Short Paper: Riduzione dell'ambiguità semantica tramite *pseudoword* multilingue

Pasquale Manfredi

Università degli studi di Torino
pasquale.manfredi@edu.unito.it

Abstract. Questo studio propone un approccio innovativo basato sulla creazione di *pseudoword* multilingue: combinazioni di termini provenienti da lingue diverse che condividono significati comuni, con l'obiettivo di ottenere unità lessicali artificiali più precise. I risultati mostrano una riduzione significativa dell'ambiguità semantica.

Keywords: Ambiguità semantica · Multilinguismo · BabelNet · Pseudoword · NLP

1 Introduzione

L'ambiguità semantica è una sfida significativa nell'elaborazione del linguaggio naturale (NLP), specialmente in contesti multilingue. Questo lavoro corredato di notebook visivo presenta un approccio innovativo per ridurre l'ambiguità semantica attraverso la creazione di *pseudoword* multilingue, combinando termini da lingue diverse per ottenere unità lessicali artificiali con significati più precisi.

2 Idea Centrale

L'idea fondamentale è sfruttare il multilinguismo per ridurre l'ambiguità semantica: si vuole infatti sfruttare la debolezza dell'omonimità di un vocabolo per accorparlo a gli altri in una (o più) lingue diverse.

Così facendo si può ricavare l'intreccio dei synset delle parole per le varie lingue.

Sia:

- x: parola in lingua L_1 con $|S_x| = N$ significati
- y: sua traduzione in lingua L_2 con $|S_y| = M$ significati, dove $M < N$

Definiamo la pseudoword x-y come l'intersezione semantica tra S_x e S_y:

$$S_{x-y} = S_x \cap S_y$$

Spesso vale:

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

3 Implementazione

Il notebook associato utilizza l'API di BabelNet per:

- Recuperare i sensi (synset) di parole in diverse lingue
- Identificare i synset comuni tra le traduzioni
- Calcolare lo score di riduzione dell'ambiguità per ogni pseudoword

La metrica di riduzione è definita come:

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

4 Risultati

Per testare lo studio è stato redatto un jupiter notebook dove sono stati testati 5 termini in inglese-italiano:

"L_1=Inglese"	L_2="Italiano"
time	tempo
right	giusto
castle	castello
bank	riva
book	libro

Table 1. campioni di test

e per ciascuno è stato ricavato mediante BabelNet ogni synset come si vede in Fig.1

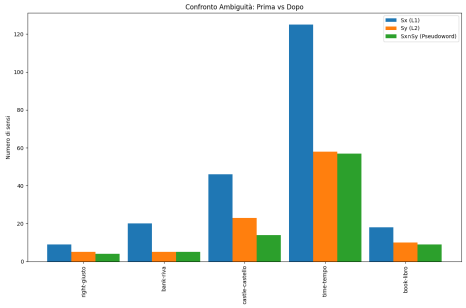

Fig. 1. Visualizzazione dell'intersezione tra synset di lingue diverse.

Index	Pseudoword	$\|S_x\|$	$\|S_y\|$	$\|S_{x-y}\|$	Ambiguity-Reduction
0	right-giusto	9	5	4	0.429
1	bank-riva	20	5	5	0.600
2	castle-castello	46	23	14	0.594
3	time-tempo	125	58	57	0.377
4	book-libro	18	10	9	0.357

Table 2. Risultati ottenuti per 5 pseudoword inglese-italiano

Dopodichè e bastato applicare la metrica precedentemente illustrata per ottenere la riduzione di ambiguità semantica. nella tabella sottostante è mostrato un riepilogo visivo del lavoro svolto. ed infine è stato fornito un grafico a barre orizzontali per una visualizzazione grafica:

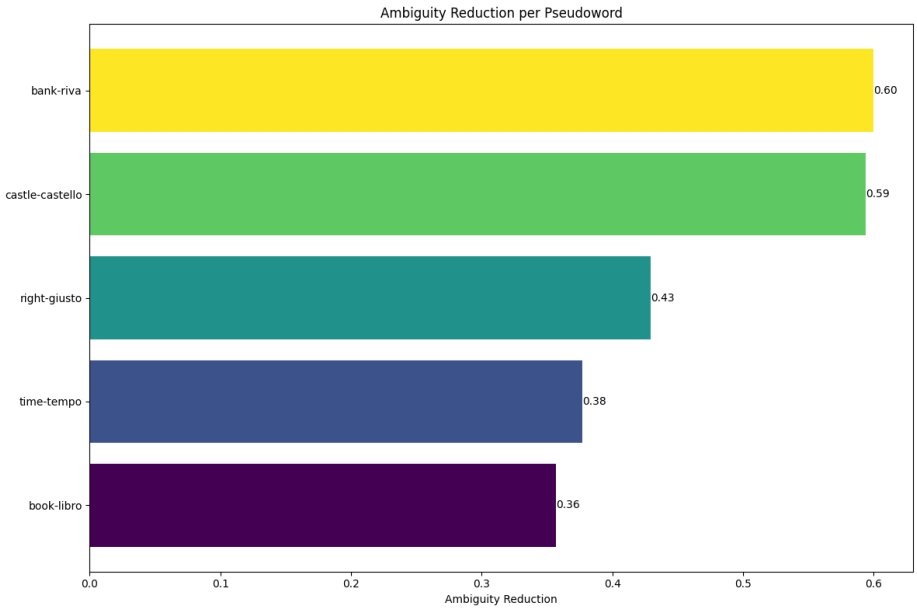

Fig. 2. Grafico a barre orizzontali in cui viene mostrata l' Ambiguity-Reduction

Remark 1. Sebbene il metodo descritto possa essere esteso a più di due lingue contemporaneamente, nella pratica l'intersezione tra i synset condivisi decresce rapidamente con l'aumentare delle lingue coinvolte. In teoria, incrociare più lingue potrebbe ulteriormente ridurre l'ambiguità semantica, ma si è osservato che tale riduzione avviene a discapito della copertura: il numero di pseudoword ottenibili con intersezione non vuota diventa molto esiguo. Per questo motivo, in questa fase sperimentale si è scelto di limitare l'analisi a coppie di lingue (in

particolare Inglese-Italiano), pur sapendo che l'approccio è generalizzabile anche a insiemi linguistici più ampi e a un vocabolario molto più esteso.

5 Vantaggi

L'approccio delle pseudoword multilingue offre numerosi benefici in ambito NLP, tra cui:

- **Riduzione della polisemia**: le pseudoword, derivando dall'intersezione dei significati condivisi tra due (o più) lingue, presentano una semantica più ristretta e precisa rispetto ai termini originali. Questo riduce le ambiguità intrinseche che ostacolano numerosi task NLP.
- **Supporto a task complessi**: il metodo si rivela particolarmente utile in attività come la Word Sense Disambiguation (WSD), la Machine Translation e la costruzione di risorse ontologiche, dove la disambiguazione semantica è cruciale.
- **Costruzione di dizionari più stabili**: le pseudoword possono fungere da unità semantiche più affidabili per la creazione di dizionari multilingue controllati, contribuendo a una base lessicale più coerente e meno ambigua.
- **Generalizzabilità del metodo**: sebbene siano stati mostrati esempi con due lingue, il framework è estensibile a più lingue simultaneamente e a un vocabolario molto più ampio. Tuttavia, come discusso, l'intersezione semantica tende a ridursi con l'aumentare del numero di lingue, portando a un minor numero di pseudoword validi.
- **Potenziale per sviluppi futuri**: un'estensione naturale dell'approccio consiste nel valutare la riduzione dell'ambiguità in contesti più ampi (frasi o documenti), oltre che a livello di singola parola. Altri sviluppi potrebbero includere l'integrazione con modelli linguistici preaddestrati e l'adattamento dinamico dei dizionari in base al dominio.

6 Conclusioni

Questo approccio dimostra come le differenze semantiche tra lingue possano essere sfruttate per ridurre l'ambiguità lessicale. I risultati sperimentali evidenziano il potenziale del metodo in ambienti multilingue, migliorando l'efficacia di varie applicazioni NLP. L'uso delle pseudoword come unità semantiche artificiali offre un'interessante direzione per la semantica computazionale.

