Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning

Stefan Trausan-Matu, Mihai Dascălu, Philippe Dessus

Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning

Stefan Trausan-Matu[1], Mihai Dascalu[1], and Philippe Dessus[2]

[1] University Politehnica of Bucharest, 313 Splaiul Independetei, Bucharest, Romania
[2] Grenoble University, 1251, Av. Centrale, BP 47, F-38040 Grenoble CEDEX 9, France
{stefan.trausan,mihai.dascalu}@cs.pub.ro,
philippe.dessus@upmf-grenoble.fr

Abstract. Computer-Supported Collaborative Learning (CSCL) technologies play an increasing role simultaneously with the appearance of the Social Web. The polyphonic analysis method based on Bakhtin's dialogical model reflects the multi-voiced nature of a CSCL conversation and the related learning processes. We propose the extension of the model and the previous applications of the polyphonic method to both collaborative CSCL chats and individual metacognitive essays performed by the same learners. The model allows a tight correlation between collaboration and textual complexity, all integrated in an implemented system, which uses Natural Language Processing techniques.

Keywords: Computer-Supported Collaborative Learning, metacognition, polyphonic model, dialogism, knowledge building, textual complexity, NLP.

1 Introduction

In recent years, Computer-Supported Collaborative Learning (CSCL) grew as an alternate solution to Intelligent Tutoring Systems (ITS) in supporting learning with computers. One of the explanations is the huge spreading of collaborative tools on the web, empowering social knowledge building: discussion forums, instant messenger (chat), social networks, and wikis. The transition from ITS to CSCL may be seen as a change of focus from learning as knowledge acquisition to learning as discourse building [1] or, from a higher abstraction level, from a cognitive to a socio-cultural paradigm. A theoretical basis for CSCL is Bakhtin's dialogism, multi-vocality and polyphony [2, 3, 4]. We further consider that these concepts are present not only in any CSCL dialogical text (e.g., forum posts or chat utterances), but also in texts written by students, in manuals read by them and even in their inner thinking and they can be used for analyzing complex assignments [4].

We propose a model and a system based on the polyphony idea, which considers both the semantic content (at the individual level related to an expert standard, like in ITSs) and the social dimension (at a collaborative level, in CSCL) by analyzing the relationships between texts in a corpus (of the considered domain), texts collaboratively written by students in CSCL chat sessions and their individual metacognitive essays written afterwards, commenting their collaborative activity. To

achieve this aim we used Natural Language Processing (NLP) techniques enabling the computation of both *distances* between voices and the overall *complexity* of threads.

2 Theoretical Considerations

Let us consider students engaged in a distance learning situation (e.g., through an Internet-based platform). Typically, their main goal is to build knowledge through two lines of activities [3], *individual* (read texts, write out notes, essays, summaries from course material) and *collective* (discussions about the course material), which can both be supported either by a teacher or computer-based feedback. All the stakeholders (the computer included) performing these activities 'say something' in natural language, in other words, emit 'utterances' [5] that may become 'voices' populating the distance learning platform, responding to each other. The way a student can, upon a given question (from herself or others), gather information from multiple textual sources (either from course material or chat utterances) in order to compose her own piece of text (mainly, summaries or syntheses) might be viewed as "contexts" in which they try to handle the polyphony of voices.

This framework allows us assume that each utterance can be analyzed by some NLP or Social Network Analysis (SNA) techniques, thus leading to the production of (semi-) automated support of learners' activities [6]. The achievement of the aim of supporting learning with computers should start from a model of how people learn. The development of any model usually begins with deciding the main ingredients to be considered as essential. The core model of ITSs was influenced by Knowledge-Based Systems, taking knowledge as major ingredient. The ITS model is centered on a knowledge base of the target domain, which may be seen as a model of what should be learned. Learners are modelled by the knowledge they acquired, either correct, usually a subset of the domain knowledge base, or erroneous, to be corrected (sometimes also described in knowledge bases). Some other types of knowledge about the particular learner may be considered, as her cognitive profile, emotional state, goals or other motivational facts.

