A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet

Bernardo Magnini, Carlo Strapparava
Fabio Ciravegna and Emanuele Pianta

IRST, Istituto per la Ricerca Scientifica e Tecnologica
I - 38050 Povo TN, Italy
e-mail: {magnini | strappa | cirave | pianta}@irst.it

Abstract

This paper describes a project that aims to develop a large-size Lexical Knowledge Base (LKB) for the Italian language. The project began because presently there is no LKB available for Italian, on an electronic support, that can be used both for natural language processing and lexicography applications. To develop the LKB techniques developed in artificial intelligence that allow acquisition and automatic data structuring will be used.

1. Motivations

The idea of a lexical knowledge base was recently proposed by the ESPRIT BRA AQUILEX [Calzolari 1991] project, to provide information, mostly of a semantic nature, consistently structured internally and available electronically. Three levels of lexical representation are proposed in AQUILEX: (a) Machine Readable Dictionary (MRD), in which one finds an electronic version of the paper dictionary; (b) Lexical Data Base (LDB), in which part of the information available in the text format dictionary has been extracted; (c) Lexical Knowledge Base (LKB), in which lexical information is structured and represented consistently. The difficulties encountered in LKB development are mostly caused by lack of consistency (e.g. circularity of the taxonomies used) and the incompleteness of the semantic information included in the machine dictionary; these facts led to the necessity of large hand integration of knowledge extracted.

In this project we propose a methodology for semiautomatically building an LKB for a language, in which knowledge described in the dictionary is integrated with the knowledge, already structured, described in WordNet [Miller 1990] a LKB developed for the English language.

The present work is part of a much larger program to develop natural language dialog systems that the natural language group at IRST has been working on for some years [Stock et al. 1993]. A need is recently emerged to increase system robustness by providing an interface with great linguistic coverage. It is in this context that the availability of large amounts of syntactic/semantic information became necessary, through which complex problems of text interpretation can be confronted.

2. WordNet

WordNet is a lexical knowledge base for English, available at no charge, electronically. Originally the project was inspired by the current psycholinguistic theory of human lexical memory. Nouns, verbs, adjectives and adverbs are organized in sets of synonyms, each of which represents a lexical concept. These sets of synonyms are interconnected by a certain number of relations and organized into taxonomies. The current version WordNet incluses about 100,000 lexical items organized into 80,000 meanings (or synsets). The correspondence among lexical forms and meanings is maintained through a bidimensional matrix in which each synset is understood to be an unambiguous designator of the meaning of the word. Often (about 70% of the time) a brief definition (gloss) is also associated to a synset. WordNet distinguishes two types of relations: *lexical relations*, such as synonomy, antonomy and polisemy, and *semantic relations*, such as hyponomy and meronomy.

