UNIVERSITÀ DEGLI STUDI DI ROMA TRE
FACOLTÀ DI SCIENZE M.F.N.

Tesi di Laurea in Matematica

di

Giulio Zabeo

Tecniche di Voting e di Co-Training per la disambiguazione semantica di parole in testi liberi

Relatore

Prof. Marco Liverani

Correlatore

Prof. Roberto Basili

Classificazione AMS : 68T50, 68Q32, 91E40.
Parole Chiave : Natural Language Processing, Computational Learning Theory.

ANNO ACCADEMICO 2002 - 2003

OTTOBRE 2003

Un Approccio al *Word Sense Disambiguation*

Durante gli anni della guerra fredda, il governo americano investì molte risorse per tentare di trovare una metodologia che permettesse a dei programmi di tradurre dal russo all'inglese. Alla fine del progetto arrivarono però alla conclusione che non era possibile fare ciò. Verso la fine degli '60 centri di ricerca sull'elaborazione del linguaggio ripresero i loro lavori, e li rianalizzarono senza soffermarsi sui pessimi risultati che ottenuti.
Una delle prime applicazioni fu la creazione di un programma di riconoscimento vocale che interpretasse gli ordini di un giocatore di scacchi, per esempio: *"Sposta il re in E4"*. Poco dopo furono ideati programmi in grado di estrapolare, da articoli di giornale, informazioni tipo: data e luogo di un disastro, numero di vittime, modalità, ecc. Tuttavia i risultati erano limitati a piccoli casi particolari. Risultò infatti subito chiaro che si riuscivano ad ottenere buone prestazioni solo quando il dominio era limitato e specifico. L'interpretazione del linguaggio naturale si mostrò da subito un problema di difficile soluzione.
In tempi moderni, con l'avvento di calcolatori sempre più potenti e l'esapnsione continua del *web*, si rendono sempre più necessari strumenti di analisi automatizzate per diversi scopi: la traduzione di testi, l'*Information Retrieval (IR)* e l'*Information Extraction (IE)*.

In tutti questi campi l'informazione che si ottiene dipende fortemente dal significato che le parole assumono all'interno delle frasi, o più in generale, all'interno di un dominio. Tale problema viene trattato nell'elaborazione dei linguaggi naturali, studiando come si possa determinare il significato che una parola assume nei vari contesti.
Nel 1991 Hanks [33] cercò di capire quali fossero gli strumenti che un lessicografo poteva utilizzare per determinare i sensi associabili alle parole, ed assertò che esistevano tre sorgenti di informazioni utili:

- un **Corpus** di testi, scelto in modo da rappresentare un campione di documenti ben formati ed omogenei;
- la propria intuizione e conoscenza della lingua;
- l'uso e/o la rivisitazione di tecniche usate precedentemente da altri, così come dizionari e grammatiche.

Nella letteratura si possono trovare diversi lavori ed articoli che si pongono il problema della determinazione delle strutture, dei meccanismi e delle regole alla base del modo con cui il nostro cervello elabora il linguaggio naturale.
Senza entrare nei meriti filosofici o biologici, gli approcci che si possono trovare in letteratura possono essere elencati nel seguente modo:

- *funzionale*, il significato è una mappatura tra le descrizioni logiche e i mondi in cui queste descrizioni possono essere vere o false;
- *connessionista*, il significato viene rappresentato tramite strutture come le reti neurali che si rifanno a quelle del cervello umano;
- basato sulla *verifica condizionale*, il significato di una frase è la condizione (espressa in modo formale) che la verifica;

- basato su *relazioni deduttive*, il significato è ciò che può implicare o essere implicato.

