Seconda parte radicioni

Introduzione

Semantica lessicale e ontologie

La semantica lessicale consiste nello studio del significato delle parole e delle loro relazioni. Uno strumento che viene in aiuto alla semantica è l'ontologia, concetto che gode particolare successo nell'ambito dell'informatica grazie al suo uso esteso nell'intelligenza artificiale.

Una ontologia è un sistema strutturato di entità organizzato in categorie e relazioni che permette di modellare un dominio di conoscenza. Ciascuna classe e relazione definisce una serie di informazioni sugli elementi che le appartengono, come ad esempio che il verbo "mangiare" deve avere un soggetto vivente e un oggetto categorizzato come "edibile".

Questa prima definizione è troppo poco restrittiva e permette di categorizzare quasi tutto come un'ontologia; si passa quindi a definire un'ontologia come specifica di una concettualizzazione.

Una concettualizzazione è una struttura formale di parte della realtà percepita da un agente, indipendente dal vocabolario utilizzato e l'occorrenza di una specifica situazione. Con questa definizione è possibile avere situazioni diverse descritte da vocabolari differenti che però condividono la stessa concettualizzazione. Il concetto torna molto utile nella semantica, perché il significato di una frase, ad esempio, rimane uguale indipendentemente dalla lingua (e il suo vocabolario) presa in esame.

Ecco quindi che si viene a creare una relazione stretta fra le ontologie e la semantica, con aspetti diversi del linguaggio catturati tramite l'uso di ontologie specifiche per il ruolo.

Le ontologie possono essere viste anche come parte di una base di conoscenza (KB). Queste infatti sono formate da due parti:

- Teminological component (T-box), rappresentata proprio dall'ontologia e indipendente da un particolare stato;
- Assertional component (A-box), rappresentante stati specifici (es. "il libro è sul tavolo") e usata per la risoluzione di problemi.

Un esempio di uso delle ontologie è per la rappresentazione del livello lessicale. Si tratta di un elenco di parole in una lingua che hanno delle relazioni lessicali fra di loro (sinonimia, iponimia, meronimia, ecc.).

Queste relazioni spesso assomigliano a relazioni tipiche delle ontologie come subclass-of e part-of, ma con una differenza: nelle ontologie sotto-categorie di una certa categoria sono in genere mutualmente esclusive, mentre nel lessico esistono invece sovrapposizioni di significati fra parole simili. Inoltre, il lessico di un certo linguaggio potrà non avere alcun termine per certe categorie ontologiche che non sono lessicalizzate in quella lingua.

Perché usare le ontologie se bisogna tenere conto di questi dettagli durante la loro costruzione? Perché permettono facilmente di condividere la comprensione della conoscenza di un certo dominio

fra persone e agenti software e permettono un riutilizzo dei dati in contesti diversi, senza dover ricreare tutta la rappresentazione della conoscenza.

Design di ontologie

In un'ontologia linguistica non è possibile utilizzare la stessa relazione per tutti gli elementi, ma serve invece differenziare in base al concetto.

Una prima, grande, distinzione fra gli elementi di un'ontologia è quella di entità ed eventi:

- entità (endurant), sono oggetti che continuano per un periodo di tempo mantenendo la propria identità (ad esempio, cellula, vaso, casa); possono cambiare nel tempo e possono essere costituiti da parti non essenziali, ma tutte quelle essenziali sono presenti nello stesso luogo in cui l'entità è presente.
- eventi (perdurant), sono oggetti che accadono, si svolgono o si sviluppano nel tempo (ad esempio, la replicazione del DNA, la caduta del vaso); sono localizzati nel tempo ma solo in relazione ad altre entità e non si può dire che mutino. Tutte le loro parti sono fondamentali.

Un'ontologia costruita su questo concetto (e pochi altri che riguardano sempre macro-distinzioni fra elementi) viene definita ontologia fondazionale (upper ontology). Si tratta di un'ontologia che cattura distinzioni di base valide in più domini; in alcuni ambiti questo tipo di ontologie tornano molto utili, in altri no. Il campo dell'elaborazione del linguaggio naturale è uno dei campi in cui tornano utili.

Durante il design di un'ontologia è anche necessario usare dei criteri per decidere se due entità sono o meno in relazione fra di loro; vediamo l'esempio della relazione basilare di sottoclasse (subclass-of) che fa uso del criterio di identità.

Questo criterio dice che serve confrontare delle proprietà necessarie delle entità coinvolte; se queste proprietà coincidono, allora un'entità è sottoclasse dell'altra. Ad esempio, una durata temporale ha come criterio di identità "stessa lunghezza", mentre un intervallo temporale ha "stesso inizio e stessa fine". Essendo i due criteri diversi, una non è sottoclasse dell'altra.

Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)

Si tratta di un'ontologia fondazionale che nasce da tre scelte:

1. Distinzione fra endurants e perdurants; la relazione principale che intercorre tra i due è quella di partecipazione degli endurant in alcuni perdurant.
2. Qualities, ovvero entità individuali che si trovano dentro agli endurants e perdurants in grado di cambiare con gli oggetti dentro cui si trovano e che prendono i propri valori da un Quality Space associato (ad esempio, il colore di una rosa è una quality che assume valore "rosso" dal quality space dei colori possibili).
3. Approccio moltiplicativo, che afferma che oggetti/eventi diversi possono trovarsi nella stessa posizione spazio-temporale mantenendo però una distinzione (ad esempio, l'argilla di cui un vaso è costituito si troverà fisicamente nello stesso posto del vaso, ma sarà considerata un'entità diversa). Le due entità restano separate perché possiedono delle caratteristiche tipiche che non si trovano nell'altra entità (il vaso, una volta rotto, cessa di

esistere, mentre l'argilla no; allo stesso tempo il vaso può essere composto da un manico, mentre l'argilla no).

Knowledge Representation (KR)

Reti semantiche

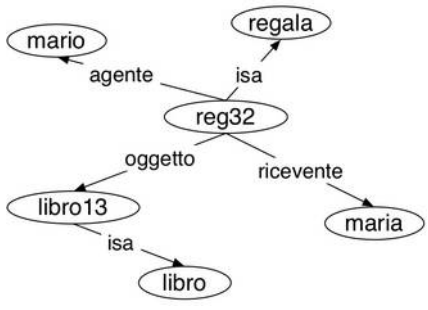
La rappresentazione della conoscenza tramite logica del prim'ordine, primo approccio storico, rende imprescindibile l'uso di regole per poter legare fra di loro le formule e derivare della conoscenza implicita (inferenza). L'inferenza però ha un costo piuttosto elevato e sono quindi nate rappresentazioni alternative più dirette.

Le reti semantiche costituiscono uno strumento alternativo alla logica classica che permette di aggregare conoscenze elementari in strutture più complesse, le quali rendono immediatamente disponibile tutte le informazioni senza la necessità di ulteriore ragionamento.

Le reti semantiche hanno una struttura a grafo in cui i nodi rappresentano concetti e gli archi le relazioni o proprietà fra concetti.

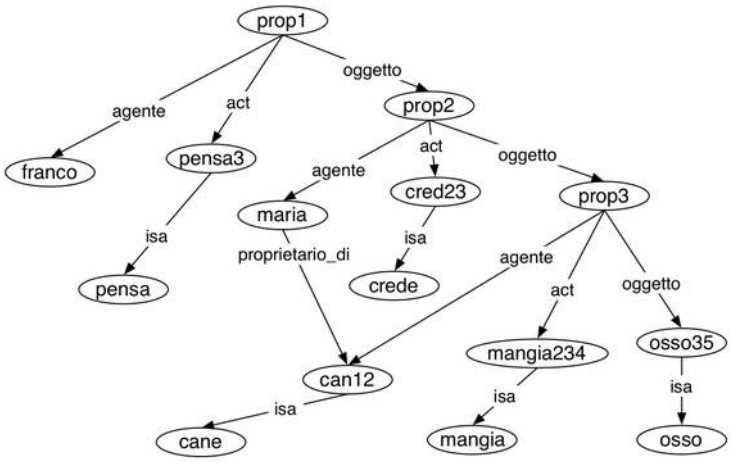
La loro versione più semplice è quella dei grafi relazionali in cui vengono descritte le relazioni fra le entità del grafo. Una delle relazioni più importanti è quella di tipo isA. È possibile vedere i grafi relazionali come sottoinsieme della FoL, dove i nodi rappresentano i termini e gli archi i predicati. Questi grafi sono adatti a rappresentare i problemi più semplici, in genere fittizi, come un mondo dei blocchi, ma hanno dei limiti nell'espressività: le relazioni rappresentate modellano soltanto la congiunzione. Tutti gli archi di un nodo sono infatti visibili come in AND fra di loro, ma manca un formalismo per rappresentare la disgiunzione, l'implicazione, i quantificatori e relazioni che non siano binarie (da un nodo all'altro).

