Repositório ISCTE-IUL

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, 1649-026 Lisboa, Portugal*
[c]*Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*

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 [18, 5] to speech [21, 39, 34] and video [1], but the application domain has been, in general, restricted to informative sources: news [2, 30, 33, 11], meetings [26, 8], 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].

---

*Corresponding author
e-mail: david.matos@inesc-id.pt (David Martins de Matos)

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, 37, 29, 30, 23]. 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; 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 Equation 1 (Q is a query; Sim_1 and Sim_2 are 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].

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

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. Equation 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 $S\left(V_i\right)$ is the score of the ith vertex.

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

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 rela-

tions 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 $U \Sigma V^T$. Summa-

rization 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 Equation 3, where Sim is a

similarity function, and ϵ_i is a threshold. The most important passages are selected by Equation 4.

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

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

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 [20, 19]. 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., $P = \lambda \tilde{P} + (1 - \lambda) \mathbf{1r}^\top$. 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).

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

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 2: Properties of plot summaries and synopses.

		AVG	MIN	MAX
#Sentences	Plot Summaries	5	1	29
	Plot Synopses	89	6	399
#Words	Plot Summaries	107	14	600
	Plot Synopses	1677	221	7110

Table 3: Properties of subtitles and scripts.

		AVG	MIN	MAX
#Sentences	Subtitles	1573	309	4065
	Script	1367	281	3720
	Script + Subtitles	2787	1167	5388
#Words	Subtitles	10460	1592	27800
	Script	14560	3493	34700
	Script + Subtitles	24640	11690	47140

3.3. Documentaries

We collected 98 documentaries. Table 4 presents the properties of their subtitles: note that the number of sentences is smaller than in films, influencing ROUGE (recall-based) scores.

Table 4: Properties of documentaries subtitles.

	AVG	MIN	MAX
#Sentences	340	212	656
#Words	5864	3961	10490

We collected 223 manual plot summaries and divided them into four classes (Table 5): 143 "Informative", 63 "Interrogative", 9 "Inviting", and 8 "Challenge". "Informative" summaries contain factual information about the program; "Interrogative" summaries contain questions that arouse viewer curiosity, e.g. "What is the meaning of life?"; "Inviting" are invitations, e.g. "Got time for a 24 year vacation?"; and, "Challenge" entice viewers on a personal basis, e.g. "are you ready for...?". We chose "Informative" summaries due to their resemblance to the sentences extracted by the summarization algorithms. On average, there are 2 informative plot summaries per documentary (minimum of 1, maximum of 3).

Table 5: Properties of the documentary plot summaries.

		AVG	MIN	MAX
	Informative	4	1	18
#Sentences	Interrogative	4	1	19
	Inviting	6	2	11
	Challenge	5	2	9
	Informative	82	26	384
#Words	Interrogative	103	40	377
	Inviting	146	63	234
	Challenge	104	59	192

4. Experimental Setup

For news articles, summaries were generated with the average size of the manual abstracts ($\approx 31\%$ of their size).

For each film, two summaries were generated, by selecting a number of sentences equal to (i) the average length of its manual plot summaries, and (ii) the length of its synopsis. In contrast with news articles and documentaries, three types of input were considered: script, subtitles, script+subtitles.

For each documentary, a summary was generated with the same average number of sentences of its manual plot summaries ($\approx 1\%$ of the documentary's size).

Content quality of summaries is based on word overlap (as defined by ROUGE) between generated summaries and their references. ROUGE-N computes the fraction of selected words that are correctly identified by the summarization algorithms (cf. Equation 5: RS are reference summaries, gram_n is the n-gram length, and $\mathrm{count}_{\mathrm{match}}(\mathrm{gram}_n)$ is the maximum number of n-grams of a candidate summary that co-occur with a set of reference summaries). ROUGE-SU measures the overlap of skip-bigrams (any pair of words in their sentence order, with the addition of unigrams as counting unit). ROUGE-SU4 limits the maximum gap length of skip-bigrams to 4.

$$\text{ROUGE-N} = \frac{\sum_{S \in \mathrm{RS}} \sum_{\mathrm{gram}_n \in S} \mathrm{count}_{\mathrm{match}}(\mathrm{gram}_n)}{\sum_{S \in \mathrm{RS}} \sum_{\mathrm{gram}_n \in S} \mathrm{count}(\mathrm{gram}_n)} \tag{5}$$

5. Results and Discussion

Subtitles and scripts were evaluated against manual plot summaries and synopses to define an optimal performance reference. The following sections present averaged ROUGE-1, ROUGE-2, and ROUGE-SU4 scores (henceforth R-1, R-2, and R-SU4), and the performance of each summarization algorithm, as a ratio between the score of the generated summaries and this reference (relative performance). Several parametrizations of the algorithms were used (we present only the best results). Concerning MMR, we found that the best λ corresponds to a higher average number of words per summary. Concerning GRASSHOPPER, we used the uniform distribution as prior.

