计算机科学2002Vol.29№.11

新闻视频、音频中的主题检测*)

Topic Detection in News Video and Audio

陈凯江 欧嘉致 黄萱菁 吴立德

(复旦大学计算机系信息论教研室 上海200433)

Abstract Topic Detection in News Video and Audio is to automatically detect snippets with a topic the user searches for, in the news streams, including video, audio and broadcasting. It is a novel research scope rises along with the rapid development of multimedia technology, automatic speech recognition and natural language processing. This technology detects the topic of the news in the semantic level and fits for most people's retrieval need.

Keywords Topic detection, Information retrieval, Video retrieval, Audio retrieval, Multimedia retrieval

1. 引言

面对日益庞大的信息量,如何有效地检索到感兴趣的内容是至关重要的。新闻视频、音频(包括电视、广播)与文字报道相比,更为生动,表达更为丰富,但也有数据量大、难以组织、索引、检索等缺点。这主要体现在两方面:

·文本有标题、段等明显的辅助标记,而视频、音频则没有。一般的浏览工具只有播放、快进、快退、拖动定位等简单手段。这对于几十、几百小时,而且还在日益增长的视频、音频数据库,是远远不能满足要求的。

·人们对新闻视频、音频感兴趣的地方在于其中的语义内容,例如,用户可能对某个主题感兴趣,需要找出有关这个主题的新闻报道。而计算机存贮的是采样点或者采样数据的编码,离语义有相当的距离。

所以,如何对新闻视频、音频进行有效的组织,使人们能方便地找到其中的语义信息是非常迫切的。

对于新闻节目而言,无论是电视还是广播,其语义信息主要通过语音来表达。所以,许多研究者致力于如何将新闻音频分段、分类,如何检索到用户需要的主题。早期的研究者主要使用较底层的特征来将音频分段、归类。例如,检测说话者的变换的地方,将说话者的编号作为索引[1~3]。文[4]使用了自动语音识别(Automatic Speech Recognition,简称 ASR)技术,将其划分为句子等有较强语义信息的单位。

近年来,随着 ASR 技术的实用化,自然语言处理(Natural Language Processing,简称 NLP)技术的发展,研究者们可以在 ASR 识别出来的文本上,进行更深入的分析,提取有关主题的信息。一方面,自然语言理解技术的发展,尤其是篇章分析、向量空间模型等技术,使自动搜寻用户感兴趣的主题成为可能。一方面,ASR 的识别率在不断提高:文[4]在96年使用的 ASR 的错误率约为40-50%,而2000年,CMU 的系统在1999年的 HUB4语料中错误率约为27.6%[5]。

因此,美国标准技术局(National Institute of Standard and Technology,简称 NIST)从1997年开始,每年举办一次"主题检测与跟踪"(Topic Detection and Tracking)评测会议[6],其目的是发展一套技术,能全自动地从新闻中,包括视频、广播等,找出相关主题的片断。参加的单位包括 BBN、CMU 等著名大学和公司。

"主题"的定义有两种方法,一种是"无导师"的方法,由机器自动聚类,每个类是一个主题,然后每个类由一些关键词来描述;一种是"有导师"的方法,由人给出大量文档,每个文档都已经归纳到某个主题中,例如政治、交通运输等,再由计算机从中训练。也有些系统不定义"主题",而要求用户提供一些例子,让计算机寻找相似的新闻。对于视频、音频,由于 ASR 识别结果是一串连续的文本,因此检测的结果是一系列起始、结束边界。

本文第二节介绍主题检测结果的评价方法,第3节介绍主题检测与分割系统的主要框架,第4节讲述主题模型的建立,包括自动聚类和人工确定两种方法,最后是结束语。

2. 评价方法

由于同一主题可能对应多个片断,这些片断可能分布在音频、视频流的多个地方,因此主题检测的结果是给出多个起始、结束边界。文[7]提出了两方面的评价:一个是查全率(Recall Rate),计算公式是

$$\frac{\text{计算机正确地找出来的边界}}{\text{所有真实的边界}} * 100\%;$$

另一个是查准率(Precision),计算公式是

$$\frac{\text{计算机正确地找出来的边界}}{\text{计算机找出来的边界}} * 100\%。$$

如果计算机找出来的边界与真实的边界有偏差,但偏差很小,用户通常是可以容忍的。所以,文[7]又使用了另一种评价:如果计算机找出来的边界与真实边界距离在50词以内,则认为是正确的。