We keep the ITS idea that students' knowledge should be compared with a 'gold standard': experts' knowledge. However, for comparing students' performance (content of chat utterances and written essays) with the desired one (content of a corpus of reference texts), we are using NLP techniques like *Tf-Idf* or Latent Semantic Analysis (LSA) [7]. We consider that a deficiency of the ITS model is its relation to the transfer of knowledge model of learning, that learning is in a very important degree also socially built [1, 3]. Therefore, in addition to keeping an ITS-type semantic based content analysis, a CSCL-like analysis is also needed, because dialog, conversation, and multi-voiced discourse in natural language have major roles: "rather than speaking about 'acquisition of knowledge,' many people prefer to view learning as becoming a participant in a certain discourse" [1].

We further assume that dialogism, multi-vocality and polyphony [2] are in any text, conversation and even thinking. The 'glue' of all these is the idea of voice in a generalized sense: as a word, a phrase, an utterance (written or thought), a discussion thread, a lexical chain, or even a whole text ('utterance' may be used for words with 'echoes', phrases and texts, as Bakhtin mentioned [5]). In our view, an utterance may become a voice if it has an impact by its emission to the subsequent utterances.

3 The Implemented Model

We implemented, using NLP tools, an evaluation model of learners' utterances derived from Bakhtin's dialogic, polyphonic model. The entire analysis process is centered on the utterance graph automatically built from the discourse and is customized for two different types of assessed text: multi-participant chat conversations, on one hand, and essays (texts in general), on the other. Utterances may be considered pieces of text whose boundaries are represented by the change of speech subject [5] and are the central unit of analysis of the discourse in our approach. Whereas in chat conversation we adopt Dong's [8] perspective of separating utterances based on turn-taking events between speakers, in texts, in general, utterances are embedded within sentences that convey relevant information, units that can be separately and independently analyzed in the first phase of the evaluation.

We start the processing with a typical NLP pipe (spell-checking, elimination of stop-words, stemming, part-of-speech tagging and lexicalized dependency parsing [9]). We seek a shallow perspective over each utterance seen individually and we provide them a quantitative mark by merging the concept of entropy from information theory with the *Tf-Idf* measure [9]. The combination of *disorder and emphasis on diversity of concepts* induced by the entropy of stems after stop words elimination, with *summing up statistical importance* of stems given a training corpus, provides a good surface indicator of the information withheld in each utterance (Eq. 1):

$$quant(u) = \left(-\sum_i p(stem_i)\log\left(p(stem_i)\right)\right)\left(\sum_i (1 + |stem_i \in u|)\left(\frac{|D|}{|stem_i \in D|}\right)\right) \quad (1)$$

where: $p(stem_i)$ expresses the probability of a stem to occur in a given utterance; $|stem_i \in u|$ denotes the number of occurrences of each stem within the utterance; $|D|$ and $|stem_i \in D|$ are related to the training corpus used also with LSA that comprises a multitude of documents closely related to the topics at hand and a general set of documents for common words. In this context, entropy is used rather as an inhibitor, where low quality or spam utterances have a lower score.

The key-step is using the Directed Acyclic Graph (DAG) of utterances reflecting the sequential ordering. Our aim is to determine the semantic cohesion between two utterances by means of similarity and degree of inter-connection. Similarity between utterances can be expressed by combining *repetitions* of stems and *Jaccard similarity* as measures of lexical cohesion, with semantic similarity computed by means of LSA. Therefore, Eq. 2 covers the general approach of measuring *cohesion*:

$$\begin{aligned} coh(u,v) &= |repetions| \times \frac{|stems\ in\ common\ u,v|}{|stems\ in\ u\ or\ v|} \times cos(vector(u), vector(v)) \\ vector(u) &= \sum_i (1 + |word_i \in u|) \times \left(\frac{|D|}{|word_i \in D|}\right) \times U_k[word_i] \end{aligned} \quad (2)$$

where $U_k[word_i]$ is the vector of $word_i$ in the U_k matrix obtained after SVD decomposition and projection over k most meaningful dimensions are performed. As a result, for a given conversation the DAG in Figure 1 is obtained automatically.

The next step in our analysis consists in determining the importance of each utterance within the discourse and two additional dimensions, besides quantitative

evaluation, are considered: a *qualitative* one centered on relevance, impact and coherence, and a *social perspective*, seen as an augmentation factor.

