Ontology-Based Multimodal High Level Fusion Involving Natural Language Analysis for Aged People Home Care Application

Olga Vybornova, Monica Gemo, Ronald Moncarey, Benoit Macq

Communications and Remote Sensing Lab, Universite Catholique de Louvain, Belgium

{vybornova, gemo, moncarey, macq}@tele.ucl.ac.be

Abstract

This paper presents a knowledge-based method of early-stage high level multimodal fusion of data obtained from speech input and visual scene. The ultimate goal is to develop a human-computer multimodal interface to assist elderly people living alone at home to perform their daily activities, and to support their active ageing and social cohesion. Crucial for multimodal high level fusion and successful communication is the provision of extensive semantics and contextual information from spoken language understanding. To address this we propose to extract natural language semantic representations and map them onto the restricted domain ontology. This information is then processed for multimodal reference resolution together with visual scene input. To make our approach flexible and widely applicable, a priori situational knowledge, modalities and the fusion process are modelled in the ontology expressing the domain constraints. Here, we illustrate ontology-based multimodal fusion on an example scenario combining speech and visual scene analysis.

Index Terms: elderly people language understanding, semantic representations, unrestricted language processing, ontology matching, multimodal fusion.

1. Introduction

Humans have a remarkable ability to exhibit and comprehend effortlessly the fully coordinated mind-and-body behaviour. Thus in order to provide natural interaction with the user(s) a system must be able to handle semantic-level input data fusion, i.e. to combine information arriving simultaneously from different modalities into one or several unified and coherent representations of the user's intention. A challenge here is to decide what information should be fused and what should not. When intending to perform an action, the user might:

- speak about it describing his/her actions or intentions (the ideal case, then the linguistic and action recognition data are complementary)
- be doing things, but speaking about something absolutely irrelevant to those actions (in this case the data from the two modalities should be analysed separately, without merging)
- the user's words might contradict the actions performed (then the actions have priority, since "actions speak louder than words")
- be just silent when performing actions (in this case the unimodal operation mode will be needed)

Everything what is said or done is meaningful only in the particular context. To accomplish the task of semantic fusion we should take into account the information obtained at least in the following three types of context [4]:

- domain context (meaning prior knowledge of the domain, semantic frames with predefined action patterns, user profiles, situation modelling, a priori developed and dynamically updated ontology defining subjects, objects, activities and relations between them for a particular person;
- conversation context (derived from natural language semantic analysis);
- visual context (capturing the user's gesture/action in the observation scene and allowing eye gaze tracking to enable salience models).

Basically, in our approach the high level fusion of the input streams can be performed in two stages: (i) early fusion that merges information already at the signal or recognition stage to provide reliable reference resolution and (ii) late fusion that integrates all the information at the final stage to give interpretation of the user's behaviour [12]. We argue that in order to make the multimodal semantic integration more efficient and practically real, special attention should be paid to the stages preceding the final fusion stage.

Following [4], we find the salience driven approach to multimodal input interpretation very promising to address the issues of elderly people's speech uptake. The general idea of the salience is that for each application domain, there is a physical world representation that captures domain knowledge (described as a set of frames within ontology). The user's language moves, elementary gestures and complex actions involving certain objects (discourse referents) mapped to the domain ontology activate the salient part of representation that includes relevant properties or tasks related to the salient objects. This salient part of the physical world is likely to be the potential content of the spoken communication, and thus can be used to tailor language models for spoken language understanding. It bridges gesture understanding and language understanding at a stage before final multimodal fusion.

The next section outlines the relevant multimodal fusion systems. Section 3 illustrates the proposed architecture for early-stage high level fusion of speech and gesture input in the scenario at hand. Sections 4 and 5 then present the ontology and the principles driving multimodal reference resolution in our approach. Finally are concluding remarks.

