Topic Segmentation of Dialogue

Jaime Arguello
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15217
jarguell@andrew.cmu.edu

Carolyn Rosé
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15217
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 and final contributions. Our evaluation shows that this hybrid approach outperforms state-of-the-art algorithms even when applied to loosely structured, spontaneous dialogue. Further analysis reveals that using dialogue exchanges versus dialogue contributions improves topic segmentation quality.

1 Introduction

In this paper we explore the problem of topic segmentation of dialogue. 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), just to name a few applications. However, most previous work on automatic topic segmentation has focused primarily on segmentation of expository text. This paper presents a survey of the state-of-the-art in topic segmentation technology. Using the definition of topic segment from (Passonneau and Litman, 1993) applied to two different dialogue corpora, we present an evaluation including a detailed error analysis, illustrating why approaches designed for expository text do not generalize well to dialogue.

We first demonstrate a significant advantage of our hybrid, supervised learning approach called Museli, a multi-source evidence integration approach, over competing algorithms. We then extend the basic Museli algorithm by introducing an intermediate level of analysis based on Sinclair and Coulthard's notion of a dialogue exchange (Sinclair and Coulthard, 1975). We show that both our baseline and Museli approaches obtain a significant improvement when using perfect, hand-labeled dialogue exchanges, typically in the order of 2-3 contributions, as the atomic discourse unit in comparison to using the contribution as the unit of analysis. We further evaluate our success towards automatic classification of exchange boundaries using the same Museli framework.

2 Defining Topic

In the most general sense, the challenge of topic segmentation can be construed as the task of finding locations in the discourse where the focus shifts from one topic to another. Thus, it is not possible to address topic segmentation of dialogue without first addressing the question of what a "topic" is. We began with the goal of adopting a definition of topic that meets three criteria. First, it should be reproducible by human annotators. Second, it should not rely heavily on domain-specific knowledge or knowledge of the task structure. Finally, it should be grounded in generally accepted principles of discourse structure.

The last point addresses a subtle, but important, criterion necessary to adequately serve downstream applications using our dialogue segmentation. Topic analysis of dialogue concerns itself mainly with thematic content. However, boundaries should be placed in locations that are natural turning points in the discourse. Shifts in topic should be readily recognizable from surface characteristics of the language.

With these goals in mind, we adopted a definition of "topic" that builds upon Passonneau and Litman's seminal work on segmentation of monologue (Passonneau and Litman, 1993). They found that human annotators can successfully accomplish a flat monologue segmentation using an informal notion of speaker intention.

Dialogue is inherently hierarchical in structure. However, a flat segmentation model is an adequate approximation. Passonneau and Litman's pilot studies confirmed previously published results (Rotondo, 1984) that human annotators cannot reliably agree on a hierarchical segmentation of monologue. Using a stack-based hierarchical model of discourse, Flammia (1998) found that 90% of all information-bearing dialogue turns referred to the discourse purpose at the top of the stack.

We adopt a flat model of topic segmentation 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 participants. We place topic boundaries on contributions that introduce a speaker's intention to shift the purpose of the discourse, while ignoring expressed intentions to shift discourse purposes that are not taken up by the other participant. We adopt the dialogue contribution as the basic unit of analysis, refraining from placing topic boundaries within a contribution. This decision is analogous to Hearst's (Hearst, 1994, 1997) decision to shift the TextTiling induced boundaries to their nearest reference paragraph boundary.

We evaluated the reproducibility of our notion of topic segment boundaries by assessing inter-coder reliability over 10% of the corpus (see Section 5.1). Three 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 for all pairs of annotators.

3 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 regions 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 proposed solution to this problem is Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997). Two LSA-based algorithms for segmentation are described in (Foltz, 1998) and (Olney and Cai, 2005). Foltz's approach differs from TextTiling mainly in its use of an LSA-based vector space model. Olney and Cai address a problem not addressed by TextTiling or Foltz's approach, which is that cohesion is not just a function of the repetition of thematically-related terms, but also a function of the presentation of new information in reference to information already presented. Their orthonormal basis approach allows for segmentation based on *relevance* and *informativity*.

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 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). Following an EM-like approach, contributions are reassigned to states until the algorithm converges.

4 Overview of Museli Approach

We cast the segmentation problem as a binary classification problem where each contribution is classified as NEW_TOPIC if it introduces a new topic and SAME_TOPIC otherwise. 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. Table 1 lists our features.

