

Word Sense Disambiguation

Tutorial

Corso di Linguaggi & Traduttori
Università degli Studi di Bari

Pierpaolo Basile (Ph.D. Student)



Outline

- Definizione del problema
- Soluzione del problema
- Un po' di storia
- Applicazioni
- Approcci
- Valutazione/Risultati
- Riferimenti



Definizione del problema



- **Word Sense Disambiguation**
 - È il problema di selezionare un significato per una parola da un insieme definito di possibilità (*sense inventory*)
- **Word Sense Discrimination**
 - È il problema di suddividere l'utilizzo di una parola in differenti significanti senza avere un insieme predefinito di possibilità

3

Come



- **Disambiguation**
 - Knowledge-based methods
 - Supervised methods (Machine Learning)
 - Bootstrapping approaches
- **Discrimination**
 - Unsupervised methods (clustering)

4

Human VS Computer



- *Polisemia* – una parola ha più di un significato
 - bank #1: a financial institute
 - bank #2: sloping land
 - bank #3: a supply or stock held in reserve for future
- Un umano è in grado di disambiguare il significato una macchina non possiede la conoscenza necessaria
- L'ambiguità *generalmente* non è un problema per gli umani ma lo è per le macchine

5

Un po' di storia



- 1960s
 - Approccio *cognitivist*
- 1970s - 1980s
 - Sistemi a regole
 - Utilizzo di risorse annotate manualmente
- 1990s
 - Approcci basati su corpora
 - Apprendimento da testi annotati
- 2000s
 - Sistemi ibridi
 - Minimizzare o eliminare l'uso di testi annotati
 - Sfruttare altre risorse come il WEB
 - Teoria dei grafi (**NEW**)

6

Conessioni con altri campi



- Scienze Cognitive e Psicologia
 - Quillian (1968), Collins and Loftus (1975) : spreading activation
- Linguistica
 - Resnik (1993): approccio statistico utilizzando corpora
- Filosofia del linguaggio
 - Teorie sull'utilizzo della semantica

7

Applicazioni



- Traduzione automatica
 - Scegliere il termine più appropriato in base al significato
- Information Retrieval
 - Ricerca semantica
- Question Answering
 - Disambiguare l'oggetto della richiesta
- Knowledge Acquisition
 - Memorizzare in maniera corretta l'informazione acquisita

8

Approcci [1/3]



- Knowledge-Based Disambiguation
 - Utilizzo di una o più risorse linguistiche esterne (Knowledge-Based), ad esempio un dizionario elettronico
 - Utilizzo di alcune proprietà del discorso: una parola mantiene lo stesso significato in un discorso
 - Utilizzo di alcune euristiche

9

Approcci [2/3]



- Supervised Disambiguation
 - Basato su training set di documenti annotati manualmente
 - Gli esempi sono classificati: etichettati con il significato corrispondente
 - Estrazione delle features
 - Classificazione di nuovi esempi

10

Approcci [3/3]



- Unsupervised Disambiguation
 - Basato su corpora non etichettati
 - Gli esempi non sono etichettati: non esiste un set di significati predefinito
 - Estrazione delle features
 - Clustering degli esempi

11

All Words Word WSD



- Cercare di disambiguare tutte le *open-class words* in un testo:
 - “He **put** his **suit** over the **back** of the **chair**”
- Approccio Knowledge-based
 - Usare le informazioni contenute in un dizionario
 - Posizione dei significati in una rete semantica
 - Uso delle proprietà del discorso
 - Significato più frequente

12

Dizionari



- Algoritmo di Lesk
 - Misurare la sovrapposizione delle glosse prendendole da un dizionario
 - Selezionare il significato la cui glossa ha maggiore *overlap* con le glosse delle altre parole nel contesto

13

Lesk



Yesterday I played basketball.

1) play (participate in **games** or sport)
2) play (act or have an effect in a specified way or with a specific effect or outcome)
3) play (play on an instrument)
...

