The Development of Lexical Resources for Information Extraction from Text Combining WordNet and Dewey Decimal Classification*

Gabriela Cavaglià
ITC-irst Centro per la Ricerca Scientifica e Tecnologica
via Sommarive, 18
38050 Povo (TN), ITALY
e-mail: cavaglia@irst.itc.it

Abstract

Lexicon definition is one of the main bottlenecks in the development of new applications in the field of Information Extraction from text. Generic resources (e.g., lexical databases) are promising for reducing the cost of specific lexica definition, but they introduce lexical ambiguity. This paper proposes a methodology for building application-specific lexica by using WordNet. Lexical ambiguity is kept under control by marking synsets in WordNet with field labels taken from the Dewey Decimal Classification.

1 Introduction

One of the current issues in Information Extraction (IE) is efficient transportability, as the cost of new applications is one of the factors limiting the market. The lexicon definition process is currently one of the main bottlenecks in producing applications. As a matter of fact the necessary lexicon for an average application is generally large (hundreds to thousands of words) and most lexical information is not transportable across domains. The problem of lexicon transport is worsened by the growing degree of lexicalization of IE systems: nowadays several successful systems adopt lexical rules at many levels.

The IE research mainstream focused essentially on the definition of lexica starting from a corpus sample (Riloff, 1993; Grishman, 1997) with the implicit assumption that a corpus provided for an application is representative of the whole application requirement. Unfortunately one of the current trends in IE is the progressive reduction of the size of training corpora: e.g., from the 1,000 texts of the MUC-5 (MUC-5, 1993) to the 100 texts in MUC-6 (MUC-6, 1995). When the corpus size is limited, the assumption of lexical representativeness of the sample corpus may not hold any longer, and the problem of producing a representative lexicon starting from the corpus lexicon arises (Grishman, 1995).

Generic resources are interesting as they contain (among others) most of the terms necessary for an IE application. Nevertheless up to now the use of generic resources within IE system has been limited for two main reasons. First the information associated to each term is often not detailed enough for describing the relations necessary for a IE lexicon; secondly the presence of a large amount of lexical polysemy.

In this paper we propose a methodology for semi-automatically developing the relevant part of a lexicon (foreground lexicon) for IE applications by using both a small corpus and WordNet.

2 Developing IE Lexical Resources

Lexical information in IE can be divided into three sources of information (Kilgarriff, 1997):

- an *ontology*, i.e. the templates to be filled;
- the *foreground lexicon* (FL), i.e. the terms tightly bound to the ontology;
- the *background lexicon* (BL), i.e. the terms not related or loosely related to the ontology.

In this paper we focus on FL only.

The FL has generally a limited size with respect to the average dictionary of a language; its dimension depends on each application needs, but it is generally limited to some hundreds of words. The level of quantitative and qualitative information for each entry in the FL can be very high and it is not transportable across domains and

*This work was carried on at ITC-IRST as part of the author's dissertation for the degree in Philosophy (University of Turin, supervisor: Carla Bazzanella). The author wants to thank her supervisor at ITC-IRST, Fabio Ciravegna, for his constant help. Alberto Lavelli provided valuable comments to the paper.

applications, as it contains the mapping between the entries and the ontology. Generic dictionaries can contribute in identifying entries for the FL, but generally do not provide useful information for the mapping with the ontology. This mapping between words and ontology is generally to be built by hand. Most of the time in transporting the lexicon is spent in identifying and building FLs. Efficiently building FLs for applications means building the right FL (or at least a reasonable approximation of it) in a short time. The right FL contains those words that are necessary for the application and only those. The presence of all the relevant terms should guarantee that the information in the text is never lost; inserting just the relevant terms allows to limit the development effort, and should guarantee the system from noise caused by spurious entries in the lexicon.

The BL could be seen as the complementary set of the FL with respect to the generic language, i.e. it contains all the words of the language that do not belong to the FL. In general the quantity of application specific information is small. Any machine readable dictionary can be to some extent seen as a BL. The transport of BL to new applications is not a problem, therefore it will not be considered in this paper.

2.1 Using Generic Lexical Resources

We propose a development methodology for FLs based on two steps:

- Bootstrapping: manual or semi-automatic identification from the corpus of an initial lexicon *(Core Lexicon)*, i.e. of the lexicon covering the corpus sample.
- Consolidation: extension of the Core Lexicon by using a generic dictionary in order to completely cover the lexicon needed by the application but not exhaustively represented in the corpus sample.

We propose to use WordNet (Miller, 1990) as a generic dictionary during the consolidation phase because it can be profitably used for integrating the Core Lexicon by adding for each term in a semi-automatic way:

- its synonyms;
- hyponyms and (maybe) hypernyms;
- some coordinated terms.

