Semantic Analysis in F-logic

No Author Given

No Institute Given

Abstract. In this paper we present the semantic analysis module in F-logic as a part of Croatian spoken dialog system developed for a limited domain of weather forecasts. Semantic analysis is implemented as an information extraction technique using the F-logic formalism. F-logic (Frame Logic) is a logic formalism with a second order syntax and an object-oriented approach. Semantic knowledge is captured through semantic categories, a semantic dictionary and output frames represented in F-logic. The developed semantic analysis is conducted in three phases. In the first phase the main semantic context for the input text is determined. In the second phase semantic units are analysed and knowledge slots in the database are filled. Since some slots of input data are missing in the third phase, incomplete data are updated with missing values. The semantic analysis is evaluated with 3 months period weather data in terms of frame rates and slot error rates. The rules for semantic analysis are defined in the F-logic language and implemented using the FLORA-2 system.

1 Introduction

This paper presents the semantic analysis in F-logic (Frame Logic) which is a part of Croatian spoken dialog system for weather information in Croatia. In such a system a user could ask questions about weather forecasts and weather conditions in Croatia and at the Adriatic Sea for different time periods. The input to the semantic analysis is the text recognized by the speech recognition system or texts collected from Web pages of the Croatian Institute of Meteorology. For fixed word order languages like French or English a semantic analysis can be performed using grammars [5] or grammars in combination with statistics and rules [7]. For free word order languages like German the rule based parsing can be more appropriate [2]. The Croatian language is a highly flective (cases, genders, numbers) and mostly free word order language, therefore instead of formal grammars the information extraction technique is used for semantic analysis implementation. The Croatian weather domain knowledge can be represented by the use of frames with slots that need to be filled with words from the input text. The slot filling approach for semantic analysis has been used in [1, 6, 8]. In our approach a deductive object-oriented (OO) logic programming language has been used for implementation. The main gain from this approach is a unified framework for semantic representation and analysis of data combined with OO paradigms, which enables the use of lattices for semantic relations definition and

the semantic context determination. F-Logic and FLORA-2 system have been used for both semantic representation of weather forecast data and for semantic analysis. The proposed semantic analysis for Croatian data is conducted through three main phases. As shown in Fig. 1, in the first phase a semantic context for the input is determined. In the second phase the input is decomposed into semantic units. Semantic units are analyzed and slots of semantic database are filled. Since some slots of input data are missing in the third phase the incomplete data is updated with missing values. Semantic data analysis is based on a previously defined dictionary, phrases, semantic categories, and output frames.

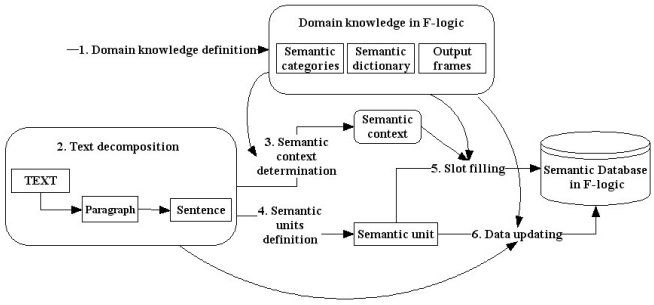

Fig. 1. Semantic analysis in F-logic

The second section of this paper describes the F-logic formalism. In the third section knowledge representation concepts in F-logic are described. Section four presents our three phase approach to semantic analysis in F-logic: determination of the semantic context of the input text, analysis of semantic units and updating of incomplete data. In the fifth section the evaluation procedure is described and experimental results are presented. Finally some possible improvements are discussed and some future work plans are presented.

2 F-logic

The frame logic (abbr., F-logic) is a formalism connecting object-oriented approach, frame languages and logical reasoning [3]. The authors of F-logic extend the classical predicate calculus aiming to define a logic that would enable inferring in object-oriented databases. At syntax level, F-logic is extended with a set of additional symbols, while at semantics level formulas of F-logic assume meaning where implementation of basic object-oriented concepts is possible. Defined in a such way, the F-logic retained some quality properties of the classical predicate logic in terms of defining derivation rule that is analogous to the resolution with unification procedure as in classical logic [3]. One of the F-logic implementations was achieved in system FLORA-2. Basic formulas of the F-logic language show objects and their respective attributes or methods ($object[attribute \rightarrow$

value]) and appurtenance of an object to a class (*object* : *class*) as well as appurtenance of a subclass to a superclass (*subclass* :: *superclass*). As an object-oriented formalism F-logic captures all important object concepts. In this work F-logic language is used for domain knowledge representation through: semantic dictionary, semantic categories and output frames definition. The rules for semantic analysis are expressed in F-logic as well.
The formal description of the Croatian weather domain is in F-logic. The meaning of the words is captured by the category in the semantic dictionary and the meaning of the units is captured in slots of the output frames.

