Componente Lessicale

- Scopi
  - Riconoscere gli elementi lessicali
  - Assegnare agli elementi lessicali informazioni sulla loro categoria grammaticale
  - Risolvere l'ambiguità grammaticale
    - Vedi lezione sull'ambiguità

Struttura e funzioni del modulo lessicale

Riconoscitore di forme (tokenizer)

- Segmenta il testo in parole e altre sequenze significative di caratteri (token), eventualmente separati da segni di interpunzione

- Categorizzatore (tagger)
  - Assegna categorie grammaticali ai token

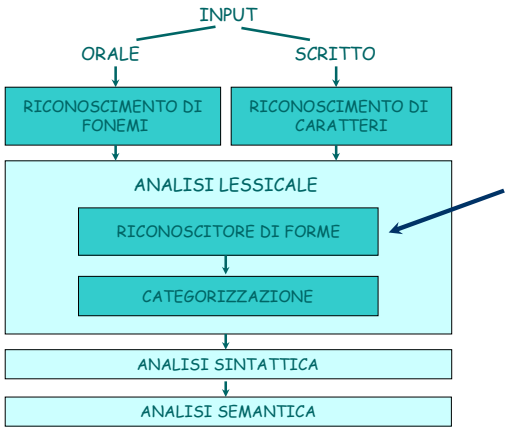
Riconoscimento e categorizzazione

- Le due fasi possono essere in parte indipendenti l'una dall'altra, ma anche interagire e sovrapporsi perché:
  - un componente da solo non è in grado di raggiungere lo scopo
  - mentre si segmentano i token, si assegnano anche le categorie

Riconoscitore di forme (tokenizer)

Scopo:

riconoscere le parole e le altre sequenze significative di un testo



Forma e lemma

- Il testo si presenta come una sequenza di "forme" grafiche, cioè un insieme di parole diverse
- Le forme grafiche possono essere ricondotte ad una voce di base, "lemma", sulla base di convenzioni lessicografiche

Lemma

Forme:		**Lemma:**
casa, case	→	casa
bello, bella, belli, belle	→	bello
mangio, mangia, mangiamo…	→	mangiare

Tokenizer

• Segmentazione di una sequenza di caratteri in sequenze di parole, simboli, segni di interpunzione, ecc.

- Parole: babbo, cane, casa, mangio…
- Polirematiche (multiwords): Banca d'Italia, a pronta presa…
- Sigle: CNR, INPS, CGIL…
- Punteggiatura
- Numeri arabi e romani
- Date: 31.12.1945, 1 gennaio 200…
- Indirizzi di posta elettronica: nerone@romaincendiata.ir
- Numeri telefonici: 39 050 666666

Tokenization

- Processo importante che permette di individuare le unità lessicali e i confini di frase necessari per la comprensione
- Dalla qualità del risultato di questo processo dipende il successo dei risultati delle operazioni successive
- Dalla qualità del risultato dipende anche il successo dell'applicazione per la quale il sistema è stato progettato

Tipi di conoscenze per il riconoscimento di forme

- (tipo)grafiche (input scritto)

Tokenization

- Considerare convenzioni grafiche e tipografiche che differiscono da lingua a lingua
  - Inglese: o'clock, Peter's, first-rate
  - Italiano: auto-analisi, nonsoché/non so che, tiremmolla/tira e molla, - Come stai? - gli chiesi
  - Francese: chemin-de-fer, as-tu

Tokenization: un problema

- Riconoscimento di date:
  - necessario tener conto dei vari stili con i quali è possibile scrivere una data:
    - 25 aprile 1945
    - 25-4-1945
    - 25/4/1945
    - 25.4.1945
    - Venticinque aprile millenovecentoquarantacinque

Esempi di selezione delle forme di un testo

- Lunedì 25 maggio u.s., la Banca d'Italia ha abbassato il tasso d'interesse di due punti, portandolo dall'8 al 6%. Soddisfazione tra i ceti produttivi che vedono nuove prospettive per il rilancio dell'economia. Entusiasta reazione della Borsa.

Segmentazione senza restrizioni

- Utilizzazione del comando di Word: "Converti testo in tabella"
- Vengono selezionate le sequenze di caratteri comprese tra due spazi bianchi
- Segni di interpunzione (punti, virgole, apostrofi, ecc.) inglobati nella sequenza selezionata

Lunedì
25
maggio
u.s.,
la
Banca
d'Italia
ha
abbassato
il
tasso
d'interesse
di
due
punti,
portandolo
dall'8
al
6%.

