**APPUNTI VERBNET**

Rappresentazione del senso dei verbi. Ci siamo avvicinati a questa rappresentazione quando abbiano introdotto framenet. L’accesso ai frame è governato dall’individuazione di verbi e dal tentativo di trovare un mapping sensato delle frasi in input e i possibili frame element dei frame in input. Questo mapping è influenzato molto dall’informazione portata dai verbi.

Partiamo da una osservazione: immaginiamo di aver a che fare con un’applicazione di question answering. L’obiettivo è estrarre informazioni, pezzi di frase in input.

* Immaginiamo la domanda: Chi ha creato il primo vaccino anti polio?
* Riusciamo ad individuare 2 possibili risposte:
  + Dickinson ha creato la prima siringa
  + Il primo vaccino è stato creato nel ‘52 da Ionas...

La risposta è nella seconda frase. Posso fare una ricerca di termini rilevanti. Restituisco la prima frase, sbagliando... In questa operazione di ricerca dei termini, un metodo è partire dal verbo. In questo caso cerco il verbo **creare**. Le 2 frasi non sono distinguibili. Si potrebbe guardare l’oggetto diretto del verbo, nel primo caso scopro che è stata creata una siringa, nella seconda no. Con un approccio shallow scopro che “**vaccino antipolio**” è citato in entrambe le frasi. Quindi la ricerca per pattern matching non va bene.

Prospettiva migliore: Individuiamo elementi che fanno parte della semantica della frase. Procediamo con un’analisi profonda in cui individuiamo un agente che **crea un qualcosa** e vediamo che nella prima frase l’oggetto creato è una siringa nel secondo è il vaccino antipolio. Quindi la seconda vince. La seconda frase è un pò più difficile da elaborare perchè è una frase passiva. Possiamo rigirare la frase dicendo che “**l’agente ha creato il qualcosa**”

Non ci muoviamo più nella ricerca di parole o di sensi (synset), stiamo cercando di individuare **“ruoli semantici”,** un task molto più difficile del **pattern matching**. Dobbiamo maneggiare la struttura sintattica. In questo caso prima facciamo analisi sintattica individuando il verbo poi prendiamo i dipendenti del verbo, capiamo che il soggetto è l’agente e l’oggetto è il tema di ciò che è stato creato. A questo punto appiccichiamo il **tag semantico** (ruolo) all’analisi dei dipendenti condotta a livello sintattico.

Soluzione della domanda?

* Analisi sintattica
* Etichettatura dei ruoli semantici di agente, paziente, tema
* Scelta della risposta che corrisponde meglio al tema creato dall’agente

Come è possibile condurre l’analisi semantica? A partire dai verbi. Sapendo il meccanismo di funzionamento dei verbi, si cerca dai dipendenti e si individuano i ruoli semantici per rispondere alla domanda.

Qual è l’ipotesi alla base di verbNET? Le possibili realizzazioni sintattiche possono essere usate come strumento base per individuare il significato del verbo. Dalla sintassi al significato. Per il verbo che cerchiamo di studiare osserviamo delle forme in cui la sintassi riflette il significato profondo di un certo verbo. Lo possiamo fare con:

* Insieme di regole che descrivono il funzionamento del verbo
* Dal punto di vista probabilistico, possibili significati del verbo più probabili

L’idea è stata quella di estendere un approccio gia presente il Framenet. Solo che in framenet il meccanismo di individuazione del significato del verbo dai suoi dipendenti era focalizzato dal punto di vista del frame. Qui non ci sono frame che fungono da meccanismi di aggregazione di vari pezzi. Qui ci concentriamo direttamente sui verbi e i loro possibili significati e i possibili modi di declinarli sintatticamente.