5.1. Newspaper Articles (TeMário)

Table 6 presents the scores for each summarization algorithm. LSA achieved the best scores for R-1, R-2, and R-SU4. Figure 1 shows the relative performance results.

Table 6: ROUGE scores for generated summaries and original documents against manual references. For MMR, $\lambda = 0.50$; Support Sets used Manhattan distance and Support Set Cardinality = 2; KP-Centrality used 10 key phrases.

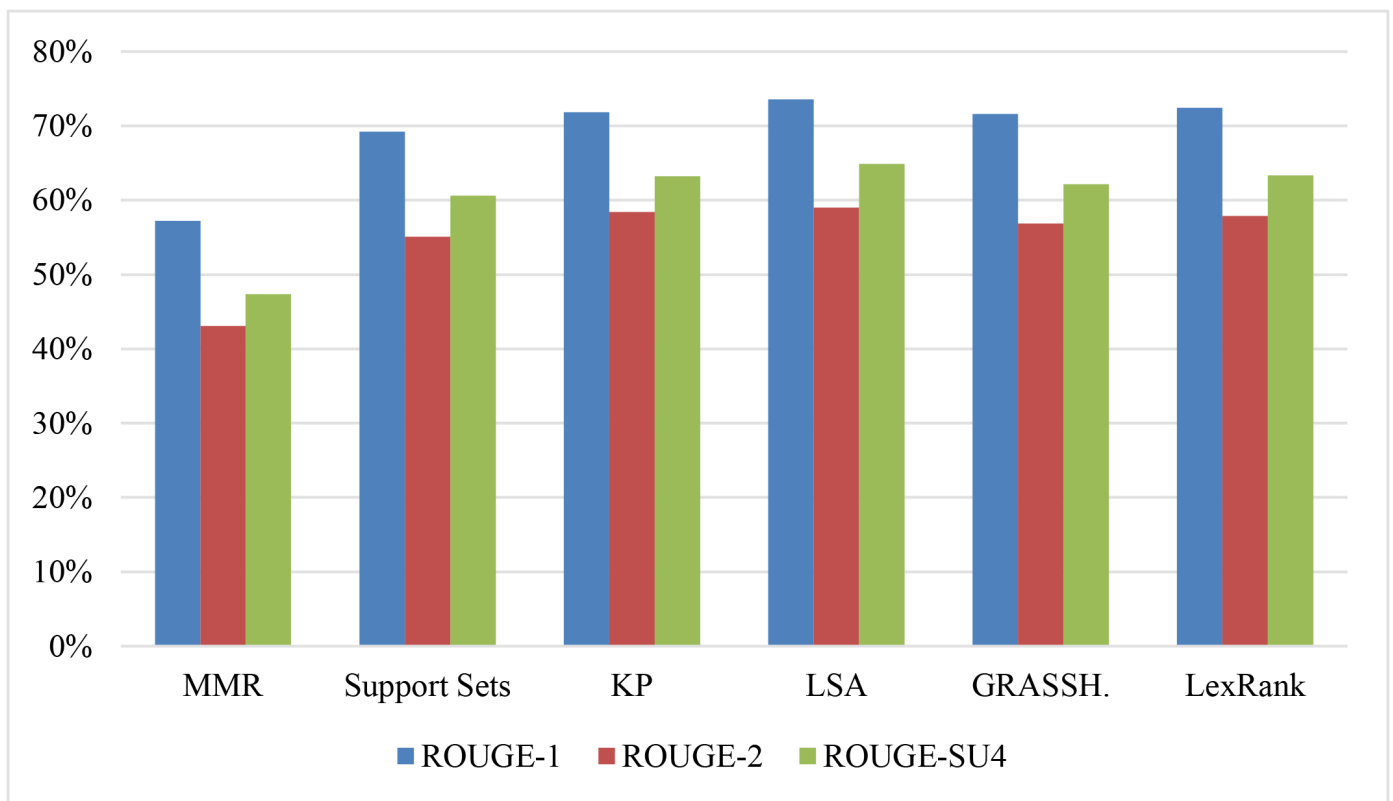|	R-1	R-2	R-SU4	AVG #Words
MMR	0.43	0.15	0.18	195
Support Sets	0.52	0.19	0.23	254
KP	0.54	**0.20**	**0.24**	268
LSA	**0.56**	**0.20**	**0.24**	297
GRASSH.	0.54	0.19	0.23	270
LexRank	0.55	**0.20**	**0.24**	277
Original Docs	0.75	0.34	0.38	608

Fig. 1: Relative performance for news articles. For MMR, $\lambda = 0.50$; Support Sets used Manhattan distance and Support Set Cardinality = 2; KP-Centrality used 10 key phrases.

