Adaptive Natural Language Interaction

Stasinos Konstantopoulos
Athanasios Tegos
Dimitris Bilidas
NCSR 'Demokritos', Athens, Greece

Ion Androutsopoulos
Gerasimos Lampouras
Prodromos Malakasiotis
Athens Univ. of Economics and Business
Greece

Colin Matheson
Human Communication Research Centre
Edinburgh University, U.K.

Olivier Deroo
Acapela Group, Belgium

Abstract

The subject of this demonstration is natural language interaction, focusing on adaptivity and profiling of the dialogue management and the generated output (text and speech). These are demonstrated in a museum guide use-case, operating in a simulated environment. The main technical innovations presented are the profiling model, the dialogue and action management system, and the text generation and speech synthesis systems.

1 Introduction

In this demonstration we present a number of state-of-the art language technology tools, implementing and integrating the latest discourse and knowledge representation theories into a complete application suite, including:

- dialogue management, natural language generation, and speech synthesis, all modulated by a flexible and highly adaptable profiling mechanism;
- robust speech recognition and language interpretation; and,
- an authoring environment for developing the representation of the domain of discourse as well as the associated linguistic and adaptivity resources.

The system demonstration is based on a use case of a virtual-tour guide in a museum domain. Demonstration visitors interact with the guide using headsets and are able to experiment with loading different interaction profiles and observing the differences in the guide's behaviour. The demonstration also includes the screening of videos from an embodied instantiation of the system as a robot guiding visitors in a museum.

2 Technical Content

The demonstration integrates a number of state-of-the-art language components into a highly adaptive natural language interaction system. Adaptivity here refers to using *interaction profiles* that modulate dialogue management as well as text generation and speech synthesis. Interaction profiles are semantic models that extend the objective ontological model of the domain of discourse with subjective information, such as how 'interesting' or 'important' an entity or statement of the objective domain model is.

Advanced *multimodal dialogue management* capabilities involving and combining input and output from various interaction modalities and technologies, such as speech recognition and synthesis, natural language interpretation and generation, and recognition of/response to user actions, gestures, and facial expressions.

State-of-the art *natural language generation* technology, capable of producing multi-sentence, coherent natural language descriptions of objects based on their abstract semantic representation. The resulting descriptions vary dynamically in terms of content as well as surface language expressions used to realize each description, depending on the interaction history (e.g., comparing to previously given information) and the adaptivity parameters (exhibiting system personality and adapting to user background and interests).

3 System Description

The system is capable of interacting in a variety of modalities, including non-verbal ones such as gesture and face-expression recognition, but in this demonstration we focus on the system's language interaction components. In this modality, abstract, language-independent system actions are first planned by the *dialogue and action manager* (DAM), then realized into language-specific text

by the *natural language generation engine*, and finally *synthesized* into speech. All three layers are parametrized by a *profiling and adaptivity* module.

3.1 Profiling and Adaptation

Profiling and adaptation modulates the output of dialogue management, generation, and speech synthesis so that the system exhibits a synthetic personality, while at the same time adapting to user background and interests.

User stereotypes (e.g., 'expert' or 'child') provide generation parameters (such as maximum description length) and also initialize the dynamic user model with *interest rates* for all the ontological entities (individuals and properties) of the domain of discourse. This same information is also provided in *system profiles* reflecting the system's (as opposed to the users') preferences; one can, for example, define a profile that favours using the architectural attributes to describe a building where another profile would choose to concentrate on historical facts regarding the same building.

Stereotypes and profiles are combined into a single set of parameters by means of *personality models*. Personality models are many-valued Description Logic definitions of the overall preference, grounded in stereotype and profile data. These definitions model recognizable personality traits so that, for example, an open personality will attend more to the user's requests than its own interests in deriving overall preference (Konstantopoulos et al., 2008).

Furthermore, the system *dynamically* adapts overall preference according to both interaction history and the current dialogue state. So, for one, the initial (static model) interest factor of an ontology entity is reduced each time this entity is used in a description in order to avoid repetitions. On the other hand, preference will increase if, for example, in the current state the user has explicitly asked about an entity.

3.2 Dialogue and Action Management

The DAM is built around the information-state update dialogue paradigm of the TRINDIKIT dialogue-engine toolkit (Cooper and Larsson, 1998) and takes into account the combined user-robot interest factor when determining information state updates.