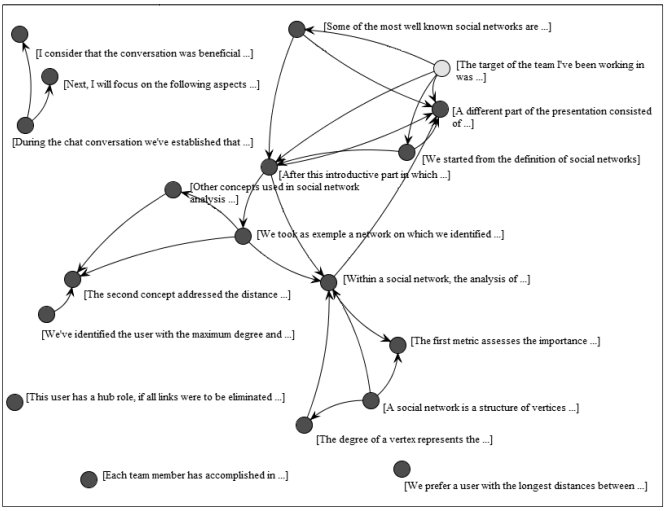

Fig. 1. Example of an utterance graph build upon a student's metacognition (texts are in Romanian)

Relevance is determined with regards to the entire discourse (practically the semantic coherence function applied between a specific utterance and the whole document) and to the vector space by means of cosine similarity between the utterance vector and the mean vector of the LSA vector space. *Completeness* is an optional factor and expresses the semantic similarity between a set of topics (manually defined by the tutor or automatically determined by the system) and the given utterance.

Thread cohesion and *future impact* express the impact of previous and of future utterances that are inter-connected to the current utterance in the utterance graph. These factors are obtained by summing up semantic cohesion values exceeding a threshold; after multiple experiments, the best empirical value of this threshold turned out to be the average minus standard deviation of all the edges of the utterance graph.

From the dialogic perspective, the *current utterance* is seen as a sum of overlapping voices materialized as concepts that are highlighted through information retrieval techniques. *Future impact* encapsulates all future echoes of the current utterance based on the coherence function that expresses both the voice, in the sense of concept repetition, and attenuation, simulated through semantic similarity. Meanwhile, *thread cohesion* acts as a memory function by referring to information previously stated and provides overall discourse coherence as a local function perpetuated through the entire discourse.

From a completely different point of view, the *social perspective* consists in applying social network analysis specific algorithms for estimating an utterances importance in the utterance graph. These metrics include degree, closeness centrality, distance centrality, eigenvector centrality, betweenness centrality and an adjusted version of the well-known Page Rank algorithm [10]. By combining all previous factors, the qualitative factor of each utterance can be expressed as follows (Eq. 3):

$$relevance(u) = \cos\big(vector(u), vector(doc)\big) + \cos(vector(u), LSA\ vector\ mean) + \cos(vector(u), vector(topics))$$

$$social(u) = \prod_{SNA\ factor\ f} (1 + \log(f(u))) \tag{3}$$

$$qualitative(u) = \left(\sum_{\substack{i=1..m \\ v_i \to u}} coh(v_i, u) + 1 + \sum_{\substack{k=1..n \\ u \to v_k}} coh(u, v_k) \right) \times relevance(u) \times social(u)$$

Regarding the social factor, a normalization induced by the logarithm function provided a smoothing of results. The factor 1 in the coherence values sum expresses internal strength in a discussion thread and was induced by the cosine similarity measure applied between utterance *u* and itself. By combining the quantitative mark with the qualitative score, the overall rating of each utterance is obtained (Eq. 4):

$$overall(u) = \left(\sum_{\substack{i=1..m \\ v_i \to u}} coh(v_i, u) quant(v_i) + quant(u) + \sum_{\substack{k=1..n \\ u \to v_k}} coh(u, v_k) quant(v_k) \right) \times relevance(u) \times social(u) \tag{4}$$

Eq. 4 clearly comprises all factors required for thoroughly evaluating an utterance: its local and individual formula, its importance within all discourse threads measured through semantic cohesion with previous and future inter-connected utterances, its relevance expressed in terms of semantic similarity with the entire document, topics of discussion and the LSA learning space, but also social networks analysis applied on the utterance graph in order to integrate centrality features in our approach.

