Improving Data Integration through Disambiguation Techniques*

Laura Po

Dipartimento di Ingegneria dell'Informazione
Universitá di Modena e Reggio Emilia
laura.po@unimore.it

Abstract. In this paper Word Sense Disambiguation (WSD) issue in the context of data integration is outlined and an Approximate Word Sense Disambiguation approach (AWSD) is proposed for the automatic lexical annotation of structured and semi-structured data sources.

1 Introduction

The focus of data integration systems is on producing a comprehensive global schema successfully integrating data from heterogeneous data sources (heterogeneous in format and in structure) [8,2]. The amount of data to be integrated can be scattered at many sources and there may not be available domain experts to perform the integration process. For these reasons and for saving time and human intervention, the integration process should be as much automated as possible. Thus, in recent years, many different data-integration tools have been improved with methods for automatic discovery of mappings and thus matching among schemata.

The biggest difficulty in schema matching lays on being able to discover the right relationships among schemata from different sources. Usually, data sources are organized by developers, according to different categorization. Therefore, it is necessary to understand the modelling logic behind structuring information (i.e the *structural relationships* among schema elements). Further, it is important to deal with the problem of how the data are "labelled"; it is often hard to understand the meaning behind the names denoting schemata elements. Annotation becomes, thus, crucial to understand the meaning of schemata.

Annotation is the inclusion of extra information on a data source. The annotation process can be performed in relation to a reference, like ontology or vocabulary. During the lexical annotation process (i.e. annotation w.r.t. a vocabulary or thesaurus) the concepts and attributes of a data sources (which in the following will be called generally terms) are annotated according to a lexical reference database (WordNet[1] in this implementation, but the method is independent on this choice), in which the terms are interconnected by *lexical relationships* that can be syntactic or semantic.

The followed approach faces the disambiguation problem in the field of integrating structured and semi-structured data sources schemata. These data have got some special

* Advisor: Prof. Sonia Bergamaschi.
[1] See http://wordnet.princeton.edu for more information on WordNet.

E. Kapetanios, V. Sugumaran, M. Spiliopoulou (Eds.): NLDB 2008, LNCS 5039, pp. 372–375, 2008.

features that distinguish them from text data. On structured or semi-structured data sources there is not as a wide context as in a text source; in addition, there are less words that concur to the definition of concepts (classes, attributes, properties). The contexts on these data can be defined considering the structural and semantic relationships over a source, like sets of attributes belonging to the same class, or classes connected by aggregations or hierarchies.

As integration system, MOMIS-Ontology Builder [2] has been used; MOMIS is a framework able to create a domain ontology representing a set of selected data sources, described with a standard W3C language wherein concepts and attributes are annotated according to the lexical reference database.

2 Related Work

A semantic approach to schema matching has been proposed in [5]. The method calculates mappings between schema elements by computing semantic relationships. The semantic relations are determined by analysing the meaning which is codified in the elements and the structures of schemas. Labels at nodes are translated into propositional formulas which explicitly codify the label's intended meaning. The use of the semantic relations in this approach is similar to what will be proposed here.

Some other authors, dealing with ontology matching, have proposed a method to solve semantic ambiguity in order to filter the appropriate mappings between different ontologies [6]. Using the semantic similarity measures, the mappings found by an ontology matching tool can be filtered, so the precision of the system improves. The limit of this method is that it does not disambiguate the label of the ontology classes, while it evaluates the possible meanings only.

3 An Approximate Word Sense Disambiguation Approach

As it is described in [7] the WSD task involves two steps: (1) the determination of all the different meanings for every word under consideration; and (2) a means to assign to each occurrence of a word its appropriate meaning. The most recent works on WSD rely on predefined meanings for step (1), including: a list of meanings such as those found in dictionaries or thesauri.

The use of a well-known and shared thesaurus (in this case WordNet) provides a reliable set of meanings and allows to share with others the result of the disambiguation process. Moreover, the fundamental peculiarity of a thesaurus is the presence of a wide network of relationships between words and meanings.

The disadvantage in using a thesaurus is that it does not cover, with same detail, different domains of knowledge. Some terms may not be present or, conversely, other terms may have many associated and related meanings. These considerations and the first tests made led to the need of expanding the thesaurus with more specific terms. On the other hand, when a term has many associated and related meanings, we need to overcome the usual disambiguation approach and relate the term to multiple meanings: i.e. to union of the meanings associated to it. Even Resnik and Yarowsky [9] ratify that there are common cases where several fine-grained meanings may be correct.

The proposed AWSD approach may associate more than one meaning to a term and this, thus, differs from the traditional approaches.

Different methods to disambiguate structured and semi-structured data sources have been developed and tested (two in CWSD [3] and one in MELIS [1]). The results of the cited methods are good, even if not totally satisfying as they do not disambiguate all the terms in a data integration scenario. In [3] it is shown that the combination of methods is an effective way of improving the WSD process performance. The problem focuses on how the different method have to be combined.

