A Feature Based Approach to Leveraging Context for Classifying Newsgroup Style Discussion Segments

Yi-Chia Wang, Mahesh Joshi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{yichiaw,maheshj}@cs.cmu.edu

Carolyn Penstein Rosé
Language Technologies Institute/
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
cprose@cs.cmu.edu

Abstract

On a multi-dimensional text categorization task, we compare the effectiveness of a feature based approach with the use of a state-of-the-art sequential learning technique that has proven successful for tasks such as "email act classification". Our evaluation demonstrates for the three separate dimensions of a well established annotation scheme that novel thread based features have a greater and more consistent impact on classification performance.

1 Introduction

The problem of information overload in personal communication media such as email, instant messaging, and on-line discussion boards is a well documented phenomenon (Bellotti, 2005). Because of this, conversation summarization is an area with a great potential impact (Zechner, 2001). What is strikingly different about this form of summarization from summarization of expository text is that the summary may include more than just the content, such as the style and structure of the conversation (Roman et al., 2006). In this paper we focus on a classification task that will eventually be used to enable this form of conversation summarization by providing indicators of the quality of group functioning and argumentation.

Lacson and colleagues (2006) describe a form of conversation summarization where a classification approach is first applied to segments of a conversation in order to identify regions of the conversation related to different types of information. This aids in structuring a useful summary. In this paper, we describe work in progress towards a different form of conversation summarization that similarly leverages a text classification approach. We focus on newsgroup style interactions. The goal of assessing the quality of interactions in that context is to enable the quality and nature of discussions that occur within an on-line discussion board to be communicated in a summary to a potential newcomer or group moderators.

We propose to adopt an approach developed in the computer supported collaborative learning (CSCL) community for measuring the quality of interactions in a threaded, online discussion forum using a multi-dimensional annotation scheme (Weinberger & Fischer, 2006). Using this annotation scheme, messages are segmented into idea units and then coded with several independent dimensions, three of which are relevant for our work, namely micro-argumentation, macro-argumentation, and social modes of co-construction, which categorizes spans of text as belonging to one of five consensus building categories. By coding segments with this annotation scheme, it is possible to measure the extent to which group members' arguments are well formed or the extent to which they are engaging in functional or dysfunctional consensus building behavior.

This work can be seen as analogous to work on "email act classification" (Carvalho & Cohen, 2005). However, while in some ways the structure of newsgroup style interaction is more straightforward than email based interaction because of the unambiguous thread structure (Carvalho & Cohen, 2005), what makes this particularly challenging

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 73–76,
Prague, June 2007.

from a technical standpoint is that the structure of this type of conversation is multi-leveled, as we describe in greater depth below.

We investigate the use of state-of-the-art sequential learning techniques that have proven successful for email act classification in comparison with a feature based approach. Our evaluation demonstrates for the three separate dimensions of a context oriented annotation scheme that novel thread based features have a greater and more consistent impact on classification performance.

2 Data and Coding

We make use of an available annotated corpus of discussion data where groups of three students discuss case studies in an on-line, newsgroup style discussion environment (Weinberger & Fischer, 2006). This corpus is structurally more complex than the data sets used previously to demonstrate the advantages of using sequential learning techniques for identifying email acts (Carvalho & Cohen, 2005). In the email act corpus, each message as a whole is assigned one or more codes. Thus, the history of a span of text is defined in terms of the thread structure of an email conversation. However, in the Weinberger and Fischer corpus, each message is segmented into idea units. Thus, a span of text has a context within a message, defined by the sequence of text spans within that message, as well as a context from the larger thread structure.

The Weinberger and Fischer annotation scheme has seven dimensions, three of which are relevant for our work.

1. *Micro-level of argumentation* [4 categories] How an individual argument consists of a claim which can be supported by a ground with warrant and/or specified by a qualifier
2. *Macro-level of argumentation* [6 categories] Argumentation sequences are examined in terms of how learners connect individual arguments to create a more complex argument (for example, consisting of an argument, a counter-argument, and integration)
3. *Social Modes of Co-Construction* [6 categories] To what degree or in what ways learners refer to the contributions of their learning partners, including externalizations, elicitations, quick consensus building, integration oriented consensus building, or conflict oriented consensus building, or other.

