Giornate di Studio
Scritto e parlato, formale e informale. La comunicazione mediata dalla rete
Torino, 29-30 Ottobre 2010

I generi del web

tra tradizione e innovazione: un'analisi linguistica sulla base del corpus PAISÀ

Claudia Borghetti (Università di Bologna) - claudia.borghetti@unibo.it
Sara Castagnoli (Università di Trento/Bologna) - scastagnoli@sslmit.unibo.it
Marco Brunello (CNR di Pisa/University of Leeds) - mlmb@leeds.ac.uk

Il progetto PAISA’

(Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati)

- Collaborazione tra
  - Università di Bologna
  - CNR di Pisa
  - Università di Trento
  - EURAC Bolzano
- Corpus dell’italiano contemporaneo – 100 milioni di parole
- Materiali testuali on-line, scaricabili con procedure automatizzate
- Fruizione libera su piattaforma web dedicata (http://www.corpusitaliano.it)
- Testi rilasciati con licenze *Creative Commons*
- Finalità di ricerca linguistica e didattiche
- Livelli multipli di annotazione:
  - Linguistica: POS tagging, lemmatizzazione, parsing
  - Metadati: argomento, funzione e genere

Perché una classificazione dei testi

- Per compensare mancanza di controllo durante la compilazione del corpus
  - capire la composizione del corpus, ovvero quanto è (s)bilanciato rispetto a argomenti e generi/tipologie testuali
  - e eventualmente intervenire per bilanciarlo attraverso lo scaricamento mirato di ulteriori testi
- Per gli utenti:
  - maggiore consapevolezza sulla provenienza delle osservazioni
  - (se bilanciato) disponibilità di maggiore varietà testuale
  - modalità di ricerca più sofisticate (visualizzazione metadati, filtri, creazione sottocorpora ecc.)

I testi di PAISA' – compilazione del corpus

- Scaricamento da web con metodo BootCat (Baroni & Bernardini 2004)
  - *query* automatizzate su *Yahoo!*
  - per coppie casuali di parole
    - estratte dal corpus *La Repubblica*
    - vocabolario di base di De Mauro
  - si ottiene lista di url
  - scaricamento contenuto url
  - controllo
    - eliminazione file vuoti
    - eliminazione file contenenti poco testo (<150 parole)
    - eliminazione pagine non *Creative Commons*

PROCEDURA AUTOMATIZZATA tranne che per
- scelta *seeds*
- controllo "qualità"

- Inclusione di testi prodotti nell'ambito dei progetti di Wikimedia Foundation (Wikipedia, Wikinews, Wikibooks ecc.)

Conseguenze su qualità e tipologia dei testi

- Scelta testi *Creative Commons*
  - Vantaggi: possibilità di condividere risorsa
  - Svantaggi:
    - esclusione di molti tipi testuali nati con il web (email, chat, social forum ecc.) non CC
    - varietà limitata in termini di genere e argomento
- Ripulitura (*KrdWrd*) – obiettivo: solo testo pulito
  - rimozione immagini, menu e *boilerplate*
  - separazione aree di testo in cui si concentra l'informazione linguistica da quelle a basso contenuto testuale

→ unità di analisi ≠ pagina web completa
→ meno elementi per classificazione automatica

L’iter di ricerca

- Definizione tassonomia (argomento, funzione, genere) sulla base della letteratura su web genres
  - ampiezza della *palette*
  - web vs. Paisà
- Test di annotazione manuale
- Modifiche/aggiustamenti tassonomia
  - inclusione / esclusione classi
  - ridefinizione classi
- Tentativo di classificazione automatica con metodi di *machine learning*
  - addestramento classificatore sulla base di *feature* estratte da testi annotati manualmente
- Verifica output
- Allargamento del *training corpus* e/o ripensamento della tassonomia

La tassonomia attuale: l’argomento (1)

- Business - economia, commercio, finanza, lavoro, ecc.
- Arti - arte e letteratura: arti visive, architettura, cinema, musica, ecc.
- Hi-tech - informatica, computer, web, telefonia, elettronica, ecc.
- MSP - medicina, salute, psicologia
- Leisure - tv, moda, astrologia, sport, videogiochi, viaggi, cucina, ecc.
- FLERS - filosofia, lingua, religione, educazione / formazione, sociologia, ecc.
- S.naturali - scienze naturali: meteorologia, astronomia, biologia, fisica, chimica, matematica, geografia, ecc.
- PSS - Politica, società e storia: istituzioni, amministrazione (trasporti, esercito, ecc.), legge, geopolitica, ecologia, etica, ecc.

