Exploring Interoperability of Language Resources: the Case of Cross-lingual Semi-automatic Enrichment of Wordnets

Francesca Bertagna, Nicoletta Calzolari, Monica Monachini and Claudia Soria
Istituto di Linguistica Computazionale del CNR, Via Moruzzi 1, 56024 Pisa, Italy

Shu-Kai Hsieh and Chu-Ren Huang
Academia Sinica, Nankang, Taipei, Taiwan

Andrea Marchetti and Maurizio Tesconi
Istituto di Informatica e Telematica del CNR, Via Moruzzi 1, 56024 Pisa, Italy

Abstract. In this paper we present an application fostering the integration and interoperability of computational lexicons, focusing on the particular case of mutual linking and cross-lingual enrichment of two wordnets, the ItalWordNet and Sinica BOW lexicons. This is intended as a case-study investigating the needs and requirements of semi-automatic integration and interoperability of lexical resources.

Keywords: distributed language resources, interoperable lexical resources, language services

1. Introduction

Enhancing the development of language resources, and in particular of lexical resources, is of foremost importance for many applications to take off. Nevertheless, large–scale multilingual lexical resources are not as widely available and are very costly to construct: the work process for manual development of new lexical resources or for tailoring existing ones is too expensive in terms of effort and time to be practically attractive. The previous trend in lexical resource development was oriented to maximization of effort by building large–scale, general–purpose lexicons. However, these lexical resources are not always satisfactory despite the tremendous amount of work needed to build them and the richness and degree of sophistication of the information contained therein. Often lexical resources are unbalanced with respect of the type of lexical information encoded, focusing on a particular type and not providing enough coverage of other aspects. In some other cases, lexical resources are too much or too little detailed for the specific purposes of an application. On the other hand, the market is increasingly calling for new types of lexical resources: lexicons that can be built rapidly, possibly by combining certain types of information while discarding other,

and tailored to specific needs and requirements. Rather than building new lexical resources, the new trend focuses on trying to exploit the richness of existing lexicons.

To meet these requirements, lexical resources need to be made available to and be constantly accessed by different types of users, who may want to select different portions of the same resource or may need to combine information coming from different resources. This scenario no longer leaves space to static, closed, and locally managed repositories of lexical information; instead, it calls for an environment where lexical resources can be shared, are reusable, and are openly customizable. At the same time, as the history of the web teaches, it would be a mistake to create a central repository containing all the shared lexical resources, if only because of the difficulties to manage it. The key has been identified in the concept of *distribution of lexical resources*, and actually the solution being constantly proposed by the lexical resource community consists in moving towards distributed *language services*, based on open content interoperability standards, and made accessible to users via web–services technology.

The paradigm of distributed and interoperable lexical resources has largely been discussed and invoked. Some initial steps are made to design frameworks enabling inter-lexica access, search, integration and operability. An example is the Lexus tool (Snijders et al., 2006), that goes in the direction of managing the exchange of data among large-scale lexical resources. A similar tool, but more tailored to the collaborative creation of lexicons for endangered language, is SHAWEL (Gulrajani and Harrison, 2002). However, the general impression is that little has been made towards the development of new methods and techniques for attaining a concrete interoperability among lexical resources.

In this paper we present a tool, based on a web–service architecture, fostering integration and interoperability of computational lexicons, focusing on the particular case of mutual linking and cross-lingual enrichment of distributed monolingual lexical resources. As a case–study, we have chosen to work with two lexicons belonging to the WordNet family, the ItalWordNet and Sinica BOW. The paper is organized as follows: section 2 describes the general architectural design of our project; section 3 describes the tool taking care of cross-lingual integration of lexical resources, while a case-study involving an Italian and Chinese lexicons is presented in Section 4.

2. An Architecture for Integrating Lexical Resources

Making the vision of shared and distributed lexical repositories a reality is a long-term scenario requiring the contribution of many different actors and initiatives (among which we only mention standardisation, distribution and international cooperation). In our work we adopted a bottom–up approach to exploring interoperability of lexical resources by developing an application dedicated to the cross–lingual enrichment of monolingual lexicons. This is intended as a case-study and a test-bed for trying out needs and requirements posed by the challenge of semi-automatic integration and enrichment of practical, large-scale multilingual lexicons for use in computer applications. We designed a distributed architecture to enable a rapid prototyping of cooperative applications for integrating distributed lexical resources. This architecture is articulated in three layers:

1. The lower layer consists of a grid of local wordnets realized as a virtual repository of XML databases residing at different locations and accessible through web services. Basic services are also necessary, such as an UDDI server for the registration of the local wordnets and other services devoted to the coherent management of the different versions of Princeton WordNet (i.e. WN1.5, WN1.6, etc.) to which the various databases are linked.

