Making Conversational Structure Explicit: Identification of Initiation-response Pairs within Online Discussions

Yi-Chia Wang
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
yichiaw@cs.cmu.edu

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

Abstract

In this paper we investigate how to identify initiation-response pairs in asynchronous, multi-threaded, multi-party conversations. We formulate the task of identifying initiation-response pairs as a pairwise ranking problem. A novel variant of Latent Semantic Analysis (LSA) is proposed to overcome a limitation of standard LSA models, namely that uncommon words, which are critical for signaling initiation-response links, tend to be deemphasized as it is the more frequent terms that end up closer to the latent factors selected through singular value decomposition. We present experimental results demonstrating significantly better performance of the novel variant of LSA over standard LSA.

1 Introduction

In recent years, research in the analysis of social media (e.g., weblogs, discussion boards, and messengers) has grown in popularity. Unlike expository text, the data produced through use of social media is often conversational, multi-threaded, and more complex because of the involvement of numerous participants who are distributed both across time and across space. Recovering the multi-threaded structure is an active area of research.

In this paper, we form the foundation for a broader study of this type of data by investigating the basic unit of interaction, referred to as an **initiation-response pair** (Schegloff, 2007). Initiation-response pairs are pairs of utterances that are typically contributed by different participants, and where the first pair part sets up an expectation for the second pair part. Types of common initiation-response pairs include question-answer, assessment-agreement, blame-denial, etc. Note that although sometimes discussion forum interfaces make the thread structure of the interaction explicit, these affordances are not always present. And even in forums that have these affordances, the apparent structure of the discourse as represented through the interface may not capture all of the contingencies between contributions in the unfolding conversation. Thus, the goal of this investigation is to investigate approaches for automatically identifying initiation-response pairs in conversations.

One of the challenges in identifying initiation-response pairs is that the related messages are not necessarily adjacent to each other in the stream of contributed messages, especially within the asynchronous environment of social media. Furthermore, individual differences related to writing style or creative expression of self may also complicate the identification of the intended connections between contributions. Identification of initiation-response pairs is an important step towards automatic processing of conversational data. One potential application of this work is conversation summarization. A summary should include both the initiation and response as a coherent unit or it may fail to capture the intended meaning.

We formulate the task of identifying initiation-response pairs as a pairwise ranking problem. The goal is to distinguish message pairs that constitute an initiation-response pair from those that do not. We believe a ranking approach, where the degree of relatedness between a message pair can be considered in light of the relatedness between each of them and the surrounding messages within the same thread, is a more suitable paradigm for this task than a discrete classification-based paradigm.

Previous work on recovering conversational structure has relied on simple lexical cohesion

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 673–676,
Los Angeles, California, June 2010.

measures (i.e., cosine similarity), temporal information (Lewis and Knowles, 1997; Wang et al., 2008), and meta-data (Minkov et al., 2006). However, relatively little work has investigated the importance of specifically **in-focus connections** between initiation-response pairs and utilized them as clues for the task. Consider, for example, the following excerpt discussing whether congress should pass a bill requiring the use of smaller cars to save the environment:

a) *Regressing to smaller **vehicles** would discourage business from producing more **pollution**.*
b) *If **CO2** emissions are lowered, wouldn't tax revenues be lowered as well? Are the democrats going to willingly give up Medicaid and social security?*

Although segment (b) is a reply to segment (a), the amount of word overlap is minimal. Nonetheless, we can determine that (b) is a response to (a) by recognizing the in-focus connections, such as "vehicles-CO2" and "pollution-CO2." To properly account for connections between initiations and responses, we introduce a novel variant of Latent Semantic Analysis (LSA) into our ranking model.

In section 2, we describe the Usenet data and how we extract a large corpus of initiation-response pairs from it. Section 3 explains our ranking model as well as the proposed novel LSA variation. The experimental results and discussion are detailed in Section 4 and Section 5, respectively.

2 Usenet and Generation of Data

The experiment for this paper was conducted using data crawled from the *alt.politics.usa Usenet* (User Network) discussion forum, including all posts from the period between June 2003 and June 2008. The resulting set contains 784,708 posts. The posts in this dataset also contain meta-data that makes parent-child relationships explicit (i.e., through the *References* field). Thus, we know 625,116 of the posts are explicit responses to others posts. The messages are organized into a total of 77,985 discussion threads, each of which has 2 or more posts.

In order to evaluate the quality of using the explicit reply structure as our gold standard for initiation-response links, we asked human judges to annotate the response structure of a random-selected medium-length discussion (19 posts) where we had removed the meta-data that indicated the initiation-reply structure. The result shows the accuracy of our gold standard is 0.89.

