Integrating Generic and Specialized Wordnets

Bernardo Magnini and Manuela Speranza
ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy
{magnini|manspera}@irst.itc.it

Paper ID: 51

Keywords: Lexical resources, Wordnet, Ontologies

Abstract

Although generic (i.e. domain independent) and specialized (i.e. domain specific) lexical resources are usually developed with different aims, an integrated consultation seems to be necessary for many NLP based applications. We describe an integration procedure based on the definition of plug-in relations that are established to manage overlaps and inconsistencies between the two resources. The approach has been experimented connecting ItalWordNet, a generic lexical database for Italian, and Economic-WordNet, a specialized wordnet for the economic and financial domain.

1 Introduction

In this paper we address the issue of integrating the information included in a generic lexical database with the information included in a specialized (i.e. domain specific) lexical database. We restrict our investigation to wordnet-like lexical resources, that is lexical databases whose model is derived from WordNet (Miller, 1990) (Fellbaum, 1998). The aim is to define a set of procedures to allow an *integrated access* of the two resources, such that overlapping senses are merged and conflicting situations are properly managed.

Our starting point are two existing wordnets for Italian: ItalWordnet (IWN)[1], a generic wordnet deriving from EuroWordNet (Vossen, 1999), and Economic-wordnet (ECOWN), a wordnet for the economic and financial domain. The current application scenario is a recommender system (Magnini and Strapparava, 2001), which includes a document processing module based on a light form of word sense disambiguation. The experimental domain is that of financial news.

As a first attempt to use the two resources in conjunction, a "specific-first" strategy was implemented, which, given a lemma in the document, first looks up in the specific wordnet and just in case of failure resorts to the generic wordnet. An evident drawback of this approach is that it does not cover cases of words belonging to both of the databases when the correct interpretation of the word is placed in the generic wordnet. For instance, in the phrase "building its share market" (taken from a financial news), the word "share" is used with the generic meaning of "portion", while in the specialized wordnet we would have the economic sense of "share" as part of the capital stock.

To overcome this limit we tried with a "union" strategy which, for a given lemma, considers the sum of the senses for that lemma in the two resources. The problem here is that senses for words belonging to both the databases tend to proliferate, making disambiguation harder.

What seems necessary is a deeper integration of the respective word senses,

[1] This research has been supported by SI-TAL (Integrated System for the Automatic Treatment of Language), a National Project devoted to the creation of large linguistic resources and software for Italian written and spoken language processing.

such that overlapping senses are merged and conflicting situations are detected and solved. Our scenario allows some significant simplifications with respect to the general problem of merging two distinct ontologies (Hovy, 1998). On the one side we have a specialized database, whose content is supposed to be more accurate and precise as far as specialized information is concerned; on the other side we can assume that the generic resource guarantees a more uniform coverage as far as high level senses are concerned. These two assumptions provide us with a powerful precedence criterion to be used for managing inheritance in the integration procedure.

The contribution of our work consists in a *plug-in* approach which allows the connecting of the generic and the specialized wordnets in a flexible and modular way. This is realized by means of a semi-automatic procedure with four main steps: (i) first, a minimum set of specialized "basic synsets" is identified; (ii) basic synsets are aligned to corresponding generic synsets and a particular plug-in configuration is selected; (iii) for each plug-in configuration a merging algorithm reconstructs the corresponding portion of the integrated wordnet; (iv) possible inconsistencies are solved.

There are two main benefits of this approach. First, already existing specialized resources can be connected to a generic resource without any change in the resource being necessary, a part from the data conversion into a wordnet-like format. Second, the inheritance of linguistic oriented information makes the specialized resources usable in existing wordnet-based applications.

The paper is structured as follows. Section 2 presents the lexical resources we have used. Section 3 introduces the basic notions of the plug-in approach with some technical details. Section 4 describes the plug-in procedure and reports the results of an application of the approach. Section 5 places our proposal in the context of related works.

2 Generic and Specialized Wordnets

In this work we assume a taxonomic wordnet-like structure of the lexicon, where nodes in the hierarchy are synsets (i.e. synonym sets), and a rather large set of conceptual relations (e.g. Part-of, Cause, Hypernymy, Pertains-to, etc.) are available to build a semantic net among synsets. The EuroWordNet model has been adopted, which is rich enough to encompass most of the relations used in existing non wordnet-like terminological databases.