2. The middle layer hosts diverse applications that exploit the wordnets grid. The so-called *Multilingual WordNet Service* (MWS, Section 3) was built as a proof of concept of the possibility to mutually enrich wordnets in a distributed environment; other, more advanced NLP applications (in particular multilingual) can be developed by exploiting the availability of the WordNet grid.

3. A higher layer, called "cooperative layer" or *LeXFlow* is intended as an overall environment where all the modules realized in the lower layers are integrated in a comprehensive workflow of human and software agents.

The figure below illustrates the general architecture. In this paper we concentrate on the description of the middle layer (see Section 3). A more detailed description of the cooperative layer can be found in (Soria et al., 2006) and (Tesconi et al., 2006).

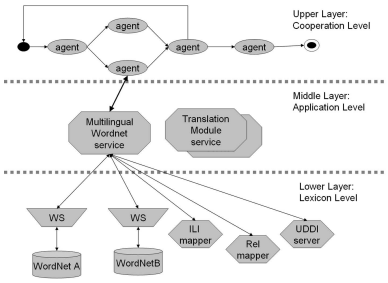

Figure 1. A three–layered architecture for integrating lexical resources

3. Multilingual WordNet Service

In this section we present a tool that addresses the issue of lexicon augmentation or enrichment focusing on mutual enrichment of two wordnets.

This tool, named "Multilingual WordNet Service" is responsible for the *automatic cross-lingual fertilization* of lexicons having a WordNet-like structure. Put it very simply, the idea behind this module is that a monolingual wordnet can be enriched by accessing the semantic information encoded in corresponding entries of other monolingual wordnets.

Since each entry in the monolingual lexicons is linked to the Interlingual Index (ILI, cf. Section 3.1), a synset of a WN(A) is indirectly linked to another synset in another WN(B). On the basis of this correspondence, a synset(A) can be enriched by importing the relations that the corresponding synset(B) holds with other synsets(B), and vice-versa. Moreover, the enrichment of WN(A) will not only import the relations found in WN(B), but it will also propose target synsets in the language(A) on the basis of those found in language(B).

The various WN lexicons reside over distributed servers and can be queried through web service interfaces.

3.1. Linking Lexicons through the ILI

The entire mechanism of the Multilingual WN Service is based on the exploitation of Interlingual Index (Peters et al., 1998), an unstructured version of WordNet used in EuroWordNet (Vossen, 1998) to link wordnets of different languages; each synset in the language-specific wordnet is linked to at least one record of the ILI by means of a set of equivalence relations (among which the most important is the EQ_SYNONYM, that expresses a total, perfect equivalence between two synsets).

Figure 6 describes the schema of a WN lexical entry. Under the root "synset" we find both internal relations ("synset relations") and ILI Relations, which link to ILI synsets.

Figure 3 shows the role played by the ILI as set of pivot nodes allowing the linkage between concepts belonging to different wordnets.

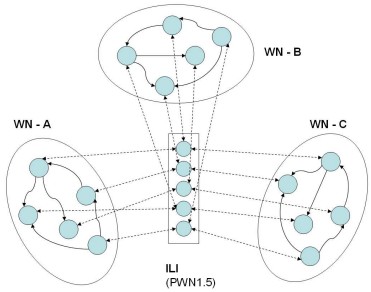

Figure 2. Interlingual Linking of Language-specific Synsets

In the Multilingual WN Service, only equivalence relations of type EQ_SYNONYM and EQ_NEAR_SYNONYM have been taken into account, being them the ones used to represent a translation of concepts and also because they are the most exploited (for example, in IWN, they cover about the 60% of the encoded equivalence relations). The EQ_SYNONYM relation is used to realize the one-to-one mapping between the language-specific synset and the ILI, while multiple EQ_NEAR_SYNONYM relations (because of their nature) might be encoded to link a single language-specific synset to more than one ILI record. In Figure 4 we represented the possible relevant combinations of equivalence relations that can realize the mapping between synsets belonging to two languages. In all the four cases, a synset "a" is linked via the ILI record to a synset "b" but a specific procedure has been foreseen in order to calculate different "plausibility scores" to each situation. The procedure relies on different rates assigned to the two equivalence relations (rate "1" to EQ_NEAR_SYNONYM relation and rate "0" to the EQ_SYNONYM). In this way we can distinguish the four cases by assigning respectively a weight of "0", "1", "1" and "2".

