Automatic annotation for mappings discovery in data integration systems (Extended abstract) ⋆

Sonia Bergamaschi, Laura Po, and Serena Sorrentino

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

Abstract. In this article we present CWSD (Combined Word Sense Disambiguation), a method and a software tool for enabling automatic annotation with lexical information of local structured and semi-structured data sources in a data integration system. CWSD is based on the exploitation of WordNet Domains and the lexical and structural knowledge of the data sources. The method extends the lexical annotation module of the MOMIS data integration system with an automatic annotation. The distinguishing feature of the method is its independence or low dependence of a human intervention. CWSD is a valid method to satisfy two important tasks: (1) the source annotation process, i.e. the operation of associating an element of a lexical reference database (WordNet) to all source elements, (2) the discover of mappings among concepts of distributed data sources/ontologies.

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) [2–4]. The amount of data to be integrated can be distributed at many sources and it is thus difficult for an integration designer to be expert of all the data source contents. 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 extended to implement methods to support automatic discovery of mappings among data source schemata.

The highest difficulty in schema mapping discovery lays on being able to discover the *right* relationships among schemata from different sources. Usually,

⋆ This work is an extended version of [1] **Acknowledgements:** 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).

data sources are organized by many developers, according to different categorization (e.g. different collections of photos might be organised in different ways: classified according to years and then place, or, as an alternative, to people and date). Therefore, it is necessary to understand the modelling logic behind structuring information. Further, it is important to deal with the problem of how the data are "labelled", i.e. it is often difficult to understand the meaning behind the names denoting schemata elements. *Annotation* becomes, thus, crucial to understand the meaning of schemata.

Annotation, in general, is the inclusion of extra information on a data source. The annotation process can be performed in relation to a reference, like an ontology or vocabulary. The use of shared vocabularies and ontologies provides a well-defined basis for automated data integration and reuse.

This paper focuses on an automatic lexical annotation (i.e. annotation w.r.t. a vocabulary or thesaurus). During the lexical annotation process the concepts and attributes of data sources (which in the following will be called generally terms) are automatically annotated according to a lexical reference database (WordNet[1] in this implementation, but the method is independent of this choice).

The automatic annotation task is related to Word Sense Disambiguation (WSD) techniques developed in the area of Semantic Web [5].

Combination methods are an effective way of improving the WSD process performance. The idea of combining the results of different WSD methods was used in most approaches to WSD in literature [6, 7].

WordNet Domains has been proven a useful resource for WSD. In fact, it has been used in different combined WSD algorithms as proposed in [8, 9].

Following the approach of combined WSD algorithms, we developed CWSD (Combined Word Sense Disambiguation), a method and a tool for the automatic annotation of structured and semi-structured data sources. Instead of being targeted to textual data sources like most of the traditional WSD algorithms, CWSD exploits the *structure* of data sources together with the lexical knowledge associated with schema elements (terms in the following). Moreover, CWSD associates more than one meaning to a term and thus differs from the traditional WSD approaches.

In [10], we developed a software tool (MELIS) for enabling an incremental process of automatic annotation of local schemas. MELIS exploits knowledge provided by the initial annotation. CWSD overcomes MELIS as no initial annotation is needed to disambiguate the source terms.

We integrated CWSD in the I^3 framework designed for the integration of data sources, MOMIS (Mediator EnvirOment for Multiple Information Sources) [11, 4], to overcome the heavy user involvement in manual lexical annotation of data source terms.

The outline of the paper is the following: Section 2 describes the CWSD tool and its components. In Section 3 we evaluate its performance. Finally we sketch out some conclusions and future works.

[1] See `http://wordnet.princeton.edu` for more information on WordNet.

