TOPIC SEGMENTATION USING MARKOV MODELS ON SECTION LEVEL

Evgeny Matusov[1,2], *Jochen Peters*[1], *Carsten Meyer*[1] *and Hermann Ney*[2]

[1]Philips Research Laboratories
Weisshausstr. 2, D-52066 Aachen, Germany
e-mail: {Evgeny.Matusov, Jochen.Peters, Carsten.Meyer}@philips.com

[2]Lehrstuhl für Informatik VI
RWTH Aachen – University of Technology
D-52066 Aachen, Germany
e-mail: ney@informatik.rwth-aachen.de

ABSTRACT

Topic segmentation, i.e. the combined task of document segmentation and topic identification, is an interesting issue both from a theoretical point of view as well as for practical applications. Previous studies have mainly focussed on applications exposing rather weak correlations regarding the topic order (e.g. Broadcast News). In this work, we concentrate on documents following a typical structure regarding the sequence and organization of the individual sections. We propose an algorithm allowing to explicitly add such structures as additional knowledge sources by modeling the document structure on the level of complete sections. Specifically, we address the issues of explicit section length modeling and modeling of typical section start phrases. On a database of dictated reports, we show significant improvements over state-of-the-art approaches both on manually and automatically transcribed text. Moreover, we show that our approach is significantly more robust against recognition errors than a phrase matching approach exploiting merely the typical section start phrases.

1. INTRODUCTION

Topic segmentation — the combined task of document segmentation and topic identification — is theoretically challenging and has a number of practical applications. For example, it helps to improve the readibility of long text documents, allows to partition them into smaller units for further processing, allows to identify related information in a pool of documents, allows a quick access to specific information in a document etc. In this work, we focus on the case of topic segmentation for documents following a typical structure regarding the sequence and organization of the individual sections. In particular, we assume significant correlations in the oder of the observed topic sequences, in the section lengths belonging to the same topic, and in typical section start or end phrases. Moreover, we assume a given, restricted set of topic labels (each of which may correspond to a multitude of typcial section start phrases). Examples of such documents are medical reports, scientific articles, meeting protocols, legal documents etc. We propose an algorithm allowing to explicitly exploit these typical document structures as additional knowledge sources, by applying a generative approach using Markov models on the level of complete sections. We aim at an optimal combination of the additional knowledge sources for significant performance improvements compared to state-of-the-art algorithms both for manually and automatically transcribed texts. Detailed evaluation results are presented on a medical report database, significantly outperforming a simulated state-of-the-art algorithm and a simple cue phrase matching approach.

The rest of the paper is organized as follows: Section 2 briefly reviews previous work on topic segmentation. Section 3 highlights the main differences of our approach compared to previous algorithms. In Section 4, we give the theoretical formulation of our algorithm. Our database is described in section 5, and the error measures used are introduced in section 6. Section 7 describes our basic system configuration. We evaluate the explicit modeling of section lengths in section 8 and report further refinements in word emission modeling in section 9. In section 10, we present evaluation results and performance comparisons on automatically transcribed texts. Finally, our main findings and future directions of work are summarized in section 11.

2. PREVIOUS WORK

Text segmentation methods were previously mostly applied to the domain of Broadcast News (topic detection and tracking, TDT) [1] where news stories appear in almost arbitrary order. Two tasks have to be distinguished: The first task is the mere segmentation, i.e. the detection of story boundaries without providing meaningful labels for the found sections. Here, most approaches concentrate on finding specific indications of a topic switch. Such indications are usually expressed as binary features (e.g. prosodical, lexical and cue word features [6]). In a statistical framework, these features are often used to train the *posterior* probability of a boundary given the presence or absence of a certain feature combination. Two important techniques to model such dependencies are Maximum Entropy [2] and Decision Trees [4].

The second, extended task includes the labeling of the found sections. Here, we have a pre-defined set of topics $t = 1, \ldots T$, and each section has to be assigned to one of them. One natural approach is a *tagging* scheme where each word or sentence has to be tagged with a certain topic label t. Here, the segmentation is implicitly determined from locally changing labels. Again, Maximum Entropy techniques (also used for Named Entity Tagging [3]) can be applied for modelling the *posterior* probability of the topic label sequence given the observed words.

