Semantic annotation of the French Media dialog corpus ⋄

H. Bonneau-Maynard1, S. Rosset1, C. Ayache2, A. Kuhn2, D. Mostefa2 and the MEDIA consortium

{maynard,rosset}@limsi.fr, {ayache,kuhn,mostefa}@elda.org, media@elda.org

(1) LIMSI-CNRS/France, (2) ELDA/France

Abstract

The French Technolangue MEDIA-EVALDA project aims to evaluate spoken understanding approaches. This paper describes the semantic annotation scheme of a common dialog corpus which will be used for developing and evaluating spoken understanding models and for linguistic studies. A common semantic representation has been formalized and agreed upon by the consortium. Each utterance is divided into semantic segments and each segment is annotated with a 5-tuplet containing the mode, attribute name representing the underlying concept, normalized form of the attribute, list of related segments, and an optional comment about the annotation. Periodic inter-annotator agreement studies demonstrate that the annotation are of good quality, with an agreement of almost 90% on mode and attribute identification. An analysis of the semantic content of 12292 annotated client utterances shows that only 14.1% of the observed attributes are domain-dependent and that the semantic dictionary ensures a good coverage of the task.

1. Introduction

The dynamic and interactive nature of dialog makes it difficult to construct a reference corpus of dialogs against which systems can be evaluated. Various influential projects have built the foundations of evaluation methodologies for spoken dialog systems, such as the ATIS [1] and COMMUNICATOR [2] projects in the USA, and the European project DISC [3]. However, presently, there are no common standard methodologies or practices agreed upon by the scientific community for the evaluation of spoken dialog systems. The aim of the MEDIA project [4], part of the French Technolangue program, is to carry out an automatic, comparative and diagnostic evaluation of the context-independent and dependent understanding capability of a dialog systems. The evaluation environment relies on the premise that, for database query systems, it is possible to construct a common semantic representation to which each system is capable of converting its own internal representation.

The first stage of the MEDIA project has been dedicated to defining, constructing and annotating a common dialog corpus. The first literal understanding evaluation test will be performed in June 2005 on a 3000 utterance test corpus. In order to enable each participant to adapt their models to the chosen evaluation domain and task, a 1250 dialogs have been recorded. A common semantic representation has been defined and notation of the corpus is underway. To the best of our knowledge it is the first time that a corpus of such a dimension will be available across different research teams. Beyond its utility for the adaptation of the participants’ systems, this work will be useful for future natural language research activities.

Table 1: Main characteristics of the 1257 dialog MEDIA corpus. The average dialog duration is 3’30.

<table>
<thead>
<tr>
<th>#utterances</th>
<th>wizard</th>
<th>client</th>
</tr>
</thead>
<tbody>
<tr>
<td>19635</td>
<td>18801</td>
<td></td>
</tr>
<tr>
<td>mean #words per utterance</td>
<td>14.4</td>
<td>8.3</td>
</tr>
<tr>
<td>vocabulary size</td>
<td>1932</td>
<td>2715</td>
</tr>
</tbody>
</table>

The remainder of this paper is organized as follows: The next section describes the French MEDIA-EVALDA project. The application domain and the recording protocol of the dialog corpus are then described in Section 3. The definition of the semantic representation and its adaptation to the domain is given in Section 4 along with a description of the annotation process. The annotation protocol and the results of the periodic inter-annotator agreement (IAGs) tests are given in Section 5 as well as an analysis of the semantic content of the annotated corpus.

2. The French EVALDA-MEDIA project

The aim of the French MEDIA-EVALDA project [4] is to define and test an evaluation methodology to compare and diagnose the context-sensitive understanding capability of spoken language dialog systems. Systems from both academic organizations (IRIT, LI A, LIMSI, LORIA, VALORIA) and industrial sites (France Telecom R&D) will be tested. ELDA is the coordinator of the Technolangue/EVALDA multi-campaign evaluation program, of which MEDIA is a sub-campaign.

Based on the PEACE paradigm [5], the MEDIA paradigm will be held in two evaluation phases: context-independent and context-dependent understanding evaluation. By context-independent understanding, we mean the building of the literal representation of the meaning of isolated spoken utterance, whereas by context-dependent, we mean the semantic representation of the query, taking into account the dialog context. In each case, the participants will have the corresponding annotated training corpus to allow the adaptation of its models to the task and the domain.

For literal evaluation, the corpus consists in a training portion of 10k requests, a development portion of 2k requests, and an unseen test corpus of 3k requests.

