Ontology-Based Analyze of Chat Conversations. An Urban Development Case[1]

Stefan Trausan-Matu[1,2], Traian Rebedea[1]

[1] "Politehnica" University of Bucharest,
Department of Computer Science and Engineering,
Splaiul Independetei nr. 313,
Bucharest, Romania
[2] Research Institute for Artificial Intelligence of the Romanian Academy,
Calea 13 Septembrie nr.13,
Bucharest, Romania
trausan@cs.pub.ro, rebedea@cs.pub.ro

Abstract. Online collaboration in communities of urbanism experts is enabled by text-based tools, such as instant messaging (chat), forums and web logs (blogs). The paper presents an ontology based system that analyses chat logs. The system integrates knowledge processing with natural language processing and discourse analysis based on Bakhtin's ideas. The system permits to detect the topics of chat, the threads of discussion and the important utterances. It also visualizes the graph of the conversation and allows the extinction of the domain ontology.

Keywords: ontology, semantic closeness, dialogism, chat visualization,

1 Introduction

The social interaction tools on the web, like discussion forums, blogs, wikis or instant messaging are basic components of the Web2.0 (the Social Web). Such tools started to be used also by communities of specialists in urbanism[2]. From the above mentioned tools, instant messaging (chat)[3], due to its specific online character, encourages multi-voiced inter-animation for collaboratively building knowledge, in a way similar to classical music polyphony or jazz improvisation [7, 8].

For developing chat analysis programs, human language processing is needed and, therefore, ontologies play a central role, because they constitute a basic knowledge representation framework for semantic analysis. In particular, in the system presented in this paper, ontologies are used for the identification of similar word threads (lexical chains), of the discussion topics and of the important utterances of a chat. The general

[1] In Teller J., Cutting-Decelle, A.F., Billen, R., (eds.) Urban Ontologies for an Improved Communication in Urban Development Projects, Ed. de l'Universite de Liege, pp. 119-127, 2009.

[2] See, for example http://www.cyburbia.org or http://www.planetizen.com

[3] An example of a chat session in the domain of urbanism may be seen at http://www.planetizen.com/node/30186 or http://www.planetizen.com/node/30813 (last accessed on 28 January 2009).

lexical WordNet ontology (http://wordnet.princeton.edu) is used together with a domain ontology.

One important feature of the system presented in this paper is that the domain ontology is extensible. For example, after new concepts are identified as a result of the analysis of the chats, the user may include them in the domain ontology and also introduce relations among them.

In addition to classical ontology-based natural language processing techniques, the polyphonic model of Bakhtin is used in order to identify inter-animation patterns among chats' discourse threads [7]. The same framework may be used also for the analysis of other social interaction tools (forums or blogs).

The paper continues with a section introducing some ideas about ontologies. The third section discusses the socio-cultural and Bakhtin's dialogism paradigms. The next section contains the description of the visualization and ontology expansion tool. The paper ends with conclusions and references.

2 Ontologies and semantic closeness

Ontologies are semantic networks modeling human conceptualization, built either manually or automatically, for example, by extracting knowledge from texts (text mining). Ontologies may be seen also as ways of sharing concepts, classifications and inter-relations in communities. Any collaboration using natural language, any discourse needs to start from a common vocabulary or, a more structured alternative, a shared ontology. WordNet or FrameNet (http://framenet.icsi.berkeley.edu) are examples of general ontologies built as extended vocabularies, offering additional linguistic data like related words, or case grammars. In addition to general ontologies, communication in communities of practice needs particular concepts from specific domain ontologies.

The word "ontology" is taken from philosophy, where it denotes the theory about what is considered to exist. Any system in philosophy starts from an ontology, from the identification of the concepts and relations considered as fundamental. Ontologies capture fundamental categories, concepts, their properties and relations. One very important relation among concepts is the taxonomic one, from a more general to a more specific concept. This relation may be used as a way of "inheriting" properties from the more general concepts ("hypernyms"). Other important relations are "part-whole" ("meronym"), "synonym" and "antonym".

Ontologies are very important in knowledge extraction from texts, in general, and from conversations, in particular. For these kind of applications, they offer the substrate for semantic analysis and, very important, the possibility of defining a measure of semantic closeness, based on the graph with concepts from ontologies as nodes and their relations as arcs [2].