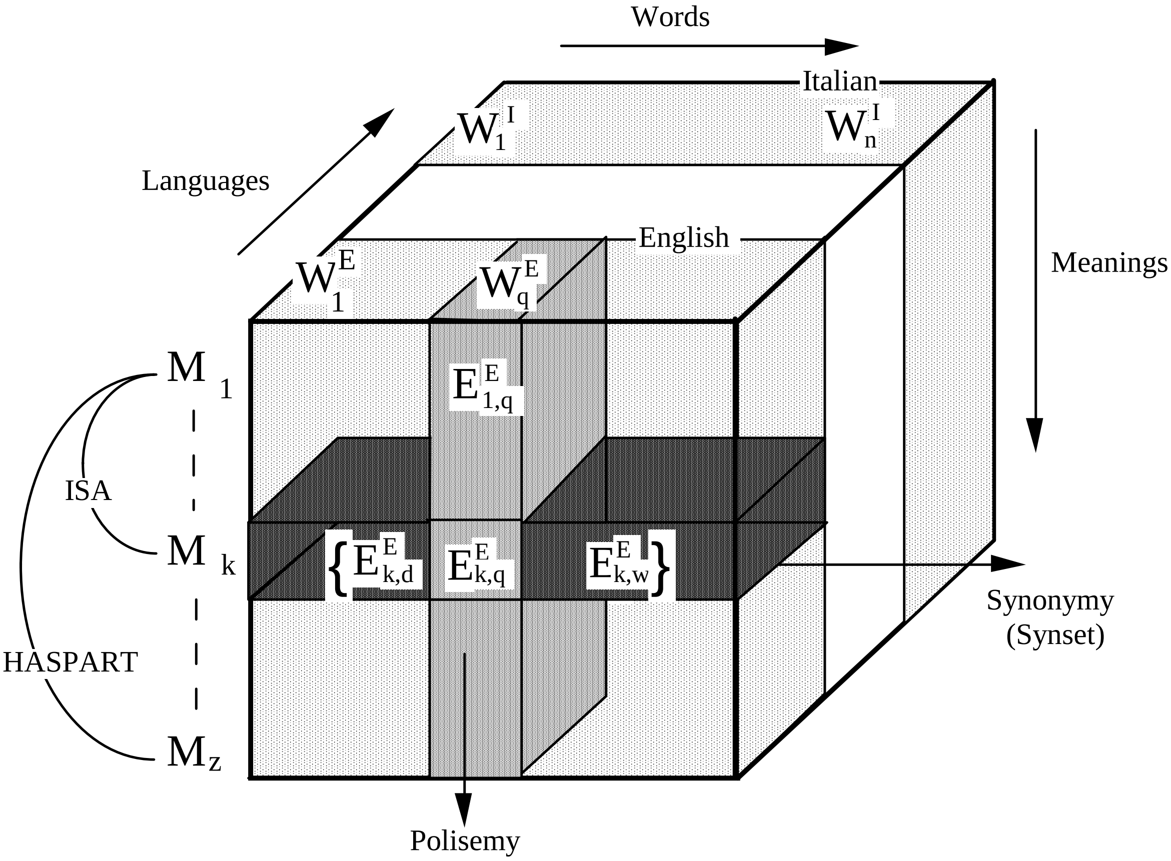

Figure 1: Multilingual lexical matrix

The most important lexical relation for WordNet is the similarity of meaning, since the ability to recognize synonomy among words is a prerequisite to build synsets and therefore meaning representation in the lexical matrix (see Figure 1).

Two expressions are synonomous if substitutivity is valid (in other words if the substitution of one with the other does not change the truth value of a phrase). Actually a weaker definition is more useful, relative to a context. Two expressions are synonomous in linguistic context C if the substitution of one with the other in C does not change the truth value. It is important to note that defining synonomy in terms of substitutivity requires partioning WordNet into nouns, verbs, adjectives and adverbs. Obviously if a word pertains to more than one synset, this gives an indication of its polysemy.

The antonym relation instead gives the main organizational characteristic for adjectives and adverbs.

WordNet is not limited to creating a set of meanings relative to lexical forms, but also indicates the semantic relations that exist among them. The hypo/hyperonomy (or ISA relation) puts subordinates and superordinates in meaning relation thereby providing a hierarchic concept structure. The meronimic relation (HAS-PART) defines, instead, a part hierarchy over the set of meanings.

3. The multilingual lexical matrix

The starting point for building a WordNet multilingual network is based on the assumption that the meaning networks already defined for the original English version may, for the most part, be reused for other languages. This may be considered plausible if we limit ourselves to the main indoeuropean languages, among which there is much cultural overlap [Miller -- personal communication].

The project foresees a multi-lingual lexical matrix (MLLM) as an extension of the bidimensional lexical matrix implemented in WordNet. A third dimension will be added to the matrix, through which it will be possible to consider different languages. The extension of the languages dimension initially will be considered for Italian. Figure 1 shows the three dimensions of the matrix: (a) words in a language, indicated by W_j; (b) meanings, indicated by M_i; (c) languages, indicated by L_k. Moreover, the main lexical and semantic relations are visualized. From an abstract point of view, to develop the multilingual matrix it is necessary to re-map the Italian lexical forms with corresponding meanings (M_i), building the set of synsets for Italian (making explicit the values for the intersections E_{ij}^I). The result will be a complete redefinition of the lexical relations, while for the semantic relations, those originally defined for English will be used as much as possible. From this point of view the dimension of meanings is considered constant in relation to the languages and words of each language. If for a certain M_k for language L one obtains $E_{ik}^L=0$, with i = 0...t, where t is the dimension of the lexicon of language L, this means that for language L there is no word that lexically realizes that meaning.

