# Which is to be master?

ambiguità e word sense disambiguation

Alessandra Zarcone
Corso di laurea in Informatica Umanistica
Università degli Studi di Pisa
Anno Accademico 2005/06

*"When I use a word," Humpty Dumpty said,*
*in rather a scornful tone,*
*"it means just what I choose it to mean -- neither more nor less."*
*The question is," said Alice,*
*"whether you can make words mean so many different things."*
*"The question is," said Humpty Dumpty,*
*"which is to be master -- that's all."*
Lewis Carroll, "Through the Looking Glass"

# Which is to be master?

**sommario**

- **Tipi di ambiguità**
- **Un approccio knowledge-based: la Selectional restriction-based disambiguation**
- **Robust word sense disambiguation: apprendimento automatico e risorse lessicografiche**
- **Valutazione: il SENSEVAL**

# Which is to be master?

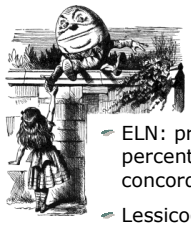
**l'ambiguità nel linguaggio**

- fenomeno pervasivo
- *il lessema è "l'associazione di una particolare forma fonologica con una forma di rappresentazione simbolica del significato"* (Jurafsky e Martin, 2000)
- ma questa associazione non è biunivoca

forma ortografica ⟺ significato

forma fonologica ⟺ significato

(associazioni non biunivoche)



# Which is to be master?

**quanti e quali significati?**

- ELN: problema dell'ITA (inter-tagger agreement: percentuale di risposte su cui gli annotatori concordano, spesso lontana dal 100%)
- Lessicografia: problema del dizionario (quali significati di una parola vanno accorpati e quali vanno invece distinti?)
- *"Not only is a lexicographer 'a lexicologist with a deadline' (Fillmore, 1988), but also a lexicologist with a page limit"* (Kilgariff, 1997)

lessicografia tradizionale ⟺ lessicografia dei corpora

# Which is to be master?

## disambiguare

- problema della disambiguazione (naturale per un parlante, ma problematica per un computer):

[ espressioni in linguaggio naturale]

(ambigue e imprecise)

⇩

[ rappresentazione del significato ]

(non ambigua)

# Which is to be master?

## diversi tipi di ambiguità lessicale

| | **Stessa categoria morfologica** | **Diversa categoria morfologica** |
|---|---|---|
| **Ambiguità Complementare** | "I am painting a double-hung **window**"<br>"She came in through the bathroom **window**"<br>(the Beatles) | "Il suono del tuo **riso** non è più lieto"<br>(E. Montale)<br>"Hai **riso** di te stesso" |
| **Ambiguità Contrastiva** | "C'era una **volta**.. - Un re! - diranno subito i miei piccoli lettori"<br>(Collodi)<br>"La navata è coperta da una **volta** a crociera" | "IL FIORELLINO: Che bella cosa essere nato vicino a te. Così tu mi ripari dalla pioggia. Ma dimmi: sei un vero ombrello o **fungi** da ombrello?<br>IL **FUNGO**: **Fungo**."<br>(A.Campanile) |

# Which is to be master?

## ambiguità strutturale

- Realizzazione della struttura argomentale ambigua:

*"la descrizione di Giorgio": Giorgio è agente o tema?*

- Struttura sintattica ambigua:

[ [Una vecchia] [legge [la regola] ] ]

⇕

[ [Una vecchia legge] [la [regola] ] ]

[ [Ho visto [Giorgio [con un binocolo] ] ] ]

⇕

[ [Ho visto [Giorgio] [ con un binocolo] ] ]

# Which is to be master?

## creatività e metafore

- metafore, idiomi, giochi di parole (sfruttare le ambiguità):

*"veniamo al succo del discorso"*

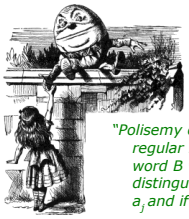
*"e poi il ladro ha tagliato la corda"*

*"Capitano! Arrivano i monsoni!"*
*"Preparatevi all'attacco!"*
*"Ma, capitano! Sono venti!"*
*"Anche se fossero cinquanta, li batteremo!"*

- creatività (creare nuove ambiguità):

*"me lo dica... il maglione rosso, là in fondo.. "*

*"il tavolo 4 ha ordinato un sandwich"*



# Which is to be master?

## polisemia regolare

*"Polisemy of the world A with the meanings $a_i$ and $a_j$ is called regular if, in the given language, there exists at least one other word B with the meanings $b_i$ and $b_j$, which are semantically distinguished from each other in exactly the same way as $a_i$ and $a_j$ and if $a_i$ and $b_i$ and $a_j$ and $b_j$ are non-synonimous"*

