TOPIC DETECTION IN READ DOCUMENTS

Rui Amaral and Isabel Trancoso

IST, Instituto Superior Técnico
INESC- Instituto de Engenharia de Sistemas e Computadores
Rua Alves Redol, 9, 1000 - LISBOA
email:ramaral@speech.inesc.pt,
Isabel.Trancoso@inesc.pt

Abstract In this paper, we address the importance and the problems involved in topic annotation in the speech retrieval domain. Identified the problem, an algorithm developed to perform automatic topic annotation of broadcast news (BN) speech corpora is described. The approach adopted is based in Hidden Markov Models (HMM) and topic language models, to solve topic segmentation and labelling tasks simultaneously. To overcome the lack of topic labelled material to train the statistical models, a two-stage unsupervised clustering was developed. Both stages are based on the nearest-neighbour search method, using the *Kullback-Leibler* as a distance measure. On-going experiments to evaluate the system performance are also described.

1 Introduction

The increasing development of multimedia technologies, computational power, and storage capacity, together with the growing and diversification of telecommunication technologies, allow us today, to receive anywhere, any kind of information supported by any kind of media. This scenery contributed to an explosion of the information available, and demanded the construction of huge multimedia databases. In order for those databases to be useful, data should be organised and stored in such a way that the access to that information be efficient and effective. Taking into account the dimension of such databases, this is only possible with the development of automatic methods to process, organise and analyse the data.

In the past, most of the research done in this area focused on textual media, obviously more prominent at the time. From that effort new research areas appeared in the information and database retrieval domains such as text retrieval and information extraction. Today, the technology allows us to extend the data storage to others media such as video, images, audio and speech. If we take into account that a great amount of information is spoken (for example: radio and TV), it is easy to understand the motivation for the recent interest in this area by the speech research community.

The speech retrieval problem cannot be equated in the same way as text retrieval because it is not yet possible to produce perfect transcriptions of speech messages automatically. For example, a state-of-the-art continuous speech recognition system for broadcast news generates automatic transcriptions with a word error rate of 30-35% [1].

In speech retrieval, the problem is to organise the recordings into a set of categories or classes predefined and known *a priori*, in order to reduce the number of documents that need to be searched in information retrieval systems. The most successful approaches, nowadays, use diverse classification techniques to process the transcriptions generated automatically by Large Vocabulary Continuos Speech Recognition systems (LVCSR)[2].

In the next section we describe a system developed to perform the topic annotation of automatic broadcast news transcriptions (radio, TV). A preliminary test of this system was done with a spoken corpus of read newspaper items. The test and the corresponding results are described in Section 3. Section 4 concludes with our perspectives for future research.

2 System Outline

Figure 1 shows a block diagram of the topic annotation system. The training process depends on the availability of topic-labelled material. If topic-labelled data is available, the topic language models can be created directly from the automatic or manual speech transcriptions. Otherwise, the unlabelled corpus is first automatically clustered into topics, and topic language models are constructed for each of the resulting clusters.

The main purpose of the clustering procedure is to join stories in clusters (topics) according to their word similarities. The clustering technique we have adopted, similar to the one described in [3], is a two-stage unsupervised clustering based on nearest neighbour search and the *Kullback-Leibler* distance measure. For each cluster, a topic language model is then built from the corresponding stories. These topic models use unigrams statistics and are smoothed versions of a global unigram model obtained with all the training text[1].

After topic models construction, the next step concerns the specification of HMM topology for each topic. In this work, each topic is modelled by a unique state with a "self-loop" and a transition probability. The estimation of these probabilities is obtained from the text corpus. Once the HMM model is defined, the last phase corresponds to the search of the best hypothesis, performed with the *Viterbi* algorithm.

3 Topic detection experiments with BD_PUBLICO

Although planned for the near future, there is still no broadcast news speech corpus for European Portuguese. To overcome this problem, we have used the BD-PUBLICO corpus [5]. The corpus contains texts from the daily Portuguese newspaper “O PUBLICO” collected almost on a daily basis from 1995 to 1999, a subset of which was selected for being recorded by 100 speakers.

---

[1] The language models were created using the CMU Cambridge Statistical Language Modelling Toolkit [4].

We have done two types of pilot experiments with this corpus: a first experiment with the textual corpus, and a second one with the recorded corpus. For the first experiment, a subset of the newspaper texts corresponding to the 28-month period from September 95 to December 97 was used to develop, train, and evaluate the system. This subset is topic-labelled using 9 broad categories: *education*, *culture*, *sports*, *economy*, *science*, *international issues*, *politics*, *media* and *society*. The selection of the stories was performed according to their length, with a maximum of 2000 words and a minimum of 100. After the selection, all the texts were pre-processed in order to solve cases such as, for example, punctuation and tags.

