638–1649, 2018.

[ESB12] Ercolessi, Philippe and Sénac, Christine and Bredin, Hervé. Toward Plot De-interlacing in TV Series Using Scenes Clustering. *Content-Based Multimedia Indexing*, 1–6, 2012.

[LCW^+17] Li, Boyang and Cardier, Beth and Wang, Tong and Metze, Florian. Annotating High-Level Structures of Short Stories and Personal Anecdotes. *LREC*, 2017.

[EF19] Eisenberg, Joshua and Finlayson, Mark. Annotation Guideline No. 1: Cover Sheet for Narrative Boundaries Annotation Guide. *Journal of Cultural Analytics*, 11199, 2019.

[Bos16] Bost, Xavier. A Storytelling Machine?: Automatic Video Summarization: the Case of TV Series. *Université d'Avignon*, 2016.

[GLV14] Garcia-Fernandez, Anne and Ligozat, Anne-Laure and Vilnat, Anne. Construction and Annotation of a French Folkstale Corpus. *LREC*, 2430–2435, 2014.

[Ash87] Ashliman, Dee L. A guide to folktales in the English language: Based on the Aarne-Thompson Classification System. *Greenwood Press New York*, 1987.

[BVS^+17] Bois, Rémi and Vukotić, Vedran and Simon, Anca-Roxana and Sicre, Ronan and Raymond, Christian and Sébillot, Pascale and Gravier, Guillaume. Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity. *Multimedia Modeling*, 185–197, 2017.

[Bre15] Bredin, Hervé. pyannote Video: a toolkit for shot detection, shot threading and face trucking. *https://github.com/pyannote/pyannote-video*, 2015.

[SZ14] Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2014.

[BDG18] Budnik, Mateusz and Demirdelen, Mikail and Gravier, Guillaume. A Study on Multimodal Video Hyperlinking with Visual Aggregation. *IEEE on Multimedia and Expo (ICME)*, 1–6, 2018.

[BGP^+15] Bois, Rémi and Gravier, Guillaume and Pascale, Sébillot and Emmanuel, Morin. Vers Une Typologie de Liens Entre Contenus Journalistiques. *TALN*, 525–521, 2015.

[OAB15] Ordelman, Roeland and Aly, Robin and Benoit, Huet. Convenient Discovery of Archived Video Using Audiovisual Hyperlinking. *Workshop on Speech, Language & Audio in Multimedia*, 23–26, 2015.

[AFM^+16] Awad, George and Fiscus, Jonathan and Martial, Michelet, and Alan F., Smeaton. Trecvid 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. *TREC Video Retrieval Evaluation (TRECVID)*, 2016.

[LM14] Le, Quoc and Mikolov, Tomas. Distributed representations of sentences and documents. *International conference on machine learning*, 1188–1196, 2014.

[GLA02] Gauvain, Jean-Luc and Lamel, Lori and Adda, Gilles. The LIMSI broadcast news transcription system. *Speech communication*, 89–108, 2002.