Feature	Description
Lexical Cohesion	Cosine correlation of adjacent regions in the discourse. Term vectors of adjacent regions are stemmed and stopwords are removed.
Word-unigram	Unigrams in previous and current contributions
Word-bigram	Bigrams in previous and current contributions
Punctuation	Punctuation of previous and current contributions.
Part-of-Speech (POS) Bigram	POS-Bigrams in previous and current contributions.
Time Difference	Time difference between previous and current contribution, normalized by: (X – MIN)/ (MAX – MIN), where X corresponds to *this* time difference and MIN & MAX are with respect to the whole corpus.
Content Contribution	Binary-valued, is there a non-stopword term in the current contribution?
Contribution Length	Number of words in the current contribution, normalized by: (X – MIN) / (MAX – MIN).
Previous Agent[1]	Binary-valued, was the speaker of the previous contribution the *student* or the *tutor*?

Table 1. Museli Features.

We found that using a Naïve Bayes classifier with an attribute selection wrapper using the chi-square test for ranking attributes performed better than other state-of-the-art machine learning algorithms on our task, 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 to avoid biasing the trained model with idiosyncratic communicative patterns associated with individual dialogue participants.

To capitalize on differences in conversational behavior between participants assigned to different roles in the conversation (i.e., student and tutor), we learn separate models for each role. This decision is motivated by observations that participants with different speaker-roles, each with different goals in the conversation, 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 be more likely to introduce new topics with a contribution that ends with an *imperative.* A student may be more likely to introduce new topics with a contribution that ends with a *wh-question*. Dissimilar agent-roles also occur in other domains such as Travel Agent and Customer in flight booking scenarios.

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 output by our chi-squared feature selection wrapper confirm our intuition that initial and final contributions of a segment are marked differently. Moreover, the highest ranked features are different for our two speaker-roles. Some features highly-correlated with student-initiated segments are *am_trying*, *should*, *what_is*, and *PUNCT_question*, which relate to student questions and requests for information. Some features highly-correlated with tutor-initiated segments include *ok_lets*, *do*, *see_what*, and *BEGIN_VERB* (the POS of the first word in the contribution is VERB), which characterize imperatives, and features such as *now*, *next*, and *first*, which characterize instructional task ordering.

5 Evaluation

We evaluate Museli in comparison to the best performing state-of-the-art approaches, demonstrating that our hybrid Museli approach outperforms 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 P_k (Beeferman et al., 1999), to about 10%.

5.1 Experimental Corpora

We used two different dialogue corpora from the educational domain for our evaluation. Both corpora constitute of dialogues between a student and

[1] The current contribution's agent is implicit in the fact that we learn separate models for each agent-role (student & tutor).

a tutor (speakers with asymmetric roles) and both were collected via chat software. The first corpus, which we call the *Olney & Cai corpus*, is a set of dialogues selected randomly from the same corpus Olney and Cai obtained their corpus from (Olney and Cai, 2005). The dialogues discuss problems related to Newton's Three Laws of Motion. The second corpus, the *Thermo corpus*, is a locally collected corpus of thermodynamics tutoring dialogues, in which tutor-student pairs work together to solve an optimization task. Table 2 shows corpus statistics from both corpora.

	Olney & Cai Corpus	**Thermo Corpus**
#Dialogues	42	22
Conts./Dialogue	195.40	217.90
Conts./Topic	24.00	13.31
Topics/Dialogue	8.14	16.36
Words/Cont.	28.63	5.12
Student Conts.	4113	1431
Tutor Conts.	4094	3363

Table 2. Evaluation Corpora Statistics

Both corpora seem adequate for attempting to harness systematic differences in how speakers with asymmetric roles may initiate or close topic segments. The Thermo corpus is particularly appropriate for addressing the research question of how to automatically segment natural, *spontaneous* dialogue. The exploratory task is more loosely structured than many task-oriented domains investigated in the dialogue community, such as flight reservation or meeting scheduling. Students can interrupt with questions and tutors can digress in any way they feel may benefit the completion of the task. In the Olney and Cai corpus, the same 10 physics problems are addressed in each session and the interaction is almost exclusively a tutor initiation followed by student response, evident from the nearly equal number of student and tutor contributions.

5.2 Baseline Approaches

We evaluate Museli against the following four 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 the same way as described in (Olney and Cai, 2005). Any parameter tuning for approaches other than our Museli was computed over the entire test set, giving baseline 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), separated by the average reference topic length (see Table 2).

As a means for comparison, we adopt two evaluation metrics: P_k and f-measure. An extensive argument in support of P_k's robustness (if k is set to ½ the average reference topic length) is presented 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?* P_k is the likelihood of misclassifying two contributions, thus lower P_k values are preferred over higher ones. It equally captures the effect of false-negatives and false-positives and favors predictions that that are closer to the reference boundaries. F-measure punishes false positives equally, regardless of their distance to reference boundaries.

