IL DIGITALE PUÒ SALVARE LA CULTURA?

RAGIONAMENTI ED ESPERIENZE SULL'INFORMATICA UMANISTICA

Internet Festival – 10 ottobre 2014

TECNOLOGIE DEL LINGUAGGIO

NUOVI MESTIERI, NUOVE RICERCHE

Maria Simi

Dipartimento di Informatica

Università di Pisa

Internet Festival – 10 ottobre 2014

INTRODUZIONE

- Il settore dell'elaborazione del linguaggio naturale
  - un settore interdisciplinare tra linguistica e informatica
- L'età della parola (IF 2013)
  - un cambio di paradigma abbastanza recente
  - esempi di successi commerciali
  - settore vitale dal punto di vista occupazionale
- NLP per la ricerca nelle DH
- Il mestiere del linguista computazionale

LINGUAGGIO NATURALE

- **Naturale**, contrapposto ad artificiale
- La forma più **naturale** di espressione, tipica dell'uomo
- Sarebbe **naturale** usarlo per interagire coi computer
- La "comprensione" da parte delle macchine difficile
  - Un problema affrontato fino dagli esordi dell'informatica
  - Fino a pochi anni fa era ancora fantascienza

TRADUZIONE AUTOMATICA

- Gli chiese di riorganizzare Forza Italia

  The churches to reorganize Italy Force (Systran)

  She asked him to reorganize Forza Italy (Google, Bing)

- Il ministro Stanca si è laureato alla Bocconi

  Minister Stanca has graduated himself to the Mouthfuls (Systran)

  The Minister Stanca is a graduate of Bocconi (Google)

  The Minister Stanca graduated from Bocconi University (Bing)

SUCCESSI RECENTI

CAMBIO DI APPROCCIO

- I bambini imparano a parlare mediante l'interazione con gli adulti … ancora prima di imparare la grammatica
- È possibile fare "imparare" il linguaggio ai computer nello stesso modo?
- Senza codificare in maniera puntuale le regole grammaticali?

SISTEMI AD APPRENDIMENTO STATISTICO

- Apprendimento "supervisionato"
  - Esperienza come collezioni di testi annotati
- Capacità di elaborazione su Big Data
  - Se fossero stati usati 10 anni fa non avrebbero ancora terminato
- Tecniche simili per il parlato, il testo, le immagini …

NECESSITÀ DI GRANDI QUANTITÀ DI DATI

- Quali dati?
  - Testi rappresentativi o n-grammi
  - Testi classificati (categorie, polarità)
  - Testi annotati (sintassi, semantica, *treebank*)
- Crowdsourcing e giochi “con scopo”
- Potenza di calcolo e di storage
- Risorse linguistiche aperte e condivise, basate su standard di annotazione
- Più sono i dati, migliori sono gli strumenti

