Passo 1
mapping
fra WordNet
e Wikipedia

Formalmente, dato l'intero set di pagine di Wikipedia $\mathbf{Senses_{Wiki}}$ e tutti i concetti di WordNet $\mathbf{Senses_{WN}}$, un mapping è una funzione **μ** del tipo:

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

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

$$\mu(w) = \begin{cases} s \in Senses_{WN}(w) & \text{se un link può essere stabilito} \\ \varepsilon & \text{altrimenti} \end{cases}$$

Nota:
Si noti la presenza di ε come valore legittimo di μ, qualora una certa Wikipage non abbia un concetto associabile in WordNet (*da non confondere la ε relativa alle relazioni semantiche non specificate*).

In sostanza la funzione μ, data una Wikipage `w` restituisce il senso `s` di WordNet ad essa associato.

Nota:
Come sappiamo, un senso in WordNet è rappresentato da un synset. D'altronde, un qualsiasi concetto non è rappresentabile senza l'utilizzo di una delle parole che sono contenute nel synset. Quando parliamo di senso `s`, ci riferiamo quindi a un determinato termine s_n^i, cioè a un preciso senso di un determinato lemma (*parola*). Talvolta ci riferiremo a questo oggetto col termine di `sense-id`.

Data una Wikipage `w` il lemma di tale pagina è dato dal suo titolo o, in un titolo composto, dal token principale (se è presente un `label`).

Es. play è il lemma per la pagina Play(theatre)

Es. la nostra funzione μ mapperà la Play(theatre) sul senso $play_n^1$, ovvero:

$$\mu(Play\ (theatre)) = play_n^1$$

Outline del problema di mapping

Entriamo ora nei dettagli implementativi dell'algoritmo che effettua il mapping fra WordNet e Wikipedia.

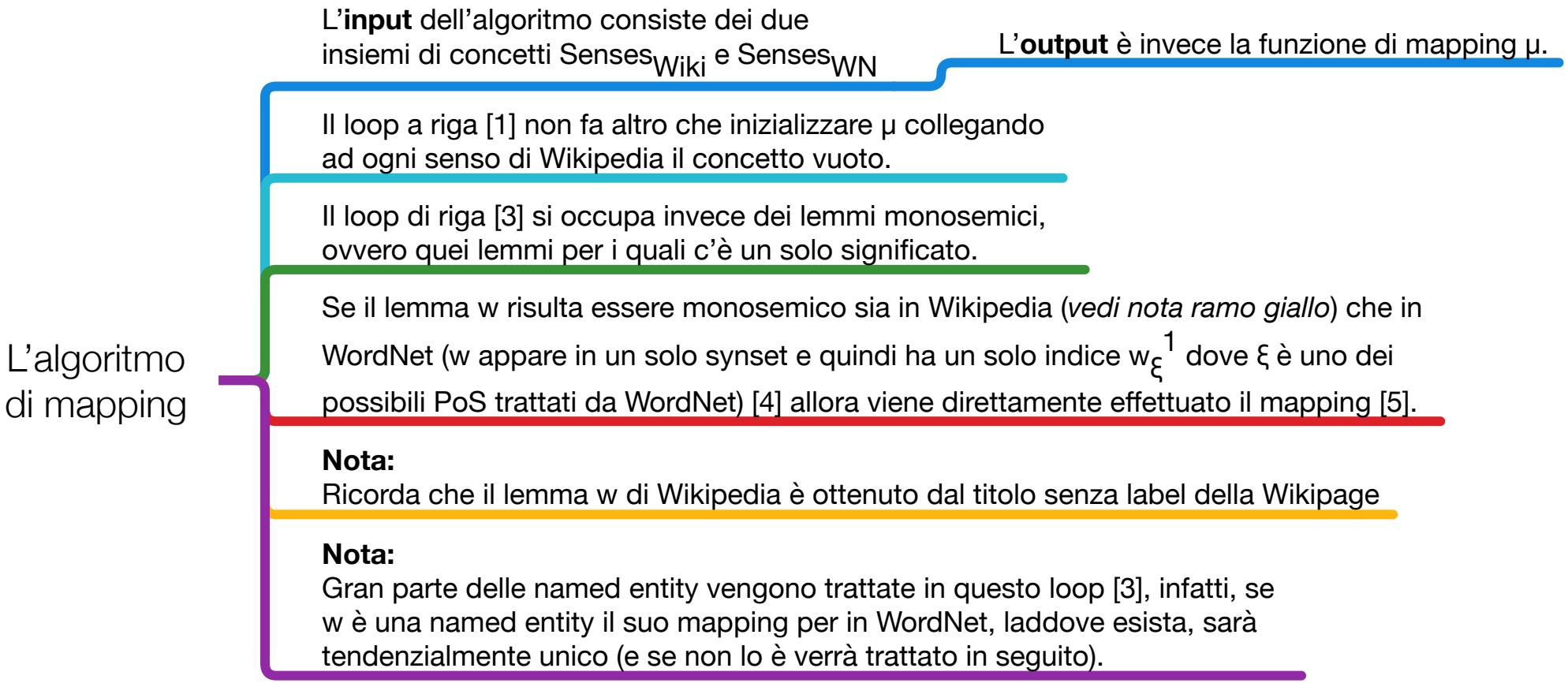
Tale algoritmo tratta prima di tutto i sensi monosemici e i link di redirection (*i link attivi nel testo di una wikipage*); dopodiché, data una Wikipage trova il senso di WordNet che massimizza la probabilità di essere un concetto adeguato corrispondente alla pagina.

