Università degli Studi di Pisa
Facoltà di Scienze Matematiche Fisiche e Naturali
Corso di Laurea Triennale in Informatica

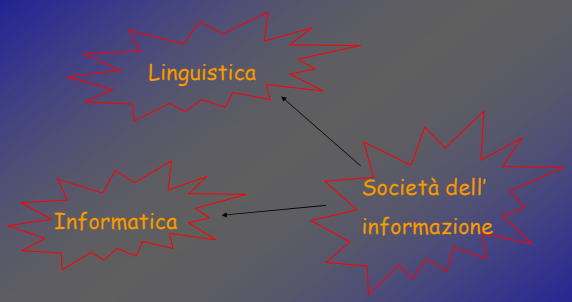
Sviluppo di Risorse Linguistiche per l'Ambiente di Scrittura Assistita (ASA)

Relazione Stage Formativo svolto presso la Synthema s.r.l.
Tutore Accademico: Paolo Mancarella

Tutori Aziendali: Amedeo Cappelli e Carlo Aliprandi

Canditato: Elisa Croci

Informatica e Linguistica…



Il ruolo del computer nella linguistica

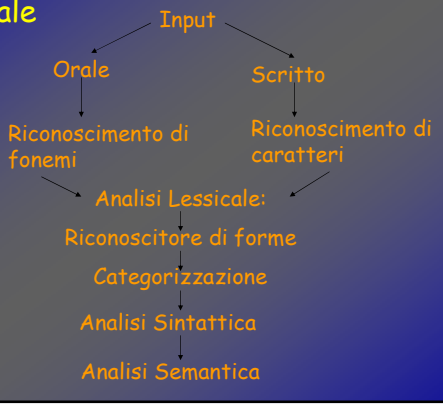
- Traduzione automatica
- Elaborazione di testi (TextProcessing)
- Classificazione di documenti
- Riconoscimento vocale (Automated Speech Recognition, ASR)
- Estrazione di conoscenza da testi (Information extraction)
- Recupero di informazione (Information retrivial)
- Interazione con gli uomini in modo naturale

Cosa è la Linguistica Computazionale

È la disciplina che si occupa dell'analisi ed elaborazione del linguaggio naturale per mezzo del computer, mettendo in luce le sue grandi capacità di elaborazione e la possibilità di un suo addestramento a fare ricerche, ordinare e calcolare dati quantitativi della lingua.

Natural Language Processing (NLP)

Processo di elaborazione del linguaggio naturale



Quali conoscenze deve avere il computer?

- FONETICHE per articolare e decodificare i suoni di una lingua
- LESSICO - MORFOLOGICHE per conoscere la struttura e l'organizzazione delle parole
- SINTATTICHE per la costruzione corretta delle frasi
- SEMANTICO - PRAGMATICHE per riuscire a dare significato alle frasi con riferimento al contesto in cui esse sono state formate

I corpora (1)

- I corpora testuali rappresentano la principale fonte di dati nella linguistica computazionale

Un corpus è una collezione di testi selezionati e organizzati in maniera tale da essere un campione rappresentativo della lingua che vogliamo descrivere.

I corpora (2)

I requisiti principali che un corpus deve soddisfare sono:

- DIMENSIONE: la quantità di documenti che costituiscono il corpus deve essere sufficiente per poter affermare che tale raccolta rappresenta in modo completo la lingua in esame
- AUTENTICITA': i documenti raccolti devono essere reali e non costruiti ad hoc per la nostra analisi
- BILANCIAMENTO: testi presi in esame devono appartenere a diverse tipologie della lingua (lingua scritta e lingua parlata trascritta da notiziari televisivi, radio, conversazioni casuali)

Il progetto…

- Creazione di un dizionario medico -radiologico (ItalRad)
- Test di copertura del linguaggio radiologico da parte del vocabolario ItalRad mediante l'Ambiente di Scrittura Assistita sviluppato presso la Synthema in collaborazione con l'Università di Pisa mediante i due tesisti Nicola Carmignani e Daniele Barsocchi

…e le risorse linguistiche

- **Corpus specialistico** composto da circa 6.000 referti radiologici di tipologia diversa e provenienti da diversi medici
- **ItalBase** è il vocabolario generale della lingua italiana
- **Terminology Wizard** e **Lexical Studio** sono applicazioni software per il trattamento del linguaggio naturale

Struttura di un vocabolario (1)

Lemma: (POS, Features, IRule, Arule)

- **Part Of Speech** indica una delle nove categorie grammaticali previste dalla lingua italiana

  Ideale: sostantivo maschile singolare

  Ideale: aggettivo maschile femminile singolare

- **Features** sono un elenco di caratteristiche sintattiche, grafiche e altre informazioni pratiche utili alla grammatica per la costruzione delle frasi

Struttura di un vocabolario (2)

- **Inflection Rule** (IRule) individuano le regole da seguire per generare il maschile e il femminile singolare, plurale di un lemma

IRule(2) amico: MALE SING → amico
FEMALE SING → amica
MALE PLUR → amici
FEMALE PLUR → amiche

Struttura di un vocabolario (3)

- **Alteration Rule** (ARule) sono regole che permettono di generare le diverse forme alterate (superlativo, diminuitivo, vezzeggiativo…) di un lemma

ARule(1) piccolo: MALE SING → piccolino
FEMALE SING → piccolina
MALE PLUR → piccolini
FEMALE PLUR → piccoline

Creazione di ItalRad (1)