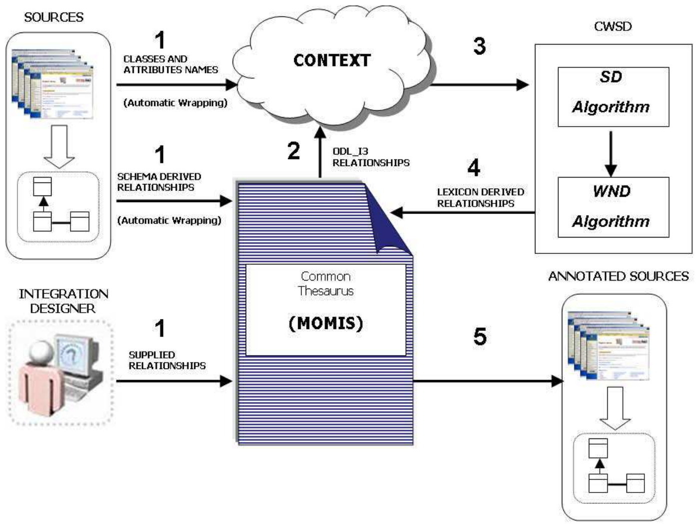


Fig. 1. Automatic annotation of local data sources with CWSD

2 The Combined Word Sense Disambiguation method

In order to disambiguate the sense of an ambiguous word, any WSD algorithm receives as input (and works in) a *context*. According to [12], many algorithms in literature represent the context as a "bag-of-words", a set of words that must be disambiguated, and sometime they insert in the context the information of the word positions in the text. Other approaches [13] consider a "window-of-context" around every target word, and submit all the words in this window as input to the disambiguation algorithm.

In CWSD the context is composed by: a set of terms (classes and attributes names) to be disambiguated, and a set of structural relationships among these terms included in a Common Thesaurus (CT) (as shown in figure 1). The default context is given by all the terms in the data source to be integrated and the structural ODL_{I^3} relationships among these terms.

The CT is a set of ODL_{I^3} relationships describing inter- and intra-schema relationships among a set of data source schemas. The ODL_{I^3} relationships can be structural or lexicon derived.

The structural ODL_{I^3} relationships are:
SYN_{EXT}: `t1` is equivalent to `t2` iff extension(t1) = extension(t2);
BT_{EXT}: `t1` subsumes `t2` iff extension(t1) $\supseteq$ extension(t2);
RT_{EXT}: `t1` is related to `t2` iff `t1` is a property of `t2` or `FK: t1 REFERENCES TABLE(t2)`.

These relationships are automatically extracted by the MOMIS wrapper and ODB-Tools [14].

The lexical ODL_{I^3} relationships are:
SYN: (Synonym-of), defined between two terms that are equivalent/ synonymous;
BT: (Broader Term), defined between two terms where the first generalizes the second (the opposite of BT is NT, Narrower Term);
RT: (Related Term) defined between two terms that are related in an aggregation hierarchy.

The use of a well-known and shared lexical database (in this case WordNet [15]) provides a reliable set of meanings and allows to be shared with others the result of the disambiguation process, especially if the lexical resource is freely and publicly available (as WordNet is). Moreover, the fundamental peculiarity of a lexical database like WordNet is the presence of a wide network of semantic relationships between words and meanings (SYN,BT,RT).

The disadvantage in using a lexical database is that it does not cover with the same detail different domains of knowledge. Some terms may not be present in the thesaurus 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 lexical database with more specific terms (in this case, the MOMIS system already includes a component, *WNEditor*, which allows adding new terms and linking them within WordNet [16]). On the other hand, when a term have 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.

CWSD is composed of two algorithms: SD (Structural Disambiguation) and WND (WordNet Domains Disambiguation). SD tries to disambiguate source terms by using semantic relationships inferred from the structure of data sources and WND tries to disambiguate the terms using domains information supplied by WordNet Domains.

Both the proposed algorithms, that we will describe in the following, may associate more than one meaning to a term and they, thus, differ from the traditional WSD approaches.