TDT 会议则使用遗漏率(P_{Miss})、虚警率(P_{FA})的加权和构成一个代价函数[8]:

$$C_{Det}=C_{Miss}\cdot P_{Miss}\cdot P_{target}+C_{FA}\cdot P_{FA}\cdot P_{non\text{-}target} \quad (1)$$

其中,C_{Miss}和 C_{FA}分别是遗漏、虚警的代价;P_{Miss}是遗漏的概率,即计算机未能找出的边界占全部边界的比率;而 P_{FA}是虚警率,即计算机找出的错误边界占计算机找出的边界的比率;P_{target}和 $P_{non\text{-}target}$分别是检测目标出现的先验概率(priori target probabilities)。($P_{non\text{-}target}=1-P_{target}$)。

在 TDT2000的主题检测评价中,以上参数分别为[8]:

*)本项目受国家自然科学基金资助(编号69935010和69873011)。陈凯江 硕士研究生,研究方向为自然语言理解、多媒体信息检索。吴立德博导,主要研究方向为大规模文本处理、自然语言理解、多媒体信息检索。

表1 TDT2000主题检测的评价参数

参数	数值
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

由于以上参数和语料有关，为了便于在不同语料和任务之间比较，TDT 又提出了归一化的代价[8]：

$$(C_{Det})_{Norm}=C_{Det}/min(C_{miss}\cdot P_{target},C_{FA}\cdot P_{non\text{-}target}) \quad (2)$$

在 TDT2000的评测中，麻省大学取得最好成绩，其系统在 TDT2000的语料上，检测代价大约为0.6[9]。

3. 基本模块

主题检测系统的主要框架如图1。其中，"人工标明主题的文本"部分是可选的；其它特征指视频、音频中，除了语音内容以外的特征，如说话人的切换、静音的长度等。

视频、音频中的语音被看作许多小段"短语"(两段静音之间的一小段语音，它们可能是一个句子，也可能仅仅是几个词)，每小段语音对应一小段由 ASR 识别出来的文本。检测、分割的大致过程如下：先确定主题类、主题模型；再用这些模型，确定每一小段语音的主题，主题变换之处即是边界。

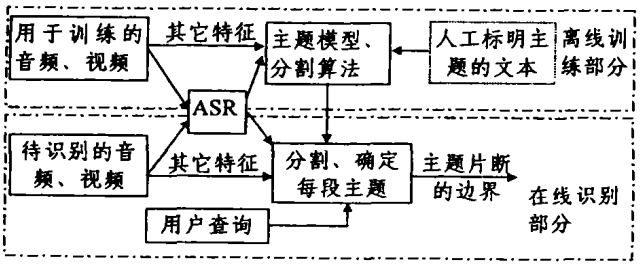

图1 主题检测系统的主要框架

4. 主题模型

主题检测系统先用大量的语料进行训练，找出各个不同主题的特征，并建立主题模型。其方法大致可以分为两类：一类是自动聚类生成主题类，另一类是由人工确定主题类别。

4.1 自动聚类

例如文[7,10]。这种方法一般要求先由人工将新闻视频、音频分段，由计算机自动将相似主题的故事聚为一类。每段是一个故事或者一则新闻(下面统一称为故事)；每段视频识别出一段文本；用多遍 k-means(Multi-pass k-means)算法进行聚类：

·先取 k 个故事作为 k 个类。遍历每个故事 Story，计算 Story 与各个类之间的距离 d_i，取 $d=\min\{d_i\}$，$j=\text{argmin}\{d_i\}$。如果 d 小于一个阈值，则将 Story 归到类 j 中；否则创建一个新的类，将 Story 归到新类中。

