Improving the topic indexation and segmentation modules of a media watch system

Rui Amaral[*], *Isabel Trancoso*[†]

[*] INESC-ID Lisboa/IPS [†] INESC-ID Lisboa/IST
L^2F - Spoken Language Systems Lab
INESC ID Lisboa, Rua Alves Redol, 9,1000-029 Lisboa, Portugal
{Rui.Amaral, Isabel.Trancoso}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract

This paper describes on-going work related to the topic segmentation and indexation module of an alert system for selective dissemination of multimedia information. This system was submitted in the past year to a field trial which exposed a number of issues that should be dealt with in order to improve its performance. Some of our efforts involved the use of multiple topics, confidence measures and named entity extraction. This paper discusses these approaches and the corresponding results which, unfortunately, are still affected by the limited amount of topic-annotated training data.

1. Introduction

The alert system was developed in the scope of the ALERT European project whose goal was to continuously monitor a TV channel, searching inside the news programs for stories that match the profile of a given user. The system may be tuned to automatically detect the start and end of a broadcast news program from the occurrence of the program jingles. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program stories. The system then searches in all the user profiles for the ones that fit into the detected topics. If any topic matches the user preferences, an email is send to that user indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a user to follow the links to the video clips referring to the selected stories.

During the last stage of the project, the system was submitted to a field trial which exposed a number of issues that should be dealt with in order to improve its performance [1]. The work described in this paper concerns our efforts in terms of the topic segmentation and indexation modules. In describing them, we shall use an almost chronological order: we shall start by briefly describing the story segmentation module and its identified problems (Section 2) and the indexation module and corresponding problems (Section 3). Next we shall describe the use of multiple topics (Section 4), word confidence measures (Section 5), and named entity extraction (Section 6). The paper concludes with a discussion of these results and our plans for future research in this area.

2. Story Segmentation

The input to the topic segmentation and indexation module is a stream of transcript segments produced by the acoustic segmentation and speech recognition modules. The corresponding XML file includes not only the transcribed text for the segments containing speech, but also additional information such as the segment duration, the acoustic background classification, the speaker gender and the identification of the speaker cluster. Special clusters are also tagged as anchors.

The segmentation algorithm uses this incoming data to split the broadcast news program into the constituent stories, through a clustering process. This is done taking into account the characteristic structure of broadcast news programs [2]. They typically consist of a sequence of segments that can either be stories or fillers (short segments spoken by the anchor announcing important news that will be reported later). Our topic segmentation algorithm is based on a very simple heuristic that assumes that all new stories start with a segment spoken by the speaker identified as anchor. Hence, it starts by defining potential story boundaries in every transition non-anchor transcript segment/anchor transcript segment. In the next step, the algorithm tries to eliminate stories that are too short (containing less than 3 spoken transcript segments), because of the difficulty of assigning a topic with so little transcribed material. In these cases, the short story segment is merged with the following one with the same speaker and background.

Using the metric adopted used in the 2001 Topic Detection and Tracking benchmark NIST evaluation, and the same cost values of miss and false alarms [3], our algorithm achieved a normalized value for the segmentation cost of 0.84 for a priori target probability of 0.8.

This experiment and the results of the field trial showed that boundary deletion is a critical problem. One of the reasons for boundary deletion is related to filler segments which usually contain speech with a specific music-like acoustic background. For certain types of news shows, fillers can be well identified and hence rejected using, for instance, neural networks techniques. Other shows, however, use fade in which renders filler detection much more difficult. When it fails, since filler segments are usually followed by a new story also introduced by the anchor and all potential story boundaries are located in transitions ”non-anchor/anchor”, the boundary mark will be placed at the beginning of the filler region and no additional boundary marks will be placed in the beginning of the following story. Filler transcriptions are often incorrect, with higher substitution and insertion rates. This was one of the main motivations for including confidence measures, in order to be able to have some information about the quality of transcriptions.