È inoltre facilmente osservabile che questo tipo di grafi è limitato alla rappresentazione di relazioni binarie: una relazione n-aria è scomponibile in più relazioni binarie introducendo dei nodi aggiuntivi per rappresentare oggetti, situazioni e azioni e etichettando le relazioni per definire il ruolo degli elementi che vi partecipano. Questo porta però ad una crescita delle dimensioni e della complessità del grafo relazionale. Nell'esempio sottostante, la rappresentazione del gesto da parte di Mario di regalare un libro a Maria.



Sebbene ci sia isomorfismo tra grafi relazionali e logica del prim'ordine, in questa traduzione si perde significatività (generalizzazione dovuta ai predicati, validi per più situazioni).

Il passaggio successivo è quello di creare delle reti in cui i nodi rappresentano non solo entità semplici, ma intere proposizioni: queste prendono il nome di reti proposizionali.



Rete proposizionale per la frase "Franco pensa che Maria creda che il suo cane stia mangiando un osso"

Una volta introdotto il concetto di nodo proposizionale è possibile aumentare l'espressività della rete introducendo i connettivi logici e i contesti dentro cui far operare i quantificatori. I sistemi di reti proposizionali più evoluti sono in grado di esprimere tutto quanto è esprimibile col calcolo dei predicati del prim'ordine.

Ad esempio, la negazione è esprimibile tramite una relazione apposita fra due nodi proposizionali: $p_1 \overset{NOT}{\rightarrow} p_2$, posizionabile nel punto della frase che si vuole negare.

La disgiunzione richiede invece l'uso delle leggi di De Morgan, trasformando $A \vee B$ in $\neg(\neg A \wedge \neg B)$ e applicando di conseguenza la relazione not.

Altri concetti come i quantificatori risultano di più complicata implementazione, ma è possibile descriverli. Il potere espressivo delle reti semantiche è quindi identico a quello della logica classica; la scelta sull'uso di un sistema piuttosto che l'altro dipende quindi da altri fattori come la leggibilità, l'efficienza e la facilità rappresentativa dei concetti.

Spesso nella linguistica bisogna rappresentare conoscenze organizzate gerarchicamente e le reti semantiche si prestano particolarmente a questa rappresentazione (la relazione isA è un esempio chiave). Inoltre, con le reti semantiche esiste il meccanismo di ereditarietà delle proprietà che permette di risalire a tutte le proprietà di cui gode un elemento risalendo il cammino gerarchico, evitando di ripetere più volte le stesse informazioni; questa operazione risulta molto più efficiente di un sistema basato su regole di inferenza. In caso di eccezioni, inoltre, esse vengono definite a livello dell'elemento interessato, venendo quindi scoperte prima della caratteristica generale (ad esempio, i pinguini si dice che non volino, seppur siano degli uccelli che hanno come caratteristica generale la capacità di volare).

Esistono però anche degli svantaggi delle reti semantiche.

Il primo è che in caso di ereditarietà multipla le cose si complicano parecchio perché si hanno grafi invece di alberi e l'efficienza del sistema di ricerca diventa peggiore di quella nello spazio degli stati. Nasce anche il problema dei conflitti di valori, in cui il risultato dipende solo dall'algoritmo di

ricerca utilizzato (profondità o ampiezza, quest'ultimo anche detto shortest-path inheritance). Le reti semantiche con ereditarietà multipla devono essere quindi trattate come ambigue, ovvero in grado di dare molteplici interpretazioni.

Inoltre, le reti semantiche non permettono di specificare per chi sia vera una certa proprietà: l'ereditarietà farà in modo che tutti i discendenti di un nodo possiedano le sue proprietà (salvo esplicita sovrascrittura), ma questo potrebbe non essere sempre desiderato. Ad esempio, associare al nodo "elefante" il fatto che sono una specie in via di estinzione, porterà a dire che Dumbo (un elefante) è una specie in via d'estinzione. Non ha senso! Risulta quindi impossibile separare la semantica della rete dal suo uso, poiché da quest'ultimo deriva il senso catturato dalla rete stessa.

Frame

Si tratta di un formalismo con aspetti in comune alle reti semantiche. Il concetto fondamentale è che le persone usano un insieme strutturato di conoscenze pregresse (frame), rappresentanti genericamente una situazione o problema, quando si trovano a dover affrontare qualcosa di sconosciuto. Il frame viene poi affinato e modificato per tenere conto dei dettagli della situazione attuale. Un esempio di frame è quello del ristorante: entrando ci si aspetta di trovare dei tavoli, di essere raggiunti da un cameriere, ecc. Questo porta a interpretare i nuovi fatti all'interno di un contesto, aiutando a ridurre le ambiguità e facilitando il recupero delle informazioni.

Non esiste un formalismo per la rappresentazione dei frame; la sua struttura varia fra implementazioni diverse. Tuttavia è possibile dire che esiste sempre una struttura gerarchica a tre livelli: quello di base, che rappresenta una rappresentazione naturale degli oggetti, il livello superordinato, che invece è costituito da una generalizzazione di quello di base e quello subordinato in cui invece si opera una specializzazione dei concetti espressi al livello di base.

L'appartenenza a una categoria non è rappresentata da un insieme di attributi ma dalla (maggiore o minore) somiglianza rispetto a un prototipo della categoria: un passero è un buon prototipo di uccello, mentre un pinguino lo è meno.

Come nelle reti semantiche, anche nei frame le espressioni isA e aKO (a kind of) permettono l'ereditarietà delle proprietà fra frame; nei livelli più bassi le proprietà possono anche essere in contrasto con quelle dei livelli superiori.

ristor_234	
isa	ristorante
denominazione	da sora lella
cucina	romana
categoria	buffet
indirizzo	...
...	...

Esempio di frame

Ogni frame è identificato univocamente da un nome e le sue proprietà sono rappresentate mediante slot. I valori degli slot possono essere noti o meno e quando non lo sono, è comunque a volte

possibile fare delle supposizioni accettabili sul valore fino a prova contraria. Il valore di uno slot può anche essere una struttura complessa (un altro frame).

Anche per i frame può esistere l'ereditarietà multipla, con diversi meccanismi di gestione dei conflitti in base al sistema a frame preso in esame (strategie di ricerca, numero massimo di livelli entro cui cercare il valore, definizione di un frame universale che viene sempre consultato, ecc.).

Oltre alla parte dichiarativa, nei frame è possibile dichiarare una parte procedurale che agevola alcune computazioni sfruttando la conoscenza del dominio specifico, come ad esempio il calcolo di un valore di uno slot quando richiesto.

È evidente che la struttura a frame rappresenti la base per il paradigma della programmazione ad oggetti, dove si raggruppano le proprietà di un entità e i rispettivi valori all'interno di un frame.

Teorie e approcci alla KR

La teoria classica nell'informatica per la rappresentazione della conoscenza è l'uso della logica formale e delle ontologie, che rappresentano un insieme di vincoli che devono essere sempre soddisfatti in un modello.

Esiste poi la teoria delle tipicalità, suddivisibile in teoria dei prototipi e degli esempi: la prima afferma che un prototipo è un'approssimazione rappresentativa di una categoria, mentre la teoria degli esempi afferma che la rappresentazione di un concetto è l'insieme delle rappresentazioni dei prototipi di quella categoria che sono stati incontrati durante la vita. Si tratta in entrambi i casi di modelli in cui si va a categorizzare un elemento sconosciuto sulla base di confronti rispetto a quanto presente nella nostra conoscenza, scegliendo la categoria a cui l'elemento più si avvicina. È quanto abbiamo visto con le reti semantiche e i frame.

È possibile unire le due teorie, apparentemente discordanti, tramite una nuova teoria, detta del processo duale, per cui l'uomo ha due capacità di ragionamento che usa in base alla necessità: la prima, rapida, naturalmente acquisita, fatta di operazioni veloci a volte approssimate e contradditorie (teoria della tipicalità) e la seconda, lenta, consapevole, legata al linguaggio, astratta e logica (teoria classica).

WordNet

Descrizione

Si tratta di un sistema di riferimenti lessicali online basato sul concetto di synonym set (synset), ciascuno dei quali rappresentanti un concetto lessicale.

WordNet suddivide il lessico in quattro categorie, ciascuna delle quali rappresentata con una struttura adeguata:

- nomi;
- verbi;
- aggettivi;
- avverbi.

La semantica lessicale separa, per ogni parola, il suo concetto lessicalizzato (il significato) dalla utterance (la stringa che forma la parola). Il concetto è definito anche "word meaning", mentre la utterance "word form".

È quindi possibile rappresentare le parole in una matrice, detta matrice lessicale.