The measures of semantic distances allow to identify groups of similar concepts and, therefore, to identify lexical chains of related words. These chains, together with repetitions and anaphors allow to further identify threads of discussions in texts. These threads may interact, according to polyphonic patterns [7].

3 Bakhtin’s polyphonic theory

In forums and chat conversations, knowledge is socially built through discourse and is preserved in linguistic artifacts whose meaning is co-constructed within group processes [4, 5]. These socio-cultural ideas are based on the work of Lev Vygotsky, who emphasized the role of socially established artifacts in communication and learning [9].

Mikhail Mikhailovici Bakhtin has extended the ideas of Vygotsky, emphasizing the role of speech and dialog in analyzing social life. He remarks that in each dialog and even in written texts there are communities of voices: “The intersection, consonance, or interference of speeches in the overt dialog with the speeches in the heroes’ interior dialogs are everywhere present. The specific totality of ideas, thoughts and words is everywhere passed through several unmerged voices, taking on a different sound in each” [1]. This dual nature of community and individuality of voices is expressed by Bakhtin also by the concept of *polyphony*, that he considers the invention and one of the main merits of Dostoevsky novels [1]. The relation of discourse and communities to music was remarked also by Tannen: “Dialogue combine with repetition to create rhythm. Dialogue is liminal between repetitions and images: like repetition is strongly sonorous” [6].

In chat conversations, different voices are obvious recognized. However, starting from Bakhtin’s ideas, in our approach the concept of voices is not only limited to the physical vocal characteristics of participants in the chat. A voice is, from our perspective, something said by a participant in a given moment and it may be reflected in many subsequent utterances. Also, each utterance may contain an unlimited number of voices.

4 Ontology-based chat analysis

The approach presented here integrates Bakhtin’s socio-cultural ideas with knowledge-based natural language processing for the identification of the topics discussed in the chat, for the detection of discussion threads and of the most important utterances in a chat. Such a system may be used, for example, for tracking the most important topics discussed by a group of experts in urbanism for solving a given problem. The chat system used in the experiments presented here was ConcertChat [3], which allows the explicit referencing of previous utterances, a facility that enables the existence of multiple discussion threads in parallel and their inter-animation.

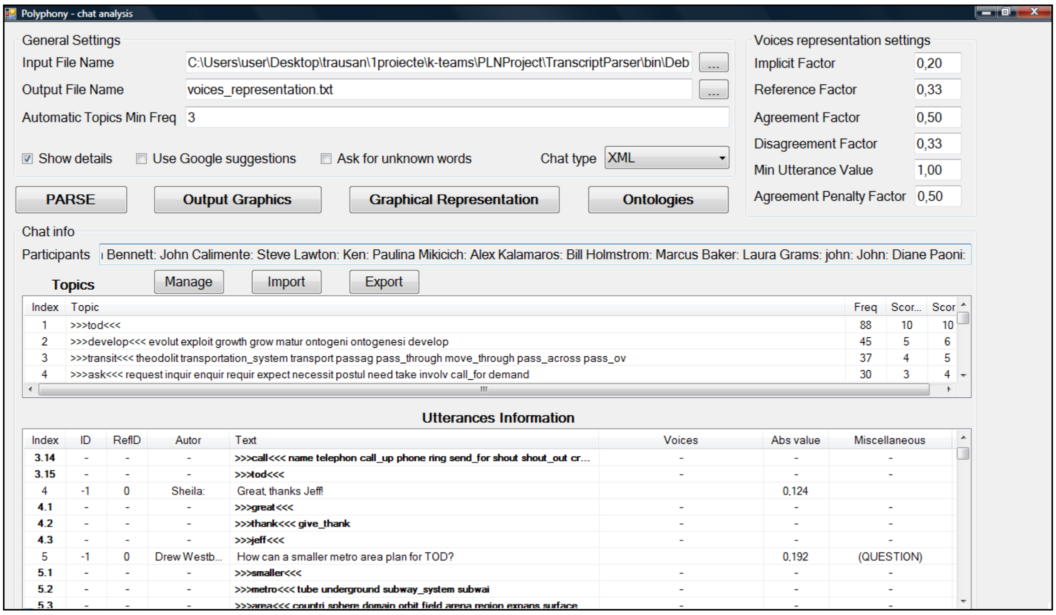
Determining the topics of a chat

The chat topics are identified as a list of concepts (words) that appeared most frequently in the conversation, by using statistical natural language processing methods. Accordingly, the importance of a subject is considered related to its frequency in the chat.

