OntoNL: An Ontology-based Natural Language Interaction Generator for Multimedia Repositories

Anastasia Karanastasi and Stavros Christodoulakis

Lab. Of Distributed Multimedia Information Systems, Technical University of Crete (MUSIC/TUC)
University Campus, Kounoupidiana, Chania, Greece,
{allegra,stavros}@ced.tuc.gr,
WWW home page: http://www.music.tuc.gr

Abstract. We propose a generalized implementation framework that can be used to easily provide natural language interactions for managing multimedia content and user profiles which are described in the information repository with metadata structured according to international standards like MPEG-7 and TV-Anytime (upper ontologies). The description of the content in the repository may also utilize metadata from domain specific ontologies (like soccer ontologies, news ontologies, etc.). Our question-answering interface takes as input queries which are natural language expressions using the upper ontologies, and performs actions or returns answers from the multimedia information repository. The system makes use of the upper and domain specific ontologies to produce the lexicon and grammar rules for parsing user input. The system also uses the user profiles to manage the ambiguities in the natural language input. This avoids lengthy disambiguation interactions which are difficult for some kinds of interaction devices like mobile devices. The architecture proposed is flexible and reusable with multimedia information repositories that may follow different multimedia standards and their content is from different application domains.

1 Introduction

Natural language interfaces to information repositories and digital libraries are known to be very important for many environments and applications which include native user environments, non-motivated users, not very structured and repetitive tasks, and complex structured data sets. In addition the proliferation of small-size mobile devices makes the traditional wide-screen interface styles inappropriate, and motivates further natural language interfaces to digital libraries possibly in combination with speech recognition.

A well known problem with natural language interfaces to information systems is ambiguities. Ambiguities often lead to complex clarification dialogues which are especially difficult to handle in small mobile devices[1].

With the proliferation of the internet great emphasis has been given in the industry to the adoption of standards, since without them companies and organizations can not automate semantic access to information and cooperation. The adoption of standards leads to more structure and semantics in the information stored in the information repository, as well as more semantic meaning included in natural language interactions with the information repository. Thus, in information repositories that use generic standards like MPEG-7 and TV-Anytime, fewer ambiguities arise, making the use of natural language interfaces even more attractive.

When the digital library also manages domain-specific ontologies to achieve higher retrieval performance, the domain-specific ontologies can also be used to further limit the ambiguities and errors in the natural language interactions. Overall then, when the upper ontologies (standards) and the domain-specific ontologies are used by the digital libraries, the natural language interfaces to them become much more attractive, because the ambiguities and the need for clarification dialogues are greatly reduced (although not eliminated) in comparison to natural language interfaces to arbitrary relational databases, for example.

In this paper, we first present a Natural Language Interface to TV-Anytime multimedia information repository that takes advantage both of the TV-Anytime standard[2], as well as the domain-specific ontologies defined to reduce ambiguities in interactions. The system also takes advantage of the user profiles (also expressed with the TV-Anytime standard) to rank the results when ambiguities arise. In this way, we avoid completely the clarification dialogues with the user. This system completely supports TV-Anytime applications and is positively evaluated by users[14].

We then present OntoNL, an Ontology - based Natural Language Interface Generator for multimedia repositories. OntoNL is a generalization of the approach that we used in Natural Language interfaces for the management of the TV-Anytime information to provide, as automatically as possible, natural language interfaces for information repositories that follow arbitrary standards (upper ontologies) and domain-specific ontologies. We show a software architecture that can support complex structured upper and domain-specific ontologies. Our first validation will be based on OWL as an ontology description language, and MPEG-7 as upper ontology.

In section 2 of this paper the multimedia and ontology description standards used by the natural language interface presented here, are described. In section 3, the natural language model for digital TV environment is presented, along with the functionality provided and the representation of the information that the system collects from the user's input. In section 4 we present the algorithm that we use instead of clarification dialogues for resolving ambiguities. In section 5 we present the analysis of the system architecture, of the modules that constitute it and of the implementation framework. In section 6 we present OntoNL, a natural language interface generator that accommodates diverse applications and ontologies for multimedia repositories. Section 7 concludes by summarizing the contents of this paper.

