# Automatic annotation for mapping discovery in data integration systems *

Sonia Bergamaschi, Laura Po, Serena Sorrentino

Dipartimento di Ingegneria dell'Informazione
Universitá di Modena e Reggio Emilia
bergamaschi.sonia@unimore.it, po.laura@unimore.it, serena.sorrentino@dbgroup.unimo.it

## 1 The Combined Word Sense Disambiguation algorithm

We propose a CWSD (Combined Word Sense Disambiguation) algorithm for the automatic annotation of structured and semi-structured data sources. Rather than being targeted to textual data sources like most of the traditional WSD algorithms found in the literature, our algorithm can exploit information coming from the structure of the sources together with the lexical knowledge associated with the terms (elements of the schemata).

We integrated CWSD in the MOMIS system (Mediator EnvirOment for Multiple Information Sources) [1], which is an I3 framework designed for the integration of data sources, where the lexical annotation of terms was performed manually by the user. CWSD combines a structural disambiguation algorithm that starts the disambiguation of the terms using the semantic relationships extracted from the schemata structural relationships with a WordNet Domains based disambiguation algorithm to refine terms disambiguation by using domains information.

Structural relationships are stored in a Common Thesaurus (CT) generated by the MOMIS system. The CT is a set of relationships describing inter- and intra-schema knowledge among the source schemas. From a source schema we extract the following relationships: SYN (Synonym-of), defined between two terms (term is the name of a class/attribute of a schema) that are considered synonyms/equivalent; BT (Broader Terms), defined between two terms such as the first one is more general than the second one (the opposite of BT is NT, Narrower Terms); RT (Related Terms) defined between two terms that are generally used together in the same context.

The extracted $ODL_{I^3}$ relationships can be used in the disambiguation process according to a lexical database (in our approach we used WordNet). The algorithm tries to find a lexical relationship when a CT relationship exists among two terms; in this case we choose the meanings connected by this relationship as the correct ones to disambiguate the terms. The same holds if we find a chain of lexical relationships that connect terms meanings.

The WordNet Domains disambiguation algorithm exploits the information from WordNet Domains. WordNet Domains [2] can be considered an extended version of

* This work was partially supported by MUR FIRB Network Peer for Business project (http://www.dbgroup.unimo.it/nep4b) and by the IST FP6 STREP project 2006 STASIS (http://www.dbgroup.unimo.it/stasis).

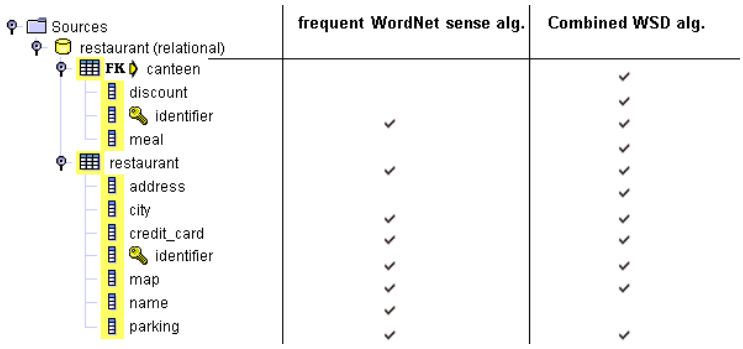


**Fig. 1.** Evaluation of the purposed WSD algorithms

WordNet, (or a lexical resource) in which synsets have been annotated with one o more domain labels. The hypotheses is that, domain labels provide a useful way to establish semantic relations among word senses, and this can be profitably used during the disambiguation process. [1]

## 2 An application of the algorithms

Figure 1 analyses an example of the application of the purposed algorithms. We have chosen a relational source composed of two different tables connected by a structural relationship (foreign key). Next to the Figure 1 we evaluated the right senses supplied by the different disambiguation approaches. The first approach is already used in MOMIS and chooses the more frequent WordNet sense as the correct meaning for a term. The second algorithm is the CWSD that combined the structural disambiguation algorithm and the WordNet Domains disambiguation algorithm. The structural disambiguation algorithm exploits the structural relationship "foreign key" to define the correct meaning of the class terms: "restaurant" and "canteen". Then the WordNet Domains disambiguation algorithm calculates the prevalent domains over the entire set of terms and compares these domains with the ones associated to each term for determine the correct meaning.

## References

1. S. Bergamaschi, S. Castano, D. Beneventano, and M. Vincini. Semantic integration of heterogeneous information sources. *Journal of Data and Knowledge Engineering*, 36(3):215–249, 2001.
2. A. M. Gliozzo, C. Strapparava, and I. Dagan. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. *Computer Speech & Language*, 18(3):275–299, 2004.

[1] A detail description of the two algorithms is available at http://www.dbgroup.unimo.it/momis/CWSD