(Apresjan, J.D., 1974)

- **Figure/Ground**: finestra, cancello
- **Container/Containee**: bicchiere, scatola
- **Count/Mass**: birra, gelato
- **Product/Producer**: Coca-Cola, Honda, l'Unità
- **Plant/Food**: caffè, pomodoro, papaya
- **Process/Result**: distruzione, costruzione
- **Place/People**: Argentina, Genova, Europa
- **Attività/Telico**: mangiare/mangiare una mela
- **Ingressivo/Stativo**: "I soldati impugnavano le armi", "Calzava un paio di sandali"
  "I soldati impugnarono le armi", "Calzò un paio di sandali"

# Which is to be master?

## una potenziale fonte di errore

- per tutte le applicazioni di ELN, l'ambiguità è una potenziale fonte di errore

Babel Fish Translation


- Traduzione automatica
- Question answering
- Topic detection
- Information retrieval
- Clustering on-line

Google News Italia BETA

Vivísimo

# Which is to be master?

## l'ambiguità in ELN

| | |
|---|---|
| A. fonematica | Speech Recognition |
| A. lessicale morfologica | Part-of-Speech Tagging |
| A. lessicale semantica | **Word Sense Disambiguation** |
| A. strutturale | Syntactic Disambiguation |
| A. dell'atto linguistico | Speech Act Interpretation |

**Word Sense Disambiguation (WSD):**

- Selectional restriction-based WSD
- Robust WSD

# Which is to be master?

## selectional restriction-based WSD

- Eliminazione dei significati non appropriati sfruttando il contesto
- È necessario usare informazioni di sottocategorizzazione

(uso di ontologie, di database relazionali, es. WordNet)

VP $\rightarrow$ V NP

VP $\rightarrow$ suonare <theme> {theme:StrumentoMusicale}
V $\rightarrow$ suonare {$\exists$ e, x, y Suonare(e) $\wedge$ Agent(e,x) $\wedge$
Theme(e,y) $\wedge$ Isa(y, StrumentoMusicale) }

# Which is to be master?

## selectional restriction-based WSD

- Restrizione di selezione argomentale:

  *Chiara suona molto bene la **tromba***

  ⇕

  *Primo Levi si è gettato dalla **tromba** delle scale*

- Restrizione di selezione reciproca:

*She likes **playing** the **bass***

(WordNet: 8 significati per *bass*, 35 per *play*, ma solo un significato di *bass* e un significato di *play* concordano)

# Which is to be master?

## selectional restriction-based WSD: limiti

Le restrizioni di selezione hanno dei limiti:

- se il contesto è troppo generico per bloccare gli altri significati di un lessema;

  *"Che tipo di piatti preferisci?"*

- se le restrizioni sono deliberatamente violate;

  *"Non posso mica mangiarmi il piatto!"*

- se vi sono delle espressioni idiomatiche di cui le restrizioni non tengono conto;

  *"Volevo nasconderle la cosa, ma lei ha mangiato subito la foglia!"*

# Which is to be master?

## Apprendimento automatico

**TARGET** = parola da disambiguare
**CONTESTO** = es. un vettore di feature salienti (collocazioni o co-occorrenze)

- **Supervisionato** (es. classificatori bayesiani, liste decisionali)
  - **Training**: **INPUT** = TARGET + CONTESTO + RISPOSTA CORRETTA
    **OUTPUT**= REGOLE
  - **Test**: **INPUT** = TARGET + CONTESTO
    **OUTPUT**= RISPOSTA CORRETTA
- **Non-supervisionato** (es. self-organizing maps)
  - **Training**: **INPUT** = TARGET + CONTESTO
    **OUTPUT**= CLUSTERING
  - **Test**: **INPUT** = TARGET + CONTESTO
    **OUTPUT**= POSIZIONAMENTO DEL TARGET IN UN CLUSTER APPROPRIATO

# Which is to be master?

## classificatori bayesiani

- scegliere il significato più appropriato dato un vettore *V* equivale a scegliere il significato più probabile:

  **per $s \in S$ $\quad i = \text{argmax}\ P(s|V)$**

  dove S sta per l'insieme dei significati della parola-target

- ma sappiamo che

$$P(s \wedge V) = P(V|s)\,P(s)$$
$$P(s \wedge V) = P(s|V)\,P(V)$$

- uguagliando i due membri destri, e dividendo per $P(V)$:

$$P(s|V) = (P(V|s)\,P(s))/P(V)$$
$$i = \text{argmax}\ (P(V|s)\,P(s))/P(V)$$

- per stimare $P(s)$ basterà la frequenza di s rispetto a *S*, $P(V)$ è costante al variare di s e può essere eliminata

$$i = \text{argmax}\ P(V|s)\,P(s)$$

$P(V|s)$: likelihood; $P(s)$: prior probability

# Which is to be master?

## decision lists

- serie di test – se un test ha successo, viene restituito il significato associato a quel test, se il test non ha successo, si verifica il successivo
- potenzialmente ogni coppia significato-feature contestuale può diventare un test della decision list
- durante la fase di training si formano le liste e i test sono ordinati per accuratezza

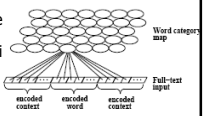
| | |
|---|---|
| *fish* within window | ⇨ bass1 |
| *striped bass* | ⇨ bass1 |
| *guitar* within window | ⇨ bass2 |
| *bass* player | ⇨ bass2 |
| *piano* within window | ⇨ bass2 |
| *tenor* within window | ⇨ bass2 |
| *sea bass* | ⇨ bass1 |
| *play/V bass* | ⇨ bass2 |
| *river* within window | ⇨ bass1 |
| *violin* within window | ⇨ bass2 |
| *salmon* within window | ⇨ bass1 |
| *on bass* | ⇨ bass2 |
| *bass are* | ⇨ bass1 |

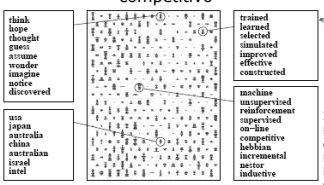
Esempio di decision list per la disambiguazione tra i due sensi di "bass" in inglese (pesce o strumento musicale) (da Jurafsky & Martin 2000)

# Which is to be master?

## self-organizing maps (SOM)

- Tipo di rete neurale (Teuvo Kohonen) basata su una griglia di neuroni artificiali, ognuno con una sua collocazione
- la mappa viene sottoposta a una fase di addestramento, durante il quale si definiscono le collocazioni dei neuroni sulla mappa, tramite un continuo adattamento di pesi, di tipo competitivo




- durante la fase di test, ogni nuovo contesto viene categorizzato (viene collocato automaticamente nella mappa di output)

Immagini da: Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen (1996): Creating an order in gigital libraries with self-organizing maps. Proceedings of WCNN'96, World Congress on Neural Networks, September 15-18, San Diego, California, Lawrence Erlbaum and INNS Press, Mahwah

# Which is to be master?

## risorse lessicografiche

- gli approcci precedenti non funzionano bene su larga scala
- uso di risorse lessicografiche machine-readable: confrontare le glosse del target con le glosse delle parole del contesto, e scegliere il significato relativo alle glosse con maggiore overlap (Lesk 1986)

| | |
|---|---|
| **pine** | 1. kinds of **evergreen tree** with needle-shaped leaves |
| | 2. waste away through sorrow of illness |
| **cone** | 1. solid body which narrows to a point |
| | 2. something of this shape wheter solid of hollow |
| | 3. fruit of certain **evergreen trees** |

Da: Lesk, M.E. (1986), Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone; in Proceedings of the Fifth International Conference on System Documentation, Toronto, CA.

# Which is to be master?

## risorse lessicografiche

- le risorse lessicografiche disponibili possono non essere adeguate (glosse troppo brevi, possono utilizzare sinonimi delle parole presenti nel contesto)
- espansione della ricerca: codici tematici delle varie voci (bot., anat., elettr.)
- sfruttare gruppi di iponimia/catene di iperonimia (WordNet)
- possibilità di arricchimento delle risorse lessicografiche (es. WordNet) con conoscenza enciclopedica (es. Wikipedia)

(*Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets* Maria Ruiz-Casado, Enrique Alfonseca and Pablo Castells - www.ii.uam.es/~ealfon/pubs/2005-awic.pdf )

# Which is to be master?

## il modello di Dutoit

- **Dicologique**: più di 100.000 parole e frasi raggruppate analogicamente e per relazioni di sinonimia

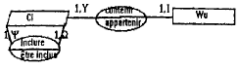
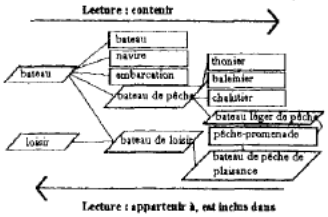
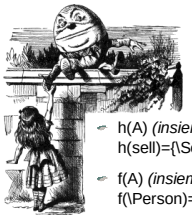


