An Ontology Based Architecture for Translation

Leonardo Lesmo, Alessandro Mazzei and Daniele P. Radicioni
Dipartimento di Informatica, Università degli Studi di Torino
{lesmo,mazzei,radicion}@di.unito.it

Abstract

In this paper we present some features of an architecture for the translation (Italian – Italian Sign Language) that performs syntactic analysis, semantic interpretation and generation. Such architecture relies on an ontology that has been used to encode the domain of weather forecasts as well as information on language as part of the world knowledge. We present some general issues of the ontological semantic interpretation and discuss the analysis of ordinal numbers.

1 Introduction

In this paper we describe some features of a system designed to translate from Italian into Italian Sign Language (henceforth LIS). The system is being developed within the ATLAS project.[1] This architecture applies a *hard* computational linguistic approach: *knowledge-based restricted interlingua* (Hutchins and Somer, 1992). We perform a deep linguistic processing in each phase of the translation, i.e (1) syntactic analysis of the Italian input sentence, (2) semantic interpretation and (3) LIS generation.[2] The main motivation to adopt this ambitious architecture is that Italian and LIS are very different languages. Moreover, LIS is a poorly studied language, so no large corpus is available and statistical techniques are hardly conceivable. We reduce our ambitions by restricting ourselves to the weather forecasts application domain.

In this paper we describe some major issues of the semantic interpretation and illustrate a case study on ordinal numbers. Our semantic interpretation is based on a syntactic analysis that is a dependency tree (Hudson, 1984; Lesmo, 2007). Each word in the sentence is associated with a node of the syntactic tree. Nodes are linked via labeled arcs that specify the syntactic role of the dependents with respect to their head (the parent node). A key point in semantic interpretation is that the syntax-semantics interface used in the analysis is based on an ontology. The knowledge in the ontology concerns an application domain, i.e. weather forecasts, as well as more general information about the world: the latter information is used to compute the sentence meaning. Indeed, the sentence meaning consists of a complex fragment of the ontology: predicate-argument structures and semantic roles are contained in this fragment and could be extracted by translating this fragment into usual First Order Logic predicates.[3]

The idea to use the ontological paradigm to represent world knowledge as well as sentence meaning is similar to the work by Nirenburg and Raskin (2004) and Buitelaar et al. (2009), but in contrast to these approaches (1) we use a syntactic parser to account for syntactic analysis; and (2) we use a recursive semantic interpretation function similar to Cimiano (2009).

2 The Ontology

The ontological knowledge base is a formal (partial) description of the domain of application. It is formal, since its primitives are formally defined, and it is partial, since it does not include all axioms that provide details about the relationships between the involved concepts. The top level of the domain ontology is illustrated in Fig. 1.[4] The classes most relevant to weather forecasts are *££meteo-status-situation*,

[1] http://www.atlas.polito.it/

[2] LIS, as all the signed languages do not have a *natural* writing form. In order to apply linguistic tools designed for written languages, in our project we developed "AEW-LIS", an *artificial* written form for LIS.

[3] However, similar to other approach (among others Bunt et al. (2007); White (2006)), our ontological meaning representation is totally unscoped.

[4] Some conventions have been adopted for ontology names: concepts (classes) have a ££prefix; instances have a £prefix; and relations and relation instances have a & prefix.

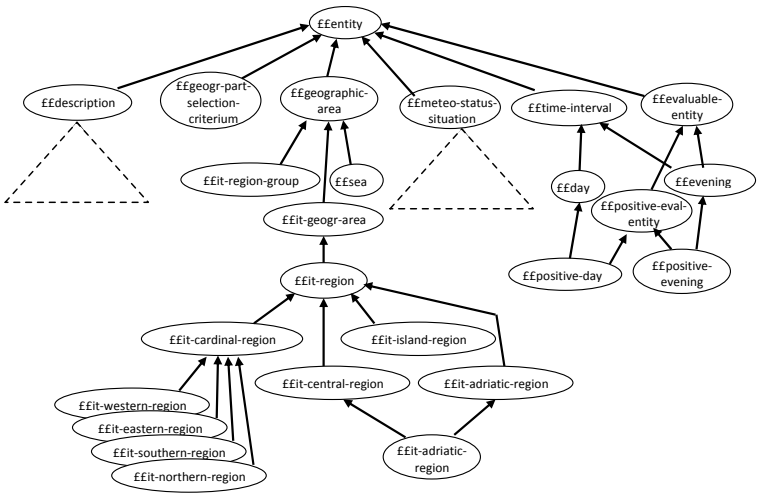

Figure 1: The top ontology used for the weather forecast domain. Dashed triangles represent collapsed regions of the hierarchy.