5.3 Results

Table 3 shows our evaluation results. Note that lower values of P_k are preferred over higher ones. The opposite is true of F-measure. In both corpora, the Museli approach performed significantly better than all other approaches ($p < .01$).

	Olney and Cai Corpus		Thermo Corpus	
	P_k	F	P_k	F
NONE	0.4897	--	0.4900	--
ALL	0.5180	--	0.5100	--
EVEN	0.5117	--	0.5131	--
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 3. Results on both corpora

5.4 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 on average 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 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 coherence is zero. TextTiling wrongly 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 Ortho approaches rely on LSA to provide strategic semantic generalizations capable of detecting shifts in topic. Following (Olney and Cai, 2005), we built our LSA space using dialogue contributions as the atomic text unit. In corpora such as the Thermo corpus, however, this may not be effective due to the brevity of contributions.

Barzilay and Lee's algorithm (B&L) did not generalize well to either dialogue corpus. One reason could be that probabilistic methods, such as their approach, 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.

6 Dialogue Exchanges

Although results are reliably better than our baseline algorithms in both corpora, there is much room for improvement, especially in the more spontaneous Thermo corpus. We believe that an improvement can come from a multi-layer segmentation approach, where a first pass segments a dialogue into dialogue exchanges and a second classifier assigns topic shifts based on *exchange initial contributions*. Dialogue is hierarchical in nature. Topic and topic shift comprise only one of the many lenses through which dialogue behaves in seemingly structured ways. Thus, it seems logical that exploiting more fine-grained sub-parts of dialogue than our definition of topic might help us do better at predicting shifts in topic. One such subpart of dialogue is the notion of dialogue exchange, typically between 2-3 contributions.

Stubbs (1983) motivates the definition of an exchange with the following observation. In theory, there is no limit to the number of possible responses to the clause "*Is Harry at home?*". However, constraints are imposed on the interpretation of the contribution that follows it: *yes* or *no*. Such a constraint is central to the concept of a dialogue exchange. Informally, an exchange is made from an initiation, for which the possibilities are open-ended, followed by dialogue contributions that are pre-classified and thus increasingly restricted. A contribution is part of the next exchange when the constraint on its communicative act is lifted.

Sinclair and Coulthard (1975) introduce a more formal definition of exchange with their Initiative-Response-Feedback or IRF structure. An initiation produces a response and a response happens as direct consequence to an initiation. Feedback serves to close an exchange. Sinclair and Coulthard posit that if exchanges constitute the minimal unit of interaction, IRF is a primary structure of interactive discourse in general.

To measure the benefits of exchange boundaries in detecting topic shift in dialogue, we coded the Thermo corpus with exchanges following Sinclair

and Coulthard's IRF structure. The coder who labeled dialogue exchanges had no knowledge of our definition of topic or our intention to do topic-analyses of the corpus. Any correlation between exchange boundaries and topic boundaries is not a bias introduced during the hand-labeling process.

7 Topic Segmentation with Exchanges

In our corpus, as we believe is true in domain-general dialogue, knowledge of an exchange-boundary increases the probability of a topic-boundary significantly. One way to quantify this relation is with the following observation. In our experimental Thermo corpus, there are 4794 dialogue contributions, 360 topic shifts, and 1074 exchange shifts. Using maximum likelihood estimation, the likelihood of being correct if we say that a randomly chosen contribution is a topic shift is 0.075 (# topic shifts / # contributions). However, the likelihood of being correct if we have prior knowledge that an exchange-shift also occurs in that contribution is 0.25. Thus, knowledge that the contribution introduces a new exchange increases our confidence that it also introduces a new topic. More importantly, the probability that a contribution does not mark a topic shift, given that it does not mark an exchange-shift, is 0.98. Thus, exchanges show great promise in narrowing the search-space of tentative topic shifts.

In addition to possibly narrowing the space of tentative topic-boundaries, exchanges are helpful in that they provide more coarse-grain building blocks for segmentation algorithms that rely on term-distribution as a proxy for dialogue coherence, such as TextTiling (Hearst, 1994, 1997), the Foltz algorithm (Foltz, 1998), Orthonormal Basis (Olney and Cai, 2005), and Barzilay and Lee's content modeling approach (Barzilay and Lee, 2004). At the heart of all these approaches is the assumption that a change in term distribution signals a shift in topic. When applied to dialogue, the major weakness of these approaches is that contributions are often times contentless: terse and absent of thematically meaningful terms. Thus, a more coarse-grained discourse unit is needed.

