Summarizing Spoken Documents: avoiding distracting content

Ricardo Ribeiro[1,3] and David Martins de Matos[2,3]

[1] Instituto Universitário de Lisboa (ISCTE-IUL)
[2] Instituto Superior Técnico - Universidade Técnica de Lisboa
[3] L2F - INESC ID Lisboa

Abstract. Driven by a cognitive perspective of the human summarization process, we address the problem of assessing the most relevant information of a single spoken language document, by minimizing the influence of distracting content, of which passages particularly affected by spoken language-related problems are major representatives. Two different approaches are considered. One, based only on the input source to be summarized, consists in a centrality-based relevance model for automatic summarization that uses support sets to better estimate the relevant content. Geometric proximity is used to compute semantic relatedness. Relevance is determined by considering the whole input source, and by assuming that information sources to be summarized comprehend different topics. A thorough evaluation shows statistically significant improvements over previous approaches. The other mimics the natural human behavior, in which information acquired from different sources is used to build a better understanding of a given topic. Information from different types of sources and of the same type is explored. A multi-document summarization framework provides the means to assess the relevant content. A perceptual evaluation shows that mixing information leads to considerably better results, both in terms of informativeness and readability. Concerning the use of information of the same type, results show that background information of the same topic clearly improves the detection of the most important content.

Keywords: Automatic summarization, Speech summarization, Text summarization, Centrality-as-relevance, Latent semantic analysis

1 Introduction

Speech summarization, and especially speech-to-text summarization, is of great importance in the nowadays context: one needs only to observe the amount of multimedia content currently produced to become aware of that importance. But in addition to that, speech specific nature also motivates the need of summarization: (summarized) text is easier to scan than speech, which can be relevant in several situations (voicemails, headline browsing, media monitoring, etc).

Human summarization is a knowledge-based task, characterized by the following aspects [4–6]: both general and specialized knowledge are used to assess

importance when analyzing an input source—*knowledge-based text analysis*; that knowledge is combined when inspecting each passage as a whole and, at the same time, detecting specific relevant information within it—*combined online processing*; passages not relevant for the utilization context of the summary are skipped and *keywords* (terms) are *the most obvious cues for attracting attention—task orientation and selective interpretation*. Pinto Molina [17] also describes the human abstraction process as a knowledge-based process, characterized by three main stages: understanding; analysis; and, synthesis. The objective of the first stage, understanding, is to achieve comprehension of the information source. This stage is a first step towards the subsequent analytical process. The analysis stage starts by a selection step which consists in eliminating repeated, not very relevant and irrelevant content, and is followed by an interpretation step guided by the summary objective. The synthesis stage aims at producing the summary. Endres-Niggemeyer [4] also explains that the information reduction techniques in the summarization process are quite close to the discourse understanding process, which, at a certain level, works by applying rules that help uncovering the macrostructure of the discourse. One of these rules, *deletion*, is used to eliminate from the understanding process propositions that are not relevant to the interpretation of the subsequent ones. This means that is common to find, in the input sources to be summarized, lateral issues or considerations that are not relevant to devise the salient information (discourse structure-based summarization builds on the relevance of nuclear segments [13, 25]), and that may affect summarization methods by leading to the selection of inadequate content. In the specific case of speech summarization, passages particularly affected by spoken language-related problems are major representatives of distracting content.

Our analysis of speech-to-text extractive summarization methods using the Portuguese language [20] underlined two important aspects: (i) human summarizers prefer well-formed passages with a low word error rate; (ii) human summarizers tend to ignore segment boundaries, joining segments that were only relevant if considered together. This shows that the human summarizers are clearly affected by spoken language-related problems, like speech recognition errors, disfluencies, and passage segmentation problems, and that they adopt strategies to minimize the influence of such problems. Computational approaches also suffer the influence of such problems: note that words are natural features for summarization models (which are negatively influenced by recognition errors) and segment definition clearly influences summarization results [15]. Moreover, given that most of the work focus on extractive approaches, resulting summaries may contain incomprehensible content. Speech-specific information (for example, acoustic/prosodic features [14] or recognition confidence scores [27]) have been used to cope with speech-related issues. To avoid the distracting content, we explore two different approaches: a new summarization model that diminishes the influence of lateral topics/noisy content; and, the use of additional related information sources to improve the assessment of the relevant content.