3. Domain description and recording protocol

The chosen MEDIA task is the reservation of hotel rooms with tourist information, using information obtained from a web-based database. The MEDIA dialog corpus (Table 1) was recorded using a WOZ system simulating a vocal tourist information phone server [4]. In this way, each user/caller believes he or she is talking to a machine whereas in actual fact he is talking to a human being (a wizard) who simulates the behavior of tourist
Figure 1: Example of the semantic attribute/value representation for the sentence “uhm yes the hotel whose price doesn’t exceed one hundred and ten euros”. The relations between attributes are given by their order in the representation and the composed attribute names. The segments are aligned on the sentences.

<table>
<thead>
<tr>
<th>r</th>
<th>word seq.</th>
<th>mode</th>
<th>attribute name</th>
<th>attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>oui</td>
<td>+</td>
<td>response</td>
<td>oui</td>
</tr>
<tr>
<td>1</td>
<td>l’</td>
<td>+</td>
<td>refLink-coRef</td>
<td>singulier</td>
</tr>
<tr>
<td>2</td>
<td>hôtel</td>
<td>+</td>
<td>BDOBJECT</td>
<td>hotel</td>
</tr>
<tr>
<td>3</td>
<td>dont</td>
<td>+</td>
<td>null</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>le prix</td>
<td>+</td>
<td>object</td>
<td>paiement-montant-chambre</td>
</tr>
<tr>
<td>5</td>
<td>ne dépasse pas</td>
<td>+</td>
<td>comparative-payment</td>
<td>inférieur</td>
</tr>
<tr>
<td>6</td>
<td>cent dix</td>
<td>+</td>
<td>payment-amount</td>
<td>110</td>
</tr>
<tr>
<td>7</td>
<td>euros</td>
<td>+</td>
<td>payment-unit</td>
<td>euro</td>
</tr>
</tbody>
</table>

Figure 2: Hierarchical representation derived from attribute/value representation of Figure 1.

in order to explicitly represent relationships between segments (section 4.2).

4.1. Semantic Dictionary

The semantic dictionary was jointly developed by the MEDIA consortium. The basic attributes are divided in several classes. The database attributes correspond to the attributes of the database tables (eg. BDOBJECT or payment-amount). The database attributes are classified in packages (eg. time or payment), which are domain-independent, and hotel which is domain-dependent). Each package is defined as a hierarchy of attributes (eg. package payment involves a sub-attribute amount which involves itself a sub-attribute integer). The modifiers attributes (eg. comparative) are linked to database attributes and used to modify the meaning of the relying database attribute (eg. in Figure 1 the comparative attribute, which value is inferieur (less than) is associated to the payment-amount attribute). General attributes are also defined as command-task which includes the different actions that can be performed on objects of the task, or command-dial with values cancellation, correction… One of the general attributes refLink is dedicated to the references annotation. Three kinds of references are represented: co-references (as in “in the same hotel”), co-domanial (as in “another hotel”), and element/set (as in “the first hotel”). The general and modifier attributes are domain independent and were directly derived from other applications[6] whereas most of the database attributes were derived from the database linked to the system.

In order to make it possible to rebuild a hierarchical representation of a query from the flat attribute-value representation, a set of specifiers are also defined in the semantic dictionary. Their combination with the database attributes specifies the ex-

information server. This enables a corpus of varied dialogs to be obtained due in part to the behavior of the wizard. 1257 dialogs were recorded, from 250 different speakers where each caller carried out 5 different hotel reservation scenarios. Several starting points were possible for the dialogs i.e. choice of town, itinerary, touristic event, festival, price, date etc. Eight scenario categories were defined each with a different level of complexity. The final corpus is on the order of 70 hours of dialogs, which have been transcribed by ELDA.
act relations between segments. In example of Figure 1, the attribute name payment-amount-integer-room results from the combination of a hierarchy of attributes from package payment-amount-integer and the specifier room. Attribute comparative-payment is also derived from the combination of the comparative attribute and the payment specifier. Two types of connectors are also defined: connect-Attr and connectProp which represent respectively the logical relations between attributes of a same object (with the default value and), and relations for complex queries (with values explanation, consequence or opposition. The example of Figure 1 can then be derived in the hierarchical representation given in Figure 2.

The semantic dictionary defined for the MEDIA project includes 83 basic attributes and 19 specifiers. The combination of the basic attributes and the specifiers, automatically generated by the annotation tool, results in a total of 1121 attributes that are able to be used during the annotation process. The 83 basic attributes includes 73 database attributes, 4 modifiers, and 6 general attributes.

