STATIC AND DYNAMIC KNOWLEDGE FOR THE EMBODIED CONVERSATIONAL AGENT MAX

- A FIRST APPROACH -

Alexa Breuing
Artificial Intelligence Group
Faculty of Technology
University of Bielefeld
33594 Bielefeld, Germany

Ipke Wachsmuth
Artificial Intelligence Group
Faculty of Technology
University of Bielefeld
33594 Bielefeld, Germany

ABSTRACT

The online encyclopedia Wikipedia is currently one of the best known collaboratively edited knowledge sources. While humans are able to inform themselves by reading the according articles in Wikipedia, machines lack the ability to understand natural language. Thus, a central task of bridging this deficit is the text technological reconstruction of the information held in Wikipedia to make this knowledge available to artificial agents. Our paper presents an approach for a connection between the embodied conversational agent Max and an ontology-based representation of encyclopedic knowledge held in Wikipedia. Thereby we strive for an automatic distinction between static and dynamic knowledge to enable a more human-like knowledge handling by the agent. The approach to infer stable knowledge proposed in this paper grounds on the so-called history pages provided by Wikipedia.

KEYWORDS

Ontologies, Collaborative Systems, Human Computer Interaction, Information Retrieval

1. INTRODUCTION

The representation and exchange of knowledge constitutes a principal task in artificial intelligence fields. Due to the possibility of making information machine-processable by a description via concepts and their relations, *ontologies* have become widely accepted for this purpose (Hesse 2002). However, ontology development is a complex process that is often hindered by the limited number of structured representations of broad knowledge available. Thereby, receiving the required information automatically is made difficult.

The web-based encyclopedia Wikipedia offers a huge amount of textual articles describing single topics (Suchanek et al. 2007). Due to its policy of letting anyone on the internet create and edit its articles, Wikipedia is among the most successful collaborative writing systems. By taking account of the categorization and the link structure of the approx. 750,000 articles of the German Wikipedia, the semi-structured topics can be arranged in a taxonomy. The additional consideration of the textual context of the articles and the lexical information, e.g. provided by Wiktionary, allows the development of an ontology-based knowledge base for an embodied conversational agent (ECA) adapted from Wikipedia.

In the initial phase of my dissertation project, this work proposes an approach for an ontology-based representation of encyclopedic knowledge for the ECA Max (Kopp et al. 2005). Thereby we distinguish between *static* and *dynamic* knowledge. In our work we define static knowledge as knowledge which is constant over an extended period, and dynamic knowledge as knowledge which changes over time. Based on these two knowledge categories, the considered agent should "know" which part of his knowledge described by the ontology forms the so-called basic knowledge, and on which information he has to check up regularly to keep himself up-to-date. This background knowledge would enable Max to evaluate the reliability of the statements made in a conversation with a human user. On this account, Max would be able to communicate his knowledge more human-like, e.g., in that he could insist on the correctness of a specific statement which is specified as static knowledge.

2. RELATED WORK

Wikipedia's open online access to a cornucopia of readily available information attracts the attention of many academics. Research in software engineering currently investigates how the semi-structured information held in Wikipedia can be extracted automatically to define a machine-readable representation of this encyclopedic knowledge. Thereby, the uniform resource identifiers (URIs) as unique identifiers for the same amount of topics are mostly used as identifiers for conceptual entities to annotate knowledge assets (e.g. implemented in (Hepp et al. 2007, Krötzsch et al. 2005)), as more than 90% of the entries show a completely stable meaning (Hepp et al. 2007). Further approaches, like the Yago ontology (Suchanek et al. 2007), take advantage of the provided classification of Wikipedia articles and construct a taxonomy based on these categories. Both, the usage of the URIs and the consideration of the categories provided by Wikipedia are supposed to establish the basis for our information extraction approach.

Ponzetto and Strube (2007) expand the development of a taxonomy by building up a subsumption hierarchy dividing the relations between related concepts into *isa* and *notisa* relations. Such a subsumption hierarchy does not settle our claims concerning the ontology-based knowledge base for the agent Max. To enable a more human-like knowledge handling and to realize natural language processing based on the represented information, an explicit distinction and definition of relations within the ontology are required. Potential approaches are described by Ruiz-Casado et al. (2006) and Suchanek et al. (2007).