VerbNET è una enorme rete semantica di verbi per la lingua inglese, che descrive i loro pattern sintattici e i loro ruoli semantici. Operazioni di verbNET:

* Costruzione di classi verbali. L’individuazione di un certo numero classi verbali avviene sulla base del significato dei verbi a partire dalla sintassi. Ogni classe è descritta da:
  + Ruoli semantici
  + Selection restriction (restrizioni semantiche)
  + Frames sintattici: I frames consistono in descrizioni sintattiche e rappresentazioni semantiche dei verbi. Permettono un mapping tra livello sintattico e semantico.
* Mapping delle classi verbali su risorse già esistenti, ovvero WordNET, frameNET e PropBANK. Ad esempio nella lista di verbi in una classe ci sono riferimenti a framenet e a wordnet
* Riferimenti alle classi di Levin per costruire le entry delle classi verbali

Esempi: Differenze tra frame sintattici

* John left the room. (oggetto diretto, leave significa uscire da..)
* John left the book on the table (lasciare)

In questo caso la sintassi ci aiuta a discriminare i significati, nella prima frase c’è un oggetto diretto, nella seconda ci sono 2 oggetti, uno diretto, uno indiretto.

* Jonh left the room
* John left a fortune

Qui la sintassi non ci aiuta perchè non possiamo distinguere i due alberi a dipendenze. Abbiamo bisogno della semantica. Vogliamo analizzare le strutture predicative, argomentali fornendo delle generalizzazioni, così come abbiamo visto per il corpus annotato di FrameNET.

**Caratteristiche di VerbNET:**

Verbnet è un lessico computazionale per i verbi, che può essere usato per cercare verbi. E’ una risorsa di larga copertura, non è dipendente da uno specifico dominio.

* Contiene un’associazione esplicita tra livello sintattico e semantico, per cui chiarisce nella descrizione delle varie classi dei verbi in cui ci si può imbattere gli elementi che appartengono a sintassi e semantica e ai rapporti trai 2 livelli.
* Fa uso esplicito di frames sintattici e di selectional restriction (vincoli selettivi). Questa cosa è importante e richiede un livello di analisi notevole. Le restrizioni selettive sono le restrizioni semantiche che un verbo impone ai suoi argomenti. Un verbo come mangiare richiede che il suo soggetto si riferisca a un'entità animata e il suo oggetto a qualcosa di concreto. Una violazione dei vincoli selettivi di una parola comporta un'anomalia, ad esempio: “in montagna si mangia sincerità”. Entrambe le restrizioni vengono violate, rendendo la frase anomala.
* Si avvale di altre risorse come WordNET, frameNET e probBank

E’ un modo che ci avvicina molto ad una riscrittura di una frase in input in un formato logico.

**Classi di Verbi:**

Che cosa sono e come sono fatte? L’intuizione dietro le classi verbali è individuare elementi e caratteristiche distintive in ogni classe per descrivere dei pattern morfo-sintattici. Quindi sono insiemi di verbi che condividono comportamenti semantici e simili pattern sintattici. Per fare ciò, i linguisti studiano **l’alternation**. In generale l’alternation è studio della variazione della forma e del suono di parole o pezzi di parole. Lo studio dell’alternation ci aiuta a capire come diverse manifestazioni sintattiche degli argomenti di un verbo possono individuare significati diversi per quel verbo. (L’alternation degli argomenti di un verbo è un insieme di espressioni diverse che, a seconda degli argomenti, possono dar luogo a significati diversi):

* Un individuo ha spruzzato dell’acqua sulle piante
* Un individuo ha spruzzato le piante con dell’acqua
* Un contadino ha caricato delle mele sul carrello
* Un contadino ha caricato il carrello con le mele

Queste coppie di esempi ci danno un esempio del funzionamento dell’alternation. Le 2 alternation sono 2 modi per usare lo stesso significato, in un caso usando un locativo nell’altro un oggetto diretto.

Le classi di verbi di VERBNET sono formate da:

* **Riferimento identificativo**
* Eventuale **classi genitore** da cui ereditano proprietà sintattiche. Le classi sono organizzate gerarchicamente. Permette la condivisione intraclasse di proprietà.
* Lista dei membri **(verbi) della classe** con riferimenti ad altre risorse (Wordnet, Framenet)
* **Insieme di pattern sintattici**, (**Frames sintattici**) ovvero le possibili alternation che caratterizzano una classe. (proprietà sintattiche)
* **Ruoli semantici**: (**Agent**, **Patient, Instrument, Experiencer**) Il sistema dei ruoli qui è diverso. I ruoli sono più semplici, e più generici come Agente, paziente, tema, location.
* **Vincoli selettivi** per ogni frame. Necessità semantiche e possibili elementi che possono taggati con uno specifico ruolo semantico.
* **Predicati semantici** nella descrizione delle varie classi che indicano come funzionare.