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					$E_{m,n}$

Le parole sulla stessa riga costituiscono dei sinonimi dello stesso concetto M, mentre parole sulla stessa colonna costituiscono forme polisemiche (ovvero la stessa scrittura rappresenta concetti diversi, come la parola "calcio" intesa come sport o come elemento chimico). I significati sono rappresentabili con l'insieme delle forme che li esprimono ($M_1 = \{E_{1,1}, E_{1,2}, ...\}$): questi insiemi sono i synsets accennati prima.

Relazioni semantiche

WordNet è organizzato in base alle relazioni semantiche, che possono essere viste come archi tra synset diversi.

Sinonimia

La similitudine di significato è la relazione più importante di WordNet; è definita come la capacità di una parola di essere sostituita ad un'altra, in un certo contesto, senza alterare il valore di verità della frase.

Antonimia

È la relazione di significati opposti fra parole. Risulta molto difficile da definire, perché non sempre equivale a una negazione netta (ricco e povero sono antonimi, ma non essere ricco non equivale perforza ad essere povero).

Iponimia

A differenza delle due relazioni precedenti che coinvolgono la word form, l'iponimia riguarda il word meaning. Rappresenta il concetto per cui due synset {x, x', ...} e {y, y', ...} sono utilizzabili in frasi del tipo "Un x è (all'incirca) un y". Ad esempio, acero è un iponimo di albero.

L'iponimia dà vita a una struttura gerarchica, dove il subordinato eredita tutte le caratteristiche del superordinato aggiungendone di proprie.

Meronimia

Costituisce una relazione part-of. Qui i due synset sono utilizzabili in frasi del tipo "Un y ha un x (come sua parte)"; si dice allora che x è un meronimo di y. Anche qui si può creare una struttura

gerarchica, ma con più eccezioni perché a volte più y possono avere un x come sua parte. La relazione inversa della meronimia è detta olonimia.

Nomi

In WordNet i nomi sono rappresentati in una struttura gerarchica sulla base delle relazioni di iponimia, in cui ogni sostantivo ha un superordinato e delle caratteristiche proprie che lo distinguono. Queste caratteristiche sono di tre tipi:

- attributi (modifiche);
- parti (meronimia);
- funzioni (predicati).

Le parole ereditano le caratteristiche dei superordinati.

Esistono 25 categorie principali dei nomi in WordNet, ciascuno rappresentante un componente semantico primitivo (ad esempio, plant, shape, ecc.). Le categorie sono gerarchie distinte, il cui insieme cattura l'interezza dei nomi in lingua inglese.

La rappresentazione di WordNet dei nomi come un albero serve a superare le limitazioni dei vocabolari tradizionali, ovvero la mancanza di riferimenti alle parti che costituiscono la parola (le pareti cellulari e il nucleo fanno parte della cellula), l'assenza di riferimenti agli altri termini dello stesso dominio (oltre alle cellule esistono altri esseri viventi), l'assenza di riferimenti agli iponimi e le definizioni circolari (l'uso di alcune parole per spiegare il significato di altre).

Caratteristiche distintive

Come detto, ciascun sostantivo eredita delle caratteristiche dal superordinato e ne introduce di proprie; è possibile individuare un livello a metà della gerarchia dove si hanno la maggior parte delle caratteristiche distintive che caratterizzano un concetto. Questo livello viene detto basic-level e al di sopra si avranno descrizioni più corte e generiche, mentre al di sotto le parole avranno caratteristiche aggiuntive che aggiungono meno al concetto basilare.

Attributi: sono caratteristiche date dagli aggettivi che modificano il nome. Non tutti gli aggettivi sono però applicabili a un nome; alcuni di essi, se usati, verranno interpretati in maniera metaforica (un canarino si può definire affamato, ma difficilmente tirchio se non come metafora).

Parti e meronimia: i meronimi sono caratteristiche distintive che gli iponimi possono ereditare, andando a creare una stretta dipendenza intrecciata. Ad esempio, becco e ali sono meronimi di uccello, e canarino (che è un iponimo di uccello) li eredita sempre come meronimi.

Funzioni e predicati: una caratteristica funzionale di un sostantivo è la descrizione di qualcosa che normalmente svolge o si può svolgere con/su di esso. Costituiscono funzioni di un nome tutti i concetti espressi da verbi applicabili con esso (coltello-tagliare, buco-scavare). Costituisce una funzione anche la capacità per una parola di svolgere il compito di qualche altra appartenente a un concetto differente: una scatola non è una sedia, ma è possibile dire che la scatola svolge la funzione di sedia.

Verbi

I verbi sono per molti linguaggi una componente imprescindibile di una frase, mentre invece i soggetti possono essere rappresentati da sostantivi generici (it, he, she). Il verbo fornisce il contesto relazionale e semantico della frase, permettendo soltanto a un sottoinsieme ristretto di nomi di riempire i buchi della frase. Il contesto semantico e relazionale ammesso dal verbo è quindi implicitamente memorizzato nella sua definizione.

Esistono molti meno verbi dei sostantivi, ma con polisemia maggiore. Questo lascia intendere che i verbi abbiano un significato più flessibile, determinato in parte dai nomi con cui vengono associati (tagliare una pianta (abbattere) e tagliare il pane (affettare)) e dalla concretezza o astrazione del concetto che li segue (avere il mal di testa o avere la Ferrari).

Anche i verbi, in WordNet, sono suddivisi in macro-categorie rappresentanti domini semantici diversi, con l'aggiunta di una categoria particolare che racchiude tutti quei verbi detti di stato che non possono essere inseriti altrove.

Le distinzioni semantiche fra verbi non sono però le stesse dei nomi: non si può, ad esempio, parlare di iponimia fra verbi, ma serve usare una relazione diversa. Essa prende il nome di troponimia e descrive il fatto che un certo verbo X descrive l'azione di un altro verbo Y in qualche modo particolare (tirare significa muovere verso di sé). Tuttavia si viene a creare una gerarchia molto meno strutturata di quella dei nomi, che non è nemmeno un albero poiché all'interno della stessa macro-categoria semantica ci possono essere più radici (esempio, give e take per i verbi di possesso). Inoltre i livelli sono meno, con concentrazioni di troponimi in un particolare livello.

BabelNet

Descrizione

BabelNet nasce dalla fusione fra WordNet e Wikipedia; creare risorse lessicali come WordNet richiede infatti un grande lavoro manuale da parte di personale esperto, che deve essere svolto da zero per ogni linguaggio. Wikipedia invece fornisce grandi informazioni semantiche su molti termini, incluse le named entities (l'azienda Apple, per esempio), ma pecca di copertura lessicografica appiattendo alcuni sensi in un'unica voce (apple il frutto e apple l'albero, ad esempio) Ha il grande vantaggio di fornire queste informazioni in più lingue.

La fusione dei due progetti porta ad avere una rete semantica molto estesa e allo stesso tempo dettagliata in un insieme di lingue molto ampio.

Come visto, WordNet si basa sul concetto di synset: ad esempio, {play$_n^1$, drama$_n^1$, dramatic play$_n^1$} è il synset di play inteso come lavoro drammatico teatrale. L'apice di ogni parola indica il particolare senso e il pedice la PoS della parola (in questo caso, nome). Parole polisemiche appaiono in più synset con apici diversi.

Wikipedia è invece formata da pagine riferite a un concetto o una named entity. Il titolo di ogni pagina ha opzionalmente un'etichetta fra parentesi che specifica il significato del lemma se ambiguo (ad esempio Play (theatre)). Certe pagine hanno una infobox che rappresenta un riassunto degli attributi principali dell'entità e sono memorizzate tramite grafi RDF. Esistono poi relazioni fra pagine di vario tipo:

- pagine di redirect: inoltrano a pagine che contengono effettivamente le informazioni sul concetto, modellando quindi la sinonimia;
- pagine di disambiguazione: raggruppano link alle pagine per i possibili concetti rappresentati da un termine, modellando quindi la polisemia;
- link interni: collegano altre pagine di concetti correlati a una o più pagine;
- link inter-lingua: collegano a pagine in lingue diverse per lo stesso concetto;
- categorie: raggruppano pagine per argomento.

BabelNet codifica l'informazione tramite un grafo direzionato dove i nodi sono i concetti e le named entities e gli archi (etichettati) collegano i nodi secondo varie relazioni semantiche descritte dall'etichetta (is-a, part-of, ecc.). Esiste anche la possibile relazione ε che indica un collegamento semantico non specificato. I nodi contengono un insieme di lessicalizzazioni in più lingue dello stesso concetto e prendono il nome di Babel synsets.

