Using Semantic Language Resources to Support Textual Inference

for Question Answering

Francesca Bertagna

Istituto di Linguistica Computazionale, CNR,
Via Moruzzi 1, 56100 Pisa, Italy
francesca.bertagna@ilc.cnr.it

Abstract

The paper presents an experiment aiming at verifying whether the rich connectivity of two Italian Language Resources, ItalWordNet and CLIPS, can be exploited to support inference for Question Answering. The methodology for the creation of primitive inferential rules from the ItalWordNet and CLIPS semantic relations is described. We also introduce some preliminary results about the usefulness of such inferential rules for QA.

1 Introduction

One of the things that makes of Question Answering a so challenging task is the necessity to go beyond the literal form of the query and of the answer. In the attempt to fill the gap between the question and the candidate answer, the system has to "*understand*" natural language, handling some representation of the meaning of the two texts and performing textual inference by extracting relevant, unstated information.

A way to tackle this challenge is resorting to lexico-semantic Language Resources (LRs), which are supposed to provide an explicit and machine understandable representation of "*word meaning*" that can be exploited by intelligent agents as source of knowledge for supporting inference. Testing activity in real tasks can thus be conceived as a possible way to evaluate the heuristic and predictive value of the *word meaning* as instantiated in various LRs.

The paper will provide an exemplification of some textual inferences emerging in the process of matching question and answer (the work has been carried out for Italian). Using, for illustration purposes, the questions and some text snippets of the CLEF2003-QA track collection (Magnini et al. 2003), we will exemplify which inferential steps are taken when identifying an answer in a particular portion of text. We will then verify whether it is possible to derive such inferences from the connections already stated in LRs by means of their expressive modalities and linguistic designs. The greater part of open-domain QA systems exploit only a small part of the information available in LRs (typically the hypo\hyperonym links during the question classification phase and the synonyms in the query expansion). We will explore the content of two language resources available for Italian: the ItalWordNet database and the CLIPS semantic lexicon. In these computational lexicons we find a significant amount of semantic knowledge and we want to evaluate if/at what extent their rich connectivity can be an answer to QA need.

2 Background

The work has been carried out following the strategy presented in (Moldovan et al. 2002 and Harabagiu et al. 1998), where *primitive inference rules* are implemented as pairs of WordNet semantic relations and are further combined to generate more complex rules. In (Harabagiu et al., 1998), in order to increase the number of links between concepts and in particular to retrieve the important cross-part of speech connections missing in WordNet, an interesting methodology as been followed to extract information from WordNet glosses. The result is a much more richer connectivity between concepts, expressed by means of 13 new relations such as AGENT, OBJECT, PURPOSE, ATTRIBUTE etc. The "classic" WordNet relations plus the new *defining features* constitute the units on which the inferential rules are built on. We adopted this same methodology to discover significant inferential paths through the large set of semantic relations of ItalWordNet and through the rich connectivity (ranging from the argument structure to the qualia roles) of CLIPS. One of the things that distinguish our experiment from Harabagiu et al.'s work is that the types of information that in (Harabagiu et al., 1998) are derived from the WordNet glosses are supposed to be already available in EuroWordNet and CLIPS, respectively represented in terms of ROLE relations and thematic roles of the predicate. For this reason, before describing the experiment of deriving inferential rules from the two language resources, we will briefly introduce the linguistic design both of ItalWordNet and CLIPS, focussing on the aspects that are more relevant for our task.

3 ItalWordNet and CLIPS

ItalWordNet (Roventini et al., 2003) represents the further development, within a National project, of the Italian component of the EuroWordNet (EWN) multilingual database (Ide et al., 1998). From EWN it derives the general linguistic design and the whole set of lexical semantic relations (few relations were added in order to better represent the semantics of adjectives). Many types of semantic relation are thus available, ranging from the ones already encoded in the Princeton WordNet (such as HYP(ER)ONYMY, HOLO(/MERO)NYMY, CAUSE-ENTAILMENT etc..) to the cross part-of-speech relations (such as ROLE/INVOLVED, XPOS_NEAR_SYNONYM, IN_MANNER/MANNER_OF etc..)[1]. ItalWordNet contains about 50.000 synsets, distributed on nouns, verbs, adjectives, adverbs and intances (Proper Names).

---

1 For a complete list of IWN relations, see (Roventini et al., 2003).

The other LR under analysis is CLIPS (Ruimy et al., 2003), a multi-layered computational lexicon consisting of about 55.000 lemmas encoded at phonological, morphological, syntactic and semantic level (the levels of linguistic description are mutually independent although connected). Each semantic entry comprehends many information, such as membership to an ontological type, domain of use, gloss, event type (for event-denoting entries), morphological derivation relation, polysemy, synonymy, distinctive features and predicative representation at the interface between syntax and semantics. A substantial part of the information is encoded by means of the Extended Qualia Structure, that enables the expression of orthogonal aspects of word meaning (instead of unidimensional inheritance conveyed by standard hyperonymy).

