Using Data Integration Techniques in a Federation of Heterogeneous Lexical Databases

Artola, X.
Dept. of Computer Languages and Systems
University of The Basque Country
649 p.k., 20080 Donostia (Basque Country)
jiparzux@si.ehu.es

Soroa, A.
Dept. of Comp. Science and Artificial Intelligence
University of The Basque Country
649 p.k., 20080 Donostia (Basque Country)
ccpsoeta@si.ehu.es

Abstract

An architecture for a federation of highly heterogeneous lexical information sources is presented. We address here the problem of querying very different existing lexical sources from the point of view of information integration, using a common query language. Each source is described as a view over a general conceptual model. The conceptual models and an algorithm that translates queries from the general model into each particular schema have been implemented.

Introduction

The purpose of this paper is to present an integrating architecture for a federation of highly heterogeneous lexical information sources. Our proposal is that lexical and dictionary stores can be brought together in an integrated federation of databases, so that we can query them by using a unique language, regardless of the particular resource to which the query will be finally targeted. As desirable as it might be, it is unrealistic to expect that the great variety of available lexical information resources could be converted into a single and standard representation schema in the near future. Hence, our approach aims for a solution that does not require any conversion of the sources, and that participates of the emerging "wrapper" technologies that enhance external communication with legacy systems (Roth & Schwartz, 1997).

We address the problem of querying very different sources of lexical information from the point of view of information integration. The so-called *local-as-view* paradigm is used for describing each lexical source as a view over a general conceptual model (GCM).

1 An architecture for a federation of heterogeneous lexical databases

The most important features of the architecture proposed in this paper are (1) the incorporation of all the different lexical information sources into a federation of databases (the term database is here used in its widest sense), and (2) the conception of a GCM that allows users and/or programs to query the system regardless of the particular source to which the query will be finally addressed.

The GCM provides the user with a global view of the information the system contains, and has to be provided in terms of representation mechanisms that allow us to abstract from the physical and logical structure of the sources.

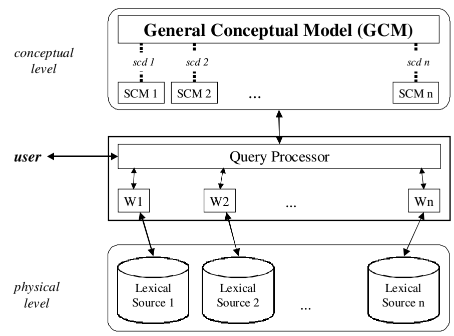

Figure 1. General view of the architecture of the system.

The sources are autonomous, and highly heterogeneous in the sense that they use very different data models and vocabularies. You could find among them from relational databases up to ontologies such as WordNet or SGML-coded MRDs, or even language-processing programs such as lemmatisers or part-of-speech taggers.

In Figure 1, a general view of the architecture of the system is given. At the top, the GCM, the conceptual model of each source (SCM), and the content descriptions of these sources (*scd*$_i$, represented in the figure by means of dotted lines) are depicted. The physical level, that is, the actual lexical resources, are illustrated at the bottom of the diagram. In the middle part, we can see the information flow (note the arrows) between the user, the main query processor, and the wrappers of each source (*w*$_i$).

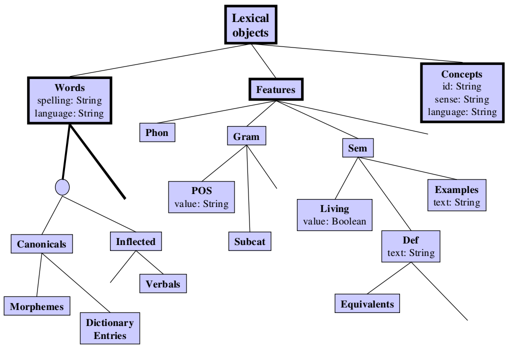

Figure 2. Partial view of the general conceptual model for describing lexical knowledge.

2 A general conceptual model for describing lexical knowledge

If you want a conceptual model for describing lexical knowledge to be general, this must be necessarily simple. But, at the same time, it must gather all the features and relationships that a lexical information source can establish.

Relation	Domain	Range	Max. Card.
`senses`	`Words`	`Concepts`	n
`synonyms`	`Concepts`	`Concepts`	n
`pos`	`Concepts`	`POS`	1

Figure 3. Some relationships established among the classes in the general conceptual model.

In our model, lexical knowledge is described by means of a taxonomy of classes and by a set of binary relationships established among them. With regard to the taxonomy, you can find at the top a main class called *Lexical Objects,* and three subclasses of it: *Words, Concepts* and *Features* (see Figure 2).

As for the relationships, a wide set of relations that link different classes in the hierarchy has been established and defined by specifying the domain, the range and the maximum cardinality of each relationship (see Figure 3).

3 Describing different lexical sources according to the local-as-view paradigm

We describe two types of information for each source: its conceptual model (3.1) and the description of its contents (3.2).

3.1 Conceptual models (CMs)

Due the "wrapper" technology we have adopted, we can make an abstraction, and conceive and model an information source as a collection of objects belonging to some source relations. So, like in the GCM, every source model will be described by means of a taxonomy and by a set of binary relationships.

