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

Yi-Chia Wang, Mahesh Joshi, Carolyn Rosé

Language Technologies Institute, School of Computer Science, Carnegie Mellon University

Motivation

Information overload in communication media (email, instant messaging, discussion boards)

Solution

Conversation summarization

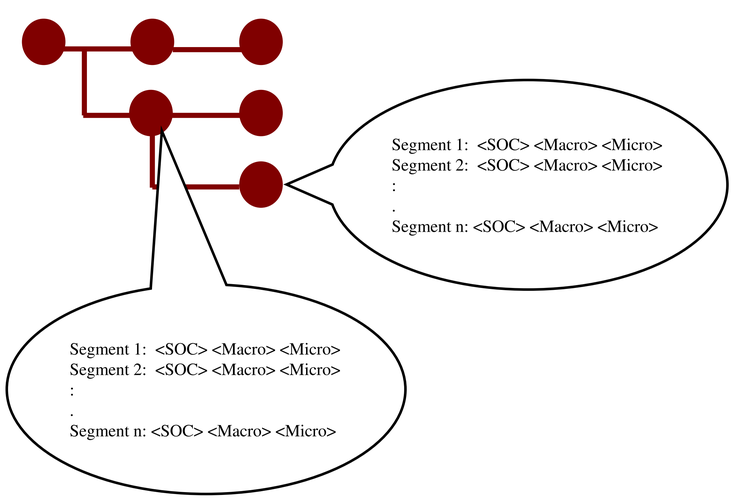
Content + Style + Structure

First Step

Classifying segments of conversation

Goal

- Assessing the quality of newsgroup style interactions using a **multi-dimensional annotation** scheme
- Utilizing the **nature** of threaded discussion forums
  - Novel thread based features
  - Sequential data



Data and Annotation Scheme

Messages are segmented into **idea units** and coded with 3 context-oriented dimensions

- **Micro-level** of argumentation (4 classes): Assessing the quality of individual arguments
- **Macro-level** of argumentation (6 classes): Examining the connection between individual arguments
- **Social Modes** of Co-Construction (6 classes): Referring to the degree of contributions of learners.

Feature Based Approach

Baseline

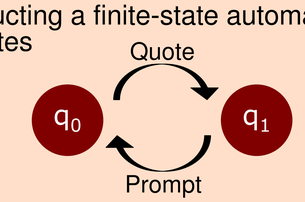
- Un-stemmed unigrams
- Punctuation
- Removing features that occurred less that 5
- Length of each segment

Thread Structure Features

- Deep: the depth in the thread where a message appears
- Parent and Child: **semantic** relation of the current message to the parent message

Sequence-Oriented Features

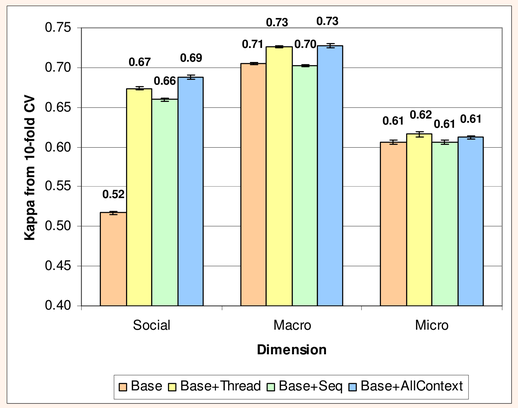
Constructing a finite-state automaton having two states



Evaluation Results

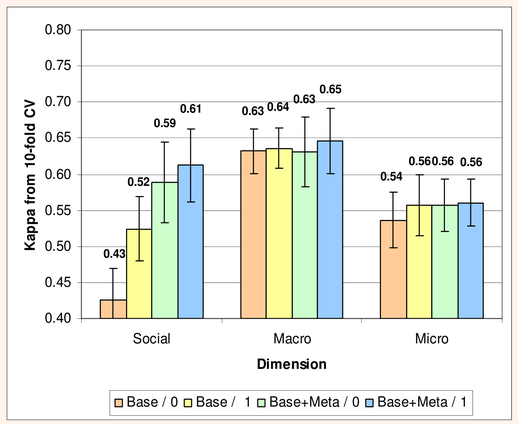
Alternative Feature Sets (SMO)

Achieving a statistically significant improvement by adding context oriented features in all 3 dimensions



Sequential Learning (Collins Perceptron)

Both context features and sequential learning provide some benefit within this task



Note: **Cohen's Kappa Statistic** is a commonly used metric in assessing inter-rater reliability on the defined analyses units among target categories.

Reference

- 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, p.89-138
- 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. Dordrecht: Springer.
- 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, ISBN 0-12-088407-0
- Zechner, K. (2001). Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains. Proceedings of ACM SIG-IR 2001.

2007 LTI Open House Student Poster Session, Mar. 23