Pattern Recognition Letters 73 (2016) 7–12

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Summarization of films and documentaries based on subtitles and scripts☆

Marta Aparício[a,b], Paulo Figueiredo[a,c], Francisco Raposo[a,c], David Martins de Matos[a,c,*], Ricardo Ribeiro[a,b], Luís Marujo[a]

[a] L2F - INESC ID Lisboa, Rua Alves Redol, 9, 1000–029 Lisboa, Portugal
[b] Instituto Universitário de Lisboa (ISCTE-IUL), Av. das Forças Armadas, Lisboa 1649–026, Portugal
[c] Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa 1049–001, Portugal

ARTICLE INFO

Article history:
Received 5 June 2015
Available online 15 January 2016

Keywords:
Automatic text summarization
Generic summarization
Summarization of films
Summarization of documentaries

ABSTRACT

We assess the performance of generic text summarization algorithms applied to films and documentaries, using extracts from news articles produced by reference models of extractive summarization. We use three datasets: (i) news articles, (ii) film scripts and subtitles, and (iii) documentary subtitles. Standard ROUGE metrics are used for comparing generated summaries against news abstracts, plot summaries, and synopses. We show that the best performing algorithms are LSA, for news articles and documentaries, and LexRank and Support Sets, for films. Despite the different nature of films and documentaries, their relative behavior is in accordance with that obtained for news articles.

1. Introduction

Input media for automatic summarization has varied from text [5,18] to speech [21,34,39] and video [1], but the application domain has been, in general, restricted to informative sources: news [2,11,30,33], meetings [8,26], or lectures [7]. Nevertheless, application areas within the entertainment industry are gaining attention: e.g. summarization of literary short stories [12], music summarization [31], summarization of books [24], or inclusion of character analyses in movie summaries [36]. We follow this direction, creating extractive, text-driven video summaries for films and documentaries.

Documentaries started as cinematic portrayals of reality [10]. Today, they continue to portray historical events, argumentation, and research. They are commonly understood as capturing reality and therefore, seen as inherently non-fictional. Films, in contrast, are usually associated with fiction. However, films and documentaries do not fundamentally differ: many of the strategies and narrative structures employed in films are also used in documentaries [27].

In the context of our work, films (fictional) tell stories based on fictive events, whereas documentaries (non-fictional) address, mostly, scientific subjects. We study the parallelism between the information carried in subtitles and scripts of both films and documentaries. Extractive summarization methods have been extensively explored for news documents [16,22,23,29,30,37]. Our main goal is to understand the quality of automatic summaries, produced for films and documentaries, using the well-known behavior of news articles as reference. Generated summaries are evaluated against manual abstracts using ROUGE metrics, which correlate with human judgements [15,17].

This article is organized as follows: Section 2 presents the summarization algorithms; Section 3 presents the collected datasets; Section 4 presents the evaluation setup; Section 5 discusses our results; and Section 6 presents conclusions and directions for future work.

2. Generic summarization

Six text-based summarization approaches were used to summarize newspaper articles, subtitles, and scripts. They are described in the following sections.

2.1. Maximal Marginal Relevance (MMR)

MMR is a query-based summarization method [4]. It iteratively selects sentences via Eq. (1) (Q is a query; Sim_1 and Sim_2 are

☆ This paper has been recommended for acceptance by Jie Zou.

* Corresponding author at: Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049–001 Lisboa, Portugal. Tel.: +351213100305.
E-mail address: david.matos@inesc-id.pt, david.matos@tecnico.ulisboa.pt (D. Martins de Matos).

http://dx.doi.org/10.1016/j.patrec.2015.12.016
0167-8655/

similarity metrics; S_i and S_j are non-selected and previously selected sentences, respectively). λ balances relevance and novelty. MMR can generate generic summaries by considering the input sentences centroid as a query [25,38].

$$\arg\max_{S_i}[\lambda \mathrm{Sim}_1(S_i, Q) - (1-\lambda)\max_{S_j}\mathrm{Sim}_2\left(S_i, S_j\right)] \quad (1)$$

2.2. LexRank

LexRank [6] is a centrality-based method based on Google's PageRank [3]. A graph is built using sentences, represented by TF-IDF vectors, as vertexes. Edges are created when the cosine similarity exceeds a threshold. Eq. (2) is computed at each vertex until the error rate between two successive iterations is lower than a certain value. In this equation, d is a damping factor to ensure the method's convergence, N is the number of vertexes, and $S(V_i)$ is the score of the ith vertex.

$$S(V_i) = \frac{(1-d)}{N} + d \times \sum_{V_j \in \mathrm{adj}[V_i]} \frac{\mathrm{Sim}\left(V_i, V_j\right)}{\sum_{V_k \in \mathrm{adj}[V_j]} \mathrm{Sim}\left(V_j, V_k\right)} S\left(V_j\right) \quad (2)$$

2.3. Latent Semantic Analysis (LSA)

LSA infers contextual usage of text based on word co-occurrence [13,14]. Important topics are determined without the need for external lexical resources [9]: each word's occurrence context provides information concerning its meaning, producing relations between words and sentences that correlate with the way humans make associations. Singular Value Decomposition (SVD) is applied to each document, represented by a $t \times n$ term-by-sentences matrix A, resulting in its decomposition $U\Sigma V^T$. Summarization consists of choosing the k highest singular values from Σ, giving Σ_k. U and V^T are reduced to U_k and V_k^T, respectively, approximating A by $A_k = U_k\Sigma_k V_k^T$. The most important sentences are selected from V_k^T.

