# Reduction of semantic ambiguity through multilingual pseudowords

Andrea Camoia[1]

University of Turin, 10149 Torino TO
andrea.camoia@edu.unito.it

## 1 Introduzione

### 1.1 Motivazione ed Obiettivi

L'ambiguità semantica rappresenta una delle principali sfide nel contesto del Natural Language Processing, con implicazioni che toccano sia la comprensione che la generazione di testo. Tradizionalmente, questo problema viene affrontato attraverso approcci "a posteriori", che cercano di identificare il significato corretto tra tutti quelli che una parola polisemica può assumere, basandosi sul contesto circostante e su altre caratteristiche linguistiche.

È importante sottolineare che l'ambiguità semantica non rappresenta una proprietà universale delle parole, ma dipende specificamente dalla lingua di riferimento. Ogni sistema linguistico codifica e organizza i significati in modo peculiare: un concetto che risulta altamente ambiguo in una lingua può essere espresso attraverso forme lessicali distinte e non ambigue in un'altra lingua.

L'intuizione alla base di questo lavoro risiede proprio nella volontà di sfruttare questa variazione cross-linguistica del significato per ottenere una disambiguazione "a priori" dei termini, riducendo alla fonte il numero di significati che un termine può assumere e di conseguenza fornendo ai sistemi di linguistica computazionale "termini" con ridotta ambiguità semantica.

### 1.2 Approccio Proposto

L'approccio proposto consente di costruire automaticamente un dizionario multilingue basato su *pseudo-word*, cioè etichette semantiche artificiali ottenute concatenando due termini che esprimono lo stesso concetto in lingue diverse. Sia dunque $x \in L_1$ una parola della lingua $L_1$, con un insieme di significati $S_x$, e sia $y \in L_2$ la sua controparte concettuale nella lingua $L_2$, con insieme di significati $S_y$. La *pseudo-word* risultante, indicata con $x$-$y$, eredita soltanto la parte di significato condivisa da $x$ e $y$, ossia $S_{x-y}$. Per costruzione vale la disuguaglianza:

$$|S_{x-y}| \leq \min\big(|S_x|, |S_y|\big),$$

con un'alta probabilità che in realtà

$$|S_{x-y}| \ll |S_x| \quad \text{e} \quad |S_{x-y}| \ll |S_y|.$$

In questo modo si ottiene una riduzione significativa dell'ambiguità semantica. Inoltre per massimizzare tale riduzione, la coppia di lingue $(L_1, L_2)$ viene scelta tra un insieme di quattro possibili lingue,

$$L = \{\text{Inglese}, \text{Italiano}, \text{Francese}, \text{Spagnolo}\}, \quad L_1 \neq L_2.$$

### 1.3 Risorse Utilizzate

Per estrapolare tutte le informazioni semantiche e lessicali necessarie alla realizzazione del progetto è stato utilizzato BabelNet[1], un dizionario elettronico multilingue che integra informazioni provenienti da diverse fonti tra cui WordNet (per varie lingue), Wikipedia, Wiktionary e altre. La scelta di BabelNet è motivata da diverse caratteristiche fondamentali che la rendono ideale per la realizzazione del progetto.

Innanzitutto la struttura di BabelNet gestisce efficientemente termini e rispettivi significati nelle oltre 200 lingue a disposizione: dato un termine è possibile ottenere i tutti i possibli synset (significati) nelle diverse lingue, e poi gestire efficientemente la loro intersezione tramite gli identificatori univoci (synset ID) associati ad ogni concetto. Ad esempio:

- synsets("Banca") = [bn:00008364n,bn:00008370n, bn:02304182n,...]
- synsets("Bank") = [bn:00008363n,bn:00008364n, bn:00008365n,...]

Inoltre BabelNet fornisce delle efficienti API che, nonostante alcune limitazioni nell'utilizzo gratuito (illimitate nel caso di ricerche accademiche), consentono l'accesso programmatico sistematico ai dati necessari per l'estrapolazione e la gestione dei termini e dei significati.

Infine, poichè la costruzione del vocabolario di *pseudo-word* parte da una lista di termini in inglese, sono state integrate le API di DeepL[2] come risorsa per ottenere le traduzioni di queste parole nelle diverse lingue target prima dell'interrogazione a BabelNet.