2 Multimedia and Ontology Description Standards

The TV-Anytime[1] metadata model[2] integrates specifications for content metadata used to describe digital TV Programs in terms of various features and specifications for user preferences used to filter program metadata. The "Filtering and Search Preferences" are used to store the preferences of the users in terms of content features so that a digital TV system can filter content and provide personalization services. The FilteringAndSearchPreferences Descriptor Scheme (DS) specifies a user's filtering and/or searching preferences for audio-visual content. These preferences can be specified in terms of creation-, classification- and source-related properties of the content.The FilteringAndSearchPreferences Descriptor Scheme (DS) is a container of CreationPreferences (i.e. Title, Creator), ClassificationPreferences (i.e. Country, Language) and SourcePreferences (i.e. DisseminationSource, DisseminationLocation). The BrowsingPreferences DS is used to specify a user's preferences for navigating and accessing multimedia content and is a container of SummaryPreferences (i.e. SummaryType, SymmaryTheme, SummaryDuration) and PreferenceCondition (i.e. Time, Place).

MPEG-7 defines a standard interface for describing multimedia content structure and features[10]. In MPEG-7, multimedia data are described by using a set of descriptors, structured into Description Schemes (DSs), which are specified by an XML Schema-based[16] Description Definition Language (DDL). The MPEG-7 MDS (Multimedia Description Schemes)[17] provides all the constructs (DSs) needed for defining metadata that describe the multimedia content and the associated multimedia content services.

MPEG-7 allows specifying much richer semantics of content. In the past mechanisms have been developed to transparently incorporate domain ontologies in TV-Anytime and MPEG-7 applications[9]. MPEG-7 allows the integration of ontologies with much richer structure than TV-Anytime which simply supports keywords.

The Web Ontology Language (OWL)[11][12], is the dominant standard in ontology definition. OWL has been developed according to the description logics paradigm, while it uses RDF(S)[18][19] syntax. The basic functionality provided by OWL can be viewed in ref. Since OWL moves to become a dominant ontology definition language, it is important to be able to integrate OWL ontologies with the multimedia standards.

An upper OWL ontology fully capturing the MPEG-7 MDS has been defined in[7] and a methodology for its extension with domain knowledge has been developed in the context of the DS-MIRF framework[**?**]. OWL/RDF metadata for audiovisual content description are produced, which are transformed, using appropriate transformation rules, to MPEG-7 and TV-Anytime compliant metadata, thus providing a basic level of MPEG-7 - TV-Anytime interoperability. Querying is also based on the (upper and domain) ontologies.

[1] The TV-Anytime Forum is an association of organizations which seeks to develop specifications to enable audio-visual and other services based on mass-market high volume digital storage in consumer platforms - simply referred to as local storage.

3 A Natural Language Model for TV-Anytime Digital Libraries

In this section we present the design of the proposed model which fulfills a certain number of requirements that also determine its final functionality. Figure 1 presents the functional requirements that the model should satisfy in the form of use cases.

A collection of sub-phrases was explored and combined to constitute the user's utterance. The categories of these sub-phrases are Introduction phrases, Search phrases, Target phrases, Temporal phrases and Summary phrases. This clustering was used to model the functionality provided by the TV-Anytime specifications.

The design of the system was made taking into account two main goals. The first goal was to determine how the system should behave according to the information gathered by the user's input utterance. Based on this behavior, the second one was, to design and define the roles of each module of the system.