*Fig 1: G, the conceptual model of the dictionary*




- **The Integral Dictionary**: versione (espansa) di Dicologique usata per applicazioni di ELN (rete semantica basata su analisi semantica composizionale)
- Francese: 185.000 significati
  Inglese: 79.000 significati
  Spagnolo, Italiano, Tedesco: 39.500 significati



# Which is to be master?

## il modello di Dutoit

- h(A) *(insieme degli antenati di A)*
  h(sell)={\Sell, \Universe}
- f(A) *(insieme dei figli di A)*
  f(\Person)={seller, florist}

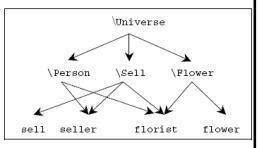
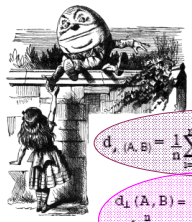



Figure 1: An example of semantic graph

- d(A, B) = d(B, A) *(n di intervalli tra i due nodi)*
  d(sell, \Sell)=1
- c(A) *(insiemi di archi tra il nodo A e la radice)*
  c(seller)={(seller, \Sell), (seller, \Person), (\Sell, \Universe), (\Person, \Universe)}
- NCA(A, B) *(più vicino antenato comune tra A e B)*
  NCA(A, B) = f ( c(A) ∩ c(B) ) – [ f ( c(A) ∩ c(B) ) ∩ h ( c(A) ∩ c(B) ) ]
- ANCA(A, B) *(NCA asimmetrico)*
  ANCA(A, B) = h ( c(A) ∩ c(B) ) che hanno un collegamento diretto a h(A) ma non ad h(b)



# Which is to be master?

## il modello di Dutoit

activation

$$d_{a\,(A,B)} = \frac{1}{n}\sum_{i=1}^{n}(d(A,NCA_i)+d(B,NCA_i))$$

$$d_{i\,(A,B)} = d_{a\,(A,B)} + \frac{1}{n}\sum_{i=1}^{n}(d(A,ANCA_i)+d(B,ANCA_i))$$

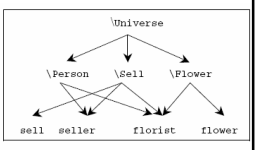



Figure 1: An example of semantic graph

proximity

$$h(M)=\bigcup_{i=1}^{n} h(m_i)$$

$$c(M)=\bigcup_{i=1}^{n} c(m_i)$$

definizione di h e c per gruppi di nodi → similarità tra parola e gruppi di parole

- Le parole del contesto che sono più vicine (proximity minima) alla parola ambigua sono descrittori migliori del significato della parola

# Which is to be master?

## valutare un sistema di WDS

- importanza di uno standard di valutazione
  - all'interno della comunità scientifica: per confrontare diversi sistemi di WSD
  - per gli sviluppatori di un sistema di WSD: per capire quando il sistema sta migliorando la propria performance
- Lo standard:
  - definire in modo chiaro e dettagliato i compiti di un sistema di WSD
  - creare un *gold standard* (corpus annotato da umani, con i sensi "corretti" di ogni parola) su cui testare il sistema

(una disambiguazione "grossolana" può essere facile, più complicato è distinguere tra due significati molto vicini di una parola polisemica)

# Which is to be master?

## valutare un sistema di WDS

- per questo motivo nasce SENSEVAL

*"an experiment designed to replace scepticism about both the reality of word senses and the effectiveness of WSD, by percentages"*

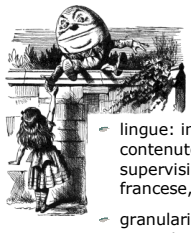
(Kilgariff, 2000)

# Which is to be master?

## SENSEVAL 1

- settembre 1998, Herstmonceux Castle (Sussex, UK)

