Automatic Summarization of Broadcast News using Structural Features

Sameer Raj Maskey, Julia Hirschberg

Department of Computer Science
450 Computer Science Building
Columbia University
New York, NY, 10027
smaskey@cs.columbia.edu, julia@cs.columbia.edu

Abstract

We present a method of summarizing brodcast news that is not affected by word errors in the transcript of broadcast news. We built a graphical model to represent the probability distribution and dependencies among the structural features. We trained the model by filling the probability table by multinomial counts on the training sentences of summary. Then we ranked the new test segments of broadcast news and extracted the highest ranked ones as a summary.

1. Introduction

Most of the recent speech summarization research has been focused on using variety of text summarization technqiues on speaker recognition transcript of the speech [1] has used statistical and linguistic method to extracts words to summarize using linguistic features that depends heavily on the semantic meaning of the segments. Our summary doesnot depend on the content of the segments but on the structural features of the segment. Hence, the presented method can handle transcripts with a worst error rate that cannot be handled by methods using linguistic features.

Structural information we use are how long was each segment of broadcast, when does the segment occur, who spoke the segment, etc. We assume that we can extract speaker information from the broadcast news using disourse cues technique of [2]. We do not need to particularly identify the speakers but we just need a measure of importance of the speaker which can be calculated by dispersion and the frequency of the speakers.

2. System Overview

The system has 4 components which are

- Feature Extraction
- Depenency Structure
- Multinomial Counts
- Probability Table for Tests

2.1. Feature Extraction

The training data is in sgml, xml format. We pre-process and extract the values for the following features for each segment. Segment here means one turn of the speaker of the broadcast. i)position of the segment ii)speaker of the segment iii) length of the segment iv) speaker before the segment v) speaker after the segment vi) length of the segment before the present one vi) length of the segment after the present one

2.1.1. Position

The feature 'position of the segment' (pos) takes account of where did the turn occur. We believe the position is vital in calculating importance of the segment. For example the first few segments of broadcast news is usally important if it is started by anchor because they usually summarize what will be discussed in the news.

2.1.2. Length

The feature 'length of the segment/turn' (len) takes account how long did the speaker speak in his/her turn. This is important because if the length is too long then we donot want to include it in the summary. If it is too short it usually does not carry enough information.

2.1.3. Speaker

The feature 'speaker of the segment/turn' (spk) does not represent the name of the speaker who spoke the segment nor does it represent it if it is anchor, reporter or interviewee. But it represents the overall weight of the importance of the speaker. For each speaker in the broadcast news we calculate the frequency of occurrence and deviation from its own mean. Then we rank speakers according to the the sum of frequency and deviation. Usually the first two anchors are always anchors and the later ones are reporters. Hence, without having to specifically label speaker as anchors or reporters we can get the general weight for them. Also using [2] we can label them as anchors and reporters with high accuracy if we want.

2.1.4. Previous and Next Speaker

The features 'speaker of previous segment' (pspk) and 'speaker of next segment' (plen) are also important. In broadcast news we notice that whenever anchor needs to provide more information on a piece of news the anchor says the news in summary and asks reporter to 'report'. In such cases usually the anchors are part are very useful to be included in summary. Previous and Next speaker feature takes account of such cases and enables us to rank the segments even finer.

2.1.5. Previous and Next Length

The features 'previous length of the segment' and 'next length of the segment' helps us to differentiate between planned, unplanned, formal and informal talks in the broadcast news. We can notice in the news that if is a reporter interviewing someone then it is usually the case that short lengths of segments are spoken back and forth. It is not only in reporter interviewee case

but in anchor-reporter, anchor-interviewee are also other cases where such patterns follow. The given feature allow us to finely give less weight to segments of such cases but not completely ignore them as sometime they might be the important ones to include summarize.

2.2. Multinomial Probability Model

We want to build a model with our given structural features that would rank segments in any new broadcast news. We will build a multinomial probability model for such purpose. In order to build multionmial probability model first we need todiscretize our data. We assign ranges of values in particular slot of distribution. For example, for the position values for segments from 0 to 5% of the whole broadcast news we assign them to a slot 0-5. We can increase or decrease the minimum and maximum of each slot as we need. For example if we think there isn't much difference in importance of news in segments occuring at 50 and 65 then we can have a slot of (max - min) = 15. But for the beginning of news where we think even a slight variation in position might effect the importance of the segment we can have a slot of (max - min) = 5.