5.2. Films

Table 7 presents the scores for the film data combinations against plot summaries. Overall, Support Sets capture the most relevant sentences for plot summaries. Nevertheless, LSA has comparable behavior according to R-2. It would be expected, for algorithms such as GRASSHOPPER and MMR, that maximize diversity, to perform well in this context, because plot summaries

are relatively small and focus on the more important aspects of the film, ideally, without redundant content. However, our results show otherwise. For scripts, LSA and LexRank are the best approaches in terms of R-1 and R-SU4.

Table 7: ROUGE scores for generated summaries for subtitles, scripts, and scripts concatenated with subtitles, against plot summaries. For MMR, $\lambda = 0.50$; Support Sets used the cosine distance and threshold = 50%; KP-Centrality used 50 key phrases.

		R-1	R-2	R-SU4	AVG #Words
MMR	Subtitles	0.07	0.01	0.02	52
	Script	0.14	0.01	0.03	53
	Script + Subtitles	0.12	0.01	0.03	71
Support Sets	Subtitles	0.23	**0.02**	**0.06**	150
	Script	0.25	0.02	0.07	133
	Script + Subtitles	0.29	**0.03**	**0.09**	195
KP	Subtitles	0.22	0.02	**0.06**	144
	Script	0.24	0.02	0.07	123
	Script + Subtitles	0.28	0.02	0.08	184
LSA	Subtitles	0.22	**0.02**	**0.06**	167
	Script	0.28	**0.03**	0.08	190
	Script + Subtitles	0.28	**0.03**	0.08	219
GRASSH.	Subtitles	0.17	0.01	0.04	135
	Script	0.21	0.02	0.06	121
	Script + Subtitles	0.22	0.02	0.06	118
LexRank	Subtitles	**0.24**	**0.02**	**0.06**	177
	Script	**0.29**	0.02	**0.09**	168
	Script + Subtitles	**0.30**	0.02	0.08	217
Original Docs	Subtitles	0.77	0.21	0.34	10460
	Script	0.74	0.23	0.36	14560
	Script + Subtitles	0.83	0.31	0.43	24640

Table 8: ROUGE scores for generated summaries for subtitles, scripts, and scripts+subtitles, against plot synopses. For MMR, $\lambda = 0.50$; Support Sets used the cosine distance and threshold = 50%; KP-Centrality used 50 key phrases.

		R-1	R-2	R-SU4	AVG #Words
MMR	Subtitles	0.08	0.01	0.02	435
	Script	0.16	0.03	0.06	745
	Script + Subtitles	0.11	0.01	0.03	498
Support Sets	Subtitles	0.25	0.04	0.08	1033
	Script	0.37	0.07	0.15	1536
	Script + Subtitles	0.42	0.08	0.16	1736
KP	Subtitles	0.24	0.04	0.08	952
	Script	0.36	0.07	0.14	1419
	Script + Subtitles	0.40	0.08	0.16	1580
LSA	Subtitles	0.31	**0.06**	0.11	1303
	Script	0.42	0.09	0.17	1934
	Script + Subtitles	0.45	**0.10**	0.18	2065
GRASSH.	Subtitles	**0.34**	**0.06**	**0.12**	1553
	Script	0.44	0.09	**0.18**	1946
	Script + Subtitles	0.47	**0.10**	**0.19**	1768
LexRank	Subtitles	**0.34**	**0.06**	**0.12**	1585
	Script	**0.45**	**0.10**	**0.18**	1975
	Script + Subtitles	**0.48**	**0.10**	**0.19**	2222
Original Docs	Subtitles	0.70	0.18	0.30	10460
	Script	0.73	0.24	0.37	14560
	Script + Subtitles	0.83	0.32	0.44	24640

Table 8 presents the scores for the film data combinations against plot synopses. The size of synopses is very different from that of plot summaries. Although synopses also focus on the major events of the story, their larger size allows for a more refined description of film events. Additionally, because summaries are created with the same number of sentences of the corresponding synopsis, higher scores are expected. From all algorithms, LexRank clearly stands out with the highest scores for all metrics (except for R-SU4, for scripts).

