Creazione di pseudoword per ridurre l'ambiguità semantica

Reineri Patric, Scicolone Alessandro

Abstract

In questo progetto viene mostrata la creazione di pseudoword tra due lingue per la costruzione automatica di un dizionario multilingue. Per gli esempi e i dati sono stati scelti l'italiano e lo spagnolo, attuando due diverse startegie di creazione delle pseudowords: la prima itera le parole del dizionario italiano, mentre la seconda strategia itera su i synset. La scelta della pseudoword viene fatta in base al calcolo della riduzione dell'ambiguità e sfruttando i synset in WordNet.

1 Introduzione

L'obiettivo del presente lavoro è la costruzione automatica di un dizionario multilingue tra una lingua sorgente L1 e una o più lingue di estensione L_i, minimizzando l'ambiguità semantica dei termini generati. Il sistema sfrutta la variazione cross-lingua nella codifica dell'ambiguità per generare pseudowords, ossia etichette semantiche artificiali non ambigue. Formalizzando meglio, sia:

- S_a significati nella lingua A,
- S_b significati nella lingua B,
- $S_{ab} = S_a \cap S_b$,

Allora:

$$|S_{a-b}| \leq min(|S_a|, |S_b|) \tag{1}$$

Nella Figura 1 vengono illustrati alcuni esempi.

L1	L2	Pseudoword	Sensi in L1	Sensi in L2	Sensi in Pseudoword
Italiano	Francese	`tempo-temps`	durata, meteo, musica	durata, meteo	durata, meteo
Inglese	Spagnolo	`right-derecho`	destra, retto, diritto, corretto	diritto, retto, direttamente	diritto, retto
Tedesco	Inglese	`Schloss-castle`	castello, serratura	castello	castello
Inglese	Italiano	`bank-riva`	banca, riva	riva	riva
Italiano	Spagnolo	`libro-libro`	opera scritta	opera scritta, registro	opera scritta

Figure 1: *Esempi incompleti di riduzione dell'ambiguitÀ semantica tramite pseudoword L1–L2*

Questo ci permette di ottenere delle pseudowords con ambiguità notevolmente ridotta utile per:

- Ridurre la polisemia nei dizionari monolingua.
- Migliorare la qualità di task downstream come Word Sense Disambiguation e Machine Translation.
- Permette la costruzione di dizionari semantici più controllati e stabili.

Per ogni coppia di lingue, è possibile costruire una base di pseudoword x-y e misurarne il grado di disambiguazione tramite uno score di riduzione dell'ambiguità, ad esempio:

$$\text{AmbiguityReduction}(x, y) = \frac{|S_x| + |S_y| - 2 \cdot |S_{x-y}|}{|S_x| + |S_y|} \tag{2}$$

Dove $|S_x|$ è il numero di sensi associati al termine x, $|S_y|$ è il numero di sensi associati al termine y, mentre $|S_{x-y}|$ è il numero di sensi della pseudoword x-y. Un valore prossimo a 1 indica una forte riduzione dell'ambiguità semantica rispetto ai termini originali.

Nelle prossime sezioni verranno presentate due metodologie per la creazione delle pseudoword. Entrambe fanno uso di OpenWordNet Multilingua (OWN-M) [Bond and colleghi 2025], una risorsa linguistica e lessicale open-source che estende il dizionario elettronico WordNet (limitato all'inglese) a più lingue, collegando ciascun synset ai corrispondenti lemmi multilingue.

2 Metodologie

Al fine di ottenere gli obiettivi citati nella sezione 1 vengono proposte due metodologie diverse per la creazione delle pseudowords: una che parte dal termine e una che parte dal significato.

2.1 Creazione delle pseudowords partendo dal dizionario

Il metodo consiste nel iterare il dizionario della lingua L1, in questo caso l'italiano, selezionando ad ogni iterazione una parola e ricercando le parole che condividono lo stesso significato nella lingua L2 utilizzando i synset multilingua presenti in OpenWord-Net Multilingua.

La creazione della pseudoword avviene scegliendo tra le parole correlate quella che riduce maggiormente l'ambiguità con la parola della lingua L1 considerata nell'iterazione corrente. Per calcolare la quantità di riduzione dell'ambiguità è stata utilizzata la formula in 2.