Costruzione

La costruzione del grafo di BabelNet avviene in tre passi:

1. Si raccoglie automaticamente l'informazione da WordNet (tutti i sensi delle parole come concetti e le relazioni lessicali e semantiche fra sysnset come relazioni) e da Wikipedia (le pagine come concetti e i link come relazioni semantiche non meglio specificate). Dato che le informazioni di WordNet e Wikipedia possono sovrapporsi sia come concetti che come relazioni, l'intersezione è fusa assieme.
2. Le realizzazioni lessicali dei concetti disponibili in più lingue sono raccolte nei Babel synset da Wikipedia (se disponibili) o da traduzioni automatiche di frasi in cui il concetto compare.
3. I Babel synset vengono infine collegati fra di loro stabilendo relazioni semantiche. Queste relazioni sono create sulla base delle relazioni recuperate da WordNet e Wikipedia valutate secondo una certa misura di affinità.

Passo 1: collegamenti fra Wikipages e sensi di WordNet

Si vuole determinare il collegamento

$\mu : Senses_{Wiki} \rightarrow Senses_{WN} \cup \{\varepsilon\}$

ovvero tale che per ogni Wikipage $w \in Senses_{Wiki}$ si abbia:

$\mu(w) = s \in SensesWN(w)$ se un collegamento esiste, oppure $\mu(w) = \varepsilon$ altrimenti.

w è dato dal titolo della pagina di Wikipedia, meno l'eventuale etichetta disambiguante.

La ricerca del collegamento avviene a sua volta in tre fasi:

1. algoritmo di mapping, che individua per una Wikipage il senso di WordNet che massimizza la probabilità di fornire un senso adeguato alla pagina;
2. creazione di un problema di disambiguazione sulla base del mapping, dove un contesto di disambiguazione è fornito sia per i sensi di WordNet che per le Wikipages;

3. stima della probabilità condizionata di un particolare senso WordNet data una Wikipage basata sui contesti di disambiguazione.

Vediamo l'algoritmo di mapping.

Algorithm 1 The mapping algorithm.

Input: $Senses_{Wiki}$, $Senses_{WN}$
Output: a mapping $\mu : Senses_{Wiki} \rightarrow Senses_{WN} \cup \{\epsilon\}$

1: **for each** $w \in Senses_{Wiki}$
2: $\quad \mu(w) := \epsilon$
3: **for each** $w \in Senses_{Wiki}$
4: $\quad$ **if** $|Senses_{Wiki}(w)| = |Senses_{WN}(w)| = 1$ **then**
5: $\quad\quad \mu(w) := w_n^1$
6: **for each** $w \in Senses_{Wiki}$
7: $\quad$ **if** $\mu(w) = \epsilon$ **then**
8: $\quad\quad$ **for each** $d \in Senses_{Wiki}$ s.t. d redirects to w
9: $\quad\quad\quad$ **if** $\mu(d) \neq \epsilon$ and $\mu(d)$ is in a synset of w **then**
10: $\quad\quad\quad\quad \mu(w) :=$ sense of w in synset of $\mu(d)$; **break**
11: **for each** $w \in Senses_{Wiki}$
12: $\quad$ **if** $\mu(w) = \epsilon$ **then**
13: $\quad\quad$ **if** no tie occurs **then**
14: $\quad\quad\quad \mu(w) := \underset{s \in Senses_{WN}(w)}{\text{argmax}} \; p(s|w)$
15: **return** μ

Il cuore dell'algoritmo è l'ultimo caso, in cui non si è individuato un mapping certo e si deve ricorrere alla massimizzazione della probabilità condizionata p(s|w) per trovare il senso di WordNet che più si avvicina alla Wikipage.

È possibile trasformare questo calcolo nel seguente modo, poiché p(w) non partecipa alla massimizzazione essendo costante indipendente da s:

$$\mu(w) = \underset{s \in Senses_{WN}(w)}{argmax} \; p(s \vee w) = \underset{s}{argmax} \frac{p(s, w)}{p(w)} = \underset{s}{argmax} \, p(s, w)$$

Il miglior senso s è quindi ottenibile massimizzando la probabilità congiunta p(s,w).

Passando al secondo passo, si usa la stessa tecnica del Word Sense Disambiguation (l'algoritmo di Lesk) per definire i contesti di disambiguazione che serviranno per il calcolo della probabilità congiunta. Si trovano i contesti, che non sono altro che insiemi di parole i cui sensi sono associati all'input secondo alcune relazioni semantiche, sia per i sensi di Wikipedia che per quelli di WordNet. Questi insiemi di parole aiutano a individuare un possibile collegamento nel calcolo della nostra μ.

Per una Wikipage si usa l'etichetta per disambiguare il termine, i nomi delle pagine che sono linkate nella pagina w, i nomi delle pagine che contengono link a w e le categorie. Il contesto di disambiguazione viene definito Ctx(w) e, per esempio, è Ctx(Play (theatre)) = {theatre, literature, comedy, ...}

La creazione di contesti di disambiguazione per un senso WordNet invece si basa sui suoi sinonimi presenti nel synset, i termini nei synset che sono iperonimi e iponimi del synset di s (quindi generalizzazioni e specializzazioni) e le parole contenute nel gloss del synset di s. Un esempio è $Ctx(play_n^{1)}$ = {drama, dramatic play, composition work, ...}.

La probabilità congiunta, dati i contesti di disambiguazione, è calcolabile come:

$$p(s,w)=\frac{score(s,w)}{\sum_{\substack{s'\in Senses_{WN}(w)\\ w'\in Senses_{Wiki}(w)}} score(s',w')}$$

Si giunge quindi al terzo e ultimo passo, ovvero il calcolo della funzione di score per poter finalmente calcolare la probabilità congiunta. Esistono due approcci:

- Bag-of-words, in cui $score(s,w)=|Ctx(s)\cap Ctx(w)|+1$ ovvero è pari alle parole in comune nei due contesti di disambiguazione più uno (fattore di addolcimento). Metodo facile, ma che non sfrutta l'informazione strutturale presente in WordNet e Wikipedia.
- Grafo, che consiste nel costruire un grafo partendo da Ctx(w) e collegando i possibili sensi di w ai sensi delle parole in Ctx(w). Questo approccio fornisce un mapping anche quando l'intersezione dei Ctx di WordNet e Wikipedia è vuota. La costruzione del grafo avviene nel seguente modo:
  - Si prendono come nodi tutti i sensi di WordNet per w (Wikipage) e i sensi di WordNet per tutte le parole contenute nel Ctx(w). Non ci sono inizialmente archi.
  - Si collegano i nodi in base ai percorsi trovati fra di loro in WordNet. Ovvero, si fa un ricerca in profondità in WordNet per ogni nodo e ogni volta che si trova un synset appartenente al nostro grafo (quindi un altro nodo), si aggiungono tutti i synset visitati e i loro collegamenti al grafo.
  - Si ottiene un sottografo di WordNet e si procede a calcolare la funzione score sulla sua base:

    $score(s,w)=\sum_{cw\in Ctx(w)}\sum_{s'\in Senses_{WN}(cw)}\sum_{p\in paths_{WN}(s,s')} e^{-(lenght(p)-1)}$ che intuitivamente significa che tanto maggiore è la distanza fra gli elementi che considero, tanto meno quegli elementi contribuiranno allo score.

Passo 2: traduzione dei Babel synset

Il passo 1 è stato svolto sulle Wikipage in inglese, collegandole ai sensi di WordNet, creando dei Babel Synset $S\cup W$ dove S è il synset di s e W è l'insieme {w, l'insieme delle pagine con link a w, le pagine in lingue diverse del senso di w e le pagine in lingue diverse con link alle pagine delle traduzioni di w}. Ad esempio, dato il mapping $\mu(Play(theatre))=play^1_n$ si ha il Babel synset {play$_{en}$, Bühnenwerk$_{de}$, opera teatrale$_{it}$}.

Esiste però il problema che alcuni concetti potrebbero essere presenti solo in WordNet o Wikipedia e che se disponibili in entrambi, le voci di Wikipedia siano disponibili soltanto in alcune lingue.

Per risolvere la mancanza di traduzioni, per ogni significato di WordNet delle parole che compongono il Babel synset si estraggono le frasi in cui compare lo stesso significato da SemCor, un corpus contenente migliaia di frasi le cui parole sono annotate con i synset di WordNet. Si estraggono anche le frasi dalle Wikipages contenenti il link al significato d'interesse.

Si ottiene così un insieme di frasi che viene poi tradotto tramite sistemi automatizzati: la traduzione più frequente del termine viene aggiunta al Babel synset. La traduzione sarà specifica del senso dato che è derivante da frasi in cui i termini del synset apparivano.

Passo 3: aggiunta delle relazioni semantiche

Il passo finale consiste nello stabilire le relazioni semantiche fra i Babel synset (che a questo punto sono multi-lingua).

Si raccolgono le relazioni da WordNet (i collegamenti fra synset) e Wikipedia (tramite la struttura a hyperlink); in seguito, si procede ad assegnare loro un peso secondo una misura di correlazione data dal coefficiente di Dice.