2. Related work

Many different systems for multimodal fusion are known in the literature and they all include some fusion mechanism to robustly integrate the separate data streams by reducing uncertainty [4, 7, 10]. However, as Oviatt et al claim in [9], the need to incorporate more natural interaction patterns in the multimodal fusion engine has emerged very recently and current systems are only starting to take into account contextual information. The major examples concern

multimodal dialogue systems. In combining speech and pen input [10], Pfleger first fuses multimodal input events with respect to their local context (i.e. previously recognized input events and expectations anticipated by the dialogue state). Subsequently [11], Pfleger et al. extend the context model to include implicitly mentioned concepts as candidates for reference resolution. In [7], Holzapfel et al. exploit contextual correlation between 3D gestures and referring words in speech to improve efficiency against falsely detected gestures. Finally, Chai uses a wider variety of contextual information including domain and conversation contexts in the MIND multimodal fusion system [4]. Our work draws on Chai's results and applies contextual information to enhance multimodal interpretation also in the stages preceding the final fusion stage.

3. Application and modalities of interest

We are developing a multimodal interface aimed at assisting elderly people living independently in their household environment. It should be noted here that apart from deliberate speaking with the system, we can anticipate other kinds of speech input from the users, since familiar surroundings and lonely life encourage elderly people to talk to themselves and to accompany their actions with corresponding utterances. Thus our application accepts spontaneous multimodal input – English speech, 3D gestures (pointing, iconic, possibly metaphoric) and user's physical action. In the near future we plan to treat eye gaze tracking modality to facilitate capturing salient objects in the scene.

Here we are facing a problem of elderly people's speech recognition and overall language processing. Experimental research [3] shows that the ability to produce spoken forms of words declines with aging. Older adults have difficulties in retrieving (familiar) words when speaking - this is often associated with age-related memory deficiency. Another problem concerns tip-of-the-tongue (TOT) states – this is a word finding failure when a person produces one or more incorrect sounds in a word, for example, uttering 'coffee cot' instead of 'coffee pot', or if a person is looking for an appropriate word, and (s)he is sure that this word is familiar, it is often the case that only one sound or syllable, most frequently the first one, is remembered and then a word resembling the correct one is uttered, for instance, a TOT word 'exotic' instead of the desired 'eccentric'. However, whereas the language production declines with aging, semantic processes are well maintained, and this fact serves a good basis for making final correct interpretation provided that speech capture is corrected with the help of our early fusion method where information from other modalities facilitates correction of speech recognition hypotheses.

We illustrate our approach on a short example of a scene where the person is moving towards the phone saying "*What is Brian's number?*" (cfr. Figure 1) Thus the system has to understand the spoken utterance meaning – extract the proper name '*Brian'*, the noun denoting the desired info (in this case) – '*number'*, identify the objects involved in the scene, recognise the action pattern – the person is moving towards the phone and finally recognise the ultimate plan – the person is going to make a phone call.

The system design is organised in a 2-stage architecture for ontology-based high level fusion (cfr. Figure 2). Such a design allows multiple interwoven relations between the acquired modalities and contextual information as each component can easily access available a priori knowledge as well as intermediate interpretations from the ontology base.

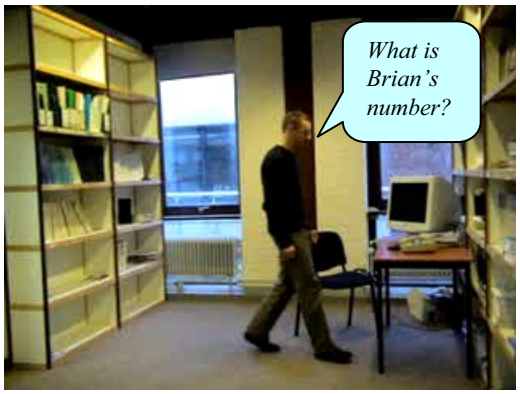

Figure 1: *An example of multimodal behaviour of the user intending to make a phone call.*

In the following paragraphs we detail the main components involved in the pipeline processing.

3.1. Spoken language analysis

Spoken input is analysed by two components, namely the speech recognition tool (ASR) and the natural language understanding module (NLU).

