Adapting and extending lexical resources in a dialogue system

Ana García-Serrano
ISYS Group
AI Department
Technical University of Madrid
Boadilla del Monte, 28660
Madrid, Spain
agarcia@ dia.fi.upm.es

Paloma Martínez
Advanced DB Group
CS Department
Universidad Carlos III de Madrid
Avda. Universidad 30
Leganés 28911, Madrid, Spain
pmf@inf.uc3m.es

Luis Rodrigo
ISYS Group
AI Department
Technical University of Madrid
Boadilla del Monte, 28660
Madrid, Spain
lrodrigo@isys.dia.fi.upm.es

Abstract

This paper presents the adaptation and customization of two lexical resources: Brill tagger, Brill (1992), and EuroWordNet, Vossen et al. (1998), to be used in the ADVICE project devoted to build an intelligent virtual reality sales and service system that uses human language technology.

1 Introduction

The work described in this paper is comprised in the ADVICE[1] project, which consists of the development of a new interface for a craftsmanship tools e-commerce system. With the aim of providing full support to customers of online shops and to the users of electronic services on the Internet during the complete customer service lifecycle; the ADVICE system will provide:

An advanced multimedia user interface supporting natural language text input and output, as well as an animated assistant, that ensures a high level of user-system interaction.
A software environment for building intelligent sales assistants for two types of selling and marketing services: sales-service and after sales service.

The virtual assistant supported by this application will be capable of interacting with the user in natural language, adapt its recommendations to the requirements and characteristics of the customers and provide explanations of the advised products as well as of the different interesting alternatives.

The dialogue is focussed on accomplishing a specific task, ordering products in an e-commerce system as well as getting advice about those products. The complexity of this domain lies in the broad product range with specialization of the crafts for particular tasks and applications.

One of the main aspects of the buying-selling interaction in the web is the capability of the web site of generating some kind of trust feeling in the buyer, just like a human shop assistant would do in a person-to-person interaction. Things like understanding the buyer needs, being able to give him technical advice, assisting him in the final decision are not easy things to achieve in a web selling site. Natural language (NL) techniques can play a crucial role in providing this kind of enhancements.

Another good motivation to integrate natural language technology in this kind of sites is to make the interaction easy to those people less confident with the Internet or even computer technologies. Inexperienced users feel much more comfortable expressing themselves and receiving information in natural language rather than through the human standard ways.

An ADVICE project objective is to build a generic language processing component following a language engineering approach, that is, to achieve maintainability, time optimization, robustness, flexibility, domain adaptability and generality. These features require profiting from the existing resources taking into account that the goal is to achieve systems that work in real domains. Therefore, resources adaptation is a

[1]Virtual Sales Assistant for the Complete Customer Service Process in Digital Markets. IST Project 1999-11305

crucial step when developing this kind of systems, because these resources are not usually compatible with domain requirements.

At current state two linguistic resources have been analyzed and adapted in order to be incorporated into the ADVICE project: Brill tagger, Brill (1992), and EuroWordNet, Vossen et al. (1998).

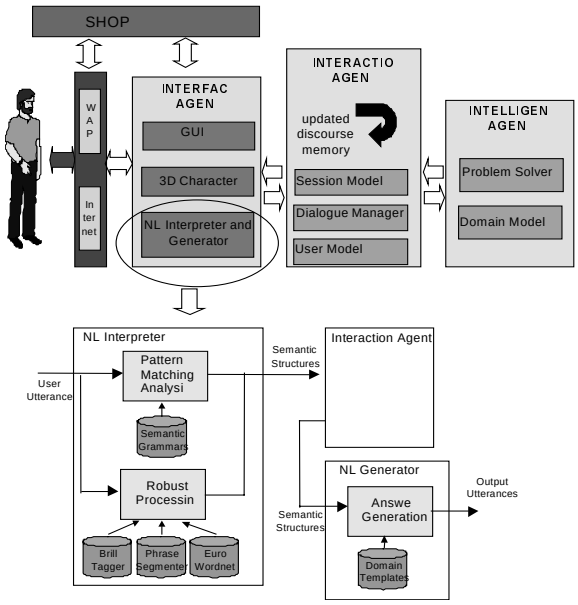

Figure 1: ADVICE architecture

The paper is structured as follows: next section focuses on a brief description of the NL module of ADVICE project; section 3 is devoted to explain how the corpus study has influenced the customization of the lexical resources; sections 4 and 5 outline the Brill tagger and EuroWordnet extension to cover specific domain peculiarities and, finally, in section 6 some conclusions and future work are presented.

