A Tool for Discourse Analysis and Visualization

Costin-Gabriel Chiru, Department of Computer Science, University Politehnica of Bucharest, Bucharest, Romania

Stefan Trausan-Matu, Department of Computer Science, University Politehnica of Bucharest, Bucharest, Romania

ABSTRACT

In this paper the authors present a system that combines the cognitive and socio-cultural paradigms in the field of discourse analysis in order to analyze both texts written by only one author (for example narrations) and those written collaboratively (chat conversations, blogs, wikis, forums). The novelty of their approach is that the majority of the existing applications are oriented on analyzing only one of these two types, an adaptation being necessary for the analysis of the other type. Another advantage of the presented system is that since it is centered on a dialogistic polyphonic model considering topics as inter-animated voices, it could show the difference between coarse- and fine-grained coherence in discourse, therefore allowing the analysis of a text from two different viewpoints: a) its intrinsic structure and cohesion and b) how well this text fits in a stream of texts (whether it is or not cohesive with the texts before and after it). The dialogistic polyphonic model was used as a starting point for a method for analyzing collaboration and social construction of knowledge in groups and communities using textual interactions, and for several implemented systems for providing computerized support to the analysis method through visualizations and feedback generation.

Keywords: Bakhtin, Data Visualization, Lexical Chains, Natural Language Processing, Polyphonic Model

INTRODUCTION

Lately, one can see a tendency toward an increased use of computers on the so-called Social Web, for both leisure and work. There is an intense use of instant messaging (chats), blogs, wikis, forums, social networks, etc. for informal talks in our spare time, for collaborative knowledge building in different kinds of virtual communities. Moreover, lots of newspapers and libraries decided to provide their content as digital documents on the Internet; the number of websites sharing information about different topics is continuously growing and so on. Consequently, very large quantities of information are available in digital format to whole communities. These documents are the result of communities interactions mediated by language in a socio-cultural knowledge building perspective (Vygotsky, 1978). We may say that
they incorporate the voices, the ideas of many persons in the community, in a dialogistic way (Bakhtin, 1981).

One of the most complex facets of text analysis is discourse and most of the applications for its analysis are biased towards one of the two types of texts: written by only one author (we will call them narrations) and those written collaboratively (we call them conversations, but we include here instant messenger chats, transcribed face-to-face conversations, forums, wikis, blogs, etc). We consider that this dichotomy is artificial. Texts, even narrations and essays are always written to be read by others; they are directed towards an implicit dialog with a community; they contain not only the voice of the author (Bakhtin, 1981). The latter should consider even implicitly the potential voices of the readers from the community.

Tools for text automatic analysis are very important for the development of the social web. Information retrieval and associated tools (for example, provided by Google) are now intensively used. Visualization means such as word clouds, tree or graph-based representations of key concepts are provided in many environments. However, they are based on a bag of words model, considering only the frequency of apparitions of these words, and not the discourse structures.

In this paper we present a unified method and a visualization tool for analyzing both kinds of texts (narrations and conversations) within the same framework. The method presented here combines the socio-cultural and cognitive paradigms using the concept of voices’ inter-animation from the polyphonic model (Trausan-Matu, 2010; Trausan-Matu & Rebedea, 2010), based on Bakhtin’s ideas (1981), and respectively the WordNet (http://wordnet.princeton.edu) linguistic database-based processing for analyzing texts. It uses Natural Language Processing (NLP) techniques (Jurafsky & Martin, 2009) (such as building lexical chains starting from the given text and a linguistic database) and the ideas related to identifying polyphonic threads (Trausan-Matu & Rebedea, 2010). Computer-Supported Collaborative Learning is one of the most important applications of the instruments of the social web, of the social construction of knowledge in small groups and in virtual communities (Stahl, 2006). They make use of instant messenger (chat), discussion forums, wikis, blogs or social networking and the participants now enter in massive open online courses. Automatic tools for analyzing discourse in the huge amount of interconnected documents and logs of interaction (in chats, wikis, blogs, forums and social networks) are obviously needed.

In the next section we will present the theoretical ideas that lead to the presented method and to the implementation of the system. The paper continues with the presentations of the system and of the possible uses of the provided visualizations. An evaluation of this system for the task of analyzing Computer-Supported Collaborative Learning (CSCL) chats (Stahl, 2006) is presented right before our conclusions regarding the proposed method and the implemented system.

**THEORETICAL IDEAS**

Jurafsky and Martin (2009) define discourse as a coherent group of sentences. We accept this definition, but we consider it is too simple and we extend it in a socio-cultural direction, according to the dialogistic view of Bakhtin (1981), considering that in the center of the analysis should be not the sentences but the socially-produced utterances. The latter may be replies in chats, posts in discussion forums and social networks, words, sentences, documents, essays, and even whole books. Moreover, we consider that discourse should be seen as an intertwining of at least two threads corresponding to dialoguing voices (Trausan-Matu, Stahl & Sarmiento, 2007). Most of the theories from the discourse analysis field can be categorized as a representative of one of two major directions. The first direction, used currently in NLP, is based on the cognitive paradigm, stating that knowledge is situated in individuals’ minds (Tannen, 1989), being focused on the way the
Social Networking and Schools: Early Responses and Implications for Practice
www.igi-global.com/chapter/social-networking-schools/39801?camid=4v1a