| **Gruppo di ricerca** | **Referente** | **Lingua** |
|---|---|---|
| Bertin, Univ Avignon | Claude de Loupy/Marc El Beze | Eng/o |
| CL Research, USA | Ken Litkowski | Eng/a |
| CNR, Pisa | Vito Pirelli | It |
| Univ Durham | Paul Hawkins | Eng/s |
| Educ Testing Service, Princeton | Claudia Leacock | Eng/s |
| EPFL, Lausanne | Romaric Besançon | Fr |
| Johns Hopkins Univ | David Yarowsky | Eng/s |
| Korea Univ | Ho Lee | Eng/s |
| Univ Sains Malaysia | Cheng Ming Guo | Eng/a |
| Univ Manitoba | Dekang Lin | Eng/a |
| Univ Manitoba | Keith Suderman | Eng/s |
| New Mex State, UNC Asheville | Tom O'Hara/R. Bruce/ J. Wiebe | Eng/s |
| Univ Ottawa | Stan Szpakowicz/Ken Barker | Eng/a |
| Univ Sunderland | Jeremy Ellman | Eng/a |
| Univ Sussex | Diana McCarthy | Eng/o |
| Tech Univ Catalonia, Univ Basque | Lluis Padro/Eneko Agirre | Eng/a |
| Univ Tilburg | Walter Daelemans | Eng/s |
| XRCE/CELI | Frédérique Segond/Luca Dini | Eng/a |
| XRCEF | Frédérique Segond | Fr |



# Which is to be master?

## SENSEVAL 1

- lingue: inglese (Eng/a=disambiguazione di tutte le parole contenuto in un testo; Eng/s=sistemi di apprendimento supervisionato; Eng/o=altri sistemi di apprendimento), francese, italiano
- granularità ammessa: può essere più o meno ampia, a seconda che i sensi da individuare siano più generici o più precisi

| | granularità fine precision (recall) | granularità media precision (recall) | granularità grossa precision (recall) |
|---|---|---|---|
| umano | **0.965** (0.963) | **0.968** (0.967) | **0.970** (0.968) |
| sistema migliore | **0.781** (0.781) | **0.804** (0.804) | **0.818** (0.818) |
| media dei sistemi | **0.639** (0.518) | **0.696** (0.555) | **0.717** (0.571) |
| sistema peggiore | **0.418** (0.127) | **0.511** (0.511) | **0.538** (0.538) |
| baseline | **0.691** (0.689) | **0.720** (0.719) | **0.741** (0.739) |

# Which is to be master?

## SENSEVAL 2

- luglio 2001, Tolosa
- lingue: basco, cinese, ceco, danese, olandese, inglese, estone, italiano, giapponese, coreano, spagnolo, svedese

**English all words**

| | granularità fine precision (recall) | granularità grossa precision (recall) |
|---|---|---|
| sistema migliore | **0.748** (0.69) | **0.748** (0.698) |
| media dei sistemi | **0.491** (0.354) | **0.499** (0.359) |
| sistema peggiore | **0.287** (0.033) | **0.294** (0.034) |

| **Italian lexical sample** | granularità fine precision (recall) | granularità media precision (recall) | granularità grossa precision (recall) |
|---|---|---|---|
| IRST Trento | **0.406** (0.389) | **0.482** (0.461) | **0.483** (0.463) |
| Johns Hopkins Univ. | **0.353** (0.353) | **0.421** (0.421) | **0.423** (0.423) |

# Which is to be master?

## SENSEVAL 3

- luglio 2004, Barcellona
- lingue: basco, cinese, inglese, italiano, spagnolo, svedese, catalano, rumeno

**Task di SENSEVAL 3:**

- English all words
- Italian all words
- Basque lexical sample
- Catalan lexical sample
- Chinese lexical sample
- English lexical sample
- Italian lexical sample
- Romanian lexical sample
- Spanish lexical sample
- Automatic subcategorization acquisition
- Multilingual lexical sample
- WSD of WordNet glosses
- Semantic Roles
- Logic Forms

**nuovi task**

***...verso SENSEVAL 4!***

***(estate 2007)***

# Bibliografia


- DUTOIT, D. [1992], **A set-theoretic approach to l.exical semantics**; in Proceedings of COLING 1992.
- DUTOIT, D. & POIBEAU, T. [2002], **Inferring Knowledge from a Large Semantic Network**; in Proceedings of COLING 2002.
- JURAFSKY, D. & MARTIN, J.H. [2000], **Speech and Language Processing**; Prentice Hall, Englewood Cliffs, New Jersey.
- KILGARIFF, A. & PALMER, M. [2000], **Introduction to the Special Issue on SENSEVAL**; in Computer and the Humanities, 34; 1-13.
- PUSTEJOVSKY, J. & BURGAEV, B. (eds.) [1996], **Lexical Semantics - The Problem of Polisemy**; Oxford, Oxford University Press.
- RAVIN, Y. & LEACOCK, C. (eds.) [2000], **Polisemy - Theoretical and Computational Approaches**; Oxford, Oxford University Press.
- http://www.itri.brighton.ac.uk - information technology research institute, Brighton (UK)
- http://www.senseval.org – SENSEVAL official site