An Incremental Architecture for the Semantic Annotation of Dialogue Corpora with High-Level Structures. A case study for the MEDIA corpus

Lina Maria Rojas Barahona, Matthieu Quignard

An Incremental Architecture for the Semantic Annotation of Dialogue Corpora with High-Level Structures. A case of study for the MEDIA corpus.[*]

Lina Maria Rojas-Barahona and **Matthieu Quignard**
LORIA/INRIA, France
{lina.rojas,matthieu.Quignard}@loria.fr

Abstract

The semantic annotation of dialogue corpora permits building efficient language understanding applications for supporting enjoyable and effective human-machine interactions. Nevertheless, the annotation process could be costly, time-consuming and complicated, particularly the more expressive is the semantic formalism. In this work, we propose a bootstrapping architecture for the semantic annotation of dialogue corpora with rich structures, based on Dependency Syntax and Frame Semantics.

1 Introduction

We propose a cooperative architecture that incrementally generates and improves the annotation of the French MEDIA dialogue corpus with high-level semantics (HLS), as a result of the cooperation of several linguistic modules. MEDIA is a French corpus that has collected about 70 hours of spontaneous speech from the task of hotel room reservation. It contains transcribed utterances[1] that have been manually segmented[2] and annotated with a **flat semantics** i.e.,concept-value pairs (Bonneau-Maynard et al., 2005).

---

[1]Utterances with ellipsis, disfluencies, false starts, reformulations, repetitions and ungrammaticalities and special characters such as the symbol '*' that indicates uncertainty due to noise in the communication channel.

[2]The term *Segment* means sequence of words in utterances.

The **HLS semantics**, namely the MultiModal Interface Language formalism (MMIL) (Denis et al., 2010), augments the expressivity of the flat semantics by representing communicative actions, predicates, arguments and fine-grained features. Communicative actions are **components** built up from two types of entity (i.e. *events* and *participants*), which are linked together by *relations* and described by sets of **features** (attribute-value pairs). It is possible to identify in entities a set of **main features**, which can be domain-specific. For the semantic annotation, components are mapped to segments in utterances. Figure 1 shows the canonical representation of an utterance in the corpus in compliance with the specifications for the annotation[3].

2 The Architecture

The architecture (Figure 2) for the automatic annotation has been formulated as a post-interpretation process that takes place after the syntactic analysis and semantic role labeling (SRL). Two linguistic resources interact within the architecture, the corpus and the Frames[4]. Four linguistic modules are involved in the annotation: the Part-Of-Speech (POS) tagger, the parsing, the semantic-role labeling (SRL) and the HLS Builder. The common knowledge base comprises two knowledge-bases (one for the domain and the other for the HLS formalism) together with a relational database management system (RDBMS). The knowledge bases assure the coherence of the an-

---

[3]http://www.port-media.org/doku.php?id=mmil_for_annotating_media

[4]Frames is the process in which the frames and frame elements (FE) are defined.
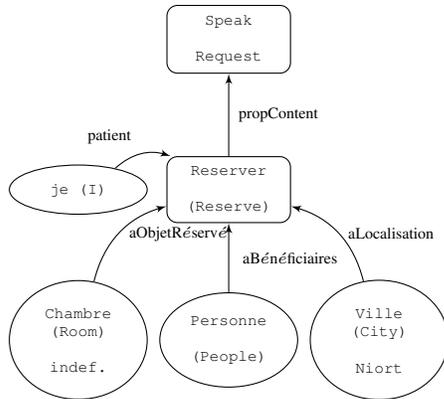
Figure 1: HLS representation for the French utterance "je voudrais faire une réservation d' une chambre pour une personne à Niort" (*So I would like to make a reservation for a room for one person in Niort*). It shows a *request to reserve*: the communicative action is *Request* the main event is *Reserve*. Note that the beneficiary and the patient are two different roles, the beneficiary is the person, not necessarily the same speaker, who will use the object reserved (e.g. rooms). The patient is the speaker. The segmentation of the HLS Component is presented in the Table, the component is mapped to the whole utterance. The fine-grained segmentation of features is shown for the Participant 2.

notation while the database assures persistence and data integrity. The database stores the corpus, the frames, the results at each level of analysis, as well as the progress in the annotation. The persistence permits progressively optimizing the algorithms until the desired annotation is obtained and integrated into the corpus files. The corpus manager is in charge of the resources management. Last but not least, two annotation tools were built: one for the SRL gold standard (web-based) and the other for the HLS gold standard (standalone).

Syntactic Analysis. We decided to employ statistical approaches that could learn the irregularities of spoken language: the French Tree-Tagger[5] and the dependency-based MALT-PARSER (Nivre et al., 2007). The parser has been trained with 1449 utterances annotated according to the annotation guidelines described in (Cerisara and Gardent, 2009).

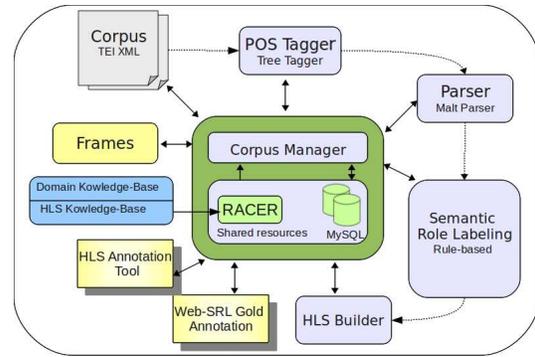[5] http://www.ims.uni-stuttgart.de/~schmid/

Figure 2: General Architecture for the HLS Annotation.