Al fine di calcolare la probabilità appena descritta il problema di mapping viene trasformato in un problema di disambiguazione nel quale sia Wikipedia che WordNet forniscono un contesto. Infine, come vedremo, per risolvere il problema di disambiguazione necessiteremo di calcolare una probabilità condizionale basata su una funzione di scoring.

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

1: **for each** $w \in Senses_{\text{Wiki}}$
2: $\quad \mu(w) := \epsilon$ — mapping vuoto
3: **for each** $w \in Senses_{\text{Wiki}}$
4: $\quad$ **if** $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$ **then** — lemma monosemico
5: $\quad\quad \mu(w) := w_n^1$
6: **for each** $w \in Senses_{\text{Wiki}}$
7: $\quad$ **if** $\mu(w) = \epsilon$ **then**
8: $\quad\quad$ **for each** $d \in Senses_{\text{Wiki}}$ s.t. d redirects to w — per ogni redirezione a *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** $w \in Senses_{\text{Wiki}}$
12: $\quad$ **if** $\mu(w) = \epsilon$ **then** — la *w* non ancora linkata
13: $\quad\quad$ **if** no tie occurs **then**
14: $\quad\quad\quad \mu(w) := \underset{s \in Senses_{\text{WN}}(w)}{\text{argmax}} \; p(s|w)$ — most likely sense to *w* based on the maximization of the conditional probabilities *p(s|w)* over the senses *s*
15: **return** μ

L’algoritmo di mapping


L'algoritmo di mapping
L'input dell'algoritmo consiste dei due insiemi di concetti $Senses_{Wiki}$ e $Senses_{WN}$
L'output è invece la funzione di mapping μ.
Il loop a riga [1] non fa altro che inizializzare μ collegando ad ogni senso di Wikipedia il concetto vuoto.
Il loop di riga [3] si occupa invece dei lemmi monosemici, ovvero quei lemmi per i quali c'è un solo significato.
Se il lemma w risulta essere monosemico sia in Wikipedia (*vedi nota ramo giallo*) che in WordNet (w appare in un solo synset e quindi ha un solo indice w_{ξ}^{1} dove ξ è uno dei possibili PoS trattati da WordNet) [4] allora viene direttamente effettuato il mapping [5].
Nota:
Ricorda che il lemma w di Wikipedia è ottenuto dal titolo senza label della Wikipage
Nota:
Gran parte delle named entity vengono trattate in questo loop [3], infatti, se w è una named entity il suo mapping per in WordNet, laddove esista, sarà tendenzialmente unico (e se non lo è verrà trattato in seguito).

L'algoritmo di mapping

A riga [6] scorriamo tutte le Wikipages `w` (i.e. (*in questo modo*) i loro titoli senza label) e se ancora non abbiamo trovato un mapping per w [7] (quindi se non siamo entrati nell'if a riga [4]) prendiamo tutte le Wikipage `d` che linkano (*hanno dei redirect*) a w [8]

Se d ha un mapping (*i.e.* un sense-id in WordNet) e quel sense-id si trova nello stesso synset di uno dei sense-id assegnati a w (ossia un certo w_{ξ}^{i} con ξ un PoS), allora il mapping che assegniamo a w è il significato che ha w nel synset di μ(d) (*i.e.* $\mu(w) = w_x{}^i{}_i$).

In questo modo riusciamo a sfruttare gli internal link dei vari documenti d che fanno riferimento a w.