## 2 Progettazione

### 2.1 Strategia

La costruzione del vocabolario di *pseudo-word* inizia da una lista di termini "sorgente", nel nostro caso in inglese, che vengono tradotti nelle rispettive lingue, ottenendo così un insieme $T$ di termini $x_i$ per ciascuna delle lingue $L_i \in L$. A causa delle limitazioni d'uso imposte dalle API di BabelNet (1000 richieste giornaliere), non è stato possibile utilizzare l'intero vocabolario di una lingua come punto di partenza. Pertanto, la lista di termini sorgente è stata creata

[1] https://babelnet.org/
[2] https://www.deepl.com/en/pro-api

combinando un insieme di parole notoriamente ambigue in inglese[3] con un altro insieme contenente parole d'uso comune[4]

Successivamente, tramite le API di BabelNet vengono estratti tutti i possibili significati, synset, associati a ciascun termine, specificando nella richiesta la lingua in cui effettuare la ricerca e la sorgente di informazioni desiderata. Per quanto riguarda quest'ultimo aspetto sono stati utilizzate le seguenti sorgenti:

- Inglese: `WordNet`, `Wikipedia`
- Italiano: `Italian Open Multilingual WordNet`, `Wikipedia`
- Francese: `French (WOLF) Open Multilingual WordNet`, `Wikipedia`
- Spagnolo: `Spanish Open Multilingual WordNet`, `Wikipedia`

A partire dalla lista di synset associati a ciascun termine nelle rispettive lingue, vengono generate tutte le possibili combinazioni $x$-$y$. Per ognuna di esse si calcola il numero di significati condivisi e il conseguente grado di riduzione dell'ambiguità. Successivamente, il vocabolario finale viene costruito selezionando, per ciascun termine di base, la *pseudo-word* che massimizza tale riduzione.

### 2.2 Metriche di Valutazione

Il grado di disambiguazione associato a ciascuna pseudo-word è quantificato attraverso uno score di riduzione dell'ambiguità, calcolato secondo la seguente formula[5]:

$$\texttt{AmbiguityReduction}(x, y) = \frac{|S_x| + |S_y| - 2 \cdot |S_{x-y}|}{|S_x| + |S_y|} \tag{1}$$

Inoltre, proponiamo una metrica alternativa progettata per privilegiare le coppie $x$-$y$ che, partendo da un elevato numero di significati iniziali $|S_x|$ e $|S_y|$ per entrambi i termini, producono un'intersezione $|S_{x-y}|$ contenuta, massimizzando così il calo di polisemia in entrambe le lingue.

$$\texttt{AmbiguityDrop}(x, y) = 2 \cdot \frac{(|S_x| + |S_y| - 2 \cdot |S_{x-y}|) \cdot \sqrt{|S_x| \cdot |S_y|}}{(|S_x| + |S_y|)^2} \tag{2}$$

Al numeratore, il termine $(|S_x| + |S_y| - 2 \cdot |S_{x-y}|)$ quantifica la riduzione assoluta di ambiguità, mentre $\sqrt{|S_x| \cdot |S_y|}$ amplifica l'impatto per coppie di parole caratterizzate da un elevato numero di significati iniziali. La metrica assume valori nel range $[0, 1]$ grazie alla normalizzazione applicata.

Vediamo un esempio delle metriche a confronto, considerando il termine $x =$ `"balance" (en)` con $|S_x| = 45$ ed $y =$ `"saldo" (es)` con $|S_x| = 4$ che producono la *pseudo-word* `"balance-saldo"` con un numero di significati in comune pari a $|S_{x-y}| = 2$ `['bn:00008044n', 'bn:00008048n']`

[3] `https://muse.dillfrog.com/lists/ambiguous`

[4] `https://www.ef.com/wwen/english-resources/english-vocabulary/top-3000-words/`

[5] (quella proposta nella traccia)

- `AmbiguityReduction`$(x, y) = 0.918$
- `AmbiguityDrop`$(x, y) = 0.502$

I risultati mostrano come la nuova metrica permetta di evidenziare casi in cui la riduzione dell'ambiguità è fortemente condizionata da una disparità del numero di sensi originali, privilegiando al contrario casi in cui si verifica una forte riduzione in entrambe le lingue.

## 3 Implementazione

Per quanto riguarda l'implementazione pratica dell'algoritmo di costruzione del nuovo dizionario, avvenuta in python, si è seguito lo schema descritto nei paragrafi precedenti. Dopo aver inizializzato la lista dei termini da disambiguare, per ciascun `lemma` l'algoritmo individua e misura tutte le possiblili *pseudo-word* tramite il metodo `[get_all_pseudoword(lemma)]`. Successivamente, viene selezionata la *pseudo-word* che massimizza la misura scelta (`AmbiguityReduction` oppure `AmbiguityDrop`), utilizzando il metodo `get_best_pseudoword(lemma)`.