The summarization model we propose is generic, language- and domain-independent, and has low computational and linguistic requirements. Based on

the centrality-as-relevance paradigm, the model relies on the use of support sets to better estimate the relevant content. Building on the ideas of Ruge [24], *[...] the model of semantic space in which the relative position of two terms determines the semantic similarity better fits the imagination of human intuition [about] semantic similarity [...]*, semantic relatedness is computed by geometric proximity. Centrality (relevance) is determined by considering the whole input source (and not only local information), and by taking into account the existence of minor topics/lateral subjects/noisy content in the information sources to be summarized. The method consists in creating, for each passage of the input source, a set containing only of the most semantically related passages (support set). Then, the determination of the most relevant content is achieved by selecting the passages that occur in the largest number of support sets.

Our other strategy to minimize the impact of distracting content is to explore the use of additional related information to cope with the difficulties posed by speech-to-text summarization. By employing contextual (prior) information, we study how to improve the assessment of the relevant content of new input sources. This informed approach to relevance assessment is envisioned either by including related solid background information from a different medium, especially if less prone to spoken language related problems (e.g., a textual source), or by using the same medium but, still using multiple sources, introducing the idea of topic evolution through time, in the summarization process. Summary generation is done using a multi-document summarization framework, Latent Semantic Analysis [10, 9], which can be used to combine multiple information sources in order to produce a summary driven by a single spoken language document.

The two following sections address our main research questions, the centrality model and the use of additional information sources. The documents ends with pertinent considerations about the impact of this work on the advance of the computational processing of the Portuguese language.

2 Centrality-as-Relevance Summarization Model

A common family of approaches to the identification of the relevant content is the *centrality* family. Although developed in the context of text summarization, current work on speech summarization has focused on improving this type of methods [8] and on its use as baseline [12, 11]. Even in text summarization, the number of up-to-date examples is significant [2, 3, 26]. Centrality-as-relevance methods base the detection of the most salient passages on the identification of the central passages of the input source(s). One of the main representatives of this family is *centroid-based summarization*. Pioneer work (on multi-document summarization) by Radev et al. [18, 19] creates clusters of documents by representing each document as a *tf-idf* vector; the centroid of each cluster is also defined as a *tf-idf* vector, with the coordinates corresponding to the weighted average of the *tf-idf* values of the documents of the cluster; finally, sentences that contain the words of the centroids are the best representatives of the topic of the cluster, thus being the best candidates to belonging to the summary. An-

other approach to centrality estimation is to compare each candidate passage to every other passage and select the ones with higher scores (the ones that are closer to every other passage). A simple approach is to represent passages as vectors using a weighting scheme like *tf-idf*; then, passage similarity can be assessed using, for instance, the cosine, assigning to each passage a centrality score. These scores are then used to create a sentence ranking: sentences with highest scores are selected to create the summary. Examples of relevant work are presented by Erkan and Radev [7] and by Mihalcea and Tarau [16]. A major problem of this relevance paradigm is that by taking into account the entire input source in this manner, either to estimate centroids or average distances of input source passages, we may be selecting extracts that being *central* to the input source are, however, not the most relevant ones. As previously mentioned, in cognitive terms, the summarization process relies on the removal of irrelevant, or of little relevance, information. This means that it is common to find, in the input sources to be summarized, inadequate content, lateral issues, or considerations that are not relevant to devise the salient information, and that may affect centrality-based summarization methods by inducing inadequate centroids or decreasing the scores of more suitable sentences.