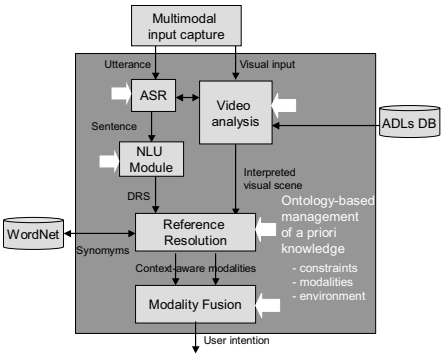

Figure 2: *Ontology-based high level fusion architecture.*

3.1.1. Speech recognition

For the task of speech recognition we use MROSDK_V0202 toolkit provided by Multitel research centre (http://www.multitel.be). General-purpose acoustic model and perfect baseline pronunciation rules (database dependent) for English enable direct speech recognition of any user without preliminary training. Speech is acquired in the voice active detection mode, voice capturing is always active and the user does not need to perform any additional interaction apart from pronouncing the words. Text prediction capabilities can be triggered by information from visual analysis. The basic configuration proved to be adequate and effective for the recognition of utterances in our example scenario.

3.1.2. Syntax and semantics

The next stage after speech recognition is syntactic and semantic analysis of the discourse. For our purposes we use the CCG (Combinatory Categorial Grammar) parser, Release 0.96, developed by S. Clark and J. Curran [5]

The grammar used by the parser is taken from CCGbank developed by J. Hockenmaier and M. Steedman [2]. CCGbank is a treebank containing phrase-structure trees in the Penn Treebank (WSJ texts) converted into CCG derivations. It allows easy recovery of long-range dependencies, provides a transparent interface between surface syntax and underlying semantic representation, including predicate-argument structure. The grammar is based

on 'real' texts, and that is why it has wide-coverage, thus making parsing efficient and robust.

The CCG parser has Boxer [1], as add-on to generate semantic representations - Discourse Representation Structures (DRSs), the box representations of Discourse Representation Theory (DRT) [8]. DRSs consist of a set of discourse referents (representatives of objects introduced in the discourse) and a set of conditions for these referents (properties of the objects).

Our initial experiments with the CCG parser together with Boxer showed that it suits well for our application in parsing speech of elderly people which is on one hand somewhat restricted to their environment, background and social relationships (the domain model and user profile help us cope with the challenge), but on the other hand is naturally broad enough to need wide-coverage tools for processing. DRSs can be generated in different output formats in Prolog or XML. To link the DRSs output with the ontology we use the XML format.

3.2. Visual scene analysis for ontology input and to provide activity pattern recognition

The visual scene analysis and human action recognition are based on predefined information about user's activities of daily living (ADLs) and activity patterns collected for recognition. A database of inside home ADLs with detailed specification of logical events (sequences of sub-activities annotated with timing and object information) of daily activities is being created to constitute the a priori knowledge of the system. The ADLs database includes patterns of various complexity, from locomotion identification (whether the person is moving or sitting still, e.g., walking, wheelchair use, etc.,) to more complex scenarios like making phone calls, preparing meals, etc. In our application we are mostly interested in extracting the ADLs for independent living and those crucial for emergency or potential risk situations detection (forgotten or incomplete actions, abnormal deviation or change of body posture, location, mismatch in action timing, etc.).

The action detection problem is phrased as inference on the underlying Dynamic Bayesian Networks representing the process of executing the actor's plan. A combination of statistical methods (online probabilistic plan recognition using the MultiAgent Abstract Hidden Markov mEmory Model (M-AHMEM) providing hierarchical action decomposition, memory for each action and multiagent support, Rao-Blackwellized Particle Filters (RBPF) allowing to reduce network complexity), are used to decide for most likely activities and detect variations of common activities [6, 12].

4. Ontology design and structure

All information involved in the high level fusion process is represented in a comprehensive ontology modelling patterns of multimodal communication, actions and a priori background knowledge about the user. We use the open-source Protégé environment (protege.stanford.edu) to manage the application metadata in frames and properties in slots. Figure 3 illustrates the developed ontology. To facilitate the best understanding of the interaction situation with the user, the ontology is organised in separate sub-domains modelling the interaction modalities, user's home environment (rooms and objects), social network (relatives, friends, doctors, nurses etc.).

