USING PRIOR KNOWLEDGE TO ASSESS RELEVANCE IN SPEECH SUMMARIZATION

Ricardo Ribeiro¹ and David Martins de Matos²

¹ ISCTE/IST/L²F - INESC ID Lisboa
² IST/L²F - INESC ID Lisboa
Rua Alves Redol, 9, 1000-029 Lisboa, Portugal

ABSTRACT

We explore the use of topic-based automatically acquired prior knowledge in speech summarization, assessing its influence throughout several term weighting schemes. All information is combined using latent semantic analysis as a core procedure to compute the relevance of the sentence-like units of the given input source. Evaluation is performed using the self-information measure, which tries to capture the informativeness of the summary in relation to the summarized input source. The similarity of the output summaries of the several approaches is also analyzed.

Index Terms— Speech processing, Text processing, Natural languages, Information retrieval, Singular value decomposition

1. INTRODUCTION

Automatic summarization aims to present to the user in a concise and comprehensible manner the most relevant content given one or more information sources [1]. Several difficulties arise when addressing this problem, but one of most importance is how to assess the relevant content. While in text summarization, up-to-date systems make use of complex information, such as syntactical [2], semantic [3] and discourse [4, 5] information, either to assess relevance or reduce the length of the output, speech summarization systems rely on lexical, acoustic/prosodic, structural, and discourse features (although in what concerns the use of discourse features, only simple approaches, such as the given/new information score presented by Maskey and Hirschberg [6] or the detection of words like decide or conclude described by Murray et al. [7], are used) to extract from the input, the most relevant sentence-like units. In fact, spoken language summarization is often considered of greater difficulty than text summarization [8, 9, 10]: speech recognition errors, disfluencies, and problems in the accurate identification of sentence boundaries affect the input source to be summarized. However, although current work strives to explore features beyond the spoken documents transcriptions, such as structural, acoustic/prosodic and spoken language features [6, 11], shallow text summarization approaches like Latent Semantic Analysis (LSA) [12] and Maximal Marginal Relevance (MMR) [13] achieve comparable performance [11].

In the current work, we explore the use of prior knowledge to assess relevance in the context of the summarization of broadcast news. SSNT [14] is a system for selective dissemination of multimedia contents, based on an automatic speech recognition (ASR) module that generates the transcriptions later used by topic segmentation, topic indexing, and title&summarization modules. Our proposal consists in the use of news stories previously indexed by topic (by the topic indexing module) to identify the most relevant passages of upcoming news stories. This approach models relevance within topics and its evolution through time, by continuously updating the corresponding set of background news stories—which can be seen as topic-directed semantic spaces—, trying to approximate the way humans perform summarization tasks [15, 16, 17]. This is in accordance with Endres-Niggemeyer [15, 16, 17], who identified the following aspects as the most important ones when observing the summarization process from a human perspective: knowledge-based text analysis; combined online processing; and task orientation and selective interpretation.

We address these issues using a statistical approach: the topic-directed semantic spaces establish the knowledge base used by the summarization process; term weighting strategies are used to combine global and local weights (which may also include information like the recognition confidence coefficient to address spoken language summarization specificities); and, finally, the use of LSA [18] provides the selective interpretation step, where terms (keywords), as described by Endres-Niggemeyer, are the most obvious cues for attracting attention.

This document is organized as follows: the next section briefly describes the broadcast news processing system; section 3 presents our approach to speech summarization, detailing how prior knowledge is included in the summarization process; section 4 presents the obtained results; before concluding, we address similar work.

2. SELECTIVE DISSEMINATION OF MULTIMEDIA CONTENTS

SSNT [14] is a system for selective dissemination of multimedia contents, working primarily with Portuguese broadcast news services. The system is based on an ASR module, that generates the transcriptions used by the topic segmentation, topic indexing, and title&summarization modules. Preceding the speech recognition module, an audio preprocessing module classifies the audio in accordance to several criteria: speech/non-speech, speaker segmentation and clustering, gender, and background conditions. The ASR module with an average word error rate of 24% [14], greatly influences the performance of the subsequent modules. The topic segmentation module [19], reported as based on clustering, groups transcribed segments into stories. The algorithm relies on a heuristic derived from the structure of the news services: each story starts with a segment spoken by the anchor. This module achieved an F-measure of 68% [14]. The main problem identified by the authors was boundary deletion: a problem which impacts the summarization task. Topic indexing [19] is based on a hierarchically organized thematic thesaurus provided by the broadcasting company. The hierarchy has 22 thematic areas on the first level, for which the module achieved a correctness of 91.4% [20, 14]. Batista et al. [21] inserted a module for recovering punctuation marks, based on maximum
3. SPEECH SUMMARIZATION USING PRIOR KNOWLEDGE TO ASSESS RELEVANCE

As introduced above, human summarization is a knowledge-based task: Endres-Niggemeyer [16] observes that both generic and specific knowledge is used to understand information and assess its relevance.