A communication data structure keeps the representation of the information gathered by the user's utterance. The structure consists of three parts namely Element, Element Type and Element Value. The first structure part (Element) is used to differentiate the TV-Anytime metadata information (modelled as TVA-properties element) from the information that directs the system to the correct management of the user's input (modelled as flags element). The TV-Anytime information about date and time is modelled as temporal element. The second structure part (Element Type) is used to further specialize the domain-specific information. The third structure part (Element Value) correspondence the value with the type of the domain-specific information of the second structure part, from the user's utterance. When a user inserts an utterance into the system, it generates a feature structure[3] that is compatible with the structure of the model.

The 'flags' element takes its value from the introduction phrases and the target phrases. The 'TVA-properties' element takes its value from the search phrases and the summary phrases and the 'temporal' element from the temporal phrases. The feature structure can contain one, two or three elements[13][14].

For the retrieval of the personalized content metadata, the management of the personalized content and the creation of the user's profile (see figure 1), the utterance contains in its body one or more search phrases. The system will create a TV-Anytime XML document, compatible with the UserIdentifier and the FilteringAndSearchPreferences Descriptor Schemes of the TV-Anytime metadata specification. For the definition of the user's preferences for the characteristics of an audio-visual content summary, the system constructs a TV-Anytime XML document, compatible with the UserIdentifier and the BrowsingPreferences Descriptor Schemes, with values in the fields of the SummaryPreferences and the PreferenceCondition (for handling the time of the summary delivery).

A user's selection list is a list containing the metadata of TV-programs that the system recommends to the user based on his preferences expressed either at his TV-Anytime profile or directly by him. Every program in this list has

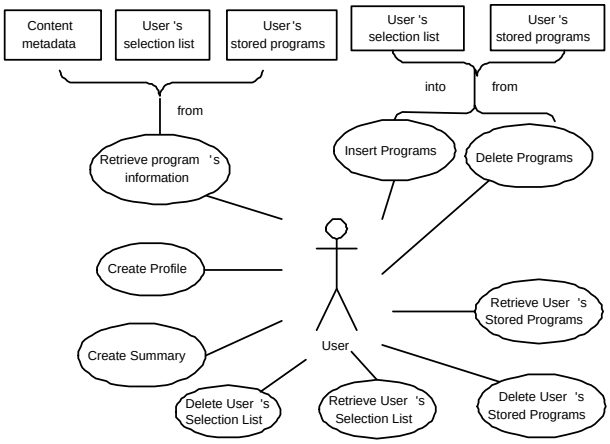

Fig. 1. The use cases that the natural language system was designed to satisfy

a status. The four possible values of this status are: undefined, toBeRecorded, recorded, toBeDeleted. The management of the contents of a user's selection list is presented in[13] .

The user may know neither the contents nor their status of his selection list, so the system undertakes to execute the proper function on behalf of the user.

4 Ambiguity Resolution

If the user does not specify in his utterance the TV-Anytime property he is referring to, the system tries to resolve any ambiguities by following specific steps. First, it collects word by word the sub-phrase with the ambiguities and creates a table of these words. Then the system assembles the remaining words, and checks their semantic values (categories that have this word as value) in its database or other existing ontologies. Then it eliminates the words that match any word in a stop list, which contains words with no semantic values, such as prepositions, pronouns, conjunctions, particles, articles, determiners and so on. However, the system retains the existence of an 'and' or an 'or' for the optimum cross-correlation of the results[15]. After the semantic resolution, it returns a TV-Anytime XML document that is compatible with the FilteringAndSearch-Preferences DS (the specifications of the standard for the structure of the filtering preferences of the user). This Description Scheme of the TV-Anytime is the one that is used for the semantic resolution of the words (TV-Anytime Semantics Resolver in figure 2). Finally, if there is a TV-Anytime user profile the system checks if the user has declared this word as a value in his TV-Anytime User Profile (User Profile Resolver in figure 2). The classification of the results is important in prioritizing the user preferences.

Several examples, that show how the algorithm works, are described in[13][14]. In the general case that there are more than two words with ambiguities the ranking of the results is done so that both the existence of words in the profile is taken into account (with the corresponding weights), and number of the words of the utterance that appear in the TV-Anytime descriptor (but not in the profile). In the case that words are separated with 'and' or 'or' the ranking only considers the number of words that are values of the same TV-Anytime property weighted according to the user profile.