·再次遍历所有故事，依照同样的距离重新归类。这样，可能有些类会消失。重复这一步 N 次(N 是预定数值)。

其中，每个故事(Story)与一个类(Cluster)之间的距离 d 使用 Kullback-Leibler 度量(简称 KL 距离，也称为交叉熵)：

$$d=\sum_n\left(\frac{s_n}{S}\log\frac{\frac{s_n}{S}}{\frac{c_n+s_n}{S+C}}+\frac{c_n}{C}\log\frac{\frac{c_n}{C}}{\frac{c_n+s_n}{S+C}}\right) \quad (3)$$

s_n、c_n 分别是词 w_n 在 Story、Cluster 中出现的次数，$S=\sum_n s_n$，$C=\sum_n c_n$。

文[11]则使用凝聚聚类(Agglomerative Clustering)。这一方法的复杂度为 $O(N^2)$，N 是语音段数目。但文[11]设置了一个阈值，当两个类之间的相似度小于这个阈值时，才进行合并，算法一直执行到没有任何两个类可以合并为止，其复杂度为 $O(N*\ln(N))$。

聚类完成之后，还必须为每个主题建立一个模型，用于确定新的文档属于这个主题的概率。最简单的是 Unigram 模型[12]。这种模型对应于每个主题 T，其词汇 w 的概率分布为 $p(w|T)$，故事 S 属于主题 T 的概率为：

$$p_T=\prod_{w\in S}p(w|T) \quad (4)$$

另一种是用 Beta-二项式混合模型(Beta-Binomial Mixture Model)来为每一个主题建模[12~14]。故事 S 属于主题 T 的概率为：

$$p_T(S)=\prod_{i=1}^{V}P_T(n_{w_i}|N,\mu_{w_i},v_{w_i}) \quad (5)$$

$$P(n_{w_i}|N,\mu_{w_i},v_{w_i})=\int_0^1 dp\,P_{BIN}(n_{w_i}|N,p)P_{BETA}(p|\mu_{w_i},v_{w_i})$$

$$P_{BETA}(p|\mu,v)=\frac{\Gamma(\mu+v)}{\Gamma(\mu)\Gamma(v)}p^{\mu-1}(1-p)^{v-1}$$

$$P_{BIN}(n|N,p)=\binom{N}{n}p^n(1-p)^{N-n}$$

其中，i 是词 w 在词汇表中的序号，N 是故事 S 的长度，V 是 S 中特征项的个数，n_w 是词 w 在故事中出现的次数，μ 是 P_{BETA} 的期望值，v 是与方差相关的一个值：$Var(P_{BETA})=\mu(1-\mu)\frac{v}{1+v}$。

4.2 人工确定主题类

这种系统要求先由人给出大量文档，并将每个文档都分好类别，然后由计算机训练。这些方法一般要用到向量空间模型，因此，下面先简略介绍向量空间模型[15]。

向量空间模型(Vector Space Model，简记为 VSM)是关于文档表示的一个统计模型。该模型以特征项作为文档表示的基本单位，特征项可由字、词或短语组成。

令 $D=\{d_i\}$，$|D|=S$($|D|$表示集合 D 中的元素个数)，为文档集，再令 $T=\{t_j\}$，$|T|=M$，为特征项集。定义特征项 t_j 在文档 d_i 中的权重 w_{ij}，如下：

$$w_{ij}=tf_{ij}/df_j,\ 1\leqslant i\leqslant S,\ 1\leqslant j\leqslant M; \quad (6)$$

其中 tf_{ij} 为特征项 t_j 在文档 d_i 中的出现频率，称为项频；df_j 则是文档集 D 中出现了特征项 t_j 的文档的数量，称为文档频率。直观地说，如果特征项 t_j 在文档 d_i 中的作用较大，必然有着较高的项频和相对较低的文档频率，故其权重 w_{ij} 也较大。

在此基础上，建立文档向量空间模型，以 $t_1,t_2,\cdots,t_M$ 为基坐标轴，把文档 d_i 表示为 M 维向量$(w_{i1},w_{i2},\cdots,w_{iM})$。再用如下的相似度来表示不同文档之间的相关程度：

$$Sim(d_i,d_j)=Cos\theta=\frac{\sum_{k=1}^{M}W_{ik}*W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}}$$

$1\leqslant i,j\leqslant S$; (7)

美国举办了两个会议系列：消息理解会议(MUC[16,17])和文本检索会议(TREC[18])。会议评测结果表明，由 Salton 提出的向量空间模型(VSM)[19,20]，是大规模语料库最佳的表示模