Ad esempio la classe “Break”:

* Membri: break, crack, crash, crush, fracture, rip, shatter, smash, snap, splinter, split, tear
* Properties:
  + **Causative/inchoative alternation**: Tony broke the window | The window broke
  + **Middle alternation**: Tony broke the crystal vase | Crystal vase break easily

1. Incohative: inizio dell’azione (smise di combattere)
2. Middle: forma impersonale

Un frame sintattico sono tutte le possibili realizzazioni che caratterizzano la semantica di una classe di verbi. si presenta sotto forma di:

* **Analisi sintattica** : Costituenti, ad esempio NP V NP (schema di funzionamento sintattico)
* Una frase esemplificativa: (Tony(**NP**) broke(**V**) the window(**NP**))
* Ruoli semantici mappati sui costituenti: (Agent V Patient)
* **Predicati semantici:** Una rappresentazione logica, con i predicati FOL per indicare che c’è un agente che è Tony che produce un evento di contatto tra pazient e un possibile strumento con esito di “degradazione materiale” di un paziente che è una finestra. Questa rappresentazione permette ad un agente di capire cosa sta succedendo.

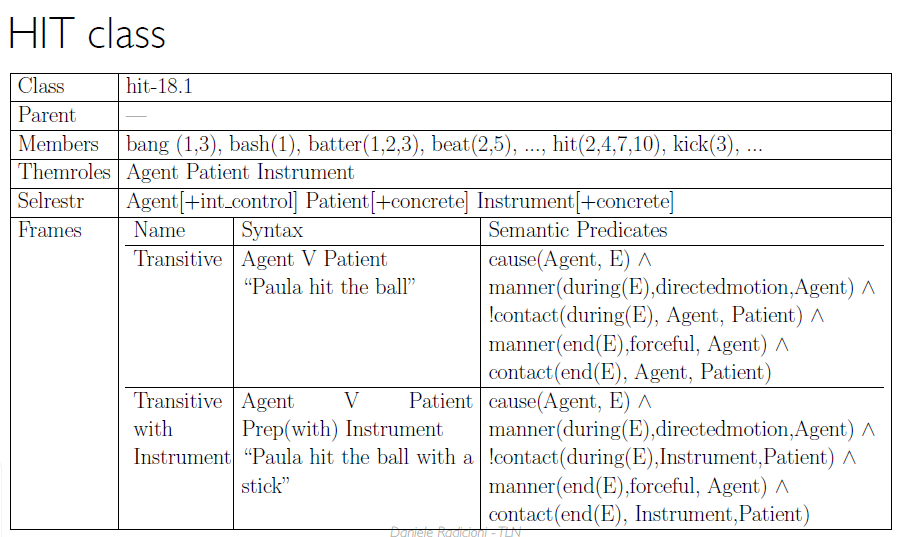
L’ipotesi è che la semantica sia governata da una sottostruttura a livello sintattico. La sintassi è utile per esprimere delle nozioni come:

* Contatto
* Movimento
* Applicazione di forze
* Cambio di stati

Ad esempio il verbo Break si può esprimere in termini del cambio di stato di un oggetto rotto. I verbi hanno un significato complesso. Alcuni elementi chiave sono:

* Lista dei partecipanti: ovvero gli elementi base degli eventi
* Hanno delle collocazioni fisiche e temporali
* Possono essere usati per esprimere eventi, stati, processi.
* In molti casi possono essere suddivisi in sotto parti. (sottospecificazioni)

I verbi possono essere descritti in termini dell’evoluzione dell’evento che rappresentano. Verbi come run, o break hanno un inizio, una durata, una fine. Run ha una fase preparatoria, break ha uno stato di conseguenza. Verbi come hit non hanno durata, iniziano e finiscono nello stesso momento. Il fatto di colpire è puntuale, il fatto di rompere è alla fine di colpire. La semantica del verbo è caratterizzata da cosa succede durante 3 fasi: inizio, evoluzione dell’evento, fine



