Acquiring and representing semantic information in a Lexical Knowledge Base

Nicoletta Calzolari

Dipartimento di Linguistica - Università di Pisa
Istituto di Linguistica Computazionale del C N R - Pisa
Via della Faggiola 32
I 56100 Pisa - ITALY
e-mail: GLOTTOLO @ ICNUCEVM

Abstract

The paper focuses on the description of the approach, taken within the ESPRIT BRA project ACQUILEX, towards: i) acquisition of semantic information from several machine- readable dictionaries (in four languages), and ii) its representation in a common Lexical Knowledge Base. Knowledge extraction is guided by a) empirical observations and b) theoretical hypotheses. As for representation, we stress the convergence of a) and b) towards the possibility of organizing the information extracted from MRDs in the form of 'meaning types' or 'templates', where a common metalanguage is used to encode conceptual and relational information. Examples taken from two Italian monolingual dictionaries and from LDOCE are given. Different uses of these templates (e.g. as guides in the semantic analysis of the definitions, as a structure for comparing, unifying, merging, integrating information coming from different sources and different languages, as a tool for correcting 'incoherences' in dictionaries, etc.) are described.

Keywords: Computational Lexicography, Lexical Knowledge Base, Lexical Semantics.

1 Large computational lexicons and the notion of "reusability"

In order to cope with the task of building large computational lexicons where, to be able to process real texts, hundreds of thousands of words are necessary and, moreover, where also semantic information is made explicit for very large portions of the lexicon, the notion of "reusability" has become a central notion in the field of computational lexicography.

This concept came out at the Grosseto Workshop (1986) on "Automating the Lexicon", sponsored by the EC (see Walker, Zampolli, Calzolari, forthcoming), where, among the set of recommendations, there was that of designing "large reusable, multifunctional, precompetitive, multilingual linguistic resources".

Reusable must be interpreted in two main senses:

reusable_1: to exploit and reuse lexical information implicitly or explicitly present in preexisting lexical resources (MRDs, terminological DBs, textual corpora, etc.) as an aid to construct large computational lexicons of the type reusable_2;

reusable_2: to construct Computational Lexicons in such a way that various users (different NLP systems - in different theoretical frameworks and for different applications - , but also human users such as lexicographers, linguists, common users) can extract - with appropriate interfaces - relevant lexical information.

Current work on Computational Lexicons can be divided into two major types, each corresponding to the two meanings above: reusable_1 and reusable_2.

ACQUILEX, an ESPRIT BRA project, can be seen as the prototype of the first of these main streams of research, linked with the notion 'reusable-1', while other projects as e.g. Eurotra-7 insert themselves in the second sense of the 'reusability' concept.

2 MRDs as implicit Knowledge Bases

ACQUILEX (see Boguraev et al. 1988) focuses its research effort in developing techniques and methodologies for utilising and interpreting existing machine-readable dictionaries (MRD) to construct components for NLP systems. The main focus of the project is in the extraction of lexical — syntactic and semantic — information from multiple machine–readable dictionaries in a multilingual context with the overall goal of constructing a single multilingual lexical knowledge base (LKB). The dictionaries we are actually using in the project are: two monolingual English, two Italian, one Dutch, one Spanish, one bilingual Italian – English, one Dutch – English.

The information extracted is not only the information which is already explicit in MRDs (word-lists, part–of–speech, etc.), but mainly the information which in MRDs is only implicitly present and not directly and immediately accessible (mostly semantic information, such as semantic taxonomies, other semantic relations, argument structures, etc.). In the final LKB prototype it will be possible to "navigate" within the lexicon with access also through concepts and semantic relations.

In this approach it is considered possible a procedural exploitation of the full range of semantic information implicitly contained in MRDs. The dictionary is therefore considered in this framework as a primary source of "basic general knowledge", and main objectives are word–sense acquisition and knowledge organization. The main sources of this information are natural language definitions. The reasons of their use can be found in the following aspects:

i) the lexicographic tradition has exerted a (usually unconscious) control over the defining vocabulary (statement made really explicit only in LDOCE) and the schemata of defining formulas;

ii) the texts of definitions do not describe singular objects or events but "typical" ones;

iii) lexicographers have translated the concepts in their mind into definitions, and we can try to move back along this path from definitions to concept acquisition;

iv) the definitions incorporate a naif view of the semantic and world–knowledge information attached to lexical entries.

The goal of ACQUILEX is the formalization of this basic general knowledge (which can also be considered as a prerequisite to domain–specific knowledge) in the form of concepts and semantic relations. The method is heuristic and mainly inductive, through progressive generalization from the common elements.