2 Overview of Natural Language component

Figure 1 shows the architecture of the ADVICE system. The Interface Agent includes the *NL Interpreter and Generator* component that is in charge of analyzing the user utterances as well as of generating the appropriate answers according to the dialogue.

The input sentence is interpreted in order to obtain a semantic representation. At this moment, two interpretation strategies are defined. Firstly, message extraction techniques useful in specific domains are used in ADVICE, implemented by means of semantic grammars reflecting e-commerce generic sentences and idioms, sublanguage specific patterns and keywords. Secondly, if the pattern matching analysis does not work successfully, a robust processor that makes use of several linguistic resources (Brill tagger, EuroWordNet and a Phrase Segmenter) integrates syntactic and semantic analysis in different ways. The aim is to attach the identified (syntactic or semantic) segments of the sentence in a partial structure to go on with the dialogue.

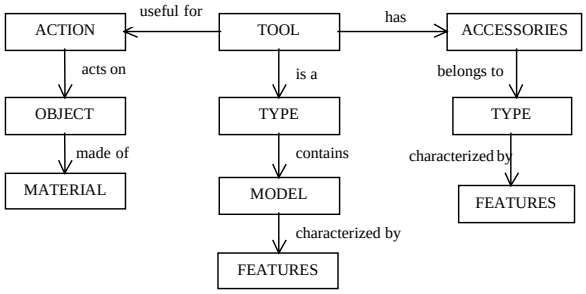

Figure 2: Partial view of craft domain ontology

Concerning the generation of answers, a domain-specific template-based approach is proposed. The templates used to generate natural language answers can be propositional (if they require arguments to fill the slots in) or not propositional (for instance, agreements, rejections and topic movements).

The information gathering for the *NL Interpreter and Generator* involves different sources: dialogues collected from users, the database of products and users and common sense about human-computer interaction.

3 Domain corpus study

When dealing with such a restricted domain, a great amount of the possibilities of success relies on the domain adaptation capability that the system proves. As long as an intelligent, open domain and human-computer dialogue is far from being realistic, the designers usually focus on the customization of the available resources so that they are as trustworthy as possible when applied to selected input, even though this implies a loss of accuracy when it has to do with out-of-the-domain texts.

The decision of directly adapting existing resources comes from the fact that the initial

amount of data available was almost non-existent. In other situation we could have considered the possibility of developing the resources from scratch or, at least, training them (specially the tagger) so that its knowledge was completely acquired for this application, but that was too time consuming.

A Wizard of Oz experiment was conducted to gain some domain specific data to work with. A corpus of 527 sentences resulted from this experiment, which were lexically and syntactically annotated. As a side effect of this experiment, the members of the development team considerably increased their knowledge and familiarity with the domain characteristics.

The corpus analysis has been performed keeping in mind the need to separate domain-independent aspects of the system from the domain-specific components that serve to define specific application domains.

Regarding the adaptation of the resources, both of them (Brill tagger and EuroWordnet) were affected by the results of the above-mentioned process. As long as more and more terminology was studied, it became evident the highly structured way in which it could be organized.

Although both lexical resources are explained in detail in sections 4 and 5, some special issues relating to domain-based customization are outlined below.

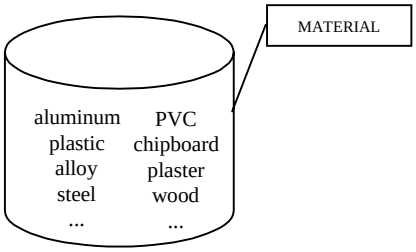

Figure 3: Detail of the content of a particular category of the ontology

After a careful study, the ontology shown in figure 2 was developed, having each of the boxes a complete set of specific terminology (see figure 3).

