Lezione 1

Semantica Lessicale

E' lo studio del significato delle parole e delle loro relazioni: cosa vogliono dire i singoli items lessicali, perché hanno quel significato e come rappresentiamo la conoscenza → uso di ontologie e' un approccio. L'ontologia e' alla base di un sistema di rappresentazione della conoscenza. L'ontologia permette di condividere questa conoscenza. Perche si usa l'ontologia?

- ci permette di **rappresentare in diversa maniera lo stesso concetto** (ie: bicchiere mezzo pieno, bicchiere mezzo vuoto, bicchiere contiene 20 cc)
- **ogni rappresentazione approssima in maniera imperfetta la realtà** (non ho un punto di vista completo con una sola rappresentazione). Quindi s**cegliere una rappresentazione piuttosto che un altra significa decidere cosa vedere e decidere cosa escludere**: **commitments**. Analogia con il blur degli occhiali, decidiamo cosa mettere a fuoco, non e' un side-effect delle ontologie, ma e' l'essenza delle ontologie(una cosa positiva).

Ontologia

Un'ontologia definisce un insieme di primitive (**classi e relazioni**) con le quali modellare un dominio. **Definizioni di classi e relazioni includono il loro significato e vincoli**.
Vengono usate per la condivisione della comprensione (descrizione) delle entità di un certo dominio, in funzione del riutilizzo dei dati e dell'informazione.

Ontology Design

entità: oggetti che continuano per un periodo di tempo mantenendo la propria identità
eventi: oggetti che accadono, si svolgono o sviluppano nel tempo
ontologie fondazionali: un'ontologia fondazionale cattura un insieme di distinzioni base valide in vari domini. Es: Dolce

DOLCE

- Enduranti (oggetti).
    - Hanno una chiara località spaziale
    - La località temporale è derivata dai perduranti in cui vivono
    - Possono cambiare nel tempo
    - Possono avere parti non essenziali
- Perduranti (eventi)
    - Hanno una chiara località temporale
    - La località spaziale è derivata dai suoi partecipanti (enduranti)
    - Non cambiano nel tempo
    - Tutte le parti sono essenziali

gli enduranti partecipano ai perduranti.

La relazione subclass-of: per valutare se le due entità sono in relazione di sottoclasse, possiamo utilizzare la nozione di **criteri di identità**. questi criteri di identità sono proprietà necessarie delle entità confrontate. Es: time-interval → time-duration. non sono in relazione di sotto classe perché hanno criteri di identità differenti. (stesso inizio e fine vs stessa durata).

Approccio Moltiplicativo
Entità diverse possono essere collocate nello **stesso punto spazio-tempo** quindi si potrebbe pensare siano la stessa cosa. Diciamo che sono entità **diverse** perché hanno **caratteristiche essenziali** diverse: vaso e argilla. diciamo che il vaso e' costituito da una quantità di argilla ma non e' una quantità di argilla.

Qualità
Le qualità sono le entità basiche che possiamo misurare o percepire: forma, colore, dimensione ecc. Le qualità sono una caratteristica individuale: due particolari non possono avere la stessa qualità. Ogni qualità ha dei certi valori: ad esempio la qualità "colore" può avere come value "rosso scuro". Il valore (**quale**) descrive la posizione della qualità all'interno di un certo **spazio concettuale** (chiamato quality space). Quando ad esempio due rose hanno lo stesso colore intendiamo che le loro qualità, che sono distinte, hanno la stessa posizione all'interno dello spazio concettuale, dato che hanno lo stesso quale.
Questo ci serve per capire la relazione tra "rosso" inteso come aggettivo e "rosso" inteso come sostantivo. Questo grazie al color space che è lo stesso.

io ho scritto cosi:
Teoria della Qualità:
Sono caratteristiche di individui specifici, **non esistono 2 individui distinti con la stessa qualità**, e non può esistere la qualità senza l'oggetto a cui si riferisce.
Es. la **qualità** è il rosso, il **valore** è lo specifico punto di rosso (questo viene chiamato **quale**, e descrive la **posizione di una qualità all'interno di uno spazio concettuale di qualità**)

Le **qualità** sono sempre diverse ma **possono puntare allo stesso quale** (nel caso in cui ad esempio due oggetti siano diversi ma dello stesso identico colore)

Lezione 2

Reti Semantiche

Formalismo nato dai primi progetti di traduzione automatica, ampiamente usato in applicazioni per l'elaborazione automatica del linguaggio naturale. Le più semplici sono i grafi relazioni.

Grafi relazionali

Struttura a grafo con **nodi e archi**. Limiti di espressività: facile esprimere **congiunzione** (and), ma la disgiunzione (or)? quantificazione universale? relazione oltre al binario?

Reti proposizionali

Estensione dei grafi relazioni con proposizione che possono essere dei nodi e non solo più degli archi. Si può esprimere il NOT, di conseguenza tramite de-morgan anche la disgiunzione. Ma rimane il problema dei quantificatori universali.

Conoscenze Gerarchiche
E' possibile strutturare gerarchicamente oggetti, ma anche azioni, eventi, stati, proprieta ecc… tramite le reti semantiche con la relazione IsA. Grazie all'ereditarietà e possibile assegnare solo legami essenziali mentre le restanti possono essere inferite (ottimizzazione spaziale ma aumento temporale). Ci possono essere problemi di ereditarietà multipla (caso di nixon), di solito si risolve considerando il shortest path, oppure si usa la **dissonanza cognitiva**: ad esempio la risposta su cosa sia Nixon deriva da cosa giudichiamo prevalente.

Frame

Un frame è una struttura che rappresenta le conoscenze di carattere generale che un individuo ha riguardo situazioni, luoghi, oggetti, personaggi stereotipati. L'utilizzo dei frame permette a un sistema di formulare previsioni ed avere delle aspettative. I frame servono a organizzare le conoscenze relative a un certo dominio in modo da facilitare

a) il reperimento delle informazioni e
b) i processi inferenziali necessari per agire in modo intelligente.

Tre livelli gerarchici: un livello di base, uno superordinato e uno subordinato.

1. i concetti di base costituiscono il modo naturale di categorizzare gli oggetti e le entità di cui è formato il nostro mondo
2. i concetti superordinati traggono la loro origine da una generalizzazione di tali categorie
3. i concetti subordinati provengono da una loro specializzazione.

Semantica Procedurale

analogamente alle reti semantiche, **i frame rappresentano le conoscenze in modo dichiarativo ma privo di una semantica formale**. pertanto, parlando di frame bisogna presupporre l'esistenza di procedure in grado di utilizzare le informazioni in essi contenute. la tradizione di ricerca sui frame si è divisa fra approfondimento degli aspetti teorici dei frame e sviluppo di linguaggi in grado di implementare in modo efficiente strutture di questo tipo

Prototipi

l'appartenenza categoriale non viene caratterizzata tramite un elenco di attributi necessari e sufficienti, ma nei termini di una **maggiore o minore somiglianza rispetto a membri tipici** della categoria, o prototipi.

Semantica lessicale (again)

È la disciplina che si occupa di stabilire cosa significano le parole.
Due principali problemi:

1. Polisemia
    a. Ho acquistato un sacco (uno zaino)
    b. Si è infilato in un sacco di guai (molti guai)
2. Semantica frasale (idiomi)
    a. Il ladro ha vuotato il sacco
    b. Anche con espressioni meno idiomatiche: 'giacca a vento', 'mulino a vento'

Semantica lessicale: analisi del contributo delle parole al significato della frase, e di come il contesto delle parole influenzi il loro significato.

Contesto: È l'insieme degli elementi adiacenti a una parola. Può essere:

- **sintattico**: insieme degli elementi adiacenti a una parola dal punto di vista delle loro proprietà sintattiche: può essere nominale, verbale, aggettivale, e
- **semantico**: elementi adiacenti alla parola, dal punto di vista delle loro proprietà semantiche: 'Saltare un fosso' vs. 'Saltare un pasto'
- **situazionale**(pragmatico o extralinguistico).
    - Nell'enunciato 'ho mangiato un sacco' il significato di sacco è disambiguato dagli elementi linguistici attigui
    - Nell'enunciato 'il tuo amico è forte', il significato di forte può essere energico, simpatico, e può divenire chiaro solo nel contesto specifico in cui è utilizzato

Teorie sulla natura del significato

1. Teoria referenziale del significato

le parole sono lo strumento attraverso il quale **facciamo riferimento a ciò che esiste: sedia è un riferimento a un oggetto; acquistare è un riferimento a un evento.**

2. Teoria mentalista o concettuale

arricchisce la precedente teoria referenziale con l'ipotesi che il riferimento fra parole e realtà non è diretto, **ma mediato dall'immagine mentale che costruiamo di queste entità**.
quando parliamo di 'sedia' non facciamo riferimento direttamente alla nozione di 'sedia', ma alla rappresentazione mentale che abbiamo della classe di oggetti che rientrano nel tipo 'sedia'

3. Teoria strutturale -> synset

il significato dei termini non consiste nel suo riferirsi per esempio a un oggetto (ipotesi referenziale), o nel rimando all'immagine mentale (ipotesi mentalista), **ma nel 'valore' che la parola assume in relazione alle altre parole presenti nella lingua che fanno parte dello stesso campo semantico e designano oggetti analoghi**. nel caso di sedia: sgabello, sedile, trono, poltrona, panca, etc

4. Teoria dei prototipi

appartiene agli sviluppi della teoria mentalista. I**l prototipo è un 'elemento esemplare di una categoria'.** Il prototipo come dispositivo fondamentale alla base della categorizzazione dell'esperienza (in particolare da dati concreti).
La rappresentazione mentale di un concetto è l'insieme di le rappresentazioni di alcuni esempi di quella categoria che noi abbiamo incontrato durante la nostra vita

5. Teoria distribuzionale

il significato delle parole è determinato in larga misura dell'insieme di altre parole con cui queste co-occorrono. **due parole sono tanto più simili quanto più sono simili i contesti in cui queste occorrono**

Calcolo del significato

una teoria del significato si propone di stabilire quale sia il significato delle parole e come si formi il significato delle frasi a partire dal significato attribuito alle parole. I due problemi fondamentali sono la contestualità del significato (tipicamente le parole del lessico, in combinazione con parole diverse, assumono significati diversi) e la polisemia delle parole.

Principio di composizionalità

spiega come il significato degli enunciati si formi a partire dal significato degli elementi lessicali che li compongono. Tranne quando abbiamo espressioni idiomatiche (la tavola rotonda), usi metaforici (il treno sbuffa) o polisemia (come assegno il senso? amico caro vs maglione caro).

Due Approcci.