Il procedimento per il calcolo consiste, per ogni relazione di WordNet fra un synset s e uno s', nella raccolta di tutti i sinonimi e le parole del gloss sia di s che di s', oltre che ai synset a loro direttamente connessi. Queste informazioni costituiscono due bag of word S e S'. Il coefficiente di Dice è calcolato dalla formula $\frac{2 * |S \cap S'|}{|S| + |S'|}$ e costituisce il peso della relazione fra i Babel synset rispettivamente contenenti s e s'.

FrameNet

Descrizione

FrameNet consiste in un ulteriore passo in avanti rispetto a WordNet e BabelNet in quanto permette di assegnare il senso non più soltanto alle singole parole ma a intere frasi. Si tratta di uno strumento per l'analisi semantica che vedremo in realtà poter essere usato per vari scopi.

Il concetto fondamentale è quello del frame, ovvero un quadro di riferimento per un certo contesto. Si tratta di quanto visto per la rappresentazione della conoscenza.

FrameNet è costituito da lessico in lingua inglese e il suo compito è quello di definire i frame, riempirli con i termini che ben si incastrano e descrivere il concetto espresso dal frame. Esiste poi una parte di frasi annotate con le notazioni di frame, correlate da informazioni di valenza (vedremo più avanti).

Ogni frame tratta una situazione più o meno precisa e i termini usati all'interno del frame sono ad esso relati. Si dice che queste parole evocano il frame.

Analogamente a quanto accade in WordNet con i synset, esiste il problema della polisemia che viene risolto raggruppando i significati analoghi delle parole in unità, qui definite **Lexical Units (LU)** [*un dict contenente tutte le LU per questo frame. Le chiavi in questo dict sono i nomi delle LU e il valore per ogni chiave è di per sé un dict contenente informazioni sulla LU (vedi la funzione lu() per maggiori informazioni.]*. Anche FrameNet effettua delle suddivisioni dei significati estremamente fini, come accadevain WordNet.

Esempio di un frame

Prendiamo un frame per studiarne le parti: scegliamo quello che esprime il concetto di "vendetta" (revenge).

Una descrizione ad alto livello della vendetta consiste in una situazione in cui un soggetto A ha compiuto uno sgarbo a B e B compie un'azione per portare a sua volta danno ad A, agendo all'infuori di ogni contesto legale o istituzionale.

Una serie di parole sono associabili al concetto di vendetta, come i nomi (revenge, vengeance, retaliation, ...), verbi (avenge, revenge, retaliate (against), get even (with), ...), aggettivi (vengeful, vidictive) e V/N Phrases (take revenge, exact retribution, ...).

Chi lavora a FN come prima cosa deve dare una definizione (una frase che spiega il concetto espresso) del frame e deve costruire un insieme di componenti, associandogli un nome adeguato: queste componenti si chiamano **Frame Elements (FE)** [*Le chiavi in questo dict sono i nomi degli FE relativi (es. 'Body_system') e i valori sono dict contenenti la seguenti info: def, name, id, semtype*] e servono per etichettare le parti di una fraseche tratta di quel particolare frame.

Nel nostro caso, il frame Revenge contiene i seguenti FE:

- avenger;
- offender;
- injury;
- injured_party;
- punishment.

Come è possibile notare, i nomi dei FE sono strettamente legati al contesto.

In seguito si passa poi a estrarre degli esempi da un corpus di frasi, privilegiando quelli che mostrano quanti più contesti sintattici del frame possibili. I costituenti di queste frasi-esempio vengono annotati con i nomi del frame element che rappresentano.

Dagli esempi annotati è possibile osservare che parole diverse nello stesso frame esibiscono una variazione di come i FE sono grammaticalmente realizzati; in altri termini, la probabilità che un particolare costituente svolga il ruolo di un certo FE nella frase dipende dalle parole presenti nella frase stessa. Questo è il concetto di variazione di valenza.

Ad esempio, si ha il 50% di possibilità (6 occorrenze su 12) che il ruolo di "avenger" e di "injured_party" siano svolte da un NP se la frase è composta dai FE avenger, injured party, offender e punishment.

12 exx TOTAL	Avenger	Injured_Party	Offender	Punishment
2 exx	-- --	NP Ext	-- --	-- --
1 exx	-- --	NP Ext	PP Comp	-- --
6 exx	NP Ext	NP Obj	-- --	-- --
1 exx	NP Ext	NP Obj	-- --	PP Comp
1 exx	NP Ext	NP Obj	-- --	PPing Comp
1 exx	NP Ext	NP Obj	PP Comp	PPing Comp

Interrogare FrameNet

È possibile sfruttare le diverse viste del database di FN per ottenere informazioni di diverso tipo.

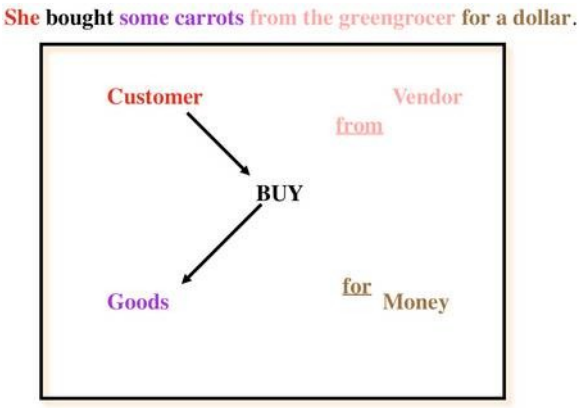
Un primo esempio è la richiesta di come sintatticamente venga realizzato il ruolo del FE “offender” (pronome you, preposizione on, against, ecc.).

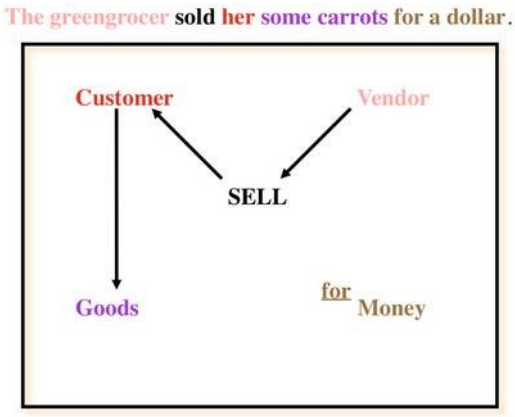
È anche possibile interrogare FN per ottenere tutte le parole associate a un dato frame o i pattern sintattici in cui ogni parola è usata per esprimere il frame.

Esistono poi tante altre query su cui non ci soffermiamo, ma va detto che al momento FN non mantiene informazioni sulla frequenza in quanto i risultati dipendono dal corpus utilizzato per estrarre le frasi di esempio (un corpus finanziario darà frequenze diverse da un corpus medico).

Esempi

Negli esempi che seguono è possibile comprendere come lo stesso contesto sia esprimibile tramite l’uso di frame diversi in base alle parole usate. Il risultato dell’azione e i FE non cambiano, ma si tratta di frame differenti. Per la sequenza completa vedere le slide.





Comprensione del testo

Un altro uso di FrameNet è estrarre informazioni sui modi in cui i termini vengono usati. Prendiamo l’esempio dell’estrazione automatizzata di dati dalla sezione di cronaca nera di un giornale: vengono selezionati i termini che appaiono più frequentemente (si presume quindi che esistano elementi a monte nella catena del NLP, come i tagger e i parser che diano informazioni corrette).

FN è utilizzabile per svolgere la disambiguazione dei termini (WSD), la composizione semantica (stabilire il ruolo dei termini in un frame, separando reggenti da dipendenti), la scelta fra analisi sintattiche alternative e la costruzione di un vocabolario di termini correlati all’argomento (più termini collegati fra di loro sono infatti presenti in un frame).

Tramite le relazioni sintattiche si vede come gli elementi sono inseriti nella struttura del frame: i dettagli temporali devono essere calcolati sulla base della data dell’articolo (ad esempio, next Tuesday) e le anafore (riferimenti impliciti ad altri termini) devono essere risolte. Questo è un compito per nulla semplice e banale in base al tipo di anafora in questione.

Un uso classico di FN è dato dalla scelta di una parola fra quelle rilevanti per comprendere quali frame vengano attivati. Dopo aver identificato i ruoli semantici si cerca di accoppiare le necessità

semantiche di ogni frame (i FE) con le parti della frase in esame: il frame che meglio si adatta è quello scelto.

Altro uso è quello della disambiguazione. FN non fornisce meccanismi diretti per questo scopo, ma offre descrizioni di frame collegati e legami semantici fra parole su cui si basano i giudizi di coerenza che decidono quale sia il significato corretto del termine.