2.1 The Structural Disambiguation algorithm

The SD algorithm exploits the structural ODL_{I^3} relationships of a data source to infer ODL_{I^3} relationships on the basis of WordNet. As described in [11] the following ODL_{I^3} relationships are automatically extracted:

- For an ISA relationship between two classes (like `T1` ISA `T2`) we extract a BT_{EXT} relationship: `T2` BT_{EXT} `T1` (`T1` NT_{EXT} `T2`)
- For a foreign key (FK) between two relations:
  `T1(A1,A2...AN) T2(B1,B2...BM) FK: B1 REFERENCES T1(A1)`
  we infer `A1` SYN_{EXT} `B1`
  and if B1 is a key on table T2: `T1` BT_{EXT} `T2` (`T2` NT_{EXT} `T1`)
  else: `T1` RT_{EXT} `T2`

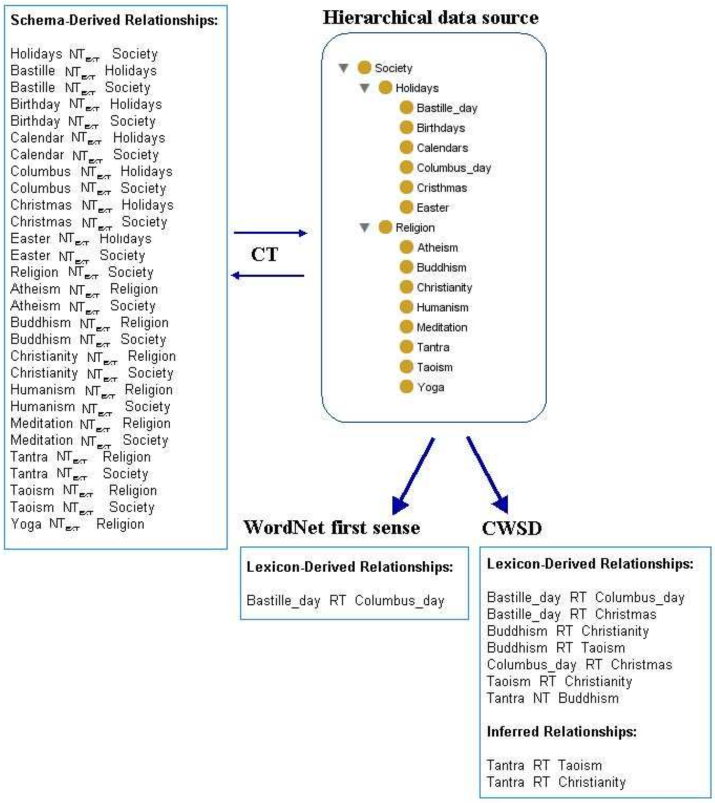


Fig. 2. Enrichment of the CT with the lexical ODL_{I^3} relationships extracted by CWSD applied to a hierarchical data source

The extracted relationships are stored in the CT to infer lexical ODL_{I^3} relationships.

SD tries to find a corresponding lexical relationship when a structural relationship holds among two terms. In practice, if we have a direct/chain of relationship between two terms, we try to find the semantically related meanings and annotate the terms with these meanings. A chain of relationships is obtained navigating through the lexical database relationships.

For all the NT_{EXT} relationships, SD finds the corresponding NT relationships in WordNet. The relation of equivalence (SYN_{EXT}) is used to find the corresponding SYN relationship in the lexical database. The RT_{EXT} relationship is used to find holonym or meronym relationships (RT) in the lexical database.

Figure 2 shows an example of the application of the SD algorithm on a hierarchical data source, i.e. a portion of the first three levels of the "society" subtree in the Google directory. The annotations generated using SD enrich the CT of new ODL_{I^3} relationships (all the lexicon-derived relationships shown in figure). Using ODB-Tools (a component of the MOMIS system [14]) the CT inferred new relationships. As you can see, the amount of relationships that are obtained by applying SD is greater than those obtained by using WordNet first sense.

2.2 The WordNet Domains algorithm

WordNet Domains [17]can be considered an extended version of WordNet, (or a lexical resource) in which synsets have been annotated with one or more domain labels. The information brought by domains is complementary with the one already present in WordNet. Besides, domains may group senses of the same word, into a thematic cluster, which has the important side effect of reducing the level of ambiguity of polysemic words.