The use of Description Logics to express domain knowledge is widely explored in the literature (Calvanese *et al.*, 1999; Levy *et al.*, 1996). We describe the different conceptual models using the *Neoclassic* Description Logic Language, since it gives us most of the representation power available in semantic networks.

```
(createConcept Lexical_Objects ClassicThing)
(createDisjoinGroup Lexical_Objects ole_dgroup)
(createRole spelling true)
(createRole senses)
(createConcept Words
  (and Lexical_Objects
       (all spelling String)
       (all senses Lexical_Objects)
       (...))
  ole_dgroup)
(...)
```

Figure 4. General Concept Model description in *Neoclassic*.

Figure 4 shows the *Neoclassic* code that describes some of the entities and relations of the general conceptual model illustrated in Figure 2.

3.2 Content description of the sources

At this level we describe the mappings between each SCM and the GCM. We adopt the *local-as-view* paradigm where each source concept (and relationship) is mapped into the general conceptual model.

For every relation (and concept) in the local schema, a conjunctive query over the GCM must be specified. This query describes the conditions the tuples in the relation must satisfy. This approach gives us two main advantages:

- The GCM does not change when a new resource is integrated into the system. This property allows us to offer the user a unique domain and vocabulary to pose the queries. It also leaves the way open to dynamically add new databases to the system.
- We can make fine-grained distinctions between the several relations and concepts of the sources, giving to the system the possibility of integrating very diverse lexical resources.

In Figure 5 we can see some conjunctive queries that map some concepts and relations of the EH dictionary resource to the GCM.

Concepts are expressed as logical predicates with arity 1, and relationships are described by means of predicates with arity 2. So, the third query states that two individuals `s` and `l` will be related via the `EH_written_form` relation in the EH conceptual model if `s` is a `Concept` individual, `l` is a `Dictionary_Entry`, and both instances are related via the `sense` relation in the GCM.

```
EH_Sense(a) <- Concept(a).
EH_sense_id(a,s) <- Concept(a),
concept_id(a,s).
EH_written_form(s,l) <-
  Dictionary_Entry(l), sense(l,s), Concept(s).
```

Figure 5. Mappings between the CM of the EH dictionary and the GCM.

4 The query translation algorithm

The query-answering algorithm guarantees that the descriptions of the sources are exploited in such a way that only sources that are relevant to a given query are effectively accessed.

The user queries are posed in terms of the GCM relations and the system must translate them into the equivalent rewritings by means of the CM of each lexical source. The queries must be rewritten in such a way that we can obtain as many answers as possible from a set of data source descriptions.

This problem is closely related to the “answering queries using views” problem (Levy *et al.,* 1995).

We have borrowed the so-called *bucket algorithm* originally developed as part of the Information Manifold system (Levy *et al.,* 1996). The bucket algorithm finds the relevant sources for a given query *Q* and reformulates them exclusively in terms of those source models. It also guarantees that this rewritings are the *maximally-contained rewritings* that are also *sound*, i.e., all the answers they produce are guaranteed to be answers to the original query.

Conclusion

A solution for the problem of integration of diverse lexical resources has been presented. This solution aims to use data integration techniques in the field of lexical knowledge.

The architecture proposed allows the user —be it human or a computer program— to pose queries to the system regardless of the actual formats and representation schemas used in the particular resources integrated in it. This is possible due to the existence of a general conceptual model of lexical knowledge. The architecture allows the integration of new resources to the system in a dynamic way.

Currently two lexical resources have been integrated in our system: a SGMLized TEI-conformant Basque monolingual dictionary (EH) and a NLP lexical database stored in a classic relational DBMS.

References

Calvanese D., de Giacomo G., Lenzerini M., Nardi D., Rosati R. (forthcoming) *Data Integration in Data Warehousing*. Int. Journal of Cooperative Information Systems. World Scientific Information Systems.

Calvanese D., de Giacomo G., Lenzerini M. (1999) *Answering queries using views in description logics.* In, “Proc. of the 6th Int. Workshop KRDB'96”, pp. 6-10. E. Franconi and M. Kifer, eds.

Levy A.Y., Mendelzon A.O., Sagiv Y. (1995) *Answering Queries using Views*. In “Proc. of the ACM Symposium on Principles of Database Systems (PODS)”, San Jose.

Levy A.Y., Rajaraman A., Ordille J.J. (1996) *Query-Answering Algorithms for Information Agents*. In “Proc. of the 13th National Conference on Artificial Intelligence (AAAI-96)”, Portland, Oregon.

Patrick J., Zhang J., Artola X. (2000) *An Architecture and Query Language for a Federation of Heterogeneous Dictionary Databases*. Computers and the Humanities 34 (4), pp. 393-407.

Roth M., Schwartz P.M. (1997) *Don't Scrap it, Wrap it!, A Wrapper Architecture for Legacy Data Stores*. In “Proc. of 23rd Int. Conference on Very Large Databases (VLDB'97)”, pp. 266-275. August 25-29, Athens (Greece).