Making Semantic Topicality Robust Through Term Abstraction*

Paul M. Heider
Department of Linguistics
University at Buffalo
The State University of New York
Buffalo, NY 14260, USA
pmheider@buffalo.edu

Rohini K. Srihari
Janya, Inc.
1408 Sweet Home Road
Suite 1
Amherst, NY 14228
rohini@cedar.buffalo.edu

Abstract

Despite early intuitions, semantic similarity has not proven to be robust for splitting multi-party interactions into separate conversations. We discuss some initial successes with using thesaural headwords to abstract the semantics of an utterance. This simple profiling technique showed improvements over baseline conversation threading models.

1 Introduction

Topic segmentation is the problem of dividing a document into smaller coherent units. The segments can be hierarchical or linear; the topics can be localized or distributed; the documents can be newswire or chat logs. Of course, each of these variables is best analyzed as continuous rather than discrete. Newswire, for instance, is a more formal, monologue-style genre while a chat log tends towards the informal register with different conversations interwoven.

We present a topic segmenter which uses semantics to define coherent conversations within a larger, multi-party document. Using a word's thesaurus entry as a proxy for its underlying semantics provides a domain-neutral metric for distinguishing conversations. Also, our classifier does not rely on metalinguistic properties that may not be robust across genres.

*The first author was partially funded through a fellowship from the SUNY at Buffalo Department of Linguistics and partially through a research assistantship at Janya, Inc. (http://www.janyainc.com, Air Force Grant No.s FA8750-07-C-0077 and FA8750-07-D-0019, Task Order 0004)

2 Background

Most work on lexical cohesion extends from Halliday and Hasan (1976). They formalize a text as any semantic unit realized through sentences. Linguistic features found to justify binding sentences together into Halliday and Hasan's notion of a text include pronouns (Hobbs, 1979; Kehler, 2000), lexical overlap (Hearst, 1997; Kozima, 1993; Morris and Hirst, 1991), cue phrases (Manning, 1998), and discourse markers (Power et al., 2003; Reynar, 1999; Beeferman et al., 1999), among others. Of course, most of this earlier work assumes the sentences constituting any text are contiguous. Thus, a document is comprised of a series of semantic units that progress from one to the next with no returns to old topics.

Multi-party interactions[1] abide by a different set of assumptions. Namely, a multi-party interaction can include multiple floors (Aoki et al., 2006). Much like at a cocktail party, we can expect more than a single conversation at every given time. These different conversational floors are the major semantic units a topic segmentation algorithm must recognize. Spoken chat models (Aoki et al., 2006; Aoki et al., 2003) can make a simplifying assumption that speakers tend to only participate in one conversation at a time. However, in text chat models, Elsner and Charniak (2008) show that speakers seem to participate in more conversations roughly as a function of how talkative they are (cf. Camtepe et al., 2005). In both modalities, speaker tendency to stay on the same topics is a robust cue for conversational coher-

[1] See O'Neill and Martin (2003) for an analysis of differences between two- and multi-party interactions.

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 46–51,
Boulder, Colorado, June 2009.

ence (Elsner and Charniak, 2008; Acar et al., 2005).

Despite the initial intuitions of Halliday and Hasan (1976), semantic similarity has not proven to be a robust cue for multi-party topic segmentation. For instance, Acar et al. (2005) and Galley et al. (2003) used word repetition in their definition of coherence but found that words common to too many conversations hurt modeling performance. Elsner and Charniak (2008) used frequency binning based on the entire document to reduce the noise introduced by high-frequency words. Unfortunately, binning requires a priori knowledge of the relative frequencies of words.[2] Additionally, those authors used an on-topic/off-topic word list to bifurcate technical and non-technical utterances. Again, this technique assumes prior knowledge of the strongest on-topic cue words.