3.1. Selecting Background Knowledge

The first step of our summarization process consists in the selection of the background knowledge that will improve the detection of the most relevant passages of the news stories to be summarized. This is done by identifying and selecting the news stories from the previous n days that match the topics of news story to be summarized. The procedure, as currently implemented, relies on the results produced by the topic indexing module, but a clustering approach could also be followed.

3.2. Using LSA to Combine the Available Information

The summarization process we implemented is characterized by the use of LSA as a core procedure to compute the relevance of the extracts (sentence-like units) of the given input source.

3.2.1. LSA

LSA is based on the singular vector decomposition (SVD) of the term-sentence frequency \( m \times n \) matrix, \( M \). \( U \) is an \( m \times n \) matrix of left singular vectors; \( \Sigma \) is the \( n \times n \) diagonal matrix of singular values; and, \( V \) is the \( n \times n \) matrix of right singular vectors (only possible if \( m \geq n \); \( M = U \Sigma V^T \)).

The idea behind the method is that the decomposition captures the underlying topics (identifying the most salient ones) of the document by means of co-occurrence of terms (the latent semantic analysis), and identifies the best representative sentence-like units of each topic. Summary creation can be done by picking the best representatives of the most relevant topics according to a defined strategy.

Our implementation follows the original ideas of Gong and Liu [12] and the ones of Murray, Renals, and Carletta [9] for solving dimensionality problems, and using, for matrix operations, the GNU Scientific Library\(^1\).

3.2.2. Including the Background Knowledge

Prior to the application of the SVD, it is necessary to create the terms by sentences matrix that represents the input source. It is in this step that the previously selected news stories are taken into consideration: instead of using only the news story to be summarized to create the terms by sentences matrix, both the selected stories and the news story are used. The matrix is defined as

\[
\begin{bmatrix}
  a_{1d_1} & \cdots & a_{1d_k} & \cdots & a_{1d_T} & a_{11} & \cdots & a_{1n_s} \\
  \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots \\
  a_{Td_1} & \cdots & a_{Td_k} & \cdots & a_{Td_T} & a_{T1} & \cdots & a_{Tn_s}
\end{bmatrix}
\]

where \( a_{ij} \) represents the weight of term \( t_i \), \( 1 \leq i \leq T \) (\( T \) is the number terms), in sentence \( s_{dk} \), \( 1 \leq k \leq D \) (\( D \) is the number of selected documents) with \( d_1^k \leq d_2^k \leq \cdots \leq d_T^k \), of document \( d^k \); and \( a_{ln_s} \), \( 1 \leq l \leq s \), are the elements associated with the news story to be summarized. The elements of the matrix are defined as \( a_{ij} = L_{ij} \times G_{ij} \), where \( L_{ij} \) is a local weight, and \( G_{ij} \) is a global weight. To form the resulting summary, only sentences corresponding to the last columns of the matrix (which correspond to the sentence of the news story to be summarized) are selected.

3.2.3. Term Weighting Schemes

In order to better assess the relevance of the prior knowledge, we explored different weighting schemes, all of them in both situations of using and not using prior knowledge. The following local weights \( L_{ij} \) were used:

- frequency of term \( i \) in sentence \( j \);
- summation of the recognition confidence scores of all occurrences of term \( i \) in sentence \( j \);
- summation of the smoothed recognition confidence scores (using the trigrams \( t_{k-2}t_{k-1}t_k, t_{k-1}t_k, t_{k+1}t_k \), and \( t_k \) is an occurrence term \( i \) in a sentence) of all occurrences of term \( i \) in sentence \( j \).

