Towards agent-based cross-lingual interoperability of distributed lexical resources

Claudia Soria
Maurizio Tesconi
Andrea Marchetti
Francesca Bertagna
Monica Monachini
Chu-Ren Huang
Nicoletta Calzolari

Istituto di Linguistica Computazionale - CNR - Pisa
Istituto di Informatica e Telematica - CNR - Pisa
Academia Sinica - Taipei
Motivation

- Ever-increasing expansion of language resources
- Inherent distributedness of LRs:
  - Locally developed and maintained
  - Strongly bounded to their natural environment
- Unsatisfactory language resources:
  - Lack of adequate breadth
  - Lack of adequate detail of linguistic information
  - Lack of wide availability
  - Time and money consuming
The answer

- A “new generation” of language resources:
  - From static, closed and locally developed resources to shared and distributed language services.
  - LRs reside over distributed places and are choreographed by agents presiding the actions that can be executed over them:
  - ..such as querying, collaborative development and validation, cross-resource integration and exchange of information.
  - This is a long-term scenario based on content interoperability standards, sovra-national cooperation and development of accessible architectures enabling accessibility.
Aims of work

- Explore new methods and techniques allowing the realization of “new paradigm” of language resources.
- Addressing integration and interoperability of computational lexicons.
- The case of semi-automatic integration and mutual enrichment of (distributed) large-scale lexical resources.
Two levels

Global focus:
• “..LRs are built as the result of controlled cooperation of different agents.”
• Development of cooperative web application for the management of lexical resources

Local focus:
• An application (actually a module of the previous one) enabling semi-automatic cross-lingual enrichment of lexical resources (cross-fertilization)
• Multilingual WordNet Service
LeXFlow

- A web-based collaborative environment for the *semi-automatic integration* of lexical resources, enabling interoperability of *distributed* lexical resources that are accessed by different types of *agents*.
LeXFlow design

- LeXFlow gets inspired from techniques of document workflows and cooperative authoring.

Xflow (Marchetti, Tesconi, Minutoli 2005), a cooperative web application for the management of document workflows.
From document to lexical flows

- Management of lexical resources as types of document workflows
  - Lexical entries are modelled as document instances
  - The behavior of a lexical entry is described by a Lexical Workflow Type
  - A Lexical Workflow Type describes:
    - The life-cycle of a lexical entry
    - The agents allowed to act over it
    - The actions to be performed by the agents
    - The order in which the actions are to be executed
Cross-fertilization and enrichment flow

Cross-fertilization agent: proposes new relations

Enrichment agent: proposes new relations extracted from corpus
Used and tested for integration of monolingual lexicons...

...With differently conceived lexical architectures and diverging formats

The same idea of *cross-fertilization*, i.e. semi-automatic induction of new information, however, can be applied in a cross-lingual perspective.
Moving to a cross-lingual perspective…

- A monolingual lexicon can be enriched by inducing the semantic information encoded in corresponding entries of other monolingual lexicons.
- To this end, the lexicons must share the same structural model.
- WordNet is the most widely spread model of semantic lexicons, with many initiatives worldwide.
- Harvesting the richness of various WordNets to enrich each of them, in a cross-breeding-like manner.
Our case-study: cross-fertilization between Italian and Chinese WordNets

- ItalWordNet (Roventini et al., 2003)
- Academia Sinica Bilingual Ontological WordNet (Sinica BOW, Huang et al., 2004)
- Both connected to Princeton WordNet (although to different versions)
- Same set of semantic relations (EWN ones)
Some basic assumptions

- **Interlingual level:**
  - There must be an Interlingua providing an indirect linkage between different WordNets, such as the Interlingual Index (ILI).

- **Synset correspondence:**
  - If there is a \( S_A \) and a \( S_B \) that point to the same ILI, they are correspondent.

- **Relation correspondence:**
  - If there are two synsets in \( WN_A \) and a relation between them, the same holds between corresponding synsets in \( WN_B \).
Linking WordNets through the ILI

- Interlingual Index (Peters et al. 1998)
- An unstructured version of WordNet used in EuroWordNet to link wordnets of different languages.
- Each synset in a WN_A is linked to at least one record of the ILI by means of a set of equivalence relations (e.g. “eq_synonym”, “eq_near_synonym”, “eq_has_hyperonym”, etc.)
Problems

- no version of the ILI can be considered a standard
- often the various lexicons exploit different version of WordNet as ILI
- Potential inaccuracy of the linking to ILI
- “Noise” induced by non-full synonymy relations.
Procedure

- Enrichment is performed on a synset-by-synset basis
- On the basis of ILI linking, a synset can be enriched by importing the relations contained in the corresponding synset(s) belonging to another wordnet.
- 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.
Cross-fertilization and enrichment flow

- MWS agent: proposes new relations
- Enrichment agent: Proposes new relations extracted from corpus
- Human agent A
- Human agent B
- Human agent C
- Human agent D
Conclusions

- We have presented a proposal for making distributed wordnets interoperable.
- This proposal lends itself to different applications in lexical resource processing:
  - Enrichment of existing lexical resources
  - Creation of new resources
  - Validation of existing resources
- If combined with LeXFlow, it can support the cooperative and collective creation and management of LRs, by providing a web-based environment for the collaboration and interaction of distributed agents and resources.
Conclusions

Prototype of a web application supporting the GlobalWordNet Grid initiative, i.e. a shared multi-lingual knowledge base for cross-lingual processing based on distributed resources over the Grid.
Links

- LeXFlow:
  http://xmlgroup.iit.cnr.it:8888/xflow/login

- MWS:
  http://xmlgroup.iit.cnr.it:88/exist/wordnet/wordnet

- GlobalWordNet Grid:
  www.globalwordnet.org/gwa/gwa_grid.html