For the two argumentation dimensions, the most natural application of sequential learning techniques is by defining the history of a span of text in terms of the sequence of spans of text within a message, since although arguments may build on previous messages, there is also a structure to the argument within a single message. For the Social Modes of Co-construction dimension, it is less clear. However, we have experimented with both ways of defining the history and have not observed any benefit of sequential learning techniques by defining the history for sequential learning in terms of previous messages. Thus, for all three dimensions, we report results for histories defined within a single message in our evaluation below.

3 Feature Based Approach

In previous text classification research, more attention to the selection of predictive features has been done for text classification problems where very subtle distinctions must be made or where the size of spans of text being classified is relatively small. Both of these are true of our work. For the base features, we began with typical text features extracted from the raw text, including unstemmed unigrams and punctuation. We did not remove stop words, although we did remove features that occured less than 5 times in the corpus. We also included a feature that indicated the number of words in the segment.

Thread Structure Features. The simplest context-oriented feature we can add based on the threaded structure is a number indicating the depth in the thread where a message appears. We refer to this feature as *deep*. This is expected to improve performance to the extent that thread initial messages may be rhetorically distinct from messages that occur further down in the thread. The other context oriented feature related to the thread structure is derived from relationships between spans of text appearing in the parent and child messages. This feature is meant to indicate how semantically related a span of text is to the spans of text in the parent message. This is computed using the minimum of all cosine distance measures between the vector representation of the span of text and that of each of the spans of text in all parent messages, which is a typical shallow measure of semantic similarity. The smallest such distance measure is included as a feature indicating how related the current span of text is to a parent message.

Sequence-Oriented Features. We hypothesized that the sequence of codes within a message follows a semi-regular structure. In particular, the discussion environment used to collect the Weinberger and Fischer corpus inserts prompts into the message buffers before messages are composed in order to structure the interaction. Users fill in text underneath these prompts. Sometimes they quote material from a previous message before inserting their own comments. We hypothesized that whether or not a piece of quoted material appears before a span of text might influence which code is appropriate. Thus, we constructed the *fsm* feature, which indicates the state of a simple finite-state automaton that only has two states. The automaton is set to initial state (q_0) at the top of a message. It makes a transition to state (q_1) when it encounters a quoted span of text. Once in state (q_1), the automaton remains in this state until it encounters a prompt. On encountering a prompt it makes a transition back to the initial state (q_0). The purpose is to indicate places where users are likely to make a comment in reference to something another participant in the conversation has already contributed.

4 Evaluation

The purpose of our evaluation is to contrast our proposed feature based approach with a state-of-the-art sequential learning technique (Collins, 2002). Both approaches are designed to leverage context for the purpose of increasing classification accuracy on a classification task where the codes refer to the role a span of text plays in context.

We evaluate these two approaches alone and in combination over the same data but with three different sets of codes, namely the three relevant dimensions of the Weinberger and Fischer annotation scheme. In all cases, we employ a 10-fold cross-validation methodology, where we apply a feature selection wrapper in such as way as to select the 100 best features over the training set on each fold, and then to apply this feature space and the trained model to the test set. The complete corpus comprises about 250 discussions of the participants. From this we have run our experiments with a subset of this data, using altogether 1250 annotated text segments. Trained coders categorized each segment using this multi-dimensional annotation scheme, in each case achieving a level of agreement exceeding .7 Kappa both for segmentation and coding of all dimensions as previously published (Weinberger & Fischer, 2006).

For each dimension, we first evaluate alternative combinations of features using SMO, Weka's implementation of Support Vector Machines (Witten & Frank, 2005). For a sequential learning algorithm, we make use of the Collins Perceptron Learner (Collins, 2002). When using the Collins Perceptron Learner, in all cases we evaluate combinations of alternative history sizes (0 and 1) and alternative feature sets (base and base+AllContext). In our experimentation we have evaluated larger history sizes as well, but the performance was consistently worse as the history size grew larger than 1. Thus, we only report results for history sizes of 0 and 1.

Our evaluation demonstrates that we achieve a much greater impact on performance with carefully designed, automatically extractable context oriented features. In all cases we are able to achieve a statistically significant improvement by adding context oriented features, and only achieve a statistically significant improvement using sequential learning for one dimension, and only in the absence of context oriented features.