Since semantic repetition is clearly useful but simple word repetition is not a reliable measure, we investigated other measures of semantic relatedness. Elsner and Charniak (2008) conceded that context-based measures like LSA (Deerwester et al., 1990) require a clear notion of document boundary to function well. Dictionary-based models (Kozima and Furugori, 1993) are a step in the right direction because they leverage word co-occurrence within definitions to measure relatedness. The richer set of connections available in WordNet models should provide an even better measure of relatedness (Sussna, 1993; Resnik, 1995). Unfortunately, these measures have unequal distribution by part-of-speech and uneven density of lemmas by semantic domain.[3] Thesaurus-based models (Morris and Hirst, 1991) provide many of the same advantages as dictionary- and WordNet-based models.[4] In addition to the hierarchical relations encoded by the thesaurus, we can treat each thesaural category as one dimension of a topicality domain similar to the way Elsner and Charniak leveraged their list of technical terms. In sum, our model focuses on the abstraction of lemmas that is inherent to a thesaurus while limiting the domain-specific and a priori knowledge required by a classifier to divide multi-party interactions into separate conversational floors.

3 Model

At a high level, our chat topic segmenter works like most other classifiers: each input line is tokenized[5], passed to a feature analyzer, and clustered with related lines. Unlike traditional topic segmentation models, each input represents a new utterance in the chat log. These utterances can range from single words to multiple sentences. Another aspect of our model (although not unique to it) is the on-line classification of text. We aim to model topic segmentation as if our classifier were sitting in a chat room, and not as a post-process.

While our feature analyzer focuses on semantic markers, interlocutor names are also recorded. Two intuitions were implemented with respect to individuals' names: continued affiliation with old conversations and naming interlocutors to focus attention. All else being equal, one would assume a speaker will continue in the conversations she has already participated in. Moreover, she will most likely continue with the last conversation she was part of. As the total number of conversations increases, the likelihood of sticking to the last conversation will decrease.

The second intuition derives from the observation in O'Neill and Martin (2003) that speakers accommodate for cocktail-style conversations by using direct mentions of interlocutors' names. We only model backward referencing names. That is, if a speaker uses the name of another user, we assume that the speaker is overtly affiliating with a conversation of the other user. Forward referencing is discussed under future work (see Section 6).

Following Budanitsky and Hirst (2006), we base our notion of semantic topicality on thesaural relations. Broadly speaking, two utterances are highly related if their tokenized words (hereafter, lemmas) co-occur in more of the same thesaural categories than not. We will defer further explanation of these features until we have explained our reference thesauri in Subsection 3.1. Unfortunately, many desirable and robust features are missing from our classifier. See Section 6 for a discussion of future work.

[2] One could use frequencies from a general corpus but that should only perform as well as a graded stop-word list.

[3] As one reviewer noted, some parts-of-speech may contribute more to a topic profile than others. Unfortunately, this empirical question must wait to be tested.

[4] Budanitsky and Hirst (2006) review the advantages.

[5] We used Semantex™ (Srihari et al., 2008).

In the final stage of our processing pipeline, we use a panel of experts to generate a simple weighted classification. Each feature described above contributes a roughly equal vote towards the final sorting decision. Barring a single strong preference or a cohort of weak preferences for one conversation, the model assumes the incoming utterance introduces a new conversational floor.

3.1 Thesauri

We chose two machine-readable and public-domain thesauri for our model: Roget's Thesaurus (1911) and Moby Thesaurus II (2002). Both are available from Project Gutenberg (*gutenberg.org*). In the compilation notes for Roget's Thesaurus, the editor mentions a supplement of 1,000+ words to the original work. A rough count shows 1,000 headwords (the basic grouping level) and 55,000 synonyms (any word listed under a headword). The second edition of Moby Thesaurus contains some 30,000 headwords and 2.5 million synonyms. Moby Thesaurus includes many newer terms than Roget's Thesaurus. Structurally, Roget's Thesaurus has a distinct advantage over Moby Thesaurus. The former includes a six-tiered category structure with cross-indexing between headwords. The latter is only organized into headword lists.