型。

许多系统为了降低向量空间维数、简化计算、防止过分拟合，不采用文档中的所有项作为表示单位，而是设置一个评价函数，来计算每个项与主题的关系，抽取关系比较密切的项作为某个主题的特征项。常用的评价函数有[21,22]：

1）词汇和类别的互信息量：$MI(W, C_i) = \log(\frac{P(W|C_i)}{P(W)})$，其中的 W 是词汇，C_i 是类别。需要选取互信息量最大的词汇作为特征词。这是因为互信息量越大，词汇和类别之间的共现概率也越大。

2）词汇和类别之间的 χ^2-统计量：用 $\bar{W}$ 表示除了 W 之外的其他词汇，$\bar{C}$ 表示除了 C 之外的其他类别。那么词汇 W 和类别 C 的共现情况就有4种：(W,C)；$(W,\bar{C})$；$(\bar{W},C)$；$(\bar{W},\bar{C})$。用 n_{11}、n_{12}、n_{21} 和 n_{22} 分别表示这4种情况的频数，总数 $n=n_{11}+n_{12}+n_{21}+n_{22}$。于是

$$\chi^2=\frac{n(n_{11}\times n_{22}-n_{12}\times n_{21})^2}{(n_{11}+n_{12})\times(n_{21}+n_{22})\times(n_{11}+n_{21})\times(n_{12}+n_{22})}$$

χ^2-统计量的值越高，词汇和类别之间的独立性就越小。同时要求 $n_{11}\times n_{22}>n_{12}\times n_{21}$，否则词汇和类别之间就是相斥的关系。

3）词汇的熵：$Entropy(W)=-\sum_i P(C_i|W)\log(C_i|W)$。词汇的熵越小，就更有可能集中在少数的类别中。

4）KL 距离，也称为交叉熵。$CE(W)=\sum_i P(C_i|W)\log\frac{P(C_i|W)}{P(C_i)}$。KL 距离反映了文本类别的概率分布和在出现了某个特定词汇的条件下文本类别的概率分布之间的距离，词汇 W 的 KL 距离越大，对文本类别分布的影响也越大。

特征项得到后，每个主题就可以用特征项及其权重矢量来构成主题模型。

但相似度公式(7)有一个缺点[23]：假设每个项是一个词，有两个不同的项 t_1、t_2 分别出现在 d_1、d_2 中，那么在上式中，t_1、t_2 对相似度的贡献为0；但实际中，两个不同的词可能是意义相近的，它们对相似度的贡献应该大于0。这个问题在新闻视频中较为突出，因为每则新闻可能比较短，两则新闻即使是同一主题的，它们之间共同的词汇也可能很少。文[23]提出了衡量两个不同词之间相似度的方法：

$$DTF(w_1,w_2,m)=TF(w_1,t_m)-TF(w_2,t_m);\ DDF(w_1,w_2)=DF(w_1)-DF(w_2);\ Di(w_1,w_2,m)=i(w_1,t_m)-i(w_2,t_m)$$

$$WD(w_1,w_2)=\frac{1}{M_T}\sum_{m=1}^{M_T}\frac{1}{\sqrt{DTF(w_1,w_2,m)^2+DDF(w_1,w_2)^2+Di(w_1,w_2,m)^2}} \quad (8)$$

其中，TF、DF、i 分别表示项频、文档频率、互信息量。M_T 是主题的个数。

文[24]结合 ASR 文本和 OCR(Optical Character Recognition)识别出来的标题、字幕等，将每则新闻的特征项的归一化频率组成一个矢量 $X=(x_1,x_2,\cdots,x_n)$，对任意两段新闻 X_I、X_A，用以下公式计算其相似度：

$$(X_I,K_A)=\sum_i\sum_j x_{I,i}\cdot x_{A,j}\cdot\frac{1}{WD(x_{I,i},x_{A,j})} \quad (9)$$

4.3 其它问题

在真实语料中，下面的各个模型都会碰到数据稀疏的问题。所以，一般还需要进行概率平滑、回推等处理。许多文献，如文[15]给出了多种概率平滑、回推等处理数据稀疏的常用方法。还有一些特别常用的词，如“the”等，出现的频率很高，但几乎在每个故事、每个主题中出现，不能提供任何和主题有关的信息，所以应当建立一个禁用词列表，将它们过滤掉。