We focus on the integration of already existing generic and specialized wordnets; both the data acquisition modalities and the evaluation of the quality of the resources do not affect our approach. As for generic lexical databases, typically they contain knowledge with no specific coverage and much attention is placed on coding linguistic-oriented information, such as subcategorization relations and fine-grained sense distinctions. There are several examples of existing generic lexical databases, including the English WordNet (Miller, 1990), monolingual wordnets for several European languages (e.g. Dutch, Spanish, Italian, Basque, etc.), the SENSUS (Hovy, 1998) and the Mikrokosmos (Mahesh, 1996) ontologies.

Specialized databases focus on a certain domain, providing sub-hierarchies of highly specialized concepts with a limited use of lexical and linguistic relations. Synset variants tend to assume the shape of complex terms (i.e. multiwords) and the role of the domain expert is crucial for establishing correct relations. In addition high level knowlegde (i.e. the top ontology) tend to be simplified and domain oriented. Many specialized lexical databases have been developed, particularly for concrete applications, including, for example, a taxonomy for the medical domain (Gangemi et *al*., 1999), the Art and Architecture Getty Thesaurus and the Getty Thesaurus of Geographical Names.

The plug-in model we present has been applied within the SI-TAL project to connect a generic wordnet and a specialized wordnet that have been created independently. *ItalWordnet* (Roventini et *al.*, 2000) created as part of the EuroWordNet project and further developed through the introduction of adjectives and adverbs, is the lexical database involved in the plug-in as a generic resource and consists of about 45,000 lemmas. *Economic-WordNet* is a specialized wordnet for the economic domain and consists of about 5,000 lemmas distributed in about 4,700 synsets. Table 1 summarizes the quantitative data of the two resources considered.

	Specialized	**Generic**
Synsets	4,687	49,108
Senses	5,313	64,251
Lemmas	5,130	45,006
Internal relations	9,372	126,326
Variants/synset	1.13	1.30
Senses/lemma	1.03	1.42

Table 1: IWN and EWN quantitative data.

3 Plug-in Tools

There are three basic intuitions underlying our view of integrating a generic and a specialized wordnet:

- *Coverage*: in the integrated wordnet all the low level synsets (i.e., the terminal nodes) of the specialized wordnet must be accessible, so that no part of the expert knowledge is omitted.
- *Precedence criteria*: in the integrated wordnet, the expert's point of view will be given precedence as far as domain specific information is concerned. This assures that no case of inconsistency may appear in the integrated wordnet.
- *Modularity*: the two resources must not be modified through a plug-in connection. This guarantees that, after an integrated consultation session, both wordnets maintain their original information and can still be used independently.

The whole apparatus to realize an integrated wordnet is based on the use of *plug-in relations* (PLUG-SYNONYMY, PLUG-NEAR-SYNONYMY and PLUG-HYPONYMY) which connect synsets of the specialized wordnet to corresponding generic synsets, and on the use of *eclipsing procedures*, which shadow certain synsets, either to avoid inconsistencies or as a secondary effect of a plug-in relation.

A plug-in relation directly connects pairs of corresponding synsets, the one belonging to the generic wordnet and the other to the specialized wordnet. The main effect of a plug-in relation is the creation of one or more new synsets (plug synsets), which will substitute the connected synsets (i.e. the two synsets directly involved in the relation).

To describe the relations inherited by a plug synset, the following classification, adapted from (Hirst & St-Onge, 1998), is used: *upward links* of a synset are those with a target synset more general than the source synset (e.g. hypernymy relations); *downward links* are those with a target synset more specific than the source synset (i.e. hyponymy and has-instance relations); *horizontal links* include all other relations, such as part-of relations, role relations, cause relations, derivation, etc.

3.1 Plug-synonymy

PLUG-SYNONYMY is used to establish connections between the generic wordnet (hereafter GWN) and the specialized wordnet (hereafter SWN) when overlapping synsets are found, i.e. synsets having the same meaning though belonging to different databases[2].

The main effects of establishing a relation of PLUG-SYNONYMY between $\{a\}^{gwn}$ and $\{a_1\}^{swn}$ is the creation of a new synset $\{a_1\}^{plug}$, which gets its variants from SWN, upward links from GWN, downward links from SWN (a typical way to give precedence to specialized

[2] In the examples, variants within a synset are enclosed in braces; "gwn" and "swn" superscripts show the database to which each synset belongs.

