Improving Knowledge Representation to Speed up the Generation of Grammars for a Multilingual Web Assistant

Marta Gatius
Department of Computer Science, Technical University of Catalonia, Barcelona, Spain
gatius@lsi.upc.edu

Abstract

This paper describes the use of a syntactico-semantic taxonomy to facilitate the generation of grammars for a multilingual web assistant. In particular, it describes the generation of grammars for two different domains: cultural events and medical specialists.

1 Introduction

Most practical conversational systems use semantic grammars adapted to a specific domain because processing results faster and more robust against errors. However, the cost of adapting those grammars to new domains and languages is usually high. To reduce this cost, many systems use semantic models representing domain entities and application specifications to facilitate the generation process. The use of semantic models representing domain concepts is especially appropriate for multilingual systems. Some of those systems use database models (Polifroni et al., 2003; D'haro et al., 2009), others use richer formalisms, such as ontologies (Dzikovska et al., 2003; Cimiano et al., 2007; Sonntag et al.,2007; Nesselrath and Porta, 2011).

In many communication systems only syntax and conceptual levels are distinguished, as in many linguistic works (Jackendoff, 1983). Our approach also distinguishes an intermediate semantic level between these two levels, as proposed in other works (Haliday, 1985; Perkings, 1989; Bateman, 1994).

Our work is on the use of a syntactic- semantic taxonomy to facilitate the generation of grammars in several languages from domain concepts. We have previously used this taxonomy for generating system messages in a dialogue system supporting English, Spanish and Catalan (Gatius et al., 2007). More recently, we have studied its possible usability for a language with a different organization, Hindi, (Gatius and Pailwal, 2013). In this paper, we describe how this taxonomy is used to generate the grammars supporting user's questions on two domains: cultural event and medical specialists.

2 Proposed Knowledge Representation

Our work is focused on the questions about specific domain information the user asks when looking for web information. For this reason, the syntactico-semantic taxonomy we use relates attributes describing domain concepts to the different grammatical structures appearing in questions about those concepts attributes. All the attribute classes distinguished in the taxonomy are necessary to reflect different surface realizations. The basic attribute classes are associated with grammatical roles: participants (**who_does, who_object, what_object**), being (**is**), possession (**has**), descriptions and relationships between two or more objects (**of**) and related processes (**does**). The class **of** is subdivided into three classes: **of_person** representing relations between persons, **of_object** representing relations between objects and **of_description** representing qualities and circumstances related to the concept. The class **of_description** has been subclassified into subclasses representing time, place, manner, cause, quantity, name and type.

Each subclass is associated with several patterns to express questions and answers about the attribute belonging to the class. Additionally, subclasses have been further subclassified if other information relevant for the linguistic realization can be considered, such as having an associated verb or preposition. For example, attributes in the class **of_name** can be realize with general patterns (i.e., *What' s <concept-name> name?*), but a new subclass **of_name_person** has been distinguished and it is is associated with the particle title (i.e., *Dr.*).

We have extended theclasses **of_time** and **of_place** by studying the descriptions of time and locations appear in the domains considered. For example, locations of equipments usually consist of a street address or a city zone.

The class **of_time** has been subclassified considering time units and the different forms of expressing them (i.e. weekdays, weekend). Patterns associated with these subclasses cover several forms of expressing time, including, for example, descriptions of intervals of time.

We have used Grammatical Framework (GF) for implementing the grammars because this framework favors the generation of grammars in several languages (Ranta, 2011). In GF, grammars are separated in two parts: **abstract syntax**, defining meaning and **concrete syntax**, mapping meanings to linguistic realization. The abstract syntax is shared across languages while concrete syntax is specific for each language.

In next subsection we describe how we have used the taxonomy to write grammars in GF representing user questions when looking for web information in two domains: cultural events and medical specialists.

2.1 The Generation of Grammars

The process of generating a semantic grammar for a new domain consists of several steps. In a first step, the domain concepts appearing in the communication have to be described by a set of attributes. Then, those attributes have to be classified according to the syntactico-semantic taxonomy. Next, for each language considered, the lexical entries related to the concepts and their attributes have to be incorporated. Using this information, grammars for several languages can be automatically generated. Although the resulting grammars have to be manually supervised and extended, the effort of generating semantic grammars for different languages from scratch is considerably reduced.

Let's see the process of generating a grammar for the domain of cultural events. There are many web sites giving information on cultural events. Although information appearing in all those web sites is not the same, in most sites there is information about the title, genre, venue and date of the cultural events. For this reason, we have represented this information as the attributes of the concept **Cultural_Event**. Then, those attributes have been classified according to the sintactico-semantic taxonomy, as shown in Figure 1. The attribute **title** represents the name of the event and is obtained at run-time from the web service. It is linked to the class **of_name**. The attribute **genre** has as value the type of the event (i.e., cinema) and is linked to the class **of_type**. The attribute **date** has as value a set of dates and is linked to the class **of_date** (a subclass **of_time**). The attribute **venue** has as value an instance of the concept **Venue** and is linked to the class **of_place**.

Cultural_Event	**Doctor**
title : of_name	**name:of_name_person**
genre : of_type	**specialist : of_type**
venue : of_place	**equipment : of_place**
days : of_date	**days : of_weekdays**

Figure 1. Classification of conceptual attributes

Next, the lexical entries associated with the concept (*cultural event* and *take place*) and the attribute values (except those set at run-time and those reused across domains) have to be incorporated. Then, the abstract syntax grammar in GF is obtained. A fragment of the event grammar is shown in Figure 2. As indicated in the header, this grammar uses the grammars **place** and **time**, that define the structures referring to time and locations.

```
abstract event = extends place, time+
flags startcat = askinf
cat askinf; converb;valtype; valplace; valname
fun
geninf : valtype → askinf
conceptualverb : valtype → converb ->valtype
geninfplace : valtype → valplace → askinf
where : valname → askinf
when : valname → askinf
whatname : valtype → askinf
music, cinema, theater, sport, circus: valtype
takes_place : converb
```

Figure 2. A fragment of the abstract grammar

From the abstract grammar, a concrete grammar is automatically generated for each language using the patterns associated with each attribute class.

The process of generating the grammar for the health domain is similar. Figure 1 shows the main domain concept, **Doctor**, and the semantic classification of the attributes describing it.

3 Conclusion

The use of a syntactico-semantic taxonomy acting as an interface between domain conceptual and general linguistic knowledge reduces the effort of generating grammars for new domains and languages. The reuse of grammars defining several forms of expressing time and locations also limits this effort.