Exploiting ItalWordNet Taxonomies in a Question Classification Task

Francesca Bertagna

Istituto di Linguistica Computazionale, National Research Council,
Via G. Moruzzi 1, I-56100 Pisa, Italy
Email: francesca.bertagna@ilc.cnr.it

Abstract. The paper presents a case-study about the exploitation of ItalWordNet for Question Answering. In particular, we will explore the access to ItalWordNet when trying to derive the information that is crucial for singling out the answers to Italian Wh-questions introduced by the interrogative elements Quale and Che.

1 Introduction

The paper describes some aspects arised during the first phase of the work carried out for a Ph.D. research[1] dedicated to the exploration of the role of linguistic resources (from now on LRs) in a Question Answering (QA) application. The leading idea of the thesis is that the testing activity can highlight potentialities, together with problems and limitations, of the bulk of information collected during the last two decades by linguists and computational linguists. Altought LRs are not conceived to meet the requirements of a specific task (but rather to represent a sort of repository of information of general interest), they are significant sources of knowledge that should allow systems to automatically perform inferences, retrieve information, summarize texts, translate words in context from a language to another etc.. Computational lexicons storing semantic information, in particular, are supposed to provide a description of the *meaning* of the lexical units they collect. It is interesting to evaluate what is the *heuristic value* of such description and to what extent it is exploitable and useful to perfom specific tasks (e.g. in matching question and answer). Tons of papers have been written about the use of WordNet in IR and in QA and the time is mature to test also resources dedicated to languages other than English, such as, for instance, the Italian component of the EuroWordNet project (i.e. ItalWordNet). The first two sections of the paper will be devoted to briefly introduce the IWN project and the preliminar steps for question analysis. The core of the paper is represented by a sort of case-study dedicate to the description of the way the QA system can access the semantic information in IWN with the goal to derive what we call the Question Focus, the information crucial to match question and answer. Unfortunately, we are not able to provide validated results yet. We are in the process of assembling the available components of the QA downstream (the search engine, the chunker and the dependency parser, as well as the LRs) and we hope to be able to provide the first results soon. The current research is not collocated within a funded project but we hope to find occasion of fundings in the future.

[1] The Ph.D is carried out within a collaboration between Pisa University (Italy) and Istituto di Linguistica Computazionale of the National Council of Research. The grant is funded by the Italian National Council of Research.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 54–59.

2 ItalWordNet

The EuroWordNet (EWN) [11] project, retaining the basic underlying design of WordNet [7], tried to improve it in order to answer the needs of research in the computational field, in particular extending the set of lexical relations to be encoded between word meanings. In the last years an extension of the Italian component of EWN was realized[2] with the name of ItalWordNet (IWN) [10], inserting adjectives and adverbs, but also nouns and verbs which had not been taken into consideration yet in EWN. IWN follows exactly the same linguistic design of EWN (with which shares the Interlingual Index and the Top Ontology as well as the large set of semantic relation[3]) and consists now of about 70,000 word senses organized in 50,000 synsets.

3 Analysis of Italian Wh-Questions and Applicability for QA

Aiming at building a benchmark for Question Answering applications, we will concentrate our attention on factoid *Wh*-questions, which are supposed to be the forms more probably submitted by a user as a query. The corpus for QA consists now of about 800 Italian factoid *Wh*-questions, the majority of which obtained translating the TREC-9 question collection. We had also the opportunity to use the question collection from the first CLEF2003 (CL and monolingual) QA track [6]. The quality of the parser output can plays an important role in a QA application so a specific set of rules for the IDEAL Italian dependency parser [1] has been written[4]. On the other hand, a shallow parser (chunker) for Italian (CHUNK-IT) [4] provides us with the possibility to individuate information crucial for the task of question classification on the basis of the expected answer (i.e. what the user is looking for with his/her question). This information is the Question Stem (QS) and the Answer Type Term (ATT) [9]. The QS is the interrogative element we find in the first chunk of the sentence, while the ATT is the element modified by the QS (e.g. *Quanto costa un kg di pane?* or *Che vestito indossava Hillary Clinton in occasione di... ?*)[5]. The convergence between these two information allows us to get closer to the expected answer type and to the text portion plausibly containing the answer. Some QSs (for example, *Quando* and *Dove*) allow the system to establish univocal correspondences between them and specific QFs. The relation between QF and QS is not bidirectional: to the same type of question can correspond different QFs (e.g. *Come si chiamava la moglie di JFK?* Vs *Come morì Janice Joplin?*[6]), and the same QF can be looked for via different QSs (e.g. *Quale poeta ha scritto la Divina Commedia?* Vs *Chi ha scritto la Divina Commedia?*[7]). We talked about *multi-strategies QA* because each QS has to be dealt with in its specificity. In what follows we will concentrate our attenction only on the interrogative elements of the Italian *Wh*-questions for handling which we have to explore information stored in LRs: the Question Stems *Che* and *Quale*.