The main themes of research connected to this goal of knowledge acquisition are the following:

- the design of procedures for the **extraction of superordinates** from natural language definitions, for their disambiguation, and for the construction of **taxonomies** all over the lexicon;
- the design of procedures for the (linguistic and computational) analysis of natural language definitions with the aim of **extracting all the implicit semantic information**;
- the study of ways of **formally representing the semantic information** which is extracted — concepts, attributes, and relations between concepts — e.g. in the form of 'typed feature structures';
- the study of how to **link and unify** taxonomies and conceptual or relational information coming **from different sources**, either monolingual or multilingual;
- the design and implementation of **basic software** for the creation, access and processing of **lexical databases and a lexical knowledge base**.

These research themes tackled within ACQUILEX are aimed at meeting one of the major bottlenecks of natural language processing, i.e. the availability of "large" computational lexicons with particular emphasis on making also semantic information explicit and accessible.

3 Semantic Information

3.1 Towards standardization at the taxonomic level

The procedural methodology for acquiring taxonomic information can be considered rather well established (see e.g. Byrd et al. 1987, Calzolari 1988). In this respect, the dictionary is considered as a "classificatory device", i.e. an empirical means of instantiating concepts. The MRD gives in fact one possible way of learning a concept (where the "learning" process assumes an inductive form): linking a concept to all its instances. All the instances of the same category/class are in fact extracted and connected together.

What is of interest with respect to taxonomies are not the leaf nodes, representing rather specific words, but middle– and top–level nodes in the IS–A hierarchy. These represent the core concepts by means of which the other words are defined via taxonomic relationship. An attempt to **normalization** is therefore being made in the project at the level of these core–nodes, in order both a) to give a more consistent structure to the hierarchy deriving from definitions, and b) to make possible the linking and merging of taxonomies extracted from different dictionaries and from different languages.

Analyses have already been performed which lead to the grouping of subsets of nodes under a same "conceptual label" representing the generalization over specific lexicalization of a similar lexical meaning. These conceptual labels are obtained through a comparative analysis of the different taxonomies for the different dictionaries, in order to create links and mappings between them, and constitute a first simple attempt of standardization at the semantic level. They should be analysed and compared with semantic primitives or features stated in other semantic systems.

3.2 More complex semantic and world–knowledge information: "types" and representation in terms of common feature structures

The aim of ACQUILEX in its second year is to extract, in addition to the simple IS–A links, more complex — and so far not really thoroughly analysed — semantic information hidden in the 'differentia' of the lexicographic definitions. In a project with several partners there is the necessity of working with similar (global) strategies of knowledge acquisition in order to reach the same result, i.e. a common LKB. We do not rely on a random extraction, but apply a knowledge acquisition strategy which, according to our views, must be guided by:

a) empirical observations,

b) theoretical hypotheses.

What is meant by a) is rather simple, being the so often observed systematic regularities and similarities of lexical items and definitional patterns.

By b) we intend the use of the theoretical approach to lexical semantics put forward by Pustejovsky (see Pustejovsky 1989, Pustejovsky and Boguraev forthcoming) in his 'qualia structures'. This approach makes use of a knowledge representation framework to express different aspects of knowledge structures concerning words. The qualia structure for a noun defines its essential attributes and "is in essence an argument structure for nouns".

In ACQUILEX we use a similar approach and similar types of structures, but in a broader sense than the qualia structures used by Pustejovsky which are made up of four main Roles (Constitutive, Formal, Telic, Agentive). These four main roles on the one side do not cover the whole range of lexical notions which characterize nominals, and on the other side do not include other pertinent world–knowledge information which can be useful in many NLP tasks or applications and can be found in MRDs.

We therefore take the underlying hypothesis of having "meaning types" and use the notion of "template" as main structuring device for semantic information, but enlarge this notion of template to include and represent: i) other semantic information not covered by the four main roles, and moreover ii) also more general or encyclopedic information concerning the concepts. An example of the template derived from the analysis of the definitions of three monolingual dictionaries (two Italian, one English) is given in Figure 1 for the concept of SUBSTANCE.

Dictionary definitions are suitable for an explicit representation in terms of feature structures as data types, which reproduce (at least partially) in an explicit way the original linear textual data (obviously not all of the definitions of a dictionary, and often not the entire definition).

This feature structure can be seen as a "meaning type", representing a maximal frame for a class of words (e.g. all the words defined by the word 'substance' or by its hyponyms). This frame, with all the potential attributes which in the definitions are found as most relevant for this subset, is inherited (as a "potential meaning type") by all the hyponyms of a "Top Node" (SUBSTANCE) and will be filled in some of its slots for each individual hyponym.

If we consider the "meaning structure" of LIQUID (see Figure 2) it is constituted by a subset of the attributes of SUBSTANCE, the same holds for GAS, and so on.