Definition of Frames. Frame Semantics, (Baker et al., 1998) arranges common background knowledge for situations by grouping verbal, nominal causative and non-causative predicates. Nevertheless, paraphrases are more used in spoken language than explicitly uttered nouns, adjectives or verbs for referring to a situation (e.g.'ask', 'request' or 'demand'). Here we introduce the term: **Frame Evoking Phrase** (FEP) for evoking frames and we include syntactic templates that mirror these phrases in frames and frame elements (FE). Table 1 summarizes the differences between PORT-MEDIA frames and FrameNet (Baker et al., 1998).

	FrameNet	PORT-MEDIA
Frames	Lexical Units	Lexical Units, POS tags and templates MEDIA Flat Semantics
Frame Elements	Lexical Units, Phrase Type and Grammatical Function Semantic Type	Lexical Units, POS tags, templates and dependency relation Semantic Type and MEDIA flat semantics

Table 1: Static Characteristics of Frames in FrameNet and in PORT-MEDIA.

Semantic Role Labeling. We built a rule-based semantic role labeling for detecting frames and FE (roles) by using dependency tree-template pattern matchers that exploit the information already compressed in frames. The SRL detects the boundaries of FEP and FE by measuring the syntactic and semantic similarity between the utterance and the frame.

HLS Builder. The HLS Builder is the last phase in the annotation process: it is rule-based and it takes utterances in the corpus with their flat semantics, de-

pendency trees and predicates-arguments and builds the HLS representation (See Figure 1), according to the specifications for the annotation and the knowledge bases. The *dialogue act* and *main event* in HLS components can be detected from the predicates. Similarly, *secondary events* and *participants* with their *features* can be detected from the roles and the flat semantics.

3 Evaluation and Discussion

For evaluating the system we separately computed the accuracy of its linguistic components. **The parser** achieved a label attachment score (LAS) (Nivre et al., 2007) of 86.16%, with a training set of 1097 utterances and a test-set of 100 utterances. **The SRL** was evaluated with metrics adapted from the CONLL 2005 evaluation (Carreras and Màrquez, 2005) f*or supporting FEP and allowing overlapped FEP for different frames*. The LAS was computed by comparing the semantic dependencies of *system's and gold's propositions*[6] and their segments. The gold standard comprises 115 utterances annotated with the major frames in the domain: Request, Reserve and Attributes. The F1-measure computed for propositions with exactly the same segments was 56.66%. When verifying whether the segments contain the same syntactic governor, the SRL achieves a better score: 71.30%. Finally, varying the number of excluded words in both segments[7] yielded a constant increase of the F1-measure until a maximum of 84.27%. **The HLS annotation** was evaluated by measuring the similarity between *gold's and system's components* with a gold standard of 330 complex utterances related to the reservation task. When rigorously measuring the *equality of components*[8], we obtained a F1-measure of 57.79%. Measuring equality of components without being so rigorous with features' segmentation, yielded a slightly higher score 63.31%. Finally, when measuring equality of components by taking

into account only the **main features** of entities, we obtained a higher score: 70.65%.

We proposed an architecture for corpus management that allows incremental updates over persistent information until a more accurate semantic annotation is obtained. The preliminary results show a general agreement when defining *the main features and the main entities in HLS components* and a disagreement when segmenting fine-grained features. We observed that the system tends to create new entities when it detects repetitions or references in long utterances. Defining a more precise segmentation policy in the manual annotation guidelines, augmenting the training data for parsing, as well as integrating reference resolution and disambiguation techniques, will enhance the annotation process. An appealing research direction would be to integrate and evaluate machine learning components in the architecture.

---

[6]A proposition is a structure containing the predicate, their arguments and the semantic relation between them.

[7]From 1 to n words not common in both segments.

[8]Two HLS components are equal if their entities and relations are equal. Two entities are equal if they have the same segment and features (feature name and feature value) and if these features are mapped to the same segments in the utterance. Two relations are equal if they have the same source and target entities as well as the same name

References

Hélène Bonneau-Maynard and Matthieu Quignard and Alexandre Denis. 2005. MEDIA: A semantically annotated corpus of task oriented dialogs in French. *Language Resources and Evaluation*.

Alexandre Denis and Lina M. Rojas-Barahona and Matthieu Quignard. 2010. Extending MMIL Semantic Representation: Experiments in Dialogue Systems and Semantic Annotation of Corpora. *In: Proceedings of the Fifth ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5), Hong Kong.*

Collin Baker and Charles Fillmore and John Lowe. 1998. The Berkeley FrameNet Project. *Proceedings of the 17th International Conference on Computational linguistics*, 86–90. Association for Computational Linguistics.

Joakim Nivre and Johan Hall and Sandra Kübler and Ryan McDonald and Jens Nilsson and Sebastian Riedel and Deniz Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007.* Prague, Czech Republic:915–932. Association for Computational Linguistics.

Christophe Cerisara and Claire Gardent. 2009. Analyse syntaxique du français parlé. *Journée thématique ATALA Quels analyseurs syntaxiques pour le français.*

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. CONLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning. 152–164. Association for Computational Linguistics.