Museli: A Multi-Source Evidence Integration Approach to Topic Segmentation of Spontaneous Dialogue

Jaime Arguello
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jarguell@andrew.cmu.edu

Carolyn Rosé
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
cprose@cs.cmu.edu

Abstract

We introduce a novel topic segmentation approach that combines evidence of topic shifts from lexical cohesion with linguistic evidence such as syntactically distinct features of segment initial contributions. Our evaluation demonstrates that this hybrid approach outperforms state-of-the-art algorithms even when applied to loosely structured, spontaneous dialogue.

1 Introduction

Use of topic-based models of dialogue has played a role in information retrieval (Oard et al., 2004), information extraction (Baufaden, 2001), and summarization (Zechner, 2001). However, previous work on automatic topic segmentation has focused primarily on segmentation of expository text. We present Museli, a novel topic segmentation approach for dialogue that integrates evidence of topic shifts from lexical cohesion with linguistic indicators such as syntactically distinct features of segment initial contributions.

Our evaluation demonstrates that approaches designed for text do not generalize well to dialogue. We demonstrate a significant advantage of Museli over competing approaches. We then discuss why models based entirely on lexical cohesion fail on dialogue and how our algorithm compensates with other topic shift indicators.

2 Previous Work

Existing topic segmentation approaches can be loosely classified into two types: (1) lexical cohesion models, and (2) content-oriented models. The underlying assumption in lexical cohesion models is that a shift in term distribution signals a shift in topic (Halliday and Hassan, 1976). The best known algorithm based on this idea is TextTiling (Hearst, 1997). In TextTiling, a sliding window is passed over the vector-space representation of the text. At each position, the cosine correlation between the upper and lower region of the sliding window is compared with that of the peak cosine correlation values to the left and right of the window. A segment boundary is predicted when the magnitude of the difference exceeds a threshold.

One drawback to relying on term co-occurrence to signal topic continuity is that synonyms or related terms are treated as thematically-unrelated. One solution to this problem is using a dimensionality reduction technique such as Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997). Two such algorithms for segmentation are described in (Foltz, 1998) and (Olney and Cai, 2005).

Both TextTiling and Foltz's approach measure coherence as a function of the repetition of thematically-related terms. TextTiling looks for co-occurrences of terms or term-stems and Foltz uses LSA to measure semantic relatedness between terms. Olney and Cai's orthonormal basis approach also uses LSA, but allows a richer representation of discourse coherence, which is that coherence is a function of how much new information a discourse unit (e.g. a dialogue contribution) adds (*informativity)* and how relevant it is to the local context (*relevance*) (Olney and Cai, 2005).

Content-oriented models, such as (Barzilay and Lee, 2004), rely on the re-occurrence of patterns of topics over multiple realizations of thematically similar discourses, such as a series of newspaper articles about similar events. Their approach utilizes a hidden Markov model where states correspond to topics, and state transition probabilities correspond to topic shifts. To obtain the desired

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 9–12,
New York, June 2006.

number of topics (states), text spans of uniform length (individual contributions, in our case) are clustered. Then, state emission probabilities are induced using smoothed cluster-specific language models. Transition probabilities are induced by considering the proportion of documents in which a contribution assigned to the source cluster (state) immediately precedes a contribution assigned to the target cluster (state). Using an EM-like Viterbi approach, each contribution is reassigned to the state most likely to have generated it.

3 Overview of Museli Approach

We will demonstrate that lexical cohesion alone does not adequately mark topic boundaries in dialogue. Nevertheless, it can provide one meaningful source of evidence towards segmenting dialogue. In our hybrid Museli approach, we combined lexical cohesion with features that have the potential to capture something about the linguistic style that marks shifts in topic: word-unigrams, word-bigrams, and POS-bigrams for the current and previous contributions; the inclusion of at least one non-stopword term (contribution of content); time difference between contributions; contribution length; and the agent role of the previous and current contribution.

We cast the segmentation problem as a binary classification problem where each contribution is classified as NEW_TOPIC if the contribution introduces a new topic and SAME_TOPIC otherwise. We found that using a Naïve Bayes classifier (John & Langley, 1995) with an attribute selection wrapper using the chi-square test for ranking attributes performed better than other state-of-the-art machine learning algorithms, perhaps because of the evidence integration oriented nature of the problem. We conducted our evaluation using 10-fold cross-validation, being careful not to include instances from the same dialogue in both the training and test sets on any fold so that the results we report would not be biased by idiosyncratic communicative patterns associated with individual conversational participants picked up by the trained model.

Using the complete set of features enumerated above, we perform feature selection on the training data for each fold of the cross-validation separately, training a model with the top 1000 features, and applying that trained model to the test data. Examples of high ranking features confirm our intuition that contributions that begin new topic segments are syntactically marked. For example, many typical selected word bigrams were indicative of imperatives, such as lets-do, do-the, ok-lets, ok-try, lets-see, etc. Others included time oriented discourse markers such as now, then, next, etc.

