LIA-RAG: a system based on graphs and divergence of probabilities applied to speech-to-text summarization

Elvys Linhares Pontes
Universidade Federal do Ceará, Campus Sobral
Fortaleza, Brasil
elvyslpontes@gmail.com

Juan-Manuel Torres-Moreno
LIA / UAPV
BP 1228, 84911 Cedex 9
Avignon, France
and École Polytechnique de Montréal (Canada)
juan-manuel.torres@univ-avignon.fr

Andréa Carneiro Linhares
Universidade Federal do Ceará, Campus Sobral
Fortaleza, Brasil
and LIA / UAPV France
andrea.linhares@ufc.br

Abstract

This paper aims to introduces a new algorithm for automatic speech-to-text summarization based on statistical divergences of probabilities and graphs. The input is a text from speech conversations with noise, and the output a compact text summary. Our results, on the pilot task CCCS Multiling 2015 French corpus are very encouraging.

Keywords: Automatic Text Summarization, Jensen-Shannon's divergence of probabilities, Speech-to-text summarization, Graph model.

1 Introduction

Nowadays, a lot of information is daily generated. It is necessary to have available memory storage because each datum must be processed and the information contained therein analyzed. The manual analysis is impossible because it is necessary a huge number of persons to analyze this information in an available time. The summary is a short text with main ideas of original text (Torres-Moreno, 2014) and reduces the read time to analyze these data.

Audio is widely used in daily life on the radio and on the internet, in news, interviews and conversations. A Call Centre Conversation creates a lot of conversations every day. These centers has issues and tasks. It is essential the control of the discussed topics and the results obtained by customers in these calls. One way to analyze and accelerate the data processing is speech summarization, that is different from traditional text summarization because there are other problems in these texts as speech errors, sentences of different sizes and colloquialisms.

"Multiling is a community-driven initiative for benchmarking multilingual summarization systems, nurturing further research, and pushing the state-of-the-art in the area"[1]. The MultiLing 2015 initiative features the following tasks: Multilingual Multi-document Summarization, Multilingual Single-document Summarization, Online Forum Summarization and Call Centre Conversation Summarization (CCCS). The CCCS pilot task consists in *"creating systems that can analyze call centres conversations and generate written summaries reflecting why the customer is calling, how the agent answers that query, what are the steps to solve the problem and what is the resolution status of the problem"* (Favre et al., 2015).

We developed the LIA-RAG summarization system based on the RAG system (Pontes et al., 2015), coupled with some post-processing rules in order to generate a final summary. LIA-RAG uses a graph model to analyze and verify a set of documents (e.g., the conversation transcription) for MultiLing'15 CCCS pilot task. LIA-RAG creates a summary computing the relevance of the words and the similarity among the sentences. The system uses a simple post-processing to improve the quality of the final summary.

The rest of the paper is organized as follows: section 2 describes related work on automatic summarization of texts and conversations. Sections 3 and 4 analyze the graph model and the system used in this work. Section 5 describes the results obtained for Multiling/DECODA French corpus and section 6 concludes this work.

[1] http://multiling.iit.demokritos.gr/pages/view/1517/multiling-2015-call-for-participation

2 Related Works

Automatic Text Summarization (ATS) aims to creates a summary containing the main ideas of a textual document (Mani and Mayburi, 1999; Mani, 2001). The summary can be an extraction or abstraction of a single document or multi-document. The extraction process identifies the most informative sentences of a document and creates a summary by assembling of these sentences (Luhn, 1958; Torres-Moreno, 2014). Extraction may be guided (by a query). In this case, the algorithm selects the most relevant information follow a particular topic. The abstraction algorithms create new (or reformulate) sentences from original texts (Seno, 2010; Seno and Nunes, 2008) and the extraction methods use the key sentences of texts (Barzilay and McKeown, 2005; Torres-Moreno, 2014).

Works about abstraction usually uses syntactic and semantic knowledge of a language to create the summary. This procedure verifies the best construction of a sentence (Barzilay et al., 1999). This type of summarization uses fusion to help the review of information. (Seno, 2010) proposed a method to fusion similar sentences in Brazilian Portuguese based on a symbolic and domain-independent approach. This method allows the fusion by union and by intersection of a document cluster. Fusion by union preserves the overall message of the cluster while fusion by intersection analyses the redundant information considered most important in the cluster. (Seno and Nunes, 2008) described how to identify common information between sentences in Brazilian Portuguese using lexical knowledge, syntactic and semantic rules of paraphrasing.