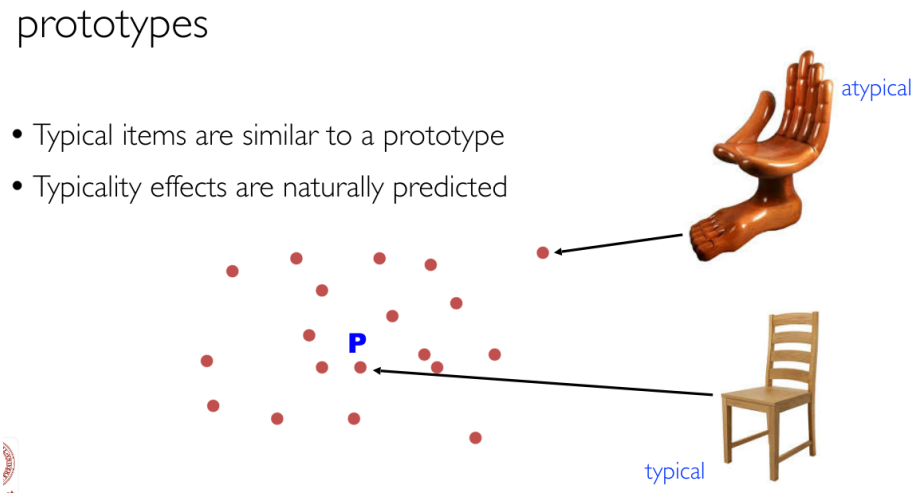
1. enumerazione dei sensi: i diversi sensi associati a un elemento lessicale sono elencati nella parola, insieme alle restrizioni lessicali che specificano i contesti in cui i diversi significati possono attivarsi
2. concezione dinamica del significato: il significato di ciascuna parola interagisce con il significato delle parole adiacenti

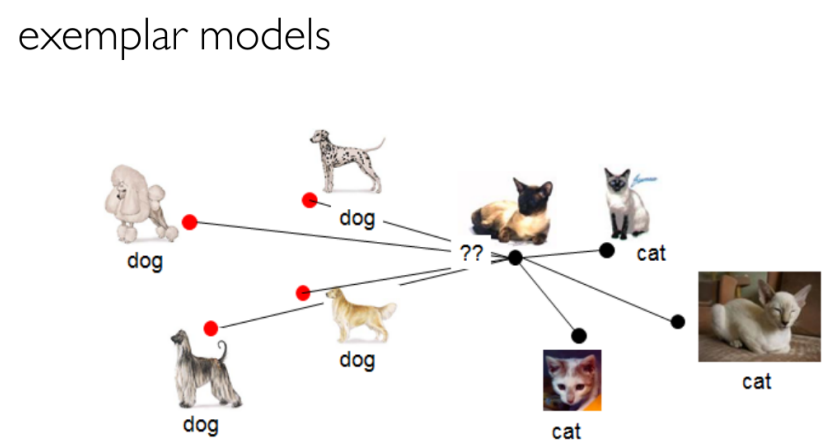
principi che consentono di illustrare diversi tipi di interazione semantica accanto alla composizionalità.

1. co-composizione: il significato del verbo potrebbe essere determinato da quello dei suoi argomenti (luca taglia il pane vs luca taglia l'erba)
2. forzatura (o conversione) del tipo semantico : un verbo in combinazione con un nome specifico lo 'spinge' a significare ciò che la semantica del verbo richiede, eventualmente variandone il tipo semantico (iniziare gli studi vs iniziare la lezione)
3. legamento selettivo: l'aggettivo può selezionare una specifica porzione del significato del nome(buon coltello)

Teoria classica vs teoria della tipicità
La prima e' quella classica basata su gerarchie, relazioni e vincoli. La seconda e' quella dei prototipi con un prototipo che rappresenta in maniera rappresentativa una categoria.

Teoria degli esemplari vs prototipi
Quella dei prototipi l'abbiamo vista prima;
Esemplari: la rappresentazione mentale di un concetto è l'insieme delle rappresentazioni di alcuni esemplari di quella categoria che noi abbiamo incontrato durante la nostra vita.





While a prototype is an abstract average of the members of a category, an exemplar is an actual member of a category, pulled from memory

WordNet (3-4)

WordNet is an on-line lexical reference system. English nouns, verbs, and adjectives are organized into synonym sets.
WordNet divide il lessico in 4 categorie: nouns, - verbs, - adjectives, and - adverbs.
WordNet organizza informazioni lessicali sulla base del loro significato.:

- nouns are organized in lexical memory as hierarchies
- verbs are organized by a variety of entailment relations (pluralità di relazioni)
- adjectives and adverbs are organized as N-dimensional hyperspaces.

‘‘word form’’ will be used here to refer to the physical utterance or inscription and ‘‘word meaning’’ to refer to the lexicalized concept that a form can be used to express.

Word Meanings	Word Forms F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
⋮				. . .	
M_m					$E_{m,n}$

• if there are two entries in the same column, the word form is polysemous; **Polisemica**
• if there are two entries in the same row, the two word forms are synonyms

Synset
Someone who knows that “board” can signify either a piece of lumber or a group of people assembled for some purpose will be able to pick out the intended sense with no more help than plank or committee. - the **synonym sets**, {board, plank} and {board, committee} can serve as unambiguous designators of these two meanings of board. these synonym sets (synsets) do not explain what the concepts are; **they merely signify that the concepts exist**

Gloss
Sarebbe la descrizione + un caso d’uso (esempio)
Since English is rich in synonyms, synsets are often sufficient for differential purposes. - sometimes, however, an appropriate synonym is not available, in which case the polysemy can be resolved by a short gloss, e.g.,
WordNet is organized by semantic relations. since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets. Le relazioni sono simmetriche.

Relazioni

Sinonimia
The most important relation for WordNet is similarity of meaning, since the ability to judge that lexical relation between word forms.
Due espressioni sono sinonime se sostituendole in una frase il valore di verità non cambia. In verità sinonimi puri sono rari. Quindi si rilassa la relazione vincolandoli ad un contesto preciso (frase) (parole nello stesso synset)

Antonimia
Denota l'opposto: ricco-povero. Si basa sulla word form!!! (a livello di significato può non essere sempre vero: se non sono ricco allora non sono necessariamente povero. posso essere benestante. nonostante ricco-povero siano in relazione di antonimia)
antonymy is a lexical relation between word forms, not a semantic one between word meanings (esiste?bo)

Iponimia
is a semantic relation between word meanings: {tree} is a hyponym of {plant}. hyponymy is transitive and asymmetrical (Word meaning (relazioni fra synset))

Meronimia
part-whole relation is transitive and asymmetrical. is a semantic relation between word meanings ad es ruota-automobile. (Word meaning (relazioni fra synset))

these and other similar relations can be represented in WordNet by pointers (labeled arcs) from one synset to another

Nouns

definitions of common nouns typically give a superordinate term + distinguishing features. three types of distinguishing features are discussed:

- **attributes** (modification),
- **parts** (meronymy),
- **functions** (predication).

Esempio: take one meaning of the noun tree, the sense having to do with trees as plants. conventional dictionaries define this sense of tree by some such gloss as: a plant (superordinate term) that is large, woody, perennial, and has a distinct trunk (distinguishing features).

Cosa manca nei dizionari ma non in Wordnet:

- Le specificazioni futili (es. non spiega che cos'è biologicamente una radice, non dice se un albero ha radici ...)

- Non c'è informazione sugli elementi coordinati (non ho puntatori ai fratelli: non so se esistano altri tipi di piante)
- Non ci sono puntatori verso le sottoclassi

superordinate-hyponym
superordinate nouns can serve as anaphors referring back to their hyponyms: He owned a rifle, but the gun had not been fired, it is immediately.

Gerarchia Multipla

the nouns are partitioned into a set of semantic primes, that select a (relatively small) number of generic concepts and each one acts as the unique beginner of a separate hierarchy. these multiple hierarchies correspond to relatively distinct semantic fields, each with its own vocabulary.

- **25 campi semantici distinti,** collettivamente dovrebbero coprire tutti i termini e i sensi dell'inglese (sono le 25 root di WordNet)

25 unique beginners (nouns)

{*act, action, activity*}
{*animal, fauna*}
{*artifact*}
{*attribute, property*}
{*body, corpus*}
{*cognition, knowledge*}
{*communication*}
{*event, happening*}
{*feeling, emotion*}
{*food*}
{*group, collection*}
{*location, place*}
{*motive*}
{*natural object*}
{*natural phenomenon*}
{*person, human being*}
{*plant, flora*}
{*possession*}
{*process*}
{*quantity, amount*}
{*relation*}
{*shape*}
{*state, condition*}
{*substance*}
{*time*}

although the overall structure of noun hierarchies is generated by the hyponymy relation, details are given by the features that distinguish one concept from another. Es: canary is a bird that is small, colorful, sing and flies.
In order to make all of this information available when canary is activated, it must be possible to associate canary appropriately with at least three different kinds of distinguishing features.

1. Attributes: small, yellow
    a. values of attributes are expressed by adjectives. adjectives are said to modify nouns.
2. Parts: beak, wings
    a. Riprendiamo dalla **relazione di meronimia** (relazione parte-tutto)
    La relazione inversa è nota come **olonimia**
    I meronimi sono i tratti distintivi che gli iponimi possono ereditare, *es becco e ala sono meronimi di uccello e se il canarino è iponimo di uccello allora becco e ala sono meronimi anche di canarino*

La relazione è simmetrica e tendenzialmente transitiva, tendenzialmente perché ad esempio non si dice che la casa ha una maniglia nonostante la maniglia sia meronima della porta che è a sua volta meronima della casa.

3. 3. Functions: sing, fly
    a. a functional feature of a nominal concept is a description of something that instances of the concept normally do, or that is normally done with or to them. all the features of nominal concepts that are described by verbs or verb phrases

Verbi

(Su cui non ci soffermiamo tanto)
Sono il pezzo più rilevante sintatticamente e lessicalmente nel linguaggio (la frase minima contiene il verbo)
E' possibile descrivere i dipendenti del verbo in termini di possibilità e preferenze tramite struttura predicato-argomentali o frame di sottocategorizzazione - **Ha un filtro su # di argomenti e anche su categorie di argomenti.**

Per argomenti del verbo si parla sia di sintassi che di semantica

Il **grado di polisemia dei verbi è maggiore** di quello dei sostantivi visto che il **significato** del verbo è più flessibile e **viene completato dai suoi argomenti**

Sono articolati in **15 alberi inizializzator**i. In un singolo campo semantico, e' frequente il caso in cui non tutti i verbi possono essere raggruppati sotto un singolo unique beginner. Quindi le gerarchie dei verbi hanno una struttura superficiale, e più bassi e larghi rispetto ai nomi

Come sono organizzati dal punto di vista strutturale (per i sostantivi avevamo a relazione di iper iponimia per costruire la gerarchia) qui si ibrida con altri principi tipo **troponimia**
es. passeggiare e camminare non fa funzionare l'iponimia perché "passeggiare è un tipo di camminare" non è accettato dal punto di vista dei parlanti..
Conflations(abstract) : Come mettiamo assieme i verbi? es. verbi motion tramite modo o causa, altri verbi tramite effetto.
Troponimia: il verbo Y è un troponimo del verbo X se nel fare l'attività Y si fa anche la X (come mormorare rispetto a parlare)

Sperimentazioni

passiamo ad un livello più sperimentale di wordnet

Word Sense similarity:

dati in input 2 termini voglio un numero che identifica quanto siano simili questi due termini
La prossimità di senso è un concetto molto “strano”, intuizioni, sono vicino o lontani in base a quanto è lungo il cammino tra i due sensi

Consegna:
Implementare 3 misure di similarità basate su wordnet e poi calcolare gli indici con le misura di Pearson e Spearman (noi le prendiamo come black-box)
L’obiettivo dell’esercitazione è utilizzare wordnet in modo da muoverci ed implementare degli algoritmi di calcolo di cammini tra synset che sono i nodi della rete di wordnet

Attenzione: l’input nel dataset è di termini, ma le funzioni da usare è sui sensi quindi per calcolare la similarità tra 2 termini si prende la massima similarità fra tutti i sensi del primo termine e tutti i sensi del secondo termine
Quindi

nella formula c sono i concetti che appartengono ai synset associati ai termini w_1 e w_2.

$$\mathrm{sim}(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [\mathrm{sim}(c_1, c_2)]$$

Per muoverci usiamo la nozione di albero tranne quando troviamo più iperonimi (si può usare solo uno di quelli dicendolo o percorrendo tutti i possibili cammini e poi scegliere il più breve)

Implementare algoritmi di cammini su alberi implementando le funzioni di similarità di

- Wu & Palmer
- Shortest Path
- Leakcock & Chodorow

Word sense disambiguation WSD

Dato un termine polisemico e un contesto in cui occorre questo termine, individuare in quale senso il termine è inteso..
Esiste un problema di finestratura del testo considerato per disambiguare un termine, qual’è la dimensione minima di questa finestra sapendo che esiste un principio di località? non si sa..

