Multilingual Lexical Knowledge Bases: Applied WordNet Prospects

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

1. Introduction

The idea of a Lexical knowledge base was recently proposed by the ESPRIT BRA AQUILEX [Briscoe 91], [Calzolari 92] project, to provide information, mostly of a semantic nature, internally consistently structured and electronically available. Three levels of lexical representation are proposed in AQUILEX: (a) Machine Readable Dictionary (MRD), i.e. an electronic version of the paper dictionary; (b) Lexical Data Base (LDB), where part of the information available in the text format dictionary has been extracted; (c) Lexical Knowledge Base (LKB), where lexical information is consistently structured and represented.

In this paper we suggest an evolution from the current monolingual lexical knowledge bases (LKB) to multilingual LKBs (MLKBs). MLKBs could be profitably used to assist Multilingual Natural Language Processing applications when both lexical and conceptual relations among words are required. There can be a number of such situations, including multilingual generation and extraction of relevant information from texts, or -on a lower level of processing - disambiguation of word senses, check of selectional restrictions, maintenance of the lexical cohesion in discourse processing.

The ideas presented are parts of an IRST project to design and implement a methodology for semi-automatically building an MLKB starting from a monolingual (Italian) dictionary, a bilingual (Italian-English) dictionary and a monolingual (English) LKB: WordNet [Miller 90]. The MLKB proposed is based on a substantial share of the semantic information and with the possibility of an automatic bootstrapping from available linguistic resources [Magnini *et al.* 1994]. Two points are in particular stressed in this paper: (a) the (semi-)automatic construction of that MLKB; (b) its usefulness in application environments.

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 organised in sets of synonyms, each of which represents a lexical concept. These sets of synonyms are interconnected by a certain number of relations and organised into taxonomies. The current version of WordNet includes about 100,000 lexical items organised into 80,000 meanings (or synsets). The correspondence among lexical forms and meanings is maintained through a bi-dimensional

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 synonymy, antinomy and polisemy, and *semantic relations*, such as hyponimy and meronomy.

3. A multilingual LKB

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].
As far as the methodology is concerned we are currently considering an Italian Dictionary as the main information source. However, dictionary knowledge could result insufficient to obtain a complete Italian LKB. This aspect has been noted by [Atkins, Levin 91] who stressed the fact that dictionaries are strongly influenced by marketing criteria, therefore having drawbacks in their linguistic organization. An improvement in the current methodology for bootstrapping an Italian WordNet could be achieved considering also the kind of lexical knowledge implicitly included in large corpora.

3.1. The multilingual lexical matrix

The project foresees a multi-lingual lexical matrix (MLLM) as an extension of the bi-dimensional 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 visualised. 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.

3.2. Automatically building the multilingual LKB

The main task to build an MLKB based on WordNet is to find the correct correspondences 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 non-English words from the available sources;
b. the correspondences between other languages and English and vice versa.

Figure 2 shows the relations between these two dimensions for a bilingual (Italian-English) LKB. 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 detected [Magnini *et al.* 94]: word comparison, glosses/dictionary definition comparison via both statistical methods (second level) and NLP techniques (third level). These three levels of analysis should not be considered mutually exclusive alternatives, but as methods to achieve complementary results.