To set up the data as a pairwise ranking problem, we arranged the posts in the corpus into instances containing three messages each, one of which is a response message, one of which is the actual initiating message, and the other of which is a foil selected from the same thread. The idea is that the ranking model will be trained to prefer the actual initiating message in contrast to the foil.

The grain size of our examples is finer than whole messages. More specifically, positive examples are pairs of spans of text that have an initiation-reply relationship. We began the process with pairs of messages where the meta-data indicates that an initiation-reply relationship exits, but we didn't stop there. For our task it is important to narrow down to the specific spans of text that have the initiation-response relation. For this, we used the indication of quoted material within a message. We observed that when users explicitly quote a portion of a previously posted message, the portion of text immediately following the quoted material tends to have an explicit discourse connection with it. Consider the following example:

> *>> Why is the quality of life of the child, mother,*
> *>> and society at large, more important than the*
> *>> sanctity of life?*
> *> Because in the case of anencephaly at least,*
> *> the life is ended before it begins.*
> *We disagree on this point. Why do you refuse to provide your very own positive definition of life? Do you believe life begins before birth? At birth? After birth? Never?*

In this thread, the reply expresses an opinion against the first level quote, but not the second level quote. Thus, we used segments of text with single quotes as an initiation and the immediately following non-quoted text as the response. We extracted positive examples by scanning each post to locate the first level quote that is immediately followed by unquoted content. If such quoted material was found, the quoted material and the unquoted response were both extracted to form a positive example. Otherwise, the message was discarded.

For each post *P* where we extracted a positive example, we also extracted a negative example by picking a random post *R* from the same thread as *P*. We selected the negative example in such a way to make the task difficult in a realistic way. Choosing *R* from other threads would make the task too easy because the topics of *P* and *R* would most likely be different. We also stipulated that *R* cannot be the parent, grandparent, sibling, or child of *P*.

Together the non-quoted text of P and R forms a negative instance. Thus, the final dataset consists of pairs of message pairs ((p_i, p_j), (p_i, p_k)), where they have the same reply message p_i, and p_j is the correct quote message of p_i, but p_k is not. In other words, (p_i, p_j) is considered as a positive example; (p_i, p_k) is a negative example. We constructed a total of 100,028 instances for our dataset, 10,000 (~10%) of which were used for testing, and 90,028 (~90%) of which were the learning set used to construct the LSA space described in the next section.

3 Ranking Models for Identification of Initiation-Response Pairs

Our pairwise ranking model[1] takes as input an ordered pair of message pairs ((p_i, p_j), (p_i, p_k)) and computes their relatedness using a similarity function *sim*. Specifically,

$$(x_{ij}, x_{ik}) = (sim (p_i, p_j), sim (p_i, p_k))$$

where x_{ij} is the similarity value between post p_i and p_j; x_{ik} is the similarity value between post p_i and p_k. To determine which of the two message pairs ranks higher regarding initiation-response relatedness, we use the following scoring function to compare their corresponding similarity values:

$$score (x_{ij}, x_{ik}) = x_{ij} - x_{ik}$$

If the score is positive, the model ranks (p_i, p_j) higher than (p_i, p_k) and vice versa. A message pair ranked higher means it has more evidence of being an initiation-reply link, compared to the other pair.

3.1 Alternative Similarity Functions

We introduce and motivate 3 alternative similarity functions, where the first two are considered as baseline approaches and the third one is a novel variation of LSA. We argue that the proposed LSA variation is an appropriate semantic similarity measurement for identifying topic continuation and initiation-reply pairs in online discussions.

Cosine Similarity (*cossim*). We choose an approach that uses only lexical cohesion as our baseline. Previous work (Lewis and Knowles, 1997; Wang et al., 2008) has verified its usefulness for the thread identification task. In this case,

$$sim(p_i,p_j) = cossim(p_i,p_j)$$

where $cossim(p_i,p_j)$ computes the cosine of the angle between two posts p_i and p_j while they are represented as term vectors.

LSA Average Similarity (*lsaavg*). LSA is a well-known method for grouping semantically related words (Landauer et al., 1998). It represents word meanings in a concept space with dimensionality k. Before we describe how to compute average similarity given an LSA space, we explain how the LSA space was constructed in our work. First, we construct a term-by-document matrix, where we use the 90,028 message learning set mentioned at the end of Section 2. Next, LSA applies singular value decomposition to the matrix, and reduces the dimensionality of the feature space to a k dimensional concept space. This generated LSA space is used by both *lsaavg* and *lsacart* later.

For *lsaavg*, we follow Foltz et al. (1998):

$$sim(p_i, p_j) = lsaavg(p_i, p_j) = \cos\left(\frac{\sum_{t_a \in p_i} \vec{t_a}}{|p_i|}, \frac{\sum_{t_b \in p_j} \vec{t_b}}{|p_j|} \right)$$

The meaning of each post is represented as a vector in the LSA space by averaging across the LSA representations for each of its words. The similarity between the two posts is then determined by computing the cosine value of their LSA vectors.