Alternatively, [5] proposed a *generative* approach for the tagging scheme which is based on Markov models and classical language modeling. Here, the joint probability of the document structure and observed text is optimized. This joint distribution is decomposed into the *prior* probability for the document structure and the *conditional* probability for the observed words given that structure. The prior probabilities are modeled with *Markov chains*: The given topics are viewed as Markov states. The topic sequence is modeled by *transition probabilities* between subsequent Markov states. (In [5] these probabilities are pooled into one global topic-switch and one topic-loop probability.) The conditional probability for the observed words is modelled by *topic-specific (unigram) language models* associated with the Markov states.

Although this approach allows an efficient training of the involved statistical models, it has some drawbacks: Due to the sentence-by-sentence tagging longer-ranging document structures on the *level of complete sections* cannot be captured. Applying transition probabilities on the sentence level results in an implicit section length distribution from the accumulated topic-loop probabilities. Unfortunately, the resulting monotonous, exponentially decaying length distribution is contrary to the experimental observation that very short sections are unlikely. Finally, since section boundaries are only implicitly given by changing tags, the typical evolution of a text within a section cannot be modelled without an extension of the algorithm.

3. MARKOV MODELS ON SECTION LEVEL

In this work we propose an algorithm that extends the generative approach based on Markov models by emphasizing typical document structures, allowing to exploit additional knowledge sources, e.g. explicit length modeling and characteristic phrases at the beginning or end of the individual sections.

Conceptually, we abandon the sentence-by-sentence tagging paradigm. Instead, we view the wanted *sections* as basic units. These sections are now specified not only by their topic but also by their size and location in the document. This allows us to model the document structure, i.e. the topic sequences, on the *level of complete sections*, optionally including longer ranging dependencies such as topic trigrams or position dependencies of topics. Furthermore, since the start and end position of each section is explicitly known, we may employ more realistic length models, thus replacing the inadequate implicit distributions used in [5]. The introduction of explicit length models will be shown to be a key advantage of the proposed method. (This comes at the cost of a more complex dynamic programming since we now perform a *two-dimensional* simultaneous optimization over the section boundaries and over the topic sequence in the text.) Finally, we may use more sophisticated emission models to capture crucial features of text evolution within each section. This is useful, if formulations in the beginning of a section usually differ from those used in the section “body”, or near its end.

4. THEORY

Our task is to find an optimal segmentation of the given word stream $w_1^N := w_1, ..., w_N$ into K sections (K is to be optimized) which are labeled by the topics $t_1^K := t_1, ..., t_K$ and characterized by the section end positions (word indices) $n_1^K := n_1, ..., n_K$. We are thus interested in:[1]

$$\underset{t_1^K, n_1^{K-1},\, K}{\operatorname{argmax}} \left\{ Pr(t_1^K, n_1^K, K | w_1^N) \right\} \tag{1}$$

Using Bayes rule and dropping the constant prior probability for w_1^N we arrive at the *generative* approach where we optimize over the product of the *prior* probability for the text structure and the *conditional* probability for the observed words given that structure:

$$\underset{t_1^K, n_1^{K-1},\, K}{\operatorname{argmax}} \left\{ Pr(t_1^K, n_1^K, K) \cdot Pr(w_1^N | t_1^K, n_1^K, K) \right\} \tag{2}$$

We will not model the distribution $p(K)$ over the number of sections explicitly and decompose the prior $Pr(t_1^K, n_1^K, K)$ into $Pr(t_1^K) \cdot Pr(n_1^K | t_1^K)$. Here, we approximate $Pr(t_1^K)$ by a product of *topic transition* probabilities $p(t_k | t_{k-1})$.[2] The probability $Pr(n_1^K | t_1^K)$ of the section end positions is decomposed into a product of topic-dependent probabilities $p(\Delta n_k | t_k)$ of the *section lengths* $\Delta n_k := n_k - n_{k-1}$. The conditional probability of the observed words is provided by topic-specific language models $p(w_n | t_k)$. They may be

[1] n_K is always $= N$ whence we optimize over n_1^{K-1}.

[2] More sophisticated models may include a dependency on the absolute section position k or longer-ranging topic M-grams.

extended to $p(w_n|t_k, n-n_{k-1})$ to include a dependency on the position of each word within its section.