After having all previous assessments completed, *textual complexity* can be evaluated and gains the focus of the entire analysis. Due to the fact that textual complexity cannot be determined by enforcing a single factor of evaluation, we propose a multitude of factors, categorized in a multilayered pyramid, from the simplest to the more complex ones, that combined provide relevant information to the tutor regarding the actual "hardness" of a text [11]. The first and simplest factors are at a *surface level* and include readability metrics, utterance entropy at stem level and proxies extracted from Page's [12] automatic essay grading technique. Slotnick's six factors [13] of fluency, spelling, diction, sentence structure, punctuation and paragraph development are the main factors we implemented in our system.

At the *syntax level*, structural complexity is estimated from the parsing tree in terms of max depth and of max width [14]. Moreover, entropy applied on parts of speech and the actual number of specific parts of speech (mostly pronouns, verbs and nouns) provide additional information at this level. *Semantics* is addressed through topics that are determined by combining *Tf-Idf* with cosine similarity between the utterance vector and that of the entire documents. The textual complexity at this level is expressed as a weighted mean of the difficulty of each topic, estimated in computations as the number of syllables of each word. The last level of *pragmatics* and *discourse* addresses textual complexity as cohesion determined upon social networks analysis metrics applied at macroscopic level. Discourse markers, co-references, rhetorical schemas and argumentation structures are also considered, but are not included in current work.

By considering the disparate facets of textual complexity and by proposing possible automatic methods of evaluation, the resulted measurement vectors provide tutors valuable information regarding the hardness of presented texts.

4 Conclusions and Future Research Directions

Borrowing from Bakhtin's dialogism and polyphony theories, we devised a framework that takes into account several dimensions of learners' activities in CSCL. Reading course materials, understanding them, discussing about them produce utterances seen as polyphonic voices interacting to each other. Our model automatically assesses these utterances at multiple levels (cognitive, metacognitive, social), and accounts for learner's comprehension of textual materials.

Acknowledgement. This research was partially supported by project 264207, ERRIC-Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1.

References

1. Sfard, A.: On reform movement and the limits of mathematical discourse. Mathematical Thinking and Learning 2(3), 157–189 (2000)
2. Bakhtin, M.M.: Problems of Dostoevsky's poetics. University of Minnesota Press, Minneapolis (1993)
3. Stahl, G.: Group cognition. MIT Press, Cambridge (2006)
4. Trausan-Matu, S., Stahl, G., Sarmiento, J.: Supporting polyphonic collaborative learning. E-service Journal 6(1), 58–74 (2007)
5. Bakhtin, M.M.: Speech genres and other late essays. University of Texas, Austin (1986)
6. Dessus, P., Trausan-Matu, S.: Implementing Bakhtin's dialogism theory with NLP techniques in distance learning environments. In: Trausan-Matu, S., Dessus, P. (eds.) Proc. 2nd Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity (NLPsL 2010), pp. 11–20. Matrix Rom, Bucharest (2010)
7. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. 104(2), 211–240 (1997)
8. Dong, A.: The language of design: Theory and computation. Springer, New York (2009)
9. Jurafsky, D., Martin, J.H.: An introduction to natural language processing. Computational linguistics, and speech recognition. Pearson Prentice Hall, London (2009)
10. Nguyen, Q.H., Hong, S.-H.: Comparison of centrality-based planarisation for 2.5D graph drawing. NICTA technical report, Sidney (2006)
11. Dascalu, M., Trausan-Matu, S., Dessus, P.: Utterances assessment in chat conversations. Research in Computing Science 46, 323–334 (2010)
12. Page, E.: The imminence of grading essays by computer. Phi Delta Kappan 47, 238–243 (1966)
13. Wresch, W.: The imminence of grading essays by computer—25 years later. Computers and Composition 10(2), 45–58 (1993)
14. Gervasi, V., Ambriola, V.: Quantitative assessment of textual complexity. In: Merlini Barbaresi, L. (ed.) Complexity in Language and Text, pp. 197–228. Plus, Pisa (2002)