ConceptNet e COVER

Il common-sense

Il senso comune è l'insieme dei fatti e delle conoscenze possedute dalla maggior parte della popolazione, come ad esempio il fatto che per aprire una porta bisogna girare la maniglia. Questa conoscenza è di vario tipo (spaziale, temporale, fisica, ecc.) e dipende dal periodo e contesto culturale. Data la sua diffusione, spesso è omessa nelle comunicazioni sociali come i testi; la comprensione di un testo richiede infatti una grande quantità di senso comune che soltanto le persone fino ad ora possiedono.

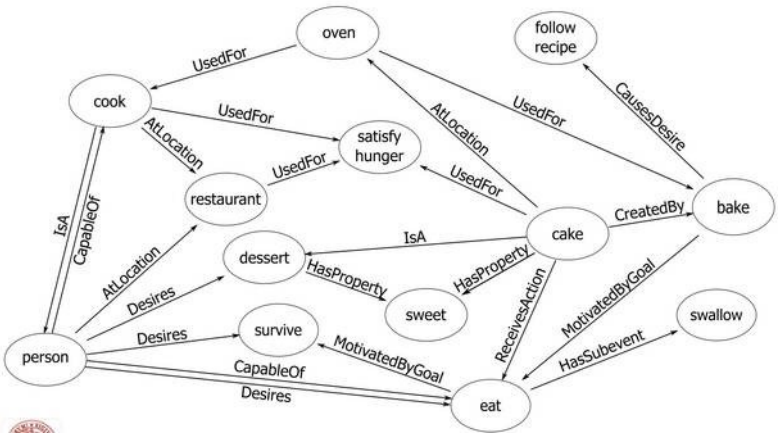
Data la vastità del commonsense (decine di milioni di fatti), per le macchine è sempre stato difficile possederla e utilizzarla.

Un primo approccio alla rappresentazione del commonsense è stato Cyc, un tentativo di organizzare la conoscenza comune in un framework logico. Le asserzioni sono manualmente create e inserite in Cyc ed è necessario rappresentare il testo da analizzare in un linguaggio proprietario, CycL, affinché sia possibile sfruttare la logica descritta in Cyc per delle assunzioni. Ma questa operazione di traduzione è complessa a causa delle ambiguità del linguaggio naturale che devono essere risolte per avere un mapping nella forma logica priva di ambiguità di CycL.

ConceptNet

Descrizione

ConceptNet invece sfrutta un approccio automatizzato per la rappresentazione del common sense, che verrà descritta in dettaglio più avanti. ConceptNet è una rete semantica i cui nodi sono frammenti di frasi in lingua inglese (ad esempio, play tennis) legati fra di loro da una ventina (nelle prime versioni) di relazioni semantiche.



ConceptNet

Si supera quindi la definizione di nodo di WordNet come elemento puramente lessicale per includere concetti composti di alto livello, permettendo di rappresentare più concetti incontrati nella vita quotidiana (gli eventi, le azioni, ecc.). Così facendo si perde però l'annotazione del senso delle parole.

Oltre ai nodi, viene estesa anche la definizione degli archi (le relazioni) rispetto a WordNet, andando ad aggiungere nessi temporali, spaziali, di capacità, ecc.

Si viene a creare una conoscenza molto più informale e di natura pratica rispetto a quella rigorosa di WordNet, riuscendo a catturare anche quella conoscenza "defeasible", ovvero spesso vera, ma non sempre. Un esempio è che un effetto della caduta dalla bici è farsi male: spesso è vera, ma in alcuni casi no.

ConceptNet eccelle nel ragionamento contestuale, ovvero nella disambiguazione di alcune parole e descrizioni sulla base del contesto della frase. È il meccanismo che permette inoltre di comprendere il sarcasmo e l'ironia, che avvengono proprio perché si superano le aspettative date dal contesto. Con il ragionamento contestuale è possibile comprendere le collocazioni spaziali e temporali di alcuni eventi descritti, o di trovare analogie per un concetto sconosciuto.

ConceptNet è particolarmente adeguato al ragionamento su di un contesto perché la maggior parte dei suoi collegamenti fra concetti sono piuttosto generici (detti k-lines), che permettono di aumentare la connessione della rete semantica e rendere più probabile il mapping di concetti con quelli presenti in ConceptNet. Inoltre, la sua abilità è anche dovuta al fatto che le parole hanno sempre dei sensi e connotazioni che invece mancano ai simboli logici.

Costruzione

La costruzione di questa risorsa linguistica parte con il sito per l'Open Mind Common Sense (OMCS), dove a volontari viene chiesto di rappresentare vari tipi di commonsense attraverso attività. Un esempio di tale attività è la compilazione di spazi vuoti in frasi del tipo "A knife is usedto_____", che fornisce informazioni sull'uso di un coltello. Vengono così raccolte frasi che rappresentano la conoscenza comune.

Da questo database di frasi vengono estratte le asserzioni che costituiscono ConceptNet attraverso un processo automatico.

Un insieme di regole di estrazione viene applicato al corpus OMCS, andando a produrre delle asserzioni di ConceptNet di tipo relazione binaria. Si fa leva del fatto che le frasi OMCS sono semi-strutturate (l'esempio precedente è un caso del template "X is used to Y"), andando a usare delle espressioni regolari per questa estrazione.

Si passa poi ad una fase di rilassamento della rete semantica per migliorare la connessione della rete e diminuire i buchi semantici. In questa fase si uniscono le asserzioni duplicate tenendo traccia della loro frequenza e si esporta verso l'alto (ovvero tramite la relazione IsA) la conoscenza che si ha per i figli ai padri. Un esempio è, dato che la maggior parte dei frutti ha proprietà "dolce", allora anche l'entità frutto la possiederà. Durante il rilassamento si creano anche generalizzazioni lessicali sfruttando i synset di WordNet (ad esempio, "buy food" è un tipo di "buy") e si uniscono differenti forme lessicali (bike con bicycle).

Ragionamento pratico sul commonsense

Un primo uso è l'individuazione di vicini contestuali, ovvero nodi di ConceptNet in relazione con un nodo di input. La relazione non è semplicemente di una distanza di un nodo da un altro, ma tiene conto anche della forza e del numero di percorsi che congiungono i nodi. I risultati sono facilmente verificabili con la propria conoscenza comune.

Un secondo uso di CommonNet è la creazione di analogie: si tratta del processo alla base dell'apprendimento in cui si decompone un'idea nelle sue parti costituenti e si individua il concetto che più parti ha in comune con quelle derivanti dalla decomposizione. CommoNet svolge il compito tramite la verifica di quali archi entranti un nodo condivide con un altro.

Si può usare CommonNet anche per attraversare il grafo a partire da un nodo seguendo un singolo tipo di relazione: ad esempio, il concetto di posizione spaziale porta dal nodo "Roma" al nodo "Lazio", e poi al nodo "Italia" e poi "Europa", ecc.

Si può anche effettuare della disambiguazione e classificazione, andando a collocare all'interno della rete di CommonNet dei documenti d'esempio; i documenti successivi vengono classificati secondo la vicinanza ad uno dei documenti d'esempio.

Valutazione

Dare un giudizio su basi di conoscenza del commonsense è generalmente difficile, perché bisogna prima stabilire cosa sia e cosa non sia conoscenza comune.

Un metodo per avere un'idea generale della bontà è la verifica di come i concetti contenuti nella rete siano espressi: se sono brevi, ovvero formati da poche parole, sono ritenuti facili. In CommonNet circa il 70% dei nodi ha una lunghezza al massimo di tre parole. Confrontando questo risultato con il fatto che un sintagma verbo-nome richiede almeno quattro parole ("take a dog for walk"), si capisce che la maggior parte dei nodi di CommonNet sono molto semplici.

Un altro metro di valutazione è la ripetizione delle asserzioni: dato che sono semplici, è legittimo aspettarsi che capitino di frequente. È dimostrabile però che la maggior parte dei concetti di ConceptNet appaiano una sola volta nelle frasi di OMCS e un altra grande percentuale non compaia proprio e sia stata quindi inferita (sono k-lines), dimostrando ancora la bontà di questa rete semantica.

COVER

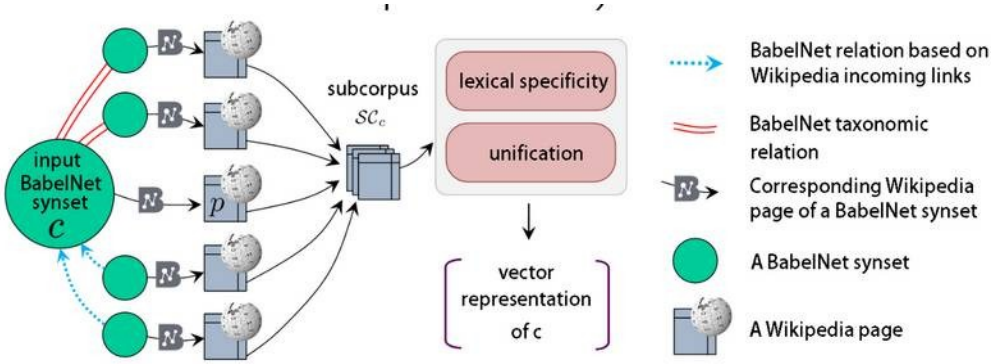
Vedere le slide.