This is the typical method for using LSA in text similarity comparisons. However, note that not all words carry equal weight within the vector that results from this averaging process. Words that are closer to the "semantic prototypes" represented by each of the k dimensions of the reduced vector space will have vectors with longer lengths than words that are less prototypical. Thus, those words that are closer to those prototypes will have a larger effect on the direction of the resulting vector and therefore on the comparison with other texts. An important consideration is whether this is a desirable effect. It would lead to deemphasizing those unusual types of information that might be being discussed as part of a post. However, one might expect that those things that are unusual types of information might actually be more likely to be the in-focus information within an initiation that responses may be likely to refer to. In that case, for our purposes, we would not expect this typical method for applying LSA to work well.

LSA Cartesian Similarity (*lsacart*). To properly account for connections between initiations and

[1] We cast the problem as a pairwise ranking problem in order to focus specifically on the issue of characterizing how initiation-response links are encoded in language through lexical choice. Note that once trained, pairwise ranking models can be used to rank multiple instances.

responses that include unusual words, we introduce the following similarity function:

$$sim(p_i, p_j) = lsacart(p_i, p_j) = \frac{\sum_{(t_a, t_b) \in p_i \times p_j} \cos(\vec{t_a}, \vec{t_b})}{|p_i||p_j|}$$

where we take the mean of the cosine values for all the word pairs in the Cartesian product of posts p_i and p_j. Note that in this formulation, all words have an equal chance to affect the overall similarity between vectors since it is the angle represented by each word in a pair that comes to play when cosine distance is applied to a word pair. Length is no longer a factor. Moreover, the averaging is across cosine similarity scores rather than LSA vectors.

4 Experimental Results

The results are found in Table 1. For comparison, we also report the random baseline (0.50).

	Random Baseline	Cos-Similarity	LSA-Average	LSA-Cart
Accuracy	0.50	0.66	0.60	**0.71**

Table 1. Overview of results

Besides the random baseline, LSA-Average performs the worst (0.60), with simple Cosine similarity (0.66) in the middle, and LSA-Cart (0.71) the best, with each of the pairwise contrasts being statistically significant. We believe the reason why LSA-Average performs so poorly on this task is precisely because, as discussed in last section, it deemphasizes those words that contribute the most unusual content. LSA-Cart addresses this issue.

To further understand this effect, we conducted an error analysis. We divided the instances into 4 sets based on the lexical cohesion between the response and the true initiation and between the response and the foil, by taking the median split on the distributions of these two cohesion scores. Our finding is that model performances vary by subset. In particular, we find that it is only in cases where the positive example has low lexical cohesion (e.g. our "vehicles-CO2" and "pollution-CO2" example from the earlier section), that we see the benefit of the LSA-Cart approach. In other cases, where the cohesion between the reply and the true initiation is high, Cos-Similarity performs best.

5 Discussion and Conclusion

We have argued why the task of detecting initiation-response pairs in multi-party discussions is important and challenging. We proposed a method for acquiring a large corpus for use to identify initiation-response pairs. In our experiments, we have shown that the ranking model using a variant of LSA performs best, which affirms our hypothesis that unusual information and uncommon words tends to be the focus of ongoing discussions and therefore to be the key in identifying initiation-response links.

In future work, we plan to further investigate the connection between an initiation-response pairs from multiple dimensions, such as topical coherence, semantic relatedness, conversation acts, etc. One important current direction is to develop a richer operationalization of the interaction that accounts for the way posts sometimes respond to a user, a collection of users, or a user's posting history, rather than specific posts per se.

Acknowledgments

We thank Mary McGlohon for sharing her data with us. This research was funded through NSF grant DRL-0835426.

References

David D. Lewis and Kimberly A. Knowles. 1997. Threading electronic mail: A preliminary study. *Information Processing and Management*, 33(2), 209–217.

Einat Minkov, William W. Cohen, Andrew Y. Ng. 2006. Contextual Search and Name Disambiguation in Email using Graphs. In *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 35–42. ACM Press, 2006.

Peter W. Foltz, Walter Kintsch, Thomas K. Landauer. 1998. Textual coherence using latent semantic analysis. *Discourse Processes*, 25, 285–307.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Schegloff, E. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*, Cambridge University Press.

Yi-Chia Wang, Mahesh Joshi, William W. Cohen, Carolyn P. Rosé. 2008. Recovering Implicit Thread Structure in Newsgroup Style Conversations. In *Proceedings of the 2nd International Conference on Weblogs and Social Media (ICWSM II)*, Seattle, USA.