By introducing the modeling assumptions into Eq. (2), we get the following optimization criterion:

$$\underset{t_1^K, n_1^{K-1}, K}{\operatorname{argmax}} \left\{ \prod_{k=1}^{K} \left(p(t_k|t_{k-1}) \cdot p(\Delta n_k \mid t_k) \cdot \prod_{n=n_{k-1}+1}^{n_k} p(w_n|t_k, n-n_{k-1}) \right) \right\} \quad (3)$$

This problem can be solved using dynamic programming in which we perform a *two-dimensional* simultaneous optimization over the section boundaries and over the topics.

Note that our formal approach allows to differentiate between two consecutive words and two consecutive sections with the same topic. We have the freedom to allow or explicitly forbid a transition from some topic to itself. This may be controlled by $p(t_k|t_{k-1})$. Such a distinction is not possible in the sentence tagging approach of [5].

Topic segmentation with this algorithm can be performed either on the word level or on the sentence level. In the second case, the allowed section end positions n_k are restricted to a set of pre-defined *sentence end* positions (e.g. exploiting punctuation). This reduces the computational complexity and makes the segmentation more robust.

5. DATABASE

We present experimental results on a medical report database. The training corpus consists of manual transcripts of 4075 dictated reports, with sections manually introduced by special formatting and inserted topic names. In order to reduce the variability in topic name formulations (e.g. singular versus plural formulations), we manually clustered the list of observed topic names with regard to similar names to arrive at a set of $T = 51$ different meaningful, coherent topic labels. The average number of sections in a report was 6, and the average section length was 65 words. An analysis of the data showed that about 85% of the sections were started with an indicative cue phrase (related to the topic label and the section content, e.g. “next is summary”). However, there was a wide variety in the formulations of these start phrases (see also section 10.1).

Two data sets with slightly different characteristics were used for testing: Test set A consists of 186 reports (with an average of 400 words and 8 sections per report), test set B of 67 reports (on average 460 words and 8 sections per report).

6. ERROR MEASURES

In order to measure the performance of the implemented algorithm, three different error metrics were used.

The *Co-Occurrence Agreement Probability* P_D is capable of assessing the placement of section boundaries independent of the assigned topic labels. This quantity is also used as an evaluation criterion in the TDT task and is described in detail in [2]. Here, for any two words drawn randomly from the text w_1^N, P_D is defined as the probability to *correctly* identify them as either belonging to the same section, or to different sections. P_D favors exact boundary matches, but also takes fairly close matches into account. In practical applications, only word pairs (w_i, w_j) inside a window of fixed length are considered. Following [2] we set the window length to half the average reference section length. In tables 1 and 2 we give the “co-occurence agreement error rate COAER”, defined as $1 - P_D$.

The *Word Labeling Error Rate* (WLER) is defined by a word-by-word comparison of the automatically assigned topic and the reference topic. One error is counted for each incorrectly labeled word, and the errors are divided by the text length N. This simple metric assesses both segmentation and labeling decisions.

Finally, we used an error metric that judges the correctness of the hypothesized sequence of topics: the *Topic Levenshtein Error Rate* (TLER). For every report, the reference and automatic segmentation each yield a topic sequence. TLER is defined as the Levenshtein editing distance between the two topic sequences, divided by the number of sections in the reference segmentation. This error metric allowed for interpretation of topic detection errors in terms of insertions, deletions, and substitutions.

The three error rates provided the basis for a versatile evaluation of our algorithm.

7. BASIC SYSTEM CONFIGURATION

The topic transition probabilities $p(t|t')$ were trained on the level of complete sections. Each report in the training corpus was converted to its topic sequence (out of 51 topic labels). An additional fictitious topic t_0 was added to represent the beginning and the end of each report. The probabilities $p(t|t_0)$ and $p(t_0|t)$, respectively, were used for initialization and final maximization in the dynamic programming algorithm. The perplexity of the topic sequence of the training corpus with respect to a topic transition bigram was 6.5.

The available training data was insufficient to train section length probabilities $p(\Delta n|t)$ for each topic t. We thus grouped all topics into several length classes and estimated *pooled* section length models from aggregated histograms for each length class.

In the basic configuration of our system, we used topic-specific word unigram probabilities $p(w|t)$ independent of the word’s section internal position.