To capitalize on differences in conversational behavior between participants assigned to different roles in the conversation (i.e., student and tutor in our evaluation corpora), we learn separate models for each role in the conversation[1]. This decision is based on the observation that participants with different agent-roles introduce topics with a different frequency, introduce different types of topics, and may introduce topics in a different style that displays their status in the conversation. For instance, a tutor may introduce new topics with a contribution that ends with an *imperative.* A student may introduce new topics with a contribution that ends with a *wh-question*.

4 Evaluation

In this section we evaluate Museli in comparison to the best performing state-of-the-art approaches, demonstrating that our hybrid Museli approach out-performs all of these approaches on two different dialogue corpora by a statistically significant margin ($p < .01$), in one case reducing the probability of error as measured by Beeferman's P_k to only 10% (Beeferman et al., 1999).

4.1 Experimental Corpora

We used two different dialogue corpora for our evaluation. The first corpus, which we refer to as the Olney & Cai corpus, is a set of dialogues selected randomly from the same corpus Olney and Cai selected their corpus from (Olney and Cai, 2005). The second corpus is a locally collected corpus of thermodynamics tutoring dialogues, which we refer to as the Thermo corpus. This corpus is particularly appropriate for addressing the research question of how to automatically segment dialogue for two reasons: First, the exploratory task that students and tutors engaged in together is more loosely structured than many task oriented domains typically investigated in the dialogue community, such as flight reservation or meeting scheduling. Second, because the tutor and student play asymmetric roles in the interaction, this corpus allows us to explore

[1] Dissimilar agent-roles occur in other domains as well (e.g. Travel Agent and Customer)

how conversational role affects how speakers mark topic shifts.

Table 1 presents statistics describing characteristics of these two corpora. Similar to (Passonneau and Litman, 1993), we adopt a flat model of topic-segmentation for our gold standard based on discourse segment purpose, where a shift in topic corresponds to a shift in purpose that is acknowledged and acted upon by both conversational agents. We evaluated inter-coder reliability over 10% of the Thermo corpus mentioned above. 3 annotators were given a 10 page coding manual with explanation of our informal definition of shared discourse segment purpose as well as examples of segmented dialogues. Pairwise inter-coder agreement was above 0.7 kappa for all pairs of annotators.

	Olney & Cai Corpus	**Thermo Corpus**
# Dialogues	42	22
Contributions/ Dialogue	195.40	217.90
Contributions/ Topic	24.00	13.31
Topics/Dialogue	8.14	16.36
Words/ Contribution	28.63	5.12

Table 1: Evaluation Corpora Statistics

4.2 Baseline Approaches

We evaluate Museli against the following algorithms: (1) Olney and Cai (Ortho), (2) Barzilay and Lee (B&L), (3) TextTiling (TT), and (4) Foltz.

As opposed to the other baseline algorithms, (Olney and Cai, 2005) applied their orthonormal basis approach specifically to dialogue, and prior to this work, report the highest numbers for topic segmentation of dialogue. Barzilay and Lee's approach is the state of the art in modeling topic shifts in monologue text. Our application of B&L to dialogue attempts to harness any existing and recognizable redundancy in topic-flow across our dialogues for the purpose of topic segmentation.

We chose TextTiling for its seminal contribution to monologue segmentation. TextTiling and Foltz consider lexical cohesion as their only evidence of topic shifts. Applying these approaches to dialogue segmentation sheds light on how term distribution in dialogue differs from that of expository monologue text (e.g. news articles).

The Foltz and Ortho approaches require a trained LSA space, which we prepared as described in (Olney and Cai, 2005). Any parameter tuning for approaches other than our hybrid approach was computed over the entire test set, giving competing algorithms the maximum advantage.

In addition to these approaches, we include segmentation results from three degenerate approaches: (1) classifying *all* contributions as NEW_TOPIC (ALL), (2) classifying *no* contributions as NEW_TOPIC (NONE), and (3) classifying contributions as NEW_TOPIC at *uniform intervals* (EVEN), corresponding to the average reference topic length (see Table 1).

As a means for comparison, we adopt two evaluation metrics: P_k and f-measure. An extensive argument of P_k's robustness (if k is set to ½ the average reference topic length) is present in (Beeferman, et al. 1999). P_k measures the probability of misclassifying two contributions a distance of k contributions apart, where the classification question is *are the two contributions part of the same topic segment or not?* Lower P_k values are preferred over higher ones. It equally captures the effect of false-negatives and false-positives and it favors near misses. F-measure punishes false positives equally, regardless of the distance to the reference boundary.

