Verso un lessico computazionale aperto per la lingua italiana

Il progetto Senso Comune

Guido Vetere
Associazione Senso Comune
IBM Centro Studi Avanzati di Roma

Pubblica Amministrazione Aperta e Libera, Pula (Cagliari), Italy 17-18 aprile 2008

Lessici computazionali

- Dizionari *machine-readable*
  - Morfosintassi
  - Accezioni
  - Relazioni lessicali
- Impiego
  - Tagging: le accezioni sono usate per caratterizzare risorse informative
  - Information Retrieval: le relazioni lessicali sono usate per potenziare le ricerche
  - Disambiguazione: attribuzione di accezioni a occorrenze di parole (token)
- Applicazioni
  - e-Commerce
  - e-Government
  - Web
  - Data mining
  - Digital libraries

WordNet

- WordNet (Princeton University): database lessicale per l'inglese, in cui i lessemi sono organizzati in gruppi di sinonimi (synset) che rappresentano concetti lessicali
  - Open Souce
- MultiWordNet (ITC): database lessicale multilingua, comprendente l'italiano, allineato con WordNet
  - Licenza

Senso Comune

- Associazione senza fini di lucro
  - Fondata nel Novembre 2006 da un gruppo interdisciplinare di studiosi

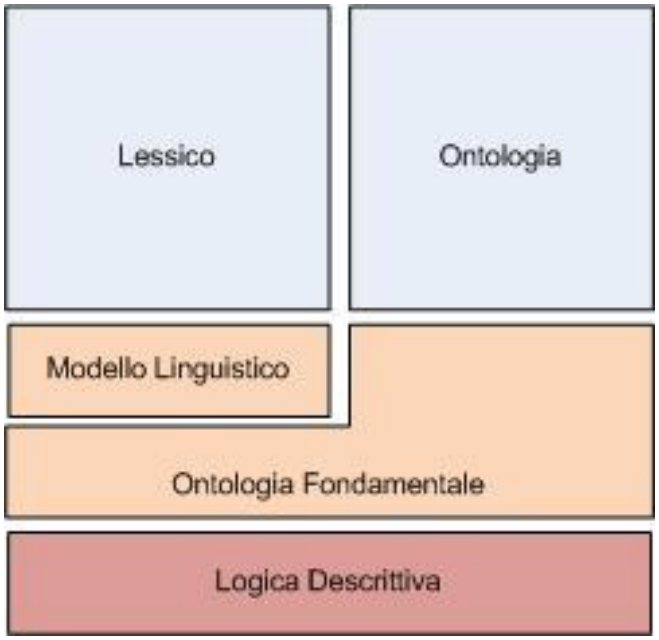
    T. De Mauro, U. Roma "La Sapienza" - A. Gangemi, CNR - N. Guarino, CNR - M. Lenzerini, U. Roma "La Sapienza" - M. Nissim, U. Bologna - G. Vetere, IBM Italia
  - Presidente: Prof. Tullio De Mauro
  - Sostenitore: Fondazione IBM Italia
  - Comitato Scientifico:

    Padre R. Busa (Emerito), Societas Jesu - N. Calzolari,CNR - G. De Giacomo,U. di Roma "La Sapienza" - R. Delmonte, U. Venezia - A. Elia,U. di Salerno - R. Rossini Favretti ,U. Bologna - A. Lenci,U. Pisa - L. Lesmo ,U. Torino - B. Magnini ,Fondazione Kessler - D. Marconi, U. Torino - M. T. Pazienza, U. Roma "Tor Vergata" - M. Poesio,U. Trento - P. Velardi,U. Roma "La Sapienza"

- Obiettivi
  - Costruzione di un lessico computazionale aperto per la lingua italiana
    - Modelli open source
    - Distribuzione Creative Commons
    - Contributi della comunità scientifica
    - Contributi della comunità di utenti (à-la wiki)

Struttura

- Linguaggio di modellazione
- Concetti di base
- Concetti linguistici
- Database lessicale
- Ontologia lessicale