5 Architecture of a Natural Language Interaction system to TV-Anytime Digital Libraries

The complete architecture of the system is presented in figure 2. This architecture follows a multi-tier approach and consists of three tiers. The lowest tier handles the metadata management. The middleware tier includes all the logic for interfacing the system with the outside world. The application tier enables the exchange of information between the server and heterogeneous clients through different communication links.

The user uses a wireless device (mobile phone, PDA) and inserts an utterance that is forwarded to the *ChartParser* module. The *ChartParser* module consists of the JavaChart parser[4] that creates a feature structure with the information from the user's input. There exist two lexicons, the stem lexicon, which contains the stems of the words used in this language and the mini lexicon, which contains the endings of the words. Finally, there is the grammar that follows a unified-based formalism.

The *Dialogue Manager* acts as the core module of the system and is responsible for communicating with all the other system modules. It takes as input a list of feature structures from the chart parser and the user's data from the application that he/she is using. It checks for specific properties for the elimination of the list of the feature structures and, in the case there are any ambiguities, it passes to the *Ambiguities Resolver* module the list of the words with the ambiguities. Finally, it creates a structure with information about the action, the target and the TV-Anytime XML document from the user input.

The *Relational Database* of the system contains the TV-Anytime Metadata information, as well as a number of ontologies (alternative classification schemes to represent genre hierarchies and creator roles as well as more advanced semantic descriptions following the Semantic Part of the MPEG-7 MDS), with information concerning digital TV. The *Relational DBMS* manages the transactions, utilizes a *Java API* (implemented for the extraction of the functionality for filtering, retrieval and summarization) and cooperates with the *XML-DB middleware*. The *XML-DB middleware* is a set of software components responsible for the management of the TV-Anytime XML documents and the correspondence of the TV-Anytime Metadata XML schema with the underlying relational schema.

The *Ambiguities Resolver* module consists of three modules that are responsible for the resolution of different kinds of ambiguities. The Date/Time Resolver

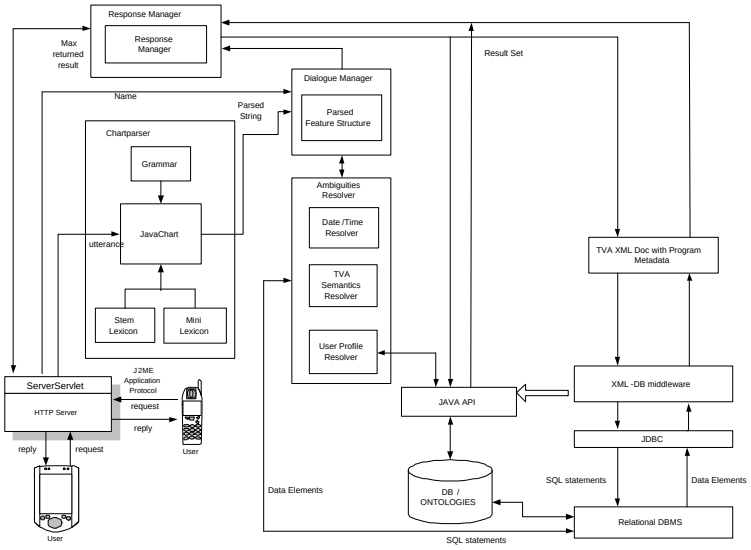

Fig. 2. The Natural Language System Architecture

is the component that converts the temporal phrases in a TV-Anytime compliant form. The TVA Semantics Resolver communicates with the relational DBMS and is responsible to attach TV-Anytime semantics to the words with the ambiguities. The User Profile Resolver filters the list of the words from the user profile (if it exists) and returns a FilteringAndSearchPreferences XML document with values from the corresponding TV-Anytime categories. Finally, it passes this document to the *Response Manager module*.