ℒℒgeographic-area, *ℒℒdescription*, *ℒℒgeographic-part-selection-criterium*.

ℒℒmeteo-status-situation It is the most relevant class in the present setting, since it refers to the possible weather situations, thus providing a starting point –in principle– to every weather forecast. It may concern the sea status, a generic weather status (either stable or not) or possible atmospheric events such as snow, rain or clouds.

ℒℒgeographic-area* and *ℒℒtime-interval Any weather situation holds in a specific place; in particular, the relevant places are geographic areas. A *ℒℒgeographic-area* can be an Italian region, a group of regions, a sea, or may be identified by specifying a cardinal direction (North, South, . . .). Yet, any weather situation holds in a specific temporal interval. Such time interval could last one or more days or a part of a day. Expression as "in the evening" are interpreted anaphorically, i.e. on the basis of current context: if the context is referring to "today", then it is interpreted as "today evening", for "tomorrow" as "tomorrow evening", etc..

ℒℒdescription The actual situation and its description are kept separated. For instance, if *today* is October 28, then "today" is a *ℒℒdeictic-description* of a particular instance (or *occurrence*) of a *ℒℒday*. "April 28, 2010" is another description (absolute) of the same instance. Particular relevance have the deictic descriptions since most temporal descriptions (*today*, *tomorrow*, but also the weekday names, as *Monday*, *Tuesday*, . . .) are deictic in nature.

ℒℒgeogr-part-selection-criterium In descriptions, a particular instance (or group of instances) can be identified by a general class term (e.g. *area*) and a descriptor (e.g. *northern*). This concept refers to the parts of the reality that can act as descriptors. For instance, the *cardinal direction* can be such a criterium for geographic parts, while a *date* is not.

The last relevant portion of the ontology concerns *relations*. Although the ontology has no axioms, class concepts are connected through relevant relations. In turn, relations constitute the basic steps to form paths (more later on). All relations in the ontology are binary, so that the representation of relations of arity greater than 2 requires that they be reified.

3 Semantic Interpretation

One chief assumption in our work is that words meaning can be expressed in terms of ontology nodes, and the meaning of the sentence is a complex path on the ontology that we call *ontological restriction*. We define the *meaning interpretation function* $\mathcal{M}_{\mathcal{O}}$, that computes the the ontological restriction of a sentence starting from the its dependency analysis and on the basis of an ontology $\mathcal{O}$.

Given a sentence S and the corresponding syntactic analysis expressed as a dependency tree $depTree(S)$, the meaning of S is computed by applying the meaning interpretation function to the root of the tree, that is $\mathcal{M}_{\mathcal{O}}(root(depTree(S)))$. In procedural terms, the meaning for a sentence is computed in two steps: (*i*) we annotate each word of the input sentence with the corresponding lexical meaning; (*ii*) we build the

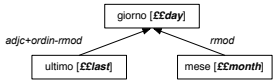

Figure 2: The dependency analysis of *ultimo giorno del mese* (*last day of the month*) enriched with lexical meaning.

actual ontological representation in a quasi-compositional way, by merging paths found in the ontology in a single representation which is a subgraph of the ontology itself. These two steps can be formalized as a meaning interpretation function $\mathcal{M}$ defined as:

$$\mathcal{M}_{\mathcal{O}}(n) := \begin{cases} \mathcal{LM}_{\mathcal{O}}(n) & \text{if } n \text{ is a leaf} \\ \dot{\cup}_{i=1}^{k}(\mathcal{CP}_{\mathcal{O}}(\mathcal{LM}_{\mathcal{O}}(n), \mathcal{M}_{\mathcal{O}}(d_i))) & \text{otherwise} \end{cases}$$

