ACM SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications
New Orleans, Louisiana, September 13, 2001

Segmenting Conversations by Topic, Initiative and Style

Klaus Ries*
Interactive Systems Labs, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
Interactive Systems Labs, Universität Karlsruhe, Fakultät für Informatik, 76128 Karlsruhe,Germany
ries@cs.cmu.edu

ABSTRACT

Topical segmentation is a basic tool for information access to audio records of meetings and other types of speech documents which may be fairly long and contain multiple topics. Standard segmentation algorithms are typically based on keywords, pitch contours or pauses. This work demonstrates that speaker initiative and style may be used as segmentation criteria as well. A probabilistic segmentation procedure is presented which allows the integration and modeling of these features in a clean framework with good results.

Keyword based segmentation methods degrade significantly on our meeting database when speech recognizer transcripts are used instead of manual transcripts. Speaker initiative is an interesting feature since it delivers good segmentations and should be easy to obtain from the audio. Speech style variation at the beginning, middle and end of topics may also be exploited for topical segmentation and would not require the detection of rare keywords.

1. INTRODUCTION

Segmenting a dialogue into meaningful units is a problem that has received considerable attention in the past and can be seen as a preprocessing stage to information retrieval [21], summarization [40], anaphora resolution and text/dialogue understanding. This paper uses keyword repetition, speaker initiative and speaking style to achieve topical segmentation of spontaneous dialogues. The intended applications are navigation support for meetings and other everyday rejoinders and preprocessing for applications such as information retrieval. This paper is also an attempt to support the authors general claim that discourse style is an important feature for information access in spoken language as also discussed in other publications [31, 32]. A clean probabilistic framework is presented which allows to formulate keyword repetition and speaker initiative as "coherence features" whereas style is modeled as a "region feature". "Coherence features" are features that have to be coherent within one topical segment – examples are the keyword distribution or speaker initiative. Speaker initiative is encoded by the speaker identity for each turn and possibly the information whether the turn is long or short. (see Sec. 6 for a discussion and experiments on the encoding of speaker initiative).

"Region features" on the other hand are designed to model properties of different regions of topical segments such as the boundary and the beginning, middle and end of a topic. Region features are used to model the change in the part of speech distribution which is a stylistic feature. Region features could also be used to encode features such as prosody and pause lengths.

The databases used in the experiments contains everyday rejoinders, meetings and (personal) telephone conversations [1]. The effective access to audio records of this nature could provide more accurate minutes, improve minutes by adding "audio citations" and increase the confidence in the minute construction process [22]. If meeting minutes are not prepared an automatically generated index may improve the access to the document. Meetings and other everyday rejoinders are fairly different from broadcast news data which has been the focus of information access to speech documents in the recent TREC-SDR (information retrieval) [6] and TDT (topic detection and tracking) initiatives. The following key properties are effecting the dialogue segmentation problem:

speech recognition performance Typical best practice Large Vocabulary Speech Recognition (LVCSR) word error rates on broadcast news have been around 20% [1] for fast decoders in 1998 whereas it is around 40% for slow systems on meeting data in 2001. The most likely explanation would be a significant difference in speaking style of everyday rejoinders from broadcasts.

domain knowledge While broadcast news seems to cover a large domain the topics seem to repeat themselves and a lot of information related to the speech document is available in electronic form. The topic repetition property allowed [38] to use only 100 topic models for segmentation while such preconstructed topic models can't be assumed for everyday rejoinders such as meetings. Keywords of everyday rejoinders may be highly ideosynchratic such that they may not be in the vocabulary of an LVCSR system. Even if some keywords are available it is unlikely that one can use online resources for document expansion [34] and vocabulary adaptation [7] which employ cooccurence information of keywords to enhance information retrieval

*I would like to thank my advisor Alex Waibel for supporting and encouraging this work and my collegues for various discoursive and practical contributions, especially Hua Yu and Klaus Zechner. The reviewers provided valuable comments for the final paper presentation. I would also like to thank our sponsors at DARPA. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, or any other party.

[1] Additionally the publicly available database of a recent publication on topic segmentation [4] is used for comparison.

and speech recognition performance.

manual cuts and genre constraints Broadcasts and especially news shows are very specific genres which are designed for mass consumption: News shows for example are cut in short but very distinct stories which are introduced or ended with very specific phrases. If video is present a "cut" is likely inserted which may be detected easily, narrowing down the number of possible topic boundaries. Everyday conversations on the other hand don't exhibit such clear topical boundaries and topic-shifts may occur gradually.

We have participated in the DoD sponsored Clarity project which dealt with dialogue processing on speech corpora. Given the information above it was unlikely that keywords detected by a speech recognizer would provide good features for topic segmentation such that other features such as speaking style have been investigated. To field products on mobile devices [37] it would be an advantage to eliminate the need for speech recognition altogether since it is expensive.