由于 ASR 识别出来的文本含有较多错误，因此有些研究者，如文[25]还考虑了 ASR 的置信度，置信度高的输出，其权重也高。

以上技术大多是基于词汇信息的，而在视频、音频中，除了词汇信息，还有很多特征可以使用。文[26]尝试了73种特征，结果发现除了词汇信息之外，还有几种特征也有用：第0共振峰 F0、静音长度、说话人变换、说话人性别等。

结束语 新闻视频、音频中的主题检测是综合了视频、音频处理、自动语音识别技术、自然语言理解技术的一项新颖的研究方向，其目的是帮助用户找出感兴趣的主题片断，节约用户的浏览时间。随着多媒体技术的进一步完善和应用领域的不断扩大，这个研究方向将会得到日益重视与发展，也将会为人们提供更多的便利。

参 考 文 献

1 Lynn Wilcox, et al. Segmentation of Speech Using Speaker Identification. ICASSP-94, I161～I164

2 Wilcox L, Kimber D, Chen F. Audio indexing using speaker identification. In: Proc. SPIE Conf. on Automatic Systems for the Inspection and Identification of Humans (San Diego, CA, July, 1994), SPIE, 149～157

3 Scheirer E, Slaney M. Construction and Evaluation of A Robust Multifeature Speech/Music Discriminator. In: Proc. 1997 IEEE ICASSP, Munich, April 1997

4 Stolcke A, Shriberg E. Automatic Linguistic Segmentation of Conversational Speech. ICSLP-96

5 Ravishankar M, et al. THE 1999 CMU 10X Real Time Broadcast News Transcription System, 2000

6 http://www.itl.nist.gov/iaui/894.01/tests/tdt/index.htm

7 Yamron J P, et al. A Hidden Markov Model Approach to Text Segmentation and Event Tracking. ICASSP-98

8 http://morph.ldc.upenn.edu/TDT/Guide/manual.front.html

9 http://www.itl.nist.gov/iaui/894.01/tests/tdt/tdt2000/Papers-n-slides/NIST-results/TDT2000_2000.11.16/index.htm

10 van Mulbregt P, et al. Segmentation of Automatically Transcribed Broadcast News Text. TDT-99

11 Eichmann D, et al. A Cluster-Based Approach to Tracking, Detection and Segmentation of Broadcast News. TDT-99

12 Yamron J P, et al. Statistical Models for Tracking and Detection. TDT-2000

13 Lowe S A. The Beta Binomial Mixture Model and Its Application to TDT Tracking and Detection. In: Proc. of the DARPA Broadcast News Workshop, Feb. 1999

14 Lowe S A. The Beta Binomial Mixture Model for Word Frequencies in Documents with Applications to Information Retrieval. In: Proc. of Eurospeech '99, Budapest, Sep. 1999

15 吴立德，等. 大规模中文文本处理. 复旦大学出版社，1997

16 Chinchor N, et al. Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). Computational Linguistics, 19(3): 409～449

17 muc.saic.com

18 trec.nist.gov

19 Salton G. Developments in Automatic Text Retrieval. Science, 1991, 253: 974～979

（下转第89页）

slope、crosses 为计算机合成图像，具有明显边缘，压缩率为 0.1bpp。由表中可以看出，对于不同图像同一方案的最佳滤波器有所不同，这表明对所有图像都表现最佳的滤波器是不存在的。但对于自然光滑图像，最佳滤波器基本是一致的。

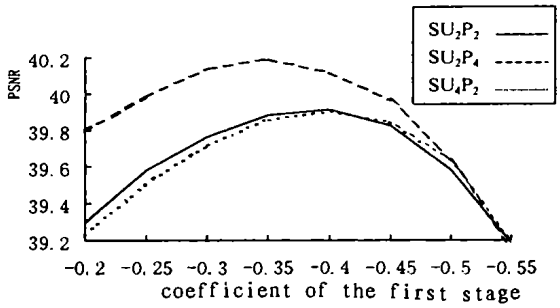

图6 不同方案在相同图像(Lenna)不同系数下的性能比较

结束语 本文通过对提升方案的空间结构进行研究，得到了几种三层提升方案。与第一代小波变换的构造方法相比，本文使用的方法简单明了，易于理解和学习。所得滤波器系数可依据数据的不同而变化，根据应用目标的不同进行适当的选择，会获得较好的结果。下一步工作中，我们将对提升方案的空间结构进行进一步的研究，同时与更多的数学指标相结合，以获得性能更好的滤波器。