First evaluation experiments showed that the algorithm performs significantly better on the sentence level than on the word level. The error rates for the basic configuration

Table 1. Performance comparison of the simulated generative approach of [5] and our section-level Markov model with explicit optimization over section boundaries in its basic and final configuration. Error rates in %, see section 6.

	TEST SET A		
ERROR RATES (%):	COAER	WLER	TLER
implicit length modeling (simulation of [5])	14.6	41.6	60.4
explicit length modeling (basic configuration)	12.0	40.4	47.9
explicit length modeling (final configuration)	5.0	27.6	25.6

of the system (sentence-level segmentation) are given in Table 1, line 2. For comparison, the first line in table 1 gives results for an algorithm without explicit section length modeling, described in the following section.

8. EVALUATION OF EXPLICIT SECTION LENGTH MODELING

A contrast experiment was performed to support the statement that the inclusion of explicit length modeling and explicit optimization over section boundaries outperforms the basic approach introduced in [5] on our database. We simulated this sentence-by-sentence labeling approach by estimating the topic transition probabilities $\tilde{p}(t|t')$ on the level of sentences (i.e. each topic index t appeared as many times in the training corpus as there are sentences in the section labeled with t). Instead of using pooled topic-loop and switch probabilities as in [5], $\tilde{p}(t|t')$ is modelled individually for each topic pair. We replaced our explicit length models $p(\Delta n|t)$ by $(\tilde{p}(t|t))^{\Delta b-1}$, where Δb is defined as the number of sentences containing the Δn words. This reflects the accumulated topic-loop probabilities.

The results of this simulation show significantly larger error rates than those obtained with explicit length models (compare lines 1 and 2 in table 1). The monotonous implicit length distributions favor short topics and result in an increased topic insertion rate (reflected in TLER). In contrast, even in its basic configuration, the section-level Markov model avoids erroneous insertions of short sections to a large extent. This can be attributed to the low probabilities for too short sections from our explicit length models. Moreover, our approach allows to explicitly model typical section start phrases allowing strong further performance improvements (line 3 in table 1), as explained in the next section. On test set B, similar results were observed.

9. WORD EMISSION PROBABILITIES

One observation in our first experiments was that simple topic-specific word unigrams did not always provide sufficient information for correct segmentation and labeling decisions. First, we lack a mechanism to place boundaries immediately before an indicative start phrase since the words' distances from the boundary are not considered by a simple unigram. Second, some words were so indicative of a certain short topic that their occurrence within a longer topic immediately triggered an erroneous topic switch. Solutions to these two problems are presented below.

9.1. Topic Start/Continue Models

We extended the word emission modeling by training separate *topic start* and *topic continue* unigrams to account for the fact that some words (e.g. from indicative cue phrases) are more likely to begin a section, while others usually represent the internal content of a section. The topic start models $p_s(w_n|t)$ were estimated on the first θ words of each section, and the topic continue models $p_c(w_n|t)$ on the remainder of the sections. Correspondingly, our dynamic programming algorithm predicted the first θ words in each assumed section by a start model and the rest of the section by a continue model:

$$p(w_n|t, n-i) := \begin{cases} p_s(w_n|t) & : \quad n-i \leq \theta \\ p_c(w_n|t) & : \quad n-i > \theta \end{cases} \qquad (4)$$

Experimentally, the optimal θ was found to be 5. Significant improvements were observed for this modeling approach: Most of the start phrases were now used by the algorithm to predict a correct topic transition.

In a refined version, we introduced a statistical modeling of θ. This was achieved by introducing a *mixture model* for the emission of the first words in a section, each mixture encoding a switch from the start to the continue state at the respective position $m \in [0, \ldots, \theta]$. Assuming that every length m of a start phrase is equally probable, we used uniform prior probabilities for all $(\theta + 1)$ paths. This resulted in further significant performance improvements, especially regarding the number of topic deletion and substitution errors. Especially in cases when either no dictated cue phrases were present in the beginning of a section or these cue phrases were short, the misleading influence of the topic start model was strongly reduced. Thus, we arrived at a flexible modeling approach effectively utilizing indicative section start phrases, if present.

