Automatic Adaptation of WordNet to Domains

Roberto Navigli, Paola Velardi

Università di Roma "La Sapienza", Dipartimento di Scienze dell'Informazione, Via Salaria 113
00198 Roma, Italy, e-mail: velardi@dsi.uniroma1.it

Abstract

The objective of this paper is to present a method to automatically enrich WordNet with sub-trees of concepts in a given language domain. WordNet is then trimmed to reduce unnecessary ambiguity and singleton nodes. The process is based on the use of statistical method and linguistic processing to extract candidate domain *terms*. Multiword terms are semantically disambiguated and interpreted using ontological and contextual knowledge stored in WordNet on singleton words.

1. Introduction

As already pointed out by many researchers, WordNet is a very useful tool, but has some important drawbacks, namely, over-ambiguity and lack of domain terminology. Several published studies attempted to solve this problem in some automatic way, for example, (Vossen, 2001) (Harabagiu et al., 1999) (Milhalcea et al., 2001) and (Agirre et al. 1999). Other studies related to the work presented in this paper deal with the more general issue of automatic ontology construction. These contributions are collected in the web proceedings of two workshops dedicated to Ontology learning, (ECAI-OL, 2000) and (IJCAI-OL, 2001).

In many described approaches for ontology learning, domain terms are firstly extracted using a variety of statistical methods; then, taxonomic relations and other types of relations between terms are detected. In the literature, the notion of domain *term* and domain *concept* are used interchangeably, though no semantic interpretation of terms takes place. For example, in (Vossen, 2001) the "concept" *digital printing technology* is considered as a kind-of *printing technology* by virtue of simple string inclusion. However, *printing* has four senses in WordNet, and *technology* has two senses. There are hence 8 possible concept combinations for *printing technology*!

In this paper we propose a method for semantic interpretation of terms, using the information available in WordNet for the individual words that appear in a terminological string. Semantic interpretation allows us to detect non-trivial taxonomic relations between *concepts*, and other types of semantic relations.

The method described in this paper is implemented in a system called OntoLearn. OntoLearn is part of an Ontology Engineering architecture, described in (Missikoff et al., 2002), developed in the context of two European projects[1], aimed at improving interoperability in the Tourism sector.

Taxonomic information is extracted from the documents available in the considered domain in 5 steps: domain terminology is identified (section 2) and structured in syntactic trees (section 3), terms are mapped to concepts (section 4), that are arranged in a domain concept forest (section 5), and then used to create a domain-specific view of WordNet (section 6).

[1] ITS – 13015 (FETISH) and ITS- 29329 (HARMONISE).

2. Identification of Relevant Domain Terminology

The objective of this phase is to extract from the available documents a domain terminology. First, we use a linguistic processor, ARIOSTO[2], to extract from a corpus of documents a list of syntactically plausible terminological patterns, e.g. compounds (*credit card*), prepositional phrases (*board of directors*), adjective-noun relations (*manorial house*).

Then, two information theory based measures are used to filter out non-terminological (e.g. *last week*) and non-domain specific terms (e.g. *world wide web* in a Tourism domain). The first measure, called *Domain Relevance*, computes the probability of occurrence of a candidate term in the application domain (e.g. Tourism), as compared with other corpora that we use for a contrastive analysis (e.g. Medicine, Economy, Novels, etc.). The second measure, called *Domain Consensus*, computes the entropy of the probability of seeing a candidate term across the documents of the application domain. The underlying idea is that only terms that are *frequently* and *consistently* referred in the available domain documents reflect some *consensus* on the use of that term. These two measures have been formally defined and extensively evaluated in (Velardi et al, 2001).

3. Generation of Syntactic Trees

From the list of filtered terminology we generate *lexicalized trees*, on the basis of a simple inclusion relation. For example, given two strings x and wx (e.g. *telephone service* and *service*), we generate $wx \rightarrow^{@} x$, where '$\rightarrow^{@}$' stands for the hyperonymy relation. Figure 1 provides an example of a generated lexicalized tree $\Im$**.** It is clear that many taxonomic relations are not captured by this simple inclusion mechanism, like *bus service* $\rightarrow^{@}$ *public transport service*.

4. Semantic Disambiguation of Terms

The process of *semantic interpretation* is one that associates to each multiword term $t = w_n \ldots w_2\ w_1$ (where w_i is an atomic word) the appropriate *concept name*.