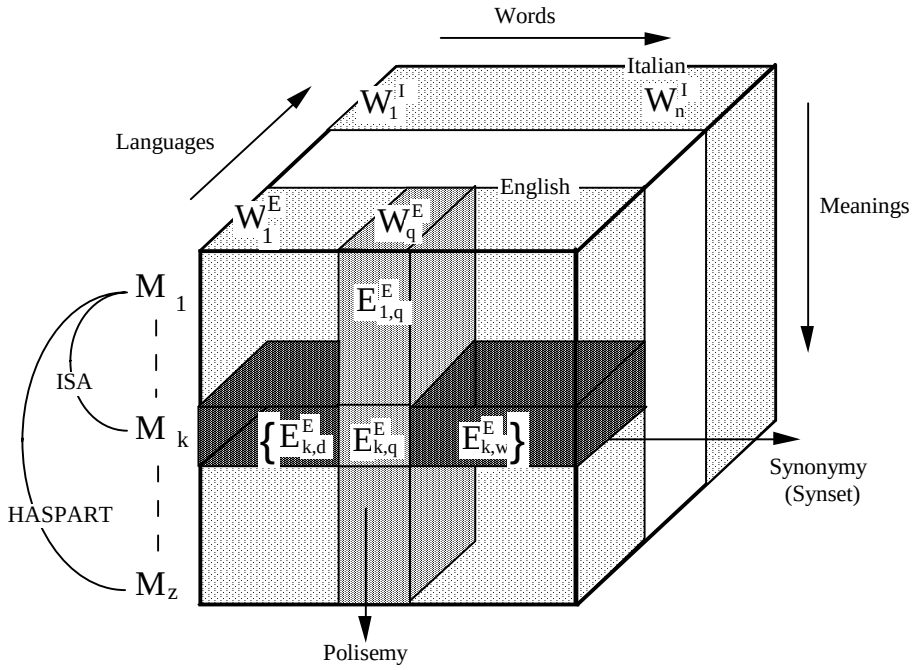

Figure 1: Multilingual lexical matrix

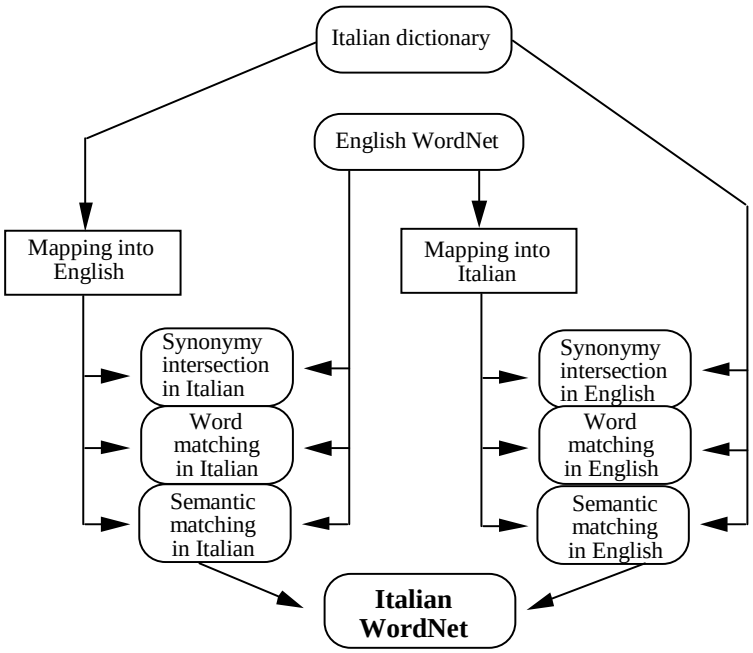

Figure 2. Extraction procedures and data comparison.

The method is based on a dual approach that starts both from Italian and English definitions; in both cases words without a corresponding meaning in the other language are expected to be found. In particular many Italian words will not have any synset to be referred to, and new nodes are expected to be inserted in WordNet taxonomy.
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 lexicographer is still a useful step.

4. Applied (M)LKBs: high level applications

Generally speaking LKBs are particularly suitable for NLP applications involving the analysis of a large amount of linguistic data on unrestricted domains that do not require complex inferential capabilities. We will investigate some relevant possibilities in the area of textual application

- **Text analysis**

Interest in Information Extraction from Texts has grown recently. The field has started to interest the market and a number of application have arisen. Unfortunately the state of the art is far from being able to cope with (even partially) unrestricted texts, as they need extended lexica and knowledge bases. Although extended lexica are available on the market (at least for some languages), knowledge bases are to be built by hands. WordNet provides a knowledge of world and language that is far beyond any available hand coded knowledge bases. It can be used both alone or in conjunction with a real KB, to cope with those parts of the input that goes beyond the KB coverage. That knowledge base could even be superimposed on WordNet as its specialisation in a particular domain.

