

Natural Language Processing

Introduzione

Corso di Linguaggi & Traduttori
Università degli Studi di Bari

Pierpaolo Basile (Ph.D. Student)



Outline

- Natural Language Processing
- Perché / Dove
- Un po' di storia
- NLP layers: dal simbolico al semantico
- Come
- Risultati



Natural Language Processing



- **Natural Language**
 - Ci riferiamo al linguaggio parlato dalle persone
- **Processing**
 - Qualsiasi applicazione che elabora il linguaggio naturale
- **Computer Linguistic**
 - tutto ciò che riguarda linguistica e computer
 - più vicina alla linguistica che al “processing”

3

Natural Language Processing



- L’NLP cerca di dotare il computer di conoscenze linguistiche allo scopo di:
 - progettare programmi e sistemi informatici che assistano l’uomo in “compiti linguistici”
 - **traduzione**
 - **gestione dei documenti e della conoscenza, ecc.**
 - sviluppare sistemi informatici che usano il linguaggio naturale per:
 - **interagire con essere umani in maniera “naturale”**
 - **estrarre automaticamente informazioni da testi o da altri media**
 - **estendere dinamicamente la propria competenza linguistica**

4

Perchè [1/2]



- Molte società spendono tempo e denaro in NLP
 - Yahoo, Google, Microsoft → Information Retrieval
 - Monster.com, HotJobs.com → Information Extraction + Information Retrieval
 - Babelfish → Machine Translation
 - Ask Jeeves → Question Answering

5

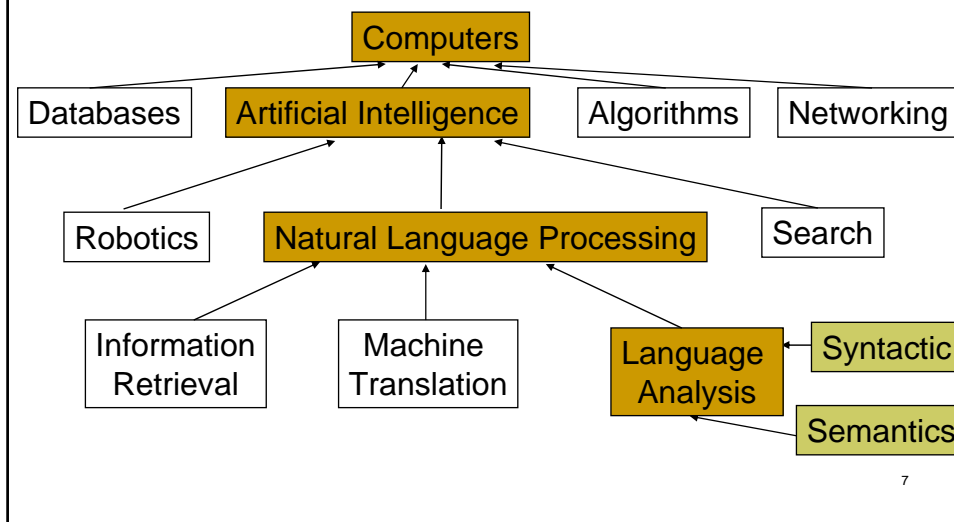
Perchè [2/2]



- Text classification
- Index & search
- Traduzione automatica
- Comprensione del linguaggio parlato
- Information extraction
- Automatic summarization
- Question answering
- Knowledge acquisition
- Text generations

6

Dove



7

Un po' di storia



Prime applicazioni dei computer ai testi letterari (1950)

Corpora elettronici (metà '60)

Nascita della statistica linguistica ('60-'80)

corpora / corpus
collezione strutturata di documenti utilizzata per:
-analisi statistiche
-calcolo delle occorrenze
-validare

8

NLP layers [1/2]



- L'elaborazione del linguaggio naturale avviene a diversi livelli
 - C'è una forte dipendenza tra i vari livelli
 - Si va da un livello puramente simbolico (fonemi, lettere) ad un livello più semantico (significati)



9

NLP layers [2/2]



- articolare e decodificare i suoni di una lingua
 - **suoni e lettere**
- conoscere le parole di una lingua, la loro struttura e la loro organizzazione
 - **lessico e morfologia**
- comporre le parole in costituenti complessi
 - **sintassi**
- assegnare significati alle espressioni linguistiche semplici e complesse
 - **semantica**
- usare le frasi nei contesti, situazioni e modi appropriati agli scopi comunicativi
 - **pragmatica**

10

Lessico & Morfologia



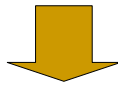
- Individuare le parole che compongono il linguaggio
- Individuazione dei lessemi e dei lemmi
 - lemma: ciascuna delle voci cui sono dedicate le singole definizioni di un dizionario
 - lessema: unità minima dotata di significato
- Analisi morfologica della parola
 - Plurare/singolare, modo e tempo verbale, ...