Esistono 2 strade fondamentali per raccogliere informazioni e condurre disambiguazioni:

- **feature collocational,** prendere un certo numero di elementi prima e dopo il termine target, un vettore di collocational feature potrebbe essere:
  [wi−2,POSi−2,wi−1,POSi−1,wi+1,POSi+1,wi+2,POSi+2] e quindi:
  [guitar, NN, and, CC, player, NN, stand, VB]

- **bag-of-word,** insieme di termini presi e destrutturati (perdiamo informazioni sulla sequenza) gli elementi sono in un set (niente duplicati), semplificatoria come modalità ma in molti casi utile. Per farlo si potrebbe prende una serie di termini frequenti: [fishing, big, sound, **player**, fly, rod, pound, double, runs, playing, **guitar**, band] e contare le frequenze nella frase di contesto: [0,0,0,1,0,0,0,0,0,0,1,0]

Lesk Algorithm (ancora usato come base per competizioni)

```
1  function SimplifiedLesk(word,sentence)
2  returns best sense of word
3  best-sense ← most frequent sense for word
4  max-overlap ← 0
5  context ← set of words in sentence
6  for all senses of word do
7     signature ← set of words in the gloss and examples of sense
8     overlap ← ComputeOverlap(signature,context)
9     if overlap > max-overlap then
10       max-overlap ← overlap
11       best-sense ← sense
12    end if
13 end for
14 return best-sense
```

Notare che si inizializza il best-sense al sense più comune
(secondo me ha comunque più senso calcolare max-overlap subito e non metterlo a 0)

Consegna:
Implementare l'algoritmo di Lesk, estrarre 50 frasi dal corpus SemCor (corpus annotato con i synset di WN) e disambiguare almeno un sostantivo per frase e calcolare l'accuratezza del sistema sulla base dei sensi annotati in SemCor
Calcolare l'errore medio che conduciamo randomizzando la selezione delle 50 frasi un certo numero di volte parametrizzato
(La percentuale dovrebbe essere circa al di sotto del 60%)

links:
https://www.nltk.org/howto/wordnet.html
http://web.eecs.umich.edu/~mihalcea/downloads.html (per SemCor)

FrameNet (5-6)

Ipotesi su questo lavoro: Abbiamo già menzionato i frame per la rappresentazione della conoscenza e la rappresentazione semantica, l'ipotesi è quindi che nel nostro modo di caratterizzare i concetti ci basiamo largamente sulla nostra esperienza personale.
I blocchi di informazioni alla base di questa ipotesi sono appunto i frame

FrameNet è un progetto che mira alla costruzione di un lessico e analogamente a wordnet tenta di mettere insieme due tipi di elementi: termini e significati, ma l'ambizione è maggiore: **wordnet associa dei possibili sensi ad un termine (la granularità è del termine singolo)**

Framenet mette insieme dei significati per dar conto di significati di espressioni ed enunciati interamente. -> Ha una granularità maggiore
Questa operazione è fatta registrando informazione su come sintagmi e frasi intere sono costruite a partire dai termini e poi utilizzando un corpus che permette di studiare come si compongono questi significati
E' quindi un oggetto bipartito che contiene da una parte la definizione di un insieme di frame e dall'altra parte un corpus annotato dove si mostra come questi frame vengono realizzati

I frame sono delle situazione stereotipate e sono descrizioni di situazioni a vario livello, all'interno di queste descrizioni di situazioni ci sono un insieme di sensi che sono definiti in relazione al dato frame.

La prima operazione è quella di individuare termini che evocano un dato frame, all'interno di framenet l'idea è che abbiamo a che fare con un oggetto (come un oggetto Java) all'interno dei cui campi desideriamo inserire in maniera ordinata i vari elementi.
La **polisemia** viene risolta non attribuendo un synset id ma individuando degli elementi detti Lexical units **LUs** (**coppia termine-senso**)
Facendo un associazione con wordnet diversi synsets apparterrebbero a diversi LUs

Wordnet restituisce molti sensi tra di loro molto vicini, anche FN fornisce delle differenze di senso abbastanza fini: ho 10 sensi estremamente vicini e devo trovarne uno da restituire è difficile..
Insieme di frame che definiscono delle situazioni stereotipate ciascuno con il suo set di ruoli semantici e di LUs e un corpus in cui questi sono annotati permette di cogliere molte sfumature
Permette quindi **di organizzare i significati in modo estremamente fine facendolo in maniera guidata dal verbo, tenta di organizzare i costituenti del verbo!**

Come dobbiamo muoverci per costruire un frame inerente alla vendetta per esempio?
Si definisce un **vocabolario del frame:**

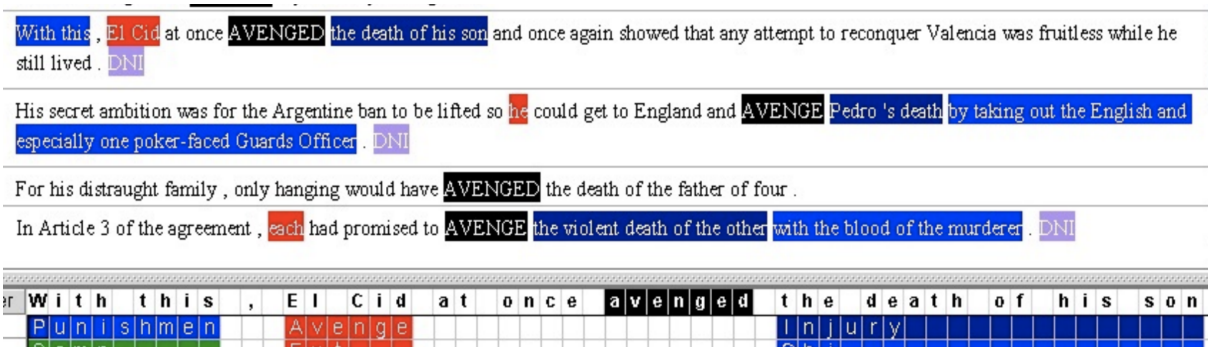
- Nouns: revenge, vengeance, reprisal, retaliation, retribution
- Verbs: avenge, revenge, retaliate (against),get back (at), get even (with), pay back
- Adjectives: vengeful, vindictive
- V+N Phrases: take revenge, exact retribution, wreak vengeance

Dopodichè si fornisce un **frame definition** (usabile come le gloss) e una **lista di frame elements** (ovviamente alcuni di questi sono core)

- FrameDefinition: because of some injury to something-or-someone important to an avenger (maybe himself), the avenger inflicts a punishment on the offender. the offender is the person responsible for the injury.
- FE List:
    - avenger,
    - offender,
    - injury,
    - injured_party,
    - punishment.

I **frame elements FE** sono come dei ruoli semantici, vengono sono usati per annotare le parole della frase

Il secondo step consiste nella costruzione di un corpus di frasi di esempio annotato con i riferimenti a questi frame, l'annotazione mette insieme evidenze sintattiche e semantiche



Quindi, i **predicati** (es 'avenged' della prima frase) sono quelle parole che evocano dei frame (come la parole vendetta) e creano quindi il contesto in cui riempire le informazione del frame, e i **fillers** (es 'the death of his son' della prima frase)**,** elementi che soddisfano le necessità semantiche dei ruoli evocati dai predicati per riempire gli slot

In nero predicato, altri colori filler -> parole che riempiono i ruoli semantici = FE

Why do we use frames
• Frame descriptions –(which some people use for situation ontologies)
• Lexical entries –(which some use for lexicon building)
• Annotations –(which some use as training corpora for machine learning)

POSSIBILI APPLICAZIONI

- **WSD** tramite ricerca diretta di lexical unit e guardando a che frame ognuno appartiene
- **composizione semantica**, cioè l'integrazione di informazioni associate con i dipendenti semantici in frames evocati dai loro reggenti semantici;

- **attivazione di un vocabolario topic-related** per il riconoscimento e la selezione del senso nelle varie parti in cui si articola un testo

ConceptNet (7)

Conoscenza di senso comune
Commonsense reasoning tool-kit
Risorsa lessicale che contiene tante asserzioni su fatti e sul mondo, la gran parte degli approcci basati su keyword o statistici hanno un approccio troppo superficiale non cogliendo alcuni aspetti salienti della realtà in cui si muovono le app NLP (posso dire che due parole co-occorrono ma non perché)

Senso comune: Insieme di elementi di conoscenza di base che sono costitutivi del nostro sistema di rapportarci col mondo (banalità come per aprire una porta ruotare la maniglia, normalmente queste banalità non sono menzionate nella KB dei sistemi)

Che tipo di conoscenza? Spaziale fisica sociale temporale aspetti psicologici quotidiani
E' intrinsecamente sfuggente e normalmente **è omessa dalle comunicazioni sociali** nonostante influenzi le comunicazioni..

OpenCyc è un'ontologia di alto livello stile enciclopedico ed è ottimizzato per il ragionamento logico (essendo un ontologia) -> consistenza del mondo asserito e verifica della classe di un nuovo elemento inserito.
To use Cyc to reason about text, it is necessary to first map the text into its proprietary logical representation, described by its own language CycL. However, this mapping process is quite complex because all of the inherent ambiguity in natural language must be resolved to produce the unambiguous logical formulation required by CycL.

ConceptNet fa inferenze semplici di natura contestuale su elementi che riguardano il mondo ed è una risorsa universale (non riguarda un singolo dominio)
ConceptNet è una rete semantica che contiene tanti archi che connettono 300 000 nodi
Questi nodi non sono esattamente specificati (cook è cuoco ma anche cucinare) e differenza di un nodo di wordnet in cui avevamo un synset con uno o più termini che caratterizzano la sua lessicalizzazioni.

