2007 Student Research Symposium

Language Technologies Institute

Using thread structure to improve context-based classification of newsgroup discussion acts

Yi-Chia Wang

Abstract:

This research is a part of the TagHelper project, which seeks to develop conversation mining technology.

The information produced by social media such as email, instant messaging, and on-line discussion forums is growing rapidly. One common characteristic across the variety of communication media is that messages or contributions are threaded together either explicitly or implicitly in order to separate conversation related to different subtopics within the discussion stream. If we can effectively leverage the information encoded in the thread structure of social media, it would be beneficial for various research applications, including text mining, educational data analysis, and sentiment recognition.

We explored different ways to extract useful features from messages by leveraging the explicit thread structure that is built into newsgroup style conversation data in order to investigate the predictive value of those features in connection with different context-oriented annotation schemes. The goal was to gain insight into how context differently influences the interpretation of conversational moves at different levels of abstraction such as argumentation structure versus consensus building style.

We developed a novel approach to leveraging context for classifying newsgroup style discussion segments. One challenge is the multi-leveled structure of context in this data where messages were composed of multiple segments, and messages were arranged into a tree-shaped thread structure.

We created context oriented features that were meant to be predictive of the role a segment of text may have played within an ongoing discussion. Our specific solution was to use the thread structure to identify pairs of potentially related segments of text and then use shallow semantic similarity metrics to compute a feature that indicated the maximum similarity value between a segment contributed by one participant and other potentially related segments contributed by other participants in the discussion. This feature was meant to indicate the extent to which the participant who contributed the segment was engaged in building upon ideas contributed by other participants in the conversation. Other context features represented how far down on a thread the message containing the segment occurred

and whether the segment occurred soon after some quoted material in its associated message. Thread depth was meant to distinguish conversational moves that are used to initiate a discussion topic with those that are meant to continue ongoing discussions. Positioning within a message with respect to quoted material is also an indicator of the extent to which a segment of text relates to what has been contributed earlier.

We added the context oriented features to the base feature space constructed using a bag-of-terms approach.

We evaluated the contribution of our designed features in comparison with baseline features only, both with a Support Vector Machine learning algorithm (SVM) and using a sequential learning approach (The Collins Perceptron Learner). Both the feature based approach (i.e., augmenting the baseline feature space with context oriented features) and the sequential learning approach are designed to leverage context for the purpose of increasing classification accuracy in connection with classification tasks where context matters. To evaluate the effectiveness of each of these approaches, We used a text categorization task where individual segments within messages were annotated by our German partners with three separate annotation schemes designed to analyze argumentation on three distinct levels. Altogether, the data set contained 1750 annotated segments of German text.

The results show that augmenting the feature space achieves a statistically significant improvement on performance across all three annotation schemes using SVM as the learning algorithm. The most dramatic improvement was an increase from .5 Kappa to .69 Kappa in connection with an annotation scheme meant to identify consensus building style. Results with sequential learning were less dramatic. A statistically significant improvement was only obtained with one of the three annotation schemes, and only when using the baseline feature space, never with the augmented feature space. Furthermore, performance overall was consistently lower with the Collins perceptron learner than with SVM. Thus, the feature based approach produced a more consistent and more dramatic improvement than sequential learning across three separate annotation schemes applied to the same data.

Wang, Y. C., Joshi, M., Rosé, C. P. 2007. (Accepted). *A Feature Based Approach to Leveraging Context for Classifying Newsgroup Style Discussion Segments*. Poster in: the 43th Conference on Association for Computational Linguistics (ACL 2007), Prague.

Rosé, C. P., **Wang, Y. C.**, Cui,Y., Arguello, J.,Fischer, F., Weinberger, A., Stegmann, K. (Submitted). *Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning*. Submmited to the International Journal of Computer-Supported Collaborative Learning (ijCSCL).