

Italian Language Resources in a Question Answering Task

Francesca Bertagna

*Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche,
Via Moruzzi 1, 56100 Pisa, Italy*

Abstract

The paper describes the first phase of a work dedicated to the exploitation of linguistic resources in a QA application for Italian. In particular, the form of Italian *Wh*-questions introduced by *Quale* and *Che* is analyzed, in order to highlight the information that is crucial for singling out the answer(s) in a document collection. A preliminar investigation of the modalities of linguistic resources exploitation will be presented.

Key words: Language Resources, WordNet, Question Answering

1 Introduction

The aim of the present paper is the description of the first phase of the work carried out during an ongoing Ph.D. research 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. LRs like i) WordNet, ii) FrameNet and iii) the SIM-PLÉ lexicons etc..are not conceived to meet the requirements of a specific task but rather to represent a sort of *repository* of information of general interest. Nonetheless, they are significant sources of linguistic knowledge that should allow systems to automatically perform inferences, distinguish senses, retrieve information, etc.. Tons of papers have been written about the use of WordNet in IR and in QA but the time is mature to test also resources dedicated to languages other than English. At the Istituto di Linguistica Computazionale of

Email address: francesca.bertagna@ilc.cnr.it (Francesca Bertagna).

the CNR, two lexical resources for Italian have been built during the past years and are now available for testing: the Italian component of the EuroWordNet project (ItalWordNet) and the Italian lexicon belonging to the SIMPLE family (CLIPS).

Question Answering allows us to face many different problems and NLP sub-tasks and to study the crucial interplay between dynamic linguistic processing modules and static lexical resources: the system has to derive specific information of syntactic nature from the question and from the corpus and to use that same information to selectively navigate and exploit the content of the LRs. Moreover, QA incorporates an IR module that can be enriched by means of consolidated techniques of morphological and lexical query expansion allowing us to try out the LRs in one of their "natural" tasks.

Section 2 of the paper will be devoted to briefly introduce the two Italian lexicons, their linguistic designs and the types of information they store. In section 3 we will present the construction schema of the QA prototype we are building in Pisa. The core of the paper (section 4) is represented by an analysis of Italian factoid *Wh*-questions introduced by *Quale* and *Che*(Which) carried out with a special attention to the task of question classification on the basis of the expected answer type. We will try to describe how the QA system interacts with semantic information in LRs, in order to match question and answer.

2 LRs content description

2.1 SIMPLE

The SIMPLE (Lenci et al. 2002) project was aimed at building wide-coverage, multipurpose and harmonised computational semantic lexicons linked to the morphological and syntactic ones which were elaborated for 12 European languages during the PAROLE project. the SIMPLE model covers a great range of information and its model and architecture offer the opportunity to deal with natural language complexity by providing a highly expressive and versatile way for language content description. In SIMPLE, the basic information unit is the *SemU* (Semantic Unit), used to encode word sense. To each SemU is assigned a *Semantic Type* involving structured information organized in the four Qualia Roles adopted in the Generative Lexicon framework. A schematic structure, the *Template*, helps the lexicographer to encode a given lexical item in a harmonized way. For each SemU many types of information can be expressed, e.g. the Semantic Type¹, the Domain, the lexicographic definition,

¹ i.e. a specific position within a structured ontology of concepts, e.g. the "Animal"- "Living Entity" Semantic Type for "dog".

the argument structure and the selectional restrictions², the Event type for verbs³, the link of the arguments to the syntactic subcategorization frames, Qualia information⁴, information about regular polisemous alternation, cross-part of speech relations⁵, synonymy relations etc.. The Italian component of SIMPLE (CLIPS) consists today of about 28.000 SemUs (nouns, verbs and adjectives)