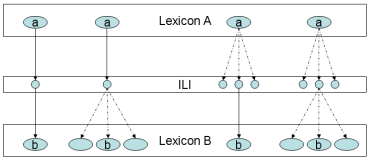

Figure 3. Possible Combinations of Relations between two Lexicons A and B and the ILI

The ILI is a quite powerful yet simple method to link concepts across the many lexicons belonging to the *WordNet-family*. Unfortunately, no version of the ILI can be considered a standard and often the various lexicons exploit different version of WordNet as ILI . This is a problem that is handled at web-service level, by incorporating the conversion tables provided by (Daudé et al., 2001). In this way, the use of different versions of WN does not have to be taken into consideration by the user who accesses the system but it is something that is resolved by the system itself . This is why the version of the ILI is a parameter of the query to web service.

On the basis of ILI linking, a synset can be enriched by importing the relations contained in the corresponding synsets belonging to another wordnet. In the procedure adopted, the enrichment is performed on a synset-by-synset basis. In other words, a certain synset is selected from a wordnet resource, say WN(A). The cross-lingual module identifies the corresponding ILI synset, on the basis of the information encoded in the synset. It then sends a query to the WN(B) web service providing the ID of ILI synset together with the ILI version of the starting WN. The WN(B) web service returns the synset(s) corresponding to the WN(A) synset, together with reliability scores. If WN(B) is based on a different ILI version, it can carry out the mapping between ILI versions (for instance by querying the ILI mapping web service). The cross-lingual module then analyzes the synset relations encoded in the WN(B) synset and for each of them creates a new synset relation for the WN(A) synset[1]. If the queried wordnets do not use the same set of synset relations, the module must take care of the mapping between different relation sets.

4. A Case Study: Cross-fertilization between Italian and Chinese Wordnets

Our case-study involves an Italian WordNet, ItalWordNet (Roventini et al., 2003), and the Academia Sinica Bilingual Ontological Wordnet (Sinica BOW, (Huang et al., 2004)).

The BOW integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED), and SUMO (Suggested Upper Merged Ontology). With the integration of these three key resources, Sinica BOW functions both as an English-Chinese bilingual wordnet and a bilingual lexical access to SUMO. Sinica Bow currently has two bilingual versions, corresponding to WordNet 1.6. and 1.7.

[1] For a more detailed description of the procedure, see (Soria et al., 2006)

Based on these bootstrapped versions, a Chinese Wordnet (CWN, (Huang et al., 2005)) is under construction with handcrafted senses and lexical semantic relations. For the current experiment, we have used the version linking to WordNet 1.6.

ItalWordNet was realized as an extension of the Italian component of EuroWordNet. It comprises a general component consisting of about 50,000 synsets and terminological wordnets linked to the generic wordnet by means of a specific set of relations. Each synset of ItalWordNet is linked to the Interlingual-Index (ILI).

The two lexicons refer to different versions of the ILI (1.5 for IWN and 1.6 for BOW), thus making it necessary to provide a mapping between the two versions. On the other hand, no mapping is necessary for the set of synset relations used, since both of them adopt the same set.

For the purposes of evaluating the cross-lingual module, we have developed a prototype WordNet grid containing just two web services that manage the two resources.

The following Figure shows a very simple example where our procedure discovers and proposes a new meronymy relation for the Italian synset passaggio,strada,via. This synset is equivalent to the ILI "road,route" that is ILI-connected with BOW synset "dao_lu, dao, lu" (Figure 7, A) . The Chinese synset has a meronymy relation with the synset "wan" (B). This last synset is equivalent to the ILI "bend, crook, turn" that is ILI-connected with Italian WordNet synset "curvatura, svolta, curva" (C). Therefore the procedure will propose a new candidate meronymy relation between the two Italian WordNet synsets (D).

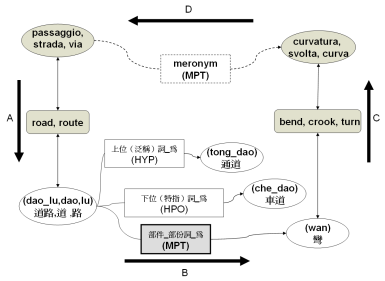

Figure 4. Example of a New Proposed Meronymy Relation for Italian

5. Conclusion

Our proposal to make distributed wordnets interoperable has several applications in processing of lexical resources. First of all, it can be used to enrich existing resources: information is often not complete in any given wordnet and, by making two wordnets interoperable, we can bootstrap semantic relations and other information from other wordnets. Second, it can be applied to the creation of new resources: multilingual lexicons can be bootstrapped by linking different language wordnets through ILI. Third, it can also be exploited for validation of existing resources: semantic relation information and other synset assignments can be validated when it is reinforced by data from a different wordnet.

In particular, our work can be proposed as a prototype of a web application that would support the Global WordNet Grid initiative (www.globalwordnet.org/gwa/gwa_grid.htm).