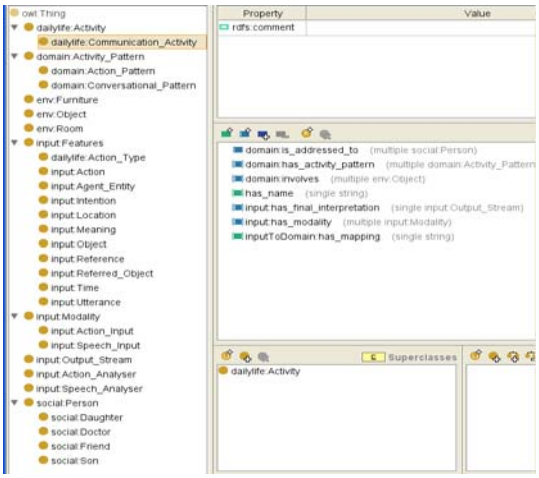

Figure 3: *Domain ontology fragment.*

ADLs of interest are defined as generic actions that the user performs himself and communicative actions involving other people. Activities are also related to the set of modalities representing the type of commands according to which they can be executed (gesture input, speech input, etc). Thanks to the programmatic API of Protégé, the ontology and the user profile can be dynamically updated with information from the linguistic input and visual scene observation. For example, the system can dynamically note the different nicknames that the user pronounces when referring to his relatives. This information is then used to perform multimodal reference resolution.

Among the slots of the activity frame there is a reference to predefined action sequences stored in an external database. This contains the user's actual or possible actions. The database further records the interpreted multimodal behaviour for validation and tuning of the system. Such collected data are analysed and used to set system parameters intervening in the early and late fusion stages (e.g. the range of subsequent captured inputs for speech and video frames requiring synchronisation). Another fundamental activity slot is a link to procedural scripts allowing reference resolution and mappings between input (e.g. DRS entities) and domain frames. Details about mappings are presented in the next section.

5. Multimodal reference resolution as early-stage fusion

Having spoken and visual input analysed and parsed, this information can be jointly processed for multimodal reference resolution (cfr. Figure 2). This component first tries to combine contextual information acquired synchronously (i.e. the conversation stream captured concurrently with gesture input) to predict the user's intention and identify possible activity frames from the ontology. For example, if the interpreted visual scene shows that the user is approaching the telephone, then the system selects '*make_a phone_call*' as candidate activity of interest. From the activity frame in the ontology, the system then learns all necessary details: it is a communicative activity involving another person (i.e. the *callee* needing to be resolved and with which *telephone number*) and the use of the object *telephone*. The activity frame also includes a slot for the procedural steps (i.e. the input to domain mapping script) that resolve missing information from contextual information of the modalities.

```
String [] word = sentence.split(" ");
OWLModel owlModel = ProtegeOWL.createJenaOWLModelFromURI(uri);
String query_str = "SELECT DISTINCT * WHERE {";

 for (int i = 0 ; i < word.length ; i++) {

    if (xdrsparse.getPostag(word[i]).equals("NN"))
    {
          query_str     += "?instance :has_nickname \"" + word[i] + "\"
     }
 }
 query_str += "}";

 results = owlModel.executeSPARQLQuery(query_str);
```

Figure 4: *Mapping script querying the ontology for nicknames of a Person frame*.

The mapping script in Figure 4 shows an excerpt of the *make_a_phone_call* activity that performs resolution of the *callee* by querying the ontology base for possible nicknames told by the user (e.g. my *son*, my *dear,* etc.). This basically consists of instructions to read DRS items encoded in XML and then to perform queries to access the ontology base. In this processing the system also checks DRS values against a list of *keywords, terms* and *verbs* specific to the activity (also stored in the ontology base). Based on the result from such queries and checks, decisions are taken to judge if all references are resolved or in case of ambiguities (e.g. more than one persons with the same name), the resolution processing continues. In this case, it widens the range of contextual information to other acquisitions preceding the one under resolution and processes them in a similar way. Another important verification concerns contradictory or unrelated contextual information. To discriminate such situations, the mapping script checks for matching similarity or better dissimilarity between contextual multimodal information of the input streams. In our initial development, this verification is performed with the help of WordNet lexical database (http://wordnet.princeton.edu/). Captured DRS values and *keywords* fetched from the activity frame are entered in the thesaurus. The thesaurus then provides the most significant synonym associated to the entered term, if the results of the two sets do not present enough communality or are disjoint, then we can conclude they are most probably unrelated. In this case the system processes only the physical context and not the conversational context.