**Ruoli semantici:**

Gli elementi che costituiscono la vera annotazione semantica. Grazie ai ruoli possiamo rispondere alla domanda iniziale, individuando chi ha fatto qualcosa. Ogni classe ha un insieme di ruoli semantici. In totale ce ne sono una trentina. Esempi di ruoli semantici sono:

* Actor: serve per caratterizzate quando c’è un secondo agent. Tipo, marco ha incontrato luca, sono entrambi degli agenti. L’actor è un altro agente che agisce, non subisce
* Agent: chi agisce con un’azione su un oggetto . In genere un umano o un oggetto animato.
* Asset: usato per alterazioni su somme di denaro
* Attribute: L’attributo di un agente o un tema indica una qualità di un qualcosa che sta per essere cambiato
* Beneficiary: L’entità che beneficia di una qualche azione
* Cause: per gestire relazioni causali
* Location: per gestire relazioni spaziali
* Experiencer: indica un partecipante che prova una qualche esperienza
* instrument: Usato per oggetto che vengono usati
* Patient: colui che subisce l’azione. (In framenet sarebbe vendicato)
* Recepient: colui che riceve qualcosa
* Theme: l’oggetto di una qualche azione (l’oggetto inventato dall’inventore dell’antipolio)

Con un sistema di ruoli molto limitato, si riesce ad incasellare tutte le necessità semantiche della lingua. In frameNET abbiamo una frantumazione della semantica che è associata alla semantica del frame. Qui anzichè distinguere per essere precisi, si è deciso di individuare ruoli generici. Questo è molto meglio. Più sintetico.

**Vincoli selettivi:**

I vincoli selettivi sono dei vincolo imposti ai ruoli semantici per raffinare ulteriormente il processo di assegnazione della semantica del verbo. Possono essere articolati su più livelli:

* Semantico
* Sintattico

Sapere che uno che corre è un essere vivente è più informativo di sapere che NP è il costituente dell’agente dell’azione di correre.

**Proposition Bank project**

PropBank è un livello ulteriore di annotazione sintattica e semantica di verbi. Si basa sull’albero sintattico contenuto nel Penn treebank. PropBank aggiunge al corpus penn treebank l’annotazione di ruoli semantici ai nodi già presenti. Quali ruoli semantici? Molto più astratti dei ruoli semantici di verbNET. Non sono dei ruoli veri, ma sono ruoli a cui sono stati dati degli indici, per renderli ancora più astratti e sintetici. Ogni verbo è annotato con uno tra i 10 possibili ruoli semantici. I ruoli indicano i diversi usi semantici di un verbo.

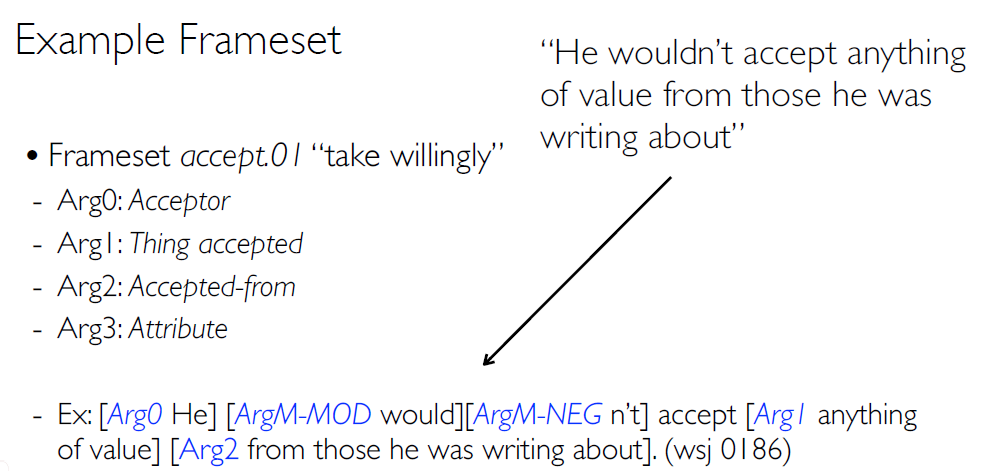
* Arg0: proto-Agent: circa (Agent, causer, experiencer)
* Arg1: proto-Patient (Patient, themes)
* Arg2: Instrument, benefactive, attribute
* Arg3: starting point
* Arg4: ending point