1) basketball, basketball **game**, hoops (a **game** played on a court by two opposing teams of 5 players; points are scored by throwing the ball through an elevated horizontal hoop)
2) basketball (an inflated ball used in playing basketball)

14

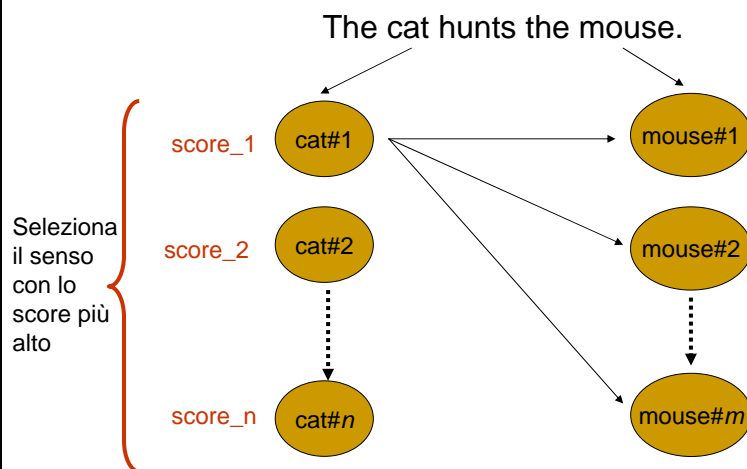
Similarità [1/2]



- IDEA: scegliere per ogni parola il significato che massimizzi la similarità
- Utilizzare una rete semantica per calcolare la distanza dei significati
- Scegliere una funzione di similarità

15

Similarità [2/2]



16

Misure di similarità



- **Length path**: minima distanza (n. di nodi) tra i due concetti nella gerarchia
- **Leacock&Chodorow**: $-\log(\text{length} / (2 * D))$
 - Length = length path
 - D = massima profondità consentita nella gerarchia
- **Resnik**
 - Information Content del Most Specific Subsumer
 - MSS è il sussuntore più specifico tra due concetti

Per fare delle prove: <http://marimba.d.umn.edu/cgi-bin/similarity.cgi>

17

Information Content



- $IC = -\log(Pr(S))$ misura la quantità di informazione collegata ad un particolare evento
- $Pr(S)$ è la probabilità che S sia il significato corretto per la parola
- Stimare S con la frequenza calcolata da un corpus annotato

18

JIGSAW



- **Knowledge-based** WSD algorithm
- Disambiguation of words in a text by exploiting **WordNet** senses
- Combination of **three different strategies** to disambiguate nouns, verbs, adjectives and adverbs
- **Main motivation:** the effectiveness of a WSD algorithm is strongly influenced by the **POS tag** of the target word

19

JIGSAW algorithm



- **Input:** document $d = \{w_1, w_2, \dots, w_h\}$
- **Output:** list of WordNet synsets $X = \{s_1, s_2, \dots, s_k\}$
 - each element s_i is obtained by disambiguating the target word w_i
 - based on the information obtained from WordNet about words in the context
 - **context C of the target word:** a window of n words to the left and another n words to the right, for a total of $2n$ surrounding words
- For each word JIGSAW adopts a different strategy based on POS tag

20

JIGSAW_nouns: the idea



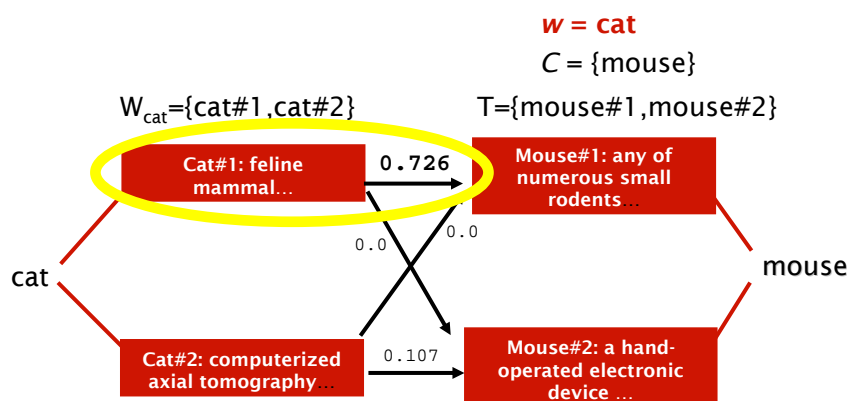
- Based on Resnik algorithm for disambiguating noun groups
- Given a set of nouns $W=\{w_1, w_2, \dots, w_n\}$ from document d :
 - each w_i has an associated sense inventory $S_i=\{s_{i1}, s_{i2}, \dots, s_{ik}\}$ of possible senses
 - **Goal:** assigning each w_i with the most appropriate sense $s_{ih} \in S_i$, according to the similarity of w_i with the other nouns in W

21

JIGSAW_nouns: semantic similarity

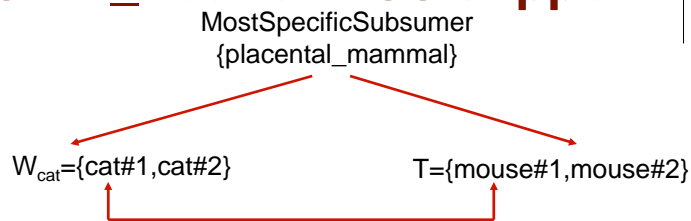


“The white **cat** is hunting the **mouse**”