The paper first presents related work (Sec. 2), the definition of topic (Sec. 3) and evaluation metrics (Sec. 4) as well as the algorithmic framework (Sec. 5). Experimental results are presented in Sec. 6 and conclusions are offered in Sec. 7.

2. RELATED WORK

2.1 Segmentation criteria

Topic segmentation has been studied by other authors previously and a variety of segmentation criteria have been suggested: [9, 10, 38] suggests that segments are assumed to contain semantically distinct elements, usually presented by lexical cohesion which is adapted in this work; [25, 2] suggest that local features indicate topic shifts; [20] proposes an approach based on rhetorical structure to derive a hierarchical representation of the dialogue; [11, 33] show how to use automatically detected prosodic features for segmentation; [36] uses initiative as a manual segmentation criterion and finally multimodal features such as gesture, movement and gaze are used by [28]. Discourse theories such as [8, 19] would also be attractive candidates for segmentation of human dialogue and indeed [20] has shown success in parsing rhetorical structure in text domains using keywords and phrases.

The author therefore decided to use the widely studied keyword repetition feature [9, 10, 38] and speaker initiative as a "coherence features". Speaker initiative has so far only been used as a manual segmentation criterion [36]. Speaking style as encoded in the part-of-speech distribution is explored as a "region feature". The suggested algorithm allows the direct integration of "coherence features" as well as "region features". So far algorithm designers have separated the two sets of features or integrated them in a less direct manner.

2.2 Keyword repetition algorithms

The part of the algorithm which handles coherence features is related to the approach of [38, 29]. [38] assumes that each segment of a conversation is generated by one of a couple of hundred pretrained topics – the algorithm is therefore domain dependent. The algorithm presented here does not make that assumption and is therefore domain independent. The domain independence is achieved by training a model for each segment on the fly instead of relying on pretrained models. An advantage of [38] is that information about semantically related terms is included implicitly. This may be achieved using other techniques such as [26]; however [38, 26] techniques rely on the availability of adequate training material which may not be available for everyday discourse or meetings. A fair comparison to [38] is not possible since there is really no topic repetition across dialogues in our databases which would disfavor their approach while the TDT database would require to add synonym handling to this algorithm.

[29] presents the domain independent probabilistic *word-frequency algorithm* for topical segmentation. It estimates the probability of a boundary for every location in the text and uses a thresholding technique to derive the actual boundaries. The drawback is that the estimation of the boundaries assumes fixed sized windows around the boundary and the boundary placement is not optimized globally unlike the Viterbi search employed by [38] and the proposed algorithm.

[10] is probably the most widely cited domain independent algorithm for topical segmentation and relies on cosine similarity measures combined with heuristic optimization criteria and optimization procedures. Similar algorithms, applying similar measures with different optimization criteria, are [29, 4]. [10, 4] where chosen to establish a comparison to existing domain independent algorithms: [10] is known widely and [4] is the most recent publication in this area which compares to [12, 29, 10].

2.3 Boundary classification algorithms

Many algorithmic approaches have used boundary classification: A classifier is trained which has the output "Boundary: Yes/No". Using "region features" the classifier can be extended to produce other outputs for larger regions such as "Begin of topic", "End of topic" and so forth. The UMass approach in [1] seems to model word type information in different regions of topical segments using an HMM model. The model presented here can be trained using a discriminative classifier but imposes a fixed structure of the topical segment.

Since news shows are a highly organized genres following specific scripts very specific topic shift indicators (such as LIVE, C. N. N.) can work very well which was used by [2, 11]. Other features studied as topic indicators are keyphrases, pauses and prosodic features such as a preceding low boundary tone or a pitch range reset [11, 33, 25]. While these may be modeled easily using region features the author hasn't been able to establish good results on the dialogues although the prosody module has been tested successfully on an emotion detection task.

Boundary classification algorithms may also integrate information about the change in the keyword distribution using features similar to most keyword repetition algorithms [2, 11]. The critique of this technique is however that it is relying on local, window based changes of the keyword distribution and that the algorithms are not applying a global optimization over all possible sequences [2]. On the other

[2] One may argue that exponential segmentation models [2] may weigh the contribution of the keyword repetition feature with the other models in a principled way. On the other hand the parameterization of the exponential models used

hand the algorithm presented in this paper as well as [4, 38] integrate keyword information over the complete topical segment.

3. DEFINITION OF TOPIC

A theoretically pleasing definition of topic that could be applied reliably in practice doesn't exist currently. A simple solution is to compose artificial data randomly picking initial segments from different documents which constitute the topics to obtain a database of different topics. This method is used by [4] in his C99-database which is also used in Tab. 1. The problem with that approach is that the modeling of topic length may be artificial and the shifts between topics may not be natural.