APPLICAZIONI POSSIBILI

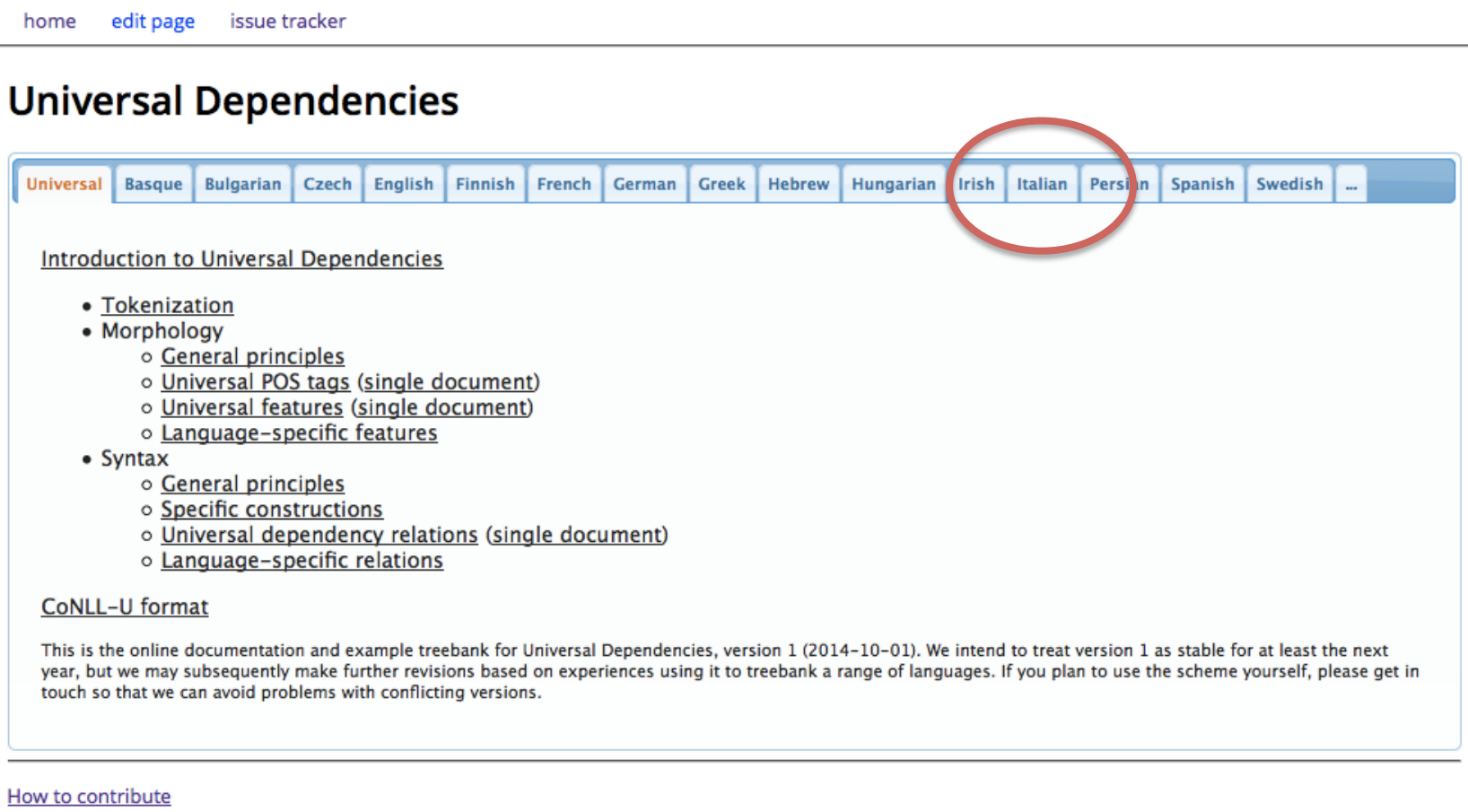
- Classificazione automatica di documenti
- Individuazione di entità semantiche (organizzazioni, persone, eventi, luoghi …)
- Analisi di sentimenti e opinioni
- Individuazione di interessi, tendenze, intenzioni di acquisto
- Sistemi di raccomandazione
- Ricerca semantica
- Sistemi di domanda-risposta (Question Answering)
- Interazione in linguaggio naturale
- Traduzione automatica
- … e molto altro

WEBSAYS + TISCALI

MA IL DIGITALE PUÒ SALVARE LA CULTURA?

- Come questo risponde alla questione iniziale?
  - La globalizzazione/tecnologia “corrompono” le lingue, l’inglese sta diventando lingua dominante nella Rete
  - In questo secolo più di 3000 lingue cesseranno di esistere (circa il 50%) [*R. G. Gordon*]
- *Le lingue che non svilupperanno adeguati strumenti di trattamento automatico sono destinate a scomparire … insieme alla cultura da esse veicolata*

ANALISI LINGUISTICA "UNIVERSALE" (GOOGLE)



home edit page issue tracker

Universal Dependencies

Universal | Basque | Bulgarian | Czech | English | Finnish | French | German | Greek | Hebrew | Hungarian | Irish | Italian | Persian | Spanish | Swedish | ...