(Jorge et al., 2010) developed a summarizer system based on the CST model (Cross-document Structure Theory). The system proposed analyses redundancy and contradiction among different information sources in Brazilian Portuguese.

(Barzilay et al., 1999) developed a method to generate automatic summaries by identifying and synthesizing similar elements in a cluster of documents. This method creates the summary based on similarity between the sentences and topic. (Barzilay and McKeown, 2005) described an approach to fusion sentences through the text-to-text technique, to synthesize repeated information from multiple documents. This method uses a syntactic alignment in sentences to identify common information. After the identification step, sentences are processed and a new text is generated with the same content.

A way to calculate the similarity between sentences is to use co-occurrence of words. (He et al., 2008) proposed a fusion method using similarity metrics, co-occurrence skip-bigram and information density to evaluate sentences and to select the most relevant ones. (Hennig and Albayrak, 2010) developed a multi-document model to summarize by analyzing the co-occurrence of sentence-term and sentence-bigram using the Jensen-Shannon (JS) divergence.

Another method to obtain relevant sentences uses compression, as reported in (Pitler, 2010). Pitler uses approaches based on syntactic trees, sentences and discourse. (Filippova, 2010) describes a multi-sentence compression method using a word-based graph.

The summarization by extraction does not have the same quality as the summaries produced by abstraction because it uses surface methods based on statistical calculations to verify the sentence relevance. However, the extraction is general and do not require deep analysis of the language (Barzilay and McKeown, 2005; Pontes et al., 2014).

(Pontes et al., 2014) use Graph theory concomitant with JS divergence to create multi-document summaries by extraction. Their system describes a text model as a graph where the sentences are represented by vertices and the edges connect two similar sentences. Their approach calculates the stable set of the graph aiming creating the summary containing sentences with general information of the cluster and without redundancy. (Linhares et al., 2013) model the text as graph model and use a heuristic (greedy algorithm) to obtain the relevant sentences in the text.

The speech summarization task is more complex and it involves other problems. It is more difficult to identify utterance boundaries because it may be fragmented, contain disfluencies and also because speech recognition introduces errors. Meetings involve multi-party conversation with overlapping speakers. The language used is informal and utterances tend to be partial, fragmentary, ungrammatical and include many ellipses and pronouns. However, the speech signal may provides additional information that emphasizes a piece of text as prosody (Murray et al., 2005).

(Mckeown et al., 2005) described some ways to use a text summarization as a speech summarization. They described some work about summarization of broadcast news and meetings. (Murray et al., 2005) analyzed extractive summarization of multiparty meetings. They described Maximal Marginal Relevance and Latent Semantic Analysis to create the summary based on prosodic and lexical features.

3 Modeling the problem

This paper aims to design a system to summarize several documents by extraction its most important sentences. Statistical techniques were used to build a language independent system. The proposed methods are based on a specific preprocessing of words, a weighting function of sentences and a bag-of-words model to represent the text content.

This model uses K matrices represented by $S^K_{[m \times n]}$ and constructed from K documents, where m_a is the number of sentences and n_a is the number of distinct words in the document a ($a \in K$). The cell s^a_{ij} of the matrix represents the frequency of word j in the sentence i (FP_{ij}) of the document a. This stage was constructed using the libraries and algorithms from Cortex summarization system (Torres-Moreno et al., 2002; Torres-Moreno et al., 2001).

$$S^a = \begin{pmatrix} s^a_{11} & s^a_{12} & \dots & s^a_{1n} \\ s^a_{21} & s^a_{22} & \dots & s^a_{2n} \\ \vdots & \vdots & & \vdots \\ s^a_{m1} & s^a_{m2} & \dots & s^a_{mn} \end{pmatrix}, a \in K$$
$$s^a_{ij} = \begin{cases} FP_{ij}, & \text{if } \exists \text{ word j in sentence i} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

3.1 Jensen-Shannon divergence

We use Jensen-Shannon (JS) divergence to measure the similarity between sentences. Let w be a words' set in P and Q. P and Q represent the probability distribution between two objects: two individuals sentences or a sentence and a set of sentences. The divergence will then calculated among these two objects. The JS divergence is symmetric and provides a stable way to measure the difference between two distributions (equation 2).

$$\begin{aligned} D_{JS}(P||Q) &= \frac{1}{2}\sum_{w \in W}\left[P_w \log\left(\frac{2 \times P_w}{P_w + Q_w}\right)\right. \\ &+ \left. Q_w \log\left(\frac{2 \times Q_w}{P_w + Q_w}\right)\right] \end{aligned} \tag{2}$$

The JS divergence value ranges from $[0, \infty+)$. It is closer to zero when the distributions are similar and they differ in another case.