3.2 Metrics

As we mentioned above, our model uses three primary metrics in classifying a new utterance: conversation affiliations of the current speaker, conversation affiliations of any explicitly mentioned interlocutors, and semantic similarity. In the end, all the conversation affiliation votes are summed with the one conversation preferred by each of the three thesaural measures.[6] The input line is then merged with the conversation that received the most votes. Details for deriving the votes follow.

Every conversation a speaker has participated in receives a vote. Moreover, his last conversation gets additional votes as a function of his total number of conversations (see Equation 1). Likewise, every conversation a named interlocutor has participated in receives a vote with extra votes given to her last conversation as a function of how gregarious she is.

[6] In the long run, a list of conversations ranked by similarity score would be better than a winner-takes-all return value.

Headword Type	Weight Change
Direct Match	1
Co-hyponymous Headword	0.25
Cross-indexed Headword	0.75

Table 1: Spreading Activation Weights.

$$Vote = \frac{3}{ln(|Conversations_{speaker}|) + 1} \quad (1)$$

Each utterance is then profiled in terms of thesaural headwords. Every lemma in an utterance matching to some headword increments the activation of that headword by one.[7] A conversation's semantic profile is a summation of the profiles of its constituent sentences. In order to simulate the drift of topic in a conversation, the conversation's semantic profile decays with every utterance. Thus, more recent headwords will be more activated than headwords activated near the beginning of a conversation. Decay is modeled by halving the activation of a headword in every cycle that it was not topical.

Moreover, a third profile was kept to simulate spreading activation within the thesaurus. For this profile, each topical headword is activated. Every cross-indexed headword listed within this category is also augmented by a fixed degree. Finally, every headword that occupies the same thesaurus section is augmented. An overview of the weight changes is listed in Table 1. The specific weights fit the authors' intuitions as good baselines. These weights can easily be trained to generate a better model.

The similarity between a new line (the test) and a conversation (the base) is computed as the sum of match bonuses and mismatch penalties in Table 2.[8] Table 3 scores an input line (TEST) against two conversations (BASE_1 and BASE_2) with respect to four headwords (A, B, C, and D). In order to control for text size, we also computed the average headword

[7] Most other models include an explicit stop-word list to reduce the effect of function words. Our model implicitly relies on the thesaurus look-up to filter out function words. One advantage to our approach is the ability to preferentially weight different headwords or lemma to headword relations.

[8] Like with Table 1, these numbers reflect the authors' intuitions and can be improved through standard machine learning methods.

Headword Present			Test		
			Yes		No
		Avg?	Above	Below	
Base	Yes	Above	+1	+0.5	-0.1
		Below	+0.5	+1	-0.05
	No		-1	-0.5	+0.0001

Table 2: Similarity Score Calculations.

	A	B	C	D	Score
TEST	high	high	low	0	–
$BASE_1$	high	low	low	high	
	+1	+.5	+1	-.1	2.4
$BASE_2$	0	high	0	0	
	-1	+1	-.5	+.0001	-.4999

Table 3: A Example Similarity Scoring for Two Conversations. 'High' and 'low' refers to headword activation.

activation in a conversation. Intuitively, we consider it best when a headword is activated for both the base and test condition. Moreover, a headword with equally above average or equally below average activation is better than a headword with above average activation in the base but below average activation in the test. In the second best case, neither condition shows any activation in a headword. The penultimately bad condition occurs when the base contains a headword that the test does not. We do not want to penalize the test (which is usually smaller) for not containing everything that the base does. Finally, if the test condition contains a headword but the base does not, we want to penalize the conversation most.

4 Dataset

Our primary dataset was distributed by Elsner and Charniak (2008). They collected conversations from the IRC (Internet Relay Chat) channel ##LINUX, a very popular room on *freenode.net* with widely ranging topics. University students then annotated these chat logs into conversations. We take the collection of these annotations to be our gold standard for topic segmentation with respect to the chat logs.