Il loop a riga [11] va poi a popolare tutti i mapping delle Wikipages ancora vuoti [12]: il mapping scelto è il senso s (fra i sensi possibili di w) che massimizza la probabilità $P(s|w)$, ovvero la probabilità di avere il senso $s \in Senses_{WN}(w)$ data la parola w [14].

P(s|w)

L'algoritmo che abbiamo descritto è abbastanza semplice, ma come calcoliamo P(s|w)?

Il senso s che massimizza la probabilità P (s |w) è calcolato mediante la formula:

$$\mu(w) = \underset{s \in Senses_{WN}(w)}{\operatorname{argmax}} P(s|w)$$

Usando la definizione di probabilità condizionata possiamo affermare che:

$$\mu(w) = \underset{s \in Senses_{WN}(w)}{\operatorname{argmax}} P(s|w) = \underset{s \in Senses_{WN}(w)}{\operatorname{argmax}} \frac{P(s \cap w)}{P(w)}$$

[3] $P(A|B) = \frac{P(A \cap B)}{P(B)}$, dove $P(A \cap B)$ è la probabililtà congiunta dell'evento A e dell'evento B, ovvero la probabilità che accadano entrambi gli eventi.

Probabilità condizionata

Giacché stiamo facendo un argmax, possiamo rimuovere P(w) dato che è un fattore costante e quindi possiamo dire che:

$$\mu(w) = \underset{s \in Senses_{WN}(w)}{\operatorname{argmax}} P(s \cap w)$$

In sostanza, il senso s più appropriato fra i sensi della parola w in WordNet è quello che massimizza la probabilità congiunta `P(s ∩ w)` del senso s e della pagina w.

Disambiguazione

Al fine di stimare la probabilità congiunta di un senso di WordNet e di una Wikipage utilizziamo la stessa tecnica già studiata per il task di WSD: andiamo a definire un contesto per ognuno dei due concetti.

Sostanzialmente, dato un concetto (*i.e.* una pagina o un senso), il contesto di disambiguazione di quel concetto è l'insieme delle parole ottenute dalla fonte corrispondente i cui sensi sono associati con il concetto in input mediante una qualche relazione semantica (che sottende dunque la possibilità che ognuna delle parole nel set dei concetti estratti sia un potenziale link per il mapping μ)

Disambiguazione Wikipage

Contesto di disambiguazione di una Wikipage

Data una Wikipage w, le seguenti informazioni sono usate come contesto di disambiguazione:

Label del titolo:
es. per la pagina Play (theatre)) la parola theatre è aggiunta al contesto di disambiguazione.

Internal link:
vengono presi i titoli delle pagine a cui w linka.
(i.e. *outgoing link, i link ad altre pagine contenuti all'interno di w*)

Redirections: vengono presi i titoli delle pagine che puntano a w. (i.e. *incoming link*)

Categories di w.
Una wikipage è tipicamente classificata in una o più categorie.

Es. La wikipage *Play(theatre*) è categorizzata come *PLAYS, DRAMA, THEATRE, etc.*

Il contesto di disambiguazione Ctx(w), di una Wikipage w, è dunque definito come l'insieme delle parole recuperate da tutte e quattro le fonti che abbiamo appena citato.

Es

$$Ctx(Play\ (theatre)) = \{theatre,\ literature,\ \ldots,\ playlet,\ \ldots,\ character\}$$

Disambiguazione WordNet

Contesto di disambiguazione di un concetto WordNet

Dato un concetto **s** di WordNet contenuto nel synset **S**, i seguenti elementi fungono da fonti per costruire il contesto di disambiguazione:

Sinonimi:
tutti i sinonimi di s (quindi tutte le parole in S).

Iperonimi/Iponimi:
tutti i sinonimi nei synsets H tali che H è un iperonimo (i.e., una generalizzazione) di S oppure un iponimo (i.e., una specializzazione) di S.

Gloss:
tutte le content word della glossa di s, e.g., dato s = play$^{1}_{n}$, definito come *"a dramatic work intended for performance by actors on a stage"*, aggiungiamo al contesto di disambiguazione di s i termini: work, dramatic work, intend, performance, actor, stage.

Proprio come per le Wikipage, l'unione delle parole ottenute dalle tre fonti che abbiamo appena citato costituiscono il contesto di disambiguazione di un significato.