However this work is concerned with the segmentation of naturally occuring dialogue in meetings and everyday rejoinders where topic shifts are not that abrupt and uninitiated. [25] discuss the topic definition problem in great detail and the most common way to establish the quality of a definition is a reasonable agreement between human coders (also called "intercoder agreement"). [10] argues that "naive" (largely untrained, linguistically inexperienced) coders generate sufficient agreements compared to trained coders when asked to place segment boundaries between topical segments. The use of naive coders may also be appropriate for work in information retrieval since it may reflect results that could be anticipated from users of an actual retrieval system. The topic definition applied in this work instructs the coders to place a boundary where the topic changes or the speakers engage in a different activity such as discussing, storytelling etc. The activities were annotated at the same time as the topic segmentation was produced [31, 32]. The primary annotation for all databases was done by semi-naive subjects. Some had considerable experience annotating the databases with dialogue features however no special linguistic training or criteria were provided for the topic segmentation task beyond the definition of activities.

The meeting database was also segmented by the author. The intercoder agreement was measured by (a) treating the second human similar to a machine using the standard evaluation metric (Sec. 4, Fig. 2), (b) measuring κ for the boundary/non-boundary distinction for each utterance ($\kappa = 0.36$) and (c) measuring κ for the distinction of links [3] as *within topic* / *across topic* ($\kappa = 0.35$). The κ-statistics [3] therefore indicates that the intercoder agreement is relatively low overall which is not surprising given the difficulty of the task. The result seems to be in the same range as other similar annotations [25].

4. EVALUATION METHODS

A standard evaluation metric for text segmentation has been suggested by [2, 1]. The metric is fairly intuitive and [2] argues that it is fairly robust against simple cheating attempts. The intuition behind the metric is that a segmentation is good if two words that belong to the same topic in the reference belong to the same topic in the hypothesis.

More specifically if two words have distance k they form a *link*. A link is called *within topic* if the two words belong to the same topic, otherwise it is across topic. If the corresponding links in the hypothesis and reference are both *within topic* or both *across topic* the hypothesis is correct, otherwise it is an error. The reported metric is the average link error rate in percent. For each database k is half the average topic length of the database.

All speech databases have been manually segmented based on the manual transcripts. The results for automatic transcripts of the meeting database have been obtained by transferring the manual topic segmentation to the results generated by the speech recognizer. The speech recognition system segments the input by pause length. Based on time stamps the next best utterance beginning is chosen as the segment boundary.

[4] calculates the link error rate differently and his technique is used when reporting results on the C99-database. The first step in his procedure is to calculate the average link error rate for every text in the database. The link length k for every text is determined as half the average topic length of the respective text. The average link error rate of a database is the average of the average link error rate of all texts in the database.

As a baseline an "equal distance segmentation" is being used, similar to B_e in [4]. The dialogue is segmented into utterances with equal sized topics of length d where d is the average length of a topic in a training set. The parameter d is estimated in a Round Robin procedure.

5. PROBABILISTIC MODELING

5.1 Introduction

The algorithm is based on a standard probabilistic modeling approach. If D is a dialogue and L is a possible segmentation the Viterbi algorithm is used to find the best segmentation L^*

$$L^* = \text{argmax}_L \quad p(L|D) = \text{argmin}_L \quad -\log p(\text{S})$$

where $\text{S} = \langle D_0, \ldots, D_n \rangle$ is the dialogue segmented into topical segments D_i. The model for $p(S)$ is assumed to be decomposable into models for the number of segments per dialogue $p(\#segments)$, the length of each segment $p(\text{length}(d_i))$ and models for the content of each segment given the segment length $p(d_i|\text{length}(d_i))$:

$$p(s) = p(\#segments) \prod_i p(\text{length}(d_i)) p(d_i|\text{length}(d_i))$$

The most crucial assumption of this model is that all segments are assumed to be independent of each other which is invalid in general, especially when a topic is resumed after a digression. The dialogue segmentation model can be simplified by assuming exponential models for $p(\#segments)$ and $p(\text{length}(d_i))$ and which allows to consolidate the two into a single penalty P in the optimization:

$$L^* = \text{argmin}_L \sum_i P \quad - \log p(d_i|\text{length}(d_i))$$

where d_i is the ith segment of d with respect to the segmentation L. Since the dialogue d is known we may call $d[l:k]$ the segment ranging from k to l and define

$$M_{k,l-k-1} := -\log p(d[k:l]|\text{length}(d[k:l]))$$

may also be interpreted as a different weighting scheme between "coherence features" and "region features". A pilot experiment using a couple of settings did not indicate any change of the segmentation accuracy.

[3] Refer to Sec. 4 for the description of links.

