1

Analisi Sintattica e Machine Learning

di Stefano Iardella

Università di Pisa – Facoltà di Informatica
Corso di Elabrazione del Linguaggio Naturale
Prof. Amedeo Cappelli

Anno Accademico 2005/2006

2

Introduzione

- Presentare brevemente la fase di analisi sintattica
- Fornire una panoramica delle modalità con le quali il machine learning può essere utilizzato per l'analisi sintattica
- Introdurre i principali approcci al machine learning

3

Sommario

- L'analisi sintattica
- Perché il machine learning ?
- NLP e Machine Learning
- Principali approcci di tipo Machine Learning

4

L'analisi sintattica

- Una fase essenziale del processo di elaborazione del linguaggio naturale
- Fornisce la rappresentazione strutturale di un enunciato in funzione delle entità linguistiche che lo compongono
- Ha un forte impatto sulla successiva fase di analisi semantica

5

L'analisi sintattica

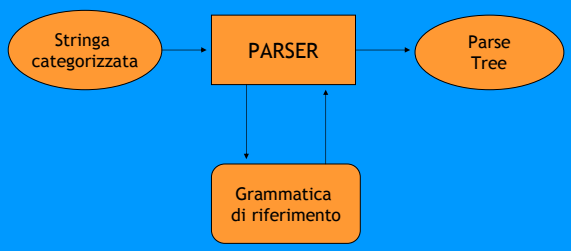
- Componenti essenziali:

  - Parser
  - Grammatica di riferimento

- Il parser effettua l'analisi in base alle regole presenti nella grammatica

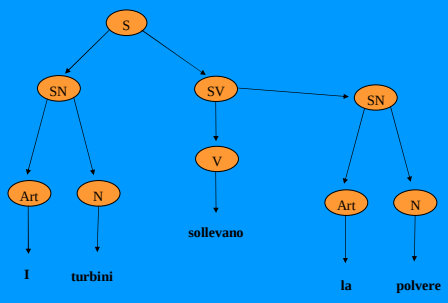
6

L'analisi sintattica (Schema)



7

L'analisi sintattica (Esempio)



8

L'analisi sintattica

- Full Parsing:
  analisi approfondita
- Shallow Parsing:
  analisi parziale, si considera un sottoinsieme delle entità linguistiche
- Chunk Parsing:
  si considerano solo entità terminali (Non ricorsive)

9

L'analisi sintattica

- Shallow parsing: maggiore robustezza, minore complessità computazionale (Lineare o polinomiale)
- Full parsing: maggiore accuratezza, maggiore complessità (Esponenziale)
- Possibile integrazione tra i due metodi

10

Perché il machine learning ?

- Difficoltà nella codifica manuale di grammatiche
- Difficoltà nella gestione di "eccezioni"
- Problema dell'ambiguità strutturale
- Maggiore copertura ➔ maggiore ambiguità
- Occorre qualcosa di più flessibile

11

Perché il machine learning ?

- Definizioni:

"A learning machine, broadly defined, is any device whose actions are influenced by past experiences" (Nilsson 1965)

"Modification of a behavioural tendency by expertise" (Webster 1984)

12

Perché il machine learning ?

- Definizioni:

"The field of machine learning is concerned with the question of how to construct computer programs that automatically learn with experience" (Mitchell, 1997)

"Changing the behaviuor in a way that makes it possibile to perfom better in the future" (Witten and Frank 2000)

13

Perché il machine learning ?

- Una "sotto-disciplina" dell'Intelligenza Artificiale
- Impiegato con frequenza in ambito NLP a partire dai primi anni '90
- Lo scopo è apprendere nuova conoscenza a partire da un insieme possibilmente ridotto di dati iniziali

14

Perché il machine learning ?

- Inizialmente attraverso metodi di tipo statistico-matematico
- Ai livelli "bassi" dell'elaborazione (Tokenization, tagging, etc.)
- Buoni risultati
- Successivamente anche metodi di tipo simbolico …

15

NLP e Machine Learning

- Supervised Learning:
  estrarre conoscenza a partire da una collezione di testo annotato
- Apprendere *regole* o *modelli* attraverso meccanismi di generalizzazione
- Dipende fortemente dalla quantità e varietà di enunciati contenuti nella collezione

16

Perché il machine learning ?