Notebook Implementation

Il notebook associato, che implementa l'intero flusso – dall'estrazione semantica tramite l'API di BabelNet all'analisi dell'intersezione dei synset e visualizzazione dei risultati – è disponibile online su GitHub e utilizzabile via Colab:

`Codice colab sul github personale`

Il notebook associato, che implementa l'intero flusso – dall'estrazione semantica tramite l'API di BabelNet all'analisi dell'intersezione dei synset e visualizzazione dei risultati – è disponibile online su GitHub e utilizzabile via Colab.

Per una corretta esecuzione del codice è necessario preparare un file di pseudoword nelle lingue desiderate, salvato come `word_pairs.csv`, e specificarlo nel secret `"WORD_PAIRS"`.

Se si utilizza Colab, le lingue vanno indicate nel secret denominato `"LANGUAGES"`, con valori come `"EN,IT"` per ottenere pseudoword di tipo inglese-italiano.

L'ultimo passaggio è ottenere una chiave BabelNet, registrandosi sul `Sito ufficiale BabelNet`, e utilizzarla nel secret `"BABELNET_API_KEY"`.

Oltre ai risultati osservati nel paper verrà creata una directory contenente per ogni parola i sensi condivisi e i synset associati reperibili nella sotto directory '/srsc'

Problemi Osservati

In pratica, i punteggi di riduzione dell'ambiguità raramente raggiungono il valore massimo di 1. Di seguito si riportano le principali ragioni:

1. Sovrapposizione semantica tra lingue

Le lingue naturali spesso condividono sensi simili, ma non perfettamente identici. Anche se due parole sono traducibili tra loro, i sensi di una possono essere **più specifici o più ampi** rispetto all'altra. Questo comporta che i synset comuni (S_{xy}) siano **più numerosi del minimo atteso**, abbassando il valore della riduzione.

2. Ricchezza (o rumorosità) semantica di BabelNet

BabelNet tende ad **associare molti sensi** anche a parole semplici, poiché aggrega fonti diverse (Wikipedia, WordNet, Wiktionary...). Ciò può causare:

- un aumento di S_x e S_y,
- la presenza di **sensi ridondanti, astratti o irrilevanti**.

Questo fa sì che il denominatore della formula cresca più rapidamente del numeratore, riducendo la metrica.

3. Uso limitato della sorgente `WIKI`

Nel codice viene specificato:

```
source = "WIKI"
```

Questo limita i sensi estratti **solo a quelli derivati da Wikipedia**, con potenziali effetti:

- **Esclusione di sensi linguistici rilevanti** (ad es. presenti in WordNet),
- **Predominanza di definizioni enciclopediche**, spesso troppo generiche.

Il risultato può essere una **sottostima dei sensi specifici** oppure una distorsione del conteggio S_{xy} verso sensi molto generali.

4. Effetti del multilinguismo

Nel passaggio tra lingue, i sensi di una parola possono:

- tradursi in **parole con campi semantici non perfettamente sovrapposti**,
- essere **mappati su synset diversi** a seconda della lingua,
- subire **perdite di granularità o ambiguità aggiuntiva** nella traduzione automatica di BabelNet.

Ciò rende meno probabile un corretto allineamento dei sensi tra le lingue, aumentando i falsi negativi nei synset comuni.

Sebbene l'algoritmo funzioni correttamente, il comportamento della metrica dipende fortemente da tre fattori principali:

- dalla **quantità e qualità dei sensi** restituiti da BabelNet,
- dalla **selezione delle fonti** tramite il parametro `source`,
- dalla **variabilità linguistica** tra le parole nei diversi idiomi.

Per ottenere valori di riduzione dell'ambiguità più elevati o meglio interpretabili, si suggerisce:

- di utilizzare `source="WIKI,WN"` o solo `"WN"` per confrontare il comportamento tra fonti enciclopediche e lessicali,
- di **filtrare synset eccessivamente generici**, come "entity" o "object",
- di valutare l'introduzione di un **peso per i sensi**.