	Plug-synonymy	Plug-hyponymy	
	$\{a_1\}^{plug}$	**$\{a\}^{plug}$**	**$\{a_1\}^{plug}$**
Variants	SWN	GWN	SWN
Upward links	GWN	GWN	$\{a\}^{plug}$
Downward links	SWN	GWN+$\{a_1\}^{plug}$	SWN
Horizontal links	GWN + SWN	GWN	SWN

Table 2: Merging rules for different plug-in relations.

information), horizontal links from SWN (and also from GWN if there is no inconsistency) (see Table 2). As a secondary effect the hypernym(s) of $\{a_1\}^{swn}$ and the hyponyms of $\{a\}^{gwn}$ will be eclipsed.

As an example, the creation of a PLUG-SYNONYMY between $\{fatturato\}^{gwn}$ and $\{fatturato\}^{swn}$ ("sales revenue") produces a new synset $\{fatturato\}^{plug}$, whose hypernym and hyponyms are, respectively, the hypernym of $\{fatturato\}^{gwn}$ and the hyponym of $\{fatturato\}^{swn}$, and causes the eclipsing of the hypernym of $\{fatturato\}^{swn}$.

3.2 Plug-near-synonymy

PLUG-NEAR-SYNONYMY is used to connect synsets that have very similar but not overlapping meanings and are not interchangeable in contexts. In practice, PLUG-NEAR-SYNONYMY is mainly used in case of *over-differentiation of GWN*, i.e. when a SWN synset has two or more corresponding synsets in GWN, which is a consequence of the different attention that domain experts and lexicographers pay to linguistic phenomena. As an example, regular polysemy (see (Apresjan 1974) and (Pustejovsky, 1995)) is generally taken into consideration by lexicographers when defining words, whereas domain experts may completely ignore it. The fact that two or more GWN synsets correspond to one single SWN synset in the integrated consultation, however, highlights the presence of a lexical relation of regular polysemy.

In a similar way, we use a PLUG-NEAR-SYNONYMY in case of *over-differentiation of SWN*, i.e. when a GWN synset corresponds to two or more SWN synsets, as a consequence of subtle distinctions made by domain experts.

Establishing a PLUG-NEAR-SYNONYMY relation has the same effects of creating a PLUG-SYNONYMY (see Table 2).

3.3 Plug-hyponymy

PLUG-HYPONYMY is used to connect a SWN synset and a GWN synset with a more generic meaning, when no corresponding synset is provided in the generic database, i.e. whenever a gap is found in GWN.

The main effect of establishing a PLUG_HYPONYMY relation between $\{a\}^{gwn}$ and $\{a_1\}^{swn}$ is the creation of two plug synsets (see Table 2):

- $\{a\}^{plug}$ gets its variants from GWN, upward links from GWN, $\{a_1\}^{plug}$ as hyponym, in addition to the hyponyms of $\{a\}^{gwn}$, and horizontal links from GWN
- $\{a_1\}^{plug}$ gets its variants from SWN, $\{a\}^{plug}$ as hypernym, downward links from SWN and horizontal links from SWN.

As a secondary effect, the hypernym of $\{a_1\}^{swn}$ will be eclipsed.

The creation of a PLUG-HYPONYMY between $\{attività\}^{gwn}$ ("activity") and $\{attività\_di\_intermediazione\_finanziaria\}^{swn}$ ("financial intermediation") for instance, would produce a new synset $\{attività\}^{plug}$, which will have the hypernym of $\{attività\}^{gwn}$ as hypernym and the hyponyms of $\{attività\}^{gwn}$ plus $\{attività\_di\_intermediazione\_finanziaria\}^{plug}$ as hyponyms, and a new synset $\{attività\_di\_intermediazione\_finanziaria\}^{plug}$, which has $\{attività\}^{plug}$ as hypernym and the hyponyms of $\{attività\_di\_intermediazione\_finanziaria\}^{swn}$) as hyponyms; moreover, it causes the

eclipsing of the hypernym of {attività_ di_intermediazione_finanziaria}^{swn}.

3.4 Eclipsing

While the plug relations described above connect synsets which share some portion of their meaning (i.e. they are, in some degree, aligned) it may be necessary to remove information that otherwise would conflict with others portions of knowledge. This is possible by the eclipsing procedure, which simply remove a certain synset as well as all the relations originating from that synset.