The first step in finding the chat subjects is to strip the text of irrelevant words (stop-words), text emoticons (e.g. “:)”, “:D”, and “:P”), special abbreviations used while chatting (e.g. “brb”, “np”, and “thx”) and other words considered of no use at this stage. Then, the resulted chat is tokenised and each different word is considered a candidate concept in the analysis. For each of these candidates, WordNet and the domain ontology are used for finding synonyms.

The last stage for identifying the chat subjects consists in unifying the candidate concepts discovered in the chat. This is done by using the synonym list for every concept: if a concept in the chat appears in the list of synonyms of another concept, then the two concepts’ synonym lists are joined. At this point, the frequency of the resulting concept is the added frequencies of the two unified concepts. This process continues until there are no more concepts to be unified. At this point, we acknowledge the subjects of the chat conversation as the list of resulting concepts, ordered by their frequency.

Figure 1 is a screenshot illustrating some topics identified in an urbanism chat (http://www.planetizen.com/node/30186).



General Settings

Input File Name: C:\Users\user\Desktop\trausan\1proiecte\k-teams\PLNProject\TranscriptParser\bin\Deb

Output File Name: voices_representation.txt

Automatic Topics Min Freq: 3

☑ Show details □ Use Google suggestions □ Ask for unknown words Chat type: XML

Voices representation settings

Implicit Factor	0,20
Reference Factor	0,33
Agreement Factor	0,50
Disagreement Factor	0,33
Min Utterance Value	1,00
Agreement Penalty Factor	0,50

Chat info

Participants: ı Bennett: John Calimente: Steve Lawton: Ken: Paulina Mikicich: Alex Kalamaros: Bill Holmstrom: Marcus Baker: Laura Grams: john: John: Diane Paoni:

Topics

Index	Topic	Freq	Scor...	Scor
1	>>>tod<<<	88	10	10
2	>>>develop<<< evolut exploit growth grow matur ontogeni ontogenesi develop	45	5	6
3	>>>transit<<< theodolit transportation_system transport passag pass_through move_through pass_across pass_ov	37	4	5
4	>>>ask<<< request inquir enquir requir expect necessit postul need take involv call_for demand	30	3	4