Introduction to Universal Dependencies

- Tokenization
- Morphology
  - General principles
  - Universal POS tags (single document)
  - Universal features (single document)
  - Language-specific features
- Syntax
  - General principles
  - Specific constructions
  - Universal dependency relations (single document)
  - Language-specific relations

CoNLL-U format

This is the online documentation and example treebank for Universal Dependencies, version 1 (2014-10-01). We intend to treat version 1 as stable for at least the next year, but we may subsequently make further revisions based on experiences using it to treebank a range of languages. If you plan to use the scheme yourself, please get in touch so that we can avoid problems with conflicting versions.

How to contribute

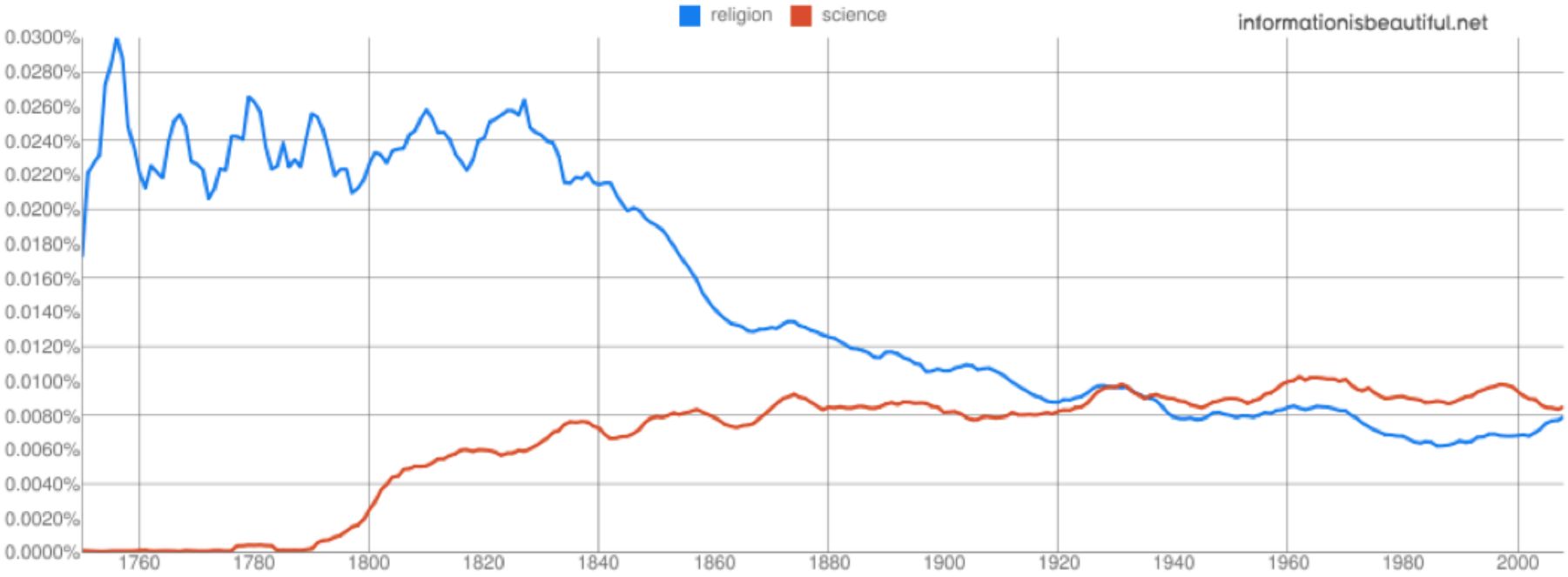
NLP E RICERCA UMANISTICA

- Franco Moretti, direttore dello Stanford Literary lab. propone una lettura “da lontano” dei testi
- Nuove pratiche di ricerca nelle Digital Humanities adottano questo paradigma e le tecnologie del linguaggo come strumento.
- Ma chiaramente NLP può contrbuire anche ad una lettura più accurata, “da vicino”.