Topics	Esperimento di Clustering
Business	**Cluster 11** - euro milione mercato mese prezzo costo circa paese pagare miliardo dollaro banca produrre prevedere produzione società
Arti	**Cluster 4** - film libro cinema personaggio protagonista pubblicare opera regista titolo attore autore romanzo mostra artista; **Cluster 6** musica canzone amico cantante cantare musicale brano concerto
Hi-tech	**Cluster 19** - sito informazione internet sistema utente rete software digitale tecnologia online dato comunicazione computer blog servizio
MSP	**Cluster 13** - medico legge ricerca umano diritto malattia salute scientifico paziente problema tale bambino embrione scienza cura
Leisure	**Cluster 1** [sport]; **Cluster 2, 10, 14** [TV]; **Cluster 5** [motori]
FLERS	**Cluster 3** [religione] signore figlio autem eius uomo chiesa terra super dominus gesù sunt popolo egli israel quae l padre santo quod
S. Nat.	**N.D.**
PSS	**Cluster 0** - lavoro comune sociale progetto politico; **Cluster 8** [politica interna]; **Cluster 15** [politica estera]; **Cluster 17** – polizia forza carcere piazza manifestazione carabiniere gruppo

Topics	Esperimento di Clustering
??	**Cluster 7** - politico paolo luca marco antonio carlo giovanni roberto blog alberto andrea poesia giorgio francesco verso franco scienza alessandro gennaio
??	**Cluster 9** - città euro centro roma fino metro zona viaggio milano strada mare piazza parco notte partire lungo luogo circa volo
??	**Cluster 12** - ragazzo figlio uomo bambino sentire amico lasciare famiglia ragazza mano vivere morte piccolo padre morto momento tornare madre occhio
??	**Cluster 16** - credere scrivere piacere sentire problema gente guardare bello niente davvero leggere buono parere magari vostro lasciare ciò visto sperare
??	**Cluster 18** - acqua animale prodotto piccolo colore usare terra bianco vino minuto circa mangiare carne cane rosso lasciare nero forma fino

La funzione (1)

- Raccomandare: raccomandare, consigliare, convincere, persuadere, ecc. [Sharoff, 2004: Recommendation]
- Informare: Informare, descrivere, presentare e raccontare, esprimere se stessi/raccontarsi, ecc. [Information]
- Argomentare: Argomentare, discutere, commentare e valutare [Discussion]
- Intrattenere: Intrattenere e divertire [Recreation]
- Istruire: Dare istruzioni, insegnare [Instruction]

La funzione (2)

“[…] Ad andare sul posto è il direttore di Italymedia.it Antonello De Pierro, nonché voce storica di Radio Roma e presidente del movimento nazionale “L’Italia dei diritti”. Il noto giornalista giunge nei locali della struttura esattamente alle ore 18,42, e quindi 18 minuti prima dell’orario canonico di chiusura al pubblico, con l’intenzione di effettuare un pagamento tramite bollettino postale. All’ingresso viene fermato da un impiegato, che con grande naturalezza lo avvisa del fatto che l’ufficio è già chiuso, cosa piuttosto singolare e assurda vista l’ora. Ignorando l’azzardato avvertimento si reca comunque presso la macchina erogatrice dei numeri progressivi che regolano l’affluenza agli sportelli, e qui lo attende un’indicibile sorpresa: dalla fessura esce un biglietto con la scritta “IL SERVIZIO NON E’ ATTIVO”. […] ccmod.4168

Il genere: problemi di definizione

Che cos'è il genere testuale?

- Genre as social action (Miller, 1984)
- A genre is a class of communicative events (Swales, 1990)
- A recurring type or category of text, as defined by structural, thematic and/or functional criteria (Duff, 2000)
- Genre has the power of predictivity (Bateman, 2008), Biber (1988), D. Lee (2001), Bruce (2008), Heyd (2008) ecc.

Che cos'è il genere (testuale sul) web?

- Cybergenres can be divided in two main classes, extant and novel (Shepherd & Watters, 1998)
- Genres are cultural products, linked to a culture, a society, a community. The Web is a new, large and heterogeneous community (Santini, 2005)

La ricerca sui generi web: questioni aperte

- Unità di analisi: problema dei documenti web
- *Palette* e annotazione: chi deve definire la *palette* dei generi web e annotare i documenti?
- Livello di generalità dei generi: a quale livello annotarli?
- Informatività delle *feature*: quali *feature* sono più informative del genere web? Solo quelle linguistiche o anche altre riconoscibili nel layout della pagina (immagini, link, ecc.), url e codice HTML?