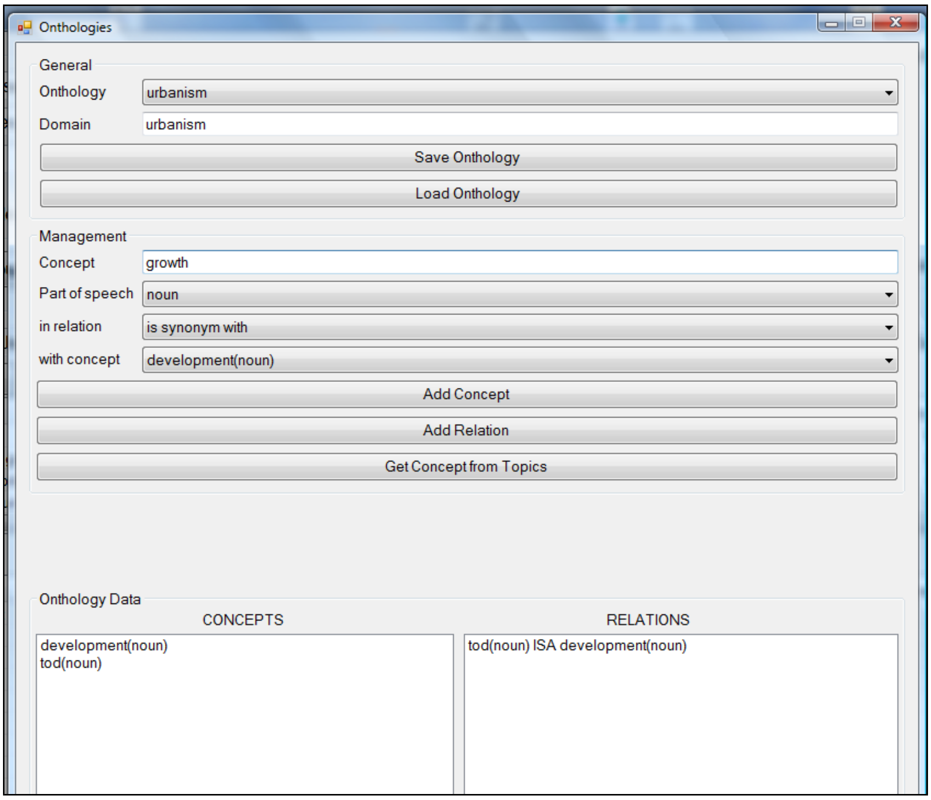
Utterances Information

Index	ID	RefID	Autor	Text	Voices	Abs value	Miscellaneous
3.14	-	-	-	**>>>call<<< name telephon call_up phone ring send_for shout shout_out cr...**	-	-	-
3.15	-	-	-	**>>>tod<<<**	-	-	-
4	-1	0	Sheila:	Great, thanks Jeff!		0,124	
4.1	-	-	-	**>>>great<<<**	-	-	-
4.2	-	-	-	**>>>thank<<< give_thank**	-	-	-
4.3	-	-	-	**>>>jeff<<<**	-	-	-
5	-1	0	Drew Westb...	How can a smaller metro area plan for TOD?		0,192	(QUESTION)
5.1	-	-	-	**>>>smaller<<<**	-	-	-
5.2	-	-	-	**>>>metro<<< tube underground subway_system subwai**	-	-	-
5.3	-	-	-	**>>>area<<< countri sphere domain orbit field arena region expans surface**	-	-	-

Figure 1. Identification of chat topics in an urbanism conversation

Extending the domain ontology with topics determined from the chat

The first three topics identified by the system (see figure 1) are tod (”transit oriented development”), develop and transit. If we don’t have these concepts in the domain ontology, the system offers the possibility of adding them. The user may add also relations among new concepts and save the ontology. In figure 2 is an excerpt of the usage of this facility.



Onthology: urbanism

Domain: urbanism

Concept: growth

Part of speech: noun

in relation: is synonym with

with concept: development(noun)

CONCEPTS	RELATIONS

development(noun)
tod(noun)
	tod(noun) ISA development(noun)