L'uso di questi approcci non sempre viene condiviso da tutti. I Ricercatori, avendo le proprie idee, cercano di modellare il problema secondo le loro conoscenze. Qualsiasi sia la scelta fatta, il problema principale è : come rappresentare i "concetti"? come poter indurre la loro conoscenza?
Il problema della conoscenza e la rappresentazione dei concetti, viene affrontato in modi differenti in letteratura. La corrente attuale porta all'integrazione, negli approcci statistici, di più sorgenti di informazione linguistica. Questo perchè il valore fornito da un modello statistico è limitato dall'ambito su cui il fenomeno viene studiato. L'uso di più informazioni dovrebbe quindi limitare le eventuali carenze.
Rifacendosi a precedenti lavori, questa tesi si propone di determinare i sensi delle parole in uno specifico dominio applicativo, ovvero su un'estesa collezione tematica di testi, apprendendo in modo *unsupervised*, cioè senza il bisogno di avere materiale annotato da un utente umano. La scelta di un approccio non supervisionato deriva dal presupposto che non sempre dati etichettati sono disponibili, sia perchè richiedono investimenti di denaro e tempo per un'annotatura, sia perchè non è sempre possibile averli, come nel caso di neologismi, forme stringate, forme gergali, ecc.
Il presupposto principale di questo lavoro è che la sintassi e la semantica siano in qualche modo correlate: data una relazione sintattica, un nome tende stabilmente ad assumere un sottoinsieme dei suoi sense nei contesti in cui compare con tale relazione.
Ci sarà quindi bisogno di uno strumento che permetta di determinare quando e quanto siano *simili* due nomi, indispensabile per cercare comportamenti affini tra parole ricoprenti stessi ruoli sintattici. Tale metrica dovrà però poggiarsi su una base di conoscenza lessico-semantica che dia qualche indicazione sulle caratteristiche delle parole che si stanno analizzando. Come in molti

altri studi, la base di conoscenza è disponibile tramite l'utilizzo di dizionari elettronici come *WordNet* ([20], [21] e [22]).
Questo dizionario ha una struttura gerarchica con la quale raggruppa i vari sensi delle parole in un grafo ad ereditarietà multipla ed aciclico. Ci sono diverse relazioni all'interno di questa gerarchia, ma le più significative sono quella di *iponimia* ed *iperonimia*, che consentono la navigazione, all'interno di questa struttura, tra i vari nodi rappresentati dai sensi, passando da concetti più specifici a concetti più generali.
In questa struttura gerarchica si possono definire diverse metriche per la valutazione della *similarità semantica* tra parole. Quelli di maggior interesse in letteratura sono dovuti a: *Agirre e Rigau* in [6] e [7], e *Resnik* in [25].
La stima di similarità che è possibile stabilire tra parole w in contesti grammaticali r identici, consente in tali contesti di privilegiare alcuni, e non tutti, tra i sensi delle parole: un senso fornirà un'interpretazione tanto più utile quanto più numerose sono le parole w che sono simili ad esso. Tale criterio consente di quantificare l'appropiatezza di un senso l per una o più parole w che condividono la proprietà di apparire nel corpus in un contesto r.
Ciò che ci si propone di apprendere con il modello presentato in questa tesi, è quindi la distribuzione di probabilità sui sensi l di una parola w, data la relazione sintattica r con cui compare nel contesto:

$$\mathrm{P}(l|w,r)$$

dove l è uno dei sensi della parola w che ricorre nella relazione sintattica r.
Calcolando tali probabilità attraverso tutte le relazioni r e tutte le parole w, possono essere quindi ricavate due ulteriori distribuzioni:

- lessicale, $\mathrm{P}(l|w)$
- contestuale, $\mathrm{P}(l|r)$

Tali informazioni forniscono quindi un metodo di inferenza probabilistica in frasi individuali, sia rispetto a contesti nuovi (r mai osservate) per w, sia per parole sconosciute (w mai incontrate).
La dimensione del problema e l'incidenza di fenomeni rari (*data sparsity*) richiedono in tale ambito una sperimentazione sistematica su vasta scala, che è quindi stata progettata per la validazione empirica dell'intera metodologia. Il modello ed i risultati di tali sperimentazioni saranno quindi presentati nei capitoli successivi in questa tesi.
Alla luce dei dati ottenuti nella fase di sperimentazione del sistema, le prestazioni del modello proposto sono globalmente soddisfacenti. Per dare comunque un giudizio più accurato si dovrebbero ripetere gli esperimenti su corpus più omogenei. Infatti il *British National Corpus* contiene un insieme di documenti (cronaca) che trattano ogni genere di concetto, ed in una tale condizione è difficile, se non quasi impossibile, privilegiare sensi a discapito di altri in maniera molto significativa. Deve quindi rimanere un certo margine di incertezza per permettere al sistema di esprimersi su un così vasto campo di concetti.
Rimane comunque un dato fondamentale a vantaggio di questa tecnica: ciò che viene costruito, alla fine dell'analisi, è un dizionario adattato al corpus, con la caratteristica molto rilevante di associare a parole e sensi una distribuzione di probabilità semantico-lessicale ($\mathrm{P}(l|w)$) ed una sintattica ($\mathrm{P}(l|r)$).
Una ulteriore limitazione degli esperimenti e misurazioni effettuati è la mappatura tra i dizionari WordNet e Longman. Dovendo infatti confrontare i dati con un corpus etichettato secondo LDOCE (Longman Dictionary Of Contemporary English), c'è sicuramente perdita di informazione, visto che non esiste un mapping diretto tra i due, e che LDOCE non ha la stessa granularità di sensi rispetto a WordNet. In molti casi le etichette prodotte dal sistema proposto nella tesi, specialmente se molto specifiche, rischiano una mappatura non precisa in nell'altro dizionario.

