Robust adaptive discourse parsing for e-learning fora

Nadine Lucas and Emmanuel Giguet
GREYC CNRS Dpt Informatique Caen University France
Nadine.Lucas@info.unicaen.fr, Emmanuel.Giguet@ info.unicaen.fr

Abstract

Discussion fora from e-learning students' platforms were parsed as collective discourse directed either by a specific task achievement goal or by a communication goal. Agora, a robust tool for French and English, provides "forum tiling" but also "discussion nesting". It uses the structure of messages along with stylistic statistical data to segment and organize the content of discourse. No external resources are needed. The output of this adaptive parser is a scalable view on collective discussion. Such views are used to browse and navigate in a very large forum, or to compare discussion progress between learners groups. Results are presented and their usefulness is discussed.

1. Introduction

Over the last decade interest has grown on the topic of new written forms of communication, particularly of discussion fora (also called web conferencing) for educational purposes. A wealth of literature deals with e-learning as related with electronic discussion groups. (See for a review [1]).

Current techniques however have the shortcoming of dissociating feedback on students' activity from contributions themselves, with few exceptions [2].

Parsing of educational fora has hardly been tempted to our knowledge. Discourse parsing is not yet developed, and newsgroup discussions are difficult to handle. Sentences are not easy to delimitate and words often misspelled. Parsers using annotations by researchers or by students themselves have been implemented [3]. A parser for un-annotated e-learning fora in French was implemented [4].

This paper further investigates robust automatic parsing of English and French discussion fora with variable granularity to handle scale. It results in a compacted expandable view of fora. We introduce the background of our research in section 2, and present the main ideas behind the adaptive parser and its implementation in section 3. We proceed to results and evaluation in section 4 and discuss the usefulness issue in section 5.

2. Background and objectives

The main issue for teachers and tutors involved in computer-based learning is to keep an eye on discussions going on without having to read all messages on all threads. Of particular importance is the "cold-start" problem.

The way a forum evolves is mainly monitored through statistical data unrelated to content. However, posting messages does not mean that an interesting discussion is going on. A widely accepted distinction holds between closed and open discussions. A closed forum is directed by a goal and has a limited span of life: typically a work group with an assignment and a deadline. An open forum relies on a community with shared (and shifting) interests and has no predefined limit in time.

A second distinction is related to the actual interface of the forum, either chronological or threaded. A third and last distinction is between fora in entirely distant learning/teaching and fora supporting partly or fully presence class. This distinction is relevant for discourse parsing.

The previous French parser ThemAgora enabled tutors to focus on the tuning time needed by different groups to actively engage in collective discourse, and for discussion to get momentum [4]. Critical comments by an educational research group (Calico project) led to further work. One requirement was to produce manageable output units, in order to facilitate monitoring and comparison. Parse on the fly and handling of English were also wished.

Contrary to the idea that a forum is made of messages with fixed characteristics, we consider that type and progression in a forum can be derived from its discourse dynamics. Related to collective discourse, this stand is backed by the enunciation theory and polyphony. This approach stresses the overall unit of the forum and enables comparisons between fora.

Agora provides a new view on the text and might be related to visualization and monitoring tools, or

used as pre-processing, prior to a theory-informed text parse.

3. Forum parsing

A forum is compared to a text, where messages are like paragraphs, arranged chronologically as a ribbon of units. The underlying hypothesis: if a collective discussion is going on, then it has discourse properties, rhythm, periods, leitmotivs and marks. These are computed by stylometrics and imply not only grouping but also nesting of messages in a hierarchy. The parser calculates the best resolution using stylistic contrastive features, (e.g. images, quotes) to manage scale.

3.1. Algorithm general principles

The goal is to detect contrasted segments in a forum. They can be coordinated (grouped) or subordinated (nested). They correspond to "moments" of discussion. The first moment matches with tuning, before elaborate discussion starts.

Discourse properties are handled as constraints, expressing relative saliency. Since the exact formulation of content is not predictable, relative contrast is computed. Evenly distributed features become constants that characterize background. Salient forms are different from background. However, they are marks if they acquire discriminative value from context, forming patterns, else they are forgotten.

Three inclusive levels are constructed by default, more if requested: forum level, discussion "moment" level and discussion "round" level. However, there are no fixed marks to define levels, because a forum has no fixed span.

Starting from *n* tokens, a population of messages in a forum, we want to create a legible structure where *s* tokens define the starting set of messages, followed by *m* moments, themselves including *r* rounds of discussion. *m* and *r* should be kept to a minimum.

Search windows proportionate to the length of the forum are defined. In (groups of) messages, higher order morphological features are, e.g., smilies or quotes [5]. Subdivision of a segment entails change of measure and of salient features.

To organize the text-file, the wrappers detection and induction principles are used [6]. The wrapper hypothesis is that valuable (though unpredictable) content can be found because wrapped by repeated substrings of text. Suppose a question message from a student is followed by many answers by students (with quotes) and by a concluding message from the tutor. The automat recognizes this as a group because both initial and last messages of that discussion round are distinctively shorter than the middle ones, and not containing any quotes: they wrap the discussion. Wrapper induction means that content is computationally wrapped although it is only marked by one border: the missing element of the pair is inferred from context.

Segments are coordinated while the borders /wrappers share the same distinctive features found in the same search window. Large segments are broken down using a finer set of features. Smaller segments are subordinated (nested inside the larger unit) when they inherit features from the parent segment.

3.2. Algorithm and implementation

Fora are formatted in a shared XML-forum format allowing export. Connectors were implemented to convert popular formats such as phpBB or specific e-learning platforms formats into XML-forum.

Input is the forum document, seen as a set of messages. Messages are ordered chronologically if need be. Messages are tokenized in paragraph blocks (delimitated by blank lines), paragraphs and sentences.

