MAXIMUM ENTROPY SEGMENTATION OF BROADCAST NEWS

Heidi Christensen†, *BalaKrishna Kolluru*†, *Yoshihiko Gotoh*†, *Steve Renals*‡

†Department of Computer Science
University of Sheffield
Sheffield S1 4DP, UK
{h.christensen, b.kolluru, y.gotoh}@dcs.shef.ac.uk

‡The Centre for Speech Technology Research
University of Edinburgh
Edinburgh EH8 9LW, UK
srenals@inf.ed.ac.uk

ABSTRACT

This paper presents an automatic system for structuring and preparing a news broadcast for applications such as speech summarization, browsing, archiving and information retrieval. This process comprises transcribing the audio using an automatic speech recognizer and subsequently segmenting the text into utterances and topics. A maximum entropy approach is used to build statistical models for both utterance and topic segmentation. The experimental work addresses the effect on performance of the topic boundary detector of three factors: the information sources used, the quality of the ASR transcripts, and the quality of the utterance boundary detector. The results show that the topic segmentation is not affected severely by transcripts errors, whereas errors in the utterance segmentation are more devastating.

1. INTRODUCTION

Applications such as summarization, news archive browsing, or query-based information retrieval rely on the availability of structured broadcast news data. The audio news stream needs to be processed in order to instate typographic cues (such as punctuations, named entity capitalization and paragraphs) and to be partitioned into coherent units (such as utterances and topics). This paper discusses a fully automated system for segmenting a news broadcast stream into utterances and topics. In particular we concentrate on statistical maximum entropy (ME) modelling of both utterance and topic boundaries. The models combine information from both audio (prosody) and textual sources (content analysis of automatic speech recognition (ASR) transcripts). Figure 1 illustrates the news broadcast segmentation system. The statistical framework used for the segmentation is based on exponential models and the ME principle [1, 2, 3], and we have incorporated the fast feature selection algorithm ("The Selective Gain Computation Algorithm") proposed in [4].

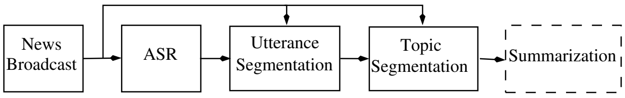

Fig. 1. Broadcast news stream segmentation system.

This research was supported by EPSRC grant GR/R42405 *S3L: Statistical Summarization of Spoken Language*

Our main focus in this paper, is on the overall performance of the system, and we present a series of experiments designed to address several issues arising from cascading ASR systems with utterance and topic segmenters. The initial stage in the broadcast news segmentation system (Figure 1) is to convert the audio to text using an ASR system. In [5] we investigated the effect the quality of the ASR transcripts have on speech summarization, and in this paper we look at the segmentation stage. We also investigate to which degree the quality of the utterance boundary detector (the second stage) affects the topic segmentation. Finally the combination of various information sources in the topic segmenter is investigated. In addition to the linguistic information which Berger *et al.*'s relies on [2], we propose the use of prosodic information which is know to contain significant structural information[6].

2. MAXIMUM ENTROPY SEGMENTATION

A maximum entropy model is a statistical model which agree with any prior set of statistical constraints, f (or feature function), concerning the target distribution, and otherwise assumes a uniform probability distribution. That is, we are looking for a model $q(y|X)$, where $y \in \{\text{YES}, \text{NO}\}$ is the boundary class, and X is the context of the hypothesised boundary. We require that the expected value of the constraints, f with respect to the model, $q[f]$ equates the expected value observed in the training data, $\tilde{p}[f]$

$$q[f] = \tilde{p}[f] \tag{1}$$

$$\sum_X \tilde{p}(X)q(y|X)f(X) = \sum_X \tilde{p}(y|X)f(X) \tag{2}$$

where $\tilde{p}(X)$ is the empirical distribution of X in the training data. To find the model that satisfies the statistical constraints and otherwise exhibits a uniform distribution, we look for the model with the maximum entropy. Solving for the maximum entropy distribution involves introducing a Lagrange multiplier, λ_i for each feature function, the solution of which is a model that belongs to a family of exponential models (we refer to [7, 2] for details).