9.2. Smoothing of Topic Continue Models

In spite of using most of the indicative cue phrases for placing correct section boundaries and topic labels, we still observed too many topic insertions significantly exceeding the number of deletions and substitutions. A manual inspection revealed that many lenghty sections are still "broken up" into several shorter topic segments. This "aggressive" segmentation behavior may be explained by the following reasoning: In our database, we observed some more general topics which are discussed in long sections, as well as highly specific topics which are typically discussed in short sections involving a highly specific vocabulary. However, some of these "highly specific" words (e.g. "education") might also appear within a longer section belonging to a more general topic (e.g. "family history") and employing a more general, less restricted vocabulary. Due to this situation, the probability of a word given a more general topic is much lower than its probability given a highly specific topic, even if the absolute frequency of the word is more or less equal in both topics. For such words, the bias *against* longer and more general topics is accumulated in the product of emission probabilities as in (3). This results in favoring many short topics by the automatic segmentation algorithm.

To overcome this problem, we tested various smoothing techniques to reduce the specificity (the discriminative dependency on the topic) of the emission models. In our system, it makes sense to strongly smooth only the topic continue models because the topic start models were introduced especially for "boosting" the influence of topic-specific start phrases.

Two straightforward smoothing techniques for $p_c(w|t)$ are (1) heavy absolute discounting and interpolation with the topic-independent language model $p(w)$ or (2) the linear interpolation of $p_c(w|t)$ with $p(w)$. Both methods reduced the topic insertion errors, but at the expense of significant increases in the deletion and/or substitution errors.

Another approach is to "scale" the product of topic continue probabilities with an exponent $\alpha < 1$ in order to reduce the accumulated bias towards too strong topic discrimination. With such a scaling, the information about the text structure (typical topic sequences, section length) gets more impact and the influence of the observed words is reduced. We have to keep in mind, however, that in certain segmentation hypotheses the scaled topic continue probabilities compete with the unscaled – and thus typically lower – topic start probabilities for the same words. This introduces a bias *against* the paths using the start models which again results in higher topic deletion error rates.

To eliminate the bias, we propose the log-linear interpolation of each topic-specific continue model with the topic-independent unigram:

$$p_c^{LOGLIN}(w|t) := \frac{p_c^{\alpha}(w|t) \cdot p^{(1-\alpha)}(w)}{\sum_{w'} p_c^{\alpha}(w'|t) \cdot p^{(1-\alpha)}(w')} \qquad (5)$$

With this approach, the discrimination between topics is reduced as with the simple scaling, but the bias against starting new sections is eliminated since $p_c^{LOGLIN}(w|t)$ is of the same order as $p_s(w_n|t)$ due to the interpolation with $p(w)$ taken to the power of $(1-\alpha)$. We thereby achieved a good balance between topic insertions and deletions.

The topic segmentation performance of our system in its final configuration is presented in Table 1, line 3. This final configuration included the topic transition bigram $p(t|t')$, length models, and the optimized emission model (start and continue models and log-linear interpolation of the continue models with $\alpha = 0.2$). In comparison with the basic configuration, the number of topic insertion errors was drastically reduced. The section boundaries are now placed quite exactly, as expressed by the low co-occurence agreement error rate of 5%. The appearance of indicative start phrases almost always results in correct topic segmentation decisions. On the other hand, sections without indicative start phrases are also detected rather well, based on the complete content of a section and on structural information.

10. PERFORMANCE ON AUTOMATICALLY TRANSCRIBED TEXTS

An important issue for a topic segmentation algorithm is its performance on automatically transcribed texts. To this end, we evaluated our section-level Markov model approach on the output of an automatic speech recognition (ASR) system and compared its performance to the simulation of the approach reported in [5] (see section 8) and a simple cue phrase matching approach (see below). Experimental results[3] are reported on test set B (table 2). The reference segmentation of the ASR texts was derived from a Levenshtein alignment with the segmented manual transcriptions. The word error rate of the ASR transcriptions on test set B was 24.5%.

10.1. Cue phrase matching approach

For comparison, we evaluated a simple cue phrase matching approach merely looking for typical section start phrases. Here, we search for longest phrase matches after sentence separator symbols using a list of indicative section start phrases extracted from the training data (each with a minimum frequency of 5). For our data, this list contained 291 phrases, with an average of 5.7 different phrases per topic. Note that this method is only able to detect sections starting with an indicative cue phrase from this list. Furthermore, this method can produce "false alarms", if a phrase from the list (e.g. "education") appears in running text without indicating the start of a new section.