The WND algorithm takes inspiration from the domain-based one proposed in [18]. First, we examine all the possible synsets connected to a term and extract the domains associated to these synsets, with this information we calculate a list of the *prevalent domains* in the chosen context. Then, we compare the *prevalent domains* with the lists of domains associated to each term. For a term we choose as the correct synsets all the synsets associated to the prevalent domains.

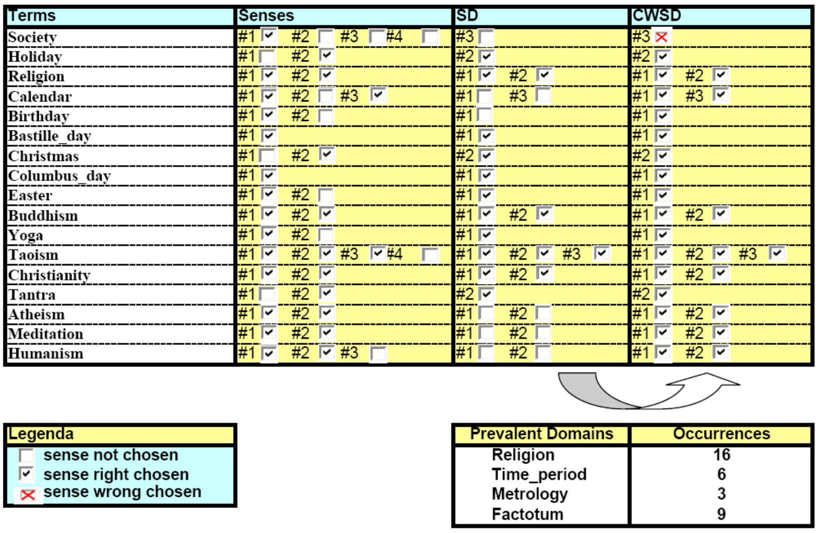
In WordNet Domains there is a particular domain called "factotum" which is the domain associated to synsets that do not belong to a specific domain and in most of the cases it is the more frequent domain in a context. In accordance with [19], we choose not to use the "factotum" domain to disambiguate the terms, because it is too generic.

We calculate the most frequent domains in a context and, if a term does not have any synset related to one of these domains, we choose the first WordNet sense.

WND results depends on the context and on the *configuration* chosen. The configuration is the maximum number of domains we select for the disambiguation. The choice of the configuration and of the context (if not default) is delegated to the user.

In Figure 3 we showed the final result of the application of CWSD to the hierarchical data source. In particular, we compare the result obtained with CWSD with the result obtained using only the SD algorithm. If we disambiguate by using only SD, we obtain the correct senses for only some terms. With the CWSD algorithm we improved the results in two directions:(1) the disambiguation of the terms is more accurate; polysemy leads to have more than one synset associated to a terms, thanks to CWSD we can assign to these terms more than one sense; (2) moreover, CWSD enriches the CT of new relationships: this is particularly important for the integration task (like showed in Figure 2). The unique term annotated in a wrong way is "Society", this is because it is associated, by the WND algorithm, to the "factotum" domain, but the correct sense is associated to the "anthropology" domain that is not present in the prevalent domains.[2]

[2] A detailed description of the CWSD algorithm is available at http://www.dbgroup.unimo.it/Momis/CWSD/