Da un'analisi molto superficiale dei dati è emerso che il sistema piuttosto sistematicamente non riesce ad etichettare correttamente alcune specifiche categorie di Longman, a causa della differente costruzione delle due risorse lessicali. Se si riuscisse a produrre un mapping ideale priovo di errore le misure di prestazioe sarebbero quasi certamente molto piu'significative.
Una estensione del sistema è la sostituzione del modulo che stima le distribuzioni di probabilità mantenendo l'architettura qui descritta. Questo consentirebbe di valutare altre metriche di similitudine (non più la *Conceptual Density*), come ad esempio, la funzione suggerita e sperimentata da *Resnink* in [25]. È possibile anche utilizzare una o più combinazioni di queste metriche, magari cambiando la definizione di CD introducendo eventuali parametri. Dalla combinazione di metriche differenti si potrebbero compensare i limiti di una con i pregi dell'altra.
Questo modello è stato oggetto di studio, assieme ad altre tecniche proposte per il *Word Sense Disambiguation*, nel WorkShop tenutosi alla John Hopkins University di Baltimora nel luglio - agosto 2003. Tra gli scopi di questa ricerca era la analisi di tecniche per il tagging semantico debolmente supervisionato e la misura della sua utilità per modelli probabilistici basati su classi semantiche piuttosto che su probabilità puramente lessicalizzate. Gli strumenti realizzati in questa tesi e i risultati, in termini di prestazioni, sugggeriscono questa come una strada praticabile in termini di accuratezza a copertura. Ne stabiliscono inoltre prospettive originali nei termini di un addestramento non supervisionato che ruslta quindi più economico, più efficiente e portabile. Lo studio di nuovi domini e collezioni testuali e la sperimentazione dei parametri diversi della corrente implemetazione saranno quindi temi critici per il definitivo successo di questa ricerca.

Bibliografia

[1] Basili, Roberto and Pazienza, Maria Teresa and Zanzotto, Fabio Massimo, *Efficient Parsing for Information Extraction*, Proc. of the ECAI98 - Brighton, UK.

[2] Steven Abney, *Part-Of-Speech Tagging and Partial Parsing*, AT&T Laboratories - Research, New York, (1996).

[3] R. Basili and M.T. Pazienza and F.M. Zanzotto, *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Customizable Modular Lexicalized Parsing,(2000).

[4] Steven Abney, *Bootstrapping*, AT&T Laboratories - Research, New York, (2002).

[5] Steven Abney, Robert E.Schapire and Yoram Singer, *Boosting applied to tagging and PP attachment*, in 'Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora', (1999).

[6] Eneko Aguirre and German Rigau, *A Proposal for Word Sense Disambiguation using Conceptual Distance*, Lengoaia eta Sistema Informatikoak saila, Euskal Herriko Universitatea, Donostia; Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, (1995).

[7] Eneko Aguirre and German Rigau, *Word sense disambiguation using Conceptual Density*, Lengoaia eta Sistema Informatikoak saila, Euskal Herriko Universitatea, Donostia; Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, (1995).

[8] K.M.Ali, *Learning Probabilistic Relational Concept Descriptions*, PhD thesis, University of California, (1996).