Linguaggio di modellazione

- Logica descrittiva DL-Lite (Università di Roma ‘La Sapienza’)
  - Miglior rapporto espressività \ computabilità per l’accesso a volumi di dati
  - Capacità di gestire larghe ontologie (ragionamenti su memoria secondaria)
  - Corrispondenza col linguaggio UML

Ontologia di base

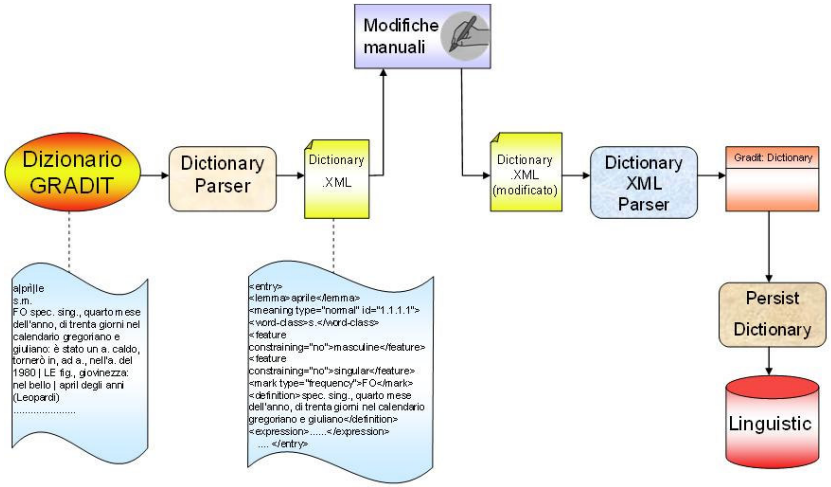
- Versione di DOLCE (CNR – LOA)
  - Distinzioni 'categoriali'
    - Entità concreta vs Qualità vs Astrazione
  - Relazioni
    - Composizione
    - Dipendenza
    - Partecipazione
    - Localizzazione
  - Mapping con categorie linguistiche
  - Mapping con WordNet
  - Metodologia di Q\A interattivo per la classificazione delle accezioni (TMEO)

Modello linguistico

- Integrazione del modello lessicografico dizionariale con un modello ‘user centered’
  - MeaningRecord := DictionaryMeaningRecord | UserMeaningRecord
- Distinzione (e relazione) tra ‘Accezione’ e ‘Concetto’
  - σ : MeaningRecord → Concept (non iniettiva)
  - Le relazioni lessicali riguardano le accezioni: le relazioni concettuali sono inferite a seguito di analisi semi-automatiche
- Mapping LMF

Il lessico di base

- Senso Comune parte dal lessico fondamentale di De Mauro
  - 2075 lemmi italiani più frequenti
  - ≈ 10000 accezioni fondamentali
  - Conversione semi-automatica dal formato dizionariale (GRADIT)
  - Disponibile da Giugno 2008



La piattaforma tecnologica

- Front-end
  - Portale \ Wiki \ Forum (Joomla)
  - Rich Internet Application (Ajax\GWT)
- Backend
  - Java API (UML\EMF)
  - DBMS (Hibernate)
  - Reasoner DL-Lite (memoria secondaria)
  - Web Services
- Distribuzione
  - Modelli: UML (XMI) – (Giugno 2008)
  - Risorsa:
    - Lessico: XML (Prima versione: Settembre 2008)
    - Ontologia: OWL (TBD)

Conclusione

- Oggi l'italiano manca di un suo dizionario-macchina di riferimento open-source
- Una risorsa di questo tipo è strategica per il web italiano – in primo luogo per la PA
- Senso Comune ha l'obiettivo di produrre questa risorsa e metterla a disposizione della comunità
- Il progetto sperimenta approcci e modelli innovativi per la rappresentazione della conoscenza linguistica

www.sensocomune.it

FONDAZIONE IBM ITALIA

IBM Centers for Advanced Studies
Centro Studi Avanzati di Roma