Segmentazione con restrizione su alcuni tipi di dato

- Riconoscimento dei seguenti tipi di dato
  - Alfanumerico
  - Numero
  - Segni di interpunzione
- Riesce a distinguere i numeri e i segni di interpunzione rispetto alle stringhe alfanumeriche

Lunedì	alfanumerico
25	numero
Maggio	alfanumerico
u	alfanumerico
.	punto
s	alfanumerico
.	punto
,	virgola
la	alfanumerico
Banca	alfanumerico
d	alfanumerico
'	apice
Italia	alfanumerico
ha	alfanumerico
abbassato	alfanumerico
il	alfanumerico
tasso	alfanumerico
d	alfanumerico
'	apice
interesse	alfanumerico
di	alfanumerico
due	alfanumerico
punti	alfanumerico
,	virgola

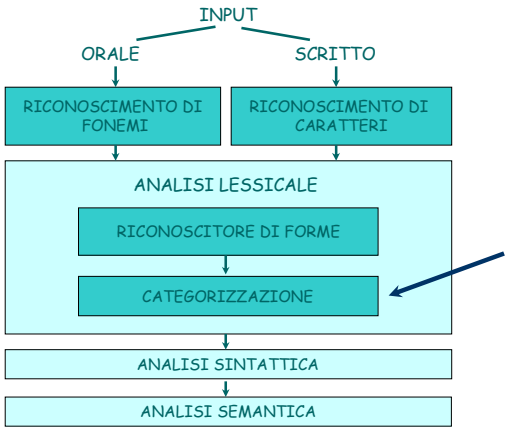
Metodi per il riconoscimento delle forme

- Ricerca delle forme all'interno di un lessico, sia generico che specialistico (dizionario di nomi, ecc)
  - Per individuare parole, multiwords, sigle
- Utilizzo di automi per ricercare schemi ricorrenti
  - Per riconoscere una data secondo le diverse convenzioni
    - Vedi Seminario De Pascalis (2002) in http://www.di.unipi.it/~cappelli/
  - Indirizzi di posta elettronica e indirizzi web

Riconoscimento di multiword con dizionario

Categorizzazione (tagging)

- Associare informazioni lessicali ad ogni forma riconosciuta
  - Lemma di riferimento
  - Genere
  - Numero
  - Persona
  - Tempo
  - Modo
  - Altri



Metodi per la categorizzazione

- Ricerca delle forme all'interno di un lessico, sia generico che specialistico
  - Vedi anche Tokenizer
- Applicazione di strumenti per risolvere problemi specifici
  - Per riconoscere sequenze ricorrenti (date, indirizzi, ecc.)
- Utilizzo di analizzatori morfologici
  - Per riconoscere o formulare ipotesi su parole non riconosciute ma che sono costruite su combinazione di pattern ricorrenti

Considerazione sui metodi

- Consultazione di un dizionario predefinito
  - Efficiente
  - Non riconosce e non può formulare ipotesi sulle forme non presenti
- Metodi specifici
  - Sono efficaci ed efficienti per il problema specifico da risolvere
- Analizzatori morfologici
  - Possono formulare ipotesi categoriali su ogni parola
  - Poco efficienti perché generano troppe ambiguità

Soluzione realistica

- Integrazione delle fasi del componente lessicale
  - Tokenizer + tagger
- Integrazione dei metodi
  - Più metodi per risolvere ciascuna fase

Procedura integrata per il riconoscimento e la classificazione di forme

- Segmenta testo in parole (stringhe di caratteri tra due spazi)
  - Input: testo
  - Output: testo suddiviso in stringhe di caratteri
- Confronta parole con dizionario di forme
  - Input: testo suddiviso in stringhe di caratteri
  - Output: testo arricchito di informazioni:
    - Parole riconosciute con parametri lessicali
    - Parole non trovate marcate come sconosciute
- Verifica con test ad hoc
  - Input: testo suddiviso in stringhe di caratteri
  - Output: testo arricchito di informazioni
    - Parole riconosciute con parametri lessicali
    - Parole non riconosciute marcate come sconosciute
- Applica analizzatore morfologico
  - Input: parole non riconosciute nella fase precedente
  - Output: testo arricchito di informazioni
    - Parole riconosciute con parametri lessicali
    - Parole non riconosciute marcate come sconosciute

Riconoscimento e classificazione utilizzando un dizionario di forme

- Sorgenti di conoscenza
  - dizionario di forme
- Procedura per confrontare il testo segmentato in parole con il dizionario delle forme
  - Confronta le parole e, se trovate, arricchisce il testo con le informazioni lessicali recuperate dal dizionario