The script+subtitles combination was used in order to determine whether the inclusion of redundant content would improve the scores, over the separate use of scripts and subtitles. However, in most cases (Figure 4), script+subtitles leads to worse scores, except for Support Sets for R-1. For plot synopses, the best scores are achieved by LexRank and GRASSHOPPER, while, for plot summaries, the best scores are achieved by LexRank and LSA. By inspection of the summaries produced by each algorithm, we observed that MMR chooses sentences with fewer words in comparison with all other algorithms (normally, leading to lower

scores). Overall, the algorithms behave similarly for both subtitles and scripts.

5.3. Documentaries

From all algorithms (Table 9), LSA achieved the best results for R-1 and R-SU4, along with LexRank for R-1. KP-Centrality achieved the best results for R-2. It is important to notice that LSA also produces the summaries with the highest word count (favoring recall). Figure 2 shows the relative performance results: LSA outperformed all other algorithms for R-1 and R-SU4, and KP-Centrality for R-2; Support Sets and KP-Centrality performed closely to LSA for R-SU4; the best MMR results were consistently worse across all metrics (MMR summaries have the lowest word count).

Table 9: ROUGE scores for generated summaries and original subtitles against human-made plot summaries. For MMR, $\lambda = 0.75$; Support Sets used the cosine distance and threshold = 50%; KP-Centrality used 50 key phrases.

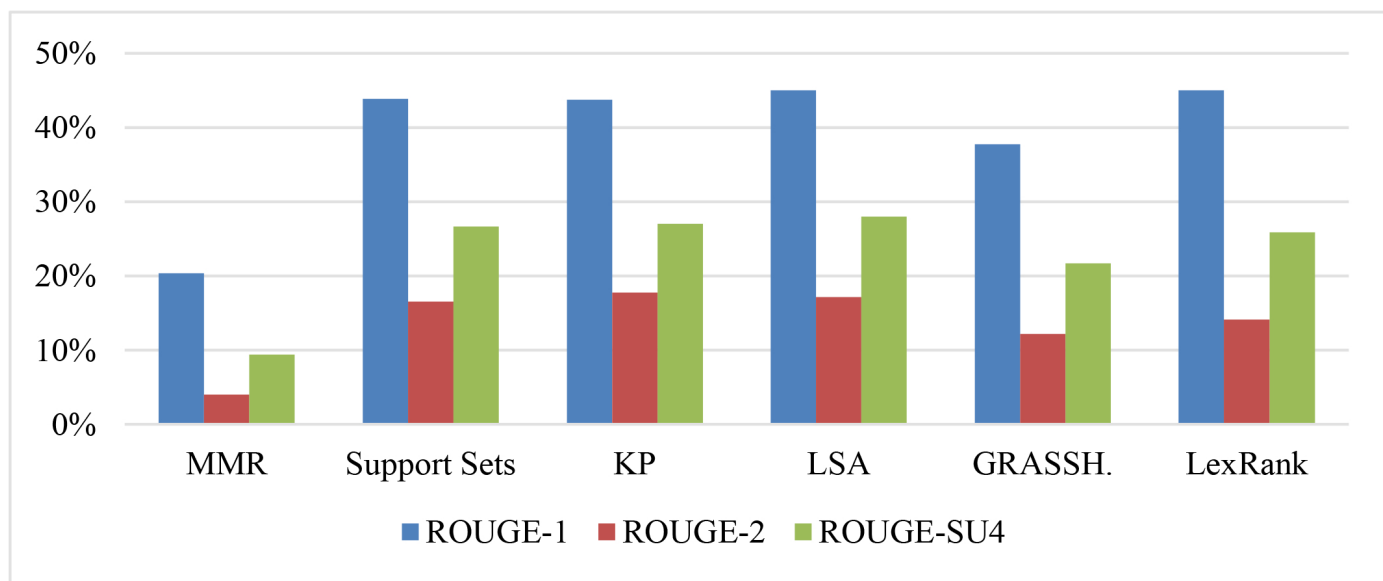|	R-1	R-2	R-SU4	AVG #Words
MMR	0.17	0.01	0.04	78
Support Sets	0.37	0.06	0.12	158
KP	0.37	**0.07**	0.12	149
LSA	**0.38**	0.06	**0.13**	199
GRASSH.	0.31	0.04	0.10	150
LexRank	**0.38**	0.05	0.12	183
Original Docs	0.83	0.37	0.46	5864

Fig. 2: Relative performance for documentaries against plot summaries. For MMR, $\lambda = 0.75$; Support Sets used cosine distance and threshold=50%; KP-Centrality used 50 key phrases.