[2] ARIOSTO is a joint effort of the Universities of Roma "La Sapienza" and "Tor Vergata".

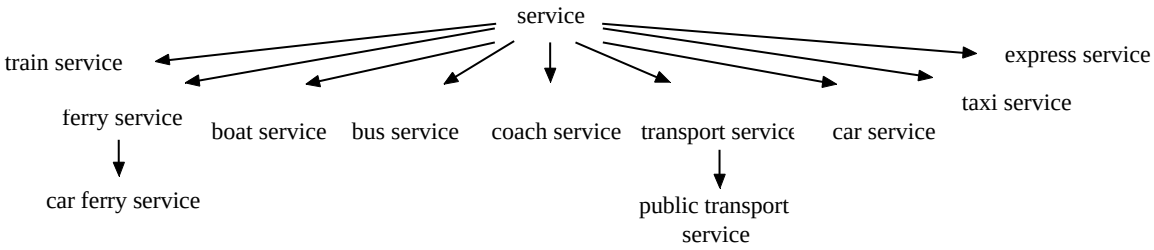

Figure 1. Example of a lexicalized tree.

Though complex terms are usually absent in WordNet, singleton words and occasionally word pairs included in a terminological string are mostly present. For example, *printing technology* as a unique term is not included, but *printing* and *technology* have an associated WordNet entry.

We derive the meaning of a complex terminological string *compositionally*, as explained hereafter.

Formally, a *semantic interpretation* is defined as follows: let $t = w_n \ldots w_2\, w_1$ be a valid term belonging to a lexicalized tree $\Im$. The process of *semantic interpretation* is one that associates to each word w_k in t the appropriate WordNet synset S_i^k, the i-th synset ($i \in \{1,...,m\}$) associated to w_k in WordNet. The *sense* of t is hence defined as:

$$S(t) = \bigcup_k S^k,\ S^k \in Synsets(w_k) \text{ and } w_k \in t.$$

where $Synsets(w_k)$ is the set of synsets each representing a sense of the word w_k.

For instance: S("transport company") = { { transportation#4, shipping#1, transport#3 }, { company#1 } } corresponding to sense #1 of *company* ("*an institution created to conduct business*") and sense #3 of *transport* ("*the commercial enterprise of transporting goods and material*").

In order to disambiguate the words in a term $t = w_n \ldots w_2\, w_1$ we proceed as follows:

a) If t is the first analyzed element of $\Im$, manually disambiguate the root node (w_1 if t is a compound) of $\Im$.

b) For any $w_k \in t$ and any synset S_i^k of w_k, create a *semantic net SN*. Semantic nets are automatically created using the following semantic relations: hyperonymy ($\rightarrow^{@}$), hyponymy ($\rightarrow^{\sim}$), meronymy ($\rightarrow^{\#}$), holonymy ($\rightarrow^{\%}$), pertonymy ($\rightarrow^{\backslash}$), attribute ($\rightarrow^{=}$), similarity ($\rightarrow^{\&}$), gloss ($\rightarrow^{gloss}$) and topic ($\rightarrow^{topic}$). The *gloss* and the *topic* relation are obtained parsing with ARIOSTO the WordNet concept definitions (*glosses*) and SemCor sentences (*topic*) including that sense. Every other relation is directly extracted from WordNet. To reduce the dimension of a SN, concepts at a distance of more than 3 relations from the SN centre, S_i^k, are removed. Figure 2a is an example of SN generated for sense #1 of *room*.

Let then $SN(S_i^k)$ be the semantic network for sense i of word w_k.

c) Starting from the "head" w_1 of t, and for any pair of words w_{k+1} and w_k (k=1,…,n-1) belonging to t, intersect alternative pairs of SNs. Let $I = SN(S_i^{k+1}) \cap SN(S_j^k)$ be one of such intersections for sense i of word w_{k+1} and sense j of word w_k. Note that, in each step k, the word w_k is already disambiguated, either manually (for k=1) or as a result of step k-1.