4.1 Feature Based Approach

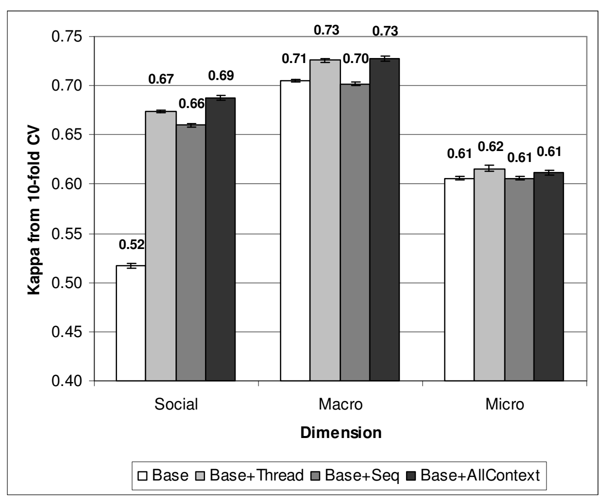

Figure 1. Results with alternative features sets

We first evaluated the feature based approach across all three dimensions and demonstrate that statistically significant improvements are achieved on all dimensions by adding context oriented features. The most dramatic results are achieved on the Social Modes of Co-Construction dimension (See Figure 1). All pairwise contrasts between alternative feature sets within this dimension are statistically significant. In the other dimensions, while Base+Thread is a significant improvement over Base, there is no significant difference between Base+Thread and Base+AllContext.

4.2 Sequential Learning

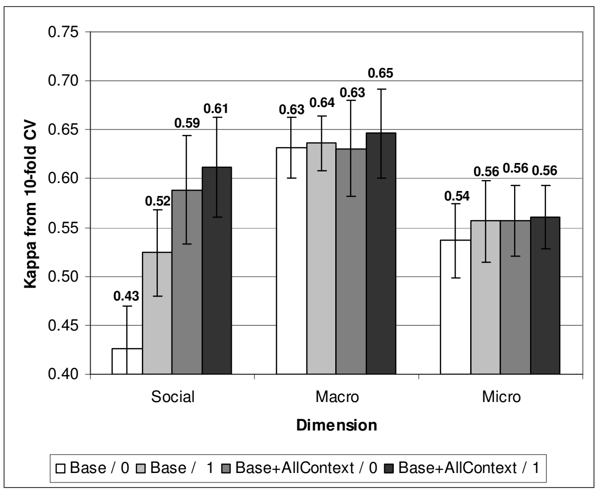

Figure 2. Results with Sequential Learning

The results for sequential learning are weaker than for the feature based (See Figure 2). While the Collins Perceptron learner possesses the capability of modeling sequential dependencies between codes, which SMO does not possess, it is not necessarily a more powerful learner. On this data set, the Collins Perceptron learner consistently performs worse that SMO. Even restricting our evaluation of sequential learning to a comparison between the Collins Perceptron learner with a history of 0 (i.e., no history) with the same learner using a history of 1, we only see a statistically significant improvement on the Social Modes of Co-Construction dimension. This is when only using base features, although the trend was consistently in favor of a history of 1 over 0. Note that the standard deviation in the performance across folds was much higher with the Collins Perceptron learner, so that a much greater difference in average would be required in order to achieve statistical significance. Performance over a validation set was always worse with larger history sizes than 1.

5 Conclusions

We have described work towards an approach to conversation summarization where an assessment of conversational quality along multiple process dimensions is reported. We make use of a well-established annotation scheme developed in the CSCL community. Our evaluation demonstrates that thread based features have a greater and more consistent impact on performance with this data.

This work was supported by the National Science Foundation grant number SBE0354420, and Office of Naval Research, Cognitive and Neural Sciences Division Grant N00014-05-1-0043.

References

Bellotti, V., Ducheneaut, N., Howard, M. Smith, I., Grinter, R. (2005). Quality versus Quantity: Email-centric task management and its relation with overload. Human-Computer Interaction, 2005, vol. 20

Carvalho, V. & Cohen, W. (2005). On the Collective Classification of Email “Speech Acts”, Proceedings of SIGIR ‘2005.

Collins, M (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of EMNLP 2002*.

Lacson, R., Barzilay, R., & Long, W. (2006). Automatic analysis of medical dialogue in the homehemodialysis domain: structure induction and summarization, *Journal of Biomedical Informatics* 39(5), pp541-555.

Roman, N., Piwek, P., & Carvalho, A. (2006). Politeness and Bias in Dialogue Summarization : Two Exploratory Studies, in J. Shanahan, Y. Qu, & J. Wiebe (Eds.) *Computing Attitude and Affect in Text: Theory and Applications, the Information Retrieval Series.*

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46, 71-95.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, second edition, Elsevier: San Francisco.

Zechner, K. (2001). Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains. *Proceedings of ACM SIG-IR 2001.*