where n is the node of a dependency tree and $d_1, d_2, \ldots, d_k$ are its dependents. $\mathcal{LM}_{\mathcal{O}}(w)$ is a function that extracts the lexical meaning of a word w accessing the dictionary: that is, a class or an individual on the ontology $\mathcal{O}$. $\mathcal{CP}_{\mathcal{O}}(y, z)$ is a function that returns the shortest path on $\mathcal{O}$ that connects y to z. The search for connections relies on the rationale that the shortest path between any two ontology nodes represents the stronger semantic connection between them. In most cases the distance between two concepts is the number of the nodes among them, but in some cases a number of constraints needs to be satisfied too (see the example on ordinal construction). Finally, the operator $\dot{\cup}$ is used to denote a particular merge operator, similar to Cimiano (2009). As a general strategy, shortest paths are composed with the union operation, but each $\mathcal{CP}_{\mathcal{O}}(y, z)$ conveys a peculiar set of ontological constraints: the merge operator takes all such constraints to build the overall complex ontological representation. In particular, a number of semantic clashes can arise from the union operation: we use a number of heuristics to resolve these clashes. For sake of simplicity (and space) in this definition we do not describe the heuristics used in the ambiguity resolution. However, three distinct types of ambiguity exist: (1) lexical ambiguity, i.e. a word can have more than one lexical meaning; (2) shortest path ambiguity, i.e. two nodes can be connected by two equal-length paths; (3) merge ambiguity, i.e. two fragments of ontology can be merged in different manners. Whilst lexical ambiguity has not a great impact due to the limited domain (and could be addressed by standard word sense disambiguation techniques), handling shortest path and merge ambiguities needs heuristics expressed as constraints that rely on general world knowledge.

A particular case of ontological constraints in merge ambiguity is present in the interpretation of ordinal numbers, so further details on the merge operator can be found in Section 4.

4 A case study: the ordinal numbers

In order to translate from Italian into LIS, we need to cope with a number of semantic phenomena appearing in the particular domain chosen as pilot study, i.e. weather forecast. One of the most frequent constructions are ordinal numbers. Consider the simple phrase *l'ultimo giorno del mese* (*the last day of the month*). The (simplified) dependency structure corresponding to this phrase is depicted in Fig. 2: the head word *giorno* (*day*) has two modifying dependents, *ultimo* (*last*) and *mese* (*month*). Since the interpretation relies heavily on the access to the ontology, we first describe the portion of the ontology used for the interpretation and then we illustrate the application of the function $\mathcal{M}$ to the given example.

The relevant fragment of the ontology is organized as shown in Fig. 3, that has been split in two parts. The upper part –labeled *TEMPORAL PARTS*– describes the reified *££part-of* relation and its temporally specialized subclasses. The lower part –labeled *ORDINALS*– is constituted by some classes that account just for ordinal numbers. In the *TEMPORAL PARTS* region of the Fig. we find the *££temporal-part-of* (reified) sub-relation, which, in turn, subsumes *££day-month-part-of*. This specifies that days are parts of months, so that *day of the month* can be interpreted as *the day which is part of the month*. The *££part-of relation* has two roles: we use the term *role* to refer to the binary relation associated with a participant in a reified relation. These roles are "value-restricted" as *&day-in-daymonth* and *&month-in-daymonth* respectively, for what concerns *££day-month-part-of*. The most relevant class in the ORDINALS part of Fig. 3 is the class *££ordinal-description*. It is the *domain* of three roles, *1*) *&ord-described-item*, *2*) *&references-sequence* and *3*) *&ordinal-desc-selector*. The range of the first relation *&ord-described-item* is the item whose position in the sequence is specified by the ordinal, that is a *££sequenceable-entity*. The range of the second relation *&reference-sequence* is the sequence inside which the position makes