NASARI

Descrizione

NASARI nasce con lo scopo di creare una rappresentazione vettoriale del senso di una parola che non dipenda dalla semantica distribuzionale, ovvero da quanto quella parola appaia vicino a delle co-occorrenze. L'approccio distribuzionale non riesce infatti a isolare il singolo senso della parola, ma forma dei vettori che racchiudono più sensi.

L'approccio di NASARI, essendo basato su BabelNet, sfrutta i gloss di WordNet per riuscire a creare vettori per un singolo senso.



Si raccoglie informazione sul BabelSynset in questione e la si raggruppa in un subcorpus: questo subcorpus contenente l'informazione contestuale viene confrontato con l'intero corpus di Wikipedia, calcolandone la specificità. Sulla base di essa, si ottiene una rappresentazione vettoriale del senso in esame che può essere di due tipi:

- rappresentazione lessicale di un concetto, dove le dimensioni del vettore sono dei lemmi ("tavolo" ha come dimensioni del suo vettore lessicale "gamba", "piano", ecc.);
- rappresentazione unificata, dove le dimensioni sono dei concetti (dei BabelSynset).

I due tipi di vettori sono memorizzati all'interno di NASARI come sequenze separate da tabulazioni, il cui primo elemento denota l'ID del BabelSynset, il secondo il titolo della pagina di Wikipedia associata al BabelSynset (se esistente) e i successivi elementi i lemmi o i synset (in base al tipo di vettore) con il loro peso associato dopo il carattere di underscore.

```
BabelSynsetId  WikipediaPageTitle  lemma1_weight1  lemma2_weight2
BabelSynsetId  WikipediaPageTitle  synset1_weight1
```

Rappresentazione unificata

Per generare un vettore della rappresentazione unificata, si raggruppano assieme le parole con un senso che condivide lo stesso iperonimo (osservando quindi la tassonomia di BabelNet).

Poi si calcola la specificità dell'insieme di tutti gli iponimi di questi sensi, compresi quelli che non compaiono nel subcorpus.

La costruzione del cluster di parole che condividono lo stesso iperonimo consiste in un processo implicito di disambiguazione.

Crane (machine)		
English	French	German
∗lifting device^1_n	∗dispositif de levage^1_n	∗hebevorrichtung^1_n
‡construction^4_n	navire^1_n	radfahrzeug^1_n
platform^1_n	limicole^1_n	†lenkfahrzeug^1_n
warship^1_n	⋄vaisseau^2_n	regler^3_n
electric circuit^1_n	spationef^1_n	reisebus^1_n
⋄vessel^2_n	‡construction^2_n	charadrii^1_n
boat^1_n	†véhicule^3_n	güterwagen^2_n

Uso: similarità semantica

L'obiettivo è stabilire un valore di similitudine tra due termini. Dato che i vettori non contengono valori numerici è necessario introdurre una metrica apposita.

Per ognuna delle due parole w_1 e w_2, si estrae l'insieme di concetti (ovvero BabelSynset) associati, denotato con C_{w1} e C_{w2}. Se w_i esiste in BabelNet, allora C_{wi} è dato dall'insieme di sensi di BabelNet. Altrimenti si usano gli hyperlink presenti nelle pagine Wikipedia per catturare i concetti per le parole non presenti in BabelNet.

Da C_{wi} si estraggono le corrispondenti rappresentazioni unificate (i vettori) v.

La similitudine fra le due parole è data dalla formula:

$$\text{sim}(w_1, w_2) = \max_{v_1 \in C_{w1}, v_2 \in C_{w2}} \sqrt{WO(v_1, v_2)}$$

dove

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

denota il Weighted Overlap. L'overlap O è costituito dalle dimensioni in comune fra i due vettori e WO consiste nell'osservare quanto distante sia ciascuna dimensione (rank fornisce infatti la posizione all'interno del vettore della dimensione q). Tanto maggiore sarà questa distanza e meno influirà nella sommatoria grazie all'elevamento alla -1. Il denominatore è un fattore di normalizzazione ed è trascurabile.

Riassunto automatizzato

Descrizione del problema

Un potenziale uso di NASARI è nel riassumere un testo. Quest'operazione, dato un testo in ingresso, ne estrae le parti più importanti, creandone una versione più breve che fornisce un'idea sul contenuto del documento originale, coprendo l'interezza degli argomenti con minor dettaglio e fornendo potenzialmente un giudizio critico sul lavoro dell'autore.

Esistono due tipi di riassunti:

- estratti, cioè riassunti formati riutilizzando parti del testo originale senza alterazioni;
- abstract, cioè riassunti creati rigenerando il testo estratto.

Inoltre, esistono altrettanti approcci per la riassunzione:

- approccio shallow, che si limita al massimo al livello sintattico estraendo le parti del testo e riorganizzandole in maniera coerente (produce in genere estratti). Gli elementi difficili delle frasi, come ellissi e anafore rischiano di produrre un riassunto incomprensibile;
- approcci profondi, che invece analizzano la semantica delle frasi per creare astrazioni del contenuto per poi ricrearlo in forma differente. Questo processo richiede però spesso una grande conoscenza sul dominio del testo da riassumere, rendendo il sistema difficilmente riadattabile per un altro contesto. Tipicamente vengono prodotti abstract.

Un riassunto è inoltre caratterizzato da un fattore di compressione (ovvero quanto più corto è il riassunto del testo di partenza), il pubblico a cui si rivolge (più o meno approfondito), la coerenza del testo prodotto (quanto bene si è riusciti a risolvere le anafore, quante frasi si ripetono, ecc.) e il numero di documenti da usare per il riassunto (documento singolo o multiplo).

Gli approcci profondi forniscono in genere riassunti con migliori caratteristiche degli approcci statistici (shallow), ma hanno lo svantaggio di richiedere regole per l'analisi del testo e la sua manipolazione, oltre che una conseguente lentezza nell'implementazione e la difficoltà ad essere adattati a contesti generici. D'altro canto, gli approcci shallow necessitano di grandi quantità di testo per fornire buoni risultati (data la necessità di training) e non sono in grado di manipolare l'informazione in modo astratto.

Criteri di rilevanza

Devono essere stabiliti dei criteri per estrapolare dal testo di partenza le parti salienti. Esistono numerosi approcci, più o meno sofisticati.

1. Posizione nel testo: consiste nel selezionare le frasi che appaiono in specifiche posizioni, come ad esempio l'introduzione e la conclusione.
2. Metodo del titolo: si estraggono dal titolo del documento le parole che costituiranno le keyword per individuare delle frasi importanti nel testo. L'approccio funziona bene per i documenti scientifici, ma peggio per i romanzi e altri documenti su internet come i blog.
3. Optimum Position Policy (OPP): si basa sull'assunzione che le frasi importanti siano posizionate in punti chiave conosciuti in anticipo (ad esempio grazie alla struttura del documento) o scovate attraverso il training.
4. Indizi: le frasi nel testo più importanti contengono delle parole/sintagmi che fungono da indizio, come l'uso dei superlativi, la parola "scopo", ecc.
5. Metodi basati sulla coesione: sfruttano approcci diversi per creare delle strutture semantiche per le frasi che indicano quanto connesse siano fra loro. Uno di questi approcci è quello basato su quante volte delle parole appaiano assieme nelle frasi. Per ogni paragrafo si

individua un insieme di altri paragrafi correlati sulla base delle somiglianze delle parole che appaiono.

Uno degli algoritmi più semplici per l'estrazione di testo è l'approccio non supervisionato. Si selezionano le frasi che contengono più parole salienti o informative.

La salienza è definita sulla base della topic signature, ovvero l'insieme dei termini più importanti del testo. Si potrebbe utilizzare la frequenza delle parole per definire i termini più importanti, ma spesso capita che quelli che appaiono maggiormente non hanno grande valore rispetto agli argomenti trattati (si pensi alle congiunzioni e agli articoli, ad esempio). Si preferisce allora osservare la specificità del lessico.

Rappresentazione del senso dei verbi

La necessità e l'approccio

In molte applicazioni reali è necessario utilizzare l'informazione semantica, perché a volte la semplice ricerca di termini in un documento porta ad ambiguità. I termini ricercati possono infatti comparire in punti della frase che però sono ininfluenti per la risposta che si vuole dare. I verbi sono l'elemento in grado di fornire una grande quantità di informazione semantica.