The *Response Manager* module interacts with the system's database, by providing it the structured information, executes the appropriate functions, retrieves the results and classifies them accordingly. Then, it creates a message and adds it to the result list.

Details concerning the implementation environment can be found in[14]

6 A Generalized Architecture for the Generation of Natural Language Interfaces to Digital Libraries that follow Upper and Domain Specific Ontologies

Given the advantages of the approach we decided to investigate how we could design generic software that will offer the same services with as small as possible adaptations in a variety of application environments. We describe here our current work which aims to generalize the above approach in order to provide a general software infrastructure that will be exploiting any existing application standards (upper ontologies) as well as the domain specific ontologies that

may be available through a natural language interface and uses ranking of results instead of disambiguation to provide application specific interaction and management functionality.

To achieve these objectives we decided to decouple as much as possible the Natural Language (NL) components from the application standards (ontologies) but at the same time allow the NL subsystem to use the different ontologies for the different application classes. For example we would like to be able to use essentially the same NL software for managing TV information following an updated (new, extended) TV-Anytime standard, or multimedia information based on the MPEG-7 standard, or even a completely different application class like tourism information and services based on the relevant tourism standards. The communication of the NL subsystem with the databases, knowledge bases, service providers, user profile agents will be based on those upper and domain specific ontologies.

Many utterances in the natural language specify a set of attribute/value pairs as well as free-standing values, where the attribute is implicit. To extract the useful information of the user utterance we propose a new categorization of the sub-phrases that constitute it. The categories of these sub-phrases are *General phrases*, *Application Specific Management phrases* and *Specific Ontology phrases*. The *General phrases* contain information that shows no specific interest for the functions of the system (e.g. 'I would like', 'I want'). The *Application Specific Management phrases* contain information for the action and the target of this action to be made from the system. The *Specific Ontology phrases* describe the information of the ontologies used by the system. This separation allows easier integration of new ontologies in the grammar.

The structure that represents the information gathered by the user's utterance remains the same, with structure parts the *Element*, the *Element Type* and the *Element Value*. The first structure part (Element) is used to differentiate the metadata information from the *Specific Ontology phrases* (modelled as *properties* element) from the information that directs the system to the correct management of the user's input from the *Application Specific Management phrases* (modelled as *flags* element). The information about date and time is modelled as *temporal* element. The second structure part (Element Type) is used for the further specialization of the aforementioned information and to obtain its corresponding Element Value (the third structure part), from the user's utterance. When a user inserts an utterance into the system, it generates a feature structure[3] that follows the structure of the model.

In the *ChartParser* module the lexicon is now updated in an automatic way from the ontologies for the specific application and domain of use, by a Syntactic Analyzer, that inserts the words (individuals) in the lexicon with the semantic value (classes) that they inherit, from the ontology used by the system. This way the system becomes flexible and reusable, in different application classes that use different ontologies. The list of feature structures that the Dialogue Manager takes as an input contains the semantic of the words from the user's input. In this way the semantic resolution becomes responsibility of the *ChartParser*,

because of the existence of a domain specific lexicon. The grammar is also been constructed in an automatic way from the ontologies for the specific application and domain of use, by a Semantic Analyzer, that creates the verb, noun and prepositional phrases, with the relations in the onto-logical structure.

A new module, the *Service Interface Manager* is responsible for the communication between the natural language interface and the application system. It contains the representation of the upper ontology in object hierarchies and formats the information from the user's utterance in a structure understood by the database. This structure contains the action, the target for the action to be made and an OWL/RDF document with the information from the user's utterance. It is also responsible to reconstruct the information retrieved from the database system and pass this information to the *Response Manager*.

The upper and domain-specific ontologies can have a hierarchical representation of their terms and relations. From the context (location) of the words in the users utterance and the type of the application the natural language interface is referring to, we can extract information that can help on the better ranking of the results. This is a responsibility of the *Semantic Resolver* module that is been added in the *Ambiguities Resolver* module.