5.4. Discussion

News articles intend to answer basic questions about a particular event: who, what, when, where, why, and often, how. Their structure is sometimes referred to as "inverted pyramid", where the most essential information comes first. Typically, the first sentences provide a good overview of the entire article and are more likely to be chosen when composing the final summary. Although documentaries follow a narrative structure similar to films, they can be seen as more closely related to news than films, especially regarding their intrinsic informative nature. In spite of their different natures, however, summaries created by humans produce similar scores for all of them. It is possible to observe this behavior in Figure 3. Note that documentaries achieve higher scores than news articles or films, when using the original subtitles documents against the corresponding manual plot summaries.
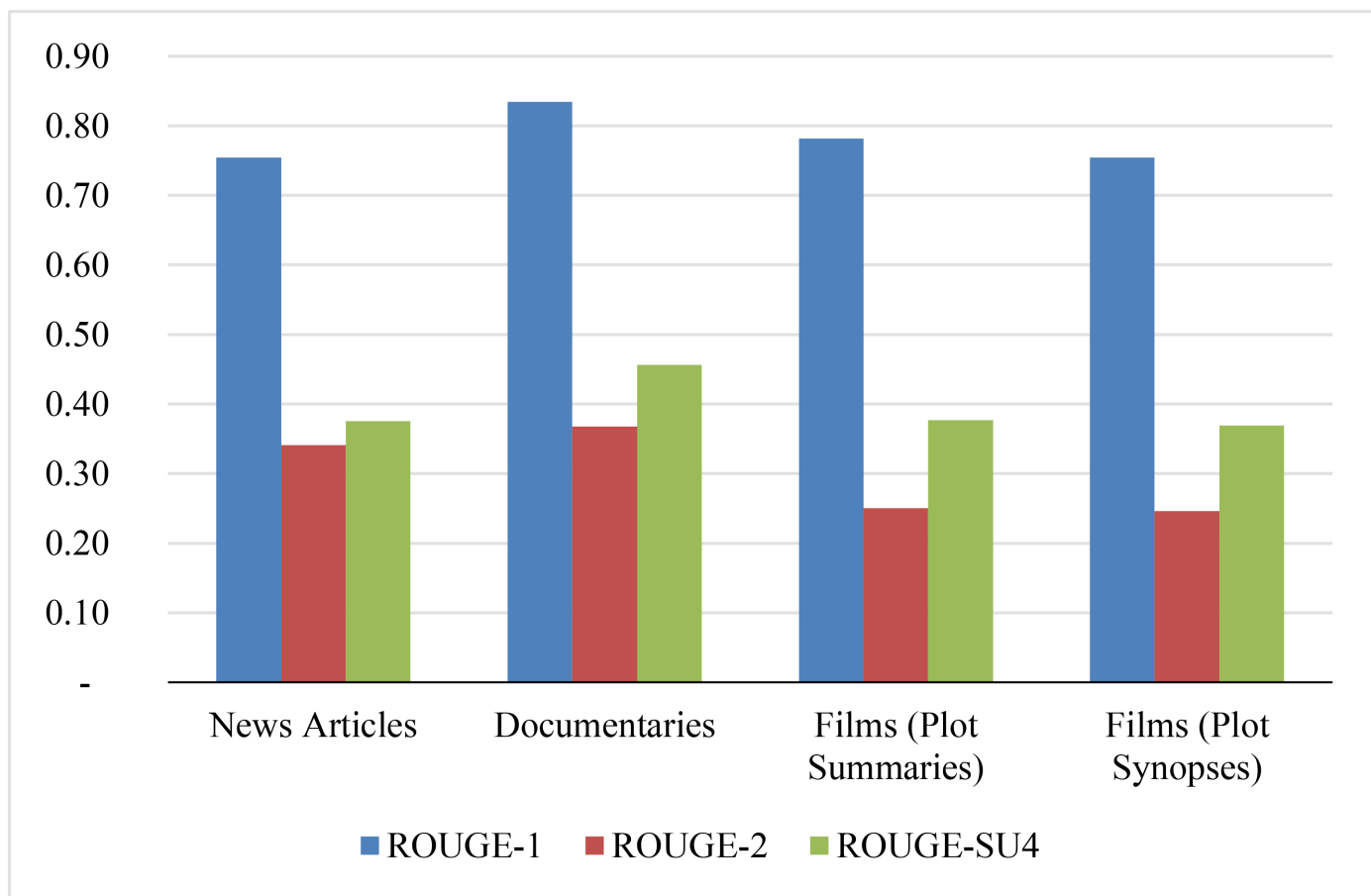

Fig. 3: ROUGE scores for news articles, films, and documentaries against manual references, plot summaries and synopses, and plot summaries, respectively.

