Lista Progetti TLN 2025

Obiettivo

L'obiettivo principale è scrivere uno "short paper" con notebook associato sui seguenti progetti proposti[1].

Soluzioni particolarmente brillanti verranno selezionate per la pubblicazione di un articolo scientifico, con l'aiuto del docente. Questa attività può essere legata eventualmente ad una tesi e/o a borse di ricerca.

Studentesse e studenti, anche in grippo, sono invitat* a scegliere uno dei progetti elencati di seguito.

1. **Creazione di una nuova lingua**: Creazione di un dizionario L1-L2 minimizzando l'ambiguità dei termini. Definizione e calcolo di uno score di disambiguazione.
2. **LLM-resource activation**: Uso di Large Language Models per l'attivazione di specifici slot semantici in risorse lessico-semantiche.
3. **Progetto aperto**, deciso da studentesse e studenti, concordato con il docente durante il corso.

[1] Il template da utilizzare è il seguente: `https://www.overleaf.com/latex/templates/springer-lecture-notes-in-computer-science/kzwwpvhwnvfj`

1 Riduzione dell'ambiguità semantica tramite pseudoword multilingue

Obiettivo

L'obiettivo di questo esercizio 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.

Idea Centrale

Quando una parola $x \in$ L1 ha N significati distinti nella lingua di partenza e la sua traduzione $y \in$ L2 ha M significati (con $M < N$), l'associazione $x \to y$ può essere utilizzata per ridurre l'ambiguità semantica del termine originario. Viceversa, anche y può essere disambiguato sfruttando il legame con x, nel caso in cui la lingua di origine codifichi più finezze semantiche.

Viene così creata una nuova unità lessicale artificiale x-y, detta *pseudoword*, che eredita una porzione condivisa dei significati di x e y, ma con ambiguità inferiore rispetto a ciascun termine singolarmente.

Formalizzazione

Sia:

- $\mathcal{S}_x$: l'insieme dei significati di x in L1
- $\mathcal{S}_y$: l'insieme dei significati di y in L2
- $\mathcal{S}_{x\text{-}y} = \mathcal{S}_x \cap \mathcal{S}_y$: significati condivisi

Allora:

$$|\mathcal{S}_{x\text{-}y}| \leq \min(|\mathcal{S}_x|, |\mathcal{S}_y|)$$

e in particolare si verifica spesso:

$$|\mathcal{S}_{x\text{-}y}| \ll |\mathcal{S}_x| \quad \text{e} \quad |\mathcal{S}_{x\text{-}y}| \ll |\mathcal{S}_y|$$

Esempio

- **L1 (Italiano)**: `banca` $\to \mathcal{S}_{\texttt{banca}} = \{\text{istituto finanziario}, \text{panchina}\}$
- **L2 (Francese)**: `banque` $\to \mathcal{S}_{\texttt{banque}} = \{\text{istituto finanziario}\}$
- **Pseudoword**: `banca-banque` $\to \mathcal{S}_{\texttt{banca-banque}} = \{\text{istituto finanziario}\}$

In questo caso, la pseudoword `banca-banque` è meno ambigua di `banca` (in L1) e anche potenzialmente più precisa di `banque` (in L2), specialmente in contesti multilingue. Altri esempi, costruiti manualmente al solo scopo di illustrare l'idea, potrebbero essere:

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

Table 1: Esempi *incompleti* di riduzione dell'ambiguità semantica tramite pseudoword L1–L2.

Vantaggi Computazionali

La creazione di pseudoword L1–L2:

- Riduce la polisemia implicita nei vocabolari monolingue.
- Migliora la qualità di task downstream come Word Sense Disambiguation e Machine Translation.
- Permette la costruzione di dizionari semantici più controllati e stabili.

Esempio di misura di riduzione di ambiguità

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:

$$\text{AmbiguityReduction}(x, y) = \frac{|\mathcal{S}_x| + |\mathcal{S}_y| - 2 \cdot |\mathcal{S}_{x\text{-}y}|}{|\mathcal{S}_x| + |\mathcal{S}_y|}$$

Un valore prossimo a 1 indica una forte riduzione dell'ambiguità semantica rispetto ai termini originali.

2 LLM-Resource activation: attivazione semantica contestuale di risorse lessico-semantiche

Obiettivo

L'obiettivo di questo esercizio è utilizzare un *Large Language Model* (LLM) per attivare, valorizzare o ponderare dinamicamente specifici slot semantici (proprietà, ruoli, attributi) all'interno di una risorsa lessico-semantica strutturata, sulla base del contesto linguistico d'uso. Il processo permette di trasformare risorse statiche come **FrameNet**, **WordNet** o **VerbNet** in versioni *context-aware*, arricchite e adattive. L'esercizio implica l'uso di diverse teorie e pratiche viste a lezione, quali:

- Ingegneria dei prompt (prompt engineering)
- Analisi semantica strutturata
- Interazione tra NLP simbolico e neurale
- Normalizzazione e rappresentazione semantica in knowledge bases

Metodo

Il task si compone dei seguenti passaggi:

1. **Selezione della risorsa semantica** tra:
   - **FrameNet**: attivazione di frame e ruoli semantici
   - **WordNet**: attivazione di relazioni e proprietà del synset
   - **VerbNet**: attivazione di ruoli sintattico-semantici
2. **Preparazione di contesti linguistici**: ogni contesto include un lemma target da analizzare.
3. **Interrogazione del LLM**: si utilizza un LLM per predire l'entry semantica attivata e i relativi slot, assegnando un peso (es. salienza, attivazione semantica) a ciascun ruolo.
4. **Parsing strutturato dell'output**: i risultati vengono raccolti in formato strutturato (es. JSON) e possono essere utilizzati per analisi successive o integrazione nelle risorse.

Esempi

Esempio 1 – FrameNet (lemma: *evacuare*)

Dopo l'esondazione del fiume, l'intero quartiere fu evacuato.

```
Frame: Removing
Ruoli attivati:
- Theme: "l'intero quartiere" (score: 0.9)
- Cause: "l'esondazione del fiume" (score: 0.8)
- Agent: null (score: 0.1)
```

Esempio 2 – WordNet (lemma: *bank*)

She sat down on the bank and watched the river.

```
Synset target: bank.n.07 (sponda del fiume)
Proprietà e score attivabili (esempi...):
- Domain: geography (ad es. da dataset CSI)
- Related synset-oriented words: river, nature
- Polysemy score: 0.25
- Most similar synsets: <synset>:<score>, ...,
- ...
```