La tassonomia per genere dei testi di PAISA’

Blog?	1° livello	2° livello
	Fiction	Prosa - Poesia - Sceneggiatura
✔	Guida	Tutorial - FAQ - Turismo - Ricetta
✔	Giornalismo	Cronaca - Editoriale - Intervista - Reportage - Recensione
	Accademia	Prosa - Lezione - Abstract
	Doc. ufficiale	Legge – Relazione - Contratto
	Scheda	Prodotto - CV - About page
	Annuncio	
	Commento	
	Lemma	

Due esempi di generi diversi su blog

“[…] Il WWF Italia valuta che, visti i rilevi riguardanti in particolare le aree di Capo Peloro – Laghi di Ganzirri e la Costa Viola (aree dove dovrebbero sorgere i piloni del ponte) e i Monti Peloritani (su cui impattano le strutture aeree del ponte), se la procedura si concludesse con il deferimento alla Corte di Giustizia europea, l’Italia sarebbe obbligata a mettere in un cassetto l’attuale progetto, che è stato posto a base di gara, e ri-elaborare un proposta radicalmente diversa da quella attuale” ccmod.1112 Annotazione: [gio.cr - pss - inf]*

“Oggi ho girato un po' di negozi di abbigliamento, e sembravano diventati tutti punti vendita di merchandising della Fiorentina. Sarà contenta mae, e lo sono pure io, ché il viola, specie scuro, è un colore che mi è sempre piaciuto. Al punto che, cosa che mi succede piuttosto di rado, ho comprato una camicia solamente perché mi piaceva, e non perché nel mio armadio non c'è più nulla che non stia cadendo a pezzi. Io non conosco i nomi dei colori che danno gli stilisti, ma direi che secondo la nomenclatura X11 è Purple gessata di DarkViolet” ccmod.048 Annotazione: [com - lei - arg]*

Due esempi di cronaca

“[…] Il Gruppo Angelucci ha versato 500.000 euro alla lista di Fitto in occasione delle elezioni regionali del 2005. Secondo il gruppo (Tosinvest), si tratta di un regolare finanziamento registrato a bilancio. Per la Procura di Bari si tratta invece di una tangente pagata per assicurarsi l’appalto da 198 milioni di euro con cui Angelucci ha ottenuto la gestione delle undici residenze sanitarie “assistite” dalla Regione Puglia[1]. Si tratta della stessa inchiesta per cui è indagato Francesco Storace. Il parlamento, tuttavia, ha respinto l’autorizzazione a procedere con l’arresto con 457 voti favorevoli (su 462 presenti), 1 contrario (Antonio Borghesi dell’Idv) e 4 astenuti” ccmod.1112 Annotazione: [gio.cr - pss - inf]*

“Il 10 di maggio, durante il suo discorso al Parlamento sullo Stato della Nazione, il Presidente Vladimir Putin ha annunciato che la Russia renderebbe il Rublo “internazionalmente convertibile”, così da poterlo utilizzare nelle transazioni riguardanti petrolio e gas naturale. Al momento, il petrolio viene esclusivamente valutato in dollari. L'annuncio di Putin risuona come un annuncio di guerra. ccmod.3667 Annotazione: [gio.cr - pss - inf]

Scritto e parlato, formale e informale

- Scritto e parlato: netta prevalenza dell'uso scritto della lingua
  - conferma dalle performance del POS-tagger:
    - POS-tagger (Pisa) è stato dell'arte per l'italiano, accuratezza del **96.34% - 97.10%**
    - Sul corpus Paisà → 95.10%

      - primo addestramento (campione casuale) → 95.70%

      - secondo addestramento (campione scelto) → **96,03%**
    - Difficilmente incrementabile:
      - (1) *Approposito del processo "Cogne bis", ma siamo seri, boja Fauss…!!!;*

      - *(2) quann è arrivat l'ora, s'add calà o sipario …*
- Formale e informale:
  - Generi tradizionali prestati al web
  - Nuovi generi
  - Generi tradizionali adattati

Work in progress e sviluppi futuri

- Annotazione automatica
  - [approccio *machine learning*] estrazione *feature* e addestramento classificatori
    - Prima fase: *feature* di tipo testuale [n° tokens, n° frasi, lunghezza media frasi, parole funzionali vs. lessicali, POS (n-grams) più frequenti ecc.]
    - Seconda fase: HTML tags (layout, links ecc.) e URL
  - Estensione esperimento *clustering* a funzione e genere
- A seconda dei risultati, validazione o disconferma/riproblematizzazione della tassonomia

Annotazione linguistica

- POS-tagger (Pisa) è stato dell'arte per l'italiano, accuratezza del 96.34% - 97.10% (senza e con dizionario)
- Sul corpus Paisà: 95.10%
- Per adattarlo ai testi del web, correzione manuale di 20.000 tokens e riaddestramento
  - Errori di segmentazione del testo
  - Nuove abbreviazioni
  - Emoticons e sequenze "strane" di caratteri
- Accuratezza al 95.70%
- Correzione manuale di ulteriori 20.000 token, selezionati in base a difficoltà
  - Scrittura non (italiano) standard
    - *Approposito del processo "Cogne bis"; ma siamo seri ; Boja Fauss…!!!*
    - *…quann è arrivat l'ora, s'add calà o sipario …*
- Accuratezza al 96.03%, difficilmente incrementabile