Questi 5 ricoprono la maggior parte dei dipendenti nella lingua. Perchè hanno usato indici? Per essere generici e per trovare un idea comune. Oltre a questi ci sono dei modificatori come:

* ArgM-Loc (Location)
* ArgM-Cause(causa)
* ArgM-Time (di tempo)

I modificatori danno ulteriori informazioni su come, quando e dove l’evento del verbo occorre. L’insieme dei ruoli che corrispondono ad un certo senso e ad un certo utilizzo del verbo costituiscono il suo Role Set. Il Role Set e insieme dei frame associati costituiscono il frame Set di un certo verbo. Un verbo polisemico può essere caratterizzato da una pluralità di frameSet. I frameSet possono essere viste come delle forme di distinzioni grossolane di sensi di verbi rispetto alla granularità.

* Ciascun frameset corrisponde a diversi sensi di wordNet. Con wordNet il fatto di avere molti sensi minimamente distanziati comportava che se i sensi sono molto vicini è più difficili distinguerli e diventa anche meno utile distinguerli. Voglio poter distinguere macro aree di senso associati ad un termine. Semplificare i frameset in propBank permette di semplificare il task di disambiguazione.
* Ogni istanza di un verbo polisemico è marcata con il frameset di appartenenza con un inter annotator agreement di 94%



Se abbiamo 2 sensi distinti di wordnet per il verbo ring possiamo trarre 2 roleSet:

* 01, make sound of bell (Abbiamo che Arg0 può essere colui che causa il ring, l’arg1 è la cosa che viene suonata e l’arg2 è il motivo per cui viene suonato)
* 02, to surround (arg1 entità che circonda, arg2 entità circondata)
* Jonh[Arg0] rings the bell[arg1]

Come si mappano questi argomenti sui ruoli tematici? Come i ruoli tematici sono mappabili su questi argomenti? Qual è il grado di copertura e quali sono i nessi tra gli argomenti di prob bank (ARG) e i ruoli semantici di verbNet?

* Arg0 ----> Agent (85% dei casi)
* ARG1 ---> Theme (47% dei casi)
* Arg2 -----> Recipient (23% dei casi)

Data una frase del penn trebank, abbiamo un albero a costituenti etichettato con gli argomenti di propBank.

**Esercitazione:**

Metodo di rocchio. Metodo usato per estendere i classificatori lineari.

Rappresentazione dei vettori: Vettori di features con bag of words. Vettori che rappresentano i documenti. Con il term frequency, contiamo il numero di termini rilevanti presenti in tutti i documenti di una collezione. Sappiamo che usiamo vettori sparsi caratterizzati con tanti 0 in base al numero di documenti nella collezione. Il vettore di un documento è un insieme di numeri

Come uso questa notazione quando faccio una query? Anche la query viene rappresentata come un vettore di n elementi. Si costruisce un vettore per la query con pochi numeri non 0. A quel punto come individuo un documento in base alla query?

Con la rappresentazione a vettori collochiamo i documenti in uno spazio multidimensionale. Punti vicini equivale a documenti simili. Distanza del coseno per la similarità! Restituiamo il documento più simile.

Altro meccanismo è correggere questo principio con un altra pesatura. Ad esempio pensando alle stop words come alle parole molto frequenti che non portano significato allo stesso modo la frequenza non va bene. L’idea è che ci possono essere termini poco frequenti ma discriminativi che con la frequenza non vengono usati. Questi termini non sono legati al POS. Ad esempio “fare” è un verbo troppo non discriminativo e molto frequente. L’idea è che i termini che sono presenti in pochi documenti sono utili per distinguere quei documenti dal resto della collezione, quindi i termini che ricorrono in tutti i documenti non sono utili. Viene introdotta la misura IDF.

IDF = N/n = Numero di documenti della collezione / numero dei documenti in cui il termine n occorre. Per questioni pratiche questo risultato è schiacciato nel logaritmo. Funziona anche senza il logaritmo. Log (N/n). **Similarità = TF \* IDF**

Non consideriamo ìnformation retrieval. Categorizziamo documenti. Definiamo un profilo di una classe a partire da un centroide. Il metodo di rocchio migliora la qualità descrittiva dei centroidi.