Finding the most likely sequence corresponds to finding the best sequence L of strictly ascending indices such that the sequence contains 0 as the first index and size of M as the last:

$$L^* = \text{argmin}_L \sum_{0<i\leq \text{size}(L)} P + M_{L[i-1],L[i]-L[i-1]-1}$$

Since very long segments are extremely unlikely our implementation uses a maximum length constraint of 300 turns. This number was chosen conservatively such that almost no mistake was made, however the win in runtime was significant since dialogues can be very long. The parameter P may be chosen to derive segmentations of different lengths. Two different strategies are used for determining P for a test utterance, the *penalty criterion* for the C99-database and *segment ratio criterion* the dialogue databases: The *penalty criterion* determines P on the training database by taking the mean of the P_i for each training utterance which generate the correct number of segments for that utterance. The *segment ratio criterion* determines the average number of utterances per topic on the training database which is used to determine the number of topics for a test utterance. P is determined for each test utterance using a logarithmic search such that the desired number of segments is obtained. Training and testing is done in a Round Robin fashion such that the whole database can be tested.

5.2 Coherence features

Keyword repetition and speaker initiative can both be modeled as coherence features by assuming that each segment follows its own language model. In the case of keyword repetition the language model describes the keywords, for speaker initiative it describes the speaker identity of an utterance and potentially an indication of the initiative such as utterance length (see Sec. 6 for details on the implementation of the features). The probabilistic model requires to define $\log p(d_i|\text{length}(d_i))$ in an appropriate fashion. In speech recognition [13] pioneered the use of so called cache models that adapt themselves over time. Cache models have been used in two flavors in the speech recognition community: Either similar to [13] and accurately following the probabilistic framework by continuously updating the context and calculating the probability for the next word on the fly (the dynamic approach) or by recognizing a segment of speech and training a static language model on the speech recognition result (the static approach). While initial experiments used both models there are no differences in the experimental results. Since the static model is simpler to implement and faster in execution it is used for all experiments reported. The static model approach seems to be somewhat counterintuitive at first but it can also be explained in the minimum description length framework [5]. The probabilistic framework can explain the static model by associating a language model with each segment boundary. The random variable L may be reinterpreted as the segmentation including the likelihood of the segment language model. If all language models are assumed to be equally likely it is modeled by another penalty that can be subsumed by P. To obtain better estimates and avoid "zero probabilities" the cache model was smoothed using absolute discounting with a fixed parameter $D = 0.5$ [23]. The discounting method and parameters used were fairly uncritical when compared to alternatives during prestudies.

5.3 Region features

Region features are an extension of the common boundary modeling approach to discourse segmentation. A region mapping is a function f which maps an integer k onto an array of k region labels. It can therefore be naturally extended to a function f' which maps a segment d_i containing k utterances to k segmentation labels. The intuition is that if the length of the segment is known it has to follow a certain fixed pattern. The simplest example is the classic *boundary modeling* approach where

$$f(k)[j] \quad := \quad \begin{cases} \text{BOUNDARY} & \text{if } j = 0 \\ \text{NONBOUNDARY} & \text{otherwise} \end{cases}$$

The *boundary modeling* assumes that there are very specific phrase or intonational events at or near the boundary (key words and phrases). The *equal size regions* approach (3 regions for Begin, Middle and End) can easily model changes in general distributions such as the part of speech distribution: At the beginning new items are introduced explicitly whereas they are referred to anaphorically towards the end. They can be combined in the *equal size + boundary* approach which features one region for the boundary and Begin, Middle and End regions. Since f' is a deterministic function of the segment length length(d_i)

$$p(d_i|\text{length}(d_i)) = p(d_i|f'(d_i), \text{length}(d_i))$$

In order to make this quantity tractable independence assumptions have to be made: All segments in a topic are independent given the segmentation labels, all segments depend only on their respective segmentation label and after those assumptions are applied there is no more dependency on the length of the segment:

$$p(d_i|\text{length}(d_i)) = \prod_j \frac{p(f'(d_i)_j|d_{i_j})}{p(f'(d_i)_j)} \cdot \prod_j p(d_{i_j})$$

where $f'(d_i)_j$ is the jth region label of $f'(d_i)$ and d_{i_j} is the jth utterance of the region d_i. Since $p(d_{i_j})$ is independent of the segmentation L it can be ignored in the search procedure. The score of the model is therefore just the probability of the region label given the dialogue segment (as determined by a classifier such as a neural network) divided by the prior of the region label.