As mentioned, there are two problems related to the use of generic dictionaries with respect to the IE needs.

First there is no clear way of extracting from them the mapping between the FL and the ontology; this is mainly due to a lack of information and cannot in general be solved; generic lexica cannot then be used during the bootstrapping phase to generate the Core Lexicon.

Secondly experience showed that the lexical ambiguity carried by generic dictionaries does not allow their direct use in computational systems (Basili and Pazienza, 1997; Morgan et al., 1995). Even when they are used off-line, lexical ambiguity can introduce so much noise (and then overhead) in the lexical development process that their use can be inconvenient from the point of view of efficiency and effectiveness.

The next section explains how it is possible to cope with lexical ambiguity in WordNet by combining its information with another source of information: the Dewey Decimal Classification (DDC) (Dewey, 1989).

3 Reducing the lexical ambiguity in WordNet

The main problem with the use of WordNet is lexical polysemy [1]. Lexical polysemy is present when a word is associated to many senses (synsets). In general it is not easy to discriminate between different synsets. It is then necessary to find a way for helping the lexicon developer in selecting the correct synset for a word.

In order to cope with lexical polysemy, we propose to integrate WordNet synsets with an additional information: a set of *field labels*. Field labels are indicators, generally used in dictionaries, which provide information about the use of the word in a *semantic field*. Semantic fields are sets of words tied together by "similarity" covering the most part of the lexical area of a specific domain.

Marking synsets with field labels has a clear advantage: in general, given a polysemous word in WordNet and a particular field label, in most of the cases the word is disambiguated. For example *Security* is polysemous as it belongs to 9 different synsets; only the second one is related to the economic domain. If we mark this synset with the field label *Economy*, it is possible to disambiguate the term *Security* when analyzing texts in an economic context. Note that WordNet being a hierarchy, marking a synset with a field label means also marking all its sub-hierarchy with such field label. In the *Security* example, if we mark the second synset with the field label *Economy* we also associate the same field label to the synonym *Certificate*, to the 13 direct hyponyms and to the 27

[1] Actually the problem is related to both polysemy and omonymy. As WordNet does not distinguish between them, we will use the term polysemy for referring to both.

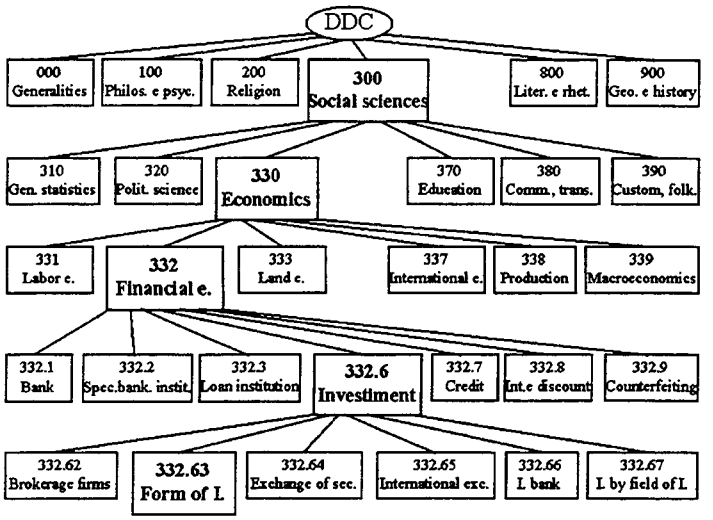

Figure 1: An extract of the Dewey hierarchy relevant for the financial field

indirect ones; moreover we can also inspect its coordinated terms and assign the same label to 9 of the 33 coordinate terms (and then to their direct and indirect hyponyms). Marking is equivalent to assigning WordNet synsets to sets each of them referring to a particular semantic field. Marking the structure allows us to solve the problem of choosing which synsets are relevant for the domain. Associating a domain (e.g., finance) to one or more field labels should allow us to determine in principle the synsets relevant for the domain. It is possible to greatly reduce the ambiguity implied by the use of WordNet by finding the correct set of field labels that cover all the WordNet hierarchy in an uniform way. Therefore we can reduce the overhead in building the FL using WordNet.

Our assumption is that using semantic fields taken from the DDC[2] , all the possible domains can then be covered. This is because the first ten classes of the DDC (an extract is shown in figure 1) exhaust the traditional academic disciplines and so they also cover the generic knowledge of the world. The integration consists in marking parts of WordNet's hierarchy, i.e. some synsets, with semantic labels taken from the DDC.

4 The development cycle using WN+DDC

The consolidation phase mentioned in section 2.1 can be integrated with the use of the WN+DDC as generic resource (see figure 2). Before starting the development, the set of field labels relevant for the application must be identified. Then the Core Lexicon is identified in the usual way.