The model for the positive event of the current context, X being a topic boundary, $C = \text{YES}$ has the form:

$$q(\text{YES}|X) = \frac{1}{Z_\lambda(X)} e^{\sum_i \lambda_i f_i(X)}, \tag{3}$$

where the normalization constant is defined as

$$Z_\lambda(X) = 1 + e^{\sum_i \lambda_i f_i(X)}. \tag{4}$$

feature type		binary question	parameter/range	# features
Cue word	- type A	*Does the word **I** occur in the W utterances before/after B?*	2500 most common words, $W \in \{0 \dots 3\}$	17500
Cue word	- type B	*Does the **I** occur/not occur in the W utterance(s) before/after B?*	four combinations of Occur/!Occur and Before/After, and $W \in \{0 \dots 3\}$	30000
pause		*Is the pause duration above T?*	Threshold [0.1 ... 3.8]	29
N-gram		*Is the N-gram probability above T?*	Threshold [0.05 ... 1.0], $C \in \{TB, notTB\}$	40

Table 1. Description of TB and UB features. I is word instance, W is window size, B is boundary context, and T is threshold.

In text segmentation the boundary context, X can be assumed to be unique at each boundary, and the feature functions take the form of *binary* questions. An example of a feature function is

$$f_j(X) = \begin{cases} 1 & \textit{if for boundary context X the word stem } \textbf{new} \\ & \textit{is in the utterance before the boundary?} \\ 0 & \textit{otherwise} \end{cases}$$

The number of such feature functions is large, and evidently not all are necessarily useful contributors of statistics, so a common practise is to precede the model training with a feature selection stage. Beeferman *et al.* [2] employs a greedy search feature selection algorithm, that includes the features exhibiting the largest gain for the model. For computational reasons, an expression for the approximate gain is used, and we have additionally implemented the fast feature selection algorithm proposed in [4]. This method, "The Selective Gain Computation Algorithm" further speeds up the feature selection stage by limiting the number of times the approximate gain attributed to each feature is recalculated.

Another issue to consider is the disproportionate number of negative to positive events in the data. To compensate for this during training, we have resampled the data so it contains a more fair distribution. Preliminary studies showed that resampling factors of 20 and 30 for the topic boundaries and utterance boundary models respectively were appropriate, and these parameter values are used throughout the work presented here.

Modelling Topic boundaries: The units of the topic boundary model is the utterance, and the model provides statistics for assigning a probability to each utterance indicating to which degree it is the last utterance before a topic boundary (TB).

Modelling Utterance boundaries: The architecture of the utterance boundary detector is in principle similar to that of the topic boundary detection. However, it operates on a word level thus hypothesising each word as a possible utterance boundary (UB).

3. DATA

We used a set of 114 ABC news broadcasts from the TDT–2 broadcast news corpus[1] totalling 43 hours of speech. Each programme spanned 30 minutes as broadcast, reduced to around 22 minutes once advert breaks were removed, and contained on average 7–8 news stories, giving 855 stories in total. In addition to the acoustic data, both manually-generated "closed-caption" transcriptions and transcriptions from six different ASR systems (with WERs ranging from 20.5% to 32.0%), are available [9].

[1] The TDT–2 [8] corpus has been used in the NIST Topic Detection and Tracking evaluations and in the TREC–8 and TREC–9 spoken document retrieval (SDR) evaluations

For the topic boundary experiments, two subsets of the data were used for training and developmental tests, containing 33.8 and 3.9 hours of speech respectively. For experiments on the utterance boundary problem, the number of units is potentially very large (working on the word level rather than the utterance level), why it was chosen to reduce the training and developmental test data with a factor of 10. Note that this in effect means a similar number of positive events (number of TBs and UBs respectively) to train each model on, as there are on average 14 words per utterance in the data set.