In the case there is a word in a sentence that is missing in another one, a smooth (different weighting) will be used to avoid null values and have a smoother distribution (Hiemstra, 2009). If a word w is not present in the sentence Q, then the smooth is calculated by the equation 3, where $\beta = 1.5 \times voc$, which voc is the number of distinct words in R, γ is the variable that controls the relevance of the missing word in the sentence and N is the number of words in R (Louis and Nenkova, 2013).

$$Q_w = \left(\frac{P_w + \gamma}{N + \gamma \times \beta}\right) \tag{3}$$

3.2 Term Frequency-Inverse Sentence Frequency (TF-ISF)

One way to verify the initial relevance of a word and a sentence to the text is through the TF-ISF. This metric is based on term frequency in the text and it is calculated by the equation 4.

$$tf\_isf(w) = tf(w) \times \log\left(\frac{n}{n_w}\right) \tag{4}$$

where $tf(w)$ is frequency of term w, n is total number of documents and n_w is number of documents that contain the term w.

4 The LIA-RAG system

In general lines, a text consists of several sentences with different topics. The text can be divided into several groups and each of them describes one step/idea in the text. If a group is large, then it is relevant to the text. It is possible to choose the sentences of the largest group and obtains the most relevant content.

The main ideas of a text are generally analyzed and discussed several times. The vertices with higher degree have more similar sentences and then, are important to the text. However, it is not necessary to have a lot of similar sentences to be a relevant one.

Résumeur Audio-texte à base de Graphes (RAG) is a summarizer system by sentence extraction, which selects the main sentences of a text and uses a post-processing to remove some errors and make the text more concise and compact.

4.1 The RAG algorithm

RAG uses Graph theory and divergence metrics to calculate the similarity and to group the sentences. Initially, the system performs a filtering process to remove the brackets. Then, it performs a segmentation, filtering and stemming processes to remove stopwords and reduce the words to their roots. RAG accomplished this preprocessing and matrix transformation based on (Torres-Moreno et al., 2001). It calculates the relevance of each sentence based on TF-ISF metric (equation 4) and removes the less relevant sentences.

The system creates a graph G which each vertex represents a sentence previously selected. The text is analyzed and modeled as a sentence graph (vertices). Based on equation 4, it calculates the similarity between sentences. If the similarity between two sentences is less than 0.16 (threshold obtained by empirical testing), then the system creates an edge between them. So, the vertices with higher degrees have the most relevant content of the text. However, some sentences may have a small degree, but they may contain important information.

RAG combines the TF-ISF and degree sentences to analyze the relevance of them. The relevance of the sentence i is defined by:

$$rel(i) = degree(i) \times tf\_isf(i) \qquad (5)$$

where $degree(i)$ is the degree of vertex i and $rel(i)$ is the relevance of the sentence i. After, the system creates a summary with the higher score sentences, excluding similar (or redundant) sentences based on Dice's coefficient (Bai et al., 2012).

The figure 1 describes the RAG system.

4.2 LIA-RAG: RAG with a specific speech post-processing

The speech recognition process produces a text that contains several grammatical problems (slang, colloquialisms, expressions and speech recognition errors). An extraction summary algorithm selects the relevant sentences, however the sentences may have some grammatical problems. So, it is necessary to perform a treatment of this summary.