The "mixed solution" seems to be the most interesting for the short term development of the field, as it can provide a good solution to cope with texts concerning partially unrestricted domain, such as those used in the MUC tests [Sundheim 92]. For example description of terrorist acts can contain descriptions of every day objects (such as suitcases, bike, dust bins, etc.), or company take-over news can contain description of company products (weather seals, iron and metal golf clubs, detergents, etc.), i.e. quite any physical and/or intellectual objects. WordNet provides such knowledge. This solution is more effective than backup lexicons as were used for example by [Mellish 92], because WordNet gives semantic information about the words it knows, whereas backup lexicons are generally very poor on this point of view.

- **(Multilingual) text generation**

In this case the MLKB is used as a kind of interlingua which allows lexical variations on the expressibility of the same content. Actually are feasible only projects operating on limited domains: this limitation is mainly due to the fact that it is both difficult and time consuming to build large knowledge bases for natural language purposes. As an example, the LRE-Gist project, whose aim is the generation of instructional texts in three languages (English, German and Italian), is currently restricted to a limited number of text and could take advantage from a MLKB of the kind we have been proposing.

- **(Multilingual) electronic publishing**

There could be many useful applications in the emerging area of electronic publishing. Very time consuming activities such as the consistency check of information included within dictionaries, the comparison and the retrieval of data from different sources, and also the realization of environments for using electronic texts, could be facilitated by means of (M)LKBs.

5. Applied (M)LKBs: low level tasks

In the next subsections we will highlight some of the advantages of using WordNet to analyse language; in particular we will focus on three technical aspects such as word sense disambiguation, selectional restriction check and lexical cohesion.

- **Word sense disambiguation**

The problem of word sense disambiguation is that of choosing the best sense for a word *w*, given the linguistic context it appears (e.g. the word senses of the sentence in which *w* is included). In recent years there has been much work aimed at the definition of statistical methodologies in which different information levels has been considered. Word collocations [Church 90] [Guthrie *et al*. 91], syntactic patterns [Smadjia 91] and also semantic relations [Zernik, Jacobs 90] [Grishman 92] have been used in statistical disambiguation. However, while the use of semantic relations is considered a crucial issue for improving disambiguation performances, less work has been done to identify a reusable set of semantic relations as a stable base for statistical experiments. The required set of semantic tags should have at least the following properties:

a. for each word in the corpus a mapping to one o more semantic tags should be provided;
b. the number of semantic tags should be quite limited, in order to allow useful generalisations, and so improving statistical results.

WordNet has much of these features. In particular, point (a) is directly provided by the WordNet implementation, being the mapping between a certain word and all the synsets the word is included in. On the other side, point (b) can be easily achieved considering the high level semantic classes of the hierarchy. In this respect abstraction algorithms have been recently proposed [Resnik 92] [Voorhees 93] which conduct a best-first search upwards in the WordNet hierarchy. In addition, if an equivalent of the WordNet taxonomy were available for languages other than English, it would be possible to compare similar disambiguation techniques over data from different languages.

- **Selectional restriction check**

WordNet can help parsers that use selectional restrictions in disambiguating ambiguous constructions. Using selectional restrictions during parsing requires three things:

1. Selectional restrictions must be specified for all the arguments that predicates select for.
2. Selectional features must be assigned to potential fillers of predicates (mostly nouns). In semantics-minded approaches, selectional restrictions are semantic classes. In this case we specify the most general class that is acceptable as argument of a certain predicate. To verify that a certain argument filler matches the expected selectional restrictions, we need component number 3.
3. A ISA hierarchy relating semantic classes with an inferential mechanism able to calculate the transitive closure of the ISA relation.

In the semantic approach to selectional restrictions WordNet can be very useful. If fact it provides components 2 and 3. In a way it gives also information about 1, although in a incomplete way.