The advantage of this approach is that it extends boundary classification to the classification of multiple regions. It is particularly useful if we assume that simple regions of topics have different properties which may provide a natural model of prosodic and stylistic difference across regions. If one would use language model classifiers using part-of-speech as features to determine $\frac{p(f'(d_i)_j|d_{i_j})}{p(f'(d_i)_j)}$ the model would burn down to the training of part-of-speech Markov models for each segment of a topic which provides an intuitive description stylistic regions. Alternatively one could train a classifier with the same parameterization discriminative – a neural network without hidden units and the softmax function as its output or exponential models are such classifiers (see also [30]).

6. EXPERIMENTS

The experiments were carried out on the CallHome Spanish, a corpus of meetings, the Santa Barbara corpus and a database used by [4] (C99-database). All experiments have been carried out on manual transcripts unless noted otherwise – only for the meeting corpus speech recognition results have been available:

CallHome Spanish The whole corpus (120 conversations, approximately 20min each) has been hand annotated with topical segmentations. The corpus features telephone calls in Spanish between family members calling from the US to their home countries. The original corpus [16] was published by [18] and recently the topical segmentation along with further dialogue annotation from our project Clarity were published as well [15].

Santa Barbara 7 of 12 English dialogues [14] have been segmented manually. The corpus features all kinds of oral interactions including meetings, evening events and kitchen table discussions.

Meetings 8 English dialogues have been annotated with topic segmentation and 2 of those have been processed using a speech recognizer. The meetings are recordings of group meetings, mostly of our own data collection group. The latest published speech recognition error rates for this corpus are around 40% word error rate and the out of vocabulary rate is about 1-2% for each meeting [35, 39]. Speech recognition results were available for two out of eight meetings and the topical segmentation has been transferred to those (see Sec. 4).

C99-database [4] used the Brown corpus to generate an artificial topic segmentation problem and the corpus is available (see [4]). A small program randomly grabbed initial portions of Brown corpus texts and concatenated them as the topics of an artificial text. The database consists of four subparts: 100 texts with topics 3-5, 6-8 and 9-11 sentences in length and 400 with topics of 3-11 sentences in length. The "all" database is the concatenation of these four databases. To maintain consistency with his results his formula for the average link error rate has been used.

The first question addressed was whether the probabilistic modeling approach for keyword repetition and speaker initiative compares well to standard algorithms. The stopwords were removed from all databases and the first four letters were retained from each word for all databases. On the C99-database Porter stemming [27] was used instead of the 4 letter stemming. Additionally the corresponding formula for the calculation of the link error rate was used on the C99-database (Sec. 4) to allow comparisons with [4]. For speaker initiative each utterance was replaced by a single token representing the speaker identity and the information whether the speaker turn was long or short (a turn was defined as short if it contains three words or less). In Tab. 1 the probabilistic approach (R01) was compared to [4] (C99) and the texttiling [10] approach (Tile). To present Tile in the best light the implementation provided on Hearst's WWW page was chosen on the dialogue segmentation tasks and the reimplementation of [4] on the C99 database [4].

	Link error rate in %			
Database	R01	C99	Tile	Baseline
Dialogue segmentation, keyword repetition				
SantaBarbara	39.0	53.7	49.2	49.0
CallHome	37.9	40.4	44.1	45.9
Meetings	37.6	45.2	44.6	47.8
Dialogue segmentation, speaker initiative				
SantaBarbara	35.3	41.6	42.3	49.0
CallHome	45.5	43.8	43.0	45.9
Meetings	38.9	39.7	39.7	47.8
C99 database, keyword repetition				
All	13.8	12.8	30.4	42
3-11	13.6	13.0	29.9	45
3-5	17.2	17.7	36.7	38
6-8	8.9	9.6	26.8	39
9-11	16.1	10.0	29.7	36

Table 1: Algorithm comparison: The proposed algorithm (R01) is compared to [4] (C99) and [10] (Tile) for keyword coherence and speaker initiative based topical segmentation. The equal distance baseline (baseline) is listed for comparison. It delivers excellent results on all database slightly worse results than C99 on C99-database. The results on the C99 database however have to be taken with a grain of salt due to the artificial construction of the database.

The results show that the new algorithm delivers excellent results on the dialogue databases: The results are always better than the other algorithms, in some cases by large margins [5]. The only exception is the speaker initiative criterion for the CallHome database which may however be a bad example since speaker initiative is likely a bad criterion for that database (see further discussion below). Tile and C99 seem to perform similar.