Figure 4 presents an overview of the performance of each summarization algorithm across all domains. The results concerning news articles were the best out of all three datasets for all experiments. However, summaries for this dataset preserve, approxi-

mately, 31% of the original articles, in terms of sentences, which is significantly higher than for films and documentaries (which preserve less than 1%), necessarily leading to higher scores. Nonetheless, we can observe the differences in behavior between these domains. Notably, documentaries achieve the best results for plot summaries, in comparison with films, using scripts, subtitles, or the combination of both. The relative scores on the films dataset are influenced by two major aspects: the short sentences found in the films dialogs; and, since the generated summaries are extracts from subtitles and scripts, they are not able to represent the film as a whole, in contrast with what happens with plot summaries or synopses. Additionally, the experiments conducted for script+subtitles for films, in general, do not improve scores above those of scripts alone, except for Support Sets for R-1. Overall, LSA performed consistently better for news articles and documentaries. Similar relatively good behavior had already been observed for meeting recordings, where the best summarizer was also LSA [26]. One possible reason for these results is that LSA tries to capture the relation between words in sentences. By inferring contextual usage of text based on these relations, high scores, apart from R-1, are produced for R-2 and R-SU4. For films, LexRank was the best performing algorithm for subtitles, scripts and the combination of both, using plot synopses, followed by LSA and Support Sets for plot summaries. MMR has the lowest scores for all metrics and all datasets. We observed that sentences closer to the centroid typically contain very few words, thus leading to shorter summaries and the corresponding low scores.

Interestingly, by observing the average of R-1, R-2, and R-SU4, it is possible to notice that it follows very closely the values of R-SU4. These results suggest that R-SU4 adequately reflects the scores of both R-1 and R-2, capturing the concepts derived from both unigrams and bigrams.

Overall, considering plot summaries, documentaries achieved higher results in comparison with films. However, in general, the highest score for these two domains is achieved using films scripts against plot synopses. Note that synopses have a significant difference in terms of sentences in comparison with plot summaries. The average synopsis has 120 sentences, while plot summaries have, on average, 5 sentences for films, and 4 for documentaries. This gives synopses a clear advantage in terms of ROUGE (recall-based) scores, due to the high count of words.

6. Conclusions and Future Work

We analyzed the impact of the six summarization algorithms on three datasets. The newspaper articles dataset was used as a reference. The other two datasets, consisting of films and documentaries, were evaluated against plot summaries, for films and documentaries, and synopses, for films. Despite the different nature of these domains, the abstractive summaries created by humans, used for evaluation, share similar scores across metrics.

The best performing algorithms are LSA, for news and documentaries, and LexRank and Support Sets, for films. Moreover, we conducted experiments combining scripts and subtitles for films, in order to assess the performance of generic algorithms by inclusion of redundant content. Our results suggest that this combination is unfavorable. Additionally, it is possible to observe that all algorithms behave similarly for both subtitles and scripts. As previously mentioned, the average of the scores follows closely the values of R-SU4, suggesting that R-SU4 is able to capture concepts derived from both unigrams and bigrams.

We plan to use subtitles as a starting point to perform video summaries of films and documentaries. For films, the results from our experiments using plot summaries show that the summarization of scripts only marginally improved performance, in comparison with subtitles. This suggests that subtitles are a viable approach for text-driven film and documentary summarization. This positive aspect is compounded by their being broadly available, as opposed to scripts.

7. Acknowledgements

References

[1] Ajmal, M., Ashraf, M., Shakir, M., Abbas, Y., Shah, F., 2012. Video summarization: Techniques and classification, in: Computer Vision and Graphics. Springer Berlin Heidelberg, pp. 1–13.

[2] Barzilay, R., Elhadad, N., McKeown, K., 2002. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research , 35–55.

[3] Brin, S., Page, L., 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine, in: Proc. of the 7th Intl. Conf. on World Wide Web, pp. 107–117.

[4] Carbonell, J., Goldstein, J., 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries, in: Proc. of the 21st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 335–336.