4.1 Metrics

Elsner and Charniak (2008) use three major measures to compare annotations: a 1-to-1 comparison, a loc_3 comparison, and a many-to-1 comparison. The 1-to-1 metric tries to maximize the global conversation overlap in two annotations. The loc_3 scale is better at measuring local agreement. This score calculates accuracy between two annotations for each window of three utterances. Slight differences in a conversation's start and end are minimized. Finally, the many-to-1 score measures the entropy difference between annotations. In other words, simplifying a fine-grained analysis to a coarse-grained analysis will yield good results because of shared major boundaries. Disagreeing about the major conversation boundaries will yield a low score.

	E&C Annotators			Our Model
	Mean	Max	Min	
Conversations	81.33	128	50	153
Avg. Length	10.6	16.0	6.2	5.2

Table 4: General statistics for our model as compared with Elsner and Charniak's human annotators. Some numbers are taken from Table 1 (Elsner and Charniak, 2008).

	Mean	Max	Min
Inter-annotator	86.70	94.13	75.50
Our Model	65.17	74.50	53.38

Table 5: Comparative many-to-1 measures for evaluating differences in annotation granularity. Some numbers are taken from Table 1 (Elsner and Charniak, 2008).

5 Analysis

Compared with the gold standard, our model has a strong preference to split conversations into smaller units. As is evident from Table 4, our model has more conversations than the maximally splitting human annotator. These results are unsurprising given that our classifier posits a new conversation in the absence of contrary evidence. Despite a low 1-to-1 score, our many-to-1 score is relatively high (see Table 5). We can interpret these results to mean that our model is splitting gold standard conversations into smaller sets rather than creating conversations across gold standard boundaries.

A similar interaction of annotation granularity shows up in Table 6. Our 1-to-1 measures are just barely above the baseline, on average. On the other

As % of ...	Misclassified			All
Error Type	Mean	Max	Min	Mean
Mismatch	25.84	23.78	28.13	11.93
Split Error	62.96	63.11	63.89	29.08
Lump Error	11.20	13.11	7.99	5.17

Table 7: Source of misclassified utterances as a percentage of misclassified utterances and all utterances.

hand, our loc_3 measure jumps much closer to the human annotators. In other words, the maximum annotation overlap of our model and any given human is poor[9] while the local coherence of our annotation with respect to any human annotation is high. This pattern is symptomatic of over-splitting, which is excessively penalized by the 1-to-1 metric.[10]

We also analyzed the types of errors our model made while holding the conversation history constant. We simulated a consistent conversation history by treating the gold standard's choice as an unbeatable vote and tabulating the number of times our model voted with and against the winning conversation. There were five numbers tabulated: matching new conversation votes, matching old conversation votes, mismatching old conversation votes, incorrect split vote, and incorrect lump vote. The mismatching old conversation votes occurred when our model voted for an old conversation but guessed the wrong conversation. The incorrect split vote occurred when our model wanted to create a new conversation but the gold standard voted with an old conversation. Finally, the incorrect lump vote occurred when our model matched the utterance with a old conversation when the gold standard created a new conversation.

Across all six gold standard annotations, nearly two-thirds of the errors arose from incorrect splitting votes (see Table 7). In fact, nearly one-third of all utterances fell into this category.

6 Future Work

The high granularity for what our model considers a conversation had a huge impact on our performance scores. The high many-to-1 scores imply that more human-like chunks will improve performance. The granularity may be very task dependent and so we will need to be careful not to overfit our model to this data set and these annotators. New features should be tested with several chat corpora to better understand the cue trading effects of genre.

At present, our model uses only a minimal set of features. Discourse cues and temporal cues are two simple measures that can be added. Our current features can also use refinement. For instance, even partially disambiguating the particular sense of the lemmas should reduce the noise in our similarity measures. Ranking the semantic similarity, in contrast with the current winner-takes-all approach, should improve our results. Accounting for forward referencing, when a speaker invokes another's name to draw them into a conversation, is also important.

Finally, understanding the different voting patterns of each feature system will help us to better understand the reliability of the different cues. Towards this end, we need to monitor and act upon the strength and type of disagreement among voters.