4 Creating Primitive Rules on EuroWordNet and CLIPS Semantic Relations.

Enabling the recognition of inferential paths could play an important role in filling the gap between the question and the answer, as it is evident in the following example: Q: "*Quale funzione ha la milza?*" ("Which is the function of the spleen?") A: "*La milza produce linfociti*" ("The spleen products lymphocytes"). In this case, since no direct relation is established between "*funzione*" (function) and "*produrre*" (to product), in order to recognize the possible answer within the paragraph, the system should resort to a complex inference. In this case the significant path through ItalWordNet would be:
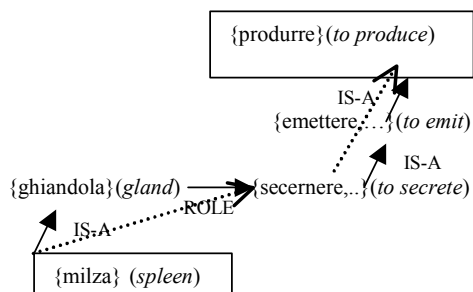

Fig. 1: Semantic path between *milza* and *produrre*

In this example, the ISA inheritance mechanism triggers the inferential rules (represented in the picture by dotted arrows) which allow us to derive:

milza ISA ghiandola+ ghiandola ROLE secernere=
milza iROLE secernere
milza iROLE secernere+secernere ISA emettere=
milza iROLE emettere
milza iROLE emettere+emettere ISA produrre=
milza iROLE produrre
The primitive rules are thus:
c_1 ISA c_2+ c_2 ROLE c_3= c_1 iROLE c_3
c_1 iROLE c_2+c_2 ISA c_3=c_1 iROLE c_3

Starting from the complete list of the almost 75 EWN semantic relations, we have studied all the possible

relation pairs[2]. Not all the available relations can be combined to generate valid primitives since some relations can be applied only to specific POSs (it is not possible to combine, in this order, a ROLE relation, which applies between nouns or between a noun and a verb, with a MANNER_OF relation, which goes from an adverb[3] to a noun or a verb). By avoiding combinations not respecting the right POS concatenation, we obtained 603 relation pairs.

Moreover, the fundamental EWN distinction between First, Second and Third Order Entities[4] prevents us from pairing relations whose concatenation doesn't respect correct entity order (in this sense a HAS_HOLONYM, which applies between first order entities, and an INVOLVED, that links a second order entity to a first order entity, cannot be combined). We found about 80 cases of this type.

At the end, about 480 formally valid relation pairs was formed and evaluated. When having to choose a name for the result of the concatenation we preserved, if possible, the name of "normal" EWN relations, preceded by the "i" of "inferential" in small letter. This allows us to more easily create complex inferential rules resulting in further concatenations of relation pairs. Moreover, we preferred to eliminate, in the resulting name, any indication of the cross-parts of speech nature of the relation. This because the primitive rules are supposed to represent a totally semantic link between not adjacent concepts and any reference to morphosyntactic features of the relation is not meaningful. For example:

accogliere (*to accept*) NEAR_SYNONYM aderire (*to agree*)+
aderire HAS_XPOS_HYPERONYM azione (*action*)=
accogliere iHAS_HYPERONYM azione

The same work has been done for the almost 75 Simple/CLIPS relations. There is a pretty good overlap between the sets of CLIPS and IWN relations, even if the relations are not exactly the same: for instance, in IWN there is no an exact equivalent of the CONCERNS relation we find in CLIPS (which links a phenomenon with the thing it effects[5]) but it can be interpreted as a particular type of the more general INVOLVED IWN relation. At the same way, the very specific KINSHIP relation (which links some animals to their "family"[6]) could be related to the more general HAS_HYPERONYM IWN relation. We could thus verified which was the most close IWN correspondent of each relation in CLIPS, making in this way inherit by CLIPS the inferential primitives built upon the IWN relations. This allowed us to not analyze every

---

2 Many relations are subsumed by more general relation types (for example the HAS_MERO_MADEOF is a special type of the MERONYMY relation) so the inferential rules are built upon classes of relations.
3 In EWN also an adjective could be the source of a MANNER_OF relation but for Italian we restricted the application of this relation to adverb.
4 The distinctions roughly corresponds to the one between abstract (1st and 3rd orders) and concrete (2nd order) entities but see (Ide et al., 1998) for more details.
5 For example: *epatite* (hepatitis) CONCERNS *fegato* (liver)
6 For example: *caprone* (he-goat) KINSHIP *capra* (goat)

CLIPS relation pairs, partially reusing the work already done on ItalWordNet.