- Importanti collezioni di testo annotato:
  le Treebank
- Annotate ad ogni livello (A noi interessa quello sintattico)
- Coppie del tipo
  <s, analisiCorretta(s)>

17

Perché il machine learning ?

- Alcune treebank:

  - Penn Tree Bank: annotazione di articoli estratti per lo più dal Wall Street Journal
  - British National Corpus: collezione di 100 milioni di parole. Enunciati provenienti da varie sorgenti.

18

Perché il machine learning ?

- Vengono utilizzate sia per la fase di training che per la fase di valutazione (Selezionando degli insiemi disgiunti…)
- Treebank differenti, grammatiche differenti
- Esistenti per un limitato numero di lingue

19

Perché il machine learning ?

- Unsupervised learning:
  la fase di training avviene su testo "puro" mediante meccanismi di generalizzazione
- Possibili vantaggi:
  - maggiore disponibilità di dati
  - diminuzione del lavoro necessario per l'annotazione

20

Perché il machine learning ?

- Il learning supervisionato è ancora quello maggiormente utilizzato
- Fino ad ora ha fornito risultati migliori nonostante alcuni svantaggi:

  - forte dipendenza dal training set
  - problemi con l'analisi di dati "sparsi"

21

Perché il machine learning ?

- Active learning:
  forma di learning supervisionato nella quale il sistema partecipa alla selezione dei testi da annotare
- Diminuzione del lavoro manuale di annotazione
- Annotazione di testo ritenuto maggiormente "informativo"

22

Principali approcci di tipo ML

- 3 differenti approcci per applicare tecniche di ML:

  - approccio tradizionale o simbolico
  - approccio statistico o stocastico
  - approccio connessionista o subsimbolico

23

Principali approcci di tipo ML

- Approccio simbolico:
  apprendimento ed estrazione di regole in funzione di aspetti strutturali dei dati
- Molto noto nell'ambito dell'IA
- Recentemente "riscoperto" e sfruttato anche in ambito NLP

24

Principali approcci di tipo ML

- Alcune tecniche di tipo simbolico:

  - Decision Trees
  - Inductive Logic Programming
  - Memory-Based Learning
  - Transformation-Based Learning
  - (Clustering)

25

Principali approcci di tipo ML

- Decision trees:
  uno dei metodi di tipo simbolico maggiormente noto nell'ambito dell'intelligenza artificiale
- A partire dal training set di riferimento si costruisce un albero di decisione

26

Principali approcci di tipo ML

- L'albero servirà in fase di analisi per ottenere il parse tree
- Ogni nodo dell'albero corrisponde ad un test da effettuare su una caratteristica o un attributo opportunamente selezionati durante la costruzione del decision tree

27

Principali approcci di tipo ML

- Gli archi uscenti da ciascun nodo rappresentano le possibili risposte al test
- La fase di learning consiste nella selezione del test da effettuare a ciascun nodo

28

Principali approcci di tipo ML

- L'obiettivo è selezionare per i test quelle caratteristiche che permettono di ottenere il maggior grado di informazione possibile
- Una volta selezionata una caratteristica il training set viene suddiviso in sottoinsiemi in riferimento a tale caratteristica

29

Principali approcci di tipo ML

- Il meccanismo viene riapplicato ad ogni sottoinsieme generato
- In fase di analisi si giunge ad ottenere il parse tree traversando i nodi dell'albero di decisione ed effettuando i test sull'enunciato da analizzare

30

Principali approcci di tipo ML

- Costruzione del decision tree:

1. Se T contiene uno o più esempi tutti appartenenti alla stessa classe Cj allora l'albero di decisione per T è una foglia etichettata con la categoria Cj.
2. Se T contiene classi differenti allora:
   - **scegli** una caratteristica e partiziona T in sottoinsiemi aventi il medesimo valore per tale caratteristica. L'albero di decisione consiste in un nodo avente il nome della caratteristica/attributo scelto, dal quale partono tanti archi quanti sono i valori possibili per l'attributo. Ciascun arco porta al sottoinsieme di T corrispondente.
   - applica la medesima procedura a tutti i sottoinsiemi di T.