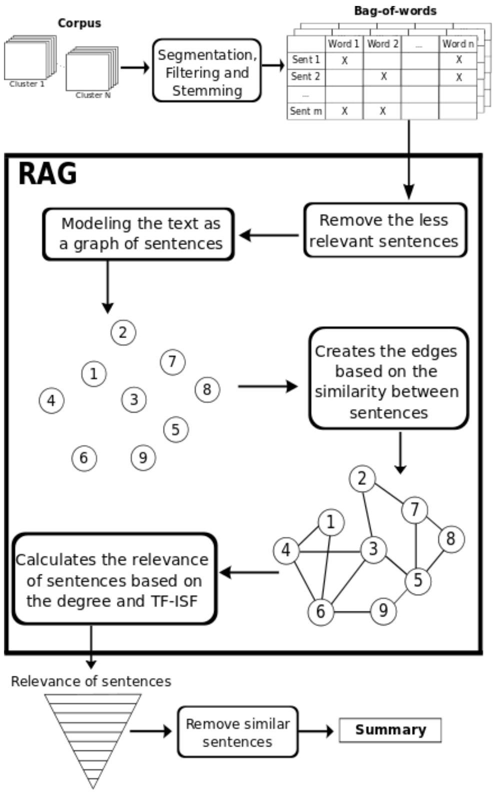
Corpus
Cluster 1
Cluster N
Segmentation, Filtering and Stemming
Bag-of-words
Word 1
Word 2
...
Word n
Sent 1
Sent 2
...
Sent m
X
RAG
Remove the less relevant sentences
Modeling the text as a graph of sentences
1
2
3
4
5
6
7
8
9
Creates the edges based on the similarity between sentences
Calculates the relevance of sentences based on the degree and TF-ISF
Relevance of sentences
Remove similar sentences
Summary

Figure 1: Architecture of the RAG system.

The main analyzed aspects in this process are:

- Colloquialisms,
- Speech expressions and
- Dates.

LIA-RAG system receives the summary as an input. In this input, some speech expressions are used to connect ideas or concepts in oral conversations. LIA-RAG removes these expressions, because often they are incorrectly transcripted (a noise source). Also, the system eliminates several colloquialisms and the duplicated words. The system replaces some mistaken words by its correct form. The figure 2 shows the architecture of the LIA-RAG system.

5 Results

The tests were carried on a computer with i5@2.6 GHz processor and 4 GB of RAM on GNU/Linux

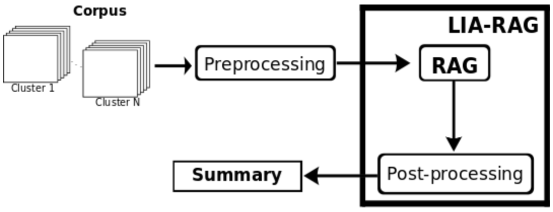

Figure 2: Architecture of the LIA-RAG system.

Debian 64-bit operating system. The algorithms of RAG were implemented using the Perl language.

We used the French DECODA corpus (Bechet et al., 2012). The systems have to generate textual summaries with the main idea of each conversation belonging to the corpus. *"The conversations topics range from itinerary and schedule requests, to lost and found, to complaints (the calls were recorded during strikes)"*(Favre et al., 2015). Each summary has 7% of the number of words of each conversation transcription. We compared LIA-RAG and RAG systems with two baseline systems (random and first lead base).

In order to evaluate the quality of the summaries, Multiling CCCS used the system Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[2], which determines the quality of an automatic summary based on the intersection of the n-grams of a candidate summary and the n-grams of a set of reference summaries. More specifically, we used ROUGE-N and ROUGE-SU measures. ROUGE-N, N $\in$ $[1, 2]$. ROUGE is an n-gram recall measure (Lin, 2004)[3]. The values of these metrics belongs to $[0, 1]$, 1 for the best result.

The table 1 shows the results obtained using the systems over the training corpus. This corpus contains 50 conversations transcription with 23,363 words and 115 summaries. Both versions of RAG provided the best results. The RAG system identified the main sentences discussed in conversations. However, the errors and speech expressions decreased the informativeness. The post-processing of LIA-RAG allowed to improve the results. This process reduces errors and generates a more informative and concise summary.

Systems	ROUGE-1	ROUGE-2	ROUGE-4
LIA-RAG:1	**0.1893**	**0.0628**	**0.0683**
RAG	*0.1833*	*0.0614*	*0.0654*
Base-first	0.1578	0.0556	0.0583
Base-rand	0.1170	0.0310	0.0371

Table 1: Evaluation of training corpus.

The French test corpus has 100 conversations transcription with 42,130 words and 212 summaries. The ROUGE-2 official performance for the systems participating to CCCS pilot task is showed in table 2 (Favre et al., 2015). The LIA-RAG system obtained the best results.

Systems	ROUGE-2
LIA-RAG:1	**0.037**
NTNU:1	0.035
NTNU:3	0.034
NTNU:2	0.027

Table 2: Evaluation of test corpus.