2.2 ItalWordNet

Since the Princeton WordNet database, a semantic network in which the meanings of words are represented in terms of their conceptual-semantic and lexical relations to other words (Miller et al. 1990) has become available, it has been the tool of choice for researchers aiming at building Natural Language Processing systems of various kinds. The basic notion around which WordNet is built is the *synset*, the set of synonymous words with the same Part-of-Speech (PoS) that can be interchanged in a certain context. The main goal of the EuroWordNet (Vossen 1999) project was to develop a multilingual lexical resource, retaining the basic underlying design of WordNet while at the same time trying to improve it in order to answer the needs of research in the computational field. In fact, whereas in WN1.5 a rigid distinction is drawn among different PoSs and each PoS forms a separate system of language-internal relations, in EWN various relations between different PoSs can be encoded. This allows to establish cross-PoS relations between words of the same semantic order referring to similar concepts (e.g., cross-PoS synonymy between *arrival* and *to arrive*, etc.). In EWN the set of lexical relations to be encoded between word meanings was extended or modified in various ways with respect to the set defined in WN1.5. For example, a set of role-involved relations can be used to keep together {musician}, {to play} and {guitar}, while a *be_in_state* link can be established between {poverty} and {poor}⁶. In the last years, an extension of the Italian component of EWN was realized⁷: ItalWordNet (IWN) (Roventini et al. 2003). The IWN database is constituted by:

- a generic wordnet which contains about 70,000 word senses corresponding to about 50,000 synsets;

² which allow us to specify if the SemU is a predicate and which are its arguments and their semantic type.

³ to characterize the actionality behaviour of the SemU, e.g. "to run" is categorized as a Process, "to have" as a State etc..

⁴ Qualia are the formal, agentive, constitutive and telic roles provided by relations between SemUs or by features.

⁵ e.g. "intelligent" - "intelligence"; "writer" - "to write".

⁶ For a complete list of the available relations see (Vossen 1999).

⁷ Within the SI-TAL project.

- an Interlingual-Index (ILI) which is an unstructured version of WN1.5, used in EWN to link wordnets of different languages;
- the Top Ontology (TO), a hierarchy of language-independent concepts, reflecting fundamental semantic distinctions (first, second and third order), built within EWN and partially modified in IWN to account for adjectives (which were not dealt with in EWN). Via the ILIs, all the concepts in the generic and specific wordnets are directly or indirectly linked to the TO;
- terminological wordnets.

3 Proposal for a QA Prototype System

What follows represents an attempt to organize a prototype QA application for Italian which exploits semantic information available in LRs. The proposed architecture, heavily inspired by the FALCON (Harabagiu et al. 2000), (Paşca 2003) and by the PIQASso (Attardi et al. 2001) systems, relies on three basic modules:

- in the first one, a detailed analysis of the question is performed in order to obtain: i) the vectorial representation of the question keywords that will be used in the IR module, ii) the Question Stem (QS) and Answer Type Term (ATT) extracted from the sentence analyzed with a shallow parsing technique, iii) the dependency representation of the question that will be compared against the dependency representation of the candidate answer, iv) the Question Focus notion that both defines the type of expected answer and provides the "semantic" type of the expected answer element. The different types of QF have been organized in a hierarchical structure called QFTaxonomy, whose nodes have been projected both on the ItalWordNet and SIMPLE data and on five Named Entity categories.
- the second module consists of a document indexing and retrieval sub-system, i.e the Paragraph Search Engine (Attardi et al. 2001) already used in the PIQASso QA system that will take in input a text collection and will provide in output a list of paragraphs matching the query vector.
- the last module represents the place where all the information collected during the first phase of question analysis will be of use. A system of filters rules out candidate paragraphs not satisfying a certain set of semantic and syntactic constraints. Only if no paragraph passes the filter series, query reformulation methods are performed

At the moment, only part of the prototype has been implemented. In particular, the prototype is now able to perform the transformation of the question in the vector of terms used by the Paragraph Search Engine to extract relevant paragraphs (we are now in the process to add a stemmer to this phase). Moreover, the prototype derives the QS and the ATT from the output of the

chunker and the grammatical relations from the analysis of the dependency parser. All the information derived from the question is saved in an *ad-hoc* XML data structure.