We are not aware of any approaches for the distinction of static and dynamic knowledge within the knowledge provided by Wikipedia. Nevertheless, there are already a number of works considering the so-called *history pages* which archive the edits of the articles and thus deliver ideas for the implementation of distinction techniques. Figure 1 shows a section of a sample history page from the German Wikipedia.

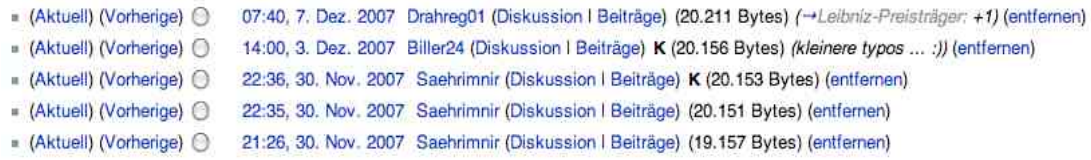

Figure 1. Section of history page of the German Wikipedia article about Bielefeld University. Each row contains a link to the current version, a link to the documented differences between the current and the previous version, a time specification of the latest change, a specification of the person who made the change and comments about the change

For instance, Viégas et al. (2004) display the changes concerning the length of a page over time via a *history flow* visualization. The resulting graph clarifies that most Wikipedia pages show a continual change in size. However, this approach excludes the possibility to define an article (and thus the contained information) as dynamic when it doesn't show a reduction in regard to the number of edits made. Instead, the approach of Buriol et al. (2006) for clustering the articles per update profile shows the correlation between an external news event and a large number of updates on a single article. This conclusion establishes an approach to define dynamic knowledge based on the history pages of Wikipedia. In conjunction with the

consideration of article changes modeled as the number of words added and removed as defined by Kittur et al. (2007), the event-update correlation forms our main starting point.

3. STATIC AND DYNAMIC KNOWLEDGE FOR MAX

Max (see Fig. 2) is a virtual agent that aims to enable natural (multimodal) conversations with human users. His verbal communication is realized by a dialog system which provides rule plans to define the conversational knowledge of Max. This rule-based knowledge enables both, the interpretation of natural language inputs via pattern matching processes and the generation of adequate answers to these user inputs. However, this way of input processing bears some weaknesses. On the one hand, the real content and topic of the verbal user expression cannot be acquired due to the fact that a machine-processable representation of the dialog knowledge is missing. On the other hand, the more rules are defined, the more complex the maintenance of the rule-based knowledge will be. Currently, the agent's rule-set contains about 2,000 unordered rules and the complexity of adapting Max's knowledge to possible changes is accordingly high.

Figure 2. The conversational agent Max at the HNF Forum in Paderborn, Germany

We aim to enhance Max by exploring the German online encyclopedia Wikipedia as the agent's primary knowledge resource via a connection between the two and, thus, to overcome the described problems of the current knowledge handling. For this purpose, the knowledge contained in Wikipedia needs to be structured in an ontology, a suitable representation formalism which meets the requirements of the agent's dialog system as shown by the experiences gained in previous work (Breuing 2007). Thereby the consideration of the URIs and the categories provided in Wikipedia will support the design of the ontology (see chapter 2). Furthermore, the additional inclusion of information from Wiktionary will equip the ontology with lexical information. Besides object and linguistic knowledge, the ontology will be endowed with information required for further IE processes. For instance, the information might consist of specifications regarding location and time which are necessary for the update of dynamic knowledge. According to this, both, the IE from Wikipedia and the development of an ontology-based knowledge base would be built upon each other. Within this knowledge base, a distinction between static and dynamic knowledge can be achieved by marking the information which has to be updated regularly. To avoid the time-consuming annotation of the dynamic information by hand, we strive for an automatic annotation and hence distinction of the two knowledge categories during the construction of the ontology. Thereby we benefit from the auxiliarily available history pages on Wikipedia. Depending on the kind and scope of the changes concerning each article, the corresponding concept of the ontology can be assigned to one of the two knowledge categories.