31

Principali approcci di tipo ML

- Inductive Logic Programming: tecnica supervisionata basata sulla programmazione logica
- Fatti e regole sono espressi secondo la logica del prim'ordine (Con clausole Horn)
- Grande espressività grazie al linguaggio utilizzato per la rappresentazione delle regole

32

Principali approcci di tipo ML

- In fase di learning si generano regole che soddisfino il maggior numero possibile degli esempi positivi contenuti nel training set
- La fase di analisi consiste nell'applicazione del principio di risoluzione da parte del programma logico

33

Principali approcci di tipo ML

Algoritmo generico per l'apprendimento:

1. Seleziona da T un esempio **e** da generalizzare; se nessun esempio è presente termina la procedura.
2. Definisci una spazio H di ricerca delle ipotesi (Regole) che soddisfano l'esempio **e**.
3. Cerca in H l'ipotesi h che massimizza la funzione obiettivo F.
4. Rimuovi da T gli esempi che sono soddisfatti dall'ipotesi h scelta.
5. Ritorna al punto 1.

(F dipende dal numero di esempi in T soddisfatti dall'ipotesi h)

34

Principali approcci di tipo ML

- Memory Based Learning:
  si distingue dai metodi simbolici tradizionali in quanto non estrae regole bensì effettua una "classificazione" basandosi su un criterio di *similitudine*
- Particolarmente utilizzato per compiti di disambiguazione

35

Principali approcci di tipo ML

- Y = l'istanza da analizzare
- M = memoria contenente m esempi del tipo:
  <Xi, Li> per i=1,…,m
  dove Xi è un esempio fornito in fase di training ed Li è l'etichetta ad esso associata

36

Principali approcci di tipo ML

- Y viene classificata in funzione degli m esempi secondo una metrica di *similitudine* oppure di *distanza*
- Si calcola l'insieme {d(Xi,Y)} per i=1,…,m dove d è la metrica scelta
- Attraverso una certa funzione F si determina quale sia l'etichetta Li da assegnare ad Y

37

Principali approcci di tipo ML

- Ad esempio F potrebbe essere così definita:

  F({d(Xi,Y)} = Li per i=1,…,m

- i = argmax{d(Xi,Y)} per i=1,…,m

oppure

- i = argmin{d(Xi,Y)} per i=1,...,m

38

Principali approcci di tipo ML

- Transformation-Based Learning: effettua un apprendimento supervisionato di tipo error-driven
- Basato su un meccanismo di trasformazioni successive mediante le quali il modello di riferimento creato a partire dal training set viene raffinato di volta in volta

39

Principali approcci di tipo ML

- Il modello viene modificato in base agli errori commessi ed ogni trasformazione mira ad eliminare il maggior numero di errori presenti nel modello
- Si prosegue finché il numero di errori non scende sotto una certa soglia prestabilita

40

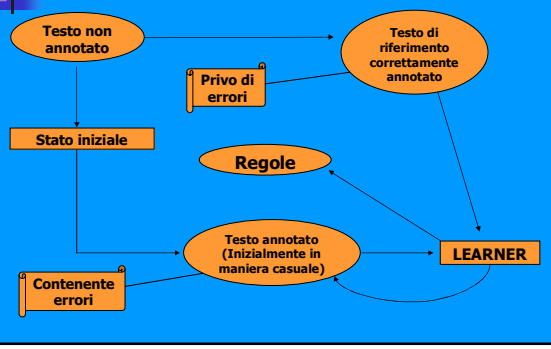
Principali approcci di tipo ML

- Algoritmo generale d'apprendimento:

1. Genera tutte le regole che correggono almeno un errore nel modello.
2. Per ogni regola:
(a) Applicala ad una copia dello stato più recente del training set
(b) Calcola lo score secondo la funzione obiettivo.
3. Seleziona la regola che ha ottenuto lo score migliore.
4. Aggiorna il modello aggiungendo tale regola.
5. Fermati se lo score raggiunto è inferiore ad una soglia prestabilita T;
altrimenti riprendi a partire dal punto 1.