8 Barzilay and Lee with Exchanges

Barzilay and Lee (2004) offer an attractive frame work for constructing a context-specific Hidden Markov Model (HMM) of topic drift. In our initial evaluation, we used dialogue contributions as the atomic discourse unit. Using contributions, our application of Barzilay and Lee's algorithm for segmenting dialogue fails at least in part because the model learns states that are not thematically meaningful, but instead relate to other systematic phenomena in dialogue, such as fixed expressions and discourse cues. Figure 1 shows the cluster (state) size distribution in terms of the percentage of the total discourse units (exchanges vs. contributions) in the Thermo corpus assigned to each cluster. In the horizontal axis, clusters (states) are sorted by size from largest to smallest.

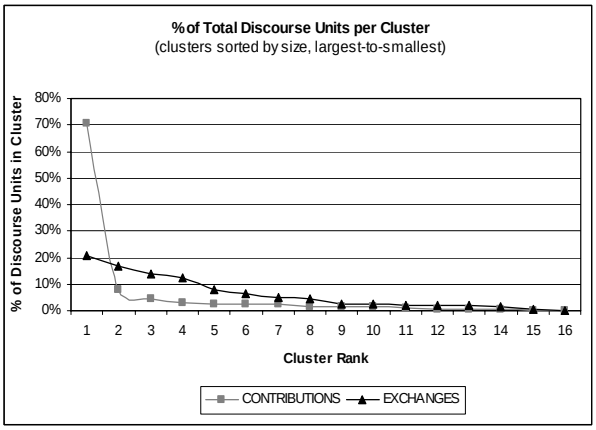

Figure 1. Exchanges produce a more evenly distributed cluster size distribution.

The largest cluster contains 70% of all contributions in the corpus. The second largest cluster only generates 10% of the contributions. In contrast, when using exchanges as the atomic unit, the cluster size distribution is less skewed and corresponds more closely to a topic analysis performed by a domain expert. In this analysis, the number of desired cluster (states), which is an input to the algorithm, was set to 16, the same number identified in a domain expert's analysis of the Thermo corpus. Examples of such topics include high-level ones such as *greeting*, *setup initialization*, and *general thermo concepts*, as well as task-specific ones like *sensitivity analysis* and *regeneration*.

A closer examination of the clusters (states) confirms our intuition that systematic topic-independent phenomena in dialogue, coupled with the terse nature of contributions in spontaneous dialogue, leads to an overly skewed cluster size distribution. Examining the terms with the highest emission probabilities, the largest states contain

topical terms like *cycle*, *efficiency*, *increase*, *quality*, *plot*, and *turbine* intermixed with terms like *think*, *you*, *right*, *make*, *yeah*, *fine*, and *ok*. Also the sets of topical terms in these larger states do not seem coherent with respect to the expert induced topics. This suggests that thematically ambiguous fixed expressions blur the distinction between the different topic-centered language models, producing an overly heavy-tailed cluster size distribution.

One might argue that a possible solution to this problem would be to remove these fixed expressions as part of pre-processing. However, that requires knowledge of the particular domain and knowledge of the interaction style characteristic to the context. We believe that a more robust solution is to use exchanges as the atomic unit of discourse.

9 Evaluation with Exchanges

To show the value of dialogue exchanges in topic segmentation, in this section we re-formulate our problem from classifying contributions into NEW_TOPIC and SAME_TOPIC to classifying exchange initial contributions into NEW_TOPIC and SAME_TOPIC. For all algorithms, we consider only predictions that coincide with hand-coded exchange initial contributions. We show that, except for our own Museli approach, using exchange boundaries improves segmentation quality across *all* algorithms ($p < .05$) when compared to their respective counterparts that ignore exchanges. Using exchanges gives the Museli approach a significant advantage based on F-measure ($p < .05$), but only a marginally significant advantage based on P_k. These results confirm our intuition that what gives our Museli approach an advantage over baseline algorithms is its ability to harness the lexical, syntactic, and phrasal cues that mark shifts in topic. Given that shift-in-topic correlates highly with shift-in-exchange, these features are discriminatory in both respects.

Of the degenerate strategies in section 5.2, only ALL lends itself to our reformulation of the topic segmentation problem. For the ALL heuristic, we classify *all* exchange initial contributions into NEW_TOPIC. This degenerate heuristic alone produces better results than all algorithms classifying utterances (Table 4). In our implementation of TextTiling (TT) with exchanges, we only consider predictions on contributions that coincide with exchange initial contributions, while ignoring predictions made on contributions that do not introduce a new exchange. Consistent with our evaluation methodology from Section 5, we optimized the window size using the entire corpus and found an optimal window size of 13 contributions. Without exchanges, the optimal window size was 6 contributions. The higher optimal window-size hints to the possibility that by using exchange initial contributions an approach based on lexical cohesion may broaden its horizon without losing precision.