3. Story Indexation

The type of indexation required by the Portuguese user partner in the ALERT project makes our system significantly different from the one developed by the French [4] and German [5] partners, and from the type of work involved in the TREC SDR Track [6]. In fact, we base our topic concept in a hierarchically organized thematic thesaurus that is used at RTP's manual daily indexing process. This thesaurus follows rules which are generally adopted within EBU (European Broadcast Union) and has an hierarchical structure with 22 thematic areas in the first level. Although each thematic area is subdivided into (up to) 9 lower levels, we implemented only 3 in our system. In fact, it is difficult to represent the knowledge associated with a deeper level of representation due to limited training data. This structure, complemented with geographic (places) and onomastic (persons and organizations) descriptors, makes our topic definition.

In our indexation module, all the segmented stories are classified in two steps. We start by detecting the most probable story topic, using the automatically transcribed text. Our decoder is based on the HMM (Hidden Markov Model) methodology and the search for the best hypothesis is accomplished with the Viterbi algorithm. The topology used to model each of the 22 thematic domains is single-state HMMs with self-loops, transition probabilities, and either unigram or bigram language models. Models were trained from automatically transcribed stories with manually placed boundaries which were post-processed in order to remove function words and lemmatize the remaining ones. Lemmatization was performed using a subset of the SMORPH dictionary with over 97k entries [7]. Smoothed bigram models were built from this processed corpus with an absolute discount strategy and a cutoff of 8.

In the second step, we find for the detected domain all the second and third level descriptors that are relevant for the indexation of the story. To accomplish that, we count the number of occurrences of the words corresponding to the domain tree leafs and normalize these values with the number of words in the story text. Once the tree leaf occurrences are counted, we go up the tree accumulating in each node all the normalized occurrences from the nodes below [8]. The decision of whether a node concept is relevant for the story is made only at the second and third upper node levels, by comparing the accumulated occurrences with a pre-defined threshold.

The field trials were conducted using only a single topic for each story, which was the cause for very frequent topic oscillations. For instance, the first part of a story on a pedophilia case that is about to go court may be classified under ethics and the second under law. On the other way, false alarms in anchor detection may cause the splitting of a story in those cases where there are topic oscillations along the story.

4. Multiple topics

The above mentioned topic oscillations motivated the use of multiple topic indexation. For each of the 22 topic domains, a non-topic language model was created using all the training stories manually classified as belonging to other topics and not related to the topic in question. The detection is based on the log likelihood ratio test (*LLR*), between the topic likelihood $p(W/T_i)$ and the non-topic likelihood $p(W/\overline{T_i})$.

$$LLR = log \frac{p(W/T_i)}{p(W/\overline{T_i})} \tag{1}$$

where W is the word sequence (story) and T_i a specific topic. The detection of any topic in a story occurs every time the correspondent score is higher than a predefined threshold. The threshold is different for each topic in order to account for the differences in the modeling quality of the topics.

In order to evaluate the indexation performance using multiple topics, we used a subset of stories from the Topic Detection Corpus with manually placed boundaries and automatically transcribed texts. We ignored all filler segments. The accuracy and correctness scores obtained were 92.0% and 94.1%, respectively [9], using only the higher-level topic. These results are significantly better then the ones obtained using single topic indexation (correctness = 73.8%). Given the better results obtained using multiple topics, all subsequent experiments related in this paper have used this strategy.

5. Word confidence measures

Together with the transcribed text, our BN speech recognition system, AUDIMUS.Media [10], produces a set of scores that reflect the matching of the acoustic and language models for each decoded word in the transcriptions. These scores were used to build a CART (Classification and Regression Tree), with the goal of tagging each word in the automatic transcription as probably: correct, substituted or inserted. All the CART experiments were done using the wagon program from the FESTIVAL toolkit [11]. In order to train and test the CART, we used subsets of the ALERT Speech Recognition Corpus (SR). The CART training set includes 30 programs of the SR Training corpus, amounting to around 244k words. For tuning the CART, we used around 47k words extracted from 10 BN programs of the SR Development Corpus. For evaluation, we used one program of the SR Final Evaluation Corpus, including around 11k words.