[5] Edmundson, H.P., 1969. New methods in automatic abstracting. Journal of the Association for Computing Machinery 16, 264–285.
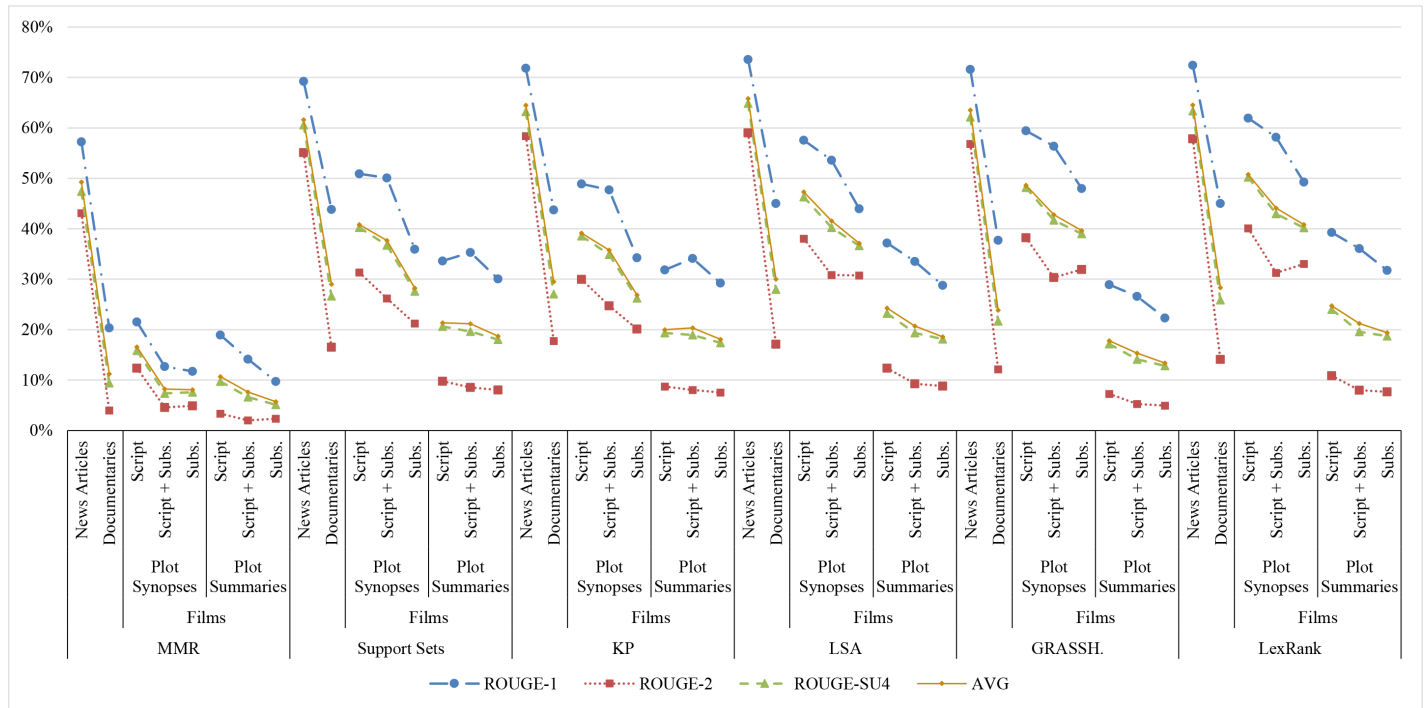
Fig. 4: Relative performance for all datasets. For films the relative performance was measured against plot synopses and plot summaries: MMR used $\lambda = 0.50$; and Support Sets used the cosine distance and threshold $= 50\%$; KP-Centrality used 50 key phrases.

[6] Erkan, G., Radev, D.R., 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research , 457–479.

[7] Fujii, Y., Kitaoka, N., Nakagawa, S., 2007. Automatic extraction of cue phrases for important sentences in lecture speech and automatic lecture speech summarization., in: Proc. of INTERSPEECH 2007, pp. 2801–2804.

[8] Garg, N., Favre, B., Reidhammer, K., Hakkani-Tür, D., 2009. ClusterRank: A Graph Based Method for Meeting Summarization, in: Proc. of INTERSPEECH 2009, pp. 1499–1502.

[9] Gong, Y., Liu, X., 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, in: Proc. of the 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 19–25.

[10] Grant, B.K., Sloniowski, J., 1998. Documenting the Documentary: Close Readings of Documentary Film and Video. Wayne State University Press.

[11] Hong, K., Conroy, J.M., Favre, B., Kulesza, A., Lin, H., Nenkova, A., 2014. A repository of state of the art and competitive baseline summaries for generic news summarization, in: Proc. of the Ninth Intl. Conf. on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014., pp. 1608–1616.

[12] Kazantseva, A., Szpakowicz, S., 2010. Summarizing short stories. Computational Linguistics 36, 71–109.

[13] Landauer, T.K., Dutnais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240.

[14] Landauer, T.K., Foltz, P.W., Laham, D., 1998. An introduction to latent semantic analysis. Discourse processes 25, 259–284.