At this stage EuroWordnet had been understood to bring robustness to the NL analysis contributing with its huge amount of words in the form of a semantic network, but the similarity of the proposed ontology with EuroWordnet built-in structure lead us to a proposal of extending the EuroWordnet semantic network with the specific vocabulary regarding our context and, even more important, the relationships among them. To carry out this proposal, it was firstly considered the option of making use of the existing relations of EuroWordnet and its domain labels. We will add the terminology and definitions that were not previously integrated, and tag all the specific terminology with a domain label, so that it could be directly identified as a word of special interest for our needs and related to other words in the domain. Unfortunately the first study of this possibility revealed that it turned to be a process much less intuitive than it appears at first sight, and a deeper study about how this process could be implemented is being carried out at the moment of writing this paper.

Concerning Brill tagger, its revision was also affected by the aforementioned corpus analysis. First, the sentences extracted from the corpus were analyzed in order to obtain a set of generic patterns (about 90 patterns). This had direct repercussions in the set of contextual rules (the rules that the tagger uses to select the correct tag for a word considering the context in which it appears) which were adapted to fit the domain peculiarities detected in the patterns, as shown in figure 4. Some new rules were developed and some of the existent were modified.

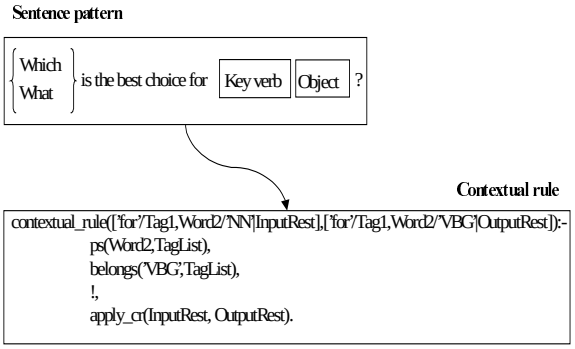

Figure 4: Influence of sentence patterns on contextual rules

The set of tags was also changed to help in the domain customization process. A new tag 'XX' was defined to tag the words not present in the lexicon. The words tagged with this tag are good candidates to be domain terminology and are stored in a special file to be analyzed. As a result of this process, after the tagging of a sufficient amount of text, we have a collection of words that are good candidates to build up a specialized lexicon. This lexicon can be used as a support for the original one, filling the gaps

that it may have when dealing with specific language, and can be further refined as more and more terms are tagged as 'XX'.

Next sections focus on the main topics of Brill tagger and EuroWordnet extension and adaptation.

4 Adapting Brill Tagger

The Brill tagger, Brill (1992), is a rule-based part of speech tagger programmed in Perl for English language. A morphological analysis of a word form produces a set of possible base forms with associated inflectional information. For each occurrence of a word form in context, a POS (Part-of-Speech) tagger discriminates which of these base forms is more likely in the context. Our objective with this tagger was twofold. Firstly, the tagger was translated (and softly modified) into Prolog, and then, an exhaustive study was carried out in order to adapt it to ADVICE domain.

The development environment used in ADVICE project, Ciao Prolog, Bueno et al. (1999), motivates the translation of resources into Prolog. Furthermore, Prolog allows rapid development and facilitates efficient and real-time processing.

The tagger has three kinds of knowledge: a base lexicon, contextual rules and lexical rules. The *lexicon* is a long list of words with one or more possible tags associated. It is used to the first assignment of tags to the input text. The original knowledge is stored in a text file where each line contains the word and the possible tags that can be associated to the word. There are 93696 different entries. This lexicon was directly translated into Prolog facts, having the following structure: `ps(Word, ListOfTags)`. The words are stored as Prolog terms and have been alphabetically sorted to take advantage of the indexing capabilities of Prolog. Here is a glance at the file contents:

```
ps('charts',['NNS','VBZ']).
ps('chary',['JJ']).
ps('chase',['NN','JJ','VB','VBP').
ps('chased',['VBN','VBD']).
ps('chasers',['NNS']).
ps('chasing',['VBG','NN']).
```

