An Approach to Automatic Summarization of Television Programs

Marco Canora, Fernando García-Granada(✉), Emilio Sanchis, and Encarna Segarra

Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
marcaal1@inf.upv.es, {fgarcia,esanchis,esegarra}@dsic.upv.es

Abstract. In this paper we present an approach to document summarization based on unsupervised techniques. We study the adequacy of these techniques to the problem of documents in which many topics of different duration are present, in our case the transcriptions of Spanish TV programs. The paper compares a classical Latent Semantic Analysis approach to a new proposal based on Latent Dirichlet Allocation. It is also studied the application of the summarization process to the different segments obtained in a previous process of topic segmentation. The topic segmentation is performed by considering distances between paragraphs, that are represented by means of continuous vectors obtained from the words contained in them. Experiments on some TV programs of political and miscellaneous news have been performed.

Keywords: Document summarization · Latent Dirichlet Allocation
Latent Semantic Analysis

1 Introduction

Multimedia content summarization is an important issue in recent years. Due to the great amount of information available in the web it is necessary to have different tools to help the users to digest that contents in an easy way. For this reason, summarization techniques are a current goal in Natural Language research [9,14]. Traditionally, summarization methods are classified in two categories: extractive and abstractive. Extractive approaches consist of detecting the most salient sentences and the summary generated is composed by those sentences, while abstractive approaches try to be more similar to human summaries and they generate new sentences that may not be in the original document. Although, logically, these last approaches are a more ambitious challenge, recent works have shown promising expectations [3,11]. In the framework of extractive approaches most systems are based on unsupervised learning models. This is the case of Latent Semantic Analysis (LSA) [7], or graph-based [4]. Other systems are based on supervised methods such as Recurrent Neural Networks [3], Conditional Random Fields (CRFs) [13], or Support Vector Machine (SVM) [5].

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 86–93, 2018.
https://doi.org/10.1007/978-3-319-99579-3_10

The organization of evaluation competitions has been an important help for the development of this area. This is the case of DUC[1] and TAC[2] conferences. They have become a forum to compare the different approaches. To do this, some evaluation corpora have been developed that can be used not only for test purposes but also for training models. Some of the most popular corpora in summarization tasks are the corpus used in DUC and the CNN/DailyMail corpus. This last corpus has widely used for learning models in Neural Networks approaches [3].

Other authors have explored the summarization considering audio documents as input [6]. This task has the additional problems of dealing with different kinds of errors, as speech recognition errors and errors in punctuation of sentences. Moreover, some expressions that appear due to spontaneous speech characteristics must be specifically processed since they could be not relevant for the summary.

In this work, we present an extractive approach to document summarization based on unsupervised techniques, in particular Latent Dirichlet Allocation (LDA) [2]. This approach can be considered as topic-based because some topics can be automatically detected and used to determine the most salient sentences according to the topics that appear in the document. Another issue of this work is that we have addressed the problem of summarization of TV programs, in particular a magazine of news. Some characteristics of this task generate specific challenges to the summarization problem. Apart from the speech recognition problems, that are not considered in this work, the most interesting problem is that this kind of programs have a very variable structure, and usually many topics of different duration are present in them. We have studied two strategies of summarization: in the first one, the transcription of the program is the input to the summarization system, and in the second one, a preprocess of segmentation of the program is done, and from the concatenation of the summaries of each segment the final summary is obtained. We have performed some experiments on Spanish TV programs in order to study the behavior of the proposed techniques.

The paper is organized as follows. In Sect. 2, the different methodologies developed are described. In Sect. 3, a description of the system architecture is presented. In Sect. 4, we show the characteristics of the corpus. In Sect. 5, we present the experimental results, and in Sect. 6, the conclusions and future works are presented.

2 System Description

Given a document, considered as a set of sentences, the objective of an extractive summarization technique consists of assigning weights to the sentences, that represent the relevance of them. From this ranked set of sentences the system selects the first ones in order to build the summary.

[1] https://duc.nist.gov/.
[2] https://tac.nist.gov/.