Il vocabolario così generato viene infine salvato in un file `csv`, che per ciascun elemento conserva le seguenti informazioni:

- `pseudoword`: la *pseudo-word* generata nella forma $x$-$y$
- `ambiguity_reduction`: valore di Ambiguity Reduction; (Formula 1)
- `ambiguity_reduction`: valore di Ambiguity Drop; (Formula 2)
- `L1`: lingua del termine $x$;
- `L2`: lingua del termine $y$;
- `|synsets_L1|`: numero di significati del termine $x$;
- `|synsets_L2|`: numero di significati del termine $y$;
- `|synsets_L1_L2|`: numero di significati in comune;
- `|synsets_intersection|`: lista dei SyntetsID in comune;

Viene riportato di seguito un estratto del vocabolario generato:

| pseudoword | ambiguity_reduction | ambiguity_drop | L1 | L2 | \|synsets_L1\| | \|synsets_L2\| | \|synsets_L1_L2\| | synsets_intersection |
|---|---|---|---|---|---|---|---|---|
| poverty-pobreza | 0,895 | 0,930 | EN | ES | 10 | 9 | 1 | ['bn:00046157n'] |
| clothes-vestiti | 0,667 | 0,497 | EN | IT | 5 | 1 | 1 | ['bn:00006125n'] |
| medicine-medicina | 0,842 | 0,686 | EN | ES | 30 | 8 | 3 | ['bn:00054133n', 'bn:00054126n', 'bn:00054128n'] |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

**Fig. 1.** Estratto del Dizionario di pseudoword risultante in formato CSV

## 4 Risultati

Il lavoro ha prodotto un vocabolario finale di *pseudo-word* contenente 1293 elementi, di cui 653 provenienti dalla lista di parole ambigue, e 640 parole comunemente usate in conversazioni, articoli e magazine inglesi. Il vocabolario è stato serializzato in un file `CSV` descritto precedentemente
`[final_pseudoword_dictionary.csv]`

A partire dal vocabolario generale, a fini di analisi, sono stati poi estratti due sotto-insiemi da 100 *pseudo-words* che massimizzano rispettivamente la misura di *AmbiguityReduction* e quella di *AmbiguityDrop*.

**Table 1.** Statistiche estratte dalle top 100 pseudowords che massimizzano il valore di Ambiguity Reduction

| | ambiguity _reduction | ambiguity _drop | \|synsets _L1\| | \|synsets _L2\| | \|synsets _L1_L2\| |
|---|---|---|---|---|---|
| **count** | 100,000 | 100,000 | 100,000 | 100,000 | 100,000 |
| mean | 0,959 | 0,684 | 43,850 | 10,740 | 1,080 |
| std | 0,008 | 0,200 | 20,813 | 8,260 | 0,307 |
| min | 0,946 | 0,274 | 8,000 | 1,000 | 1,000 |
| 25% | 0,951 | 0,561 | 31,000 | 5,000 | 1,000 |
| 50% | 0,957 | 0,729 | 40,000 | 9,000 | 1,000 |
| 75% | 0,965 | 0,848 | 52,000 | 13,250 | 1,000 |
| max | 0,983 | 0,954 | 115,000 | 38,000 | 3,000 |

**Table 2.** Statistiche estratte dalle top 100 pseudowords che massimizzano il valore di Ambiguity Drop

| | ambiguity _reduction | ambiguity _drop | \|synsets _L1\| | \|synsets _L2\| | \|synsets _L1_L2\| |
|---|---|---|---|---|---|
| count | 100,000 | 100,000 | 100,000 | 100,000 | 100,000 |
| mean | 0,924 | 0,903 | 20,586 | 16,293 | 1,333 |
| std | 0,025 | 0,023 | 9,264 | 7,047 | 0,606 |
| min | 0,875 | 0,868 | 7,000 | 7,000 | 1,000 |
| 25% | 0,902 | 0,883 | 14,000 | 11,000 | 1,000 |
| 50% | 0,926 | 0,901 | 19,000 | 15,000 | 1,000 |
| 75% | 0,944 | 0,923 | 24,000 | 18,500 | 2,000 |
| max | 0,966 | 0,954 | 57,000 | 38,000 | 4,000 |

Dalle statistiche emerge chiaramente il fenomeno già discusso nella presentazione delle metriche di valutazione. Come riportato nella Tabella 2, valori elevati di `Ambiguity Drop` sono naturalmente associati a livelli altrettanto alti di `Ambiguity Reduction`. Andamento confermato anche dal grafico in figura 2.