The *contextual rules* manage the knowledge relative to the suitable tags for a word regarding the context in which it appears. It considers both the words and the tags that surround the word under consideration and decides about the correctness of the preassigned tag. The original file contains 284 rules, stored in a text file in the following format: `NN VB PREVTAG TO`, which means "change a NN (noun, singular or mass) tag to VB (verb, base form) if the previous tag is TO". These contextual rules have been translated into a set of Prolog rules plus a predicate that goes through the input and applies the appropriate rule if necessary until it reaches the end of the text to tag. Every time it changes a tag, it checks that the new tag is contained in the set of allowed tags for that word. In other case, the rule is not applied. Below, the main predicate and an example rule are shown:

```
apply_cr([],[]).
apply_cr(List1,List2):-
      contextual_rule(List1,List2).

contextual_rule(
[Word1/'TO',Word2/'NN'|InputRest],
[Word1/'TO',Word2/'VB'|OutpRest]):-
      ps(Word2,TagList),
      belongs('VB',TagList),
      !,
      apply_cr(InputRest,OutpRest).
```

As in the case of the original Brill, the order of the rules is relevant to the final result, as one of the early rules may affect the context of a word so that a later rule may (or may not) be applicable.

The *lexical rules* are used to infer the most possible tag for a word considering its lexical shape, specially its prefixes and suffixes. From the 148 original rules, only the 63 regarding to suffixes have been translated at this moment. The structure of the original rules is: `ly hassuf 2 RB` representing "if the word has the suffix, of length 2, "-ly", assign the tag RB (adverb)". The structure of the Prolog predicates containing these lexical rules is equal to the one from the contextual rules; we have an "apply_lr" predicate and several "lexical_rule", combined in the same way.

Finally, once all the resources have been translated, the last part was the simulation of the tagging process. This has been carried out considering the information in Brill (1992), Brill (1994) and Brill (1995), taken as a reference starting point. The designed process works as follows: first of all, the input is pre-processed to

decompose the possible contractions. Next, each word in the input is attached its most probable tag, taking as the most probable the first one in the list of the lexicon. If a word is not present in the lexicon, it is assigned a special tag XX representing *unknown word* instead of the noun tag that it was assigned in the original tagger. Then, lexical rules are applied (if possible) trying to disambiguate the words that were not found in the lexicon. Once every word has a tag, we apply the contextual rules, resulting in the definitive tagged text.

Apart from the already mentioned modification of the general process of tagging, some modifications in the knowledge bases have been made, basically in order to adapt the resource to our domain (craft tools). Regarding the *lexicon*, we have adapted the punctuation mark's tags to our needs, changing the tags that the original Brill tagger used. The objective was to assign a different tag to every punctuation mark, giving them a treatment equal to the words. These symbols are relevant when we are looking for a sentence syntax analysis or a sentence pattern. Another change to the lexicon is related with the order of the tags in certain words, and is further explained in the next paragraph.

The original set of contextual rules was firstly translated as is. With this set of rules, a corpus of 20 dialogues restricted to the domain language obtained by the wizard-of-Oz experiments was tagged. We compared the results to the manually tagged text and we extracted some common mistakes that the tagger was making (Table 1). These mistakes could lead us in two directions. If it was a mistagging of a particular word frequently repeated, the best way to fix it was to change the order of the tags associated to that word in the lexicon.

Common mistakes	Solutions
-Mistagging of a particular word.	-Change the tag preference in the lexicon.
-Problems tagging word categories.	-Modifications in the set of rules.
-Lexical rules applied to already known words.	-Assign tag 'XX' to unknown words.

Table1: Tagger errors

For example, in our domain the word "saw" appears most of the times as a noun while the preferred tag in the lexicon for that word was verb. In this case, placing the noun tag before the verb one in the tag list turns to be a good and simple solution. If the problem we were facing was not related to particular words, but to word categories, the way to fix it was through the adding/deleting/modification of the contextual rules. After the study of the errors, we developed several new rules. For instance, a rule for "changing a tag 'IN' (meaning preposition or subordinating conjunction) to RB (adverb) in the collocation As ... as".

Finally, concerning the lexical rules, the main change comes from the fact that the words that were not in the lexicon, in the original Brill tagger were annotated as 'NN', while now they have an special tag 'XX'. As these lexical rules try to disambiguate these kind of words, whenever an original rule was applied to a word tagged as 'NN', it has been adapted to act on the words tagged as 'XX'. Finally, only if none of the rules has been able to disambiguate it, it is automatically changed to 'NN'.