We hypothesize that input sources comprehend different topics (lateral issues beyond the main topic, or distracting content, of which passages particularly affected by spoken language-related problems are major representatives) and propose as a possible solution a centrality-based generic relevance model, which is language and domain independent [21]. The model detects the most relevant content of a given information source by creating for each passage a support set consisting only of the most semantically related passages. Then, the estimation of the most relevant content is performed by selecting the passages that occur in the larger number of support sets. This is an important difference from previous centrality models: centrality is influenced by the groups of related passages that are uncovered by the introduction of the support sets layer and not by passages directly. We ground semantic similarity on geometric proximity, exploring how the different distances influence the estimation of the relevant content. The model, thoroughly evaluated using both written text and speech transcriptions, performs consistently better (we report significance levels using adequate statistical tests where appropriated) than previous summarization approaches, including more complex models [11, 12]. The obtained results indicate that our model is robust, being able to detect the most relevant content without specific information of where it should be found and performing well in the presence of noisy input. The proposed model is unsupervised, has low computational requirements, and identifies the most salient passages of an input source, based exclusively on information drawn from the used input source.

3 The Use of Additional Related Information Sources

Our other approach to diminish the influence of distracting content is inspired by the natural human behavior, in which information acquired from different sources

is used to build a better understanding of a given subject. The integration of all sources highlights the most important content of the single information source to be summarized. This idea was explored in two different strategies.

In the first strategy [22], we further advance the goal of minimizing the influence of speech-related phenomena: we select from solid related background information of a different type (e.g. a textual source) passages similar to the ones of the input source we intend to summarize. The selection of information at the passage level means that the selected passages are strongly related to the main information source and can be used to substitute the corresponding noisy passages. In this sense, the additional information is used to improve the assessment of relevant content by reinforcing the most important passages and to improve the quality of the summary by substituting passages affected by recognition errors, disfluencies, and segmentation problems by high quality passages free from such problems. To select the related passages, diminishing the influence of speech-related problems, we propose a **method for selecting related passages based on phonetic information**. To reduce the influence of the speech-related problems, we use the alignment at the phonetic level of SUs of the main information source and sentences of the additional information sources (sentences are selected if they *sound* like the SUs). The alignment costs are based on a model of phone production, that uses several features to define the distance between phones. To decide if a SU is similar to a sentence, we estimate a threshold using the average distance between automatic transcriptions and the corresponding manual transcriptions. We build on the presumption that the average alignment costs between automatic and manual transcriptions are sufficiently similar to the alignment costs between automatic transcriptions and related textual information. This summarization strategy was human evaluated and performed consistently better than the baseline. Human evaluators were asked to select best summaries, and score the informativeness and the readability of five different summaries for each news story (human extracts, human abstract, automatic summaries containing textual passages only; automatic summaries containing textual and transcribed passages; automatic summaries containing transcribed passages only). This enabled us to perform an **analysis of the reaction of the human evaluators to the content of summaries**. Several interesting findings were observed: human extractive summaries were preferred over human abstractive summaries, suggesting that if recognition errors do not affect intelligibility, humans tend to disregard them; the readability of summaries containing both textual and transcribed passages achieved a average score closer to the ones containing only written text passages, while maintaining a low standard deviation; automatic summaries generated using this strategy achieved considerably better results than the baseline concerning informativeness, although with high standard deviation values (this suggests that the inclusion of new content in the summary, although being closely related to the input source, is controversial).

In the second strategy [23], the additional information was selected from the same medium (speech documents, specifically, additional broadcast news stories), and both passages and full documents were explored. The method builds