[2] Within the SI-TAL project.

[3] For a complete list of the available semantic relations cf. [10]

[4] A detailed description of this phase and the results are in [2]

[5] *How much does a kg of bread cost?* Or *Which dress did Hillary Clinton wear when.....?*

[6] *What is JFK's wife name?* Vs *How did Janice Joplin die?*

[7] *Which poet wrote the Divina Commedia?* Vs *Who wrote the Divina Commedia?*

3.1 (Che|Quale)-questions

In capacity as interrogative adjective, *Che* is ambiguous between an interpretation selecting individuals and classes: when it is used to ask about an individual to be chosen among a group it overlaps, especially in North Italy, to the interrogative element *Quale*. For both, it is true the same consideration: generally, the QF refers to the entity belonging to the type of the noun modified by the interrogative adjective. For example, the answer of a question like: *Quale mammifero vive in mare?*[8] can be extracted from sentences like: *la balena vive nell'Oceano Atlantico*[9] where the informative links allowing the recognition of the answer are:

{Balena 1} –HAS_HYPERNYM → {cetaceo 1} –HAS_HYPERNYM → {mammifero 1};
{Atlantico 1} –BELONGS_TO_CLASS → {oceano 1} –HAS_HYPERNYM → {acque 1} –HAS_HYPONYM → {mare 1};

In this case we can lexically single out the QF searching among the hyponyms of the noun. This type of question is one of the most complex since the system has to resort to an additional lexical-semantics analysis module and the exploitation of language resources can make the difference. The need of an information stored in a lexical-semantics resource is also evident when we find questions like: *Quale stretto separa il Nord America dall'Asia?*[10] and *Quale parco nazionale si trova nello Utah*?[11].

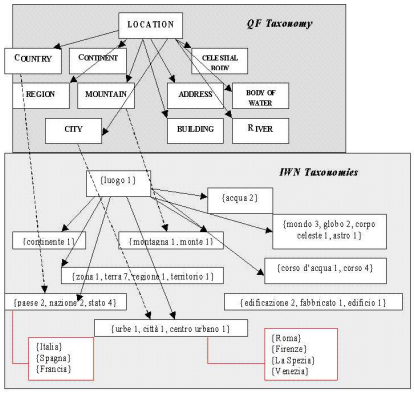

Fig. 1. Mapping the node Location of the QfTaxonomy on IWN

The semantic type of the noun modified by the interrogative adjective is the only thing able to tell us that we have to look for a named entity of the type *location* in the candidate

[8] *Which mammal lives in the sea?*
[9] *Whales live in Atlantic Ocean.*
[10] *Which strait separates North America and Asia?*
[11] *Which national park is in Utah?*