The DAM combines various interaction modalities and technologies in both interpretation/fusion and generation/fission. In interpreting user actions the system recognizes spoken utterances, simple gestures, and touch-screen input, all of which may be combined into a representation of a multi-modal user action. Similarly, when planning robotic actions the DAM coordinates a number of available output modalities, including spoken language, text (on the touchscreen), the movement and configuration of the robotic platform, facial expressions, and simple head gestures.[1]

To handle multimodal input, the DAM uses a fusion module which combines messages from the language interpretation, gesture, and touchscreen modules into a single XML structure. Schematically, this can be represented as:

```
<userAction>
  <userUtterance>hello</userUtterance>
  <userButton content="13"/>
</userAction>
```

This structure represents a user pressing something on the touchscreen and saying *hello* at the same time.[2]

The representation is passed essentially unchanged to the DAM, to be processed by its update rules, where the ID of button press is interpreted in context and matched with the speech. In most circumstances, the natural language processing component (see 3.3) produces a semantic representation of the input which appears in the `userUtterance` element; the use of 'hello' above is for illustration. An example update rule which will fire in the context of a greeting from the user is (in schematic form):

```
if
 in(/latest_utterance/moves,  hello)
then
 output(start)
```

Update rules contain a list of conditions and a list of effects. Here there is one condition (that the latest moves from the user includes 'hello'), and one effect (the 'start' procedure). The latter initiates the dialogue by, among other things, having the system utter a standardised greeting.

As noted above, the DAM is also multimodal on the output side. An XML representation is created which can contain robot utterances and robot movements (both head movements and mobile platform moves). Information can also be presented on the touchscreen.

[1] Expressions and gestures will not be demonstrated, as they can not be materialized in the simulated robot.

[2] The precise meaning of 'at the same time' is determined by the fusion module.

3.3 Natural Language Processing

The NATURALOWL natural language generation (NLG) engine (Galanis et al, 2009) produces multi-sentence, coherent natural language descriptions of objects in multiple languages from a single semantic representation; the resulting descriptions are annotated with prosodic markup for driving the speech synthesisers.

The generated descriptions vary dynamically, in both content and language expressions, depending on the interaction profile as well as the dynamic interaction history. The dynamic preference factor of the item itself is used to decide the level of detail of the description being generated. The preference factors of the properties are used to order the contents of the descriptions to ensure that, in cases where not all possible facts are to be presented in a single turn, the most relevant ones are chosen. The interaction history is used to check previously given information to avoid repeating the same information in different contexts and to create comparisons with earlier objects.

NaturalOWL demonstrates the benefits of adopting NLG on the Semantic Web. Organizations that need to publish information about objects, such as exhibits or products, can publish OWL ontologies instead of texts. NLG engines, embedded in browsers or Web servers, can then render the ontologies in natural language, whereas computer programs may access the ontologies, in effect logical statements, directly. The descriptions can be very simple and brief, relying on question answering to provide more information if such is requested. This way, machine-readable information can be more naturally inspected and consulted by users.

In order to generate a list of possible follow up questions that the system can handle, we initially construct a list of the particular individuals or classes that are mentioned in the generated description; the follow up questions will most likely refer to them. Only individuals and classes for which there is further information in the ontology are extracted.

After identifying the referred individuals and classes, we proceed to predict definition (e.g., 'Who was Ares?') and property questions (e.g., 'Where is Mount Penteli?') about them that could be answered by the information in the ontology. We avoid generating questions that cannot be answered. The expected definition questions are constructed by inserting the names of the referred individuals and classes into templates such as 'who is/was *person X*?' or 'what do you know about *class or entity Y*?'.

In the case of referred individuals, we also generate expected property questions using the patterns NaturalOWL generates the descriptions with. These patterns, called *microplans*, show how to express the properties of the ontology as sentences of the target languages. For example, if the individual templeOfAres has the property excavatedIn, and that property has a microplan of the form '*resource* was excavated in *period*', we anticipate questions such as 'when was the Temple of Ares excavated?' and 'which period was the Temple of Ares excavated in?'.

Whenever a description (e.g., of a monument) is generated, the expected follow up questions for that description (e.g., about the monument's architect) are dynamically included in the rules of the speech recognizer's grammar, to increase word recognition accuracy. The rules include components that extract entities, classes, and properties from the recognized questions, thus allowing the dialogue and action manager to figure out what the user wishes to know.