11

Lessico & Morfologia



The dog ate my homework.



- The:** articolo determinativo
Dog: nome singolare
verbo (to dog) *pedinare*
Ate: verbo tempo=passato infinito=to eat
My: aggettivo possessivo
esclamazione (utilizzato nelle forme esclamative)
Homework: nome singolare

12



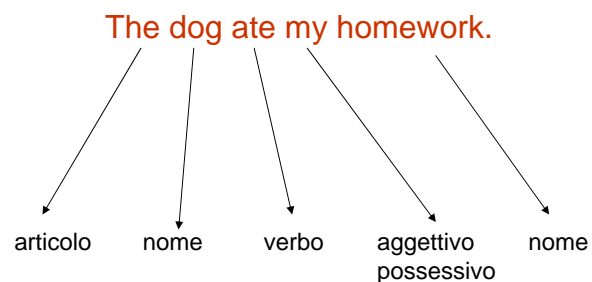
Sintassi

- Individuazione delle parti del discorso (Part-of-speech)
 - Verbi, nomi, aggettivi, avverbi, preposizioni, pronomi
- Identificare gruppi di parole
 - Hot dog, look for
- Shallow parsing: identificare le parti elementari (parte nominale e parte verbale)
- Full parsing: derivazione dell'albero sintattico completo

13



Part-of-speech

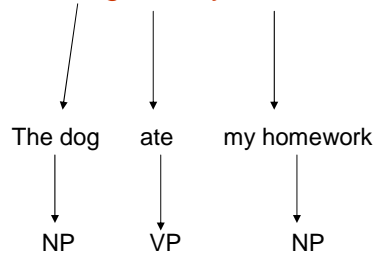


14

Shallow parsing



The dog ate my homework.

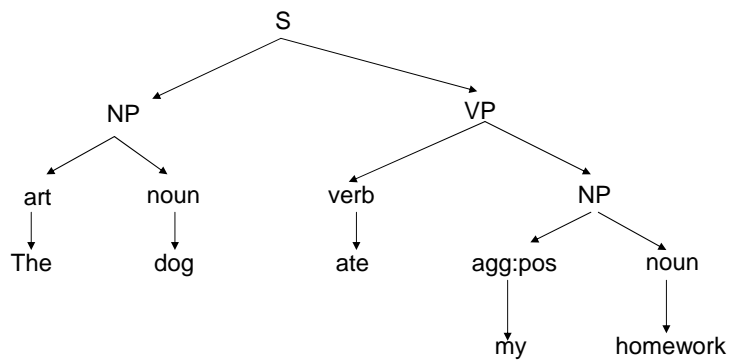


15

Full parsing



The dog ate my homework.



16

Entity recognition [1/2]



- Subtask dell'Information Extraction
- **Individuazione** di parole o gruppi di parole che possono individuare **entità**
- **Classificazione** dei gruppi individuati in **categorie**:
 - Persone
 - Luoghi geografici
 - Eventi
 - Espressioni temporali
 - ...

17

Entity recognition [2/2]



PARIGI - Cadono gli dei. Gli All Blacks, i mostri sacri, escono di scena ai mondiali già nei quarti: non era mai successo, in sei edizioni. È la Francia ad abbattere i mostri sacri, con la forza del carattere, nel match giocato in trasferta, al Millennium di Cardiff, per esigenze geopolitiche. Erano i grandi favoriti neozelandesi, i dominatori della scena da anni.

organization

location

location

18



Come?