on the idea of topic evolution through time, so broadcast news programs from the previous days were used as additional information sources. As previously mentioned, summarization is a knowledge-based task, and the idea of incorporating topic-related information sources explores that direction: note that, in this case, there is no substitution of the passages of the main input source, since the relation between SUs of the main input source and SUs from the additional sources is not as strong as in the previous method. There are several approaches that allow the retrieval of topic-related documents. Since this step is based on document similarity, information retrieval techniques can be used to address the problem. Moreover, for specific contexts of application, specific solutions may be selected (as it happens in our case study [1], in which we depend on a module for topic segmentation and indexing). As previously mentioned, in addition to using full stories, we explored a coverage metric to select passages from the topic-related documents. To diminish the influence of the problems that affect speech transcriptions, we explore several term representation strategies, combining local (e.g., frequency and recognition confidence scores) and global weights (e.g., *idf* and *self-information*). To test our method, we selected a corpus composed by excerpts of broadcast news programs where is possible to find topic related news stories in a chronological frame close to the information source to be summarized. As no reference summaries were available and there were several summary alternatives, this approach was automatically evaluated using an information-theoretic measures which does not require reference summaries. Results show that the approaches using additional information sources achieved the best results when employing a global weighting strategy (namely, *idf*). Weighting strategies using recognition confidence scores achieved worse results than similar approaches using term frequency.

In both strategies, the relevant content was estimated using the Latent Semantic Analysis framework.

4 Final Remarks

To the best of our knowledge, this is the first exhaustive study on speech summarization of the Portuguese language. We started by comparing three methods for extractive summarization of Portuguese broadcast news: feature-based, Maximal Marginal Relevance, and Latent Semantic Analysis [20]. The main goal was to understand the level of agreement among the automatic summaries and how they compare to summaries produced by non-professional human summarizers. Beyond the proposed approaches to speech summarization, not language specific, we carried out in-depth evaluations of that methods using the Portuguese language. One of the main dilemas we faced during this work was whether we should build a corpus for summary evaluation consisting of information sources and respective human reference summaries or not. Note that, although we propose language independent methods, we opted to work using the Portuguese language, contributing also to advance the computational processing of Portuguese. And, for that reason, no data collection was available to allow an evaluation similar

to the ones performed in campaigns like the ones of the Document Understanding Conferences or the more recent Text Analysis Conference. Moreover, from our review of the literature, it is possible to notice that research in summary evaluation is an active research area and that there is no established evaluation method that is considered completely adequate by the research community (even considering the wide adoption of ROUGE, both in text and speech summarization). In addition, most evaluation methods require significant human labour. In fact, one the most recent research trends concerning the automatic evaluation of summaries proposes the use of information-theoretic measures, using as reference the information sources themselves. Given these facts, instead of developing a larger collection with several reference summaries, we opted, when possible, to use human evaluation, or to include human participation, in order to attain reliable results (despite the problems that affect human evaluations): this means that our collections were developed in accordance with evaluation setups and are of reduced size when compared to campaign evaluations. Nevertheless, the work here described constitutes a first step to further advance the research on the summarization of Portuguese spoken language documents.

References

1. Amaral, R., Meinedo, H., Caseiro, D., Trancoso, I., Neto, J.P.: A Prototype System for Selective Dissemination of Broadcast News in European Portuguese. EURASIP Journal on Advances in Signal Processing 2007 (2007)
2. Antiqueira, L., Oliveira Jr., O.N., da Fontoura Costa, L., das Graças Volpe Nunes, M.: A complex network approach to text summarization. Information Sciences 179(5), 584–599 (2009)
3. Ceylan, H., Mihalcea, R., Özertem, U., Lloret, E., Palomar, M.: Quantifying the Limits and Success of Extractive Summarization Systems Across Domains. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. pp. 903–911. ACL (2010)
4. Endres-Niggemeyer, B.: Summarizing Information. Springer (1998)
5. Endres-Niggemeyer, B.: Human-style WWW summarization. Tech. rep., University for Applied Sciences, Department of Information and Communication (2000)
6. Endres-Niggemeyer, B.: SimSum: an empirically founded simulation of summarizing. Information Processing and Management 36(4), 659–682 (2000)
7. Erkan, G., Radev, D.R.: LexRank: Graph-based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22, 457–479 (2004)
8. Garg, N., Favre, B., Reidhammer, K., Hakkani-Tür, D.: ClusterRank: A Graph Based Method for Meeting Summarization. In: Proceedings of INTERSPEECH 2009. pp. 1499–1502. ISCA (2009)
9. Gong, Y., Liu, X.: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In: SIGIR 2001. pp. 19–25. ACM (2001)
10. Landauer, T.K., Dumais, S.T.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review 104(2), 211–240 (1997)
11. Lin, S.H., Chen, B.: A Risk Minimization Framework for Extractive Speech Summarization. In: Proceedings of the 48^{th} annual meeting of the Association for Computational Linguistics. pp. 79–87. Association for Computational Linguistics (2010)