[15] Lin, C.Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summ. Branches Out: Proc. of the ACL-04 Workshop, pp. 74–81.

[16] Lin, C.Y., Hovy, E., 2000. The automated acquisition of topic signatures for text summarization, in: Proc. of the 18th Conf. on Computational Linguistics - Volume 1, pp. 495–501.

[17] Liu, F., Liu, Y., 2010. Exploring correlation between rouge and human evaluation on meeting summaries. IEEE Transactions on Audio, Speech & Language Processing 18, 187–196.

[18] Luhn, H.P., 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2, 159–165.

[19] Marujo, L., Gershman, A., Carbonell, J., Frederking, R., Neto, J.P., 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization, in: Chair), N.C.C., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources

Association (ELRA).

[20] Marujo, L., Viveiros, M., Neto, J.P., 2011. Keyphrase cloud generation of broadcast news., in: INTERSPEECH, ISCA. pp. 2393–2396.

[21] Maskey, S.R., Hirschberg, J., 2005. Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization, in: Proc. of the 9th EUROSPEECH - INTERSPEECH 2005, pp. 621–624.

[22] McKeown, K., Hirschberg, J., Galley, M., Maskey, S., 2005. From text to speech summarization, in: Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE Intl. Conf. on, pp. v/997–v1000 Vol. 5.

[23] McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B., Sigelman, S., Summarization, M., 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster, in: Proc. of HLT 2002, pp. 280–285.

[24] Mihalcea, R., Ceylan, H., 2007. Explorations in automatic book summarization, in: EMNLP-CoNLL'07, pp. 380–389.

[25] Murray, G., Renals, S., Carletta, J., 2005a. Extractive Summarization of Meeting Recordings, in: Proc. of the 9th European Conf. on Speech Communication and Technology, pp. 593–596.

[26] Murray, G., Renals, S., Carletta, J., 2005b. Extractive Summarization of Meeting Records, in: Proc. of the 9th EUROSPEECH - INTERSPEECH 2005, pp. 593–596.

[27] Nichols, B., 1991. Representing Reality: Issues and Concepts in Documentary. Bloomington: Indiana University Press.

[28] Pardo, T.A.S., Rino, L.H.M., 2003. TeMario: a corpus for automatic text summarization. Technical Report. Núcleo Interinstitucional de Linguística Computacional (NILC).

[29] Radev, D.R., Blair-goldensohn, S., Zhang, Z., Raghavan, R.S., 2001. NewsInEssence: A System For Domain-Independent, Real-Time News Clustering and Multi-Document Summarization, in: Proc. of the First Intl. Conf. on Human Language Technology Research, pp. 1–4.

[30] Radev, D.R., Otterbacher, J., Winkel, A., Blair-Goldensohn, S., 2005. NewsInEssence: Summarizing Online News Topics. Communications of the ACM 48, 95–98.

[31] Raposo, F., Ribeiro, R., de Matos, D., 2015. On the Application of Generic Summarization Algorithms to Music. IEEE Signal Processing Letters , 26–30.

[32] Ribeiro, R., Marujo, L., de Matos, D., Neto, J.P., Gershman, A., Carbonell, J., 2013. Self reinforcement for important passage retrieval. Digital. URL: http://dl.acm.org/citation.cfm?id=2484028.2484134.

[33] Ribeiro, R., de Matos, D., 2007. Extractive summarization of broadcast news: Comparing strategies for european portuguese., in: TSD, pp. 115–122.

[34] Ribeiro, R., de Matos, D., 2012. Summarizing speech by contextual reinforcement of important passages, in: Proc. of PROPOR 2012, pp. 392–402.

[35] Ribeiro, R., de Matos, D.M., 2011. Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. Journal of Artificial Intelligence Research , 275–308.

[36] Sang, J., Xu, C., 2010. Character-based movie summarization, in: Proc. of the Intl. Conf. on Multimedia, pp. 855–858.

[37] Spärck Jones, K., 2007. Automatic summarising: The state of the art. Inf. Process. Manage. 43, 1449–1481.

[38] Xie, S., Liu, Y., 2008. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization, in: Proc. - ICASSP, IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, pp. 4985–4988.

[39] Zhang, J.J., Chan, R.H.Y., Fung, P., 2010. Extractive Speech Summarization Using Shallow Rhetorical Structure Modeling. IEEE Transactions on Audio Speech and Language Processing 18, 1147–1157.

[40] Zhu, X., Goldberg, A.B., Gael, J.V., Andrzejewski, D., 2007. Improving Diversity in Ranking using Absorbing Random Walks, in: Proc. of the 5th NAACL - HLT, pp. 97–104.