To identify common semantic patterns several heuristic rules are used, e.g.:

$$G, M \in Synset_{wn} : S_1 \overset{gloss}{\rightarrow} G \overset{@}{\rightarrow} M \overset{\leq 3}{\rightarrow} S_2 \overset{\leq 3}{\rightarrow} @$$

The heuristic (named "gloss+parallelism") reads: "*given two central concepts S_1 and S_2, there exist two concepts G and M such that G appears in the gloss of S_1 and both G and S_2 reach the concept M in $SN(S_1) \cap SN(S_2)$ through a hyperonimy path.*

An example is the bold pattern in Figure 2b:

transport#3 $\overset{gloss}{\rightarrow}$ *enterprise*#2 $\overset{@}{\rightarrow}$ *organization*#1 $\overset{@}{\rightarrow}$ *company*#1.

5. Creating a Domain Concept Forest

Initially, all the terms in a tree $\Im$ are independently disambiguated. Subsequently, taxonomic information in WordNet is used to detect *is-a* relations between *concepts*, e.g. *ferry service* $\rightarrow^{@}$ *boat service*. In this phase, since all the elements in $\Im$ are jointly considered, some interpretation errors produced in the previous disambiguation step are corrected. In addition, certain concepts are *fused* in a unique concept name on the basis of pertonimy, similarity and synonymy relations (e.g. respectively: *manor house* and *manorial house*, *expert guide* and *skilled guide*, *bus service* and *coach service*).

Notice again that we detect semantic relations between *concepts*, not words. For example, *bus#1* and *coach#5* are synonyms, but this relation does not hold for other senses of these two words. Each lexicalized tree $\Im$ is finally transformed in a *domain concept* tree Υ.

Figure 3 shows the concept tree obtained from the lexicalized tree of Figure 1.

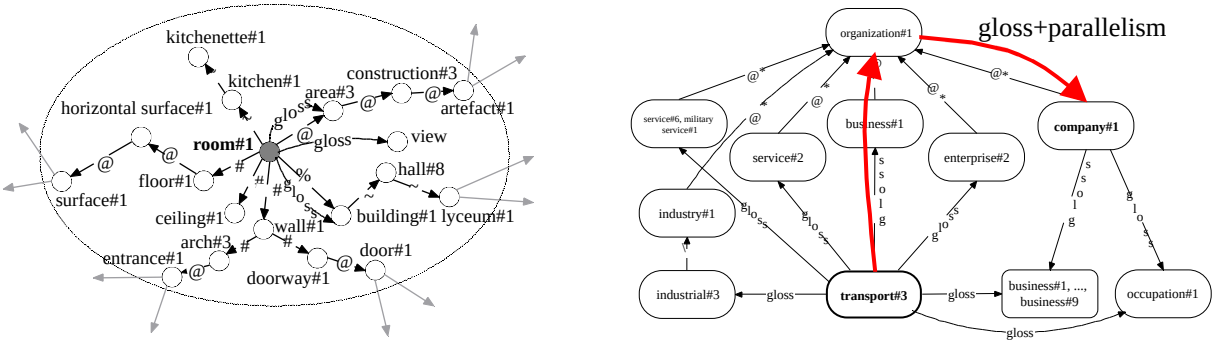

Figure 2. a) example of semantic net for *room#1*; b) example of intersecting semantic patterns for transport#3 and *company#1*.

For clarity, in Figure 3 concepts are labeled with the associated terms (rather than with synsets), and numbers are shown only when more than one semantic interpretation holds for a term, as for *coach service* and *bus service* (e.g. sense #3 of "bus" refers to "old cars").

6. Pruning and Trimming WordNet

The final phase consists in creating a domain-specialization of WordNet. In short, WordNet pruning and trimming is accomplished as follows:

1. The Domain Concept trees are attached under the appropriate nodes in WordNet.
2. An intermediate node in WordNet is pruned whenever the following conditions hold together:

   i. it has no "brother" nodes;
   ii. it has only one direct hyponym;
   iii. it is not the root of a Domain Concept tree;
   iv. it is not at a distance ≤ 2 from a WordNet *unique beginner* (this is to preserve a "minimal" top ontology).

Figure 4 shows an example of pruning the nodes located over the Domain Concept tree with root *wine#1*. Appendix A shows an example of domain-adapted branch of WordNet in the tourism domain.

7. Evaluation

OntoLearn is a knowledge extraction system aimed at improving human productivity in the time-consuming task of building a domain ontology. Our experience in building a tourism ontology for the European project Harmonise reveals that, after one year of ontology engineering activities, the tourism experts were able to release the most general layer of the tourism ontology, comprising about 300 concepts.