We started by aligning the manual and automatic transcriptions using the SCLITE Scoring program from the NIST Speech Recognition Scoring Toolkit (SCTK) Version 1.2c, which is a tool for scoring and evaluating the output of speech recognition systems.[1] Table 1 shows the results of this alignment. Since for training the CART we obviously ignored deletions, the percentages of correct, substituted and inserted words were computed over a lower number of words, indicated in the second line.

Table 1: *Tag distribution between the train, development and evaluation corpus.*

	Train 221k	Development 42k	Test 10k
Correct	78.8%	73.8%	49.1%
Substituted	13.2%	17.6%	40.8%
Inserted	8.0%	8.6%	10.1%

The CART assigned a ”YES” label to the correctly transcribed words in the training set (78.8%) and a ”NO” label for the substitution and insertion cases (21.2%). The CART was created by incrementally choosing the best features and the minimum number of examples for leaf nodes for the tuning corpus. After several experiments to find the best tuning values, a CART was created with the minimum of 25 examples for leaf nodes, yielding a correctness score of 76.4% for the training set. In the evaluation experiment, the CART was used to output the probabilities of a certain word being correctly transcribed. The result is shown in the ROC plot of figure 1.

5.1. Application to topic indexation

In order to conduct our first indexation experiments involving confidence measures, we started by choosing a working point in the above ROC plot. We selected a threshold of 0.61, corresponding to 20.0% false alarms (i.e. considering correct a word erroneously transcribed by AUDIMUS.Media) and 84.9% correct detection, and rejected all words with lower confidence scores.

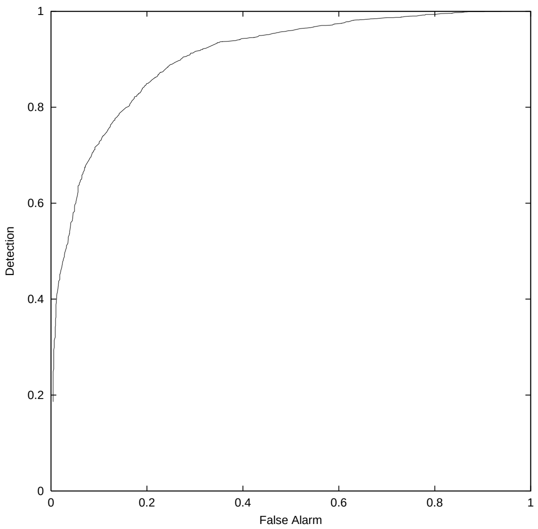

Figure 1: *ROC plot of the CART experiment.*

In terms of indexation, the new topic models created with the non-rejected words yielded an accuracy and correctness of 90.0% and 95.3%, respectively. We observed a slight decrease in the number of misses and an increase in the number of hits. However, we have also observed an increase in the number of false alarms, which we have attributed to the significant reduction in the amount of topic training material, and higher OOV probability.

6. Named entity experiment

Besides the thematic classification described above, our alert system also provides the possibility of the user specifying a free string search in his/her profile. As expected, free string matching, which can also involve Onomastic and Geographic indexation, is more prone to speech recognition errors, specially when it involves a single word This can cause, for instance, the detection of the onomastic descriptor *Pena* in a sentence like *vale a pena* (it is worth), since our algorithm is based on simple word matching, with no attention to context. In order to overcome this problem an experiment using name entities was performed, using the SR corpus.

This corpus was manually annotated according to the Linguistic Data Consortium conventions used for the Hub4 Broadcast News Corpus [2]. Following these conventions, both proper names and place names were manually tagged. All the occurrences of those names were collected automatically, rejecting those with only a single word such as *figo*, *pena*, *guarda*, etc. From a list of 2K distinct names annotated in the corpus, 655 names were

[1] http://www.nist.gov/speech/tools/index.htm

[2] $http://www.ldc.upenn.edu/Projects/Corpus\_Cookbook /transcription/broadcast\_speech/english/index.html$

used in the construction of a named entity list. Although we did not quantify the improvements of this list in the Onomastic and Geographic classification, we observed a significant reduction of the false alarm rate, as expected.