2.4. Support sets

Documents are typically composed by a mixture of subjects, involving a main and various minor themes. Support sets are defined based on this observation [35]. Important content is determined by creating a support set for each passage, by comparing it with all others. The most semantically-related passages, determined via geometric proximity, are included in the support set. Summaries are composed by selecting the most relevant passages, i.e., the ones present in the largest number of support sets. For a segmented information source $I \triangleq p_1, p_2, \ldots, p_N$, support sets S_i for each passage p_i are defined by Eq. (3), where Sim is a similarity function, and ϵ_i is a threshold. The most important passages are selected by Eq. (4).

$$S_i \triangleq \{s \in I : \mathrm{Sim}(s, p_i) > \epsilon_i \wedge s \neq p_i\} \quad (3)$$

$$\arg\max_{s \in U_{i=1}^n S_i} |\{S_i : s \in S_i\}| \quad (4)$$

2.5. Key Phrase-based Centrality (KP-Centrality)

Ribeiro et al. [32] proposed an extension of the centrality algorithm described in Section 2.4, which uses a two-stage important passage retrieval method. The first stage consists of a feature-rich supervised key phrase extraction step, using the MAUI toolkit with additional semantic features: the detection of rhetorical signals, the number of Named Entities, Part-Of-Speech (POS) tags, and 4 n-gram domain model probabilities [19,20]. The second stage consists of the extraction of the most important passages, where key phrases are considered regular passages.

2.6. Graph Random-walk with Absorbing StateS that HOPs among PEaks for Ranking (GRASSHOPPER)

GRASSHOPPER [40] is a re-ranking algorithm that maximizes diversity and minimizes redundancy. It takes a weighted graph W ($n \times n$: n vertexes representing sentences; weights are defined by a similarity measure), a probability distribution r (representing a prior ranking), and $\lambda \in [0, 1]$, that balances the relative importance of W and r. If there is no prior ranking, a uniform distribution can be used. Sentences are ranked by applying the teleporting random walks method in an absorbing Markov chain, based on the $n \times n$ transition matrix $\tilde{P}$ (calculated by normalizing the rows of W), i.e., $P = \lambda\tilde{P} + (1-\lambda)\mathbf{1}\mathbf{r}^\top$. The first sentence to be scored is the one with the highest stationary probability $\arg\max_{i=1}^n \pi_i$ according to the stationary distribution of P: $\pi = P^\top\pi$. Already selected sentences may never be visited again, by defining $P_{gg} = 1$ and $P_{gi} = 0, \forall i \neq g$. The expected number of visits is given by matrix $N = (I-Q)^{-1}$ (where N_{ij} is the expected number of visits to the sentence j, if the random walker began at sentence i). We obtain the average of all possible starting sentences to get the expected number of visits to the jth sentence, v_j. The sentence to be selected is the one that satisfies $\arg\max_{i=|G|+1}^n v_i$.

3. Datasets

We use three datasets: newspaper articles (baseline data), films, and documentaries. Film data consists of subtitles and scripts, containing scene descriptions and dialog. Documentary data consists of subtitles containing mostly monologue. Reference data consists of manual abstracts (for newspaper articles), plot summaries (for films and documentaries), and synopses (for films). Plot summaries are concise descriptions, sufficient for the reader to get a sense of what happens in the film or documentary. Synopses are much longer and may contain important details concerning the turn of events in the story. All datasets were normalized by removing punctuation inside sentences and timestamps from subtitles.

3.1. Newspaper articles

TeMário [28] is composed by 100 newspaper articles in Brazilian Portuguese (Table 1), covering domains such as "world", "politics", and "foreign affairs". Each article has a human-made reference summary (abstract).

3.2. Films

We collected 100 films, with an average of 4 plot summaries (minimum of 1, maximum of 7) and 1 plot synopsis per film (Table 2). Table 3 presents the properties of the subtitles, scripts, and the concatenation of both. Not all the information present in the scripts was used: dialogs were removed in order to make them more similar to plot summaries.

Table 1
TeMário corpus properties.

		AVG	MIN	MAX
#Sentences	News story	29	12	68
	Summary	9	5	18
#Words	News story	608	421	1315
	Summary	192	120	345