The same idea can be expressed in natural language in many ways using various lexical items and syntactical constructions. In our example with the user willing to call his son we tested the meaning of utterances in 4 different ways. The person said: (a) What is Brian's number? (b) What is my son's number? (c) I want to call my son, (d) *Brian*. The (d) variant was tested because people often do not utter complete sentences when doing something, but fragments, or just single words, thinking, hesitating. Here we are not interested in every single word, but in correct capturing of the meaningful words denoting people, reference objects and their quality attributes as well as location and temporal information. We discovered that even one key word in this case – proper name '*Brian'* is captured by the NLU module and thus is correctly merged with ontology evoking the frame about the user's son *Brian*. In this and other three synonymous variants the fusion process gave correct result.

6. Conclusions

We have presented a method of contextual knowledge-based approach to multimodal high level fusion and illustrated it on integration of data from spoken input and visual scene analysis. We described the role of ontology in our approach – a knowledge base that can be dynamically updated during the interaction developed for the application of assisting elderly people living alone in their homes. Thus we have a restricted domain to work with, but we deal with unrestricted natural human behaviour – spontaneous spoken input and gesture. At present we are busy with implementation of multi-stage crossmodal fusion that is seen promising from the point of view of reference ambiguity resolution before the final fusion. It is exactly crossmodal fusion that helps us cope with problems of speech recognition for elderly people (caused by age-related decline of language production ability) because information from other modalities refines the language analysis at the early stage of recognition. This gives us an efficient reference resolution mechanism that improves further processing. Language semantic representations (DRSs) generated from the NLU module provide an excellent input to the ontology. This handles contextual information within the domain and also serves as a metamodel for Bayesian networks used to analyse and combine the modalities of interest. With the help of Bayesian networks we obtain robust contextual fusion thanks to non-deterministic weighting of modalities.

7. Acknowledgements

This work has been supported by the European FP6 NoE SIMILAR (http://www.similar.cc). We are very grateful to Johan Bos and Stephen Clark for providing us the tools for primary tests. Special thanks to Multitel research centre.

8. References

[1] Bos J. "Towards wide-coverage semantic interpretation." Proceedings of IWCS-6, 2005.

[2] Bos J. et al., "Wide-coverage semantic representations from a CCG-parser" Proc. of Int. Conf. COLING, 2004.

[3] Burke D. et al., "Aging and Language Production," Current Directions in Psychological Science, 13, 2004.

[4] Chai J., Pan S. and Zhou M., MIND: A Context-based Multimodal Interpretation Framework, Kluwer ed., 2005.

[5] Clark S. and Curran J. "Parsing the WSJ using CCG and Log-Linear Models" Proc. of ACL-04 meeting, 2004.

[6] Gaitanis K., Correa P. and Macq B., "Modelization of limb coordination for human action detection," Proc. of the IEEE Int. Conf. on Image Processing, 2006.

[7] Holzapfel H., et al., "Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gesture", Proc. of the Int. Conf. ICMI, 2004.

[8] Kamp H. and Reyle U., "From Discourse to Logic. Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory", Kluwer, Dordrecht, 1993.

[9] Oviatt S. et al., "Toward a theory of organized multimodal integration patterns during human-computer interaction", Proc. of the Int. Conf. ICMI, 2003.

[10] Pfleger N., "Context based multimodal fusion", Proc. of the Int. Conf. ICMI, 2004.

[11] Pfleger N. and Alexandersson J. "Towards Resolving Referring Expressions by Implicitly Activated Referents in Practical Dialogue Systems" In: Proc. of the Workshop on Dialogue Semantics and Pragmatics, 2006.

[12] Vybornova O., Gaitanis K. and Macq B., "Plan recognition using multimodal integration," 2006, Proc. of Int. Conf. CogSci, 2006.