Open Mind Common Sense (OMCS) initiative ha portato alla costruzione di questo corpus con 700 000 frase in inglese contenenti senso comune

In ConceptNet ci sono in totale 20 relazioni semantiche (rispetto a synonym, is-a and part-of di wordent)

Tutti i sensi (che in wordnet erano più divisi possibili) qua invece sono tutti assieme, e a ciascun senso è associato un termine che varia per ciascun senso
e.g bank -> money oppure bank -> river

Per le relazioni invece si catturano anche concetti appartenenti ad una sfera emotiva (?) oltre a classiche causa, effetto, gerarchia ecc..

La KB è di tipo **rivedibile** (collochiamo nel mondo in cui la conoscenza è rivedibile, parzialmente rumorosa, mancante come la conoscenza umana)
Es. meaning it describes something that is often true, but not always, e.g. EffectOf('fall off bicycle', 'get hurt')

- Wordnet è comoda per fare **lexical categorization** and **word-similarity determination**
- Cyc (ontologia senso comune con strutturazione logica) **deductive reasoning** in contesti specifici e non ambigui
- ConceptNet fare **ragionamento di tipo contestuale** over real-world texts

applicazioni di contextual-reasoning:
- dato un testo fare delle previsioni sul mood della storia, **continuazioni possibili** ecc..
- disambiguazione: vedo in che regione si colloca un termine quali sono gli elementi adiacenti per condurre una disambiguazione elementare
- analogie, la conoscenza di conceptnet è associativa

Fase di estrazione

Per costruire ConceptNet hanno usato una 50ina di regole di estrazione dal OMCS corpus, tramite regex essendo OMCS semi-strutturato.
I nodi di ConceptNet come risultato hanno una struttura garantita ed è una combinazione dei 4 costruttori sintattici: verbs, noun phrases, prepositional phrases, adjectival phrases

Fase di rilassamento

le asserzioni duplicate venivano merged (fornendo anche una frequenza di quella relazione) e lifting della conoscenza estratta in modo di portarla più in alto possibile della gerarchia (es banana è dolce, fragola è dolce ecc. -> frutto è dolce)

Cosa possiamo fare con questi nodi

- **Contextual neighborhoods** Relatedness misurabile tramite link path ma anche la densità di nodi nel loro percorso (tramite # di nodi e peso dei possibili percorsi che connettono due nodi).
- **Analogia**, due nodi possono essere considerati analoghi se i loro archi entranti sono sovrapposti (es mela e ciliegia hanno stessi archi)
- **Projection**: LA in california che è in america che è nel mondo (segui un arco dello stesso tipo che esprime una relazione transitiva)
- **Disambiguazione & Classificazione**
    - An exemplar document is fed into a function that computes the **contextual-regions** they occupy in the ConceptNet semantic network.
    - New documents are classified or disambiguated into the exemplars by calculating the nearest neighbor

- **Analisi di affetti**: individuazione delle categorie emotive, dividi conceptnet in queste categorie e fai più o meno quello di prima

(Le ultime due non sono troppo efficaci)

Compound word (ciò che c'è in un nodo): unico termine a cui sono associati più termini (cook), più è lungo e informativo il compound word e meno è ambigua l'espressione

Cover (7)

Sviluppata da UnitTo per garantire descrizioni di senso comune ai sostantivi a partire da BabelNet + Nasari

- BabelNet

BabelNet è una risorsa che è stata proposta per portare in un contesto multilinguistico e corrisponde ad un porting di wordnet unito a wikipedia
Ha quindi una marea di Named Identity (nomi propri di persone e aziende)
E' stato acquisito in maniera quasi del tutto automatica -> tanti più elementi ma meno puliti…
La struttura è la stessa di Wordnet!!

- Nasari

Vettorializzazione dei BabelSynsets Id
Ci sono tre tipi: **lexical**, **unified** e **embedded**
Unified: vettori fisionomia a: testa (synset che si sta definendo), il synset di wordnet corrispondente, il nome su wikipedia e un elenco di altri vettori avente un peso che indica quanto è in relazione con la testa

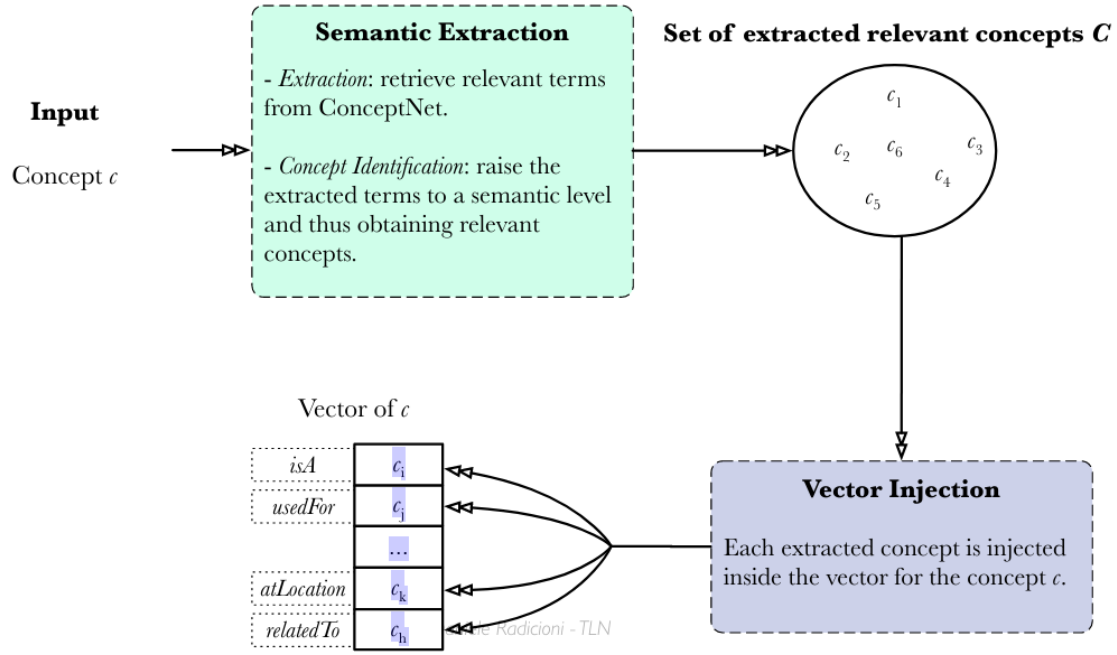
```
bn:00011639n    {board, plank, Plank (wood), ...}
wn:15101854n
Plank (wood)

bn:00052293n    [343.35]    {Timbered, 2x4 wood, ...}
bn:00011639n    [289.82]    {board, plank, ...}
bn:00081492n    [249.42]    {wood, sapwood, ...}
bn:00013077n    [201.57]    {bridge, span, ...}
bn:00074531n    [126.16]    {Strake, Wale, ...}
bn:00077259n    [112.47]    {timber}
bn:00008691n    [104.7]     {barrel, cask, ...}
```

[2]Collados, Pł

- ConceptNet

In ConceptNet tutti i possibili sensi di un termine sono assieme a differenza di BabelNet e Nasari

Obtaining Common-sense Representations



In Cover abbiamo un vettore per un concetto dove ad ogni entry del vettore abbiamo un synset inerente alla relazione a quella entry

Example of the 'fork' vector

relatedTo	*tool, food, utensil, cutlery, eating*
isA	*tool, cutlery, utensil*
atLocation	*table, desk, plate*
usedFor	*eating*
madeOf	*metal*

Le relazioni derivano da ConceptNet (le più utilizzate)

Procedimento

0) Input of the system
The input of the system is the concept c bn:00035902n, the BabelSynsetId of fork intended as "a utensil used for eating or serving food" This is a sample of its NASARI vector:

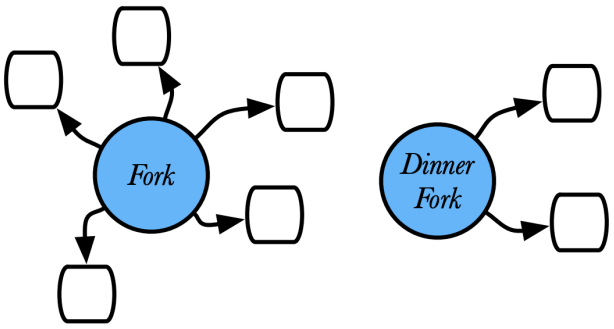
bn:00024649n {tableware, …}
bn:00049322n {knife, …}
bn:00073547n {spoon, …}

1.1) Semantic Extraction - Extraction step (nodes retrieval)

bn:00035902n

{Fork, King of utensils, Pickle fork, Fish fork, Dinner fork, Chip fork, Beef fork}

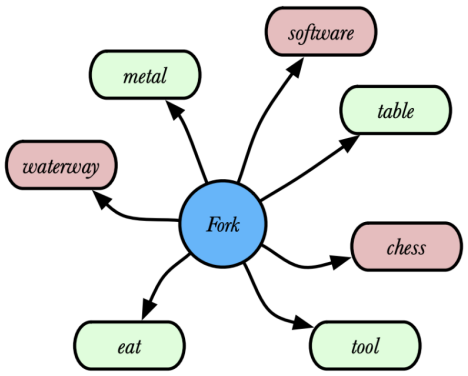
Retrieve all ConceptNet nodes and relationships for each lexicalization in the input's synset:



1.2) Semantic Extraction - Extraction step (relevance detection)

For each retrieved node:

Determine if the term is relevant or not:

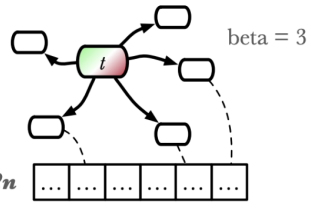


A term t in $\mathcal{T}$ is relevant if either:

t is a synset term of one of the elements inside the input's NASARI vector:

bn:00035902n | … | … | bn:_n{*t*, …} | … | … |

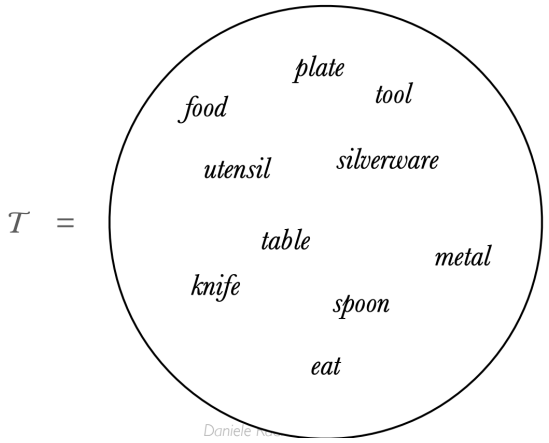
at least beta elements around t in ConceptNet node are elements inside the input's NASARI vector:



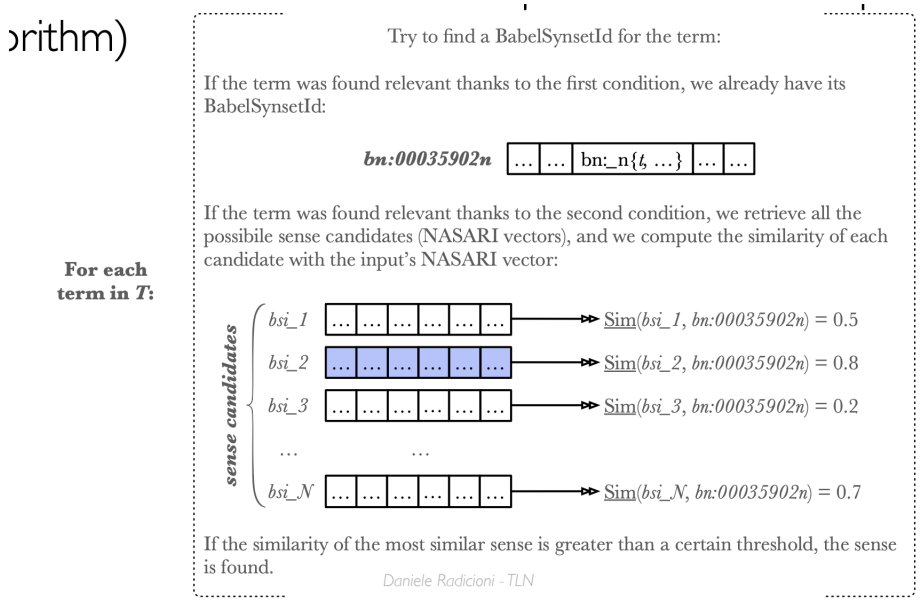
bn:00035902n | … | … | … | … | … | … |

1.3) Semantic Extraction - Extraction step (output)

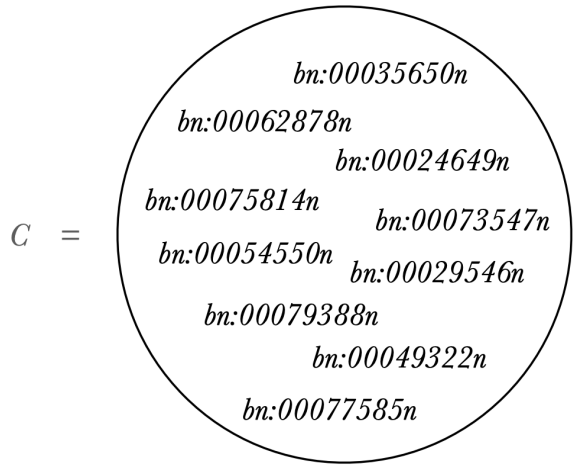
The output of this step is the set T of relevant extracted terms:



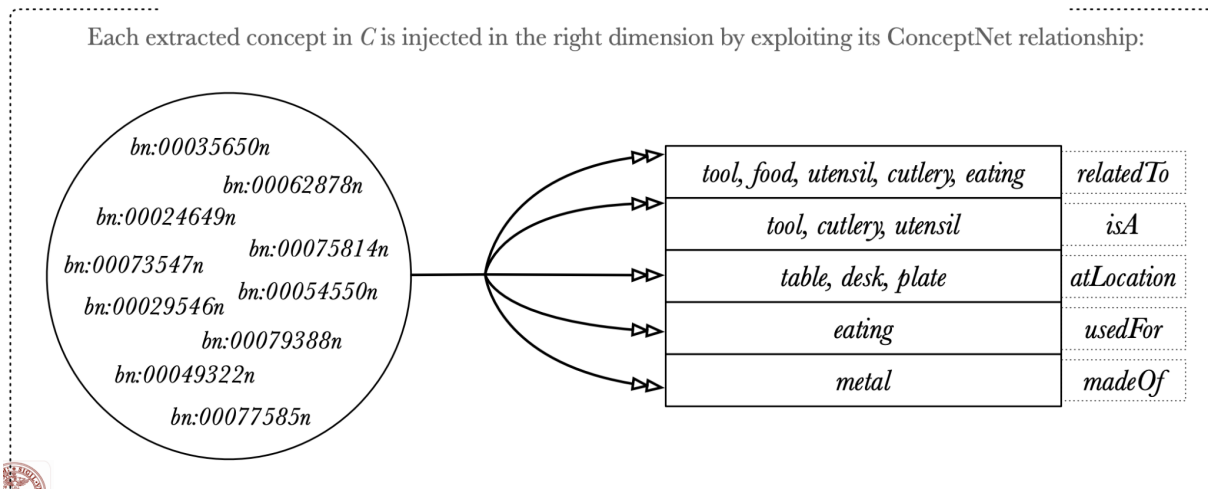
2.1) Semantic Extraction - Concept Identification step (algorithm)



2.2) Semantic Extraction - Concept Identification step (output)



3) Vector Injection



Il concetto di similarità può passare invece che dal percorso passare da quanti elementi in comune hanno quei due vettori associati ai synset.
Il sistema può anche quindi fornire una spiegazione associata al punteggio di similarità restituito

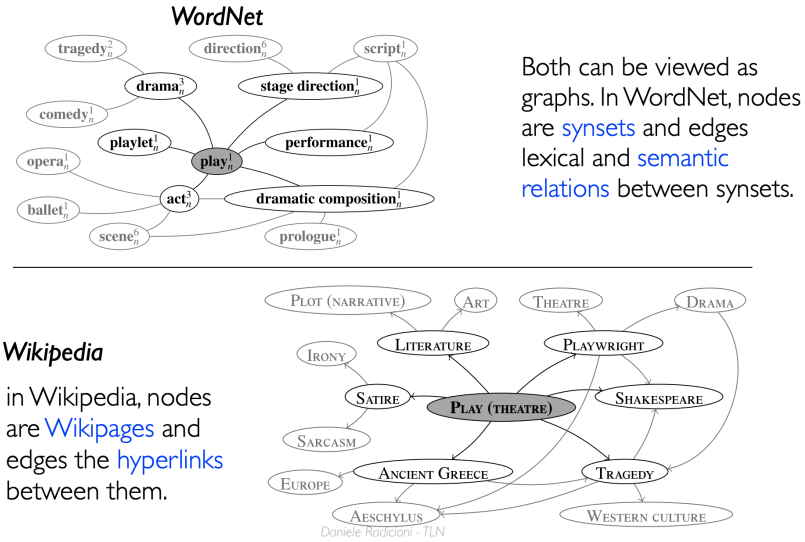
BabelNet (8)

E' un allargamento di wordnet e consiste nel porting multilinguistico dei synset
L'idea: c'è un sacco di conoscenza lessicale nella rete in forme diverse (tesauri, glossari, dizionari online, lessici e ontologie computazionali) oltre a wordnet e cyc, l'altra osservazione è che la costruzione di risorse come wordnet (sviluppata all'uni di Prinston) richiedeva un sacco di cash. Oltre a questi problemi, ogni lingua richiede di ripartire da capo usando la struttura definita e sviluppata e costi ulteriori per connettere le varie risorse.

L'idea è di sovvertire questo meccanismo introducendo dei processi automatici partendo da wikipedia, dizionario enciclopedico facendo un merge tra wordnet e wikipedia

Il titolo di una pagina in wikipedia può contenere informazione per disambiguare il termine titolo in caso sia ambiguo (e.g. Play(theater))
Il testo in wikipedia è parzialmente strutturato, la parte strutturata di wikipedia (le info box) costituisce DBpedia
I links di wikipedia possono essere delle:

- redirezioni
- disambiguazione, contiene dei link per dei concetti
- internal links
- inter-language links per saltare nelle pagine di wikipedia in altri linguaggi
- categoria



Babelnet è un grafo diretto costituito da un insieme di nodi e archi che connettono nodi
Ogni nodo (Bebel synset) contiene lessicalizzazioni del concetto per lingue differenti.

BabelNet encodes knowledge as a labeled directed graph G = (V,E) where
- V is the set of nodes (i.e., **concepts** such as play and **named entities** such as Shakespeare)
- $E \subseteq V \times R \times V$ is the set of edges connecting pairs of concepts (e.g., play is-a dramatic composition). Each edge is labeled with a semantic relation from R, i.e., {is-a, part-of, ..., ε}, where ε denotes an unspecified semantic relation.
- each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., {playen, Theaterstückde, drammait, obraes, ... , pièce de théâtrefr}.
- We call such multilingually lexicalized concepts Babel synsets

Building the graph
to build the BabelNet graph, the information collected is:
- From WordNet, all available word senses (as concepts) and all the lexical and semantic pointers between synsets (as relations);

- From Wikipedia, all encyclopedic entries (i.e., Wikipages, as concepts) and semantically unspecified relations from hyper-linked text.
WordNet and Wikipedia can overlap both in terms of concepts and relations - in order to provide a unified resource, the intersection of these two knowledge sources is merged.
Next, to enable multilinguality, the lexical realizations of the available concepts in different languages are collected.
Finally, multilingual Babel synsets are connected by establishing semantic relations between them.

NASARI (8)

Una delle prima risorse vettoriali che hanno modificato il funzionamento della semantica lessicale negli ultimi tempi.
Mette insieme significati diversi di un termine in un vettore unico
due modalità di rappresentazione dei vettori:

- vettori lessicali: le feature sono termini
- vettori unified: le feature sono costituite da id di senso di BabelNet

Semantic similarity: Weighted Overlap
se due vettori contengono un certo elemento ed è inserito all'inizio del vettore -> più rilevante nella descrizione dell'elemento, rango ^-1 coglie questa relazione semantica

$$WO(v_1, v_2) = \frac{\sum_{q \in O} \left(rank(q, v_1) + rank(q, v_2) \right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}$$

v1 e v2 sono rappresentazioni di senso indicizzate sulla base del babel synset
La similarità tra due termini è data da:

$$sim(w_1, w_2) = \max_{v_1 \in \mathcal{C}_{w_1}, v_2 \in \mathcal{C}_{w_2}} \sqrt{WO(v_1, v_2)}$$

Per ogni termine oppure synset associati ad un vettore viene specificato il peso

Riassunto automatico - Summarization

2 tipi:

- RIassunto di tipo **estrattivo**: vengono prelevati pezzi del documento che vengono riutilizzati
- Riassunto **astrattivo**: c'è un passaggio in più di rappresentazione interna del significato del testo e a partire da questa rappresentazione si ricrea il testo tramite uno step generativo

Può essere user-focused, variare in base alla conoscenza preliminare del lettore

Approcci:

- Superficiali, estrattivo
- Profondi, mirano a cogliere la semantica della frase a livello profondo

- Single document
- Multiple document (es diverse news che trattano dello stesso argomento e si vuole generare un'unica summarisation per descrivere tutte le news lette)

Parametri:

- Compression rate (summary length/source length)
- Audience (user-focused vs. generic)
- Relation to source (extract vs. abstract)
- Function (indicative vs. informative vs. critical)
- Coherence: the way the parts of the text gather together to form an integrated whole

Criteri per stabilire la rilevanza di elementi di un testo (paragrafi, frasi o insieme di frasi)

- **posizione all'interno del testo**, frasi importanti sono collocati in punti specifici, **lead-based summary**
- **titolo**, il titolo dei documenti generalmente ci dicono di cosa tratta, si può creare una lista di termini presenti nel titolo e utilizzare questi elementi come keyword per cercare frasi rilevanti
- Optimum Position Policy, le frase più importanti sono in luoghi dipendenti dal genere (noti a priori o oggetto di addestramenti)
- Cue phrases, sono locuzioni/modi di dire -> the purpose of this article is .. (sappiamo già che quello che c'è dopo è saliente, lo prendiamo o ci diamo un peso)
- Cohesion-based methods, si vuole enfatizzare e riconoscere come più salienti elementi più connessi, bisogna studiare attraverso il lessico quanti paragrafi sono connessi agli altri e analizzare quei paragrafi altamente connessi oppure misure come word similarity

Esercitazione 3...