2.1 The weather domain description

The Croatian weather forecasts domain has a limited vocabulary (2300 words). The main reason lies in the fact that exact meteorological phrases are often used in weather forecasts and weather forecasts have predefined structures consisting of different weather related sections: meteorological situation, current temperature, forecast for the following day, expected temperature, etc. The weather forecasts related texts were collected from transcriptions of spoken television and radio weather reports and from Web sites of the Croatian Meteorological Institute in a period of four years. According to the content and the structure of collected texts the formal description of the domain was specified and represented in F-logic. In the domain knowledge representation the weather forecast is a part of more complex maritime weather structure which captures all forecast features: the meteorological situation, the warning, the visibility, the sea waviness, the wind condition, the weather forecast and the temperature. The maritime weather forecast is very detailed because it is mainly prepared for all naval activities: professional or leisure sailing, surfing, diving, fishing, etc. Since the weather forecast has plain structure, it is considered a part of the maritime weather structure. This makes it possible for the same weather domain structure to be appropriate for weather representation of maritime and continental regions of Croatia.

2.2 The weather knowledge representation

In order to find out the meaning of the sentence, words from the sentence have to be associated with some semantic representation of the domain knowledge. The limited weather domain knowledge is captured in three basic concepts: semantic categories, semantic dictionary and output frames. Each word from the weather domain is associated with a semantic category, that is represented in the F-logic language with classes and subclasses. There are 10 main semantic categories in the dictionary (semantic categories of the first level). Each of these categories consists of semantic subcategories. Overall there are 36 semantic subcategories at the second level. Some semantic categories are referring to relative terms (relative time, relative place, relative forecast). Relative terms are related to the other parts of the sentence which contain information necessary to capture the meaning of the relative term. The relative terms are analyzed in the last phase

of semantic analysis by special rules that can handle missing data values.
According to all semantic categories of the words in a sentence or in a part of the sentence (sentence unit), it is possible to define the semantic context of a whole sentence. The semantic context extracts the general meaning from the sentence. For example the semantic context of the sentence can be wind, temperature, the level of the river water etc. The word dictionary was prepared from collected texts and contains almost 2300 different words with the appropriate semantic category. The semantic dictionary is in F-logic language. Since Croatian language is highly flective the dictionary comprises all word formats that occur in the data. For each word an adequate basic word form (lema) is given. For example a noun in genitive is transformed into basic word form as a nominative noun. One of the goals of semantic analysis is preparation of a semantic database. The semantic database captures weather knowledge and weather data used for answer generation by the dialog manager. The semantic database schema is organized with hierarchy of output frames. Each frame is described with slots. Concepts of frames, slots and values are naturally captured in F-logic formalism because it has an object-oriented approach. Semantic analysis can be considered as a process in which each input sentence (unit) is mapped into an adequate semantic representation (frame). All relevant information are extracted from written input and transformed into slot values of the output frames [1]. In this work we used three levels of output frames for capturing a meaning of a sentence (unit). Figure 2 shows three levels of output frames. The first level is the most general one and corresponds to the semantic context. The second level contains more detailed semantic frames. Finally at the lowest level there are the six most detailed frames: weather, temperature, wind, sea, visibility and meteorology with slots. Figure 2 presents only one segment of the domain. Besides sea weather forecast and weather forecast, the biometorological forecast, the level of river water and other types of forecasts are included in the domain as well. Slots time and place are common for every output frame. Other slots like weather_info and wind_name depend on the output frame type. Some main semantic categories are at the same time output frames as well.

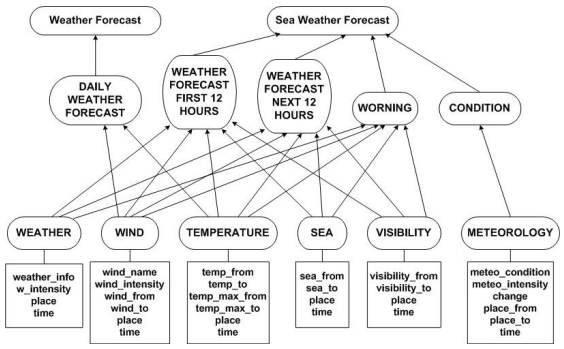

Fig. 2. Output frames with slots

3 Semantic analysis phases

This section describes three phases of the semantic analysis. First the semantic context of the input text is determined and input text is decomposed into semantic units. Each unit is analyzed and slots of the output frame are filled. Finally incomplete data is updated with missing values.