- Con Terminology Wizard abbiamo estratto dal corpus una lista di circa *2.700 parole sconosciute* rispetto al vocabolario generale italiano
- Lemmatizzazione delle parole appartenenti alla lista
- Assegnazione delle POS a ciascun lemma ottenendo così una lista di coppie (lemma: POS) del genere

alveolitico: ADJ
anecogeno: ADJ
carcinoma: NOUN

Lemmatizzare significa portare ciascuna parola alla propria forma base chiamata *lemma*

Creazione di ItalRad (2)

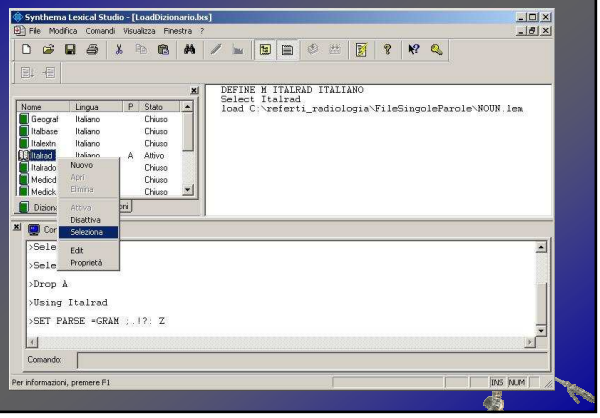
- Con Lexical Studio abbiamo analizzato un file composto da coppie (LEMMA: POS) mediante uno script ottenendo come risultato dell'elaborazione due file con estensione

.LEM contenente la lista dei lessemi con associata la propria IRule

.WRD contenente le espansioni di ciascun lessema applicando la regola di flessione ad esso assegnata

- Il file .LEM è necessario per inserire i lessemi nel vocabolario che possiamo creare usando Lexical Studio e riempire mediante un file di script

Creazione di ItalRad (3)



Creazione di ItalRad (4)

Dopo il caricamento di tutti i file .LEM abbiamo ottenuto un vocabolario composto da:

7.362 lemmi

226.296 forme

308.523 classificazioni

Testing: l'applicazione utilizzata (1)…

L'*Ambiente di Scrittura Assistita* (ASA) effettua la predizione della parola che l'utente sta scrivendo completandola automaticamente risparmiando così

il numero di tasti da digitare

evitando possibili errori ortografici

I componenti principali di questa applicazione sono:

- *l'interfaccia utente* formata da una finestra di editing la cui finalità è quella di mostrare le elaborazioni effettuate per la predizione della parola

…l'applicazione utilizzata (2)

- il *software di predizione* il quale entra in azione ad ogni carattere digitato, mostrando un elenco di parole che potrebbero completare quella digitata; i suggerimenti proposti sono ortograficamente corretti e rispettano le concordanze morfo-sintattiche. La previsione è effettuata a livello di frase, distinguendo tre casi: la prima parola, la seconda e le successive
- le *risorse linguistiche* sono costituite dai dizionari e dalle grammatiche

Testing: i criteri

- numero di caratteri risparmiati *(keystroke saving)* grazie alla predizione
- numero medio di digitazioni per visualizzare nella lista dei suggerimenti, la parola che stiamo scrivendo
- tempo di composizione necessario a scrivere un referto

Velocizzare la scrittura in ambienti dove le periferiche di input sono molto piccole

Aiutare le persone disabili nell'interazione con il computer

Testing: i risultati

I documenti su cui abbiamo svolto i test sono stati prelevati in modo casuale dal corpus medico di partenza cercando di ricoprire tutte le tipologie di referti e le diverse fonti

40 referti

34 parole

238 caratteri

Al termine di questa fase di testing abbiamo ottenuto, in media un

keystroke saving del 25%

Esempio referto

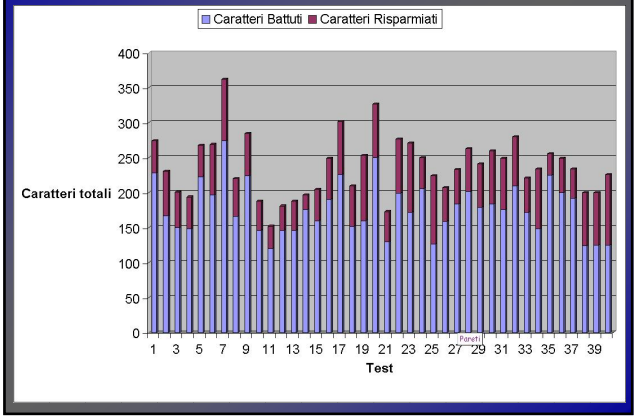
(medico2torace.txt)

Ipertrofia compensatoria del rene di sinistra. La dissezione sembra coinvolgere in parte l'origine dell'arteria renale di sinistra. L'arteria destra è ipoplasica e sembra nascere dal lume vero.

Caratteri totali: 200
Caratteri digitati: 125

Keystroke Saving: 38%

Grafico dei risultati



Conclusioni (1)…

Nella fase di testing, per alcuni referti abbiamo ottenuto ottimi risultati raggiungendo un

keystroke saving del 43%

Il risultato ottenuto è estremamente positivo e incoraggiante, visto che il modello del linguaggio (grammatica e il software di predizione) usato per predire i referti radiologici non è stato addestrato per un linguaggio specialistico come quello medico il quale ha una struttura grammaticale diversa dall'italiano generale

…Conclusioni (2)

Sviluppi futuri:

- creazione di un modello del linguaggio che descriva la struttura del sottolinguaggio medico