Lezione 9

Rappresentazione del senso dei verbi, tratteremo VerbNet e PropBanks

Perchè rappresentazione semantica?
Es. un sistema che risponde a delle domande non costruendo una rappresentazione dei documenti che rappresentano la sua KB ma cercando termini in questi documenti, è meglio cercare dei pezzi di informazione inerenti alla domanda, noi stiamo cercando chi ha creato cosa, dove il cosa è il primo vaccino contro la poliomielite

- Q: Who created the first effective polio vaccine?

- A1: [Becton Dickinson$_{agent}$] created [the first disposable syringe$_{theme}$] for use with the mass administration of the first first effective polio vaccine

- A2: [The first effective polio vaccine$_{theme}$] was created in 1952 by [Jonas Salk$_{agent}$] at the University of Pittsburgh

Individuare il tema e l'agent di una azione è lo scopo delle risorse che vedremo oggi

Ipotesi: il comportamento di un verbo è in larga misura determinato da suo significato, il livello sintattico riflette l'insieme dei significati e la semantica complessiva di quel verbo.
Il fatto che la sintassi rifletta la semantica è un punto di partenza largamente condivisa.

VerbNet

E’ una rete di verbi che collega pattern sintattici e aspetti di significato, una rete in cui gli elementi verbali sono rappresentati gerarchicamente e indipendente dal dominio. I lessici sono mappati in altre risorse lessicali come WordNet PropBank e FrameNet.
E’ organizzata in classi verbali (insieme di verbi) e ciascuna classe è descritta da:

- **Ruoli tematici**, i vari ruoli che caratterizzano il verbo: il tema
- **Selectional preferences**, ciascun ruolo ha degli elementi che possono/devono/non devono caratterizzarlo (es mangiare richiede un essere vivo e senziente e un oggetto commestibile)
- **frames** consisting of a syntactic description and a semantic representation with subevent structure

Caratteristiche di VerbNet:

Lessico verbale computazionale e domain-independent
Fornisce un associazione chiara tra livello sintattico e livello semantico, si prende come argomento la struttura degli argomenti del verbo, i frame sintattici e le selectional restriction che caratterizzano i vari argomenti, elenca i predicati semantici e poi potenzialmente può essere legata a wordnet, framenet ecc..

Classi

Lexical verb class: set of verbs that exhibit shared semantic behavior and similar morpho-syntactic patterns

- Useful to generalize properties to a cross-linguistic setting

Alternation

Variazioni nell’espressione degli argomenti con implicazioni semantiche o meno

- Locative alternation (riguarda i complementi di luogo)
    - Sharon sprayed water on the plants
    - Sharon sprayed the plants with water
    - The farmer loaded apples into the cart
    - The farmer loaded the cart with apples
- causative/inchoative Alternation
    - Tony broke the window | The window broke
- Middle alternation
    - Tony broke the crystal vase | Crystal vase break easily

Alcuni verbi con significati simili possono non supportare le stesse alternations

Verbs have complex meaning: key components can be made explicit

- have participants
- space
- verbs represent processes/events/states which are located in time
- can be subdivided into sub-parts to capture during, end, results

Levin classes

- Verbs are grouped into classes
- Each class is characterized by a set of syntactic patterns
    - John broke the jar / The jar broke / Jars break easily
    - - John cut the bread / *The bread cut / Bread cuts easily
    - - John hit the wall / *The wall hit / *Walls hit easily
- Hypothesis: syntax reflects implicit semantic components
- Levin classes are not semantically homogeneous
- classes are not completely syntactically homogeneous
- verbs can be in multiple class listings
- alternation contradictions
    - Carry verbs disallow conative but include {push, pull, shove, etc.} also in Push/pull class which does take conative

Hypothesis: **syntax reflects implicit semantic components**

- contact, directed motion,
- exertion of force, change of state

Le classi non sono semanticamente o sintatticamente omogenee e i verbi possono essere presenti in diversi classi

Tripartizione della struttura eventiva a cui è possibile rapportare un certo numero di verbi. Questa tripartizione considera 3 tipologie di eventi:

- processi di tipo preparatorio
- elementi di culminazione (tipicamente sono elementi puntuali)
- stato conseguenza all'azione

Nel primo tipo abbiamo verbi come run,jog,hop.. hanno a che fare con attività rappresentate come processi
Nel secondo verbi come hit che hanno una caratteristica puntuale (uno che colpisce o calcia è un tipo di azione che non ha una durata, è un punto nel tempo)
Nel terzo verbi che hanno a che fare con uno stato che segue (es il rompersi di qualcosa)

VerbNet class entries

- Verb classes based on Levin's classification (quindi divide le classi in base al pattern sintattico)
- classes defined by syntactic properties (alternations)

- capture generalizations about verb behavior
- for each verb class
    - thematic roles
    - selectional restrictions for the arguments in each frame
    - syntactic frames
    - each frame includes semantic predicates with a time function

Predicati semantici

Ogni frame include predicati semantici con una funzione temporale e può essere espressa grazie ad E, variabile eventiva associata a ciascun predicato semantico associato a verbNet e permette di specificare quando all'interno dell'evento il predicato è vero

- **start(E)** for preparatory stage,
- **during(E)** for the culmination stage, and
- **end(E)** for the consequent stage.

HIT class

Class	hit-18.1		
Parent	—		
Members	bang (1,3), bash(1), batter(1,2,3), beat(2,5), ..., hit(2,4,7,10), kick(3), ...		
Themroles	Agent Patient Instrument		
Selrestr	Agent[+int_control] Patient[+concrete] Instrument[+concrete]		
Frames	Name	Syntax	Semantic Predicates
	Transitive	Agent V Patient "Paula hit the ball"	cause(Agent, E) ∧ manner(during(E),directedmotion,Agent) ∧ !contact(during(E), Agent, Patient) ∧ manner(end(E),forceful, Agent) ∧ contact(end(E), Agent, Patient)
	Transitive with Instrument	Agent V Patient Prep(with) Instrument "Paula hit the ball with a stick"	cause(Agent, E) ∧ manner(during(E),directedmotion,Agent) ∧ !contact(during(E),Instrument,Patient) ∧ manner(end(E),forceful, Agent) ∧ contact(end(E), Instrument,Patient)

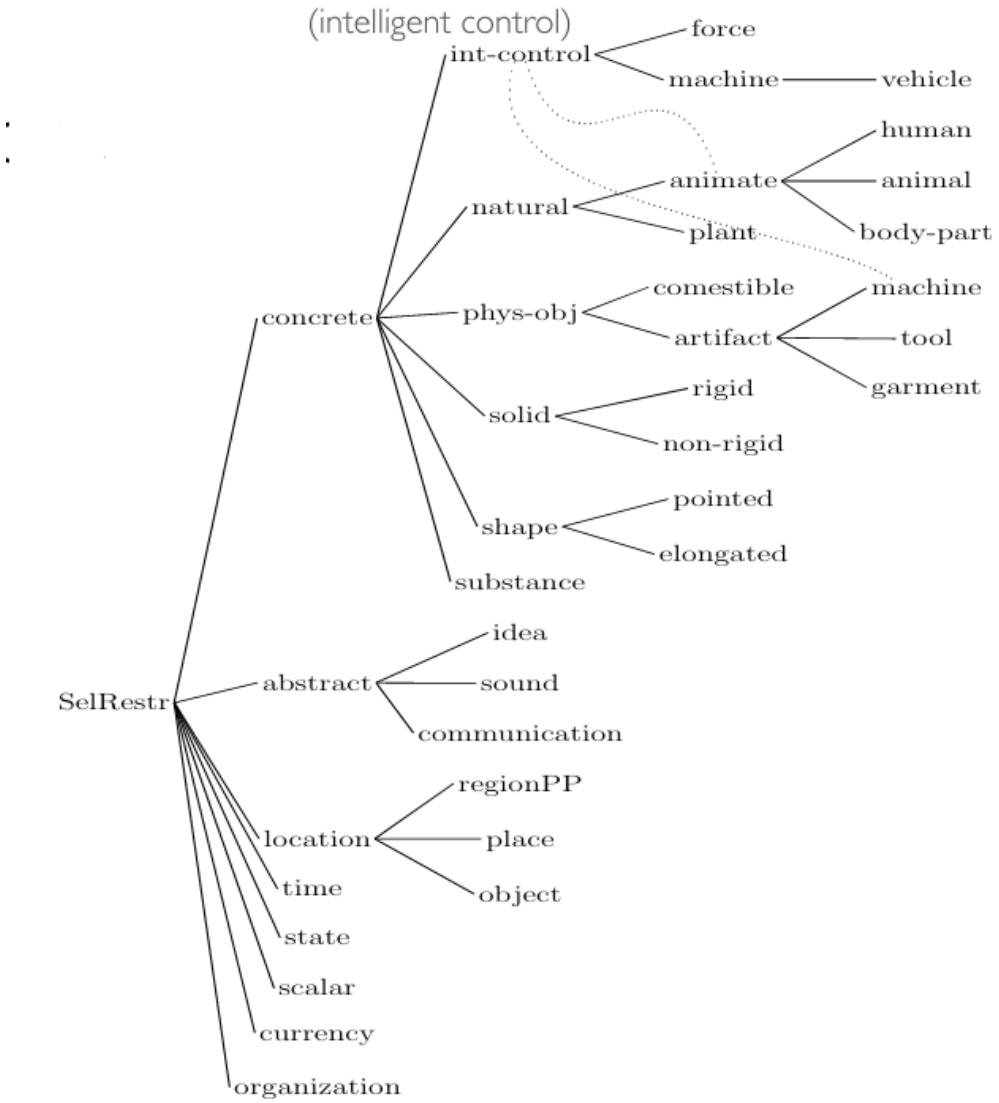
Struttura delle classi

Ruoli Tematici:

Servono per precisare il comportamento e gli elementi partecipanti coinvolti nei vari tipi di azioni, possono aiutare a differenziare le classi.
In VerbNet sono una 40ina , sono utilizzati attraverso le classi e alcuni richiedono delle caratteristiche specifiche per essere presenti (es patient è quello che subisce un certo tipo di cambiamento). *Actor Agent Attribute Beneficiary Cause Location Destination Source*.. sono i più importanti Actor è uno pseudo-agent

Selectional Restrictions:

Funzionano come una tassonomia, permette di specificare preferenze e/o vincoli



Syntactic Frame:

Strumenti che permettono di descrivere le realizzazioni superficiali per i vari verbi che appartengono ad una certa classe, alcuni esempi di costruzioni sono transitive intransitive e resultative e un **set di Levin's alternations**

Examples:

1. Agent V Patient (John hit the ball)
2. Agent V at Patient (John hit at the window)
3. Agent V Patient[+plural] together (John hit the sticks together)

Semantic Predicates:

Congiunzione di stemants che ci permettono di indicare la caratterizzazione semantica, relazioni che coinvolgono i ruoli tematici e sono temporalmente definite attraverso l'evento E
Aspect captured by the temporal function present in the predicates:

- activities (e.g., run) have during(E)

- bounded activities (e.g., hit) have during(E) and end(E)
- accomplishments (e.g., break) have result(E)

PropBank

RoleSet: insieme di ruoli che corrispondono ad un utilizzo distinto di un verbo e può essere associato ad un insieme di frame sintattici che permettono di individuare delle possibili variazioni sintattiche consentite nell'espressione di quel set di ruoli.