6 Conclusion and perspectives

Divergence of probabilities in a graph model to extract key sentences in French speech-to-text summarization was very interesting. LIA-RAG system uses very few language resources (stopwords and stemming) and has achieved good results. Nevertheless, the system is easily adaptable to other languages with only some modifications in the preprocessing stage.

An interesting perspective of this work consists in the utilization of the speech TAGs markers to improve the computation of the sentences score. In addition, it is necessary to improve the post-processing in order to increase the quality of the final summary. Finally, the verification of the grammaticality and readability of the extracted key sentences can help to produce more realistic abstracts.

Acknowledgments

This project was partially founded by a scholarship from FUNCAP-CE (Brazil).

References

Ming-Hong Bai, Yu-Ming Hsieh, Keh-Jiann Chen, and Jason S. Chang. 2012. Domcat: A bilingual concordancer for domain-specific computer assisted translation. In *Proceedings of the ACL 2012 System*

[2] The options for running ROUGE 1.5.5 are -a -l 10000 -n 4 -x -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0

[3] `http://www.berouge.com/Pages/default.aspx`

Demonstrations, pages 55–60, Jeju Island, Korea, July. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31(3):297–328, September.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL'99, pages 550–557, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot. 2012. Decoda: a call-centre human-human spoken conversation corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015)*. Prague, Czech Republic.

Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 322–330, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tingting He, Fang Li, Wei Shao, Jinguang Chen, and Liang Ma. 2008. A new feature-fusion sentence selecting strategy for query-focused multi-document summarization. In Cheolyoung Ock, JeongYong Byun, YuDe Bi, and Hongfei Lin, editors, *ALPIT*, pages 81–86. IEEE Computer Society.

L. Hennig and S. Albayrak. 2010. Personalized multi-document summarization using n-gram topic model fusion. In *Proceedings of LREC'10, 1st Workshop on Semantic Personalized Information Management (SPIM 2010)*, pages 28–34, Valletta, Malta. European Language Resources Association.

D. Hiemstra. 2009. Probability smoothing. In *Encyclopedia of Database Systems, pp. 2169-2170, Springer*.

Castro Jorge, Maria Lucía del Rosario, and Thiago Alexandre Salgueiro Pardo. 2010. Experiments with cst-based multidocument summarization. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-5, pages 74–82, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Andréa Carneiro Linhares, Juan-Manuel Torres-Moreno, and Javier Ramirez. 2013. Résumé automatique 4-lingue avec un algorithme glouton. In *ROADEF'13*.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April.

I. Mani and M. Mayburi. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Publishing Co.

Kathleen Mckeown, Julia Hirschberg, Michel Galley, and Sameer Maskey. 2005. From text to speech summarization. In *ICASSP. 2005. Philadelphia, PA*.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *in Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596.

E. Pitler. 2010. Methods for sentence compression. Technical report, University of Pennsylvania, Technical Report MS-CIS-10-20.

Elvys Linhares Pontes, Andréa Carneiro Linhares, and Juan-Manuel Torres-Moreno. 2014. Sasi: sumarizador automático de documentos baseado no problema do subconjunto independente de vértices. In *Proceedings of the XLVI Simpósio Brasileiro de Pesquisa Operacional*.

Elvys Linhares Pontes, Juan-Manuel Torres-Moreno, and Andréa Carneiro Linhares. 2015. Rag : un système de résume automatique à base de graphes.

Eloize Rossi Marques Seno and Mariadas Graças Volpe Nunes. 2008. Some experiments on clustering similar sentences of texts in portuguese. In António Teixeira, VeraLúciaStrube de Lima, LuísCaldas de Oliveira, and Paulo Quaresma, editors, *Computational Processing of the Portuguese Language*, volume 5190 of *Lecture Notes in Computer Science*, pages 133–142. Springer Berlin Heidelberg.

Eloize Rossi Marques Seno. 2010. *Um método para a fusão automática de sentenças similares em português*. Ph.D. thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

Juan-Manuel Torres-Moreno, Patricia Velázquez-Morales, and Jean-Guy Meunier. 2001. Cortex : un algorithme pour la condensation automatique des textes. In *ARCo'01*, volume 2, pages 365–366. Lyon, France.

Juan-Manuel Torres-Moreno, Patricia Velázquez-Morales, and Jean-Guy Meunier. 2002. Condensés de textes par des méthodes numériques. In *JADT*, volume 2, pages 723–734.

Juan-Manuel Torres-Moreno. 2014. *Automatic Text Summarization*, volume 1. John Wiley & Sons.