Instead of concentrating in determining a unique best meaning of a term, AWSD associates a set of related meanings to a term, each one with its own reliability degree. Our idea is supported by Renisk and Yarowsky that have introduced [9] that the problem of disambiguation is not confined to search for the best meaning. They thought it is significant that a method reduces all possible meanings associated to a term, and that, within this set assigns a probability to the correct meanings.

Let us suppose that a disambiguation method, called `A`, provides two possible meanings for a term `t` (`t#1` and `t#2`[2]) while a second method, called `B`, gives `t#2` and `t#3`. In a combined approach, both methods are applied, and, eventually, the list of all its associated meanings with their probabilities is proposed. The probabilities will be associated to the methods on the basis of their reliability. So if a method is trusted, the method will have a high probability. Moreover, if a WSD method is iterative, it could obtain different meanings with changing probabilities during iteration.

In the previous example, let us suppose that the first method is more reliable than the second one and that the meanings extracted by both methods are different but equiprobable. Let us suppose, method `A` has 70% probability, while method `B` has 30%. The final result will be as shown in table 1.

So far the probability associated with a method is defined by the user who can interpret the experimental results provided by methods (for example, in [3], CWSD combines a structural disambiguation algorithm (SD), with a WordNet Domains based disambiguation algorithm (WND); the SD method has a higher precision then WND, then the user can associate to SD method a higher probability).

Table 1. Example of the AWSD output

term	**meaning**	**probability** - method `A`	**probability** - method `B`	**probability** - ASWD
t	`t#1`	50%	-	35%
t	`t#2`	50%	50%	50%
t	`t#3`	-	50%	15%

Choosing more meanings for a term means that the number of discovered lexical relationships connecting a term to other meanings in the thesaurus increase.

Let us assume that it is necessary to analyze the relationships among the terms `t`, `s` and `r`. Filtering the result by a threshold, `t#1` and `t#2` are maintained as the correct synset for the term `t`. If there is a lexical relationship between `t#1` and `s#2`,

[2] `t#2` is the meaning associated to the second sense of the word occurring in the label “t”.

and another lexical relationship between `t#2` and `r#1`, choosing both `t#1` and `t#2` for disambiguate the term `t` means that two relationship among the three terms are maintained.

The set of lexical relationships together with the set of structural relationships in a dynamic integration environment are the input of the mapping process. Enriching the input of the mapping tool means to improve the discover mappings and so to refine the global schema.

Moreover widening the use of probability to the discover relationships allows computing probabilistic mappings. Finding probabilistic mapping is the basis of managing uncertainty in data integration system. And this leads to new method of query answering like suggested in [4].

4 Conclusion

In a dynamic integration environment, annotation has to be performed automatically, i.e. without human intervention. This leads to uncertain disambiguation results i.e. more than one meaning is associated to a term with a certain probability. In this paper an approximate approach is proposed that associates to each term its meanings with a certain probability. In a lexical database the meanings are related with lexical relationships. By exploiting lexical relationships among the meanings of the terms extracted by AWSD, we can evaluate a similarity degree between the terms and compute probabilistic mappings.

References

1. Bergamaschi, S., Bouquet, P., Giacomuzzi, D., Guerra, F., Po, L., Vincini, M.: An incremental method for the lexical annotation of domain ontologies. Int. J. Semantic Web Inf. Syst. 3(3), 57–80 (2007)
2. Bergamaschi, S., Castano, S., Vincini, M.: Semantic integration of semistructured and structured data sources. SIGMOD Record 28(1), 54–59 (1999)
3. Bergamaschi, S., Po, L., Sala, A., Sorrentino, S.: Data source annotation in data integration systems. In: Proceedings of the Fifth International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P) (2007)
4. Dong, X.L., Halevy, A.Y., Yu, C.: Data integration with uncertainty. In: VLDB, pp. 687–698 (2007)
5. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Semantic schema matching. In: OTM Conferences, vol. (1), pp. 347–365 (2005)
6. Gracia, J., Lopez, V., D'Aquin, M., Sabou, M., Motta, E., Mena, E.: Solving semantic ambiguity to improve semantic web based ontology matching. In: Proceedings of the Workshop on Ontology Matching (OM 2007) at ISWC/ASWC 2007, Busan, South Korea (2007)
7. Ide, N., Véronis, J.: Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics 24(1), 1–40 (1998)
8. Lenzerini, M.: Data integration: A theoretical perspective. In: Popa, L. (ed.) PODS, pp. 233–246. ACM, New York (2002)
9. Resnik, P., Yarowsky, D.: Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. Natural Language Engineering 5(2), 113–133 (2000)