The named entity list was also used to build new topic language models. All the names were merged with the ”_” symbol, giving origin to a ”single” word, and all the topic training data was filtered in order to identify occurrences of those named entities. After the creation of new topic language models based on unigram statistics, an experiment was done with the topic evaluation corpus. The results are presented in table 2. As we can see, the use of named entities caused a slightly improvement in terms of accuracy and correctness, due to the increase of the number of hits and the decrease of the number of deletions.

Table 2: *Topic detection scores without and with Named Entity extraction.*

	without NE	with NE
accuracy	92.0%	93.0%
correctness	94.1%	95.2%

7. Conclusions and Future Work

This paper presented the on-going work toward the improvement of the story segmentation and indexation modules of our alert system. In parallel with this work, we are also currently working on extending lexical and language models, on using more robust jingle detection strategies, and on improving sentence boundary detection by using prosodic information.

Although not reported in this paper, we also attempted to build better topic models using a discriminative training technique based on the conditional maximum likelihood (CML) criterion for the implemented Naïve Bayes classifier [12]. After several experiments, the results did not show any significant improvement in the detection results, which may be due to the small amount of manually topic-annotated training data. The availability of additional data is therefore of crucial importance to all our work on topic indexation.

8. Acknowledgments

The present work is part of Rui Amaral’s PhD thesis, initially sponsored by a FCT scholarship.INESC-ID Lisboa had support from the POSI program of the ”Quadro Comunitário de Apoio III”. The authors would like to acknowledge the continuing support of our former ALERT project colleagues - João Neto, Hugo Meinedo and Diamantino Caseiro.

9. References

[1] I. Trancoso, J. Neto, H. Meinedo, and R. Amaral, “Evaluation of an alert system for selective dissemination of broadcast news,” in *Proc. Eurospeech ’2003*, Geneva, Switzerland, Sept. 2003.

[2] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, “The rules behind roles: Identifying speaker role in radio broadcast,” in *Proc. AAAI 2000*, Austin, USA, July 2000.

[3] N. S. Group, “The 2001 topic detection and tracking (tdt2001) task definition and evaluation plan,” 2001.

[4] Y. Lo and J. Gauvain, “The LIMSI topic tracking system for tdt 2002,” in *Proc. DARPA Topic Detection and Tracking Workhsop*, Gaithersburg, USA, November 2002.

[5] S. Werner, U. Iurgel, A. Kosmala, and G. Rigoll, “Tracking topics in broadcast news data,” in *Proc.ICME’2002*, Lausanne, Switzerland, September 2002.

[6] J. Garofolo, G. Auzanne, and E. Voorhees, “The TREC spoken document retrieval track: A success story,” in *Proc. RIAO’2000*, Paris, France, April 2000.

[7] S. Ait-Mokhtar, “L’analyse présyntaxique en une seule étape,” Ph.D. dissertation, Université Blaise Pascal, GRIL, Clermont-Ferrand, 1998.

[8] A. Gelbukh, G. Sidorov, and A. Guzmán-Arenas, “Document indexing with a concept hierarchy,” in *Proc. NDDL 2001 - 1st International Workshop on New Developments in Digital Libraries*, Setúbal, Portugal, July 2001.

[9] R. Amaral and I. Trancoso, “Topic indexing of tv broadcast news programs,” in *Proc. PROPOR ’2003*, Faro, Portugal, June 2003.

[10] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, “Audimus.media: a broadcast news speech recognition system for the european portuguese language,” in *Proc. PROPOR ’2003*, Faro, Portugal, June 2003.

[11] A. Black and K. Lenzo, “Building voices in the festival speech synthesis system,” 2000.

[12] C. Chelba, M. Mahajan, and A. Acero, “Speech utterance classification,” in *Proc. ICASSP ’2003*, Hong Kong, Apr. 2003.