- **Lexical cohesion.**

As stated by [Halliday *et al*. 76] the property of cohesion is a way of getting the text to hang together as a whole. There are various types of cohesion: back-references, ellipsis, conjunction and in particular for our purposes lexical cohesion. Lexical cohesion is the cohesion that arises from semantic relationship between words. In any NLP system the use of a LKB as WordNet could be fundamental to give 'a measure' of lexical cohesion and to provide for the determination of coherence and discourse structure. In such a scenario WordNet could be used along with a traditional KB (frames or semantic networks).
The advantage of using WordNet are mainly its dimensions and the relations provided, for example entailment and coordinated terms. Here follow two examples on how it is possible to maintain cohesion using WordNet:

- Here is a book by Dick
- OK. Read the first chapter to me .

Book -HAS-PART- chapter. The chapter mentioned is the first chapter of Dick's book. There is lexical cohesion. Moreover 'the first chapter' is an implicit anaphoric reference to 'Dick's book'.

Conclusions

In this paper we have suggested the evolution of LKBs towards MLKBs. We have shown the impact of MLKBs on NLP both from a technical and applicative point of view. We have shown some of the details of an IRST internal project that aims at the semi-automatic construction of a bilingual LKB. The presented project 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*. 93]. 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.

References

[Atkins, Levin 91] Atkins Beryl T. and Levin Beth: "Admitting Impediments", in Zernik Uri (ed.) *Lexical Acquisition - Exploiting On-line Resources to Build a Lexicon*, Lawrence Erlbaum Associated, Hillsdale, NJ, 1991.

[Briscoe 1991] Briscoe Ted, "Lexical Issues in Natural Language Processing", in Klein E. and Veltman F. (eds.): *Esprit Symposium on Natural Language and Speech*, Berlin, Springer-Verlag, 1991.

[Calzolari 92] Calzolari N.: "Acquiring and Representing Semantic Information in a Lexical Knowledge Base", in Pustejovsky J. and Bergler S. (eds.) *Lexical Semantics and Knowledge Representation* , Springer-Verlag, Berkeley, 1992.

[Church 90] Church Kenneth W.: Word Association Norms, Mutual Information, and Lexicography, Computational Linguistic, Vol. 16, No. 1, 1990.

[Grishman *et al*. 92] Grishman R., Sterling J.: "Acquisition of Selectional Pattern", Proc. COLING-92.

[Guthrie *et al*. 91] Guthrie Joe, Guthrie Louise, Wilks Yorick, Aidinejad Homa: "Subject dependent co-occurrence and word sense disambiguation", Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991.

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

[Magnini *et al.* 94] Bernardo Magnini, Carlo Strapparava, Fabio Ciravegna, Emanuele Pianta: "A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet", IRST Technical Report # 9406-15, June 1994.

[Miller 90] Miller, G. A. (ed.), WordNet: "An on-line lexical database". *International Journal of Lexicography* (special issue), 3 (4), 235-312, 1990.

[Stock *et al*.. 93] Stock, O. and the AlFresco Project Team, "AlFresco: enjoying the combination of NLP and hypermedia for information exploration", in M.T. Maybury (ed.), *Intelligent Multimedia Interfaces*, AAAI and MIT Press, 1993.

[Halliday, Hasan 76]: Halliday M. A. K., Hasan R.: *Cohesion in English*, English Language Series, Longman, New York , 1976.

[Resnik 92] Resnik Philip: "WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery", Proc. AAAI-92.

[Smadjia 91] Smadjia Frank A.: "From N-Grams to Collocations: an Evaluation of Xtract", Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991.

[Sundheim 92] Sundheim, B. (ed.) Proc. of the 4th Message Understanding Conference (MUC-4), McLean, Virginia, June 1992, distr. Morgan Kaufmann Publishers Inc., San Mateo California.

[Mellish *et al*. 92] Mellish C., Allport D., Hartley A.F., Evans R., Cahill L.J., Gaizauskas R., Walker J.: "The TIC Message Analyser", draft paper.

[Voorhees 93] Voorhees Ellen M.: "Using WordNet to Disambiguate Word Senses for Text Retrieval", Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993.

[Zernik, Jacobs 90] Zernik U, Jacobs P.: "Tagging from Learning: Collecting Thematic Relations from Corpus", Proc. COLING-90.