Any multilingual process, such as cross-lingual information retrieval, must involve both resources and tools in a specific language and language pairs. For instance, a multilingual query given in Italian but intended for querying English, Chinese, French, German, and Russian texts, can be send to five different nodes on the Grid for query expansion, as well as performing the query itself. In this way, language specific query techniques can be applied in parallel to achieve best results that can be integrated in the future. As multilingualism clearly becomes one of the major challenges of the future of web-based knowledge engineering, WordNet emerges as one leading candidate for a shared platform for representing a lexical knowledge model for different languages of the world. This is true even if it has to be recognized that the wordnet model is lacking in some important semantic information (like, for instance, a way to represent the semantic predicate). However, such knowledge and resources are distributed. In order to create a shared multi-lingual knowledge base for cross-lingual processing based on these distributed resources, an initiative to create a grid-like structure has been recently proposed and promoted by the Global WordNet Association, but until now has remained a wishful thinking. The success of this initiative will depend on whether there will be tools to access and manipulate the rich internal semantic structure of distributed multi-lingual WordNets. We believe that our work on LeXFlow offers such a tool to provide inter-operable web-services to access distributed multilingual WordNets on the grid.

This allows us to exploit in a cross-lingual framework the wealth of monolingual lexical information built in the last decade.

References

Calzolari, N. and C. Soria. A New Paradigm for an Open Distributed Language Resource Infrastructure: the Case of Computational Lexicons. In *Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (KCVC05)*. Stanford, CA, USA, pp. 110–114.

Calzolari, N. Technical and Strategic issues on Language Resources for a Research Infrastructure. In *Proceedings of the International Symposium on Large-scale Knowledge Resources (LKR2006)*. Tokyo Institute of Technology, Tokyo, Japan, pp. 110–114.

Daud, J., Padr, L. and G. Rigau. A Complete WN1.5 to WN1.6 Mapping. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Association for Computational Linguistics, Pittsburg, PA, USA, pp. 83–88.

Gulrajani, G., and D. Harrison. SHAWEL: Sharable and Interactive Web-Lexicons. In *Proceedings of the LREC2002 Workshop on Tools and Resources in Field Linguistics*. Las Palmas, Canary Islands, Spain, pp. 1–4.

Huang, C., Chang, R. and S. Lee. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In *Proceedings of LREC2004*. Lisbon, Portugal, pp. 1553–1556.

Huang, C., Chen, C., Weng, C., Lee, H., Chen, Y. and K. Chen. The Sinica Sense Management System: Design and Implementation. *Computational Linguistics and Chinese Language Processing*, 10(4): 417–430.

Kemps–Snijders, M., Nederhof, M. and P. Wittenburg. LEXUS, a web-based tool for manipulating lexical resources. In *Proceedings of LREC2006*. Genoa, Italy, pp. 1862–1865.

Peters, W., Vossen, P., Diez-Orzas, P. and G. Adriaens. Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index. In N. Ide, D. Greenstein and P. Vossen, editors, *Special Issue on EuroWordNet. Computers and the Humanities*, 32(2-3): 221–251.

Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Girardi, C., Magnini, B., Marinelli, R. and A. Zampolli. ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. In A. Zampolli, N. Calzolari and L. Cignoni, editors, *Computational Linguistics in Pisa*, IEPI, Pisa-Roma, pp. 745–791.

Ruimy, N., Monachini, M., Gola, E., Calzolari, N., Del Fiorentino, C., Ulivieri, M. and S. Rossi. A Computational Semantic Lexicon of Italian: SIMPLE. In A. Zampolli, N. Calzolari and L. Cignoni, editors, *Computational Linguistics in Pisa*, IEPI, Pisa-Roma, pp. 821-864.

Soria, C., Tesconi, M., Bertagna, F., Calzolari, N., Marchetti, A. and M. Monachini. Moving to Dynamic Computational Lexicons with LeXFlow. In *Proceedings of LREC2006*. Genoa, Italy, pp. 7–12.

Soria, C., Tesconi, M., Marchetti, A., Bertagna, F., Monachini, M., Huang, C. and N. Calzolari. Towards Agent-based Cross-lingual Interoperability of Distributed Lexical Resources. In *Proceedings of the COLINGACL 2006 Workshop on Multilingual Language Resources and Interoperability*. Sydney, Australia, pp. 7–12.

Tesconi, M., Marchetti, A., Bertagna, F., Monachini, M., Soria, C. and N. Calzolari. LeXFlow: a System for Cross-fertilization of Computational Lexicons. In *Proceedings of the COLINGACL 2006 Interactive Presentation Sessions*. Sydney, Australia, pp. 9–12.

Vossen, P. Introduction to EuroWordNet. In N. Ide, D. Greenstein and P. Vossen, editors, *Special Issue on EuroWordNet. Computers and the Humanities*, 32(2-3): 73–89.