[9] Eric Bauer and Ron Kohavi, *An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants* in 'Machine Learning', vol. 36, pp. 105-142 (1999), Computer Science Department, Stanford University, Stanford; Blue Martini Software, San Mateo, CA, (1998).

[10] Avrim Blum and Tom Mitchell, *Combining labeled and unlabeled data with Co-Training*, School of Computer Science, Carniage Mellon University, Pittsburg, (1998).

[11] L. Breiman, *Bagging predictors*, University of California, Berkeley, (1996).

[12] William W.Cohen and Yoram Singer, *A simple, fast and effective rule learner*, in 'Proceedings of the Sixteenth National Conference on Artificial Intelligence', (1999).

[13] Jerome Friedman, Trevor Hastie and Robert Tibshirani, *Additive logistic regression: a statistical view of boosting*, Technical Report, (1998).

[14] Sally Goldman and Yan Zhou, *Enhancing Supervised Learning with Unlabeled Data*, Department of Computer Science, Washington University, St.Louis, MO (2000).

[15] Frederick Jelinek, *Estimation of Probabilities from Counts and the Back-Off Model*, in Statistical Methods for Speech Recognition, MIT Press (1997).

[16] S. Katz, *Estimation of probabilities from sparse data for the language model component of a speech recognizer*, in IEEE Transaction on Acoustics, Speech, and Signal Processing (1987).

[17] Michael J. Kearns and Umesh V.Vazirani, *An Introduction to Computational Learning Theory*, MIT Press, (1994).

[18] S. Klink, T. J'ager, *A Comprehensive Voting of Commercial OCR Devices*, German Research Center for Artificial Intelligence GmbH, Kaiserslautern, Germany (1994).

[19] Erika F. de Lima, *Assigning Grammatical Relations with a Back-Off Model*, GMD - German National Research Center for Information Technology, Darmstadt, Germany (1997).

[20] George A. Miller, *Wordnet: a Dictionary Browser*, in 'Information in Data, Proceedings of the first cConference of the UW Center for the New Oxford Dictionary', Waterloo, Canada: University of Waterloo, (1985).

[21] George A. Miller, *Dictionaries in the Mind*, Language and Cognitive Process 1: 171-185, (1986).

[22] George A. Miller, *Introduction to WordNet: An On-Line Lexical Database*, (revised on August 1993).

[23] David Pierce and Claire Cardie, *Limitations of Co-Training for Natural Language Learning from Large Datasets*, Department of Computer Science, Cornell University, Ithaca NY, (2001).

[24] R.Rada, H.Mili, E.Bicknell and M.Blettner, *Development an application of a metric on semantic nets*, in IEEE Transactions on Systems, Man and Cybernetics, pp. 17-30, (1989).

[25] Philip Resnik, *Selectional Preference and Sense Disambiguation*, Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland, College Park, MD, (1997).

[26] Robert E.Schapire, *The Boosting Approach to Machine Learning*, ATT Labs - Research, Shannon Laboratory, Florham Park, NJ (2001).

[27] Yoav Freund and Robert E. Schapire, *A decision-theoretic generalization of on-line learning and application to boosting*, in 'Proceedings of the Second European Conference on Computational Learning Theory' (Springer-Verlag, pp. 23-37), ATT Labs - Research, Shannon Laboratory, Florham Park, NJ (1996).

[28] Michael Sussna, *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network*, in 'CIKM 93, Proceedings of the Second International Conference on Information and Knowledge Management, Washington, DC, USA', (1993).

[29l] L.G.Valiant, *A Theory of the learnable*, Communications of the ACM, 27(11):1134-1142, (1984).

[30] Vladimir N.Vapnik, *The Nature of Statistical Learning Theory*, Spinger, (1995).

[31] David Yarowsky, *One Sense per Collocation* in ARPA Workshop on Human Language Technology, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, (1993).

[32] David Yarowsky, *Unsupervised word sense disambiguation rivaling supervised methods*, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, (1995).

[33] Patrick Hanks, *Evidence and intuition in lexicography*, In Jerzy Tomaszczyk and Barbara Lewandowska-Tomaszczyk, editors, Meaning and Lexicography. John Benhamins Publishing Company, Amsterdam/Philadelphia, (1990).