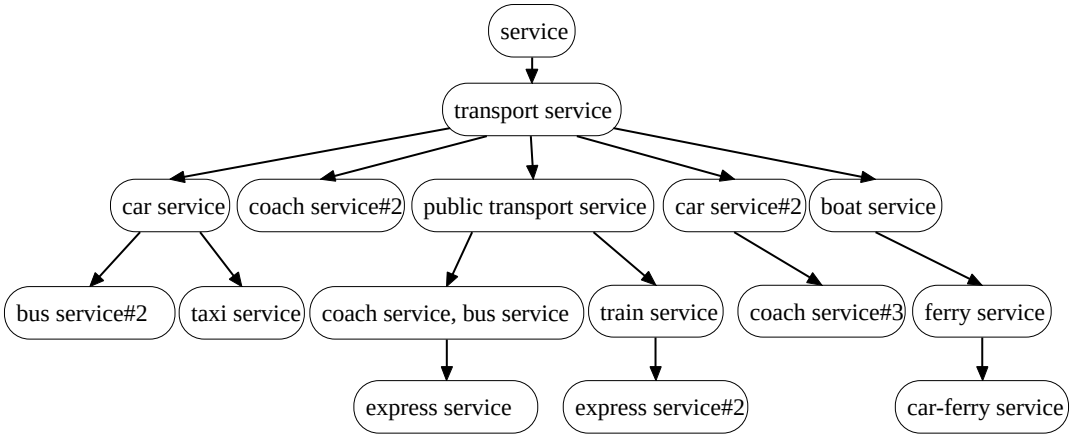

Figure 3. A Domain Concept Tree.

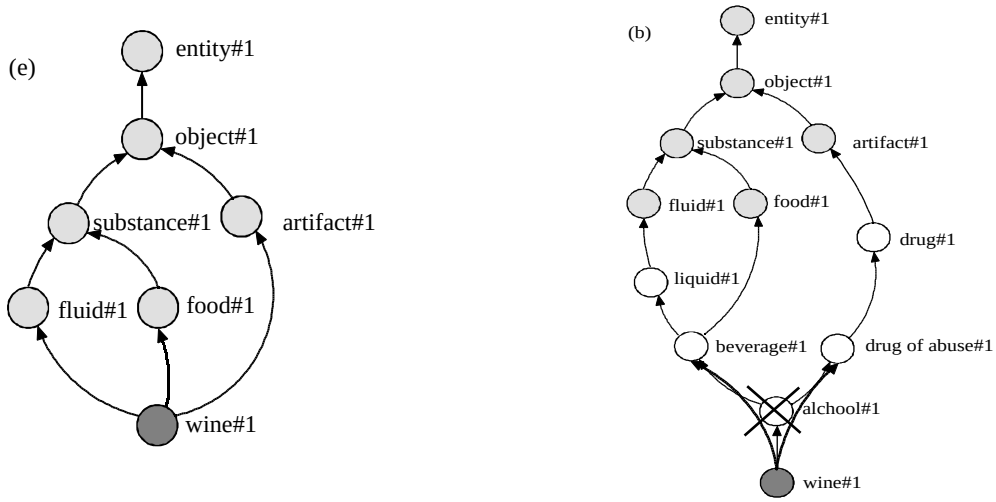

Figure 4. An intermediate step and the final pruning step over the Domain Concept Tree for "wine#1".

Then, we decided to speed up the process developing the *OntoLearn* system, aimed at supporting the ontology engineering tasks. This produced a significant acceleration in ontology building, since in the next 6 months[3] the tourism ontology reached about 3,000 concepts.

The OntoLearn system has been also evaluated independently from the ontology engineering process. We extracted from a 1 million-word corpus of travel descriptions (downloaded from Tourism web sites) a terminology of 3840 terms, manually evaluated[4] by domain experts participating in the Harmonise project. We obtained a precision ranging from 72.9% to about 80% and a recall of 52.74%. The precision shift is motivated by the well-known fact that the intuition of experts may significantly differ.

After this expert evaluation, we added few *ad hoc* heuristics that brought the precision to 97%. However, the use of heuristics limits the generality of the method.

The recall has been estimated by submitting a list of 6000 syntactic candidates to the experts, requiring them to mark truly terminological entries, and then comparing this list with that obtained by our statistical filtering method described in section 2.