answer. These questions are not introduced by the interrogative adverb *Dove* (*Where*), but they are indeed used to ask about a location. But how do we derive the information that maps the *stretto* or the *parco nazionale* of the questions into the QF Location? In IWN, {parco nazionale 1} is a hyponym of {territorio 1, regione 1, zona 1, terra 7} while {stretto 1} is a hyponym of {sito 1, località 1, posto 1, luogo 2} and these areas of the IWN taxonomies can easily be mapped onto the Question Focus Location. The problem is that, when we want to project the QF Location on the IWN taxonomies, we have to address it on scattered and different portions of the semantic net. The node Location of the Question Focus taxonomy is mappable on the synset {luogo 1 – *parte dello spazio occupata o occupabile materialemte o idealmente*}, that has 52 first level hyponyms and that can be further organized in other sub-nodes, such as: country, river, region, etc. The major part of these taxonomies is leaded by the same synset {luogo 1}, which circumscribes a large taxonomical portion that can be exploit in the QF identification. To this area we also have to add other four sub-hierarchis {corso d'acqua 1, corso 4 – *l'insieme delle acque in movimento* }, {mondo 3, globo 2, corpo_celeste 1, astro 1}, {acqua 2 – *raccolta di acqua*}, {edificazione 2, fabbricato 1, edificio 1 – *costruzione architettonica*}. Figure 1 gives an idea of this situation: the circumscribed taxonomical portion includes the nodes directly mapped on the QFs, all their hyponyms (of all levels) and all the synsets linked to the hyerarchy by means of the BELONGS_TO_CLASS/HAS_INSTANCE relation. A different way to group the IWN lexical items together is recurring to the EWN Top Ontology (TO). The EWN architecture allows us to select and circumscribe wide lexicon portions, kept together by: i) the links between the monolingual database and the ILI portion hosting the Base Concepts, ii) the links between the Base concepts and the TO, iii) the ISA relations linking the synset corresponding to the Base Concept with its conceptual subordinates of n level, from the top to its leaf nodes. In the case of QF Location, for example, we can extract all the synsets belonging to the Top Concept PLACE. The problem is that *River*, *Celestial_Body* and *Building* belong to other ontological portions ((*River* and *Celestial_Body* are classified as *Object/Natural* while *Building* as *Artifact/Building/ Object*) (see Figure 2). The Top Concepts *Object* and *Artifact* are too generic and not discriminating in the selection of the lexical area pertinent to the respective QFs. Thus the exploitation of the Top Ontology nodes can not be the default methodology for indivituating the relevant synsets[12]. The case of Location is only an example of the necessity to (manually) link the highest and most pertinent nodes of the lexical resources to the QFtaxonomy. We are now in the process[13] of adding a new module containing the almost 50 nodes of the QFTaxonomy to the IWN data structure, specifying, when possible, the subsumption links between the synsets and the type of expected answer. The internal ontological structure of ItalWordNet is obviously very different from the QFTaxonomy and it seems that the above mentioned strategy is much more practicable when working with concrete entities than with abstract entities. In *Quali conseguenze ha la pioggia acida?*[14], the candidate answer *L'impoverimento del terreno deriva dalle pioggie acide*[15] contains the answer element *impoverimento*, which is a

[12] The hypothesis of a hybrid strategy which uses both the Top Concepts and the lexical nodes has to be evaluated.

[13] Using the ItalWordNet tool.

[14] *Which are the consequences of the acid rain?*

[15] *the impoverishment of the soil derives from acid rain*

direct hyponym of the abstract noun *conseguenza*.[16] But in the question-answer pair: *Quale funzione ha la milza? La milza produce linfociti*[17] there is no hyponymy relation between *funzione* and *produrre*. In this case we should be able to resort to more complex inferences, as we see in Figure 3.