Eclipsing is mainly used as a side effect of establishing a plug-in relation, in that it removes, for instance, the hyponyms of a generic synsets which will be substituted by a specialized sub-hierarchy.

Eclipsing, however, is also used as an independent procedure to avoid the presence of synsets which overlap semantically but are placed inconsistently in the taxonomies. A typical example for this last situation is that of "whale", which may be placed under a "fish" sub-hierarchy in a common sense ontology and under the "mammal" taxonomy in a scientific ontology.

4 Integration Procedure

The *plug-in* approach described in the previous section has been realized by means of a semi-automatic procedure with the following four main steps.

(1) Basic synsets identification. The domain expert identifies a preliminary set of informative synsets ("basic synsets") of the specialized wordnet. These synsets are highly representative of the domain and typically are also present in the generic wordnet. In addition, it is required that basic synsets are disjoint among each other and that they assure a complete coverage of the specialized wordnet (i.e. all terminal nodes have at least one basic synset in their ancestor list).

(2) Alignment. This step consists in aligning each basic synset with the more similar synset of the generic wordnet, on the basis of the structural and lexical properties of the synsets. Then, for each pair a plug-in configuration is decided among those described in Section 3.

(3) Merging. For each plug-in configuration an integration algorithm reconstructs the corresponding portion of the integrated wordnet. If no inconsistency is detected by the integration algorithm, the next plug-in configuration is considered, otherwise step 4 is called.

(4) Resolution of inconsistencies. An inconsistency occurs when the implementation of a plug-in configuration is in contrast with an already realized plug-in. In this case the domain expert has to decide which configuration has the priority and consequently modify the other configuration, which will be passed again to step 2 of the procedure.

The integration procedure described above has been tested for the integration of IWN and EWN (see Section 2). As a first step, about 250 basic synsets (5.3% of the resource) of the specialized wordnet were manually identified by a domain expert, including, for instance "azione" ("share"), and excluding less informative synsets, such as "action". Given that basic sysnsets represent relevant concepts in the domain, they are also quite stable, i.e. they do not change with new versions of the resource. Alignment with respect to the generic wordnet (step 2 of the procedure) is carried out with an algorithm that considers the match of the variants. Candidates are then checked by the domain expert, who also chooses the proper plug-in relation. In case of gaps, a synset with a more generic meaning was selected and a PLUG-HYPONYMY relation was chosen.

At this point the merging algorithm takes each plug-in relation and reconstructs a portion of the integrated wordnet. In total, 4,662 ECOWN synsets were connected to IWN: 577 synsets (b-area in Figure 2) substitute the synsets provided in the generic to represent the corresponding concepts (b_1-

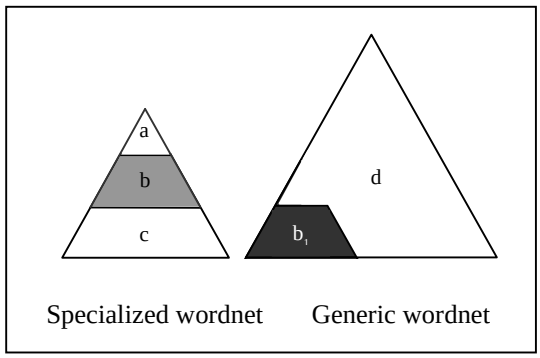

Figure 1: Separated specialized and generic wordnets. Overlapping is represented in colored area.

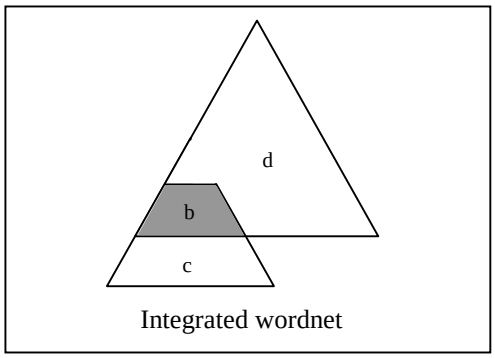

Figure 2 Integrated wordnet. As for overlapping, precedence is given to the specialized wordnet.

area in Figure 1); 4085 synsets, corresponding to the most specific concepts of the domain (c-area in Figure 2) are properly added to the database. 25 high level ECOWN synsets (a-area in Figure 1) were eclipsed as the effect of plug-in relations.