Struttura del dizionario delle forme

Lemma	Categoria Grammaticale	Forma	Parametri Morfologici
porto	Sostantivo Maschile	*porti*	Maschile Plurale
porto	Sostantivo Maschile	*porto*	Maschile Singolare
porto	Aggettivo Qualificativo	*porte*	Femminile Plurale
porto	Aggettivo Qualificativo	*porta*	Femminile Singolare
porto	Aggettivo Qualificativo	*porti*	Maschile Plurale
porto	Aggettivo Qualificativo	*porto*	Maschile Singolare
porto	Sostantivo Maschile	*porti*	Maschile Plurale
porto	Sostantivo Maschile	*porto*	Maschile Singolare
porto	Sostantivo Maschile	*porto*	Maschile Mobile

lunedì	lunedì	Sostantivo Maschile Mas.Mob.
25		Non Trovato
maggio	maggio	Sostantivo Maschile Mas.Sing.
u.s.		Non Trovato
la	la	Pronome Personale Femm.Plur.
la	la	Articolo Femm.Sing.
la	la	Sostantivo Maschile Mas.Sing.
banca	banca	Sostantivo Femminile Femm.Sing.
d'	di	Preposizione
italia		Non Trovato
ha	avere	Verbo Trans.Intrans.3 Pers.Sing.Ind.Pres.
abbassato	abbassare	Verbo Trans.Pron.Intrans.Rifl. Mas.Sing.Part.Pass.
abbassato	abbassato	Aggettivo Qualificativo Mas.Sing.
il	il	Pronome Personale Mas.Sing.
il	il	Articolo Mas.Sing.
tasso	tassare	Verbo Trans.Rifl. 1 Pers.Sing.Ind.Pres.
tasso	tasso	Sostantivo Maschile Mas.Sing.
d'	di	Preposizione
interesse	interesse	Sostantivo Maschile Mas.Sing.
di	di	Preposizione
due	due	Numerale Cardinale
due	due	Sostantivo Maschile Mas.Mob.
punti	pungere	Verbo Trans. Mas.Plur.Part.Pass.
punti	pungere	Verbo Trans. Mas.Plur.Part.Pass.
punti	puntare	Verbo Trans.Intrans. 2 Pers.Sing.Ind.Pres.
punti	puntare	Verbo Trans.Intrans. 1 Pers.Sing.Cong.Pres.
punti	puntare	Verbo Trans.Intrans. 2 Pers.Sing.Cong.Pres.
punti	puntare	Verbo Trans.Intrans. 3 Pers.Sing.Cong.Pres.
punti	punto	Aggettivo Qualificativo Mas.Plur.
punti	punto	Aggettivo Indefinito Mas.Plur.
punti	punto	Sostantivo Maschile Mas.Plur.
portandolo		Non Trovato
dall'	da	Preposizione Femm.Plur.
dall'	da	Preposizione Femm.Sing.
dall'	da	Preposizione Mas.Sing.
8		Non trovato
al	a	Preposizione Mas.Sing.
6		Non trovato
%		Non trovato

Limiti della categorizzazione con uso di un dizionario

- Ad ogni forma vengono associate le informazioni grammaticali se trovate
- Forme omografe vengono ricondotte a più lemmi, non risolvendo l'ambiguità
  - Vedi "punti" nell'esempio presentato nella diapositiva precedente
- Alcune forme non vengono riconosciute e non viene formulata alcuna ipotesi
  - Vedi "portandolo" nell'esempio, che viene semplicemente marcato come "non trovato"

Applicazione di strumenti specifici per riconoscere parole non presenti nel dizionario delle forme

- Numeri
  - Numeri romani
  - Ordinali
  - Frazioni
- Alfanumerici
- Iniziali
- Numeri telefonici
- Multiwords
- Indirizzi di posta elettronica e siti web
- Date
- Nomi propri

lunedì 25 maggio u.s.	Data
la	Pronome Personale Femm. Plur.
la	Articolo Femm. Sing.
la	Sostantivo Maschile Singolare
Banca d'Italia	Multiword - Sostantivo Maschile Singolare
ha	Verbo Trans. Intrans. 3ª Pers. Sing. Ind. Pres.
abbassato	Verbo Trans. Pron. Intrans. Rifl. Mas. Sing .Part. Pass.
abbassato	Aggettivo Qualificativo Maschile Singolare
il	Pronome Personale Maschile Singolare
il	Articolo Maschile Singolare
tasso d'interesse	Multiword - Sostantivo Maschile Singolare
di	Preposizione
due	Numerale Cardinale
due	Sostantivo Maschile Mobile
punti	Verbo Trans. Mas. Plur. Part. Pass.
punti	Verbo Trans. Mas. Plur. Part. Pass.
punti	Verbo Trans. Intrans. 2ª Pers. Sing. Ind. Pres.
punti	Verbo Trans. Intrans. 1ª Pers. Sing. Cong. Pres.
punti	Verbo Trans. Intrans. 2ª Pers. Sing. Cong. Pres.
punti	Verbo Trans. Intrans. 3ª Pers. Sing. Cong. Pres.
punti	Aggettivo Qualificativo Maschile Plurale
punti	Aggettivo Indefinito Maschile Plurale
punti	Sostantivo Maschile Plurale
portandolo	Non Trovato
dall'	Preposizione Femminile Plurale
dall'	Preposizione Femminile Singolare
dall'	Preposizione Maschile Singolare
8	Numero Cardinale
al	Preposizione Maschile Singolare
6 %	Numero Percentuale

ancora non classificato

Analizzatore morfologico

(si veda lezioni su morfologia e seminari di De Pascalis (2002) e Utzeri in http://www.di.unipi.it/~cappelli/)