So, the CLIPS relations IS_THE_ACTIVITY_OF, IS_THE_ABILITY_OF, IS_THE_HABIT_OF, AGENTIVE_PROG and AGENTVERB behave as the same way as the group of the ROLE relations in IWN, allowing us to create the primitive inferential rule:

C1 (IS_THE_ACTIVITY_OF| IS_THE_ABILITY_OF| IS_THE_HABIT_OF| AGENTIVE_PROG|AGENTVERB) C2+
C2 (KINSHIP|ISA|AGENTIVE_CAUSE|AGENTIVE_EXPERIENCE) C3=
C1 I(IS_THE_ACTIVITY_OF| IS_THE_ABILITY_OF| IS_THE_HABIT_OF| AGENTIVE_PROG|AGENTVERB) C3

It's obvious that the various restrictions we imposed on the application of the IWN relations (the not allowed concatenations of POSs and entities of certain orders) in this way become valid also for the CLIPS relations.

5 Using inferential rules for QA

Endless possible ways to navigate along the relations in IWN and CLIPS can be discovered, but the key is to find only fundamental concatenations that support inference. We tried, then, to verify if the paths based on this large set of primitive rules can be of any help in the QA task. We have to specify that we haven't implemented yet an automatic procedure to extract the resulting semantic paths: we have worked manually on question-answer pairs of the CLEF QA campaign, extracted using a specific search engine[7]. Potentially, the linguistic design seems adapt to support text inference but the number of available links and connections is too low to be useful on an extended, open-domain task. An example is question_#4: *Quando e' stato stipulato il Trattato di Maastricht?* (When was the Maastricht Treatry draw up?). The three keywords (*Trattato* AND *Maastricht* AND *stipulare* (stemmed)) are not enough to retrieve any passages, while with only (*Trattato* AND *Maastricht*) we obtain a high recall of about 300 paragraphs. But how can the system pinpoint the "answer" among this large set of paragraphs? The presence in the paragraph of a named entity of the type "Date" is not enough to discriminate (since in almost all the paragraphs there is at least one temporal expression).

Searching among all the candidate answers we found 4 interesting paragraphs:

"*...ratifica del Trattato di Maastricht...nell'autunno del 1992*"[8]
"*...conclusione del Trattato di Maastricht nel 1991*"[9]
"*..secondo referendum di ratifica dopo quello..del settembre '92....del Trattato di Maastricht*"[10].

These three paragraphs contains important clues for an answer. They contains keywords that we know being in some way related to the event "a certain Treaty was firmed and its conditions became operative". The paragraphs can be represented, in the same way of the question, as a vector of keywords and a possible way to retrieve the exact answer could be the evaluation of the path connecting the keywords of the question and of the candiate answer. In this case, we recognize that the clues of the answer are the keywords *ratifica* and *conclusione*. Is there any correlation in our LRs between *stipulare* and *ratifica* and *conclusione*?

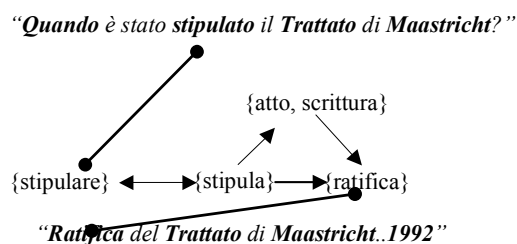The following picture shows the path connecting *stipulare-ratifica* in IWN:

Fig. 2: *ratifica-stipulare* path.

Two different inferential paths are traced by the primitives.
The first path:

stipulare XPOS_NEAR_SYNONYM stipula+ stipula HAS_HYPERONYM atto=
stipulare iHAS_HYPERONYM atto
stipulare iHAS_HYPERONYM atto+atto HAS_HYPONYM ratifica=
stipulare iPLAU_NOT ratifica

In this case, the resulting rule says that *stipulare* and *ratifica* are plausibly mutually exclusive, since they are co-hyponyms of the same synset {atto, scrittura}.
The second path:

stipulare XPOS_NEAR_SYNONYM stipula+ stipula IS_SUBEVENT_OF ratifica= stipulare iIS_SUBEVENT_OF ratifica