A detailed design and implementation of this generalized architecture is currently pursued in MUSIC/TUC, in order to be used in different application environments. We are currently focusing in the satisfaction of MPEG-7 applications transparently utilizing domain specific ontologies. This way we will produce an environment where natural language interactions are used to manage both TV-Anytime and MPEG-7 applications. In a second step we will be evaluating if this architecture can be used in environments with arbitrary upper and domain specific ontologies with as little adaptation as possible.

7 Summary - Conclusions

In this paper we described the design and implementation of a natural language interface system, to design libraries that manage multimedia documents following the TV-Anytime standard. We argued that application environment that support standards and domain specific ontologies reduce ambiguities and presented our approach that uses user profiles (also supported by the standards) to rank the potential answers, thus avoiding completely the clarification dialogues.

We believe that this approach is useful and generally applicable in digital library environments that support different standards (upper ontologies) and different domain-specific ontologies. We, therefore, proceeded in the design of a system that generates natural language interfaces to digital library applications that utilize various standards and ontologies. Our implementation currently focuses into supporting multimedia applications that follow MPEG-7 or TV-Anytime and OWL ontologies. The design however owes to be generic so that it can accommodate other standards and ontologies in the future.

References

1. Maier, E.: Clarification dialogues in VERBMOBIL. In Proceedings of EuroSpeech97 (1997) 1891-1894, Rhodes
2. The site of the TV-Anytime Forum, http://www.tv-anytime.org
3. Jurafsky, D., Martin, J.H.: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, (2000)
4. Degerstedt, L. JavaChart User Manual (2002), http://nlpfarm.sourceforge.net/javachart/
5. MySQL: The World's Most Popular Open Source Database, http://www.mysql.com/
6. Java Technology - J2ME, http://java.sun.com/ , http://java.sun.com/j2me/
7. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing, in Proceeding of CAiSE (2004)
8. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support for Ontology-based Video Retrieval Applications, in Proceedings of CIVR (2004)
9. Tsinaraki, C., Fatourou, E., Christodoulakis, S.: An Ontology-Driven Framework for the Management of Semantic Metadata describing Audiovisual Information. In Proceedings of CAiSE (2003) 340-356
10. Salembier, P.: MPEG-7 Multimedia Description Schemes, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6 (2001)
11. Mc Guiness, D., Van Harmelen, F.: OWL Web Ontology Language Overview, W3C Can-didate Recommendation (2003) http://www.w3.org/TR/owl-features/
12. OWL Web Ontology Language, http://www.w3.org/TR/2004/REC-owl-features-20040210/
13. Karanastasi, A., Kazasis, F., Christodoulakis, S.: A Natural Language Model and a System for Managing TV-Anytime Information in Mobile Environments, in Proceedings of UMICS (2004)
14. Karanastasi, A., Kazasis, F., Christodoulakis, S.: A Natural Language Model for Managing TV-Anytime Information from Mobile Devices, in Proceedings of NLDB (2004)
15. Karanastasi, A., Kazasis, F., Christodoulakis, S.: A Natural Language Model and a System for Managing TV-Anytime Information in Mobile Environments, ACM/Verlag Personal and Ubiquitous Computing, Vol. 4 (2005)
16. Fallside, D.: XML Schema Part 0: Primer,W3C Recommendation (2001) http://www.w3.org/TR/xmlschema-0/
17. Chang, S.F., Sikora, T., Puri, A.: Overview of the MPEG-7 standard, IEEE Transactions on Circuits and Systems for Video Technology 11, pp. 688-695 (2001)
18. Manola, F., Milles, E. (eds.): RDF Primer. W3C Recommendation, (2004) http://www.w3.org/TR/rdf-primer
19. Brickley, D., Guha R. V.(eds.): RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation (2004) http://www.w3.org/TR/rdf-schema