4. FEATURE FUNCTIONS

Three distinct types of feature functions are used: cue word feature functions similar to the example given in section 2, feature functions related to the prosody of the speech, and feature functions derived from tri-gram models trained on utterance boundary annotated data. Table 1 gives an overview of the different feature function types. The inherently non-binary features (such as the pause and the N-gram probabilities) are converted into binary feature functions by the introduction of a threshold, T; that is asking *"is the boundary pause above $T > 0.5$"*.

The **Cue word** feature functions are concerned with the occurrence of words around a boundary [2]. Version A describes the cue word occurrence either before (window length, $W < 0$) or after ($W > 0$); version B questions the word occurrence *across* a boundary. The cue words themselves are stemmed, and the utterances are stemmed and filtered for stop words before cue word occurrences are extracted. To speed up computational costs, the initial gains for the cue word features are calculated offline.

Prosodic features are known to convey structural information, which is ignored by systems relying solely on the linguistic information. In previous work we have used prosodic cues (pause duration and pitch information) for structuring broadcast news data [10, 11]. At present, prosodic features in the news broadcast segmentation system are limited to the inclusion of pause information, obtained from the ASR outputs. However, we would expect pause duration to be the single most significant prosodic cue for utterance and topic segmentation.

The N**-gram** based features are obtained from a tri-gram language model trained on utterance boundary annotated transcripts from the a subset of the Hub-4 acoustic data[10]. Feature functions are derived by thresholding $P(TB|\mathcal{H})$ or $P(!TB|\mathcal{H})$, that is the probability of the current boundary being/not being a topic boundary given the history $\mathcal{H}$.

rank	type/word	W	L	rank	type/word	W	L
1	A/new	-1		12	A/new	-3	
2	A/new	0		13	A/abc	-2	
3	A/abc	-1		14	B/todai	2	01
4	A/abc	0		15	B/abc	2	10
5	B/abc	1	10	16	A/abc	-3	
6	A/todai	2		17	B/new	2	10
7	A/todai	1		18	B/abc	3	10
8	B/todai	1	01	19	B/new	3	10
9	A/todai	3		20	B/todai	3	01
10	A/new	-2		21	B/just	3	00
11	B/new	1	10	22	B/think	2	00

Table 2. Selected cue word features in ranked order, where 'W' is the size of the window parameter, and 'L' is the across boundary logic ('00' - word to occur !Before & !After boundary, '01' - word to occur Before & !After boundary etc.

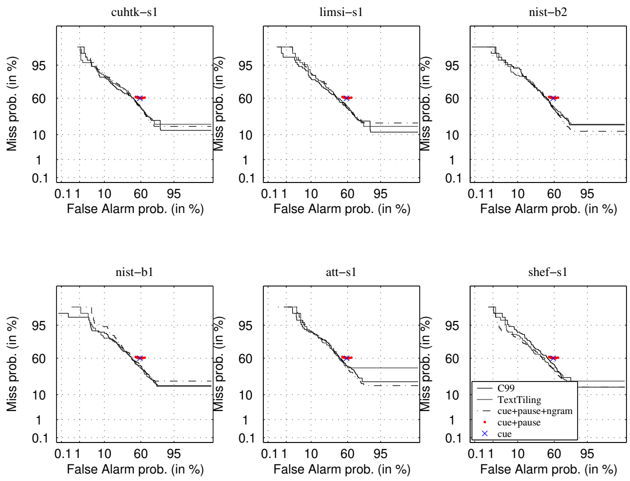

Fig. 2. Effect of using of cue word, prosodic and N-gram features.

4.1. Feature selection

The feature selection algorithm was used to reduce the large number of cue features (47500 in total) to a more manageable number. Preliminary experiments on the closed caption transcripts showed that selecting around 100 cue word features was reasonable. This is the same number of features that Berger *et al.* [2] use. Table 2 shows a ranked list of the first 20 cue word features as output by the feature selection module.