There will obviously be different "meaning types" for different categories of words. Attributes for verbs usually represent thematic roles which are relevant for a given "Action Type" and, where possible, also aspectual information is now being semi-automatically extracted from the definitions (see Alonge 1991). An example of the structures which result from dictionary definitions for the verbs of "hitting" and "dividing" is given in Figure 3. When an argument slot is filled for a hyponym of "to hit", this is an inherently specified argument (e.g. in "to hammer" the instrument role is lexically specified). However, as seen above, the view of assigning to nouns descriptions in terms of "frames" is also taken, with attributes or slots (and fillers), which are, at least partly, acquired from a procedural analysis of the definitions.

Different types of templates with different attributes are typical of "derivatives", which constitute a very large portion of a lexicon. They exhibit very special patterns and relationships with respect to their bases, with very interesting properties from a linguistic point of view. Their conceptual templates contain, among many others, attributes such as: AGENT, ACT_OF, PROPERTY_NAME, LOCATION, SET_OF, etc., but also attributes of a more encyclopedic nature such as: INHABITANT_OF, FOLLOWER_OF, etc.

We can associate to each of these relational patterns, which contribute to defining a very large amount of lexical items, conceptual templates (sets of properties) which are then inherited by default by all their defined words. As an example, we can associate to the AGENT attribute, among the others, the following set of attributes:

IS_A : human
TELIC : verb

where 'verb' is a variable which takes as value the base– verb for each derivative. E.g.: *lavoratore*: [AGENT: lavorare] by default also inherits: [IS_A: human, TELIC: *lavorare*].

It is with data of these types that we are beginning to feed the common LKB, a network consisting not only of IS_A relations, but of all the different types of semantic relations and semantic features implicitly present in the "differentia" part of all the definitions from all the available sources.

3.3 The Templates

The templates, or feature structures, in which we represent the semantic information, will serve many purposes in the whole process of knowledge acquisition, organization and representation.

They can be used in the following tasks:

a) As a *guide* both in the automatic or semi-automatic *parsing* of all the definitions of a same lexical field, from the top to leaf-nodes in the taxonomy, and in the *interpretation* of the results of the parsing process (e.g. to predict and constrain the interpretation of certain types of structures, or the proper attachment of PPs, etc.). In this syntactic/semantic parsing process the appropriate attribute/slots in the template are filled with the pertinent values.

b) As a common structure to be filled *independently* of the actual *lexical/surfacerealizations* of the semantic features, both in different dictionaries for the same language and in different languages. They act therefore as a scheme which makes *uniform* the interpretation process throughout all the different sources.

c) As a tool for *comparing* information coming from different sources, while making possible a semi- automatic and partial mapping of word-senses. When a link or a mapping is established, the data coming from different dictionaries are combined, according to the results of the comparison, either:

   i) by *merging*, when different surface lexicalizations are found for the same underlying concept (and in this case a score can be given to reinforce that information), or

   ii) by *integrating* different types of information on the same lexical item, e.g. filling different attributes.

d) As a tool for *correcting* errors or *incoherencies*, e.g. in the use of superordinates in the dictionary definitions.

These tasks can be summarized as follows:

- disambiguation
- knowledge acquisition
- knowledge uniformization
- knowledge representation
- knowledge comparison
  - merging
  - integration

These templates are inherited as potential "meaning types" by all the hyponyms, and the taxonomies are the vehicles by means of which this information is inherited. Obviously some of the values can be overridden by specific information.

4 Conclusions

With this common method of representing the information, the goal of sharing data and establishing correspondances among different sources is achieved. In this approach taxonomies and conceptual templates constitute in fact the point of convergence between different sources and languages, and between the empirical and the theoretical approaches.

The taxonomies and the templates — as developed within ACQUILEX — already constitute a first degree of normalization or standardization in the representation of semantic and world-knowledge information, both across many (about 10) dictionaries and (4) languages, and between the lexicographic approach to semantics and theoretical approaches. This is the first time that a project of semantic and world-knowledge information encoding for a very large part of the lexicon is carried out in such an extensive way.

5 References

Alonge A. (1991), "Extraction of information on Aktionsart from Verb Definitions", in Proceedings of Avignon Conference on NLP, Avignon.

Byrd R.J., N. Calzolari, M. Chodorow, J. Klavans, M. Neff and O. Rizk (1987), "Tools and Methods for Computational Lexicology", Computational Linguistics. 13(3-4), 219-240.

Boguraev B.K., E. Briscoe, N. Calzolari, A. Cater, W. Meijs, A. Zampolli (1988), "Acquisition of Lexical Knowledge for Natural Language Processing Systems", proposal for ESPRIT BRA, Cambridge (UK).