4. Automatically building the LKB

The main task to build an LKB based on WordNet is to find the correct corespondences between Italian words and synsets defined for English. For automatically building the LKB there are two main problems: a) the ability to extract data relative to Italian words from the available sources; b) the correspondences between Italian and English and viceversa. Figure 2 shows the relations between these two dimensions. As for the sources, the lexical knowledge for Italian will be acquired from a dictionary of the Italian language in electronic format. As for translations, a bilingual dictionary Italian/English and English/Italian is used.

At least three levels of depth of the data are distinguished. The first level operates only on WordNet synsets seeking to build the italian equivalent; at the next level definitions from the italian dictionary are compared with the WordNet glosses using statistical methods. The third level compares the definitions again but it uses natural language processing techiniques. These three levels of analysis, at the present status of the informatiion extraxction techniques, should not be considered mutually exclusive alternatives, but as methods to achieve complementary results.

The method uses a dual approach: a) starting from italian definitions, in which case italian words without a corresponding english meaning are expected and new meanings must be added to the Wordnet taxonomy; b) starting from english definitions, in which case words that do not have a corresponding italian meaning are expected to be found.

Given the relative simplicity of the english definitions contained in WordNet, quantitatively better results are expected when starting from english definitions.

After automatically building the LKB, the validation of the proposed choices by a lexographer is still a useful step.

In the following paragraphs, for each level mentioned above, a brief description of the procedures that may be used to obtain data will be given as well as algorithms that may be applied to compare homogeneous information.

4.1. Synonyms intersection

At this level the intersection of synonyms of Italian and English words is considered, using techniques derived from computer science. The idea is to start from the following information sources:

1. the English synsets and their relation with the taxonomy;
2. a machine readable bilingual Italian-English dictionary
3. a machine readable Italian dictionary and one of synonyms.

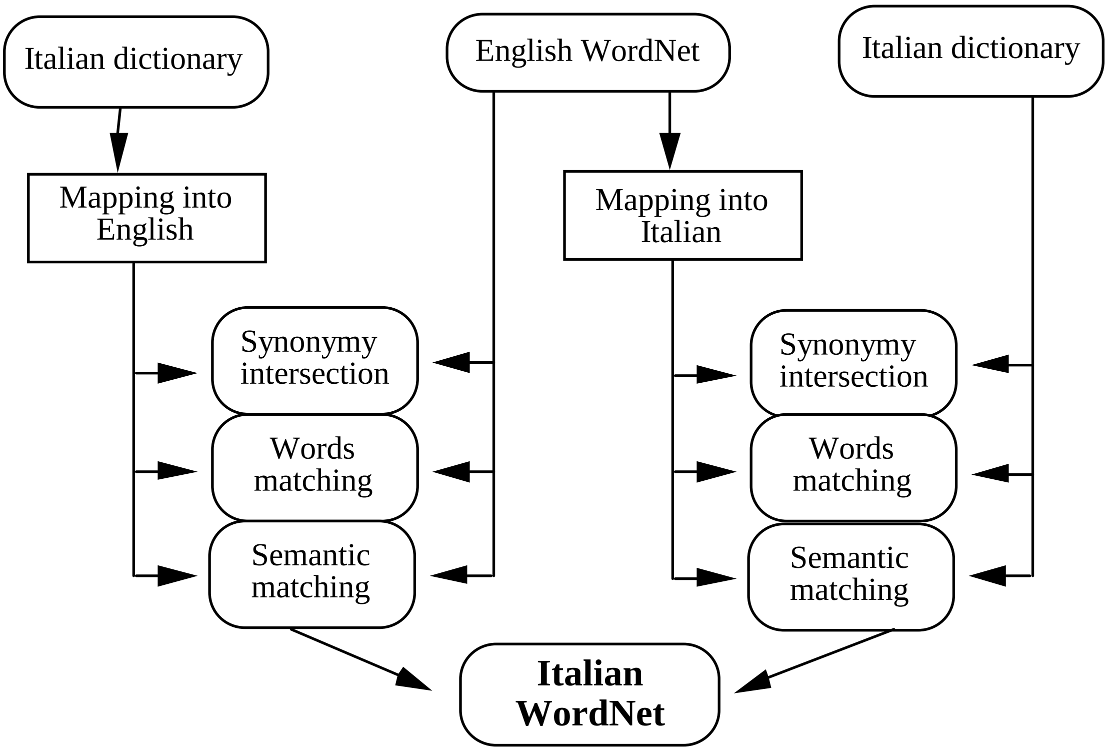

Figure 2. Extraction procedures and data comparison.