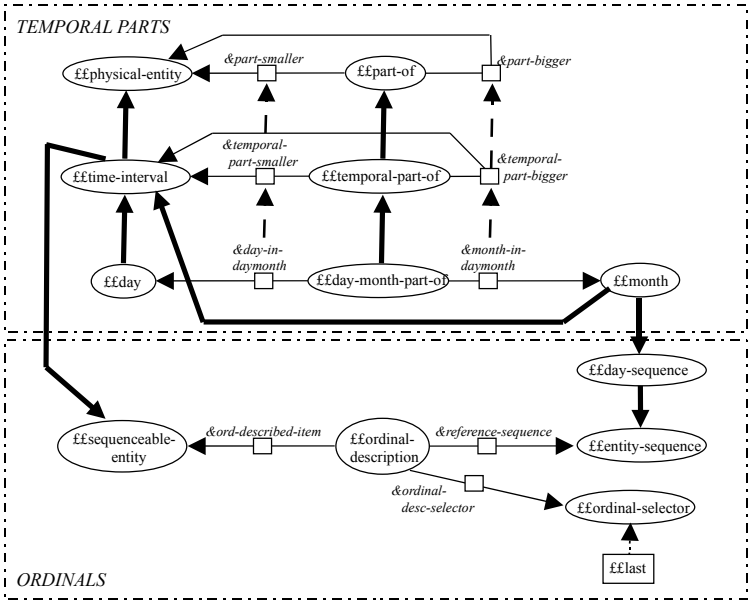

Figure 3: The fragment of the ontology accounting for ordinals.

sense, that is an *££entity-sequence*. The range of the third relation *&ordinal-desc-selector* is item that specifies the position, that is a *££ordinal-selector*. In the example, *£last* is an instance of *££ordinal-selector*. Of course, any (true) ordinal (first, second, thirtythird) can fill that role. The two portions of the ontology are connected by two arcs. The first arc specifies that a *££time-interval* is a subclass of *££sequenceable-entity* (so that one can say *the fourth minute*, *the first year*, and so on). The second arc specifies that *££month* is subclass of *££day-sequence*, which in turn is subclass of *££entity-sequence*. As a consequence it can play the role (can be the range) of the *&reference-sequence*.

We now describe how the meaning interpretation function is applied on the considered example. It consists of three steps: *1.* we compute the connection path between the concepts *££day* and *£last*; *2.* we compute the connection path between *££day* and *££month*; *3.* we merge the connection paths previously computed. In details:

1. By computing $\mathcal{CP}(££day, £last)$ we obtain the connection path in Fig 4-a. Note that this ontological restriction contains the concept *££ordinal-selector*.

2. By computing $\mathcal{CP}(££day, ££month)$ we obtain the connection path in Fig 4-b. In this case the shortest path is not actually the "shortest" one, i.e. the presence of the preposition *del* (*of*) constraints the value returned by $\mathcal{CP}$. Moreover, this ontological restriction contains the concept *££day-month-part-of*, which is a sub-concept of *££part-of*.

3. The last step consists of the application of the meaning composition function to $\mathcal{CP}(££day, £last)$ and $\mathcal{CP}(££day, ££month)$. The *££ordinal-description* concept is detected in the first ontological restriction; moreover *££day* is recognized as (subclass of) a possible filler for *££ordinal-description*. At this point we need establishing how *££day* fits as the smaller part of a *&part-of* relation. We scan the remaining ontological restriction(s) looking for a bigger part involved in a *&part-of* relation or in any of its sub-relations. The resulting representation (Fig. 4-c) is built by assuming that the larger entity (here *££month*, since *&month-in-daymonth* restricts *&part-bigger*) is the reference sequence for the ordering. So, the direct *££day-month-part-of* of the second ontological restriction is replaced by a path passing through *££ordinal-description*. In such final ontological restriction *££day* is the *&ord-described-item* and *££month* is the *&reference-sequence*.

5 Conclusions and future work

In this paper we illustrated the analysis component of a knowledge-based restricted interlingua architecture for the translation from Italian into LIS. The structure produced by the semantic interpretation of the

