Extracting Keywords from Movie Subtitles

Matúš KOŠÚT*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova2, 842 16 Bratislava, Slovakia
matuskosut@gmail.com

In our work we aim at keyword extraction from movie subtitles. Keywords and key phrases although missing the context can be found very helpful in finding, understanding, organising and recommending the content. Generally they are used by search engines to help find the relevant information. With the rising amount of information available on the Web, keywords are becoming more and more important, though it is even harder now to determine keywords for all content by person, so we target on automatic keyword retrieval.

Movies and video content are becoming massively available and widespread. The ability to automatically describe and classify videos has a vast domain of application, which seems to be more efficient compared with video and audio analysis. The main goal of our work is to design a method able to use of specifics of subtitles. First part of method focuses on the pre-processing. Pre-processing tries to process timings, information for hearing impaired persons (closed captioning) and tags included in subtitles. Second part divides subtitles into conversations according to the speed of speech (words per minute) and the gaps detected between the conversations. Scored conversations are used for keyword extraction.

The first idea to enrich the keywords extraction with metadata included in subtitles is to recognise individual parts of movie. We call them conversations. We suppose that we could divide dialogues in subtitles into conversations and get the scenes severally, approaching the natural distribution of scenes as it is seen by viewers. In the sense of when there are different characters talking in dialogues or scene is changed we would start a new conversation.

We want to explore possibilities for rating these conversations according to relevance. The higher the relevance is, the higher the ratings of keywords extracted from the conversations become. These conversations could also possibly help us with joining subtitles created by different authors, assuming different authors use different sentences and on condition that it is not a transcript.

---

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 73-74

To split the subtitles into conversations, we use timings included in subtitles. Using timings we want to detect the gaps between individual subtitles, supposing that if there is a gap bigger than are the gaps in the surrounding titles the conversation has changed into next one. We also experiment with the speech rate (words per minute) in individual titles and conversations as way to differ the conversations and to rate them supposing there is a relation between the speed and the importance of conversation.

Subtitles for hearing impaired viewers are a special category of subtitles created to help disabled people understand what is happening in the scenes. Based on our experience with various types of subtitles we concluded that subtitles for hearing impaired viewers contain descriptions of the most necessary sounds from scenes and backgrounds. We propose to use these metadata to substitute audio analysis. We assume precision increase of the results, because of a better representation of scenes and connecting the sounds with surrounding keywords is a sign of their importance [1].

We also experiment on speech rate (words per minute) and subtitles rate (titles showed per minute) as a helper tools for a stress, action or aggression detection in scenes.

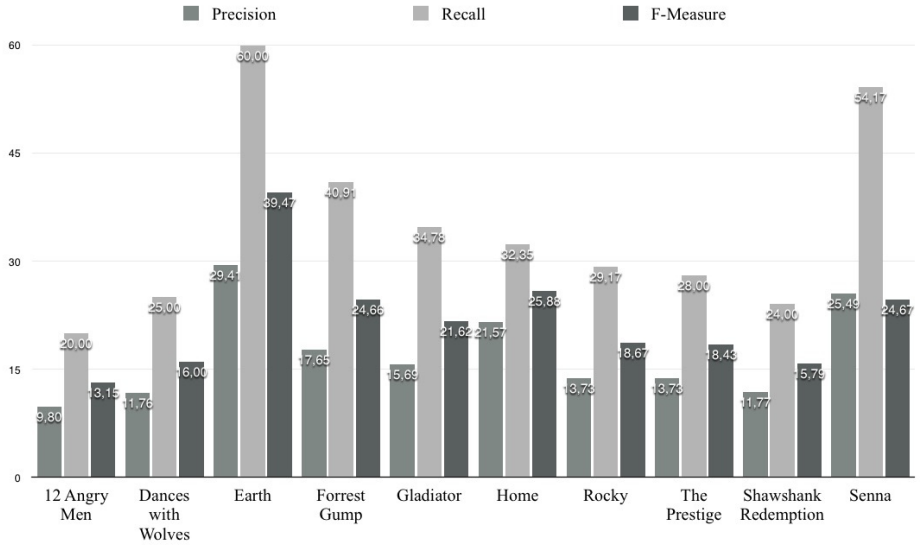

Figure 1.Evaluation overview of TextRank performance with subtitles.

Extended version was published in Proceedings of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, pp. 20-24.

Acknowledgement: This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] Langlois, T. et al.: VIRUS: video information retrieval using subtitles. In: Proc. of the 14th Int. Academic Mind Trek Conf.: Envisioning Future Media Environments, ACM, (2010), pp. 197–200.