5 Adapting EuroWordnet

The first step in adapting EuroWordNet database was to translate it into Prolog. EuroWordNet stores its knowledge in file texts that are organized in a very structured way. As a first approach, the information extracted from the databases has been the synonymy, hyponymy and hyperonymy basic relations. This is the minimum information needed to preserve the semantic structure that lies beyond EuroWordNet. The resulting lexical database stores the information in the predicate `ewn/5` as Prolog facts. The first field references the word of interest, the second stores the grammatical category of the word, and the following, the synonyms, hyperonyms and hyponyms, respectively, as is shown below:

```
ewn('phonograph', n,
     sin(['record player']),
     hyper(['machine']),
     hypo(['acoustic gramophone',
           'jukebox'])).
```

From the analysis of domain application, an ontology of concepts incorporating the terminology has been obtained (a partial view was shown in Figure 2). Intensive work is being made in order to merge EuroWordNet and domain ontology structures, resulting in an enlarged resource fully adapted to our needs and

useful both as a lexicon and as a semantic network.

Figure 4 shows an EuroWordNet partial structure from Vossen et al. (1998) on which the vocabulary corresponding to two kinds of concepts from the domain ontology (OBJECT and ACTION) have been integrated (colored rounded boxes with a dotted line). Moreover, a new semantic relation named *Acts on* is also shown to represent the association between OBJECT and ACTION. However, some of the relationship provided by EuroWordNet (apart from synonymy, hiponymy, and meronymy) could probably cover some of the relations present in the domain ontology.

In the same way, TOOL terminology is inserted as *Instrument—>Function—>1stOrderEntity*; terminology MATERIAL *as Substance—> Form—>1stOrderEntity*, etc.

6 Conclusion

The most important aspects of Brill tagger and EuroWordNet adaptation to a specific application domain and a development platform have been explained. These resources are going to be used in a dialogue system in order to give robustness to the NL interpretation process.

We are currently involved in the process of implementation of the enhancements proposed in the paper. Concerning Brill tagger, only preliminary tests have been carried out. The first translation of the tagger was used to tag the available corpus, showing an accuracy of 0.9163. Although this result is far from the state of the art taggers, it is hopeful result, as the improvements concerning the full terminology and complete set of contextual rules had still been incorporated. Results up to state of the art levels are expected as the enhancements are implemented.

Related EuroWordNet, we are studying if the semantic relations (apart from synonymy, hiponymy, meronymy) supported by EuroWordNet are enough to cover the relationships shown in the domain ontology of Figure 2. For instance, there are semantic relationships such as *is derived from*, *has mero made of*, *involved instrument*, *has instance* and many others that are useful in the application domain of crafts selling.