The first step is diagnosis of the forum to assess its size. The approximate measure for borders is set accordingly. In a small forum, a message can signal a transition between two distinct moments, while in a large forum a group of messages separates moments.

Next tests are made to detect morphological features: inclusion of images, links, quotes, code or whatever was labeled in the original files; at a finer grain character strings (smiley, uppercase strings, multiple punctuations). The most frequent forms of tokens are considered as background. Relative saliency is calculated against the background.

The second step is segmentation of the forum in two: tuning and discussion proper. The first division is made with the strongest contrast at the beginning of the forum (not exceeding 1/5 of its length).

Discussion is segmented in a reasonable number of very large chunks, delimitated by remarkable (groups of) messages. Contrast is checked on relative length of messages and salient message features (e. g. no quotes from other messages included).

The third step reiterates the segmentation and diagnosis process with finer marks, e.g. structure of messages or distribution of smilies or emphatic multiple punctuation marks. It may be described as subdividing the largest chunks obtained at level 2 (moments), into subparts (rounds). If no contrast is found, moment is not subdivided.

The fourth step is extraction of wrappers, shrinking of the segments content and visualization.

The program is in php and runs on-line.

3.3. Output

Results are shown in a compact view fitting in a screen and showing only the start and moments of discussion (Fig. 1). The start corresponds to tuning. The list of moments at level 2 is numbered from group 1 to *n*.

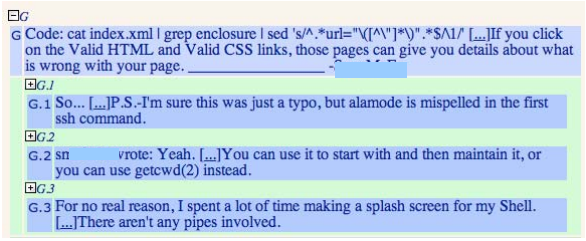

Figure 1. Compact view of a forum (OS Projects)

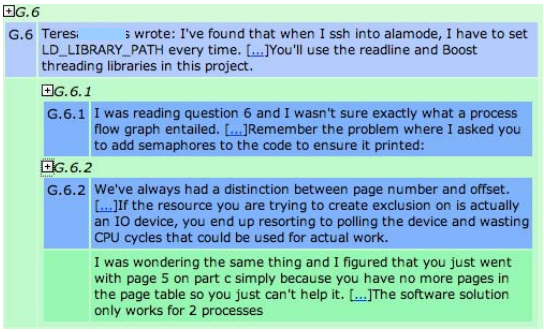

Figure 2. Zoom on a discussion moment, last round expanded (OS Concepts)

To expand the compact view, a cascading zoom allows opening nested segments at levels 2 and 3 (Fig. 2). A fisheye view shows the contents of a cell in full detail at any level (otherwise, only wrappers are shown).

4. Experimental results and discussion

French and English were the two languages dealt with. 11 French-language fora came from the corpus studied by the Calico educational research group. Settings are distance education, presence class and mixed. Formats of the original fora were tabulated and threaded. Three on-going English-language fora were downloaded from an educational platform on the web and parsed on the fly.

All fora were parsed, showing robustness of the tool. Teachers and tutors assessed results on their students' discussion fora and on unknown external ones. Users valued the high granularity segmentation. It gives a good idea on the forum at a glance. Results were deemed sensible. Usually in distance teaching, tuning is longer than in presence class. Closed fora show more nested rounds of discussions. Comparing fora parsed by the same tool fostered comments on the class behavior and tutoring style as perceived through discussion rounds and moments.

Critics bore on the snippets of text shown in the compact view. Extract of wrappers are not quite indicative of the content of the discussion. No ready solution was found to select representative extracts to be shown in a view still fitting in a screen.

The Agora parser is designed for avoiding tedious manual annotations on large data, so reliability cannot be assessed against manual annotation as gold standard. The added value is to signal when collective discourse gets momentum, which is a common concern with educators. If run at time *t,* the system returns a view of an ongoing forum, provided there are contrastive features. An interesting issue is to compare on-the-fly results with past data to try and predict the outcome of a given discussion forum.

The Agora parser needs no external resources. It achieves segmentation and nesting of discussion fora in French and English. It achieves adaptation by computing the best measure as related to the input data, to keep the output at three granularity levels.

References

[1] A. Dimitracopoulou, *State of the Art on Interaction Analysis for Metacognitive Support and Diagnosis,* Kaleidoscope Network of Excellence, Report JEIRP. D.31.1.1, 2005, pp. 6-62. Online www.noekaleidoscope.org.

[2] S. George, "Contextualizing discussions in distance learning systems", *4th IEEE International Conference on Advanced Learning Technologies (ICALT 2004)*, Joensuu, Finland, 2004, pp. 226-230.

[3] S. Corich, Kinshuk and L. M. Hunt, "Measuring Critical Thinking within Discussion Forums using a Computerised Content Analysis Tool", *5th International Conference on Networked Learning* 2006, Lancaster University, UK.

[4] M. Sidir, N. Lucas and E. Giguet, "De l'analyse des discours à l'analyse structurale des réseaux sociaux: une étude diachronique d'un forum éducatif", *Revue STICEF* 13, 2006, pp. 289-316. On line http://sticef.univ-lemans.fr/num/vol2006/sticef_2006_som.htm#special

[5] M. Marcoccia, "On-line Polylogues: conversation structure and participation framework in Internet Newsgroups", *Journal of Pragmatics*, 36 (1), 2004, pp. 115-145.

[6] N. Kushmerick, "Wrapper Induction: Efficiency and Expressiveness" *Artificial Intelligence* (118), 2002, pp. 15-68.