We personally evaluated the semantic disambiguation algorithm using a test bed of about 650 extracted terms, which have been manually assigned to the appropriate WordNet concepts. These terms contributed to the creation of 90 syntactic trees. The entire process of semantic disambiguation and creation of domain trees has been evaluated, leading to an overall 84.5% precision. The precision grows to about 89% for highly structured sub-trees, as those in Figure 3. In fact, the phase described in section 5 significantly contributes at eliminating disambiguation errors (in the average, 5% improvement). We also analyzed the individual contribution of each of the heuristics mentioned in section 4 to the performance of the method, but a detailed performance report is omitted here for sake of space. The results of this performance analysis led to a refinement of the algorithm and the elimination of one heuristic.

[3] The time span includes also the effort needed to test and tune OntoLearn. Manual verification of automatically acquired domain concepts actually required few days.

[4] Here manual evaluation is simply deciding whether an extracted term is relevant, or not, for the tourism domain.

8. References

Agirre E., Ansa O., Hovy E. and Martinez D. *Enriching very large ontologies using the WWW*, in (ECAI-OL 2000).

Harabagiu S., Moldovan D. *Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text*. AAAI/MIT Press, 1999.

Milhalcea R., Moldovan D. I. *eXtended WordNet: progress report.* NAACL 2001 Workshop, Pittsburg, June 2001.

Missikoff M., Velardi P. and Fabriani P. *Using Text Processing Techniques to Automatically enrich a Domain Ontology*. Proc. of ACM Conf. On Formal Ontologies and Information Systems, ACM_FOIS, Ogunquit, Maine, October 2002.

Velardi P., Missikoff M. and Basili R. *Identification of relevant terms to support the construction of Domain Ontologies*. ACL-EACL Workshop on Human Language Technologies, Toulouse, France, July 2001.

Vossen P. *Extending, Trimming and Fusing WordNet for technical Documents*, NAACL 2001 workshop on WordNet and Other Lexical Resources, Pittsbourgh, July 2001.

ECAI 2000, workshop on Ontology Learning http://ol2000.aifb.uni-karlsruhe.de/

IJCAI 2001, workshop on Ontology Learning http://ol2001.aifb.uni-karlsruhe.de/

Appendix A: A fragment of trimmed WordNet for the Tourism domain

{ activity%1 }
{ work%1 }
{ project:00508925%n }
{ tourism_project:00193473%n }
{ ambitious_project:00711113%a }
{ service:00379388%n }
{ travel_service:00191846%n }
{ air_service#2:00202658%n }
{ air_service#4:00194802%n }
{ transport_service:00716041%n }
{ ferry_service#2:00717167%n }
{ express_service#3:00716943%n }
{ exchange_service:02413424%n }
{ guide_service:04840928%n }
{ restaurant_service:03233732%n }
{ rail_service:03207559%n }
{ maid_service:07387889%n }
{ laundry_service:02911395%n }
{ customer_service:07197309%n }
{ guest_service:07304921%n }
{ regular_service#2:07525988%n }
{ outstanding_customer_service:02232741%a }
{ tourism_service:00193473%n }
{ waiter_service:07671545%n }
{ regular_service:02255650%a,scheduled_service:02255439%a }
{ personalized_service:01703424%a,personal_service:01702632%a }
{ secretarial_service:02601509%a }
{ religious_service:02721678%a }
{ church_service:00666912%n }
{ various_service:00462055%a }
{ helpful_service:02376874%a }
{ quality_service:03714294%n }
{ air_service#3:03716758%n }
{ room_service:03250788%n }
{ car_service#3:02384960%n }
{ car_service#4:02385109%n }
{ car_service#5:02364995%n }
{ hour_room_service:10938063%n }
{ transport_service#2:02495376%n }
{ car_service:02383458%n }
{ bus_service#2:02356871%n }
{ taxi_service:02361877%n }
{ coach_service#2:02459686%n }
{ public_transport_service:03184373%n }
{ bus_service:02356526%n,coach_service:02356526%n }
{ express_service#2:02653414%n }
{ local_bus_service:01056664%a }
{ train_service:03528724%n }
{ express_service:02653278%n }
{ car_service#2:02384604%n }
{ coach_service#3:03092927%n }
{ boat_service:02304226%n }
{ ferry_service:02671945%n }
{ car-ferry_service:02388365%n }
{ air_service:05270417%n }
{ support_service:05272723%n }