NLP E DH:LETTURA DA LONTANO

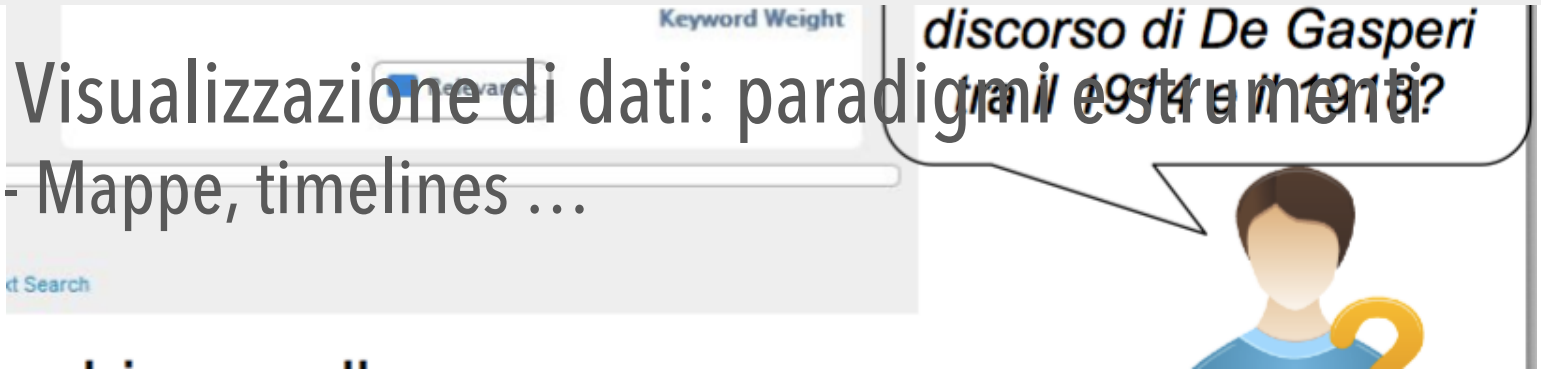
- Collezioni di testi storici o letterari
  - Riconoscimento di entità: nomi di persone, luoghi, date
  - Estrazione di eventi e loro correlazioni temporali
  - Analisi di “tendenze”
  - Analisi dei grafi di interazione sociale tra i personaggi
  - Analisi dei generi letterari
  - …
  - Georeferenziazioni e linee del tempo

ESEMPIO 1: GOOGLE NGRAM VIEWER



ESEMPIO 2: PROGETTO ALCIDE (FBK)

- Visualizzazione di dati: paradigmi e strumenti
  - Mappe, timelines …



NLP E DH: LETTURA DA VICINO

- Analisi delle strutture linguistiche del testo
  - complessità e leggibilità dei testi
  - valutazione delle competenze linguistiche
  - Identificazione del genere testuale
- Importanti per …
  - comunicazione PA – cittadino
  - campo didattico
  - recupero dei documenti sulla base del registro linguistico

LINGUISTA "COMPUTAZIONALE"

- Un mestiere interdisciplinare in cui sono richieste competenze linguistiche e informatiche, ma anche matematiche, di visualizzazione di dati …
  - Lavoro di squadra e dialogo difficile per le persone che si sono formate in ambiti disciplinari classici
- Pratiche di *apertura* dei dati, aderenza agli standard e condivisione indispensabili per massa critica
- *Apertura, standard, condivisione, interdisciplinarietà*

PROSSIMAMENTE A PISA

CLiC it

Prima conferenza italiana di Linguistica Computazionale

9-10 dicembre 2014

Evalita 2014. *Evaluation of NLP and Speech Tools for Italian*

11 dicembre 2014