Boguraev B. K., R.J. Byrd, J. L. Klavans, M. S. Neff (1989), "From Structural Analysis of Lexical Resources to Semantics in a Lexical Knowledge Base", presented at the Workshop on Lexicon Acquisition, IJCAI, Detroit.

Calzolari N. (1988), "The dictionary and the thesaurus can be combined", in M. Evens (ed.), Relational Models of the Lexicon, Studies in Natural Language Processing, Cambridge University Press, Cambridge, 75-96.

Pustejovsky J. (1989), "Current Issues in Computational Lexical Semantics", Invited Lecture, Proceedings of the Fourth Conference of the European Chapter of the ACL, Manchester, England.

Pustejovsky J., B. Boguraev (forthcoming), "A richer Characterization of Dictionary Entries: The Role of Knowledge Representation", in S. Atkins, A. Zampolli (eds.), Computational Approaches to the Lexicon: Automating the Lexicon, OUP.

Walker D., A. Zampolli, N. Calzolari (eds.) (forthcoming), Automating the Lexicon: Research and Practice in a Multilingual Environment, OUP.

SUBSTANCE

FUNCTION
- USED_FOR (TELIC):
- USED_IN:
- USED_AS:
- USED_BY:

PROPERTY
- NATURE:
- STRUCTURE:
- ORIGIN:
- STATE:

- TASTE:
- SMELL:
- COLOUR:
- SHAPE:
- ASPECT:

- LACKING:

- SIMILAR_TO:

CONSTITUENCY
- CONSTITUTED_BY:
  - MAIN:
  - MUCH:

- CONSTITUENT_OF:
  - MAIN:

SOURCE
- DERIVED_FROM:
- PRODUCED_BY:
- PRODUCED_BY_MEANS_OF:

LOCATION:
CAUSE_OF:
TYPIC_EXAMPLE:
NAME:

Figure 1: Example of the template for SUBSTANCE Nouns.

L I Q U I D

FUNCTION
- USED_FOR (TELIC):
- USED_IN:
- USED_AS:

PROPERTY
- SMELL:
- COLOUR:

CONSTITUENCY
- CONSTITUTED_BY:
  - MAIN:
  - MUCH:
- CONSTITUENT_OF:
  - MAIN:

SOURCE
- DERIVED_FROM:
- PRODUCED_BY:
- PRODUCED_BY_MEANS_OF:

LOCATION:
CAUSE_OF:

Figure 2: Example of the template for LIQUID Nouns.

C O L P I R E (to hit)	**D I V I D E R E** (to divide)
WITH_INSTR:	OBJECT:
MANNER:	IN_PARTS:
OBJECT:	PART_NAME:
ITERATIVITY:	PART_NUMBER:
LOCATION:	WITH_S.ONE:
	WITH_INSTR:

Figure 3: Examples of templates for Verbs derived from dictionary definitions.

References

Alonge A. (1991), "Extraction of information on Aktionsart from Verb Definitions", in *Proceedings of Avignon Conference on NLP*, Avignon.

Byrd R.J., N. Calzolari, M. Chodorow, J. Klavans, M. Neff and O. Rizk (1987), "Tools and Methods for Computational Lexicology", *Computational Linguistics*. 13(3-4), 219-240.

Boguraev B.K., E. Briscoe, N. Calzolari, A. Cater, W. Meijs, A. Zampolli (1988), "Acquisition of Lexical Knowledge for Natural Language Processing Systems", proposal for ESPRIT BRA, Cambridge (UK).

Boguraev B. K., R.J. Byrd, J. L. Klavans, M. S. Neff (1989), "From Structural Analysis of Lexical Resources to Semantics in a Lexical Knowledge Base", presented at the Workshop on Lexicon Acquisition, IJCAI, Detroit.

Calzolari N. (1988), "The dictionary and the thesaurus can be combined", in M. Evens (ed.), *Relational Models of the Lexicon*, Studies in Natural Language Processing, Cambridge University Press, Cambridge, 75-96.

Pustejovsky J. (1989), "Current Issues in Computational Lexical Semantics", Invited Lecture, *Proceedings of the Fourth Conference of the European Chapter of the ACL*, Manchester, England.

Pustejovsky J., B. Boguraev (forthcoming), "A richer Characterization of Dictionary Entries: The Role of Knowledge Representation", in S. Atkins, A. Zampolli (eds.), *Computational Approaches to the Lexicon: Automating the Lexicon*, OUP.

Walker D., A. Zampolli, N. Calzolari (eds.) (forthcoming), *Automating the Lexicon: Research and Practice in a Multilingual Environment*, OUP.