41

Principali approcci di tipo ML



42

Principali approcci di tipo ML

- Si confronta l'annotazione di riferimento con l'annotazione eseguita dal sistema
- Si aggiornano le regole in modo da aumentare il grado di concordanza con l'annotazione di riferimento
- Funzione obiettivo: minimizzare gli errori

43

Principali approcci di tipo ML

- (Clustering):
  tecnica ancora poco sfruttata nell'ambito dell'analisi sintattica
- Degna di citazione:
  una delle poche ad effettuare un learning completamente non supervisionato

44

Principali approcci di tipo ML

- I dati contenuti nel training set vengono suddivisi in classi (Clusters) in funzione di certe proprietà
- Ancora una volta a guidare l'apprendimento è una metrica di similitudine

45

Principali approcci di tipo ML

- Approccio statistico:
  si considerano aspetti di tipo statistico anziché caratteristiche strutturali
- Maggiore grado di robustezza rispetto ai metodi di tipo simbolico
- Particolarmente utilizzato per shallow e chunk parsing

46

Principali approcci di tipo ML

- A differenza dei metodi simbolici la fase di disambiguazione è incorporata nella tecnica stessa di analisi
- Essa viene effettuata automaticamente in funzione di parametri statistici

47

Principali approcci di tipo ML

- In fase di analisi il parse tree selezionato è quello avente la maggiore probabilità di corrispondere all'analisi "corretta"
- Varie tecniche disponibili
- Si distinguono essenzialmente per il modo in cui assegnano ed aggiornano le statistiche

48

Principali approcci di tipo ML

- Due importanti tecniche (Citate soltanto, senza entrare nel merito):

  - Hidden Markov Models

  - Maximum Entropy

49

Principali approcci di tipo ML

- Hidden Markov Models (HMM):

  tecnica ampiamente utilizzata in fase di tagging e successivamente introdotta anche in fase di analisi sintattica, soprattutto per parsing parziale

50

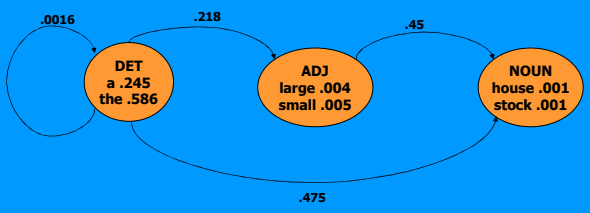
Principali approcci di tipo ML

- Si crea un automa a stati finiti a più "livelli" (Stati di input, stati di output, stati nascosti) che tenga conto di frequenze estratte dai dati iniziali del training set
- L'analisi viene effettuata attraverso una complessa "navigazione" tra gli stati del modello costruito

51

Principali approcci di tipo ML

- Un esempio di HMM per il tagging:



52

Principali approcci di tipo ML

- Notiamo ad esempio la probabilità che un articolo (DET) sia 'a' (24.5%) oppure 'the' (58.6%) e che dopo un articolo vi sia un nome (NOUN) (47.5%)
- Da notare la bassa probabilità che l'etichetta DET possa seguire un'altra etichetta DET
- Questo perchè…

53

Principali approcci di tipo ML

- Maximum Entropy:
  la creazione del modello probabilistico cui fare riferimento viene effettuata selezionandolo da una classe di ipotesi fatte a partire dai dati a disposizione per il training

54

Principali approcci di tipo ML

- In funzione dell'osservazione di nuovi dati il modello viene modificato in modo da essere consistente con la conoscenza ulteriormente acquisita

55

Principali approcci di tipo ML

- Spesso approcci di tipo simbolico e statistico vengono integrati
- In tal modo è possibile mantenere la "precisione" dell'approccio simbolico e avvalersi della robustezza e della capacità di disambiguazione dell'approccio statistico

56

Principali approcci di tipo ML

- Approccio di tipo connessionista:
  cerca di avvicinarsi a quello che è il "ragionamento" proprio degli umani
- Per certi aspetti similare all'approccio di tipo statistico in quanto considera caratteristiche di tipo probabilistico

57

Principali approcci di tipo ML

- Si differenzia da esso in quanto manipola formule logiche mediante la cui trasformazione implementa il meccanismo di inferenza

58

Principali approcci di tipo ML

- Una tecnica di tipo connessionista:
  - Reti Neurali (NNs)
- Una rete di nodi rappresentanti concetti o entità
- Gli archi sono relazioni di dipendenza tra concetti

59

Principali approcci di tipo ML

- A ciascun arco è associato un peso che rappresenta un "pezzo" di conoscenza
- Il processo di learning consiste nell'aggiornare opportunamente il peso degli archi e le connessioni tra i nodi della rete in modo da migliorare il "ragionamento" della rete

60

Conclusioni

Alcuni possibili sviluppi:

- Aumentare il numero di tecniche capaci di implementare un learning non supervisionato
- Migliorare la capacità di generalizzazione delle diverse tecniche
- Integrare in maniera efficiente gli approcci considerati

61

Bibliografia

[1] L. Marquez (2000) "Machine Learning and Natural Language Processing"

[2] A. Roberts (2003) "Machine Learning in Natural Language Processing"

[3] Liddy "Natural Language Processing" in *Enciclopedia of Library and Information Science, second edition*

[4] X. R. Hu, E. Atwell "A survey of machine approaches to analysis of large corpora"

[5] J. Hammerton, M. Osborne, S. Armstrong, W. Daelemans "Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing"

[6] C. Cardie, R.J. Mooney "Guest Editor's Introduction: Machine Learning and Natural Language"

62

Bibliografia

[7] C.A. Thompson, R.J. Mooney, L.R. Tang (1997) "Learning to Parse Natural Language Database Queries into Logical Form" in *Proceedings of the ICML-97 Workshop on Automata Induction, Grammatical Inference and Language Acquisition, Nashville, TN, July 1997*

[8] S. Lappin (2005) "Machine Learning and the Cognitive Basis of Natural Language"

[9] E. Charniak (1997) "Statistical Techniques for Natural Language Parsing"

[10] W. Daelemans "Machine Learning of Natural Language"

[11] L.R. Rabiner "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition"

[12] J.M. Zelle "Using Inductive Logic Programmino to Automate the Construction of Natural Language Parsers"

63

Bibliografia

[13] A. McCallum, D. Freitag, F.Pereira "Maximum Entropy Markov Models for Information Extraction and Segmentation"

[14] D. Roth (1999) "Memory Based Learning in NLP"

[15] E. Brill "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging"

[16] P. Sebillot "Symbolic Machine Learning: A Different Answer to the Problem of the Acquisition of Lexical Knowledge from Corpora"

[17] M. Dickinson "Natural Language Processing"

[18] N. Calzolari, A. Lenci "Linguistica Computazionale – Strumenti e Risorse per il trattamento automatico della lingua" in *Mondo Digitale n. 2, Giugno 2004*

[19] C. Siefkes "Transformation-Based Learning"

64

Bibliografia

[20] E. Charniak (1997) "Statistical Techniques for Natural Language Parsing"

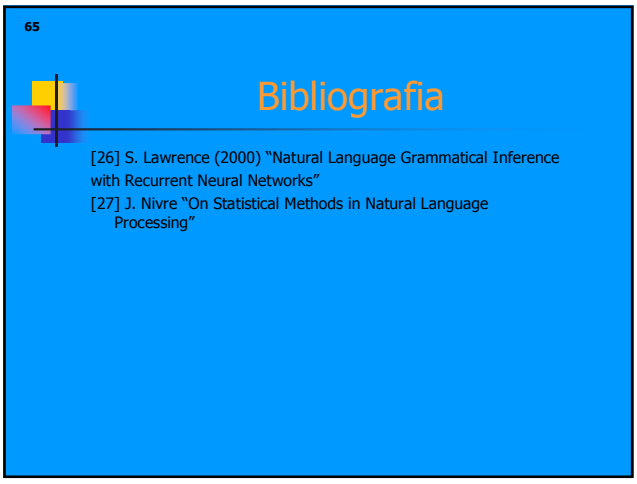
[21] R. J. Mooney (2003) "Oxford Handbook of Computational Linguistics", *Oxford University Press*, cap. 20, pp. 376-394

[22] J.M. Balfourier, P. Blache, T. Van Rullen "From Shallow to Deep Parsing Using Constraint Satisfaction"

[23] A. Roberts (2001) "Automatic Acquisition of Word Classification Using Distribution Analysis of Content Words with Respect to Function Words"

[24] C.A. Thompson, M.E. Califf, R.J. Mooney (1999) "Active Learning for Natural Language Parsing and Information Extraction"

[25] J.M. Zelle, R.J. Mooney (1995) "A Comparison of Two Methods Employing Inductive Logic Programming for Corpus-based Parser Construction"


65
Bibliografia
[26] S. Lawrence (2000) "Natural Language Grammatical Inference with Recurrent Neural Networks"
[27] J. Nivre "On Statistical Methods in Natural Language Processing"