22

JIGSAW_nouns: MSS support



- *MostSpecificSubsumer* between words
- Give more importance to senses that are hyponym of MSS
- **Combine MSS support with semantic similarity**

23

Difference between JIGSAW_nouns and Resnik



- **Leacock-Chodorow measure** to calculate similarity (instead Information Content)
- a **Gaussian factor G**, which takes into account the distance between words in the text
- a **factor R**, which takes into account the synset frequency score in WordNet
- a **parameterized search** for the MSS (*Most Specific Subsumer*) between two concepts

24

JIGSAW_verbs: the idea



- Try to establish a **relation between verbs and nouns** (distinct IS-A hierarchies in WordNet)
- Verb w_i disambiguated using:
 - nouns in the *context* C of w_i
 - nouns into the *description* (*gloss* + *WordNet usage examples*) of each candidate synset for w_i

25

JIGSAW_verbs: algorithm [1/4]



- For each candidate synset s_{ik} of w_i
 - computes $\text{nouns}(i, k)$: the set of nouns in the *description* for s_{ik}
 - for each w_j in C and each synset s_{ik} computes the highest similarity max_{jk}
 - max_{jk} is the highest similarity value for w_j wrt the nouns related to the k -th sense for w_i (using Leacock-Chodorow measure)

26

JIGSAW_verbs: algorithm [2/4]



I play basketball and soccer $w_i=play$
 $C=\{basketball, soccer\}$

1. (70) play -- (participate in games or sport; "We played hockey all afternoon"; "play cards"; "Pele played for the Brazilian teams in many important matches")
2. (29) play -- (play on an instrument; "The band played all night long")
3. ...

nouns(play,1): game, sport, hockey, afternoon, card, team, match

nouns(play,2): instrument, band, night

...

nouns(play,35): ...

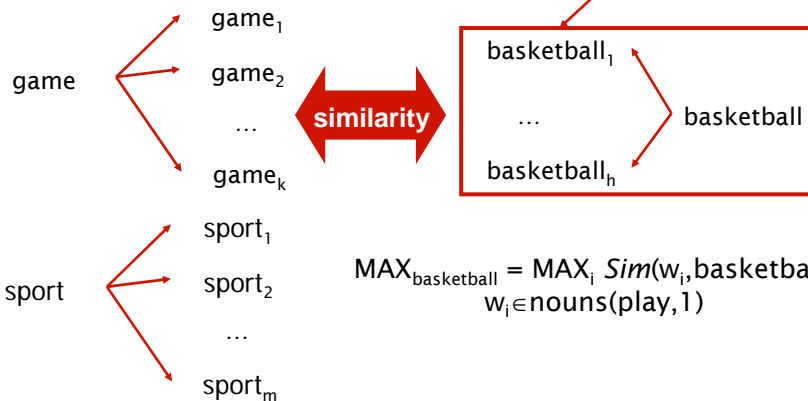
27

JIGSAW_verbs: algorithm [3/4]



$w_i=play$
 $C=\{basketball, soccer\}$

nouns(play,1): game, sport, hockey, afternoon, card, team, match



28

JIGSAW_verbs: algorithm [4/4]



- finally, an **overall similarity score**, $\varphi(i, k)$, among s_{ik} and the whole *context* C is computed:

$$\varphi(i, k) = R(k) \frac{\sum_{w_j \in C} G(\text{pos}(w_i), \text{pos}(w_j)) \cdot \max_{jk}}{\sum_h G(\text{pos}(w_i), \text{pos}(w_h))}$$

- the synset assigned to w_i is the one with the highest φ value

29

JIGSAW_others



- Based on the WSD algorithm proposed by Banerjee and Pedersen (inspired to Lesk)
- **Idea:** computes the **overlap** between the *glosses* of each candidate sense for the target word to the *glosses* of all words in its context
- assigns the synset with the highest overlap score
 - if ties occur, the most common synset in WordNet is chosen

30

Approccio supervised



- Supervised
 - Apprendere da corpora annotati
 - SemCor è un corpus dove tutte le *open-class word* sono annotate semanticamente
- Buoni risultati in precisione
- Possibile basso richiamo: non è detto che gli esempi coprano tutte le open-class words