3.1 Semantic context

Semantic analysis starts with determining a semantic context of input text. Semantic context is determined by choosing adequate output frame of the first level. There are three levels of output frames (Fig. 2) but only the first level corresponds to the semantic context of the input text (weather forecast, sea weather forecast). The input text is organized into paragraphs (Fig. 1). The paragraph context corresponds to the second level of a frame and finally the sentence (unit) context is defined, which corresponds to the third frame level. This way each input sentence is connected with three semantic contexts at different levels. Strategy of semantic context determination depends of the text format and text source (TV news, radio news, Web pages). For Web pages texts semantic context can be determined using key words, headers, tags and similar marks. Situation is more complex with recognized text of spoken utterances. In that case we use lattices as mathematical formalism for choosing an adequate semantic context. In the implemented solution the context is determined by generating the set of all possible semantic categories in text paragraph and to define a superclass of that set as a semantic context. All semantic categories are hierarchically organized and represented as classes in the F-logic language. We form a lattice structure in a way such that there is exactly one class that is a subclass of all classes and there is exactly one class that is a superclass of all classes given by semantic categories. With class inclusion as a partial order relation, this set of classes forms a lattice. Lattices are implemented in FLORA-2 system.

3.2 Analysis of semantic units

Semantic units are generated from the input text as parts of sentence that contains necessary information to fill values of output frame slots. After that output frames instances are generated from semantic units, relevant slots are filled with adequate values extracted from semantic units [1, 6]. The input text is divided into paragraphs, sentences and sentence parts like the one shown at Fig. 1. Sentence parts are detected on the delimiters places: commas and conjunctions. The speech recognition subsystem output is text with natural pauses noted in angle brackets (for example <breath>, <sil>) instead of punctuation marks. All events in angle brackets are used as delimiters as well. Sentence parts are further transformed into semantic units according to a general principle: a sentence part represents a semantic unit if it contains new information about any weather aspect. Sentence parts that contain no new information are concatenated to previous or next semantic units according to some additional rules implemented in

FLORA-2 system. If a sentence part contains new information, a new output frame instance with slots is inserted into the semantic database. The type of the frame is determined by the semantic context of the sentence part. Each semantic unit is mapped into exactly one output frame instance and semantic unit content is transformed into slots of a frame. Sometimes it is impossible to fill all data slots at once. Therefore some frame instances have missing slot values. In this case the incomplete data is updated in the last phase of the process.

3.3 Updating incomplete data

Some output frame slots can have missing data because semantic unit is referred to a previous or next semantic unit with some relative term. In the third phase of semantic analysis the incomplete data of relative terms are updated with missing values by a set of rules implemented in FLORA-2. Each output frame instance should contain three mandatory values: general information, place and time. After the second phase there are some frame instances that miss this crucial data. So place, time and weather slots are filled from previous or next semantic units. Besides missing slot values there are some slot values filled with relative data. The relative data should be transformed into adequate value contained in previous or next units. Figure 3 presents an example of incomplete data updating.

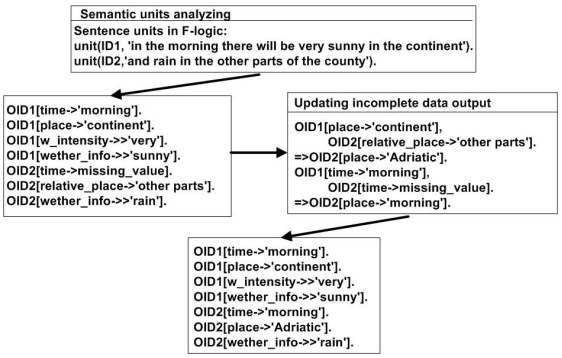

Fig. 3. Updating incomplete data

4 Evaluation and results

The semantic analysis process described in previous section was evaluated with Web weather data collected in a three month period. The texts for evaluation consist of one general weather forecast for Croatia and three maritime weather forecasts per day. For the test text we manually prepared a reference semantic database that includes all correctly determined slots and slots values [4]. A

test semantic database was generated by the F-logic system from the same evaluation texts. The evaluation is performed on the frames and slot value level. From the test text 2696 slot values of 398 frame instances were generated. First the number of correctly determined frames corresponding to semantic units is evaluated. The semantic unit evaluation results in terms of number of correct frames and error rates are shown in Table 1. The slot values are evaluated by