The distinction of different knowledge categories (i.e. static vs. dynamic) would provide Max with the ability to handle knowledge in a similar way humans handle knowledge. For instance, the update of his

dynamic knowledge might happen once a day at a specific time, like most humans update their knowledge once a day by watching the news or reading the daily newspaper. Furthermore, a spontaneous update can be actuated at any time if necessary. During the conversation with a human user Max might recognize, e.g. as a result of an advice or a hint from his dialog partner, that some of his information is out-dated. In this case Max is able to update this specific information immediately by checking the corresponding article of Wikipedia or another source of information. This idea much resembles how humans inform themselves.

Our approach for the extraction of information required for the ontology-based representation of the Wikipedia knowledge will be based on previous experiences with the online encyclopedia (Mehler 2008). Thereby the ontology will be defined in OWL DL and serialized in RDF/XML to achieve a better machine comprehensibility. A starting point for the technical realization forms an approach realized in a former work (Breuing 2007). To enable the update of the dynamic knowledge, the corresponding concept of the ontology will contain the necessary information, i.e. a specification of the source of information, a description of the method to get up-to-date information, etc. for a particular update process.

4. CONCLUSION

Humans are able to inform themselves about certain topics by reading the according articles in Wikipedia. Due to the dependency of machines on machine-readable representations of concepts and their relations, the connection of an ECA to the online encyclopedia requires the reconstruction of the information held in Wikipedia.

Our paper presents an approach for such a connection to enhance the ECA Max by exploring the German Wikipedia as the agent's primary knowledge resource. For this purpose, we will develop an ontology-based knowledge base for Max adapted from Wikipedia. By considering additional (taxonomical and lexical) information and by distinguishing between static and dynamic knowledge we aim to enable a more human-like knowledge handling by the agent. To realize the distinction of the two knowledge categories, conclusions will be drawn from the information held in the history pages of Wikipedia which store the edits of each article. With these first steps we pursue our long-term objective in enabling Max to be more topic and situation aware and to communicate his ontological knowledge during interactions with human users.

ACKNOWLEDGEMENT

This work is supported by the DFG in the context of the KnowCIT research project in the Center of Excellence Cognitive Interaction Technology.

REFERENCES

Breuing, A., 2007. Eine ontologiebasierte Wissensbasis für den konversationalen Agenten Max mit Anbindung an das Semantic Web. Diploma Thesis.

Buriol, Luciana S. et al. 2006. Temporal Analysis of the Wikigraph. *Proceedings of the Web Intelligence Conference*. Hong Kong.

Hepp, M. et al, 2007. Harvesting Wiki Consensus: Using Wikipedia Entries as Vocabulary for Knowledge Management. *IEEE Computer Society*, Vol. 11, No. 5, pp. 54-65.

Hesse , W., 2002. Ontologie(n). *Informatik Spektrum,* Vol. 16, No. 6, pp. 477-480.

Kittur, A. et al, 2007. Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. *CHI 2007.* San Jose, California, USA.

Kopp, S. et al, 2005. *A Conversational Agent as Museum Guide -- Design and Evaluation of a Real-World Application*. Springer, Berlin.

Krötzsch, M. et al, 2005. Wikipedia and the Semantic Web - The Missing Links. *Proceedings of Wikimania 2005 - The First International Wikimedia Conference*. Frankfurt am Main, Germany.

Mehler, A., 2008. Structural Similarities of Complex Networks: A Computational Model by Example of Wiki Graphs. Applied Artificial Intelligence, Vol. 22

Ponzetto, S.-P. and Strube, M., 2007. Deriving a Large Scale Taxonomy from Wikipedia. *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*. Vancouver, B.C., pp. 1440-1447.

Ruiz-Casado, M. et al, 2006. From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. *ESWC 2006*.

Suchanek, F. M. et al, 2007. Yago: A Core of Semantic Knowledge. *Proceedings of 16th international World Wide Web conference (WWW 2007)*. Banff, Canada.

Viégas, F. B. et al, 2004. Studying Cooperation and Conflict between Authors with *history flow* Visualizations. *CHI 2004*. Vienna, Austria, pp. 575-582.