12. Lin, S.H., Yeh, Y.M., Chen, B.: Extractive Speech Summarization – From the View of Decision Theory. In: Proceedings of INTERSPEECH 2010. pp. 1684–1687. ISCA (2010)
13. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. The MIT Press (2000)
14. Maskey, S.R., Hirschberg, J.: Comparing Lexical, Acoustic/Prosodic, Strucural and Discourse Features for Speech Summarization. In: Proceedings of the 9^{th} EUROSPEECH - INTERSPEECH 2005 (2005)
15. Maskey, S.R., Rosenberg, A., Hirschberg, J.: Intonational Phrases for Speech Summarization. In: Proceedings of INTERSPEECH 2008. pp. 2430–2433. ISCA (2008)
16. Mihalcea, R., Tarau, P.: A Language Independent Algorithm for Single and Multiple Document Summarization. In: Proc. of the 2^{nd} IJCNLP: Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts. pp. 19–24. Asian Federation of Natural Language Processing (2005)
17. Pinto Molina, M.: Documentary Abstracting: Toward a Methodological Model. Journal of the American Society for Information Science 46(3), 225–234 (1995)
18. Radev, D.R., Hatzivassiloglou, V., McKeown, K.R.: A Description of the CIDR System as Used for TDT-2. In: Proc. of the DARPA Broadcast News Wksp. (1999)
19. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: NAACL-ANLP 2000 Wksp.: Automatic Summarization. pp. 21–30. ACL (2000)
20. Ribeiro, R., de Matos, D.M.: Extractive Summarization of Broadcast News: Comparing Strategies for European Portuguese. In: Matoušek, V., Mautner, P. (eds.) Text, Speech and Dialogue – 10^{th} International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings. Lecture Notes in Computer Science (Subseries LNAI), vol. 4629, pp. 115–122. Springer (2007)
21. Ribeiro, R., de Matos, D.M.: Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. Journal of Artificial Intelligence Research 42, 275–308 (2011)
22. Ribeiro, R., de Matos, D.M.: Multi-source Multilingual Information Extraction and Summarization, chap. Improving Speech-to-Text Summarization by Using Additional Information Sources. Theory and Applications of Natural Language Processing, Springer (2012)
23. Ribeiro, R., de Matos, D.M.: Summarizing Speech by Contextual Reinforcement of Important Passages. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigo, F. (eds.) Computational Processing of the Portuguese Language: 10*th* International Conference, PROPOR 2012, Coimbra, Portugal, April 10-20, 2012. Proceedings. Lecture Notes in Computer Science (Subseries LNAI), vol. 7243. Springer (2012)
24. Ruge, G.: Experiments on linguistically-based term associations. Information Processing and Management 28(3), 317–332 (1992)
25. Uzêda, V.R., Pardo, T.A.S., das Graças Volpe Nunes, M.: A comprehensive comparative evaluation of RST-based summarization methods. ACM Transactions on Speech and Language Processing 6(4), 1–20 (2010)
26. Wan, X., Li, H., Xiao, J.: EUSUM: Extracting Easy-to-Understand English Summaries for Non-Native Readers. In: SIGIR 2010. pp. 491–498. ACM (2010)
27. Zechner, K., Waibel, A.: Minimizing Word Error Rate in Textual Summaries of Spoken Language. In: Proceedings of the 1^{st} conference of the North American chapter of the ACL. pp. 186–193. Morgan Kaufmann (2000)