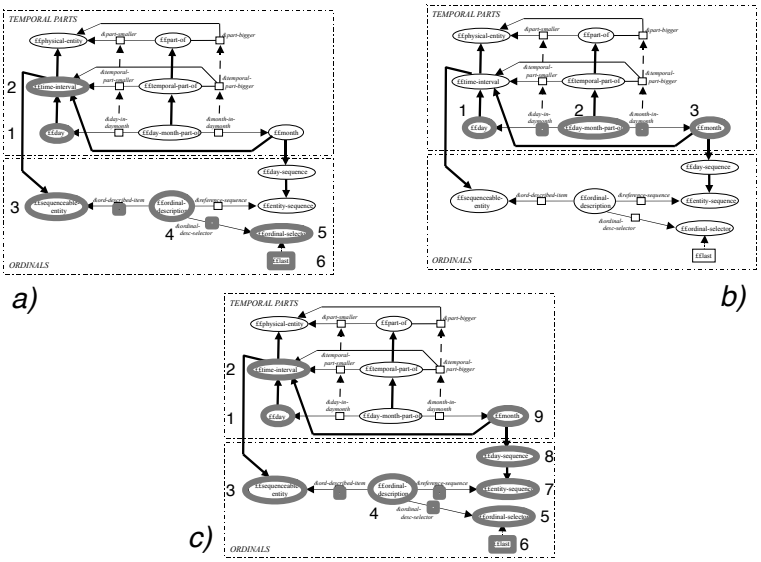

Figure 4: The ontology fragment computed by the semantic interpretation function.

source sentence is a complex ontology fragment obtained by the application of the function $\mathcal{M}_{\mathcal{O}}$. As case study we showed how this function uses the ontology $\mathcal{O}$ to interpret the ordinal numbers. The decision to use an ontology fragment as semantic representation is motivated by theoretical assumptions and has some practical appeals. From a theoretical point of view, we represent language semantics as part of the world knowledge in ontologies (Buitelaar et al., 2009; Galanis and Androutsopoulos, 2007; Nirenburg and Raskin, 2004). From an applicative point of view the ontology restriction produced by the semantic interpretation is used (in logical form) as input of the OpenCCG tool, in the generation component of the translation architecture (White, 2006). As a consequence, similar to Nirenburg and Raskin (2004), we use ontologies in all components of our architecture (cf. Galanis and Androutsopoulos (2007); Sun and Mellish (2007)).

We have currently implemented the main features of the $\mathcal{M}_{\mathcal{O}}$ and the ontology is being developed. Our working hypothesis is that the weather forecast sub-language is characterized by plain and short sentences and this guarantees scalability of our approach. In the next future we plan to broaden the coverage of linguistic phenomena, so to unify ordinals, superlative and comparative adjective analyses.[5]

References

Buitelaar, P., P. Cimiano, P. Haase, and M. Sintek (2009). Towards linguistically grounded ontologies. In *Proceedings of the 6th Annual European Semantic Web Conference (ESWC)*.

Bunt, H., R. M. M. Dzikovska, M. Swift, and J. Allen (2007). *Customizing Meaning: Building Domain-Specific Semantic Representations From A Generic Lexicon*, Volume 83. Springer.

Cimiano, P. (2009). Flexible semantic composition with DUDES. In *Proceedings of the 8th International Conference on Computational Semantics (IWCS'09)*.

Galanis, D. and I. Androutsopoulos (2007). Generating multilingual descriptions from linguistically annotated OWL ontologies: the naturalOWL system. In *In Proceedings of the 11th European Workshop on Natural Language Generation, Schloss Dagstuhl*.

Hudson, R. (1984). *Word Grammar*. Oxford and New York: Basil Blackwell.

Hutchins, W. and H. L. Somer (1992). *An Introduction to Machine Translation*. London: Academic Press.

Lesmo, L. (2007, June). The Rule-Based Parser of the NLP Group of the University of Torino. *Intelligenza Artificiale 2*(4), 46–47.

Nirenburg, S. and V. Raskin (2004). *Ontological Semantics*. The MIT Press.

Sun, X. and C. Mellish (2007). An experiment on "free generation" from single RDF triples. In *Proceedings of ENLG '07*, pp. 105–108. Association for Computational Linguistics.

White, M. (2006). Efficient realization of coordinate structures in combinatory categorial grammar. *Research on Language and Computation 2006*(4(1)), 39—75.

[5]**Acknowledgement:** This work is partly supported from the ATLAS project, that is co-funded by Regione Piemonte within the "Converging Technologies - CIPE 2007" framework (Research Sector: Cognitive Science and ICT).