- Idea: c'è una firma del significato del verbo, e questa firma è sintattica, noi guardando la sintassi e come è realizzato sintatticamente una frase con un certo verbo riusciamo a capire da quella firma sintattica (roleSet) il significato attivato da quel verbo.

FrameSet: RoleSet a cui è stato associato un frame (sono stati specificati dei ruoli per gli arg)
Verbo polisemico = possiede diversi frameset, es.

Example

- [John$_{ARG0}$] rings [the bell$_{ARG1}$] ← **ring.01**
- [Tall aspen trees$_{ARG1}$] ring [the lake$_{ARG2}$] ← **ring.02**

ring.01	Make sound of bell
Arg0	Causer of ringing
Arg1	Thing rung
Arg2	Ring for

ring.02	To surround
Arg1	Surrounding entity
Arg2	Surrounded entity

Motivi per cui 2 frame set vengono considerati distinti:

- Syntactic-semantic criteria go into this
- Alternations which preserve verb meanings, such as causative/inchoative or object deletion are considered to be one frameset only.
- Verb-particle combinations are always distinct framesets (ad es. set-in set-off è meglio metterli in due frameset distinti)

Hanno una granularità meno fine rispetto a wordnet, un frameset può corrispondere a più synsets.

Differenze con framenet:

- Costruire alternation simmetriche-asimmetriche in propbank, essendo descritti da proto-ruoli (arg0, arg1) questi sono più intercambiabili come valori che assumono rispetto a i FE di Framenet che sono più specifici
- qua ci sono dei trattamenti per le tracce per le ellissi (elem che non vengono menzionati esplicitamente ma fanno parte del contesto) (?)

Mega riassunto di propbank:

E’ un tentativo diverso rispetto a VerbNet, invece di menzionare esplicitamente i tematic role il set di elementi che caratterizzano ciascun verbo è costituito da un insieme di dipendenti -> no classi verbali ma verbi che possono essere raggruppati sulla base di sistemi di dipendenti, ma in linea di principio sono individuati ciascuno sulla base dell’insieme dei dipendenti. Questi dipendenti sono una decina di argomenti (arg0-arg4 per gli arg del verbo e poi altri modificatori per specificare luogo durata tempo e negazioni)

Lezione 10

Semantica lessicale
Ramo di NLP che studia la rappresentazione e l’utilizzo dei sensi delle parole (cerca di capire quali sono i sensi delle parole nelle frasi)
Perchè è utile studiare i sensi? Alta interazione con le macchine
Ovviamente ambiguità (White house) e polisemia (bank) rendono complicato questo task

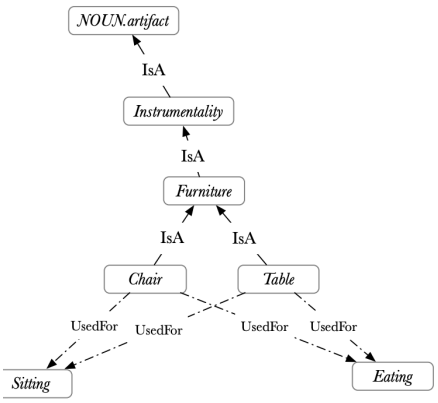
Risorse lessicali
Composizione di uno o più db che definiscono il nostro vocabolario, contengono lexical items, se sono multilingue questi items la risorsa è multilingua
Con le risorse noi risolviamo il problema (Bah)

Reti Semantiche VS Distributional resource (Word embeddings)

Distributional resource
Termini che occorrono in un contesto simile hanno significato simile (nozione di senso basato sulle co-occorrenze)
(Tutti i sensi di una parola in un singolo vettore.. bad..) <- **Problema**
Entra in gioco la frequenza di un significato della parola (sedia vista molte più volte come sedia classica che come sedia elettrica, il vettore tenderà a cogliere il significato di sedia classica)

Knowledge Representation

Semantic Networks



chair#4, death chair#1, hot seat#1; (an instrument of execution by electrocution; resembles an ordinary seat for one person)

Distributional

Distributional hypothesis: terms that occur in a simila context tend to assume a similar meaning

I forgot the keys on the *chair*

I forgot the keys on the *table*

Word embeddings						
chair	0.5	0.7	0.4	0.1	…	0.9
table	0.2	0.9	0.1	0.4	…	0.7

Pro and Cons

Semantic Networks

- High specificity
- Multi-language
- Difficult comparison between senses
- Difficult integration in downstream applications

Distributional

- Meaning conflation deficiency
- (mostly) Single language
- Easy to compare senses
- Easy integration in downstream table applications

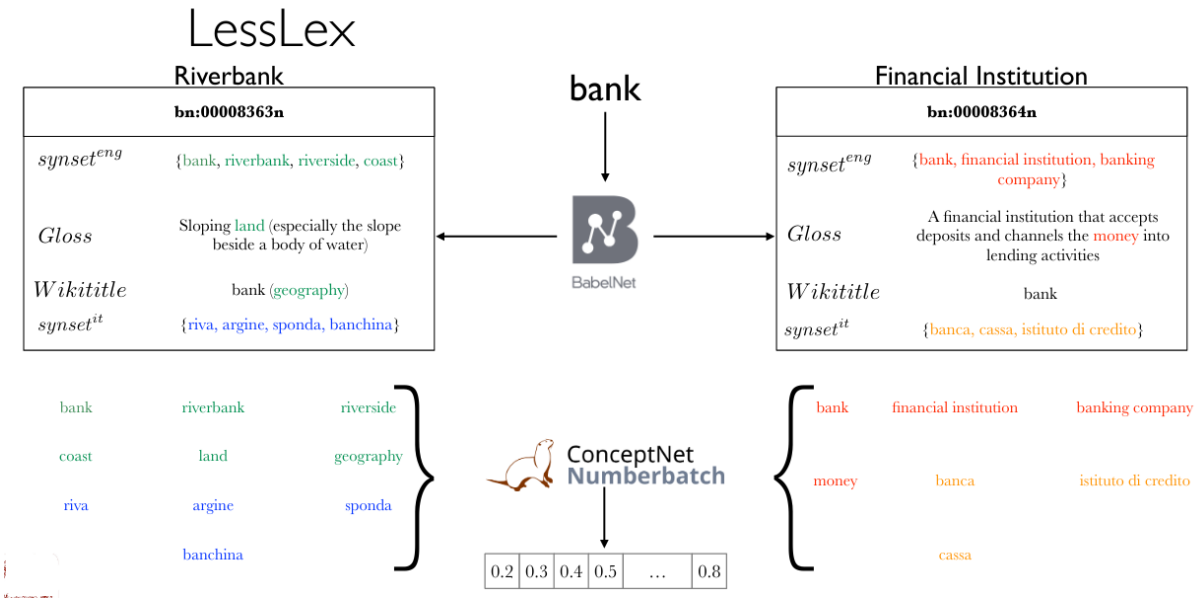
- BN is a wide-coverage multilingual semantic network built by integrating WordNet and Wikipedia.
- ConceptNet Numberbatch is a set of precomputed word embeddings, obtained by integrating Word2Vec, GloVe and the common-sense knowledge provided by CN
- CN is a multilingual, domain-independent knowledge graph that connects words and phrases of natural language;

Lesslex

Vorremmo avere una risorsa in cui le unità lessicali sono i synset ma la rappresentazione definita in vettori (specificità di BabelNet + facilità di confronto di Vettori distribuzionali)

Come fare?
Accedo a BabelNet con il termine *Bank* e ottengo i suoi significati (synset)

Si estraggono i termini interessanti dai significati, cerchiamo il vettore per ognuno di questi termini, otteniamo una serie di vettori per ogni senso
Facciamo la media di tutti questi vettori per cogliere il vettore inerente a quel senso



Semantic Similarity

Semantic Similarity task: where two terms or senses are compared and systems are asked to provide a numerical score expressing how close they are; systems' output is compared to human ratings (nine datasets in mono, multi, cross-lingual settings were used).

Dati due termini eg teacher and student, associamo un punteggio di similarità andando a selezionare un senso per ogni termine.

$$\text{Similarity}(s_i, s_j) = \cos(\theta) = \frac{s_i \cdot s_j}{||s_i|| \cdot ||s_j||}$$

MAX

1 - bn:02193088n - The Teacher (film)	1 - bn:02935389n - Student (film)	0.81
2 - bn:00046958n - teacher (person)	2 - bn:00029806n - student (learner)	0.61
3 - bn:17408133n - Teacher (song)	3 - bn:03259763n - Student (magazine)	

Quando noi essere umani facciamo similarity task per dare un valore di similarity associamo a questi dei significati!
Nel dataset questo non viene specificato
Obiettivo di questo task è individuare il significato associato ai termini, non emulare lo score di similarity

Ranked Semantic Similarity

$$\text{rnk-sim}(t_1, t_2) = \max_{\substack{\vec{c}_i \in s(t_1) \\ \vec{c}_j \in s(t_2)}} \left[\left((1-\alpha) \cdot (\text{rank}(\vec{c}_i) + \text{rank}(\vec{c}_j))^{-1} \right) + \left(\alpha \cdot \text{cos-sim}(\vec{c}_i, \vec{c}_j) \right) \right]$$

Nuovo Task

Sense Identification Task

• The Sense identification task: two terms are compared (e.g., <greeting, wave> , and <weather, wave>) and systems are asked to indicate which senses are mostly intended when considering the word pair; systems' output is then compared to judgements.

Max Similarity

given a term pair <t,u>, the classical approach is to compute the semantic similarity as the maximal cosine similarity featuring all sense combinations <s^t,s^u>

$$\mathcal{M}\text{-sim}(t, u) = \max_{\substack{s^t \in S^t, \\ s^u \in S^u}} \left(\text{cos-sim}(s^t, s^u) \right)$$

retrieving the selected sense pair <s^t,s^u> involves finding the sense pair that maximizes this expression.

$$\langle s^t, s^u \rangle \leftarrow \arg\max_{\substack{s^t \in S^t, \\ s^u \in S^u}} \left(\text{cos-sim}(s^t,\ s^u) \right)$$

Si usa semeval che dato un termine comprende una lista di bn_id che ne rappresentano i vari sensi di quel termine.