	Thermo Corpus (Contributions)		**Thermo Corpus (Exchanges)**	
	P_k	**F**	**P_k**	**F**
NONE	0.4900	--	N/A	--
ALL	0.5100	--	0.4398	0.3809
EVEN	0.5132	--	N/A	--
TT	0.5353	0.1614	0.4328	0.3031
B&L	0.5086	0.1512	0.3817	0.3840
Foltz	0.5058	0.1180	0.4242	0.3296
Ortho	0.4898	0.2111	0.4398	0.3813
Museli	**0.4043**	**0.3693**	**0.3737**	**0.3897**

Table 4. Results using perfect exchange boundaries

In this version of B&L, we use exchanges to build the initial clusters (states) and the final HMM. B&L with exchanges significantly improves over B&L with contributions, in terms of both P_k and F-measure ($p < .005$) and significantly improves over our ALL heuristic (where all exchange initial contributions introduce a new topic) in terms of P_k ($p < .0005$). Thus, its use of exchanges goes beyond merely narrowing the space of possible NEW_TOPIC contributions: it also uses these more coarse-grained discourse units to build a more thematically-motivated topic model.

Foltz's and Olney and Cai's (Ortho) approach both use an LSA space trained on the dialogue corpus. Instead of training the LSA space with individual contributions, we train the LSA space using exchanges. We hope that by training the space with more contentful text units LSA might capture more topically-meaningful semantic relations. In addition, only exchange initial contributions where used for the logistic regression training phase. Thus, we aim to learn the regression equation that best discriminates between exchange initial contributions that introduce a topic and those that do not. Both Foltz and Ortho improve over their non exchange counterparts, but neither improves over the ALL heuristic by a significant margin.

For Museli with exchanges, we tried both training the model using only exchange initial contributions, and applying our previous model to only exchange initial contributions. Training our models using only exchange initial contributions produced slightly worse results. We believe that the reduction of the amount of training data prevents our models from learning good generalizations. Thus, we trained our models using contributions (as in Section 5) and consider predictions only on exchange initial contributions. The Museli approach offers a significant advantage over TT in terms of P_k and F-measure. Using perfect-exchanges, it is not significantly better than Barzilay and Lee. It is significantly better than Foltz's approach based on F-measure and significantly better than Olney and Cai based on P_k ($p < .05$).

These experiments used hand coded exchange boundaries. We also evaluated our ability to automatically predict exchange boundaries. On the Thermo corpus, Museli was able to predict exchange boundaries with precision = 0.48, recall = 0.62, f-measure = 0.53, and $P_k = 0.14$.

10 Conclusions and Current Directions

In this paper we addressed 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 term distribution alone fails to provide adequate evidence of topic shifts in dialogue.

We have presented a supervised learning algorithm for topic segmentation of dialogue called Museli that combines linguistic features signaling a contribution's function with local context indicators. Our evaluation on two distinct corpora shows a significant improvement over the state-of-the-art algorithms. We have also demonstrated that a significant improvement in performance of state-of-the-art approaches to topic segmentation can be achieved when dialogue exchanges, rather than contributions, are used as the basic unit of discourse. We demonstrated promising results in automatically identifying exchange boundaries.

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*, 113 - 120.

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*.

Giovanni Flammia. 1998. *Discourse Segmentation of Spoken Dialogue, PhD Thesis*. Massachusetts Institute of Technology.

Peter Foltz, Walter Kintsch, and Thomas Landauer. 1998. The measurement of textual cohesion with LSA. *Discourse Processes*, 25, 285-307.

Michael 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.

Thomas Landauer and Susan Dumais. 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.*

Andrew Olney and Zhiqiang Cai. 2005. An Orthonormal Basis for Topic Segmentation of Tutorial Dialogue. In *Proceedings of HLT/EMNLP*. 971-978.

Rebecca Passonneau and Diane Litman. 1993. Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proceedings of ACL*, 148 – 155.

John Rotondo, 1984, *Clustering Analysis of Subject Partitions of Text*. Discourse Processes, 7:69-88

John Sinclair and Malcolm Coulthard. 1975. *Towards an Analysis of Discourse: the English Used by Teachers and Pupils*. Oxford University Press.

Michael Stubbs. 1983. *Discourse Analysis. A Sociolinguistic Analysis of Natural Language*. Basil Blackwell.

Klaus Zechner. 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. Thesis. Carnegie Mellon University.