4.3 Results

Results for all approaches are displayed in Table 2. Note that lower values of P_k are preferred over higher ones. The opposite is true of F-measure. In both corpora, Museli performed significantly better than all other approaches ($p < .01$).

	Olney & Cai Corpus		**Thermo Corpus**	
	P_k	**F**	**P_k**	**F**
NONE	0.4897	--	0.4900	--
ALL	0.5180	--	0.5100	--
EVEN	0.5117	--	0.5132	--
TT	0.6240	0.1475	0.5353	0.1614
B&L	0.6351	0.1747	0.5086	0.1512
Foltz	0.3270	0.3492	0.5058	0.1180
Ortho	0.2754	0.6012	0.4898	0.2111
Museli	**0.1051**	**0.8013**	**0.4043**	**0.3693**

Table 2: Results on both corpora

4.4 Error Analysis

Results for all approaches are better on the Olney and Cai corpus than the Thermo corpus. The Thermo corpus differs profoundly from the Olney and Cai corpus in ways that very likely influenced the performance. For instance, in the *Thermo corpus* each dialogue contribution is an average of 5 words long, whereas in the *Olney and Cai corpus*

each dialogue contribution contains an average of 28 words. Thus, the vector space representation of the dialogue contributions is much more sparse in the Thermo corpus, which makes shifts in lexical coherence less reliable as topic shift indicators.

In terms of P_k, TextTiling (TT) performed worse than the degenerate algorithms. TextTiling measures the term-overlap between adjacent regions in the discourse. However, dialogue contributions are often terse or even contentless. This produces many islands of contribution-sequences for which the local lexical cohesion is zero. TextTiling wrongfully classifies all of these as starts of new topics. A heuristic improvement to prevent TextTiling from placing topic boundaries at every point along a sequence of contributions failed to produce a statistically significant improvement.

The Foltz and the orthonormal basis approaches rely on LSA to provide strategic semantic generalizations. Following (Olney and Cai, 2005), we built our LSA space using dialogue contributions as the atomic text unit. However, in corpora such as the Thermo corpus, this may not be effective because of the brevity of contributions.

Barzilay and Lee's algorithm (B&L) did not generalize well to either dialogue corpus. One reason could be that such probabilistic methods require that reference topics have significantly different language models, which was not true in either of our evaluation corpora. We also noticed a number of instances in the dialogue corpora where participants referred to information from previous topic segments, which consequently may have blurred the distinction between the language models assigned to different topics.

5 Current Directions

In this paper we address the problem of automatic topic segmentation of spontaneous dialogue. We demonstrated with an empirical evaluation that state-of-the-art approaches fail on spontaneous dialogue because word-distribution patterns alone are insufficient evidence of topic shifts in dialogue. We have presented a supervised learning algorithm for topic segmentation of dialogue that combines linguistic features signaling a contribution's function with lexical cohesion. Our evaluation on two distinct dialogue corpora shows a significant improvement over the state of the art approaches.

The disadvantage of our approach is that it requires hand-labeled training data. We are currently exploring ways of bootstrapping a model from a small amount of hand labeled data in combination with lexical cohesion (tuned for high precision and consequently low recall) and some reliable discourse markers.

Acknowledgments

This work was funded by Office of Naval Research, Cognitive and Neural Science Division, grant number N00014-05-1-0043.

References

Regina Barzilay and Lillian Lee (2004). Catching the drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of HLT-NAACL 2004*.

Doug Beeferman, Adam Berger, John D. Lafferty (1999). Statistical Models for Text Segmentation. *Machine Learning* 34 (1-3): 177-210.

Narjès Boufaden, Guy Lapalme, Yoshua Bengio (2001). Topic Segmentation: A first stage to Dialog-based Information Extraction. In *Proceedings of NLPRS 2001*.

P.W. Foltz, W. Kintsch, and Thomas Landauer (1998). The measurement of textual cohesion with latent semantic analysis. *Discourse Processes*, 25, 285-307.

M. A. K. Halliday and Ruqaiya Hasan (1976). *Cohesion in English*. London: Longman.

Marti Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragragh Subtopic Passages. *Computational Linguistics*, 23(1), 33 – 64.

George John & Pat Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of UAI* 2005.

Thomas Landauer, & Susan Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104, 221-240.

Douglas Oard, Bhuvana Ramabhadran, and Samuel Gustman (2004). Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. In *Proceedings of SIGIR* 2004.

Andrew Olney and Zhiqiang Cai (2005). An Orthonormal Basis for Topic Segmentation of Tutorial Dialogue. In *Proceedings of HLT-EMNLP* 2005.

Rebecca Passonneau and Diane Litman (1993). Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proceedings ACL* 2003.

Klaus Zechner (2001). Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains. In *Proceedings of SIGIR 2001*.