10.2. Performance comparison

Table 2 shows evaluation results for the section-level Markov model (in its final configuration), the simulation of [5] and the cue phrase matching approach, both for

[3] As before, all models were trained on the manual transcriptions.

Table 2. Performance comparison of the section-level Markov model algorithm with the simulated sentence-by-sentence labeling approach and the cue phrase matching method (manually / automatically transcribed reports). Error rates in %, see section 6.

Manual transcriptions of test set B			
Error Rates (%):	COAER	WLER	TLER
Section Markov model	6.9	31.8	34.4
Simulation of [5]	16.7	47.7	77.5
Cue phrase matching	6.5	35.3	43.6

ASR transcriptions of test set B			
Error Rates (%):	COAER	WLER	TLER
Section Markov model	8.5	35.0	42.2
Simulation of [5]	18.2	48.2	80.4
Cue phrase matching	9.5	49.2	58.0

the reference and the ASR transcriptions. As can be seen in the upper part of table 2, our segmentation algorithm outperforms the cue phrase matching method already for the reference transcriptions. This is due to the inability of the cue phrase matching method to detect sections with no indicative start phrase (accounting for more than 15% of all sections) and to learn the great variety of start phrase formulations.

As we go to the ASR transcriptions, the phrase matching method shows a considerable degradation since many indicative start phrases were not recognized properly and thus could not be found by a simple string matching. Therefore, the number of topic deletions increased dramatically. Under these considerations, the application of a cue phrase matching method does not seem appropriate even if the majority of the sections start with an indicative cue phrase.

Contrary to this severe degradation, the section-level Markov model showed only a moderate deterioration. This is remarkable since the ASR error rate was quite high. These findings support the expectation that the utilization of several knowledge sources (in particular, of the whole section content) renders our topic segmentation algorithm quite robust against recognition errors.

For both the manual and automatic transcriptions, the simulated sentence-by-sentence tagging method performed worst on our database, since it does neither utilize indicative start phrases nor explicit section lengths.

Similar findings were drawn from results on test set A.

11. CONCLUSIONS

We proposed a novel topic segmentation algorithm effectively utilizing additional knowledge sources related to section internal structures, like explicit section lengths and indicative start phrases. In a generative approach, these knowledge sources were effectively integrated by performing a (two-dimensional) simultaneous optimization over the section boundaries and the topic sequence in the text (“section-level Markov model”), allowing to utilize statistical information about typical topic sequences, section lengths and typical text phrases at any position within a section. On a database consisting of dictated reports, we obtained strong performance improvements compared to state-of-the-art algorithms both on manually and automatically transcribed texts. We showed that this is due to the flexibility of our algorithm to effectively use the additional knowledge sources related to section lengths and typical section start phrases. Moreover, our algorithm was found to be remarkably robust against recognition errors, due to the utilization of section content *and* typical start formulations which are modeled in a flexible way.

Our results strongly motivate to exploit all available statistical information on typical text structures for significant performance improvements in topic segmentation and text processing tasks, integrated in a profound theoretical framework.

We are currently working on extensions of our algorithm, allowing to include the dependency of a topic on its absolute position k into the topic prediction probabilities $p(t_k|t_{k-1}, k)$ and to use a topic trigram $p(t|t', t'')$.

12. REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, “Topic Detection and Tracking Pilot Study: Final Report,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 194–218.

[2] D.Beeferman, A. Berger, and J. Lafferty, “Statistical Models for Text Segmentation,” *Machine Learning*, vol. 34, pp. 1–34, 1999.

[3] A. Borthwick, *A Maximum Entropy Approach to Named Entity Recognition*, Ph.D. thesis, New York University, 1999.

[4] S. Dharanipragada, M. Franz, J. S. McCarley, S. Roukos, and T. Ward, “Story Segmentation and Topic Detection in The Broadcast News Domain,” in *Proc. of the DARPA Broadcast News Workshop*, 1999.

[5] P. van Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron, “Segmentation of Automatically Transcribed Broadcast News Text,” in *Proc. of the DARPA Broadcast News Workshop*, 1999, pp. 77–80.

[6] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, “Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation,” *Computational Linguistics*, vol. 27(1), pp. 31–57, 2001.