The results for the C99 database are very different, the C99 and R01 algorithms perform similar with the exception of the 9-11 part of the database where C99 performs a lot better. The situation changes if the algorithms for determining the number of segments are changed: If the number of segments for R01 is chosen to be the number of C99 the result of R01 is not much worse. If both algorithms are given the number of segments from the reference R01 performs better. As noted above R01 worked a lot better on the C99-database using the *penalty criterion* unlike the *segmentation ratio criterion* used on the dialogue databases. Given these results the author cautions the interpretation of the results on the C99-database since it has been artificially constructed. Specifically the length distribution of the segments seem to be unnatural and may place too much weight on the algorithm determining the number of segments. Overall R01 is slightly worse than C99 on this database yet much better

[4] Note that the same stemming algorithms were used for all algorithms – [4] didn't use Porter stemming in the tiling implementation which was used here and Hearst's algorithm was also fed with the exact same input as the others. The author replaced the stopword removal and stemming from the external algorithms and replaced them with his own implementation. The native implementation of the Porter algorithm of C99 delivered identical results to the reimplementation used here.

[5] The results for Tile and C99 improve when their native criterion for determining the number of segments is replaced by the *segmentation ratio* criterion presented here – however R01 still performs significantly better. These results are not shown in the tables since they are secondary and would require much more space.

	Link error rate in %					
Features	None	4 letters		No mapping		Trigram
Stopwords		No	Yes	No	Yes	Yes
Santa Barbara (baseline 49.0%)						
		39.0	38.6	41.1	41.0	43.8
Speaker + ls	35.3	38.5	38.7	35.9	41.8	41.8
Speaker	39.1	36.4	39.4	36.2	40.5	41.7
CallHome Spanish (baseline 45.9%)						
		38.6	38.4	39.4	38.6	37.2
Speaker + ls	45.6	38.8	39.6	39.3	38.3	37.3
Speaker	45.3	38.4	38.3	39.1	38.8	37.2
Meetings,topic segmented database (8 meetings) manual transcript (baseline 47.8%) Second human 32.3%						
		37.6	33.1	37.6	34.3	35.3
Speaker + ls	38.9	35.6	33.1	36.9	32.9	33.7
Speaker	42.6	36.0	32.9	37.9	33.8	34.9

Table 2: Dialogue segmentation: Topical segmentation was tested on the Santa Barbara corpus, CallHome Spanish and the meeting corpus, all corpora manually transcribed and annotated with speakers. Two types of features are being compared, keyword repetition and speaker repetition.

than Tile and the other algorithms tested in [4].

Tab. 2 compares coherence features. For word repetition the following choices can be made: (a) should stopwords be modeled as well and (b) should a word be mapped onto some baseform (stemming). The inclusion of stopwords may model the speaker identity implicitly or it may model general speaking style. The stemming algorithms tested were *No mapping* which doesn't perform any stemming, the *4-letter* stemming which maps a word onto its first 4 letters and the *trigram* method which maps each word onto the trigrams that occur in it. The *4 letter* stemming seems to be effective. Additional attempts to use Porter stemming [27] on the English database did not show improved results. The *trigram* stemming may capture endearments or other morphological features in Spanish which may explain its effectiveness on CallHome. The inclusion of stop words is typically improving the performance if speaker initiative is not modeled.

For speaker initiative each utterance can either be replaced by the speaker identity itself (*Speaker*) or the speaker identity plus the information whether the utterance was long or short (*Speaker+LS*). An utterance is called short if it contains three words or less. This definition is designed to capture the information whether a speaker issued a dominant dialogue act or a non-dominant dialogue act. Short utterance tend to be non-dominant dialogue acts such as backchannels or answers. A strong correlation of dominance and the dialogue act type has been shown empirically by [17] and the results indicate that the *Speaker+LS* feature performs significantly better than the *Speaker* feature by itself. The long/short criterion has the advantage that it may also be implemented easily without having access to a speech recognition engine. Other encodings of speaker initiative did not improve the results.

The speaker initiative approach doesn't seem to be very successful on CallHome Spanish. The reason for that fact may be seen in the familiarity of the speakers and their established (dominance) relationship as well as in the fact that one speaker is abroad whereas the other is "back home".

Meetings, LVCSR database		
	Link error rate in %	
Features	manual	machine
baseline	42.4	42.1
words,no stopwords	34.3	38.7
words+stopwords	32.5	35.2
speaker+ls	36.9	36.5
speaker+ls and words, no stopwords	34.0	39.4
speaker+ls and words+stopwords	30.4	33.6

Table 3: Speech Recognition: Two of the meetings have been fully decoded by an LVCSR system with a word error rate of approximately 40% [35]. The *4 letter* word normalization has been used (see Tab. 2).

Both properties may lead to dialogues where the dominance is rarely shifting between topics. For the multi-party dialogues in the Santa Barbara and meeting corpus however speaker initiative outperforms the keyword based approach. On meetings the combination of the two delivers the best results.

Tab. 3 demonstrates the effect of speech recognition on the segmentation accuracy. While the result for keywords information is worse using speech recognition it is not as bad as one might assume. This result may also be due to consistent misrecognitions that might be produced by a speech recognizer due to keywords that are missing from the vocabulary. Using stopwords additionally to words resulted in a significant improvement in link error rate with no degradation introduced by the speech recognizer. Speaker initiative can be used by itself and it can be combined successfully with word and stopword information. The results have to be taken with caution due to the small size of the database available and the manual annotation of speaker identity.