- Strumenti per riconoscere, suffissi, prefissi e composti lessicali
  - Suffissi
    - It. - bellissimo
  - Pronomi personali atoni
    - It. - dirtelo, mangiamocela
    - Sp. - digame
  - Parole composte
    - It. - antigovernativo

Risultati della procedura

- Non si ottiene in pieno lo scopo del componente lessicale
- Alcune parole restano ambigue
- Necessità di introdurre altri strumenti di analisi

Altri strumenti di analisi Analizzatore morfo-sintattico

- Scopo: risolvere l'ambiguità grammaticale di alcuni termini lessicali ambigui (vedi esempio precedente):

punti	pugnere	Verbo Trans. Mas.Plur.Part.Pass.
punti	pungere	Verbo Trans. Mas.Plur.Part.Pass.
punti	puntare	Verbo Trans.Intrans. 2 Pers.Sing.Ind.Pres.
punti	puntare	Verbo Trans.Intrans. 1 Pers.Sing.Cong.Pres.
punti	puntare	Verbo Trans.Intrans. 2 Pers.Sing.Cong.Pres.
punti	puntare	Verbo Trans.Intrans. 3 Pers.Sing.Cong.Pres.
punti	punto	Aggettivo Qualificativo Mas.Plur.
punti	punto	Aggettivo Indefinito Mas.Plur.
punti	punto	Sostantivo Maschile Mas.Plur.

Analizzatore morfosintattico

- Utilizza alcune conoscenze sintattiche
- Analizza il contesto locale della parola da riconoscere, basandosi sulle categorie grammaticali delle parole che precedono o seguono e sulla loro compatibilità sintattica
- Utilizza regole:
  - Sintagmatiche
  - Fonosintattiche

Regole sintagmatiche

- Permettono di escludere alcune combinazioni di categorie
- Ambiguità articolo/pronome
  - Esempio "La notte"
  - L'ambiguità di "la", articolo o pronome, viene risolta in articolo in virtù di una regola sintagmatica che esclude la combinazione pronome + sostantivo

Rappresentazione (quasi)formale di una regola sintagmatica

SE

la forma da analizzare è ambigua,

E l'ambiguità è tra articolo e pronome,

E la forma seguente è un sostantivo non ambiguo,

E concorda con esso per genere e numero;

ALLORA

la categoria della forma da analizzare è articolo.

Regole fonosintattiche

- Permettono di escludere alcune combinazioni di categorie utilizzando informazioni sulla compatibilità fonetico-fonologiche tra parole
- Ambiguità articolo/pronome e sostantivo/verbo
  - Esempio "Lo cambio"
  - Le ambiguità di "lo", articolo o pronome, e di "cambio", sostantivo/verbo, vengono risolte in 'pronome + verbo' in virtù di una regola fonosintattica che esclude la presenza della forma "lo" dell'articolo davanti a sostantivi che non inizino per "z", "s" impura , "x", "ps", "pn", "gn" e "sc" e "i" semiconsonante.

Regole sintagmatiche e fonosintattiche Precedenza

- Per risolvere l'esempio precedente, "lo cambio", prima vengono applicate le regole sintagmatiche che forniscono i due esiti che vengono risolti con la successiva applicazione delle regole fonosintattiche

Ambiguità grammaticale e componente lessicale

- Non sempre si ottiene la risoluzione dell'ambiguità grammaticale
- Alcune parole resteranno grammaticalmente ambigue perché la combinazione delle loro categorie ammette diverse categorizzazioni, tutte sintatticamente compatibili tra loro
  - Es. La vecchia porta la sbarra

Componente lessicale e ambiguità lessicale

- Con gli strumenti presentati fino ad ora, non è possibile risolvere l'ambiguità lessicale tra parole con la stessa categoria, ma con significato diverso
  - Porto - sostantivo maschile
    - **spesa di trasporto**
    - **spazio di mare protetto dove le navi possono sostare in sicurezza**
    - **meta ultima** [Figurato]
    - **rifugio sicuro e tranquillo** [Figurato]
    - **vino portoghese**