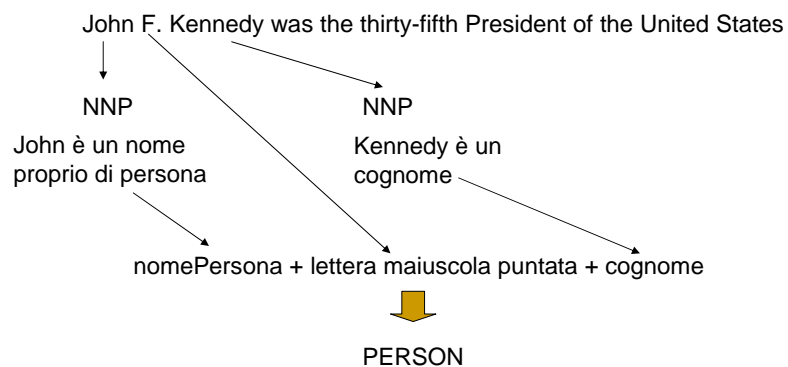
- Regole (primi approcci)
- Machine Learning (approccio statistico)
 - Alberi di decisione
 - Hidden Markov Model
 - Support Vector Machine (SVM)
 - Latent Semantic Analysis

19



Come - Regole

- Un esempio nell'Entity Recognition



20

Come - Machine Learning [1/2]



- Fornire un numero consistente di esempi
 - Individuare le **features**
- Scegliere un algoritmo di **learning**
 - Apprendere dagli esempi
 - Classificare nuovi esempi

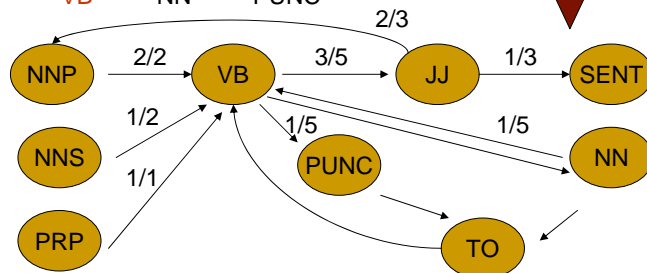
21

Come - Machine Learning [2/2]



PRP\$	NNS	VB	JJ	SENT
NNP	VB	JJ	NNP	NNP
NNP	VB	JJ	NNP	NNP
PRP	VB	PUNC	TO	NN
NNS	TO	VB	NN	PUNC

ESEMPI
MODELLO



22

Latent Semantic Analysis [1/2]



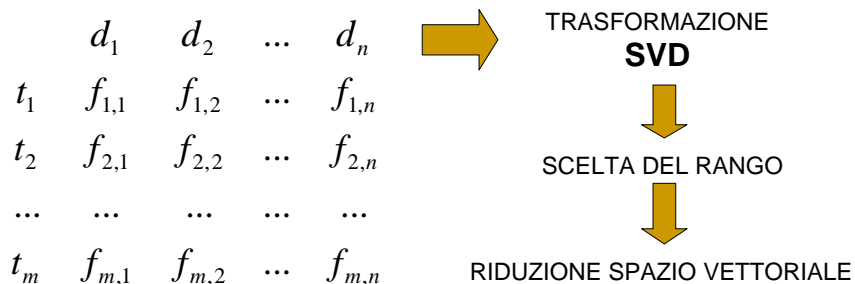
- Creazione di uno spazio vettoriale delle features
- Riduzione dello spazio vettoriale attraverso la scomposizione SVD della matrice
- Analisi della matrice nello spazio ridotto

23

Latent Semantic Analysis [2/3]



Es.: matrice terminixdocumenti



Se due termini risultano **simili** vengono **rappresentati** da un'unica riga nello **spazio ridotto**.

24

Latent Semantic Analysis [3/3]



- E' un processo matematico efficace
- Ha un'alta complessità in tempo e spazio soprattutto perché la matrice iniziale è sparsa
- Rende difficile l'update di documenti
- Ha una logica nascosta frutto di un processo matematico: non fornisce spiegazioni del perché 2 (o più) termini risultano essere simili

25

Valutazione



- Corpus di valutazione
 - per tecniche di ML è necessario un corpus di addestramento dal quale estrarre le features differente da quello di valutazione
- Precisione
 - $P = \frac{\text{\#istanze_correte}}{\text{\#istanze_classificate}}$
- Richiamo
 - $R = \frac{\text{\#istanze_correte}}{\text{\#tot_istanze}}$

26

Risultati

- POS-tagging -> P = 90%
- Parsing -> P = 80%
- Entity Recognition -> P = [80%-90%]
- Semantic -> ???



27

Riferimenti

- *Daniel Jurafsky & James H. Martin* – “Speech and Language Processing” – An introduction to Natural Language Processing, Computational Linguistic and Speech Recognition
- *C. Manning and H. Schutze* - Foundations of Statistical Natural Language Processing



28