31

Approccio supervised



SEMCOR

```
<wf cmd=done pos=NN lemma=committee wnsn=1  
lexsn=1:14:00::>Committee</wf>  
<wf cmd=done pos=NN lemma=approval wnsn=1 lexsn=1:04:02::>approval</wf>  
<wf cmd=ignore pos=IN>of</wf>  
<wf cmd=done rdf=person pos=NNP lemma=person wnsn=1 lexsn=1:03:00::  
pn=person>Gov._Price_Daniel</wf>  
<wf cmd=ignore pos=POS>'s</wf>  
<punc>` `</punc>  
<wf cmd=done pos=JJ lemma=abandoned wnsn=1  
lexsn=5:00:00:uninhabited:00>abandoned</wf>  
<wf cmd=done pos=NN lemma=property wnsn=2 lexsn=1:21:00::>property</wf>
```

32

Approccio supervised



		features			
label		f_1	f_2	...	f_n
	<i>bank#1</i>	$v_{1,1}$	$v_{1,2}$		$v_{1,n}$
esemi	<i>bank#1</i>	$v_{2,1}$	$v_{2,2}$		$v_{2,n}$

	<i>bank#2</i>	$v_{m,1}$	$v_{m,2}$...	$v_{m,n}$

33

Approccio supervised



- Features
 - Pos-tag delle parole vicine
 - Lemma/stemming delle parole vicine
 - Informazioni sintattiche utilizzando un parser
- Classificazione
 - Classificazione del nuovo vettore di feature
 - Assegnazione della classe
 - **Metodi:** K-NN, Bayes, SVM, algoritmi genetici, ...

34

Approccio supervised



- Se non ci sono esempi per una determinata word?
 - First sense
 - Knowledge approach per le parole senza esempi
 - Altre euristiche

35

Bootstrapping approach



- Approccio misto supervised/unsupervised
- Partendo da pochi esempi si classificano nuovi esempi
- I nuovi esempi possono rientrare nel training set
- Scelta di un valore di soglia per scegliere gli esempi da includere nel training set

36

Target WSD



- Disambiguare solo una parola target
 - “Take a seat on this **chair**”
 - “The **chair** of the Math Department”
- WSD è visto come un tipico problema di classificazione
 - Utilizzo di tecniche di machine learning
- Training:
 - Corpus di tanti esempi della target word annotati con il relativo significato
 - Costruzione del vettore delle features
- Disambiguation:
 - Disambiguare la parola target in un nuovo esempio

37

Target WSD



- Prendere una finestra di n parole attorno alla parola target
- Costruire il vettore delle features
 - words, root forms, POS tags, frequency, ...
- Costruire il modello di training
- Classificare le nuove istanze

38

Unsupervised Disambiguation



- Disambiguare i significati di una parola:
 - Senza alcun dizionario
 - Senza training
 - Senza alcuna risorsa, il set dei significati non è predefinito
- E' impossibile distinguere tra "chair/furniture" e "chair/person" ma è possibile:
 - Raggruppare i contesti simili in cui compare la parola
 - Discriminare i significati in base ai gruppi

39

Unsupervised Disambiguation



- IPOTESI: i significati delle parole compaiono in contesti simili
- Metodo
 - Identificare il vettore dei contesti per tutte le occorrenze della parola che si vuole disambiguare
 - Partizionare gli esempi
 - Ogni gruppo di esempi rappresenta un significato

40

Valutazione di WSD



- Metriche
 - Precisione= $\frac{\#istanze_corrette}{\#istanze_classificate}$
 - Richiamo= $\frac{\#istanze_corrette}{\#tot_istanze}$
- Esempio test su 100 parole
 - Il sistema riesce a disambiguare 75 parole
 - Disambigua correttamente 50 parole
 - precision=50/75; recall=50/100
- Confrontare con uno standard
 - SEMCOR corpus, **SENSEVAL** corpus, **SEMEVAL** corpus

41

Valutazione di WSD



- Difficoltà nella valutazione
 - La natura delle parole che si vuole disambiguare rende complicata la valutazione
 - Coarse VS Fine-grained
- Sense maps
 - Raggruppare significati simili
 - Garantisce sia una fine-grained che coarse-grained evaluation

42

Links



- **SENSEVAL/SEMEVAL:** <http://www.senseval.org/>
- Rada Mihalcea: <http://www.cs.unt.edu/~rada/>
- Ted Pedersen: <http://www.d.umn.edu/~tpederse/>
- **WordNet:** <http://wordnet.princeton.edu/>
- **MultiWordNet:**
<http://multiwordnet.itc.it/english/home.php>