Looking at the word identity, it is evident how closely the selected features match to data. This is an example of a typical "lockout"/"lead" sequence from the ABC news stories:

... American strike against Saddam Hussein.
David Ensor ***ABC news*** *Riyadh Bahrain.*
< NewSection >
In New York ***today*** *the UN. secretary ...*

Of the first 20 highest ranked features 14 are either based on the cue word 'abc' or 'new'. The third most important cue word is 'todai'. It is also interesting to note, that the remaining cue word features in the list are all of type 'B_00'; ie. a certain word appeared neither before nor after a boundary.

5. TOPIC BOUNDARY DETECTION RESULTS

The experimental work presented in this section is concerned with the effect various factors have on the performance of the topic boundary detection achieved by the broadcast news stream segmenter. These factors are 1) the type of feature functions used, 2) the quality of the ASR transcripts, and 3) the quality of the utterance boundary segmentation.

The results are presented in the form of DET curves[2], displaying the relationship between the rate of missed and spurious boundaries.

5.1. Effect of feature combinations

The ME model is suitable for combining the three different types of features investigated in this work: the cue word features, pause features and N-gram based features. Figure 2 presents the DET curve illustrating the performance of different types of features in combination and applied to the six ASR systems. For reference the performances of two baseline text based topic segmentation systems, the TextTiling[12]and the C99[3][13] system run on the closed caption transcripts, are also shown. For all ASR systems, the relative merits of using the different feature types is constant; 'cue' $<$ 'cue+pause+ngram' $<$ 'cue+pause', and in all cases the ME systems outperform the baseline systems.

All combination of feature types are based on the most 100 parameters selected from the total pool of features, eg. all the 29 pause features plus all the 47500 cue word features. Because the feature selection is very time consuming, it was chosen to run these selection experiments only on the closed caption transcripts, and then adopt this list of 100 best features when training the ME model for each of the ASR transcripts.

5.2. Effect of ASR quality on topic segmentation

A fully automatic news stream segmenter would include an ASR system in the initial stages, and understanding the cascading effects of any transcripts errors on the following systems is important. The experiments in this section aims at analysing how the performance of the topic segmenter is affected by the quality of the ASR system.

The topic boundary detector is run on transcripts of six different recognizers and various features, the results of these experiments are illustrated in Figure 4.

The DET curves show relatively little effect of the different transcript quality. The left hand plot for the cue word feature model shows the least variation; introducing the pause features (which are derived from the ASR output, and so also affected by the ASR implementation) gives more variations, and the largest variation is seen when the N-gram features are introduced. This is presumably due to these features being very dependent on the correct recognition of three word sequences.

[2]A DET curve depicts the relation between the false alarm probability and the the miss probability for every possible classifier output threshold.

[3]The TextTile and C99 implementations are both available from www.cs.man.ac.uk/ choif.

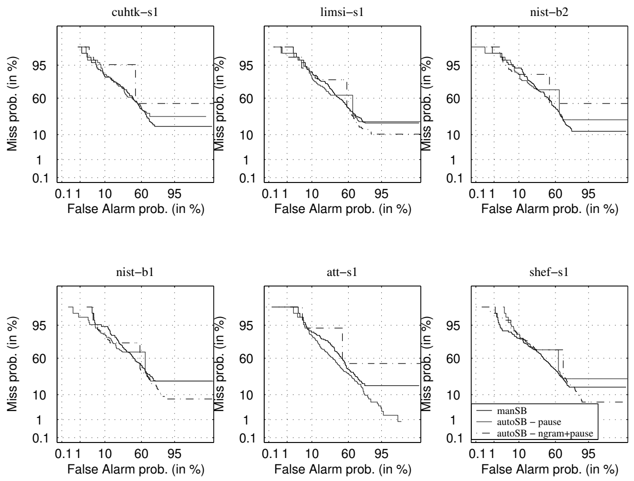

Fig. 3. Effect of utterance boundary segmentation - utterances are manually ('manSB') or automatically ('autoSB') segmented using two different features setups for the utterance boundary detector.