The number of plug-in relations established is 269 (92 PLUG-SYNONYMY, 36 PLUG-NEAR-SYNONYMY and 141 PLUG-HYPONYMY relations), while 449 IWN synsets with an economic meaning were eclipsed, either as a consequence of plug-in relations (when the two taxonomic structures are consistent) or through the independent procedure of eclipsing (when the taxonomies are inconsistent). Each relation connects averagely 17,3 synsets.

5 Related Works

The work we have presented can be considered as a particular case of merging ontologies, a problem that has received increasing attention in the last years. Hovy (1998) uses a semi-automatic methodology to merge SENSUS and Mikrokosmos, two concept based ontologies. Our scenario poses a number of simplifications that allow us to overcome the combinatorial problems of the general task. In particular we have assumed a sort of precedence of the specialized wordnet with respect to the generic one, which dramatically reduces the cases of conflict.

A fully automatic approach has been proposed in (Daude et *al*., 2000), where a constraint relaxation algorithm is used to map two different versions of WordNet. Although the final goals of our work are quite different (i.e. we are interested in the integration of the taxonomies) this mapping techniques could be used in the alignment phase of our procedure to automatically propose candidate pairs for the plug-in phase.

As far as WordNet is concerned, some of the plug-in configurations discussed in Section 3 (e.g. overlapping, over-differentiation, etc.) have been also addressed in the EuroWordNet project (Vossen, 1998) to establish equivalence relations between a monolingual wordnet and correspondent synsets of the Interlingual Index.

Finally, within the field of terminology it has been emphasized the inadequacy of current terminological resources for NLP applications (Maynard & Ananiadou, 2000). According to this, the integration of a specialized wordnet into a generic resource enriches, via inheritance, domain specific information with the linguistic information present in the generic wordnet.

6 Conclusions

We proposed an approach which aims at building an integrated version of pairs of generic and specialized wordnet-like lexical databases. We presented an integration procedure and the relations needed to cope with cases of overlapping, differentiation and absence of information in the two databases, in order to access them in conjunction with each other. We also reported a concrete experience where the plug-in approach has been used to connect a generic wordnet for the Italian language and a specialized wordnet for the economic-financial domain.

In the next future, we plan to work following two main directions. We are interested to test the validity of the plug-in approach by applying the model to specialized lexical resources for different domains and for different languages. Second, practical experiments in which the advantages of the integration of two resources for NLP based applications are evaluated are still needed. In particular, it would be of great interest to quantify the amount of linguistic information usefully inherited by the specialized resource from the generic one.

The amount of overlapping information through a generic and a specialized database may be significant: a first question is the alignment of this information, which also requires a structural alignment to guarantee that two hierarchy fragments actually reflect the same point of view.

7 References

Apresjan, JU. Regular Polysemy. *Linguistics*, 142, 1974.

Daude J., Padro L. and Rigau G. Mapping Wordnets Usinf Structural Information. Proceedings of ACL-2000, pp. 504-511, Honk Kong, 2000.

Fellbaum, C. Wordnet, an Electronic lexical database. MIT Press, 1998.

Gangemi, A., Pisanelli, D.M. and Steve G.. Overview of the ONIONS project: Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering,* vol.31, 1999.

Hirst G. and St-Onge, D. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Fellbaum (ed.) *Wordnet, an Electronic lexical database*. MIT Press, 1998.

Hovy, E.H. 1998. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). Granada, Spain.

Magnini B. and Strapparava C. . Improving User Modeling with Content-Based Techniques. To appear in Proceedings of the 8th International Conference of User Modeling UM-2001.

Mahesh K. Ontology development for Machine Translation: Ideology and Methodology. CLR Report MCCS-96-292, 1996.

Maynard D., and Ananiadou S. Creating and Using Domain-Specific Ontologies for Terminological Applications. Proceedings of LREC-2000, *Second International Conference on Language Resources and Evaluation*, pp. 869-874, Athens, Greece, 2000.

Miller G.A. WordNet: An Online Lexical Database. International Journal of Lexicography 3(4) (special issue), 1990.

Pustejovsky, J. *The Generative Lexicon.* MIT Press, 1995.

Roventini A., Alonge A., Bertagna F., Magnini B. and Calzolari N. ItalWordNet: a large semantic database for Italian. Proceedings of LREC-2000, *Second International Conference on Language Resources and Evaluation*, pp. 783-790, Athens, Greece, 2000.

Vossen, P. EuroWordNet: a multilingual database with lexical semantic networks. Kluver Academic Publishers, 1998.