In Tab. 4 the effect of part-of-speech features for region modeling is shown. A neural network classifier was trained without hidden units, the softmax output function was used as the output function. The vocabulary for the neural network (NN) and language model (LM) classifier were the most frequent 500 word/part of speech pairs while the remaining words are mapped on their part of speech. The effects are clear especially for the equal size+boundary region model and the improvements can also be confirmed when combining the model with repetition modeling, especially on CallHome Spanish. It is therefore clear that there are changes in the word and part of speech distributions in different topical regions. However the combination of word based region modeling with the best repetition model didn't always yield better results for the other databases. Neural network performed significantly better than language models as region classifiers on some segmentation tasks but are slightly worse on some others.

7. CONCLUSION

A probabilistic framework for dialogue segmentation is presented and applied. The algorithm proposed has a clean probabilistic interpretation and performs well compared to [10, 4], especially on dialogue databases. There is still room for improvement, especially information about cooccurence of words could be included in the model as suggested by [2, 26, 38] and more work on prosodic features could be attempted. The algorithm was tested on a variety of spon-

Coherence feature	Link error rate in %						
	No region	Segmentation					
		boundary		equal size		both	
		NN	LM	NN	LM	NN	LM
CallHome Spanish							
none	45.9	43.4	42.6	38.3	39.3	36.5	37.8
keyword	38.6	35.8	34.1	36.2	35.6	34.7	33.6
speaker+ls	45.6	42.8	42.3	43.2	42.2	41.9	41.4
both	38.8	36.5	34.3	37.4	35.7	35.6	34.9
Meeting							
none	47.8	38.9	37.4	42.1	45.1	41.5	46.0
keyword	37.6	36.2	37.9	37.9	36.9	37.1	39.5
speaker+ls	35.6	38.0	36.7	40.7	36.9	39.7	40.1
both	36.0	37.7	36.1	35.6	36.3	36.6	35.8
Santa Barbara							
none	49.0	40.1	41.0	43.4	48.9	44.1	48.2
keyword	39.0	40.0	38.1	39.7	40.4	39.5	40.8
speaker+ls	38.5	37.7	38.0	38.8	42.9	38.9	41.0
both	36.4	36.2	38.9	37.5	38.7	37.4	37.1

Table 4: Segmentation using word based regions: A neural network (NN) and language model classifier (LM) were trained to discriminate between different regions of topical segments, either just boundary vs. non-boundary, equal sized regions (begin/middle/end) or a combination of the two (both). The table shows the combination of these features with keyword repetition (keyword) and speaker initiative (speaker+LS) and the combination of the two (both).

taneous speech corpora and (stemmed) keywords, character n-gram and speaker initiative were used as features. Speaker initiative was found to perform almost as well as keyword repetition: This finding confirms the intuition that topical change is correlated with the activity the speaker are engaging in and their speaking rights which is encoded in their speaker initiative distribution. The results however also show that speaker initiative may fail in certain situations such as CallHome Spanish where only one speaker is dominant while the topic may be changing. Determining speaker initiative according to the definition here should be very tractable since speaker identity may be available trivially or it can be determined very effectively and reliably in meeting situations [24]. Modeling begin/middle/end as well as the boundary of a topical segment it was possible to exploit changes in the word and part of speech distribution.

Dialogue segmentation can therefore be done with a couple of features with similar performance. These features include lexical cohesion, speaker initiative and changes in the part of speech profile. The results presented here therefore fit the general claim of the author that dialogue style has to be an important feature in information access systems for spoken interactions. Speech recognition – even on hard corpora – didn't have a disastrous impact on the segmentation performance but resulted in significant degradation. Speaker initiative is a very powerful criterion which can likely be detected reliably without the need for expensive LVCSR.

8. REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. P. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, USA, February 1998.

[2] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34:177–210, 1999. Special Issue on Natural Language Learning (C. Cardie and R. Mooney, eds).

[3] J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31, March 1997.

[4] F. Choi. Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, Seattle, USA, 2000. Available with software at: http://www.cs.man.ac.uk/~choif/ http://xxx.lanl.gov/abs/cs.CL/0003083.

[5] D. Cook and L. B. Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, 1:231–255, 1994.

[6] J. Garofolo, C. Auzanne, and E. Voorhees. The TREC spoken document retrieval track : A success story. In E. Voorhees, editor, *Text Retrieval Conference (TREC) 8*, Gaithersburg, Maryland, USA, 1999. November 16-19.