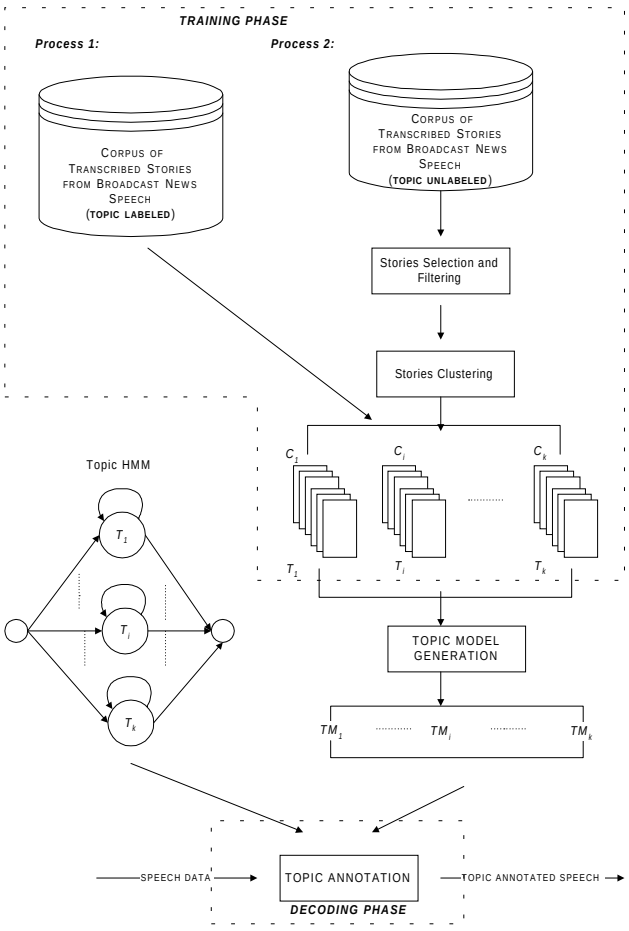

Fig. 1 – Architecture of the automatic topic annotation system, for speech messages

The first 16 months of the corpus were used to train the system. This training material has about 23 million words spread across 42000 articles and a 560-word per story average length. The remainder of the corpus was equally divided into development and evaluation sets, each containing about 3 million words.

The topic annotation results on the evaluation corpus showed 89,9% correctness and an average of 260 out-of-vocabulary words (OOV). The topic confusion matrix is presented in Table 1.

Table 1 - Topic confusion matrix

	B	C	D	E	H	I	P	R	S	Total	Diag
B	88.7%	3.2%	0.0%	0.0%	0.0%	0.2%	2.1%	0.0%	5.8%	433	89%
C	0.4%	95.7%	0.0%	0.6%	0.0%	0.9%	0.2%	0.5%	1.7%	846	96%
D	0.3%	0.2%	95.2%	0.4%	0.0%	0.0%	1.6%	0.6%	1.9%	1985	95%
E	0.4%	0.3%	0.0%	88.2%	0.0%	0.7%	3.3%	0.0%	7.0%	987	88%
H	1.9%	1.9%	0.0%	1.5%	87.3%	0.0%	0.4%	0.0%	6.9%	259	87%
I	0.0%	0.9%	0.1%	0.6%	0.0%	93.5%	1.8%	0.3%	2.8%	679	94%
P	0.2%	0.7%	0.0%	2.2%	0.0%	3.8%	88.2%	0.0%	5.0%	583	88%
R	1.4%	17.0%	1.7%	4.0%	1.1%	1.7%	2.6%	59.5%	10.9%	348	59%
S	0.4%	2.9%	0.1%	1.7%	3.5%	2.7%	2.1%	0.5%	86.0%	801	86%

The confusion matrix shows us that the *media* topics is the most confusable one. He is often confused with the *culture* and *society* topics, which is reasonable since they are strongly related.

The second experiment was done with a subset of the spoken corpus. To do this, the corresponding transcriptions were automatically generated by a Portuguese LVCSR system, with a vocabulary of 5k words and a word-error-rate (WER) of 16.4%, [6]. This relatively small subset has only 36 stories, which is the average size of a television broadcast news program. However, they are all very short, corresponding to partial versions of the original stories, with less than 100 words each.

On this subset, the topic annotation system achieved only 61% of correctness. The low score obtained is mainly due to the small size of the lexicon and the large OOV rate. This limitation, together with the small size of the stories, explains the results achieved.

Conclusions and future work

This paper described our first experiments in the area of topic detection for spoken documents. On-going work is mainly focussed on adjusting the referred lexicon and generating a significant number of automatic transcriptions for the full story versions. Our long-term goal, however, is to work on topic detection for broadcast news. This will be done in the scope of the recently started European project ALERT (Alert System for Selective Dissemination of Multimedia Information). For that purpose, we have already started the collection of a pilot corpus of news related programs in cooperation with RTP, the national TV broadcaster, amounting for the time being to approximately 7 hours of video and audio data.

Acknowledgements

This work was done in the scope of project REC - Speech Recognition Applied to Telecommunications, sponsored by FCT. The authors would like to thank their colleagues Diamantino Caseiro and João Paulo Neto for many helpful contributions.

References

1. Fiscus, J., Doddington, G., Garofolo, J., Martin, A., "NIST'S 1998 Topic Detection and Tracking Evaluation (TDT2)", in Proceedings of DARPA Broadcast News Workshop, USA, February 1999.
2. Ng, K., "Survey of Approaches to Information Retrieval of Speech Messages" Technical report, Spoken Language Systems Group, Massachusetts Institute of Technology, February 1996.
3. Yamron, J. P., Carp, I., Gillick, L., Lowe, S., "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", in Proceedings of ICASSP-98, Seattle, May 1998.
4. Clarkson, P., Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit" Technical report, Spoken Language Systems Group, Massachusetts Institute of Technology, February 1996.
5. Neto, J., Martins, C., Meinedo, H., and Almeida, L., "The Design of a Large Vocabulary Speech Corpus for Portuguese", in Proceedings of EEUROSPEECH 97, Rhodes, Greece, 1997.
6. Caseiro, D., Trancoso, I., "A Decoder for Finite-State Structured Search Spaces" submitted to ISCA ITRW International Workshop on "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, September 18-20, 2000.