4.2. Semantic unit segmentation and annotation

Semantic annotation is done on the transcriptions of the dialogs. In order to decrease the annotation cost, the annotation tool described in [6] was used. It helps for both the definition of the semantic representation and the annotation process. Semantic disambiguation may require to listening to the signal. The annotation tool provides compatibility with Transcriber [7], which is becoming a standard for speech transcription. The formalization of the semantic dictionary and the assistance provided by the tool to the annotators increases the consistency of the annotations. For literal annotation the tool presents the dialog turns in a random order to avoid the use of the dialog context. The attribute name is selected from the list generated from the semantic dictionary. Automatic completion of attribute names speeds up the process significantly and is well appreciated by the annotators. An on-line verification is performed on the attribute value constraints. The tool ensures that the provided annotation respects the semantic representation defined in the semantic dictionary.

5. Results

Based on the semantic representation described above, the literal semantic annotation of the MEDIA corpus has been performed by two ELDA annotators. The MEDIA consortium decided not to use semi-automatic techniques in order not to bias the evaluation process in the favor of one or two participant systems. The MEDIA corpus has been split by ELDA into randomly generated packages of 200 dialogs. Three packages have already been annotated and delivered to the participants, resulting in a total of 12,292 user utterances. The mean annotation duration time decreased from about 22 minutes per dialog for the first package, to 19 minutes per dialog for the second package and 15 minutes for the last package, resulting approximately in 5 times real time.

5.1. Inter-Annotator agreement

In order to verify the quality of the annotations, periodic evaluations are performed. For the attribute identification the kappa statistic[8], which normalizes for task complexity, can be computed by $k = \frac{P(A) - P(E)}{1 - P(E)}$, where P(A) is the proportion of time that annotators agree and P(E) is the probability that an annotator chooses the correct concept name by chance. A kappa value exceeding 0.8 is commonly considered in the literature as good. In this case, this corresponds to a correct agreement of more than 80% (P(E) is equal to 1/145 - where 145 is the total number of dictionary attributes which were observed).

Therefore, before beginning the annotation of the first package, some preliminary inter-annotator agreement (IAg) evaluations were performed, with the objective to obtain an agreement of more than 80% for the mode and attribute identification. The different annotation manuals were updated after each IAg in order to take into account the observed annotation inconsistencies. In order to ensure the quality of the annotation, periodic IAg scores were done after each package delivering. Figure 3 gives the IAg scores. It can be observed that annotators were able to obtain a good agreement after 3 steps. For the moment, the participants dispose of the first 3 packages. IAg 4, 5, and 6 were performed respectively after the annotation of package 1, 2 and 3. Table 3 describes the content and results of all the IAg scores. A slight decrease of agreement can be observed after IAg4. This decrease was mostly explained by a desagreement about the use of the mode + and ?. After a revision of the rules, annotators were able to get good score (IAg7). One of the most common sources of desagreement across the annotators is also the use of the specifiers. As far as specifiers don’t affect the meaning of an attribute, but are only useful for the building of a hierarchical structure from the attribute/value structure, this errors is not very important. However, this problem will have to be dealt with for the evaluation process. The attribute inter-annotator agreement is always greater than 80% (resulting in a kappa of more tha 0.8) and in the best case (IAg4) it is almost 90%. This agreement is correct enough to deliver the adaptation corpus for the participants to adapt their understanding models to the task. However it is very difficult to obtain a perfect agreement, so the consortium plans to perform a double annotation for the test portion of the corpus.

5.2. Analysis of the Semantic content of the corpus

An analysis of the semantic content of the first 617 annotated dialogs allows a good description of the global semantic content of the corpus. Figure 4 gives the variation of the number of
different attributes observed as a function of the number on annotated client utterances. Whereas the total number of attributes grows linearly, the attribute vocabulary size, after a rapid increase (in 2000 utterances 100 of the 145 attributes were observed), grows very slowly (10,000 utterances are necessary to observe the remaining 45 attributes).

Figure 5 shows the repartition of the different types of attributes. As it could have been predicted, the most frequent attribute is the yes/no response. It is interesting to note that the most frequently encountered attributes are task-independent (localization, time, ...) and that task-dependent attributes (hotel, room...) represent only 14.1% of the observed attributes. The semantic dictionary ensures a good coverage of the task as far as only 0.1% segments annotated with the unknown attribute.

### 6. Conclusion

The first success of the MEDIA project is that the consortium, which involves several teams, with very different speech understanding backgrounds for able to define a common semantic representation. The literal annotation of the MEDIA corpus will end in May 2005. The resulting annotated corpus will be useful not only for the MEDIA evaluation campaign but also for future work in natural language processing. Thanks to the annotated corpus, which contains the alignment of the annotations with the corresponding word sequences, each team is now adapting its understanding model to the task. The literal evaluation campaign is planned in June 2005. While the literal annotation is being performed, propositions for contextual annotations have been done. A contextual understanding evaluation campaign is planned for September 2005. An evaluation package which includes this corpus along with protocols, scoring tools, and evaluation results that were produced during the campaigns will be available and distributed by ELDA.

### 7. References


