Ipotesi sull'applicazione del modello relazionale a corpora linguistici

Simone Torsani

XXI Convegno Ass.I.Term, Roma, 26 maggio 2011

Terminologie e ontologie: definizioni e comunicazione fra norma e uso

Obiettivi

- Impiego di metadati in corpora linguistici (soprattutto dal web)
- A supporto di ricerche (in questo caso terminologiche e lessicologiche):
  - Analisi comparative
  - Analisi diacroniche

Indice

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- Metadati e Web
- Definizione di strumenti
  - Estrazione/creazione
  - Analisi

Problematiche della ricerca

•**Problematiche della ricerca**
•Metadati e corpora
•Metadati e modello relazionale
•Modello relazionale, basi di dati e corpora
•Metadati e Web
•Definizione di strumenti
    •Estrazione
    •Analisi

- Due contributi "esemplari"
- Lüderling et al. (2007): più generale (web data for linguistic purposes), notano l'importanza dell'uso di metadati per analisi:
  - Comparative (metadati per categorizzare)
  - Diacroniche (metadati per ordinare cronologicamente)
- Claridge (2007): più specifico (corpus from the web: message boards), utilizzo (limitato) di metadati in messaggi di forum (poco standard, a basso livello…)
- Entrambi analizzano il "Web come corpus" (vol. Corpus Linguistics and the Web)

Emergono i punti deboli degli strumenti a disposizione e/o nell'uso dei metadati

Metadati e corpora

- Problematiche della ricerca
- **Metadati e corpora**
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- Metadati e Web
- Definizione di strumenti
  - Estrazione
  - Analisi

- Metadatazione a diversi livelli (Burnard, 2004) → "alto" (sul corpus), "intermedio" (sui componenti/testi), "basso" (su porzioni del testo – annotazioni)
- In Claridge (2007) si nota un utilizzo poco efficiente della metadatazione

Per es. non uniformità (es. date in formati diversi 05:06pm Nov 22, 2004 vs. 2 Feb 2005 16:33) → sarebbe difficile utilizzarli per categorizzare / ordinare

Importanza nella scelta del **sistema più efficiente / produttivo di metadatazione** (per es. il sistema relazionale)

Metadati e modello relazionale

- Problematiche della ricerca
- Metadati e corpora
- **Metadati e modello relazionale**
- Modello relazionale, basi di dati e corpora
- Metadati e Web
- Definizione di strumenti
  - Estrazione
  - Analisi

- Modello relazionale = tabella di base di dati
- Ogni relazione è un insieme (tupla) di attributi
- "Testo" = {id , anno, categoria, autore, testo}
- Ogni oggetto è un'istanza di questa relazione → ogni oggetto è un aggregato di attributi
- Se a un testo si aggiungono metadati, questi possono essere considerati come attributi e il corpus "diventa" una tabella: 1 testo = 1 record → 1 record = attributi (di cui una parte è testo, un'altra sono metadati)
- Funzione **query** → interrogazione della BD utilizzando criteri basati sugli attributi (e l'algebra relazionale): SQL (structured query language) → "SELECT * FROM testi WHERE anno>2008 AND anno<2011"

La distinzione tra Base dati e Corpus non è sempre chiara…

Modello relazionale, basi di dati e corpora

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- **Modello relazionale, basi di dati e corpora**
- Metadati e Web
- Definizione di strumenti
  - Estrazione
  - Analisi

- L'associazione base dati / corpus non è nuova e il confine tra i due non è netto, come si nota anche da alcuni termini utilizzati
  - Diversi *Corpus Query Language* (per interrogare corpora con *query*), per es. basati su espressioni regolari → [lemma = "bias"] [word = "towards|toward"] []{1,3}[tag= "NN."] (cerca collocazioni)
  - Corpus DWDS (tedesco) integra metadati cronologici (ordina per anno di pubblicazione)
  - Barbiers et al. (2007) corpus/database di dialetti olandesi: usano diverse tabelle per immagazzinare metadati (più base dati che corpus)
- Si tende ad oscillare verso uno o l'altro estremo (base dati vs. corpus e relativi strumenti di analisi)

Metadati e Web I

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- **Metadati e Web**
- Definizione di strumenti
  - Estrazione
  - Analisi

- Un maggior uso dei metadati rende necessaria la definizione di specifici ambiti di competenza (tra corpus/analisi linguistica e database/query).
- Il Web, i cui testi sono ricchi in metadati e oggi sempre più utilizzato come fonte per i corpora, rende urgente chiarire ruolo e procedure e dell'uno e dell'altro sistema anche in relazione agli obiettivi

Metadati e Web II

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- **Metadati e Web**
- Definizione di strumenti
  - Estrazione
  - Analisi

- Il Web è uno degli elementi più innovativi nel settore corpora
- Nel Web è facile che i testi siano associati / contengano metadati (come in Claridge) anche in diverse forme:
  - Metatag (es. "date")
  - Nel testo vero e proprio
- Diversi sistemi di estrazione dei metadati:
  - *Scraping*
  - *Information Extraction*
- Es. 1 siti di aziende: es. Sito di Intel → pagine "seriali" (nella struttura) di schede madri
- Es. 2 forum di discussione specializzato : es. www.tomshw.it

Testo e metadato sono separati (estrazione) e associati (nella relazione)

Metadati e Web III

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- **Metadati e Web**
- Definizione di strumenti
  - Estrazione
  - Analisi

In un'ottica di analisi di corpus, per combinare al meglio le due funzioni (ricerca e categorizzazione) è importante mantenerle distinte:

1) analisi testuale (attributi testo) e
2) *query* di selezione (attributi metadato) – v. Barbiers et al. (2007)

Definizione di strumenti – Estrazione

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- Metadati e Web
- Definizione di strumenti
  - **Estrazione**
  - Analisi

- Sistema di estrazione deve individuare e separare metadati dal testo
- Deve normalizzare (standardizzare) i metadati (per es. le date)
- Realizzabile con piattaforme che esistono già (GATE)

Definizione di strumenti- Analisi

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- Metadati e Web
- Definizione di strumenti
  - Estrazione
  - **Analisi**

- Il sistema di analisi deve permettere di eseguire una ricerca testuale "classica"
- Deve essere flessibile e adattarsi ai vari metadati
- Deve integrare le funzioni di ricerca testuale (concordanza, collocazioni ecc…) con funzioni di *query* sui metadati che:
  - filtrino / categorizzino i risultati
  - ordinino i risultati cronologicamente
- L'esempio mostrato in Barbiers et al. (2007) è più una base dati che un corpus

Conclusioni

- Problematiche della ricerca
- Metadati e corpora
- Metadati e modello relazionale
- Modello relazionale, basi di dati e corpora
- Metadati e Web
- Definizione di strumenti
  - Estrazione
  - **Analisi**

- Il Web è una fonte sempre più utilizzata per la costruzione di *corpora*
- L'utilizzo di metadati permette di svolgere molte più operazioni
- Il recupero di metadati da pagine web è più facile (in percentuale) rispetto a testi normali
- Lo sviluppo di sistemi in grado di estrarre / convertire / associare metadati sarà sempre più importante in questo ambito
- L'uso di un sistema che combini le due funzioni appare il più efficiente per ottenere risultati migliori, più standardizzati e applicabili a un numero illimitato di documenti