4 Analysis of Italian Wh-Questions and Applicability for QA

CHUNKIT (Lenci et al. 2001) is a shallow parser which analyzes sentences as non recursive, flat syntactic structures (the *chunks*). We used CHUNKIT to individuate the Question Stem and the Answer Type Term (Paşca 2003). The QS is the interrogative element we find in the first chunk of the sentence (*Cosa, Chi, Quando, etc.*⁸), while the ATT is the element modified by the QS (e.g. *Quanto costa un kg di pane?*⁹ or *Che vestito indossava Hillary Clinton in occasione di...?*¹⁰). In what follows we will concentrate only on the interrogative elements of the Italian *Wh*-questions for handling which we have to explore information stored in LRs: the Question Stem *Che* and *Quale*¹¹. 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* (Renzi et al. 1995). 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?*¹² can be extracted from sentences like: *la balena vive nell'Oceano Atlantico*¹³, where the informative links allowing the reconstruction of the answer are: {balena 1} ISA {cetaceo 1} ISA {mammifero 1} {Atlantico 1} BELONGS_TO_CLASS {oceano 1} ISA {acque 1} ISA {mare 1}. In this case, as well in many others, we can lexically single out the QF searching among the hyponyms (of all levels) of the noun. The need of a information stored in a lexical-semantic resource is also evident when we find questions like: *Quale stretto separa il Nord America dall'Asia?*¹⁴ and *Quale parco nazionale si trova nello Utah?*¹⁵. 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 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

⁸ What, Who, When etc..

⁹ *How much does a kg of bread cost?*

¹⁰ *Which dress did Hillary Clinton wear when..?*

¹¹ Which

¹² *Which mammal lives in the sea?*

¹³ *Whales lives in Atlantic Ocean.*

¹⁴ *Which strait separates North America and Asia?*

¹⁵ *Which national park is in Utah?*

we derive the information that map 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, localit 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}¹⁶, that can be further organized in at least 10 sub-nodes, such as i) *country* (mappable on {paese 2, nazione 2, stato 4}¹⁷, ii) *river*, {fiume 1}¹⁸, iii) *region*, {zona 1, terra 7, regione 1, territorio 1}¹⁹, etc..

The major part of these taxonomies is leaded by the same synset {luogo 1}, which circumscribes a large taxonomical portion that can be exploited for the identification of QF. To this area we added other four sub-hierarchies {corso d'acqua 1, corso 4²⁰}, {mondo 3, globo 2, corpo celeste 1, astro 1}²¹, {acqua 2}²², {edificazione 2, fabbricato 1, edificio 1}²³. The resulting area includes the nodes directly mapped on the QFs, all their hyponyms (of all levels) and all the synsets linked to the hierarchy by mean of the BELONGS_TO_CLASS/HAS_INSTANCE relation. A different way to group the IWN lexical items together is resorting to the EWN Top Ontology, which selects lexical units 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 QFs for the question about Location, for example, we can extract all the synsets belonging to the Top concept Place. But, only the QFs Country, Region, Mountain, Continent, City and Body of water can be projected on this wide category, while River, Celestial Body and Building belong to other ontological portions (River e Celestial Body are classified as Object/Natural while Building as Artifact/Building/Object). Thus, the Top Concepts Object and Artifact are too generic and not discriminating in the selection of the lexical area pertinent to the respective QFs. The exploitation of the Top Ontology nodes can not be the default methodology for individuating the relevant synsets, but it is possible to hypothesize a hybrid strategy which uses both the TCs and the lexical nodes. Establishing links between the QFTaxonomy and the ontological structures of the lexicon seems to be a highly recommended

¹⁶ Place.

¹⁷ State, Nation.

¹⁸ River.

¹⁹ Region, Zone.

²⁰ Water flow.

²¹ Celestial body.

²² Body of water.