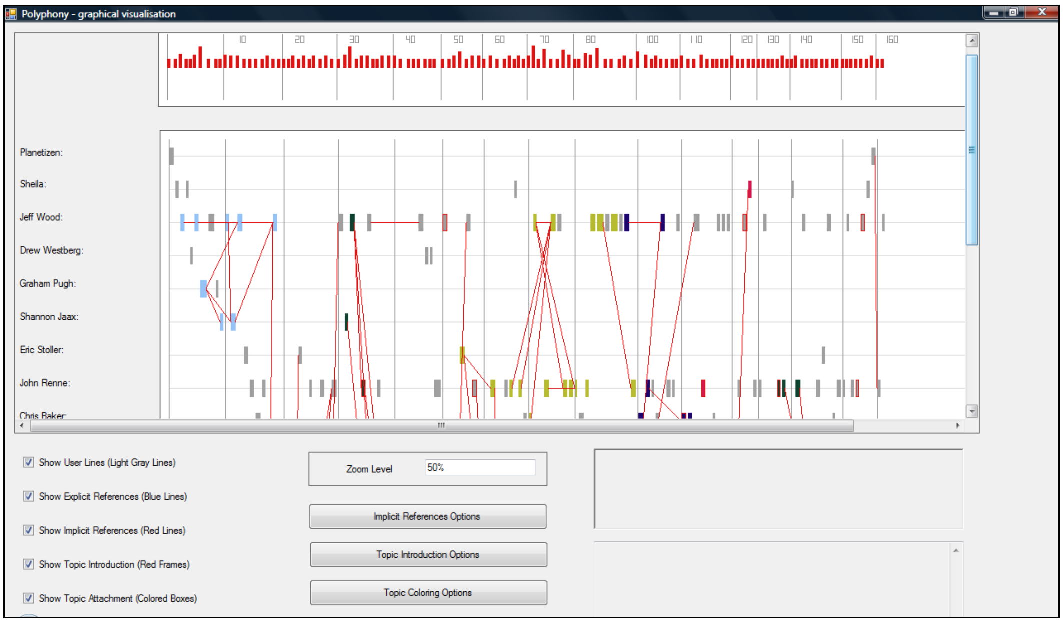
Figure 2. Extending the domain ontology

The graphical representation of the conversation

Starting from existing references within the analyzed conversations, both those explicit, allowed by the chat environment (ConcertChat [3]), as well as those implicit, determined by the program, a graph that visualizes the conversation is built. Within this graph, each utterance from the chat is a vertex, while the references between utterances (either explicit or implicit) represent the edges. The output is a directed graph specific to the conversation.

The graphical representation of the chat was designed to permit the best visualization of the conversation, to facilitate an analysis based on the polyphony theory of Bakhtin, and to maximize the straightforwardness of following the chat elements. For each participant in the chat, there is a separate horizontal line in the representation and each utterance is placed in the line corresponding to the issuer of that utterance, taking into account its positioning in the original chat file – using the timeline as an horizontal axis. Each utterance is represented as a rectangle aligned according to the issuer on the vertical axis and having a horizontal axis length that is proportional with the dimension of the utterance. The distance between two different utterances is proportional with the time passed between the utterances. Of course, there is a minimum and a maximum dimension for each measure in order to restrict anomalies that could appear in the graphical representation due to extreme cases or chat logging errors.

The relationships between utterances are represented using colored lines that connect these utterances. The explicit references that are known due to the use of the ConcertChat software are depicted using blue connecting lines, while the implicit references that are deduced using the method described in this paper are represented using red lines. The utterances that introduce a new topic in the conversation are represented with a red margin.



☑ Show User Lines (Light Gray Lines)

☑ Show Explicit References (Blue Lines)

☑ Show Implicit References (Red Lines)

☑ Show Topic Introduction (Red Frames)

☑ Show Topic Attachment (Colored Boxes)

Zoom Level 50%

Fig. 3. The threads of references in the chat from figure 1

The graphical representation of the chat has a scaling factor that permits an attentive observation of the details in a conversation, as well as an overview of the chat. The different visual elements determined by our application – such as utterances in the same topic, topic introducing utterances and relationships between topics – can be turned on and off in the graphical representation by use of checkboxes.

At the top of the graphical representation of the conversation, there is a special area that represents the importance of each utterance, considered as a chat voice, in the conversation. How this importance is determined is presented further in this paper. Moreover, all the details of an utterance in the chat – the content of the utterance, the implicit and explicit references and other details – can be visualized by clicking the rectangle representing the utterance.

Determining the importance of an utterance

The importance of an utterance in a conversation can be calculated through its length and by the correct selection of the words in the utterance – they should contain as many possible key (important) words. This approach could prove useful in chat summarization. Nevertheless, in a social context, another approach is also possible: an utterance is important if it influences the subsequent evolution of the conversation. Using this definition as a starting point, we may infer that an important utterance will be that utterance which is a reference for as many possible subsequent utterances.

Even if this approach could be extended to include the types of subsequent references (implicit or explicit, agreements or disagreements), in the present case we have preferred a more simplistic approach, without making allowances for the types of references to the utterance.

Consequently, the importance of an utterance can be considered as a strength value of an utterance, where an utterance is strong if it influences the future of the conversation (such as breaking news in the field of news). When determining the strength of an utterance, the strength of the utterances which refer to it is used. Thus, if an utterance is referenced by other utterances which are considered important, obviously that utterance also becomes important.

As a result, for the calculation of the importance of every utterance, the graph is ran through in the opposite direction of the edges, as a matter of fact in the reverse order of the moment the utterance was typed. Utterances which do not have references to themselves (the last utterance of the chat will certainly be one of them) receive a default importance – taken as the unit.

Then, running through the graph in the reverse order of references, each utterance receives an importance equal to that of the default plus a quota (subunit) from the sum of the importance of the utterances referring to the current utterance. Another modality to calculate could be 1 plus the number of utterances which refer to the present utterance, but this choice seemed less suitable.