Terms	Senses	SD	CWSD
Society	#1 ☑ #2 ☐ #3 ☐ #4 ☐	#3 ☐	#3 ✗
Holiday	#1 ☐ #2 ☑	#2 ☑	#2 ☑
Religion	#1 ☑ #2 ☑	#1 ☑ #2 ☑	#1 ☑ #2 ☑
Calendar	#1 ☑ #2 ☐ #3 ☑	#1 ☐ #3 ☐	#1 ☑ #3 ☑
Birthday	#1 ☑ #2 ☐	#1 ☐	#1 ☑
Bastille_day	#1 ☑	#1 ☑	#1 ☑
Christmas	#1 ☐ #2 ☑	#2 ☑	#2 ☑
Columbus_day	#1 ☑	#1 ☑	#1 ☑
Easter	#1 ☑ #2 ☐	#1 ☑	#1 ☑
Buddhism	#1 ☑ #2 ☑	#1 ☑ #2 ☑	#1 ☑ #2 ☑
Yoga	#1 ☑ #2 ☐	#1 ☑	#1 ☑
Taoism	#1 ☑ #2 ☑ #3 ☑ #4 ☐	#1 ☑ #2 ☑ #3 ☑	#1 ☑ #2 ☑ #3 ☑
Christianity	#1 ☑ #2 ☑	#1 ☑ #2 ☑	#1 ☑ #2 ☑
Tantra	#1 ☐ #2 ☑	#2 ☑	#2 ☑
Atheism	#1 ☑ #2 ☑	#1 ☐ #2 ☐	#1 ☑ #2 ☑
Meditation	#1 ☑ #2 ☑	#1 ☐ #2 ☐	#1 ☑ #2 ☑
Humanism	#1 ☑ #2 ☑ #3 ☐	#1 ☐ #2 ☐	#1 ☑ #2 ☑

Legenda
☐ sense not chosen
☑ sense right chosen
✗ sense wrong chosen

Prevalent Domains	Occurrences
Religion	16
Time_period	6
Metrology	3
Factotum	9

Fig. 3. Evaluation of the CWSD algorithm on a hierarchical data source

3 Evaluation: experimental result

We experimented CWSD over a real set of data sources. In particular, we selected the first three levels of a subtree of the Yahoo and Google directories ("society and culture" and "society", respectively), which amounts to 327 categories for Yahoo and 408 for Google.

In table 1 we compare the disambiguation of the subtree of the Google and Yahoo directories obtained with different algorithms: only SD, only WND, CWSD and MELIS.

The MELIS algorithm is incremental, so the evaluation is done after a number of runs until a fixed point has been reached. We compared CWSD results with the ones in MELIS that start with no annotations at all.

The annotation results have been evaluated in terms of recall (the number of correct annotations made by the algorithm divided by the total number of annotations, i.e. one for each category, as defined in a golden standard) and precision (the number of correct annotations retrieved divided by the total number of annotations retrieved). In the table, the recall and precision values are obtained by considering an element as properly annotated if the annotation given by the user is included in the set of annotations calculated by the WSD approach evaluated.

The application of SD over the web directories exploits the ISA relationships (792) and allows to obtain 60 annotations of which 58 are correct annotations, so we deduce a high precision but a very low recall (8.0%). For our experience this is caused by the incompleteness of semantic WordNet relationships.

WSD approach	**Recall**	**Precision**
SD	8.00%	97.00%
WND	66.62%	69.97%
CWSD	74.18%	74.18%
MELIS	53.03%	58.85%

Table 1. Comparing the different WSD approaches on the Google and Yahoo directories

The results remark that a combined algorithm outperforms the single algorithm of which it is composed[3]. Moreover the results gained by CWSD improve the ones obtained by MELIS.

4 Conclusion and future work

In this paper we presented a combined algorithm for the automatic annotation of structured and semi-structured data sources. CWSD exploits structural knowledge of a set of data sources together with the lexical knowledge supplied by WordNet & WordNet Domains lexical databases, to automatically annotate data source schemata.

We automatically extracted schema-derived relationships from the data sources using the ODB-Tools component of the MOMIS system and inserted them in the MOMIS Common Thesaurus. In the first step, the SD algorithm infers lexical meanings for terms from WordNet and the structural ODL_{I^3} relationships stored in the Common Thesaurus. In the second step, the WND algorithm refines terms disambiguation using domain information supplied by WordNet Domains. The experimental results showed the effectiveness of CWSD. Moreover, the structural knowledge of data sources significantly improves the disambiguation results (i.e. enhances the WND algorithm results).

Future work will be devoted to inserting a data cleansing step before the application of CWSD. In fact, CWSD cannot be used for sources that include acronyms/abbreviations or compound terms.

Other research will investigate the role of the context choice in our algorithm and determine a criteria to choose the best number of domains during the configuration of WND.