5.3. Effect of quality of utterance boundary detection

The final set of experiments are concerned with how robust the topic boundary detector is to mistakes in the utterance segmentation. All the previously described experiments have been carried out on ASR transcripts where utterance boundaries have been imposed through an alignment (manually adjusted) with the closed caption transcripts which contain hand-segmented utterances. These manual boundaries are only near perfect, but are considerably superior to any automatic method of obtaining utterance boundaries. So how is the performance of the topic boundary detector affected by a non-perfect utterance segmentation? Figure 3 shows the DET curves for a cue word based topic segmentation based on three cases of utterance segmentations: 1) manual, 2) automatic based on pause features, and 3) automatic based on pause and N-gram features. Although the DET curves are close for all ASR transcripts, the manual boundary segmentations curves have smoother characteristics (ie. generally better performance for more operating points) than the automatic segmentations.

6. CONCLUSIONS

A system for the fully automatic preparation of a news broadcast stream for applications such as news story browsing, information retrieval or summarization has been presented. The cascading of non-perfect systems has been investigated, in particular the effect on the performance of the topic boundary detector from three different factors: the type of information used, the quality of the ASR transcripts, and the quality of the utterance boundary detector.

Both the utterance and the topic boundary detector was implemented using a statistical model trained on the ME principle. Cue word, prosodic and N-gram features were employed, after a feature selection algorithm was used to reduce the number of features to a manageable size. A positive effect was found from combining information extracted directly from the audio stream (ie. pause duration) with content information obtained from the ASR transcripts. Transcripts from six different ASR systems was processed and it was found that WERs ranging from 20.5 % to 32.0 % have

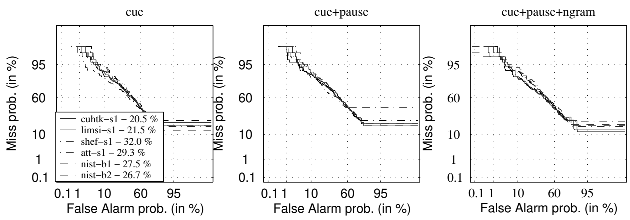

Fig. 4. Effect of different ASR quality.

little effect on the topic boundary detection. Degradations in the utterance segmentations were shown to have more severe effects on the topic segmentation.

In the future we plan to use out automatic news stream system in our work on automatic speech summarization.

7. REFERENCES

[1] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, 1996.

[2] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1-3, 1999.

[3] S. Dharanipragada, M. Franz, J.S. McCarley, and K. Papineni, "Statistical models for topic segmentation," in *ICSLP-2000*, 2000.

[4] Y. Zhou, F. Weng, L. Wu, and H. Schmidt, "A fast algorithm for feature selection in conditional maximum entropy modeling," *EMNLP-2003*, 2003.

[5] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "From text summarisation to style-specific summarisation for broadcast news," in *ECIR'04*, UK, 2004.

[6] G. Tür, A. Stolcke, D. Hakkani-Tür, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Comput. Linguist.*, vol. 27, no. 1, 2001.

[7] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, 1997.

[8] C. Cieri, D. Graff, and M. Liberman, "The TDT-2 text and speech corpus," in *DARPA BN Workshop*, 1999.

[9] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *RIAO-2000*, Apr. 2000.

[10] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *ASR2000*, 2000.

[11] H. Christensen, S. Renals, and Y. Gotoh, "Punctuation annotation using statistical prosody models," in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, New York, US, 2002.

[12] M. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, 1997.

[13] F. Choi, "Advances in domain independent linear text segmentation," in *NAACL'00*, 2000.