After we have made our data discrete such that value for each feature can only be from discrete distribution we can do multinomial counts on the features to fill the probability table that represents the importance of features. The important thing to remember here though is we are doing counts (multinomial counts) on the segments which are good.

We first summarize by human a few number of broadcast news to get training data. Each of the segments in the summary has values for each of our features. Using these values of our training data we want to build a probabilistic model that can rank new unseen segments of broadcast. Since our features now can take only certain discrete values we can represent the distribution of the values as a multinomial distribution.

$$p(x) = \prod_{m=1}^{M} \alpha_m^{x_m} \tag{1}$$

Here alphas sum to 1 and x's are vectors of 1 by M dimension where it is 1 in one dimension and 0s in other dimensions. With out given traning data we to estimate alphas so that best represents the distribution. Since this distribution is for segments of manually segment-extraced summaries, it can be used as a model to rank new segments. In order to find alphas we find maximum likelihood and set it to zero and we get alphas as follows.

$$\alpha_q = \frac{N_q}{N} \tag{2}$$

where Nq is total number of occurrences of feature taking value q and N is total number of occurences of that feature.

Hence, now to build our probability model for summarizer we could build a probability table filled with alphas that we can compute by using eq. 2. If we build such a table we will be able calculate what

$$P(x_1 = c, x_2 = b, ...x_n = z)$$

equals where xns are features and a,b,..z are discrete values of the features. But if we find alphas by counting for each such possible configuration for our structural features where 4 features can each take 7 different values and 3 features each can take 4 different values we will get a probability table of size 153664. This is a problem because we will never find enough data to fill the probability table of this size to take account of our distribution.

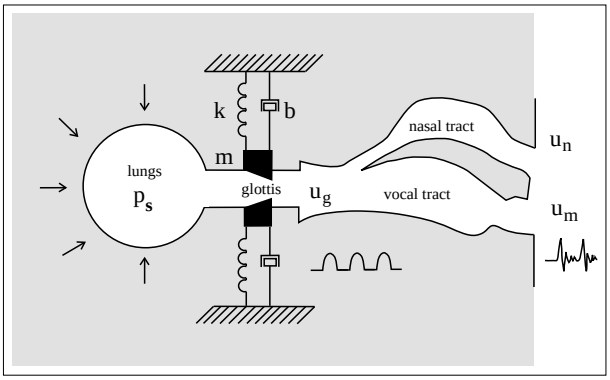

Figure 1: *Bayesian Network for Speech Summarization*

Therefore we implement a graphical model that enables us to reduce the size of the probability table by assuming some independencies. Particulary we use Bayesian Network.

2.3. Bayesian Network

We implement bayesian network where nodes represent the probability tables calculated by doing multinomial counts for each feature and the arcs represent dependencies among feature nodes. In total we have 7 features hence we have total of 7 nodes.

The dependencies we have included in the network represented by symbol (x-y, where y depends on x) are as follows : position, speaker - turn length; position - speaker; speaker - previous speaker; speaker - next speaker; turn length, previous speaker - previous turn length; turn length, next speaker - next turn length.

From the figure we can deduce that the total probability for a particular segment with a given set of values is the product of probability of the feature getting those values in each node. Hence we can calculate the probability that the segment is worth including in the summarization is given by equation 3.

$$P(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = P(x_2|x_1, x_3)$$
$$P(x_3|x_1)P(x_4|x_3)P(x_5|x_3)P(x_6|x_2, x_4)P(x_7|x_2, x_5) \tag{3}$$

When we use the following equation our probability table reduces from $7^4 * 3^4 = 153664$ to $3 * (4 * 7^2) + 2 * (4^2) + 7 + 7 * 4 = 687$ entries. This is a huge reduction in number of parameters we have estimate taking care of data sparsity problem.

3. Results

4. Discussion/Evaluation

5. Conclusions

6. Acknowledgements

7. References

[1] Hori, C., Furui, S., Malkin, R., Yu, H. and Waibel, A. "Automatic Speech Summarization Applied to English Broadcast News", HLT2002 Conference, March 2002, p 241

[2] Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S., ”The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts”, AAAI/IAAI 2000: 679-684

[3]