The algorithm is designed to obtain a synset formed of Italian words with meanings comparable to those in English through a series of successive passes, as in the following example:

SYNSET WORDNET = {registration enrollment}

Italian translations taken from the two words of the synset:

registration : 1. registrazione, iscrizione, immatricolazione;
enrollment: : 1. elencazione, registrazione, arruolamento, iscrizione;
2. numero degli arruolati;

With a simple intersection of the sets of Italian words provided for each English form we can derive that registration[1] and enrollment[1] have the same meaning; from here we find that the:

SYNSET ITALIANO = {iscrizione, registrazione}

corresponds to that in English {registration, enrollment}, and that therefore it refers to the same level of WordNet taxonomy. In more complex cases it may be necessary to use the dictionary of Italian synonyms to correctly define the mapping between different meanings of a single word. In the same way the dictionary of synonyms may be used to enlarge the Italian synset.

4.2 Word matching

Comparing the WordNet glosses and the definitions of the Italian dictionary is another method of matching. A similar approach was already successfully tried to match the same WordNet glosses and the (English) definitions in the LDOCE dictionary. The algorithm of similarity between the definitions is based on a statistic rate of the presence of common words in the two definitions [Knight 93]. In our case, though, there is a further complication: definitions are written in two different languages. However, by using correctly the data contained in WordNet (synset of words, concepts, etc.), as well as some of the techniques mentioned in the previous point, this problem also may be resolved. Note that Knight himself is working on an analogous system for definitions derived from a Spanish dictionary.

4.3 Semantics matching

With this method it is necessary to extract semantic information both from the definitions of the Italian dictionary and the WordNet glosses. The task for the glosses is simplified because they use a limited vocabulary (about 7,000 roots, excluding proper nouns) and simplified syntactic constructions.

A first level of semantic extraction regards the "genus" of the definition. Additional information (the “differentia”), will also be considered using extraction techniques that consist of filling slots with known frames. Another analysis level consists of making a superficial syntactic analysis of the definitions producing simplified semantic representations (for example, without quantifiers). For this type of analysis plans are to use only data on the syntactic category of the words and the type of subcategorization of the verbs available in the paper dictionaries.

In a typical situation, given an Italian frame, the first step will be to select a set of English frames potentially corresponding (using data on the genus or correspondences indicated by the bilingual dictionary - see preceeding techniques). Then the English frame (which corresponds to a synset) must be chosen most similar to the Italian frame. The algorithm must therefore compare and establish a degree of similarity between the two frames, containing data derived, respectively from Italian and English. The matching algorithm will use a bilingual dictionary, through which it must verify the correspondence between two terms in two different languages. It is not therefore a translation from one language to another, but more simply a verification that, given an English word and an Italian one, these are compared by a bilingual dictionary. The comparison may be of various types (in both senses, in one sense only, through connected terms, etc.) and the algorithm must be able to evaluate it.

References

[Calzolari 1991] Calzolari N.: “Acquiring and representing semantic information in a Lexical Knowledge Base”, in Pustejovsky J. and Bergler S. (eds.) *Proceedings of the Lexical Semantics and Knowledge Representation Workshop,* Berkeley, CA, June 1991.

[Knight 93] Knight K.: "Building a Large Ontology for Machine Translation", proceedings of the *DARPA Human Language Conference*, March 1993.

[Miller 1990] Miller, G. A. (ed.), WordNet: “An on-line lexical database”. *International Journal of Lexicography* (special issue), 3 (4), 235-312, 1990.

[Stock et.al. 1993] Stock, O. and the AlFresco Project Team, “AlFresco: enjoing the combination of NLP and hypermedia for information exploration”, in M.T. Maybury (ed.), *Intelligent Multimedia Interfaces*, AAAI and MIT Press, 1993.