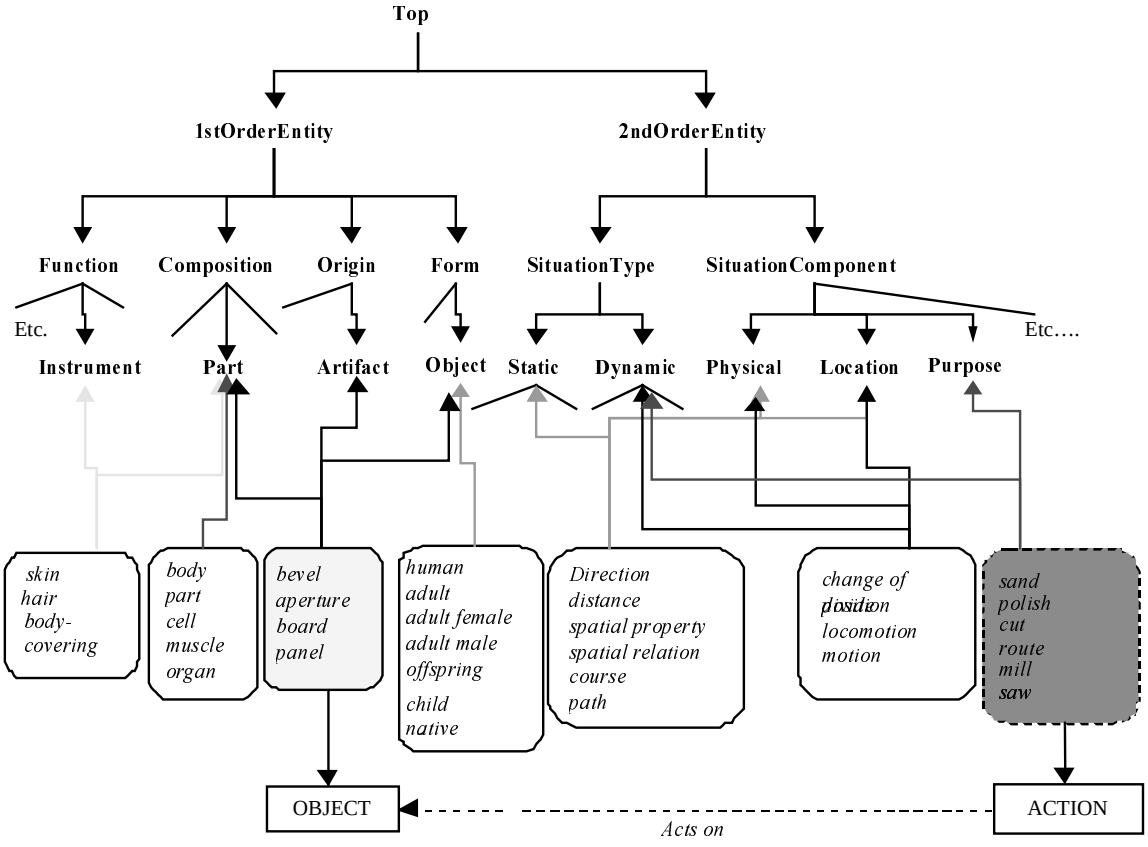

Figure 5: Proposal of extension adapted from Vossen,

References

Alonge (1996) Alonge, A., *Definition of the links and subsets for verbs in the eurowordnet project.* Technical report, Deliverable D006.

Bloksma (1996) Bloksma, L., Díez-Orzas, P., Vossen, P., *User requirements and functional specification of the eurowordnet project.* Technical report, Deliverable D001.

Brill (1992) Brill, E*., A simple rule-based part of speech tagger.* Proceedings of the Third Conference on Applied Natural Language Processing, ANPL, ACL, Trento, Italy, pp. 152-155.

Brill (1994) Brill, E., *Some advances in rule-based part of speech tagging.* Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa. , pp. 722-727.

Brill (1995) Brill, E., *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging.* Computational Linguistics, 21(4): 543-565

Bueno et al. (1999), F. Bueno, D. Cabeza, M. Carro, M. Hermenegildo, P. López, and G. Puebla. *The Ciao Prolog System: A Next Generation Logic Programming Environment, REFERENCE MANUAL.* The Ciao System Documentation Series Technical Report CLIP 3/97.1, The CLIP Group School of Computer Science Technical University of Madrid.

Climent (1996) Climent, S., Rodríguez, H., Gonzalo, J., *Definition of the links and subsets for nouns in the eurowordnet project.* Technical report, Deliverable D005.

Gonzalo et al. (1998) Gonzalo, J., Verdejo, M.F., Chugur, I., López, F., Peñas, A., *Extracción de relaciones semánticas entre nombres y verbos en EuroWordnet.* Revista SEPLN nº 23, 1998.

Hepple (2000) Hepple, M., *Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers.* 38th Annual meeting of the Association for Computational Linguistics (ACL-2000), pp 278-285, Hong Kong, October 2000.

Kilgarriff (1997) Kilgarriff, A., *Foreground and background lexicons and word sense disambiguation for information extraction.* Proc. Workshop on Lexicon Driven Information Extraction, Frascati, Italy, July 1997.

Martínez et al. (2000), Martínez, P.; García-Serrano, A., *The role of knowledge-based technology in language applications development.* Expert Systems with Applications 19 , 31-44, 2000.

Miller et al. (1990), Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., *Introduction to WordNet: An On-line Lexical Database.* (Revised August 1993). Princeton University, New Jersey.

Vossen et al. (1998), Vossen, P., Bloksma, L., Rodríguez, H., Climent, H., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., Peters, W., *The EuroWordNet Base Concepts and Top Ontology. Version2.* EuroWordNet (LE 4003) Deliverable.