[9] Elsner and Charniak (2008) found their annotators also tended to disagree on the exact point when a new conversation begins.

[10] Aoki et al. (2006) present a thorough analysis of conversational features associated with schisming, the splitting off of new conversations from old conversations.

Acknowledgments

Harish Srinivasan was great help in tokenizing the data. The NLP Group at Janya, Inc., Jordana Heller, Jean-Pierre Koenig, Michael Prentice, and three anonymous reviewers provided useful feedback.

References

Evrim Acar, Seyit Ahmet Camtepe, Mukkai S. Krishnamoorthy, and Blent Yener. 2005. Modeling and multiway analysis of chatroom tensors. In Paul B. Kantor, Gheorghe Muresan, Fred Roberts, Daniel Dajun Zeng, Fei-Yue Wang, Hsinchun Chen, and Ralph C. Merkle, editors, *ISI*, volume 3495 of *Lecture Notes in Computer Science*, pages 256–268. Springer.

Paul M. Aoki, Matthew Romaine, Margaret H. Szymanski, James D. Thornton, Daniel Wilson, and Allison Woodruff. 2003. The Mad Hatter's cocktail party: A social mobile audio space supporting multiple simultaneous conversations. In *CHI 03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 425–432, New York, NY, USA. ACM Press.

Paul M. Aoki, Margaret H. Szymanski, Luke Plurkowski, James D. Thornton, Allison Woodruff, and Weilie Yi.

	Other Annotators	E&C Model	Our Model	E&C Best Baseline
Mean 1-to-1	52.98	40.62	35.77	34.73
Max 1-to-1	63.50	51.12	49.88	56.00
Min 1-to-1	35.63	33.63	28.25	28.62
Mean loc_3	81.09	72.75	68.73	62.16
Max loc_3	86.53	75.16	72.77	69.05
Min loc_3	74.75	70.47	64.45	54.37

Table 6: Metric values for our model as compared with Elsner and Charniak's human annotators and classifier. Some numbers are taken from Table 3 (Elsner and Charniak, 2008).

2006. Where's the "party" in "multi-party"? Analyzing the structure of small-group sociable talk. In *ACM Conference on Computer Supported Cooperative Work*, pages 393–402, Banff, Alberta, Canada, November. ACM Press.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. In *Machine Learning*, pages 177–210.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Seyit Ahmet Camtepe, Mark K. Goldberg, Malik Magdon-Ismail, and Mukkai Krishnamoorty. 2005. Detecting conversing groups of chatters: A model, algorithms, and tests. In *IADIS AC*, pages 89–96.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41:391–407.

Micha Elsner and Eugene Charniak. 2008. You talking to me? A corpus and algorithm for conversation disentanglement. In *The Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008)*, Columbus, Ohio.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the ACL*, pages 562–569.

Michael Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman Group, New York.

Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

J. R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

A. Kehler. 2000. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

Hideki Kozima and Teiji Furugori. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL–93)*, pages 232–239, Utrecht.

Hideki Kozima. 1993. Text segmentation based on similarity between words. In *Proceedings of ACL'93*, pages 286–288, Ohio.

C. D. Manning. 1998. Rethinking text segmentation models: An information extraction case study. Technical Report SULTRY–98–07–01, University of Sydney.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Jacki O'Neill and David Martin. 2003. Text chat in action. In *GROUP '03: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, pages 40–49, New York, NY, USA. ACM Press.

R. Power, D. Scott, and N. Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

Jeffrey C. Reynar. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 357–364, Maryland, USA, June.

Peter Mark Roget, editor. 1911. *Roget's Thesaurus*. Project Gutenberg.

R. K. Srihari, W. Li, C. Niu, and T. Cornell. 2008. Infoxtract: A customizable intermediate level information extraction engine. *Journal of Natural Language Engineering*, 14(1):33–69.

Michael Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKMA–93)*, pages 67–74, Arlington, VA.

Grady Ward, editor. 2002. *Moby Thesaurus List*. Project Gutenberg.