Table 1. The frame evaluation results

	Weather forecast	Maritime forecast	OVERALL
Reference	267	131	398
Generated	265	129	394
Correct	264	129	393
Error rate	1.49%	1.52%	1.51%

slot error rate [4]. The slot error rate is analogous to the word error rate in speech recognition performance. Slot error rate combines the deleted, inserted and substituted types of error. An algorithm implemented in FLORA-2 system is used to align generated against reference slot values. The corresponding slots are then matched and scored as either correct or not. If not correct, the error is marked as a substitution (incorrect slot), deletion (missing slot), or insertion (spurious slot). According to a number of errors measure of performance can be computed using the formula:

$$SER = \frac{N_S + N_I + N_D}{N_C + N_S + N_D} \tag{1}$$

where N_S is a number of substituted slots, N_I is a number of inserted slots, N_D is a number of deleted slots and finally N_C is number of correct slots. The slot error rate (SER) results are shown in Table 2. The 20% range of slot error rate was expected due to the flective nature of Croatian and due to existence of referenced terms in the input texts. It is reasonable to assume that additional improvement can be achieved in modification of relative terms manipulation.

Table 2. The slot value evaluation results

	Weather forecast	Maritime forecast	OVERALL
Reference	1655	1041	2696
Generated	1655	1038	2703
Correct	1655	918	2421
SER	18.97%	23.35%	20.66%

5 Conclusion

This paper presents semantic analysis of Croatian weather data in object-oriented logic programming language F-logic. Domain knowledge representation and semantic analysis is implemented in F-logic using FLORA-2 system. The domain data semantics is captured in a semantic dictionary, semantic categories and output frames. The proposed semantic analysis of Croatian language implements a slot filling technique in three phases. Initially the semantic context of the input text is determined. Further the input text is divided into semantic units. Then the slots of output frames are filled with values. Finally, since some missing data value can occur in slot filling phase, the incomplete data is updated.
Proposed approach for semantic analysis is used in the spoken dialog system for Croatian weather forecast. The aim of semantic analysis in a spoken dialog system is two folded. First it has to decompose Web weather data into the semantic database and second it has to analyze the text recognized by the speech recognition subsystem. Since the semantic analysis is the first step in the spoken dialog management subsystem activities towards speech understanding and answer generating should be considered. Presented semantic analysis was evaluated using test texts collected from Web pages. The achieved results of 20% slot error rate are quite promising for further development of Croatian semantic analysis for spoken dialog systems. In the future we will consider the use of formal grammars in F-logic in order to improve updating of the incomplete data with missing values. Further we will expand the domain of interest with some specific kind of forecasts, such as forecasts for agriculture. That means that we will need to add some new data sources and consider heterogeneous data integration.

References

1. Cardie, C.: Empirical Methods in Information Extraction. AI Magazine. **18(4)** (1997) 65–79
2. Engel, R.: Robust and Efficient Semantic Parsing of Free Word Order Languages in Spoken Dialog Systems. Interspeech 2005, Portugal (2005) 3461–3464
3. Kifer, M., Lausen, G. and Wu. J.: Logical Foundations of Object-Oriented and Frame-Based Languages. Journal of the ACM **42(4)** (1995) 741–843
4. Makhoul, J., et al.: Performance Measures For Information Extraction. in DARPA Broadcast News Workshop (1999)
5. Rayner, M., Carter, D.: Fast Parsing Using Pruning and Grammar Specialization. In 34th Annual Meeting of the Association for Computational Linguistics, Morristown, New Jersey (1996)
6. Souvignier, B., et al.: The Thoughtful Elephant. Strategies for Spoken Dialog Systems. IEEE Trans. Speech and Audio Processing **8** (2000) 51–62
7. Wang, Y.-Y., et al.: Combination of Statistical and Rule-based Approaches for Spoken Language Understadning. International Conference on Spoken Language Processing (ICSLP) (2002) 609–612
8. Žibert, J., et al.: Development of a Bilingual Spoken Dialog System for Weather Information Retrieval. EUROSPEECH 03. **1** (2003) 1917–1920