References

1. Bergamaschi, S., Po, L., Sorrentino, S.: Automatic annotation in data integration systems. In Meersman, R., Tari, Z., Herrero, P., eds.: OTM Workshops (1). Volume 4805 of Lecture Notes in Computer Science., Springer (2007) 27–28

[3] In this evaluation we do not discuss the chosen configuration, because in general this is delegated to the user; however the showed results have been obtained on a limited context that considers together the terms of the classes that are related with an ISA relationship and with the best choice for the number of domains.

2. Lenzerini, M.: Data integration: A theoretical perspective. In Popa, L., ed.: PODS, ACM (2002) 233–246
3. Halevy, A.Y.: Data integration: A status report. In Weikum, G., Schöning, H., Rahm, E., eds.: BTW. Volume 26 of LNI., GI (2003) 24–29
4. Bergamaschi, S., Castano, S., Vincini, M.: Semantic integration of semistructured and structured data sources. SIGMOD Record **28(1)** (1999) 54–59
5. Bechhofer, S., Carr, L., Goble, C.A., Kampa, S., Miles-Board, T.: The semantics of semantic annotation. In Meersman, R., Tari, Z., eds.: CoopIS/DOA/ODBASE. Volume 2519 of Lecture Notes in Computer Science., Springer (2002) 1152–1167
6. Rigau, G., Atserias, J., Agirre, E.: Combining unsupervised lexical knowledge methods for word sense disambiguation. CoRR **cmp-lg/9704007** (1997)
7. Mihalcea, R., Moldovan, D.I.: An iterative approach to word sense disambiguation. In Etheredge, J.N., Manaris, B.Z., eds.: FLAIRS Conference, AAAI Press (2000) 219–223
8. Gliozzo, A.M., Giuliano, C., Strapparava, C.: Domain kernels for word sense disambiguation. In: ACL, The Association for Computer Linguistics (2005)
9. Novischi, A.: Combining methods for word sense disambiguation of wordnet glosses. In Barr, V., Markov, Z., eds.: FLAIRS Conference, AAAI Press (2004)
10. 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) (2007) 57–80
11. Bergamaschi, S., Castano, S., Beneventano, D., Vincini, M.: Semantic integration of heterogeneous information sources. Journal of Data and Knowledge Engineering **36**(3) (2001) 215–249
12. Pahikkala, T., Pyysalo, S., Ginter, F., Boberg, J., Järvinen, J., Salakoski, T.: Kernels incorporating word positional information in natural language disambiguation tasks. In Russell, I., Markov, Z., eds.: FLAIRS Conference, AAAI Press (2005) 442–448
13. Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordnet. In Gelbukh, A.F., ed.: CICLing. Volume 2276 of Lecture Notes in Computer Science., Springer (2002) 136–145
14. Bergamaschi, S., Sartori, C., Beneventano, D., Vincini, M.: Odb-tools: A description logics based tool for schema validation and semantic query optimization in object oriented databases. In Lenzerini, M., ed.: AI*IA. Volume 1321 of Lecture Notes in Computer Science., Springer (1997) 435–438
15. Fellbaum, C., Miller, G., eds.: WordNet: An electronic lexical database. The MIT Press (1998)
16. Benassi, R., Bergamaschi, S., Fergnani, A., Miselli, D.: Extending a lexicon ontology for intelligent information integration. In de Mántaras, R.L., Saitta, L., eds.: ECAI, IOS Press (2004) 278–282
17. Gliozzo, A.M., Strapparava, C., Dagan, I.: Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. Computer Speech & Language **18**(3) (2004) 275–299
18. Magnini, B.: Experiments in word domain disambiguation for parallel texts (November 20 2000)
19. Buscaldi, D., Rosso, P., Masulli, F.: Integrating conceptual density with wordnet domains and cald glosses for noun sense disambiguation. In González, J.L.V., Martínez-Barco, P., Muñoz, R., Saiz-Noeda, M., eds.: EsTAL. Volume 3230 of Lecture Notes in Computer Science., Springer (2004) 183–194