2.1 Latent Semantic Analysis

Many unsupervised summarization systems are based on LSA. This technique permits to extract the representative sentences for the automatically detected topics in the documents. This is done applying the singular value decomposition of word-sentences matrices. That is, given the word-sentence matrix C the Singular Value Decomposition generates the U, Σ, and V^T matrices, where V^T represents the association of underlying topics to sentences.

$$C = U\Sigma V^T$$

From this decomposition there are different ways of assigning weights to sentences and then selecting those ones to appear in the summary. Some of them are based on the most salient sentence for each topic, others are based on the combination of the results of the matrix decomposition. We have chosen the Cross method that permits to extract more than one sentence associated to the most important topics [12].

2.2 Latent Dirichlet Allocation

Another way for discovering hidden topics in documents is the LDA approach. This methodology has been successfully used for topic identification, and can also be used for summary purposes. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes a generative process for each document in a corpus, given the a-priori parameters α and β, that characterize the probabilistic distributions. It assumes that for each word in a document, a topic is chosen given the multinomial distribution of topics, and then a word is chosen given multinomial probability of words conditioned by the selected topic. In order to use LDA, it is necessary to compute the posterior distribution of the hidden variables given a document, and to do this, one of the most popular approach is the Gibbs Sampling. Once the process is done for a fixed number of topics, two matrices are obtained: one of them represents the probability that a concept appears in a document, and the other one represents the probability that a word belongs to a topic (word-topic matrix).

Once these matrices have been obtained, we used the word-topic matrix to assign a weight to each word in a sentence. From this information we obtain a sentence-topics matrix that is the input to an adaptation of the Cross method used in the LSA approach.

2.3 Document Segmentation

Sometimes, as in our case, the documents to be summarized are long and heterogeneous, that is, they are composed by different sections, each one focused on a different subject. For this reason it could be convenient to split the document in different pieces, that is know as topic segmentation.

The approach that we have developed consists of obtaining vector representations of two consecutive paragraphs and defining a distance between vectors to decide if they belong to the same or to different topics. Then, an overlapped sliding window of paragraphs across the document provides the distances between two pairs of consecutive paragraphs. That is, we calculate at the end of each sentence the distance between the previous n sentences and the following n sentences. The length of the sliding window is experimentally determined.

In order to represent the paragraphs a semantic-based approach was done, in particular a Word2vec representation [10]. To do this, it was necessary to learn the Word2vec values from a large corpus. This was done from Wikipedia articles in Spanish.

Once the word representation was obtained, the way to represent the paragraphs was done by the addition of vectors of the words contained in them. The measure used to determine the distance between consecutive paragraphs was the cosine distance.

3 System Architecture

We have explored different approaches to the problem of summarization. Figure 1 shows the architecture of the first system. In it, the documents are the input to the LSA or LDA processes, and the matrix obtained is the input for the Cross method process.

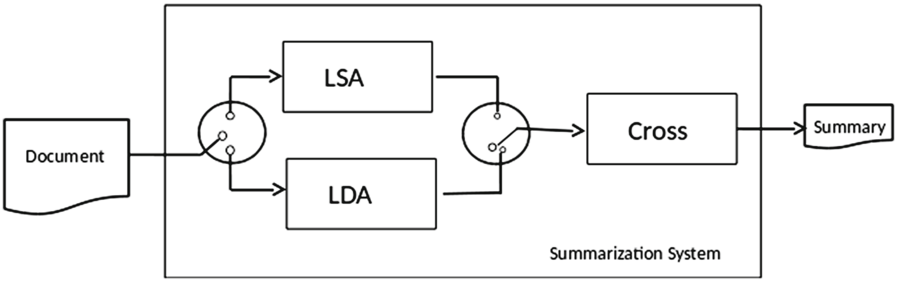

Fig. 1. Architecture of the system.

Figure 2 shows the architecture for the summarization system when a previous phase of topic segmentation is performed. That is, first of all, the documents are segmented, and each segment is summarized. Then, a concatenation of this topic-dependent summaries is performed in order to generate the final summary.