Using WN+DDC it is possible for each term in the Core Lexicon to:

- identify the synsets the term belongs to; ambiguities are reduced by applying the intersection of the field labels chosen for the current application and those associated to the possible synsets.
- integrate the Core Lexicon by adding, for each term: synonyms in the synsets, hyponyms and (maybe) hypernyms and some coordinated terms.

The proposed methodology is corpus centered (starting from the corpus analysis to build the Core Lexicon) and can always be profitably applied. It also provides a criterion for building lexical resources for specific domains. It can be applied in a semiautomatic way. It has the advantage of using the information contained in WordNet for expanding the FL beyond the corpus limitations, keeping under control the ambiguity implied by the use of a generic resource.

5 Conclusion

Up to now experiments have been carried on in the financial domain, and in particular in the domain of bonds issued by banks. Experiments are continuing. The construction of WN+DDC is a long process that has to be done in general. Up to now we have just started inserting in WordNet the field labels that are interesting for the domain

[2]The Dewey Decimal Classification is the most widely used library classification system in the world; at the broadest level, it classifies concepts into ten main classes, which cover the entire world of knowledge.

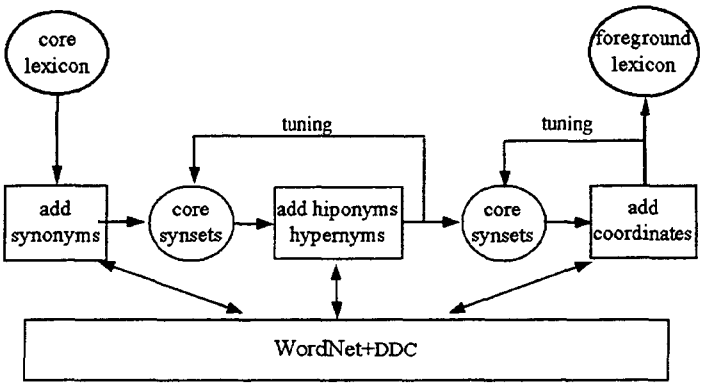

Figure 2: Outline of the final Consolidation phase.

under analysis. If the final experiments will confirm the usefulness of the approach, we will extend the integration to the rest of the WordNet hierarchy. The final evaluation will include a comparison of the lexicon produced by using WN+DDC with a normally developed lexicon in the domain of bond-issue (Ciravegna et al., 1999). The evaluation will consider both quality and quantity of terms and development time of the whole lexicon. One of the issues that we are currently investigating is that of choosing the correct set of field labels from DDC: DDC is very detailed and it is not worth integrating it completely with WordNet. It is necessary to individuate the correct set of labels by pruning the DDC hierarchy at some level. We are currently investigating the effectiveness of just selecting the first three levels of the hierarchy.

References

Roberto Basili and Maria Teresa Pazienza. 1997. Lexical acquisition for information extraction. In M. T. Pazienza, editor, *Information Extraction: A multidisciplinary approach to an emerging information technology.* Springer Verlag.

Fabio Ciravegna, Alberto Lavelli, Nadia Mana Luca Gilardoni, Silvia Mazza, Massimo Ferraro, Johannes Matiasek, William Black, Fabio Rinaldi, and David Mowatt. 1999. Facile: Classifying texts integrating pattern matching and information extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99).* Stockholm, Sweden.

Melvil Dewey. 1989. *Dewey Decimal Classification and Relative Index. Edition 20.* Forest Press, Albany.

Ralph Grishman. 1995. The NYU system for MUC-6 or where's syntax? In *Sixth message understanding conference MUC-6.* Morgan Kaufmann Publishers.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In M. T. Pazienza, editor, *Information Extraction: a multidisciplinary approach to an emerging technology.* Springer Verlag.

Adam Kilgarriff. 1997. Foreground and background lexicons and word sense disambiguation for information extraction. In *International Workshop on Lexically Driven Information Extraction,* Frascati, Italy.

G.A. Miller. 1990. Wordnet: an on-line lexical database. *International Journal of Lexicography,* 4(3).

Richard Morgan, Roberto Garigliano, Paul Callaghan, Sanjay Poria, Mark Smith, Agnieszka Urbanowicz, Russel Collingham, Marco Costantino, Chris Cooper, and the LOLITA Group. 1995. University of Durham: Description of the LOLITA system as used for MUC-6. In *Sixth message understanding conference MUC-6.* Morgan Kaufmann Publishers.

MUC-5. 1993. *Fifth Message Understanding Conference (MUC5).* Morgan Kaufmann Publishers, August.

MUC-6. 1995. *Sixth Message Understanding Conference (MUC-6).* Morgan Kaufmann Publishers.

Ellen Riloff. 1993. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence,* pages 811–816.