表1 不同图像在方案 SU_2P_4 下第一层系数与信噪比的关系

图像	−0.55	−0.5	−0.45	−0.4	−0.35	−0.3	−0.25	−0.2
lenna	39.1589	39.63	39.962	40.1202	40.1905	40.1393	39.9833	39.7913
goldhill	34.6613	35.3789	35.9505	36.1979	36.3187	36.3079	36.2091	36.0314
boat	38.1483	38.533	38.8098	38.9254	38.9259	38.7647	38.6404	38.4753
barbara	34.7272	35.359	35.8336	36.0583	36.147	36.025	35.7705	35.3671
camera	34.795	35.7481	36.1308	36.1309	36.3716	36.2602	36.0515	35.7106
montage	41.0301	41.5123	42.0451	42.2955	42.353	42.2366	42.0715	41.7944
squares	45.4506	45.2311	46.3892	46.2184	45.8686	45.1266	40.5775	39.6563
Circles	22.7287	24.7538	25.4911	26.1634	26.3965	26.4088	26.2906	26.2373
Horiz	33.4489	35.3641	36.0342	36.7213	36.3617	35.1695	35.4249	35.7689
slope	29.3052	30.9485	32.6504	33.5015	33.7399	33.7054	33.2101	32.0418
crosses	20.7124	20.951	21.7264	22.1978	22.4553	22.4652	22.286	21.2438

参考文献

1 Daubechies I, Sweldens W. Factoring wavelet transforms into lifting steps. J. Fourier Anal., 1998, 4: 245～267

2 Wang Guoqiu. Matrix methods of constructing wavelets filters and discrete hyper-wavelet transforms. Optical Engineering, 2000, 39 (4): 1080～1084

3 Wu Yu, Wang Guoyin, Nie Neng. Adaptive lifting scheme of wavelet transforms for image compression. Proceedings of SPIE on Wavelet Application Ⅱ, 2001, 4391: 154～160

4 Roger L, Claypoole Jr, Baraniuk R G. Flexible Wavelet Transforms Using Lifting. In: Proc. of the 68th SEG Meeting, New Orleans, Louisiana, USA, 1998

5 Calderbank A R, et al. Wavelet transforms that map integer to integer. Applied and computation harmonic analysis, 1998, 5: 332～369

6 Said A, Pearlman W A. A new fast and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans. Circuits and System for Video Technology, 1996, 6: 243～250

7 Fernandez G, Periaswamy S, Sweldens W. LIFTPACK: a software package for wavelet transforms using lifting. Proc. SPIE 2825, Wavelet Applications in Signal and Image Processing, 1996, 4: 396～408

8 Sweldens W. The lifting scheme: a construction of second generation wavelets. SIAM J. Math. Anal., 1997, 29: 511～546

9 Wu Yu, Li Gang, Wang Guoyin. Spatial model of lifting scheme in wavelet transforms and image compression. In: Proc. of SPIE's AeroSence 2002 on Wavelet Applications IX, preprint

(上接第100页)

20 Salton G, et al. Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Text. Science, 1994, 264: 1421～1426

21 黄萱菁，吴立德. 独立于语种的文本分类方法. 中文信息学报，2000 (6)

22 Takao S, Ariki Y, Ogata J. Segmentation and Classification of TV News Articles Based on Speech Dictation. IEEE Region 10 Intl. Conf. on Electrical and Electronic Technology (TENCON 2001)

23 Takao S, Ogata J, Ariki Y. Study on New Term Weighting Method and New Vector Space Model Based on Word Space in Spoken Document Retrieval. The RIAO (Computer-Assisted Information Retrieval)2000 in Paris

24 Takao S, Ogata J, Ariki Y. Expanded Vector Space Model based on Word Space in Cross Media Retrieval of News Speech Data. ICSLP-2000

25 van Mulbregt P, et al. Text Segmentation and Topic Tracking on Broadcast News Via a Hidden Markov Model Approach. ICSLP-98

26 Stolcke A, et al. Combining Words and Speech Prosody for Automatic Topic Segmentation. TDT-99