4 Corpus Description

The corpus consists of seven Spanish TV programs of news including some miscellaneous topics, such as music, gastronomy, culture, etc. We used the correct

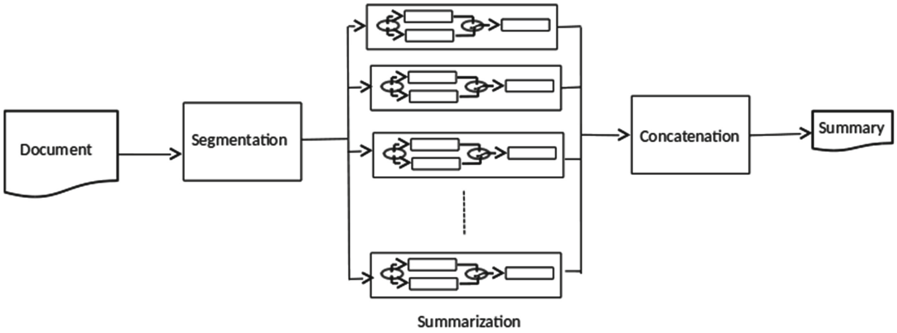

Fig. 2. Architecture of the system with a previous topic segmentation phase.

transcriptions of the speech, in particular the screenplay of the program presenter. It should be noted that the structure of these programs is very heterogeneous. Sometimes a sequence of short news, one or two sentences, of different topics is followed by a long sequence of sentences related to one topic (for example a musical group that presents a new disc, even including interviews with the musicians). Some characteristics of this corpus are shown in Table 1. In order to evaluate the results, a summary of a 20% of the original document was performed for each document. They were manually built by an expert.

Table 1. Corpus characteristics.

Total number of words	27,881
Average number of words per TV program	3,983
Number of words of the shortest TV program	2,924
Number of words of the longest TV program	4,980

5 Experiments

Two series of experiments were done. The first one consisted in the application of both methodologies, LSA and LDA to the set of documents, and the second one was the application of the same methodologies with the previous topic segmentation process.

We have used different ROUGE [8] measures to evaluate the summaries. The ROUGE metrics include: the ROUGE-n that measures the overlap of n-grams between the system and reference summaries, the ROUGE-L based on the Longest Common Subsequence (LCS), the ROUGE-W that is a Weighted LCS-based statistic, the ROUGE-S that is a skip-bigram based co-occurrence statistic, and finally, the ROUGE-SU that is a skip-bigram plus unigram-based co-occurrence statistic. The most widely used in the literature are the ROUGE-1, ROUGE-2, and the ROUGE-L.

The results of applying LDA and LSA directly to the transcriptions of the programs are shown in Tables 2 and 3 respectively. Results show that both methods have a good behavior and there is not a relevant difference between them. This can be explained by the fact that both approaches are based on the underlying topics of the documents, although each one of them has its particular way to model the semantics of the document.

Tables 4 and 5 show the results when a previous segmentation was done. The p_k value [1] of the segmentation was 0.59. It should be noted that the systems with a previous segmentation do not outperform the direct application of the proposed methodologies to the whole document. This could be explained by the fact that the topic segmentation approach is based on a decoupled architecture. That kind of decoupled architecture is very sensitive to the errors in the first phase of the process. This way the errors are transmitted to the following phases, the summarization in our case.

Table 2. Evaluation using LDA.

	Recall	Precision	F_1
ROUGE-1	0.57134	0.59537	0.58298
ROUGE-2	0.28718	0.29915	0.29299
ROUGE-3	0.22941	0.23884	0.23399
ROUGE-4	0.21471	0.22352	0.21899
ROUGE-L	0.53478	0.55706	0.54558
ROUGE-W-1.2	0.13903	0.27932	0.18561
ROUGE-S*	0.29909	0.32546	0.31145
ROUGE-SU*	0.29976	0.32615	0.31213

Table 3. Evaluation using LSA.

	Recall	Precision	F_1
ROUGE-1	0.58019	0.60525	0.59232
ROUGE-2	0.27962	0.29257	0.28588
ROUGE-3	0.20853	0.21844	0.21333
ROUGE-4	0.18838	0.19743	0.19275
ROUGE-L	0.52826	0.55124	0.53938
ROUGE-W-1.2	0.13183	0.26544	0.17612
ROUGE-S*	0.30823	0.33603	0.32124
ROUGE-SU*	0.30890	0.33672	0.32192

Table 4. Evaluation using LDA when a previous topic segmentation is done.