Global weights are entropy-based (more specifically, are based on the self-information measure), considering the probability of a term \( t_i \) characterizing a sentence \( s_j \):

\[
P(t_i) = \frac{\sum_j C(t_i, s_j)}{\sum_i \sum_j C(t_i, s_j)}
\]

where

\[
C(t_i, s_j) = \begin{cases} 
1 & t_i \in s_j \\
0 & t_i \notin s_j 
\end{cases}
\]

The global weights \( G_{ij} \) were defined as follows:

- 1 (not using global weight);
- \(-log(P(t_i))\);
- same as previous, but calculating \( P(t_i) \) using \( CR(t_i, s_j) \) (eq. 4) instead of \( C(t_i, s_j) \)—used only when using the recognition confidence scores in the local weight;
- same as previous, but using the smoothed recognition confidence scores as described for the local weights—used only when using the smoothed recognition confidence scores in the local weight.

\[
CR(t_i, s_j) = \begin{cases} 
\sum_{t_i \in s_j} \text{recog. conf. score of occurrences of } t_i \text{ in } s_j \\
0 & t_i \notin s_j
\end{cases}
\]

4. EXPERIMENTAL SETUP

Our goal was to understand whether the inclusion of prior knowledge increases the relevant content of the resulting summaries.

4.1. Data

In this experiment, we tried to summarize the news stories of 2 episodes, 2008/06/09 and 2008/06/29, of a Portuguese news program (their properties are detailed in the top part of table 1). As source of prior knowledge, we used for the 2008/06/09 show the episodes from 2008/06/01 to 2008/06/08, and for the 2008/06/29 show, the episodes from 2008/06/20 to 2008/06/28.

\(^1\)http://www.gnu.org/software/gsl/
4.2. Evaluation

As previously mentioned, our goal is to understand if the content of the summaries produced using prior knowledge is more informative than using baseline LSA methods. To do so, we use self information \( I(t_i) = -\log(P(t_i)) \) as the measure of the informativeness of a summary: this meets our objectives and facilitates an automatic assessment of the improvements. The probabilities of term \( t_i \), are estimated using their distribution in the document to be summarized.

In this experiment, we generated 3 sentence summaries (which corresponds to a compression rate of about 5% to 10%) of each news story (similar to what is done, for example, in CNN website http://edition.cnn.com).

4.3. Results

The obtained results are depicted in figures 1, that shows the overlap between the output summaries, and figure 2, that shows the global performance of all approaches.

As can be seen in figure 1, the use of global weight increases the differences between using and not using prior knowledge. This is due to the fact that global weight for each term is computed using not only the news story to be summarized, but also the prior knowledge: that influences also the elements of the term by sentences matrix related to the news story, whereas when not using global weights, the prior knowledge does not affect those elements. That is shown on the top graphic of figure 1, where all summaries have more overlap.

Figure 2 shows the number of times, where each method produced the most informative summary. In general, the methods using global weight achieved the worst results. This can be explained in part by the noise present both in the transcriptions and in the topic indexing phase, since when not using global weights, the use of prior knowledge has a considerably better performance.

5. RELATED WORK

Furui [10] identifies three main approaches to speech summarization: sentence extraction-based methods, sentence compaction-based methods, and combinations of both. Sentence extractive methods comprehend, essentially, methods like LSA [9, 22], MMR [23, 9, 22], and feature-based methods [9, 6, 7, 22]. Sentence compaction methods are based on word removal from the transcription, with recognition confidence scores playing a major role [24]. A combination of these two types of methods was developed by Kikuchi et al. [25], where summarization is performed in two steps: first, sentence extraction is done through feature combination; second, compaction is done by scoring the words in each sentence and then a dynamic programming technique is applied to select the words that will remain in the sentence to be included in the summary.

Closer to our work is the way topic-adapted summarization is explored by Chatain et al. [26]: one of the components of the sentence scoring function is an n-gram linguistic model that is computed from given data. However, in the two experiments performed, one using talks and the other using broadcast news, only the one using talks used a topic-adapted linguistic model and the data used for the adaptation consisted of the papers in the conference proceedings of the talk to be summarized. CollabSum [27] also explores document proximity in single document text summarization: by means of a clustering algorithm, documents related to the document to be summarized are grouped in a cluster and used to build a graph that reflects the relationships between all sentences in the cluster; then informativeness is computed for each sentence and redundancy removed to generate the summary.

6. FUTURE WORK

Future work should address the influence of the modules that precede summarization in its performance; and, assess the correlation of human preferences in summarization with the informativeness measure used in the current work.

7. REFERENCES

Fig. 2. Comparative results using all variations of the evaluation measure. A: plain LSA; B: plain LSA + GW; C: plain LSA + GW + RC; D: plain LSA + GW + RC smoothed; E: plain + RC; F: plain LSA + RC smoothed; G through L, the same but with PK.