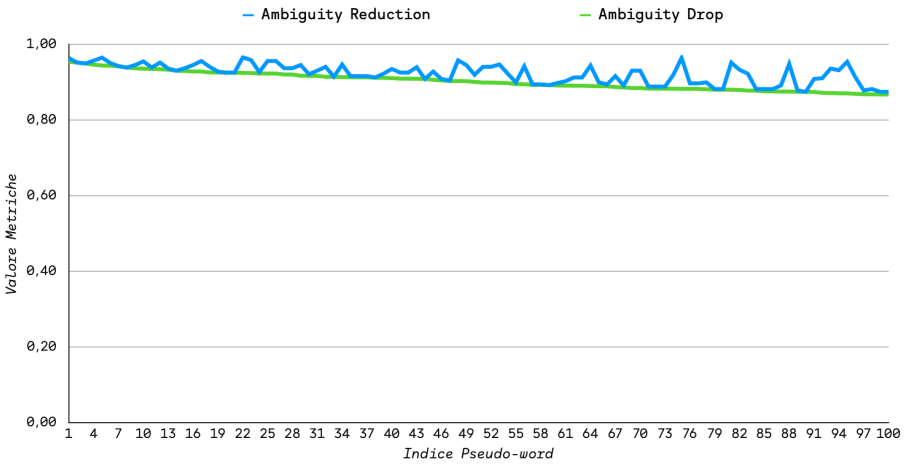


**Fig. 2.** Andamento della metrica di Ambiguity Reduction confrontato a quella di Ambiguity Drop per le top 100 *pseudo-word* che massimizzano l'Ambiguity Drop

Al contrario, i risultati mostrati nella Tabella 1 evidenziano come le *pseudo-word* che massimizzano la metrica di `Ambiguity Reduction` presentano valori decisamente più bassi di `Ambiguity Drop`. Questo comportamento, mostrato nel grafico in figura 3, è dovuto alla presenza di *pseudo-word* generate a partire da termini con un numero di significati fortemente sbilanciato tra le lingue, producendo quindi una riduzione della polisemia limitata a una sola di esse.

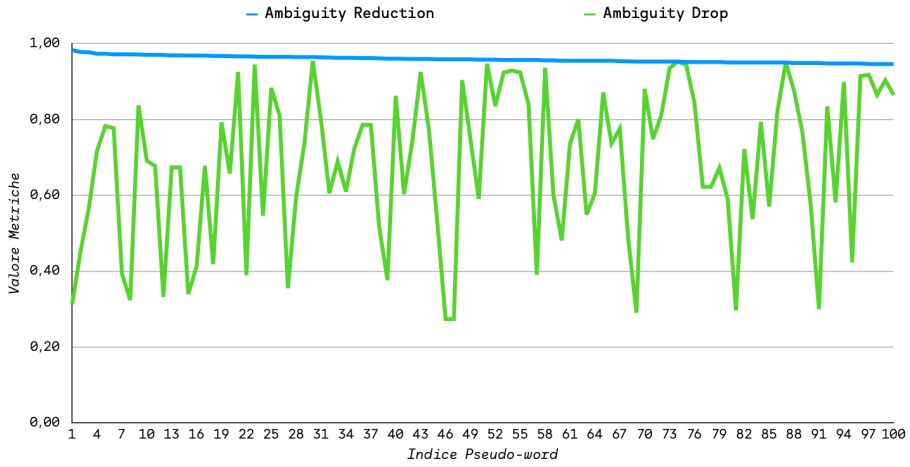


**Fig. 3.** Andamento della metrica di Ambiguity Reduction confrontato a quella di Ambiguity Drop per le top 100 *pseudo-word* che massimizzano l'Ambiguity Reduction

# 5 Conclusioni

## 5.1 Limitazioni e Problematiche

Nonostante i risultati complessivamente soddisfacenti del progetto, bisogna evidenziare alcune limitazioni che hanno influenzato la portata e la precisione dell'analisi condotta:

La prima problematica riguarda i vincoli delle API di BabelNet, che hanno limitato la copertura lessicale e le dimensioni del dizionario di pseudo-word costruito. Tuttavia, le API offrono accesso illimitato per progetti di ricerca autorizzati, rendendo questa limitazione superabile in contesti di ricerca strutturati.

La seconda criticità riguarda invece i probabili errori ottenuti nelle traduzioni utilizzate per il mapping interlinguistico, che hanno occasionalmente compromesso la qualità delle pseudo-word generate. Questa problematica potrebbe essere risolta adottando sistemi specializzati per la traduzione lessicale termine per termine.

## 5.2 Sviluppi Futuri

I limiti appena presentati non compromettono però il risultato ottenuto, con un approccio efficace che fornisce una base solida per future estensioni della ricerca. Gli sviluppi futuri potrebbero concentrarsi sull'estensione dell'insieme delle lingue utilizzate, dato che combinazioni linguistiche più ampie o tipologicamente diverse potrebbero rivelare variazioni cross-linguistiche più marcate ed una conseguente disambiguazione più efficace.

## 5.3 Data Availability

Il codice e tutti i risultati del progetto sono disponibili nella repository GitHub al seguente link:
`https://github.com/AndCamo/TLN_3_Projects/tree/main/Progetto`