3.4 Speech Synthesis and Recognition

The natural language interface demonstrates robust speech recognition technology, capable of recognizing spoken phrases in noisy environments, and advanced speech synthesis, capable of producing spoken output of very high quality. The main challenge that the *automatic speech recognition* (ASR) module needs to address is background noise, especially in the robot-embodied use case. A common technique used in order to handle this is training acoustic models with the anticipated background noise, but that is not always possible. The demonstrated ASR module can be trained on noise-contaminated data where available, but also incorporates *multi-band acoustic modelling* (Dupont, 2003) for robust recognition under noisy conditions. Speech recognition rates are also substantially improved by using the predictions made by NATURALOWL and the DAM to dynamically restrict the lexical and phrasal expectations at each dialogue turn.

The *speech synthesis* module of the demonstrated system is based on *unit selection technology*, generally recognized as producing more nat-

ural output that previous technologies such as diphone concatenation or formant synthesis. The main innovation that is demonstrated is support for emotion, a key aspect of increasing the naturalness of synthetic speech. This is achieved by combining emotional unit recordings with run-time transformations. With respect to the former, a complete 'voice' now comprises three sub-voices (*neutral*, *happy*, and *sad*), based on recordings of the same speaker. The recording time needed is substantially decreased by prior linguistic analysis that selects appropriate text covering all phonetic units needed by the unit selection system. In addition to the statically defined sub-voices, the speech synthesis module implements dynamic transformations (e.g., emphasis), pauses, and variable speech speed. The system combines all these capabilities in order to dynamically modulate the synthesised speech to convey the impression of emotionally modulated speech.

3.5 Authoring

The interaction system is complemented by ELEON (Bilidas et al., 2007), an authoring tool for annotating domain ontologies with the generation and adaptivity resources described above. The domain ontology can be authored in ELEON, but any existing OWL ontology can also annotated.

More specifically, ELEON supports authoring *linguistic resources*, including a domain-dependent *lexicon*, which associates classes and individuals of the ontology with nouns and proper names of the target natural languages; *microplans*, which provide the NLG with patterns for realizing property instances as sentences; and a partial ordering of properties, which allows the system to order the resulting sentences as a coherent text.

The *adaptivity* and profiling resources include *interest rates*, indicating how interesting the entities of the ontology are in any given profile; and *stereotype parameters* that control generation aspects such as the number of facts to include in a description or the maximum sentence length.

Furthermore, ELEON supports the author with immediate previews, so that the effect of any change in either the ontology or the associated resources can be directly reviewed. The actual generation of the preview is relegated to external generation engines.

4 Conclusions

The demonstrated system combines semantic representation and reasoning technologies with language technology into a human-computer interaction system that exhibits a large degree of adaptability to audiences and circumstances and is able to take advantage of existing domain model created independently of the need to build a natural language interface. Furthermore by clearly separating the abstract, semantic layer from that of the linguistic realization, it allows the re-use of linguistic resources across domains and the domain model and adaptivity resources across languages.

Acknowledgements

The demonstrated system is being developed by the European (FP6-IST) project INDIGO.[3] INDIGO develops and advances human-robot interaction technology, enabling robots to perceive natural human behaviour, as well as making them act in ways that are more familiar to humans. To achieve its goals, INDIGO advances various technologies, which it integrates in a robotic platform.

References

Dimitris Bilidas, Maria Theologou, and Vangelis Karkaletsis. 2007. Enriching OWL ontologies with linguistic and user-related annotations: the ELEON system. In *Proc. 19th Intl. Conf. on Tools with Artificial Intelligence* (ICTAI-2007).

Robin Cooper and Staffan Larsson. 1998. Dialogue Moves and Information States. In: *Proceedings of the 3rd Intl. Workshop on Computational Semantics (IWCS-3)*.

Stéphane Dupont. 2003. Robust parameters for noisy speech recognition. U.S. Patent 2003182114.

Dimitrios Galanis, George Karakatsiotis, Gerasimos Lampouras and Ion Androutsopoulos. 2009. An open-source natural language generator for OWL ontologies and its use in Protégé and Second Life. In this volume.

Stasinos Konstantopoulos, Vangelis Karkaletsis, and Colin Matheson. 2008. Robot personality: Representation and externalization. In *Proc. Computational Aspects of Affective and Emotional Interaction* (CAFFEi 08), Patras, Greece.

[3] http://www.ics.forth.gr/indigo/