By using this method of calculating the importance of an utterance, the *utterances* which have started an important conversation within the chat, as well as those *utterances* which begin new topics or mark the passage between topics, are more easily emphasized. If the explicit relationships were always used and the implicit ones could be correctly determined in as high a number as possible, then this method of calculating the importance of a voice would be successful.

Identifying discussion threads

Using an algorithm for determining the connected components from the conversation graph, we were able to find the *utterances* connected through at least one relationship. It is normal to assume that all these *utterances* are part of a single discussion topic.

This method can be used for successfully finding the conversational threads. We have considered that the important topics are those consisting of at least four *utterances*. This minimum number of *utterances* in a topic should be parameterized according to the length of the chat, but 4 *utterances* is considered to be a minimum. For each determined topic, we have highlighted the most frequent concepts (as a synset list) in that topic. This way, each topic is described by the most relevant concepts found in the *utterances* present in that topic.

An interesting observation to be made is that this method to determine the topics of the conversation produces some remarkable results. Thus, the discussion can have more than one topic at a moment in time – the participants being involved in different topics at the same time. Inter-crossings between different topics can be easily observed on the chat graphics as well as topics started and finished whereas other more important topics are abandoned for a while and then continued.

This method can be improved by considering the similarities between the *utterances* in closely related topics. We can also combine this solution with an analysis of the time passed between the *utterances*. Two similar *utterances* that are separated by a great distance in time can be considered part of different topics.

5 Conclusions

The paper presents an application that detects the topics of chat, the threads of discussion and the important utterances. It also visualizes a graph of the conversation. The application may be used for inspecting what is going on and in what degree participants are implied in a chat conversation, for example, a group of urbanism specialist discussing how to solve a given problem.

The application uses the WordNet ontology and domain ontologies. Natural language technology is used for the identification of discussion topics, for segmentation and for identifying implicit references. The domain ontology may be extended as a result of new topics identified by the system. For this aim, an editing interface has been implemented.

Further work will consider more complex semantic distances (than only synonymy). Machine learning techniques will be used for the identification of discourse patterns. Moreover, a completely automated version for discovering new rules for the implicit relations is in progress.

References

1. Bakhtin, M.M., Problems of Dostoevsky's Poetics, Ardis (1973)
2. Hirst, G., St-Onge, D., "Lexical chains as representations of context for the detection and correction of malapropisms". in Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database, chapter 13, pages 305 - 332. The MIT Press, Cambridge, MA, (1998).
3. Holmer, T., Kienle, A., Wessner, M. "Explicit Referencing in Learning Chats: Needs and Acceptance," in Innovative Approaches for Learning and Knowledge Sharing, First European Conference on Technology Enhanced Learning, EC-TEL 2006, Nejdl, W., Tochtermann, K., (eds.), LNCS, 4227, pp. 170-184, Springer (2006)
4. Koshmann, T., Toward a Dialogic Theory of Learning: Bahtin's Contribution to Understanding Learning in Settings of Collaboration, in C.Hoadley and J. Roschelle (eds.), Proceedings of the Computer Support for Collaborative Learning 1999 Conference, Stanford, Laurence Erlbaum Associates.
5. Schegloff, E., Discourse As An Interactional Achievement: Some Uses Of 'Uh huh' And Other Things That Come Between Sentences, in Tannen, D. (ed.), Georgetown University Roundtable on Languages and Linguistics 1981; Analyzing Discourse: Text and Talk, Georgetown University Press, Washington D.C.
6. Tannen, D., Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse, Cambridge University Press (1989)
7. Trausan-Matu, S., Stahl, G., Sarmiento, J., Supporting Polyphonic Collaborative Learning, E-service Journal, vol. 6, nr. 1, Indiana University Press, 2007, pp. 58-74.

8. Trausan-Matu, S., Rebedea, T., Polyphonic Inter-Animation of Voices in VMT, in Stahl, G. (ed.), Studying Virtual Math Teams, Springer, to appear in 2009.

9. Vygotsky, L. (1978). Mind in society, Cambridge, MA: Harvard University Press