Es

$$Ctx(play^{1}_{n}) = \{drama,\ dramatic\ play,\ composition,\ work,\ \ldots,\ actor,\ stage\}$$

Ora che sappiamo come ottenere i contesti di disambiguazione per w e per s possiamo calcolare la probabilità di congiunta `P(s ∩ w)` mediante la seguente stima:

$$P(s \cap w) \approx \frac{\text{score}(s, w)}{\sum_{\substack{s' \in Senses_{WN}(w) \\ w' \in Senses_{Wiki}(w)}} \text{score}(s', w')}$$

P(s ∩ w)

Andiamo cioè a contare il punteggio della coppia (s, w) rispetto alla somma di tutti i punteggi possibili ottenuti dalle combinazioni di tutti i sensi s' di w in WordNet con tutte le pagine w' aventi lo stesso significato di w, ovvero pagine che condividono lo stesso titolo di w in Wikipedia (ricordiamo che non viene considerato il label per costruire $Senses_{Wiki}$).

Calcolare la funzione di scoring

Stiamo ancora cercando di calcolare P(s|w) e abbiamo visto che l'uso di contesti di disambiguazione può aiutarci per definire la probabilità condizionata in un modo più semplice da calcolare.

Per ora abbiamo assunto la presenza di una funzione score che si basi sui contesti e ci restituisca un adeguato punteggio data una coppia di sensi s e w. In questa sezione vediamo brevemente due diversi modi di implementare la funzione di scoring.

Scoring mediante bag-of-words

Lo score viene computato seguendo la formula:

$$score(s, w) = | Ctx(s) \cap Ctx(w) | + 1$$

Ossia, andiamo a contare quante sono i lemmi comuni nei due contesti. Il valore 1 è aggiunto per ragioni di smoothing (onde evitare di avere probabilità pari a 0).

Questo semplicissimo metodo usa soltanto le parole nei contesti e dunque non sfrutta le informazioni strutturali disponibili in WordNet o Wikipedia.

Scoring graph-based

L'idea di questa funzione di score è quella di partire dal contesto di disambiguazione della Wikipage Ctx(w) e di trasformarlo in un grafo diretto con archi etichettati.

I nodi (*V vertici*) di tale grafo sono costituiti da tutti i sensi di WordNet possibili per il titolo della Wikipage w e da tutti i sensi di ogni parola contenuta in Ctx(w)

In formule $V = Senses_{WN}(w) \cup \left(\cup_{cw \in Ctx(w)} Senses_{WN}(cw) \right)$

Una volta definiti i nodi, questi vengono connessi tra loro basandosi sulle relazioni presenti in WordNet.

Più formalmente, per ogni vertice $v \in V$ si effettua una ricerca in profondità sul grafo di WordNet e ogni volta che troviamo un $v' \in V$ (ossia un nodo già presente nel nostro grafo di disambiguazione) che sia diverso da v su un path di lunghezza minore di un certo L impostato a priori, aggiungiamo tutti i nodi e gli archi del path.

Formalmente

$v, v_1, v_2, \ldots, v_k, v'$

Pertanto, avendo un path del tipo

$$V = V \cup \{v, v_1, v_2, \ldots, v_k, v'\}$$
$$E = E \cup \{(v, v_1), (v_1, v_2), \ldots (v_k, v')\}$$

Il nostro insieme di nodi V e il nostro insieme di archi E diventeranno

Scoring graph-based

Al termine della costruzione avremo quindi un grafo di disambiguazione che è una porzione di WordNet, ma che contiene necessariamente tutte le parole del contesto di disambiguazione di w e tutti gli archi fra quelle parole che siamo riusciti ad ottenere da WordNet.

Non resta che calcolare lo score come segue:

$$\text{score}(s, w) = \sum_{cw \in Ctx(w)} \sum_{s' \in Senses_{WN}(cw)} \sum_{p \in paths_{WN}(s,s')} e^{-(length(p)-1)}$$

dove:

$paths_{WN}(s,s')$ è l'insieme di tutti i percorsi fra s ed s' in WordNet (o meglio, nel grafo di disambiguazione che ne è un sottoinsieme)

`length(p)` è la lunghezza del path p in termini del numero di archi che contiene.

Come possiamo notare, questa metodologia per il calcolo dello score è assai più complessa dell'approccio bag-of-words, ma ci permette di sfruttare appieno le relazioni semantiche contenute in WordNet.

Ricapitolando *scoring*

Ricapitolando, possiamo utilizzare uno dei due metodi che abbiamo descritto per calcolare la funzione di score (che necessita dei contesti di disambiguazione), grazie alla quale stimiamo la probabilità congiunta di una Wikipage w ed un senso di WordNet s.

Questa operazione viene ripetuta per ogni senso s' associabile a w in WordNet e quello che massimizza tale probabilità (e quindi che ottiene il miglior score) viene scelto come mapping definitivo per w.