Algorithm 1 Pseudowords generation

Require: Language L1 and L2, Dictionary D in L1
Ensure: pseudoword list P

```
shuffle D randomly
P ← ∅
for word w ∈ D do
    relatedwords ← related words of w from L2
    pseudo ← init pseudo-word from w
    for c ∈ relatedwords do
        pseudo ← max_ambiguity(pseudo, c)
    P ← P ∪ {pseudo}
return P
```

Un vantaggio dell'algoritmo è che va a controllare ogni parola correlata "passando" da ogni synset della parola L1, confrontando tutte le coppie possibili, ottenendo così la parola in L2 migliore. Inoltre garantisce che ongi parola del dizionario in L1 venga considerata.

Un svantaggio è che la completezza nella generazione delle pseudoword dipende da WordNet, non è infatti strano nel caso visionato (utilizzando l'italiano e lo spagnolo), che ad un synset in una lingua non corrisponda alcun synset nelle seconda lingua (a causa di una mancanza di dati e non una vera mancanza del significato nella lingua stessa).

2.2 Creazione delle pseudowords partendo dai synset in WordNet

Il secondo metodo scorre l'intera lista dei synset di WordNet, selezionandone uno a ogni iterazione. Da ciascun synset è quindi

possibile ricavare l'insieme dei lemmi della lingua L1 e quello dei lemmi corrispondenti nella lingua L2. Successivamente per ogni coppia possibile (w_1, w_2) dove $w_1 \in L1$ e $w_1 \in L2$ si controlla che la pseudoword w_1-w_2 non sia già presente nell'insieme delle pseudoword restituito dalla funzione. Nel caso di esito positivo, la coppia viene scartata e si passa all'iterazione successiva, altrimenti si procede calcolando il valore di riduzione dell'ambiguità utilizzando la formula in 2. Se il valore della pseudoword è maggiore del valore presente in *maxAmbiguityScore* allora la pseudoword corrente viene considerata come nuovo massimo. Infine una volta considerate tutte le coppie di termini, la pseudoword che riduce maggiormente l'ambiguità viene aggiunta all'insieme delle pseudoword da restituire.

Algorithm 2 Pseudoword generation from Synsets

Require: Number of Pseudowords n, Language L1 and L2
Ensure: List of pseudowords P

$P \leftarrow \emptyset$
$WordNetSynsets \leftarrow$AllWordNetSynset()
for $Synset \in WordNetSynsets$ **do**
    $LemmasL1 \leftarrow$ all lemmas from S
    $LemmasL2 \leftarrow$ all lemmas from S
    if $LemmasL1 = \emptyset$ OR $LemmasL2 = \emptyset$ **then**
        skip
    $maxAmbiguityScore \leftarrow -1$
    $maxPseudoword \leftarrow null$
    for $w_1 \in LemmasL1, w_2 \in LemmasL2$ **do**
        if w_1-$w_2 \notin$ P **then**
            $score \leftarrow$ AmbiguityReduction(w_1, w_2)
            if $score > maxAmbiguityScore$ **then**
                $maxAmbiguityScore \leftarrow score$
                $maxPseudoword \leftarrow$ PseudoWord(w_1, w_2)
    $P \leftarrow P \cup \{maxPseudoword\}$
    if $|P| = n$ **then**
        stop
return P

Questa metodologia presenta il vantaggio per cui una volta ottenuti gli insiemi $LemmasL1$ e $LemmaL2$ associati ai synset vi è la garanzia che questi condividano almeno 1 significato evitando il caso in cui la formula della ambiguity reduction restituisce 1 , ovvero quando i due termini sono completamente disgiunti dal punto di vista semantico (non condividono alcun synset). Inoltre con questo metodo viene esplorato tutto lo spazio semantico di WordNet, considerando ogni synset una sola volta.

Tra gli svantaggi vi è il poco controllo lessicale su quali termini vengono utlizzati per generare le pseudoword. Inoltre il numero di pseudowords dipende dal numero di synsets in WordNet. Infine la granularità con cui vengono generate le pseudoword dipende dal grado di completezza di WordNet rispetto alle due lingue considerate. Infatti, a uno stesso synset potrebbero essere associati ulteriori lemmi che, tuttavia, non sono presenti in WordNet.

3 Risultati

In questa sezione le due metodologie verranno analizzate e messe a confronto sulla base di due aspetti principali:

- L'efficienza computazionale dei metodi considerati;
- La qualità delle pseudoword generate, valutata in termini di riduzione dell'ambiguità.

Per ottenere questo risultati sono state scelte le lingue italiano e spagnolo, poichè hanno circa la stessa quantità di lemmi associati in WordNet.