²³ Building.

strategy in the case of the use of SIMPLE(/CLIPS). The SIMPLE Ontology is more detailed than the IWN one (157 Templates Vs 68 Top Concepts in IWN) and seems to be adapt to select and circumscribe rather homogeneous subsets. But, again, in order to represent Celestial Body, the manual selection of a common and shared hyperonym in the lexicon is necessary. Also in the hypothesis of using CLIPS, a hybrid strategy that allows the system to reach the lexical items (the SemUs) via both the ontological information and the lexical nodes is needed. At the moment, we are mapping all the nodes of the Qftaxonomy onto the lexical units in IWN (using the tool for editing data) and onto the SIMPLE Ontology. Many times, the above mentioned strategy is not practicable, as we can see in the question-answer pair: *Quali sono le conseguenze della pioggia acida? - L'impoverimento del terreno deriva dalle piogge acide*²⁴. In this case the only possible link is represented by a certain semantic contiguity between the verb *derivare* and the noun *conseguenza*. In IWN a XPOS_NEAR_SYNONYM link exists between the synsets {derivare 1, conseguire 3,..., risultare 1} and {risultato 1, esito,..., conseguenza 1}. The problem is to find a way to exploit information conveyed by relations different from hyperonymy.

5 Concluding Remarks

In this paper, we provided an overview of the modalities of gathering and processing information from Language Resources for a Question Answering system. We saw that IWN and SIMPLE have to be manually inspected and linked to the QFTaxonomy in order to provide the system with useful paths of navigation and exploration of the data. The possibility to exploit a larger part of the rich connectivity of language resources is something that should be carefully evaluated and studied in the future. The various expressive modalities of LRs convey a rich amount of semantic information but, although the connectivity in LRs can be an answer to reasoning needs, we have to carefully consider the problems arising when automatically handling information that can lead us on wrong paths, far way from the desired target. The work on lexical chains by (Moldovan et al. 2002) is very interesting under this point of view. Moreover, following (Hermjakob et al. 2002) and (Lin et al. 2001), we think that much help can derive from the dynamic extraction of paraphrases and inferential rules from texts. Dynamically boosting the "inferential" potentialities of static, hand-generated LRs could play an important role in filling the gap between question and answer.

²⁴ *Which are the consequences of the acid rain? The impoverishment of the soil derives from acid rain.*

References

- [Attardi et al. 2001] Attardi G., Cisternino A., Formica F., Simi M., Tommasi A., Zavattari C., *PIQAsso: Pisa Question answering System*, in Proceeding of the 10th TREC Conference, 2001.
- [Harabagiu et al. 2000] Harabagiu S., Moldovan D., Paşca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus R. and Morarescu P., *FALCON: Boosting Knowledge for Answer Engines*, in Proceedings of the Text Retrieval Conference (TREC-9), 2000.
- [Hermjakob et al. 2002] Hermjakob U., Echihabi A., Marcu D., *Natural Language Based Reformulation Resource and Web Exploitation for Question Answering*, Proceeding of TREC-2002, 2002.
- [Lenci et al. 2002] Lenci A., Bel N., Busa F., Calzolari N., Gola E., Monachini M., Ogonowsky A., Peters I., Peters W., Ruimy N., Villegas M., Zampolli A., *SIMPLE: A General Framework for the Development of Multilingual Lexicons*. In International Journal of Lexicography, XIII (4), 249-263, 2000.
- [Lenci et al. 2001] 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.
- [Lin et al. 2001] Lin D., Pantel P., *Discovery of Inference Rules for Question Answering*. In Natural Language Engineering 7(4):343-360. 2001.
- [Miller et al. 1990] 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, 235-244, 1990.
- [Moldovan et al. 2002] 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.
- [Paşca 2003] Paşca M., *Open-Domain Question Answering from Large Text Collections*, CSLI Studies in Computational Linguistics, USA, 2003.
- [Renzi et al. 1995] Renzi L., Salvi G., Cardinaletti A. (eds.), *Grande grammatica italiana di consultazione*, Bologna, Il Mulino, 1995.
- [Roventini et al. 2003] 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.
- [Vossen 1999] Vossen, P. (ed.), EuroWordNet General Document, 1999.