Sense Identification Task

Abbiamo bisogno di costruire un nuovo DataSet.. (partendo da quello che già esisteva e aggiungendo i sensi con cui i termini sono intesi nella coppia considerando il punteggio di similarità)

$$\langle s^t, s^u \rangle \leftarrow \underset{\substack{s^t \in S^t, \\ s^u \in S^u}}{\arg\max} \left(\text{cos-sim}(s^t, s^u) \right)$$

Posso fare così ma non ho informazioni sullo score
Inoltre due sensi vicini darebbero score alto anche se non sono quelli venuti in mente all'annotatore

Anche qui Ranked similarity ma non ho di nuovo capito
E più raffinatamente **Neighborhood similarity**: una volta trovato un senso (che è un vettore ormai) si fa cluster con i vettori vicini e si considera il vettore centroide per calcolare la cos-sim)

Lezione 11

Principio di composizionalità

Già spesso evocato, dice che **il significato delle espressioni linguistiche è una funzione dei costituenti e delle regole sintattiche che si utilizzano nelle loro composizione**
Atomi + regole di composizione per combinare questi atomi in enunciati complessi = significato

Perchè la composizionalità?
Vorremmo avere un meccanismo semplice che bottom up spiega molti fenomeni

- Esiste un gran numero di frasi da costruire in linguaggio naturale pur avendo una memoria finita (procedure finite per generare un numero infinito di frasi) **principio della produttività,** posso esprimere significati mai espressi in precedenza

Non spiega esempi come 'flat tyre', 'flat beer', 'flat note' etc (la macchina ha mangiato la strada)
In quesi casi di adj-noun il nome è l'argomento di una funzione dell'aggettivo

Argomento della **sistematicità**: se siamo in grado di capire degli enunciati allora siamo in grado di capirne altri -> **la lingua è sistematica**
however, would everyone who understands 'within an hour' and 'without a watch' also understand 'within a watch' and 'without an hour'?
la lingua non è sistematica..

Tema dell **modularità:**
due forme di modularità:

- Chomsky e Fodoror: il p**otere generativo della grammatica può essere catturato da un solo modulo** avente un insieme di conoscenze lessicali + un repertorio di regole sintattiche
- modularità rilassata: **primo modulo per l'analisi sintattica più un modulo con significati e regole di combinazione**

Composizione incrementale

The meaning of a complex expression at some processing stage σ is computed based on the constituent expressions processed at σ and of the syntactic structure built up at σ

Composizione semplice

- The **meanings** of elementary expressions are the only constraints on content in the computation of the meaning of a complex expression.
- The **syntax** of elementary expressions is the only constraint on structure in the computation of the meaning of a complex expression.

Le implicazioni sono:

- tesi dell'indipendenza dal contesto, il significato delle espressione non dovrebbe dipendere dal contesto in cui occorre
- principio inside-out, l'analisi semantica procede in maniera bottom up

Se abbiamo una frase s posso dire io penso s, e se abbiamo s1 ed s2 possiamo comporle tramite operatori

Nelle frasi progressive il significato dipende dal contesto in realtà..
a.She resembles her mother.
b.*She is resembling her mother. (sbagliata secondi i nativi inglesi)
c.She is resembling her mother more and more every day

Dipende dal contesto perchè the meaning of 'resemble' is recomputed when the adverbial phrase 'more and more every day' is processed
Quindi la composizione semplice non funziona

Principio del contesto Elementary expressions do not have meaning in isolation, but only in the context of (as constituents of) complex expressions
Il principio del contesto è opposto a quello di composizionalità.. dobbiamo memorizzare le informazioni perchè il significato non può essere costruito e dedotto da significati di partenza come nella composizionalità

Evidenze sperimentali

Esistono evidenze sperimentali per avere delle idee su come funziona la composizionalità e quali sono i meccanismi che la mettono in sofferenza

Semantic illusions

when asked ‘How many animals of each sort did Moses put on the arch-linux?’, subjects tend to respond... ?

I soggetti hanno risposto in larga maggioranza “2 animali”
Siamo attratti da un meccanismo contestuale più che dal significato analitico, facciamo un’operazione tipicamente non composizionale, prendiamo il contesto del diluvio universale (non evocato direttamente) e lo proiettiamo su Mosè che NON è Noè
Abbiamo fatto una falsa deduzione confondendo Mosè con Noè

Misinterpretations

While Anna dressed the baby played in the crib

Questa frase è cortocircuitante
Il bambino all’inizio è parsificato come il soggetto di dressed e poi parsificato come soggetto di played in the crib

Event-related brain potentials (ERPs)

Risposte che diamo a certi tipi di stimoli, sono degli impulsi elettrici interpretati e messi in relazione con stimoli precedenti

I neuroscienziati utilizzano 2 componenti (sono onde):

- **N400**, componente di questi impulsi che produciamo dopo una stimolazione che viene prodotta da tutte le content-word. L’ampiezza di questa componente dipende dal termine che l’ha prodotta con il contesto in cui compare. Complessità dei meccanismi che noi usiamo per integrare il significato di un termine
- **P600**, è associata a violazioni di vincoli sintattici, aventi a che fare sulla struttura della locuzione, principi di sottocategorizzazione etc.. E’ interpretata come un indice di risorse spese dal sistema per attribuire una rappresentazione sintattica che tiene conto della frase

Sempre in questo filone di esperimenti neuroscientifici sono stati studiati elementi che hanno a che fare con il **fictional discourse,** tramite l’onda N400

A.The peanut was salted.
B.The peanut was in love

Violazione di context independent e inside-out (perchè non è costruito in maniera bottom up)

Semantic attraction
a.The hearty meal was devouring the kids.
b.The hearty meal was devoured by the kids.

devouring produce una P600 (violazioni sintattiche) ampia, quindi frase a e quel verbo in particolare produce una violazione sintattica più importante

Coercion
Ha luogo quando ci sono entità la cui interpretazione sintattica viene forzata a diventare eventi (il giornalista ha iniziato l'articolo, l'articolo non si inizia, sarebbe come dire ha iniziato a scrivere)
Sintassi della frase + conoscenze che da qualche parte devono essere memorizzate

Conclusioni

Non esistono meccanismi semplici per spiegare la variabilità del linguaggio e la pluralità di fenomeni che caratterizzano l'incontro di vari pezzi lessicali

Fenomeni che stanno fuori dai tentativi di simple composition
Composizione e modularità con meccanismi che possono essere rilassati per integrare conoscenza

Esercitazione 4

Prima fase

Dobbiamo annotare 50 coppie di termini:

- 4 coppie di termini estremamente simili
- 3 2 termini condividono molte idee fondamentali del loro significato ma con dettagli che li differenziano
- 2 non hanno un significato simile ma condividono un topic domain (casa finestra)
- 1 chiaramente differenziati ma con lontane relazioni o dominio comune e reperibili nello stesso documento o trattazione (software e keyboard)
- 0 totalmente diversi e privi di collegamenti (pencil e frog)

Prendiamo 50 coppie di termini da it.test.data.txt e annotiamo
Le 50 coppie (sul totale di 500 coppie presenti nel file) sono da individuare sulla base del cognome, tramite la funzione definita nel notebook semeval_mapper.ipynb.

L'output è quindi un file csv con coppie e punteggi

Tutti i membri del gruppo in realtà annotato le stesse 50 coppie di termini e si calcola il val medio.
E' necessario calcolare l'agreement degli annotatori utilizzando gli indici di Pearson e Spearman

Per confrontare i risultati prendi mini_Nasari.tsv che sono una lista di vettori e calcola la cos-similarity dei vettori

Se non ci sono dei termini presenti nella nostra 50 coppie escludiamo le coppie

$$\mathrm{sim}(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [\mathrm{sim}(c_1, c_2)]$$

dove sim è il cos-sim

Fra i membri del gruppo facevamo inter-rate agreement, qui invece facciamo valutazione tra il risultato calcolato tra il giudizio umano nostro e cos-similarity

Seconda Fase

La domanda che ci poniamo è la seguente: quali sensi abbiamo effettivamente utilizzato quando abbiamo assegnato un valore di similarità a una coppia di termini (per esempio, società e cultura)?

- NB: questa annotazione, sebbene svolta successivamente a quella della prima consegna, deve essere coerente con l'annotazione dei punteggi di similarità.

L'output qui deve contene il 2 babel synset id dei due termini in questione e i termini associati a quel synset:

#Term1 Term2 BS1 BS2 Terms_in_BS1 Terms_in_BS2
macchina bicicletta bn:00007309n bn:00010248n auto,automobile,macchina bicicletta,bici,bike

Quindi abbiamo 6 campi separati da tab e virgola come separatore interno tra i termini del singolo synset
Valutare se si fa in gruppo la Kappa di Cohen cohen_kappa_score della libreria sklearn.metrics

Per Valutare invece facciamo sempre con mini_Nasari.tsv

$$c_1, c_2 \leftarrow \arg\max_{c_1 \in s(w_1), c_2 \in s(w_2)} [sim(c_1, c_2)]$$

e sempre cos-sim(c1,c2) per la sim(c1,c2)

Misuriamo sia l'accuratezza sui singoli elementi sia sulle coppie

Domande Radicioni

- Wordnet: cos'è un synset? (e poi annessa discussione su Wordnet)

- Wordnet è uno strumento per una annotazione funzionale? Nel senso, per i task svolti nelle esercitazioni, non sarebbe meglio qualcos'altro?

- Come si usa Framenet in un contesto applicativo?
- eventuali problemi dell'alta granularità di FrameNet. Differenze con quella di WordNet
- Come fare disambiguazione con FrameNet

- Inizia sempre da un'esercitazione. O a scelta sua o a scelta dello studente
- Spiegazione dell'esercitazione e tool usato (Wordnet, Framenet ecc.)
- Nel testo dell'esercitazione vi ho chiesto di produrre un output particolare, discutiamolo un momento, cosa c'è dentro a questo file?

- Se io le dico "spazi concettuali" gli viene in mente qualcosa? Cosa sono? Dove li abbiamo visti? (lo studente non se lo ricordava ed è passato oltre) -- DOLCE
- Mi definisca in modo sintetico cos'è Verbnet.
- Partenza dall'esercitazione sulla summarization: Come valutare un riassunto?
- Cos'è Nasari?
- Nasari è una perfetta copia vettoriale di Babelnet? _No, Babelnet è più ricco, (iponimi, iperonimi...) è qualsi__un'ontologia, mentre Nasari è una rappresentazione degli oggetti_.
- Cos'ha BabelNet in più di Nasari? _I verbi_.
- Cos'è un endurante? Un perdurante?
- Com'è strutturato FrameNet? A cosa serve?
- Com'è strutturato VerbNet? A cosa serve?
- Quali sono le differenze fra FrameNet, VerbNet e ConceptNet?
- Che differenza c'è in Babelnet tra Named Entities e Concetti? (trabocchetto, in un nodo Babelnet possono essere
presenti entrambi)
- Mi parli della teoria dei prototipi.
- Cosa le viene in mente se le dico "ruolo semantico" e "funzione sinattica"? _Verbnet_. Perfetto, mi parli di Verbnet allora.
- Verbnet: cos'è l'alternation?
- Cos'è l'approccio moltiplicativo -- DOLCE
- Cos'è la semantica procedurale
- Differenza tra teoria dei prototipi e teoria degli esemplari