	Recall	Precision	F_1
ROUGE-1	0.51899	0.54040	0.52937
ROUGE-2	0.22402	0.23387	0.22879
ROUGE-3	0.16544	0.17291	0.16905
ROUGE-4	0.15117	0.15808	0.15452
ROUGE-L	0.48046	0.50050	0.49017
ROUGE-W-1.2	0.12027	0.24191	0.16062
ROUGE-S*	0.25231	0.27433	0.26264
ROUGE-SU*	0.25297	0.27501	0.26331

Table 5. Evaluation using LSA when a previous topic segmentation is done.

	Recall	Precision	F_1
ROUGE-1	0.51915	0.54059	0.52954
ROUGE-2	0.22379	0.23363	0.22856
ROUGE-3	0.16549	0.17296	0.16911
ROUGE-4	0.15133	0.15825	0.15468
ROUGE-L	0.48154	0.50137	0.49114
ROUGE-W-1.2	0.11991	0.24106	0.16012
ROUGE-S*	0.25372	0.27567	0.26401
ROUGE-SU*	0.25437	0.27635	0.26468

6 Conclusions

In this paper we have presented an approach to summarization of Spanish TV programs. It is based on unsupervised methods, and it is specially oriented to documents with heterogeneous structures, that is, documents that contain many topics with very different durations. Two approaches based on underlying topic detection have been explored. The first one consists in the application of the methods directly to the document and the second one has a previous phase of topic segmentation. Results show that both approaches provide good results, and they have a similar behavior.

As future work, we will try to improve the segmentation based approach developing some mechanisms to transmit more than one segmentation hypothesis to the summarization phase. This way, the errors generated by the first phase could be recovered during the summarization process. It can be also interesting to develop another way to combine the summaries of the detected segments, instead of a straight forward concatenation of them.

Acknowledgments. This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC: Affective Multimedia Analytics with Inclusive and Natural Communication (TIN2017-85854-C4-2-R).

References

1. Beeferman, D., Berger, A., Lafferty, J.: Statistical models for text segmentation. Mach. Learn. **34**(1), 177–210 (1999)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). http://dl.acm.org/citation.cfm?id=944919.944937
3. Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, 7–12 August 2016, Berlin, Volume 1: Long Papers (2016)
4. Erkan, G., Radev, D.R.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res. **22**(1), 457–479 (2004)
5. Fuentes, M., Alfonseca, E., Rodríguez, H.: Support vector machines for query-focused summarization trained and evaluated on pyramid data. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL 2007, pp. 57–60. Association for Computational Linguistics, Stroudsburg (2007). http://dl.acm.org/citation.cfm?id=1557769.1557788
6. Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C.: Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Trans. Speech Audio Process. **12**(4), 401–408 (2004)
7. Gong, Y., Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 19–25. ACM, New York (2001). https://doi.org/10.1145/383952.383955
8. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: MarieFrancine Moens, S.S. (ed.) Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81. Association for Computational Linguistics, Barcelona (2004)
9. Lloret, E., Palomar, M.: Text summarisation in progress: a literature review. Artif. Intell. Rev. **37**(1), 1–41 (2012)
10. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
11. Nallapati, R., Zhai, F., Zhou, B.: Summarunner: a recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, , San Francisco, 4–9 February 2017, pp. 3075–3081 (2017)
12. Ozsoy, M.G., Cicekli, I., Alpaslan, F.N.: Text summarization of turkish texts using latent semantic analysis. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010, pp. 869–876. Association for Computational Linguistics, Stroudsburg (2010). http://dl.acm.org/citation.cfm?id=1873781.1873879
13. Shen, D., Sun, J.T., Li, H., Yang, Q., Chen, Z.: Document summarization using conditional random fields. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI 2007, pp. 2862–2867 (2007)
14. Tur, G., De Mori, R.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Wiley, New York (2011)