Per quanto riguarda il primo aspetto sono stati effettuate due sperimentazioni illustrate in figura 3 dove si confrontando i due metodi monitorando tempo di esecuzione e numero di coppie di termini confrontate, all'aumentare del numero di pseudoword create (100, 500, 1000, 3000, 5000, 7000, 10000).

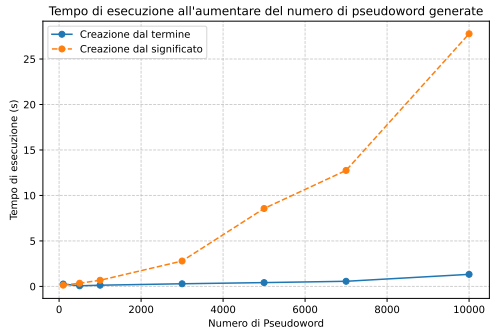

Figure 2: *Confronto dei due metodi sul tempo di esecuzione.*

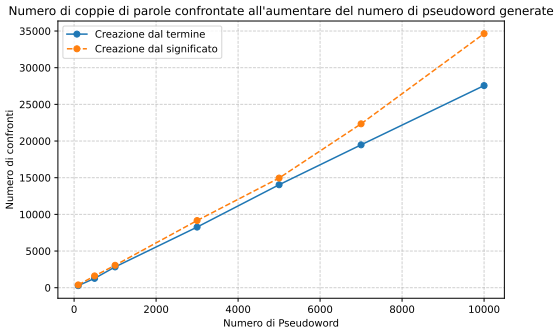

Figure 3: *Confronto dei due metodi sul numero di confronti di termini per calcolo dell'ambiguity reduction.*

I risultati ottenuti evidenziano una notevole differenza nel tempo di esecuzione del metodo 2 che risulta molto più alta. La spiegazione di questo comportamento è data dal numero di confronti effettuati, che risulta molto più alto per il secondo metodo.

Per confrontare i due metodi in termini di riduzione dell'ambiguità delle pseudoword create, viene effettuato un esperimento in cui ogni metodo genera 10.000 pseudoword e per ciascuna pseudoword create si tiene traccia dello score di riduzione. Questi score vengono salvati in un file xls e utilizzati per creare due gaussiane calcolando media e deviazione standard calcolati a partire dagli score ottenuti:

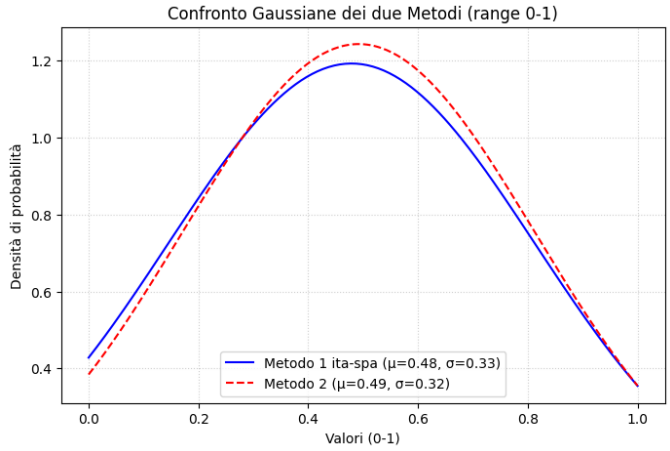

Figure 4: *Confronto tra metodo 1 e metodo 2 su riduzione dell'ambiguità*

Confrontando le due gaussiane si osserva che:

1. La curva rossa risulta più a destra (verso i valori alti di ambigity reduction) rispetto alla curva blu. Il metodo 2 ha una media leggermente più alta (0.49) del metodo 1 (0.48). Questo significa che il Metodo 2 tende a produrre valori mediamente più alti di riduzione dell'ambiguità rispetto al Metodo 1.

2. Il metodo 2 ha una deviazione standard leggermente inferiore ripetto a quella del metodo 1. Questo significa che nel caso del metodo 1 i valori sono maggiormente concentrati intorno alla media.

4 Conclusione

Nel corso delle varie sezioni sono state mostrate le due metodologie per la creazione delle pseudowrod: creazione partendo dai singoli termini del dizionario e creazione a partire dalla lista di synset in WordNet.

Dai risultati ottenuti il secondo metodo consente di raggiungere in media valori poco più elevati rispetto al primo metodo, ma a fronte di un significativo aumento dei tempi di esecuzione.

References

BOND, F., AND COLLEGHI, 2025. Open multilingual wordnet (omw). `https://omwn.org/`. Accessed: 2025-09-07.