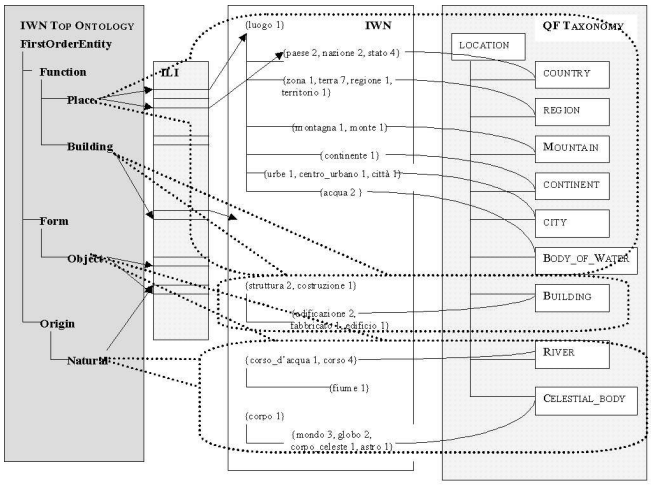

Fig. 2. Projection of the nodes of the QF Location on the EWN TO

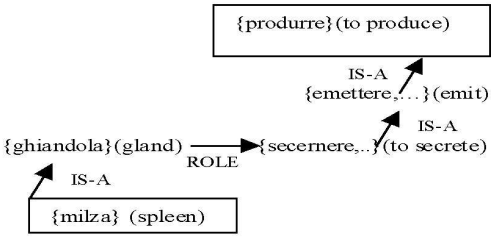

Fig. 3. An inferential path through the IWN synsets

[16] Another informative link is the semantic relatedness between the verb *derivare* (*to derive*) and the noun *conseguenza* (*consequence*), expressed in IWN by mean of a XPOS_NEAR_SYNONYM link between the synsets {derivare 1, conseguire 3,. . . , risultare 1} and {risultato 1, esito,. . . ., conseguenza 1}.

[17] *Which is the function of the spleen? The spleen products lymphocytes*

4 Future work

In the next step of our work we will try to provide a systematic analysis of the types of inference needed in the task of matching question and answer (very insightful in this sense is the work on *lexical chains* by [8]). We will verify whether it is possible to derive such inferences from the connections already stated in IWN by mean of the large set of semantic relations. It has to be avaluated also the impact of dynamic extraction of paraphrasis and inferential rules from texts [3,5], which constitutes a bottom-up approach leading to a notion of *meaning* inspired by distributional criteria. The idea is that dynamically boosting the "inferential" potentialities of static, hand-generated LRs can plays an important role in filling the gap between question and answer and, more generally, that the interplay between static lexical information and dynamic information acquired from text via processing is one of the way LRs could be improved and renewed in the future.

References

1. Bartolini R., Lenci A., Montemagni S., Pirrelli V., *Grammar and Lexicon in the Robust Parsing of Italian: Towards a Non-Naïve Interplay*, in Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan (2002).
2. Bertagna F., *Parsing Italian Wh-Questions*, ILC Internal Report, in prep.
3. Hermjakob U., Echihabi A., Marcu D., *Natural Language Based Reformulation Resource and Web Exploitation for Question Answering*, Proceeding of TREC-2002, (2002).
4. Lenci A., Montemagni S., Pirrelli V., *CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation*, in Linguistica Computazionale, Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN 0392–6907 (2001).
5. Lin D., Pantel P., *Discovery of Inference Rules for Question Answering*. In Natural Language Engineering 7(4):343–360 (2001).
6. Magnini B., Romagnoli S., Vallin A., Herrera J., Penas A., Peinado V., Verdejo F., de Rijke M., *The Multiple Language Question Answering Track at CLEF2003*, Working Notes for the CLEF2003 Workshop, Norwey (2003).
7. Miller, G., Beckwith R., Fellbaum C., Gross D., Miller K. J., *Introduction to WordNet: An On-line Lexical Database*. In International Journal of Lexicography, Vol.3, No.4 (1990) 235–244.
8. Moldovan D., Harabagiu S., Girju R., Morarescu P., Lacatusu F., Novischi A., Badulescu A., Bolohan O., *LCC Tools for Question Answering*, Proceeding of TREC-2002 (2002).
9. Paşca M., *Open-Domain Question Answering from Large Text Collections*, CSLI Studies in Computational Linguistics, USA (2003).
10. Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A., *ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian*. In Zampolli A., Calzolari N., Cignoni L. (eds.), Computational Linguistics in Pisa, Special Issue of Linguistica Computazionale, Vol. XVIII-XIX, Istituto Editoriale e Poligrafico Internazionale, Pisa-Roma (2003).
11. Vossen, P. (ed.), *EuroWordNet General Document*, 1999.