[7] P. Geutner, M. Finke, and P. Scheytt. Adaptive vocabularies for transcribing multilingual broadcast news. In *ICASSP*, 1998.

[8] B. Grosz and C. Sidner. Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):172–204, 1986.

[9] M. Halliday and R. Hasan. *Cohesion in English.* Longman Group, 1976.

[10] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March 1997.

[11] J. Hirschberg and C. Nakatani. Acoustic indicators of topic segmentation. In *ICSLP*, Sidney, Australia, 1998.

[12] M.-Y. Kan, J. Klavans, and K. R. McKeown. Linear segmentation and segment significance. In *Proceedings of the 6th International Workshop on Very Large Corpora (WVLC-6)*, pages 197–205, Montreal, Canada, August 1998.

[13] R. Kuhn and R. de Mori. A cache-base natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and machine Intelligence*, 12(6):570–583, June 1990.

[14] Santa barbara corpus of spoken american english part-i, 2000.

[15] Callhome spanish dialogue act annotation, 2001. catalogue number LDC2001T61.

[16] Callhome spanish, lexicon, speech and transcripts, 1996. catalogue number LDC96L16, LDC96S35 catalogue number LDC96L16, LDC96S35 .

[17] P. Linell, L. Gustavsson, and P. Juvonen. Interactional dominance in dyadic communication: a presentation of initiative-response analysis. *Linguistics*, 26:415–442, 1988.

[18] Linguistic Data Consortium (LDC). Catalogue, 2000. http://www.ldc.upenn.edu/.

[19] W. C. Mann and S. Thomson. Rhetorical structure theory: Towards a functional theory of text

organization. *TEXT*, 8:243–281, 1988.

[20] D. Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* PhD thesis, Department of Computer Science, University of Toronto, December 1997. Also published as Technical Report CSRG-371, Computer Systems Research Group, University of Toronto.

[21] E. Mittendorf and P. Schäuble. Document and passage retrieval based on hidden markov models. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994.

[22] T. P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. van Melle, and P. Zellweger. "i'll get that off the audio": A case study of salvaging multimedia meeting records. In *CHI 97*, 1997.

[23] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependencies in stochastic language modeling. *Computer Speech and Language*, 8:1–35, 1994.

[24] Y. Pan and A. Waibel. The effects of room acoustics on MFCC speech parameters. In *Proceedings of the ICSLP*, Beijing, China, 2000.

[25] R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103, March 1997. 139.

[26] J. M. Ponte and B. W. Croft. Text segmentation by topic. In *Proceedings of the first European Conference on research and advanced technology for digital libraries*, 1997. U.Mass. Computer Science Technical Report TR97-18.

[27] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[28] F. Quek, D. McNeill, R. Bryll, C. Kirbas, H. Arslan, K. E. McCullough, and N. Furuyama. Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of the Computer Vision and Pattern Recognition CVPR*, 2000.

[29] J. C. Reynar. *Topic segmentation: Algorithms and applications.* PhD thesis, Computer and Information Science, University of Pennsylvenia, 1998. Institute for Research in Cognitive Science (IRCS), University of Pennsylvenia, Technical report: IRCS-98-21.

[30] K. Ries. HMM and neural network based speech act classification. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 497–500, Phoenix, AZ, March 1999.

[31] K. Ries, L. Levin, L. Valle, A. Lavie, and A. Waibel. Shallow discourse genre annotation in callhome spanish. In *Proceecings of the International Conference on Language Ressources and Evaluation (LREC-2000)*, Athens, Greece, May 2000.

[32] K. Ries and A. Waibel. Activity detection for information access to oral communication. In *Human Language Technology Conference*, Sand Diego, CA, USA, March 2001.

[33] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody modeling for automatic sentence and topic segmentation from speech. *Speech Communication*, 32(1-2):127–154, 2000. Special Issue on Accessing Information in Spoken Audio.

[34] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *In Proceedings of SIGIR*, 1999.

[35] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. In *ICASSP*, Salt Lake City, Utah, USA, 2001.

[36] M. A. Walker and S. Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *In Proc. 28th Annual Meeting of the ACL*, 1990.

[37] S. Whittaker, P. Hyland, and M. Wiley. Filochat: handwritten notes provide access to recorded conversations. In *In Proceedings of CHI94 Conference on Computer Human Interaction*, pages 271–277, 1994.

[38] Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden markov model approach to text segmentation and event tracking. In *Proceedings of ICASSP*, volume 1, pages 333–336, Seattle, WA, May 1998.

[39] H. Yu, T. Tomokiyo, Z. Wang, and A. Waibel. New developments in automatic meeting transcription. In *Proceedings of the ICSLP*, Beijing, China, October 2000.

[40] K. Zechner and A. Waibel. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING*, Saarbrücken, Germany, 2000.