In this case, the resulting inferential path confirm the common knowledge that "*something that has been ratified it has first been stipulated*". This type of information is due to the presence of an IS_SUBEVENT_OF relation.
On the contrary, we cannot derive any useful information from the path connecting *stipulare* and *conclusione* (see Fig. 3):

stipulare HAS_HYPERONYM realizzare + realizzare HAS_HYPERONYM far divenire= stipulare iHAS_HYPERONYM far divenire
stipulare iHAS_HYPERONYM far divenire+far divenire HAS_HYPONYM concludere= stipulare iPLAU_NOT concludere
stipulare iPLAU_NOT concludere +atto XPOS_NEAR_SYNONYM conclusione= stipulare iPLAU_NOT conclusione
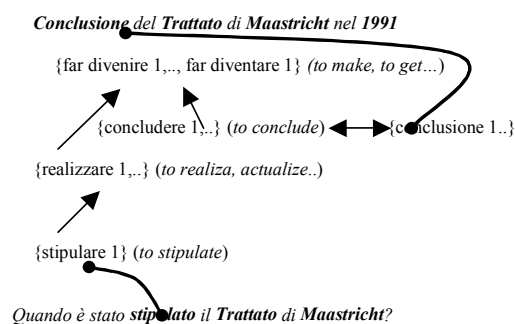

Fig. 3: *Stipulare-conclusione* path

The problem is that in this case the most significant paragraph is the one talking about the conclusion of the Maastricht Treaty in 1991.

6 Conclusions

Results are not incouraging: only a small number of question-answer pairs can be identified using semantic paths since very rarely we found the connections we need in our task.

The two LRs under analysis are very expressive but in general it seems that the high number of semantic relation types is not accompanied by a high number of actually encoded connections of the same type. As a matter of facts, altough the richness of the expressive modalities, the information is not consistently distributes (many relations are very rare, in particular the transitive relations of the type CAUSES/CAUSED_BY and SUBEVENT_OF/HAS_SUBEVENT which would be very useful in a reasoning task).

It doesn't seem a problem of the two specific resources and of the way the lexicographers encoded the specific lexical entries. The case of CLIPS is emblematic: while in IWN the only one non-optional relation is HYPERONYMY, in CLIPS one of the core elements is the *Template*, introduced to enhancing the general consistency of the lexicon by providing structured sets of information which is associated with the semantic type of a SemU. The idea was to provide the lexicographer with guidelines saying which are the minimally required, the recommended and the optional information for the lexical entry at hand. This means that the lexicographers had to encode the relations of telic type when handling lexical entries belonging to the ontological type [Instrument], while they had to encode constitutive relations (part_of etc..) when handling lexical entries about body_part etc..This strategy really plays an important role in make the distribution of information types more consistent and equilibrate. Notwithstanding, rarely we found the relations we need to support complex inferences.

The promising results presented (Moldovan, 2002) seems indicate that the problem could be quantitative: in that case it could be subdued by increasing the number of actual connections, for example by means of an automatic strategy as the one presented in (Harabagiu et al., 1998).

But the problem could also be more *qualitative* than quantitative, connected to the nature itself of the lexical meaning instantiated in LRs. It could be difficult to discover inference chains in a hand-crafted, static LRs, which presuppose a certain notion of word meaning, i.e. static, relational, discrete, in some way context-independent. For (Lin et al. 2001) it is very difficult for humans to encode word meaning with such an awareness to work as basis for sound, robust and effective inference. The next step in our work plan is the preparation of an experiment following the Lin and Pantel's methodology that broadens the scope of Harris' Distributional Hypothesis from the word to the dependency trees of parsed corpus with the aim to automatically discover so-called Inference Rules.

We think that the comparison of the contribution of the two approaches (i.e. the one exploiting static resources and the one based on distributional criteria) to the inferential capability of QA systems may be an input to the open debate about the modalities of representation and conception of the word meaning.

References

Attardi G., Costernino A., formica F., Simi M., Tommasi A., Zavattari C. (2001). PIQAsso: Pisa Question answering System. In Proceeding of the 10th TREC Conference.

Harabagiu S. and Moldovan D. (1998). Knowledge Processing on Extended WordNet, in WordNet: An Electronic Lexical Database and Some of its Applications, Editor C. Fellbaum, (pp. 379-405). MIT Press.

Nancy Ide, Daniel Greenstein, Piek Vossen (eds.)(1998) Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3.

Lin D., Pantel P. (2001). Discovery of Inference Rules for Question Answering. In Natural Language Engineering 7(4) (pp. 343-360).

Magnini B., Romagnoli S., Vallin A., Herrera J., Penas A., Peinado V., Verdejo F., de Rijke M. (2003). The Multiple Language Question Answering Track at CLEF2003, Working Notes for the CLEF2003 Workshop, Norwey.

Moldovan D., Harabagiu S., Girju R., Morarescu P., Lacatusu F., Novischi A., Badulescu A., Bolohan O. (2002). LCC Tools for Question Answering, Proceeding of TREC-2002.

Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A. (2003). 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.

Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri M., Rossi S. (2003). A Computational Semantic Lexicon of Italian: *SIMPLE*. 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.