L'idea chiave alla base degli strumenti che vedremo è che gran parte del comportamento di un verbo, ossia la forma e l'interpretazione dei suoi argomenti, è determinato dalla sintassi. Usando la realizzazione sintattica si potrà investigare il significato di un verbo.

VerbNet

È una rete che unisce la sintassi e il significato semantico dei verbi inglesi. È indipendente dal dominio e contiene collegamenti anche ad altre risorse lessicali come WordNet, FrameNet e PropBank (vedere più avanti).

I verbi sono organizzati in classi all'interno di VerbNet, le quali sono caratterizzate da un insieme di informazioni come i ruoli tematici (l'agente, il paziente, ecc.), le preferenze di selezione per gli argomenti (ovvero le restrizioni semantiche poste a ciascun ruolo) e dei frame sintattici. Questi frame descrivono ogni possibile forma sintattica utilizzabile per i verbi nella classe con associata una rappresentazione semantica.

Le differenze nella sintassi aiutano a disambiguare il senso di un verbo: ad esempio, "left" in presenza di un complemento di luogo avrà significato di "lasciare", mentre in sua assenza potrà significare "uscire". A volte però queste differenze non bastano e serve analizzare le relazioni tra il predicato e i suoi argomenti; per questo in VerbNet sono presenti tutti gli elementi appena elencati, che vanno proprio a stabilire un nesso tra sintassi e semantica del verbo.

Le classi di VerbNet sono basate sulla classificazione di Levin.

Classi di Levin

Si trattò del primo approccio alla classificazione dei verbi, raggruppandoli in classi in base al loro comportamento semantico e ai pattern morfo-sintattici.

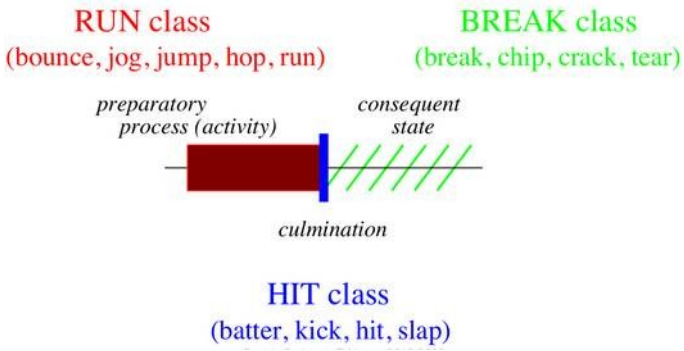
Ogni classe ha un riferimento univoco, un insieme di verbi membri e delle proprietà. Queste proprietà sono le alternation possibili per la classe, ovvero le forme diverse con cui gli argomenti del verbo possono essere espressi. A volte esistono delle alternation particolari, contraddistinte da un *, che denotano delle realizzazioni lessicali errate per affermare che un certo verbo della classe non può usare quell'alternation.

Le classi di Levin però presentavano dei problemi, fra cui la mancanza di omogeneità semantica fra i verbi contenuti e il fatto che lo stesso verbo potesse essere contenuto in più classi.

Classi di VerbNet

Come accennato, le classi di VerbNet si basano su quelle di Levin apportando opportune modifiche per risolvere i problemi. Introducono anche una rappresentazione della parte dell'evento a cui quella classe si riferisce.

Infatti, una teoria dice che i verbi si riferiscano a degli eventi che possono essere suddivisi in tre parti: un processo preparatorio, la culminazione e lo stato conseguente.



La struttura di una classe di VerbNet è la seguente:

- ruoli tematici
- frame sintattici
  - realizzazione sintattica
  - significato semantico
  - predicati semantici con funzione temporale (le parti dell'evento)
- preferenze di selezione (restrizioni selezionali) degli argomenti in ogni frame

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)

I ruoli tematici sono usati per fornire quanta più informazione possibile per la classe e sono volutamente specifici per aiutare nella differenziazione fra classi. Sono in numero limitato (circa 20). Tra i più importanti sicuramente Actor, Agent, Cause, Instrument, Patient, Theme, Time.

Le restrizioni selezionali provengono da una tassonomia basata su una variante di WordNet, con unica relazione di tipo Is-A. Pongono dei limiti a cosa può svolgere un certo ruolo tematico della classe.

I frame sintattici descrivono le realizzazioni possibili per i verbi della classe, come transitiva, intransitiva e molte alternation derivanti dalle classi di Levin. Contengono anche una congiunzione di predicati semantici che catturano la semantica della realizzazione, compresa la posizione temporale all'interno dell'evento (start, during, end, result). I predicati semantici possono essere generali o specifici per la classe.

Le classi di verbi sono organizzate in maniera gerarchica, andando a ereditare predicati semantici, ruoli tematici e frame sintattici dal genitore.

Proposition Bank (PropBank)

Si tratta di informazione aggiuntiva di tipo predicato-argomento (ovvero etichette semantiche) alle strutture sintattiche (alberi) del Penn Treebank. Per ogni nodo di un albero, PropBank assegna un ruolo semantico.

PropBank usa il concetto di RoleSet, che descrive un insieme di ruoli derivanti da un uso particolare di un verbo, associabile a sua volta a un insieme di frame sintattici che descrivono le variazioni sintattiche ammesse per quell'insieme di ruoli.

I ruoli di un RoleSet sono detti argomenti e sono etichettati con un numero da 0 a 9. Si usa un numero e non una descrizione testuale per evitare il problema del consenso fra annotatori. In genere ARG0 denota l'agente e ARG1 il paziente, ma non è una certezza e più sale il numero dell'argomento e meno è possibile fare questa generalizzazione.

Esistono poi degli argomenti particolari che svolgono il ruolo di modificatori per aggiungere informazione sull'evento (come, dove, quando si è svolto). Un esempio è ArgM-TMP che denota il tempo dell'evento.

Come detto, a un RoleSet è possibile associare dei frame sintattici: in tal caso il costrutto prende il nome di Frameset. È importante tenere presente che un verbo polisemico può avere più Frameset se

le sue differenze di significato sono abbastanza diverse e che i Frameset di PropBank sono diversi da quelli di FrameNet.

Un Frameset infatti rappresenta una distinzione di senso molto grezza, a cui corrispondono molteplici sensi di WordNet. Questo senso molto ampio consente però ai verbi polisemici di essere inseriti in Frameset ritenuti corretti.

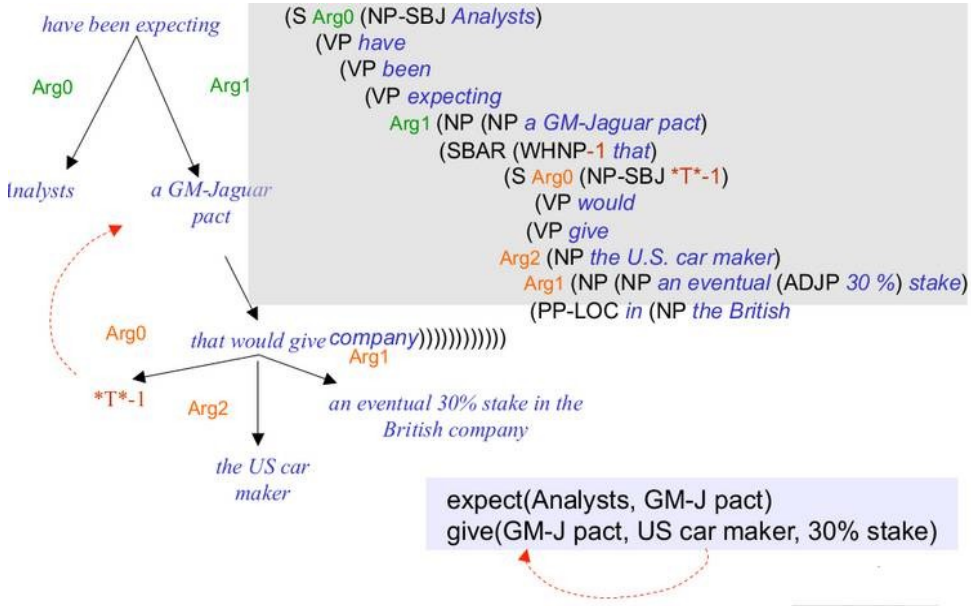
- Frameset *accept.01* "take willingly"
  - Arg0: *Acceptor*
  - Arg1: *Thing accepted*
  - Arg2: *Accepted-from*
  - Arg3: *Attribute*

Esempio di Frameset

Un esempio di frase annotata con il contenuto di questo frameset è il seguente:

Ex: [*Arg0* He] [*ArgM-MOD* would][*ArgM-NEG* n't] accept [*Arg1* anything of value] [Arg2 from those he was writing about]. (wsj 0186)

Nella realtà la frase è rappresentata ad albero (deriva da un albero sintattico del Penn Treebank) e, nel caso di più verbi, sono necessari più Frameset.



Esempio di frase più compless