

Relazione esercizio “Mapping di Frame in WN Synsets”

Il secondo esercizio affronta un task di wordsense disambiguation attraverso l'uso WordNET. I termini da disambiguare provengono da alcuni elementi lessicali tratti dalla risorsa di FrameNET che consistono di un insieme di nomi di frame, di frame element e di lexical unit associati agli studenti componenti del gruppo. Tale insieme di elementi viene definito frameSet. I frameSet considerati sono stati ricavati dall'invocazione della funzione “**getFrameSetByStudent**” della libreria “student” passando per parametro il cognome dello studente. Ad ogni studente viene associato un frameSet formato da 5 frame differenti.

L'invocazione della funzione per lo studente “colombino” ha prodotto i seguenti frame: **Confronting_problem, Hostile_encounter, Choosing, Entering_of_plea, Piracy**. Alcuni dei precedenti frame non trovavano riscontro in alcun synset di wordnet, per tale motivo sono state applicate diverse combinazioni tra nome e cognome (Michele e Colombino) e ne sono stati selezionati altri. Il frameSet restituito per lo studente “Akhlaghinia” non ha riscontrato anomalie.

- **Colombino:** [*Piracy, Emotions, Departing, Usefulness, Degree*]
- **Akhlaghinia:** [*Existence, Mental_activity, Education_teaching, Immobilization, Response_scenario*]

Alcuni dei termini restituiti si presentano sotto forma di multiword expression o di più parole separate da spazi bianchi. Non tutti i termini composti sono stati considerati nella successiva fase di annotazione, infatti analizzando la composizione dei precedenti frame su FrameNET si nota che ci sono dei frame element o delle lexical unit mancanti rispetto a quelli trattati in questo esercizio. La scelta è stata fatta sulla base di 3 fattori:

1. Esistenza di un synset su WordNET;
2. Quantità di informazione nei contesti di disambiguazione;
3. Incoerenza con il significato del frame

In assenza di un synset associato al termine, al suo reggente nel caso di multiword expression o al suo lemma, sarebbe stato impossibile procedere nel task di disambiguazione, similmente avendo a disposizione un contesto formato di poche parole di contenuto oppure avendo synset il cui significato si discostava di molto con il significato del termine dato il suo frame. In tutti questi casi si è deciso di scartare tali termini.

La seconda parte dell'esercizio ha visto l'annotazione dei termini selezionati, come accennato in precedenza. In questa fase è stato creato un file JSON, (“**annotazioni.json**”) riportando in formato aggregato per cognome dello studente e per nome del frame tutti i termini con i loro synset; informazioni che verranno usate nella fase di disambiguazione. L'associazione termine – synset è stata fatta considerando le loro definizioni recuperate dalle diverse risorse lessicali. Particolare attenzione è stata posta all'assegnazione del synset più adeguato per un certo termine in accordo

al suo significato e a quello del frame di appartenenza. A titolo di esempio vengono riportate le annotazioni sui frame degli studenti evidenziando alcune scelte che sono state prese relativamente all'annotazione di multiword expression e di termini composti.

Esempio di annotazione di alcuni termini

Frame	Termine	Definizione	Synset	Definizione synset
Piracy	Piracy	The words in this frame describe situations in which a Perpetrator forcibly seizes control over a Victim's Vehicle to gain some end.	Piracy.n.01	Hijacking taking a ship or plane away from the control of those who are legally entitled to it. "air piracy"
Usefulness	Usefulness	An Entity aids in the successful completion of a Purpose.	Utility.n.02	the quality of being of practical use
Departing	Containing_event (event)	This FE denotes an event that occurs or state of affairs that holds at a time that includes the time during which the event or state of affairs reported by the target occurs and of which it is taken to be a part.	event.n.01	something that happens at a given place and time
Immobilization	lock up.v (lock)	make or become rigidly fixed or immovable by locking'	lock.v.03	become rigid or immoveable
Education_teaching (teaching)	teaching	An Entity aids in the successful completion of a Purpose.	education.n.01	This frame contains words referring to teaching and the participants in teaching....

Come si può notare, parole composte come "lock up" sono state troncate prelevando il termine reggente, (lock) e assegnando ad essi il synset rispettando la definizione su frameNET del termine originale.

Per il processo di disambiguazione, come algoritmo di mapping è stato considerato l'approccio a "**bag of words**" con scelta del senso che permette di massimizzare l'intersezione fra i contesti: da una parte la definizione del termine e dall'altra glosse ed esempi del synset e di tutti i suoi iponimi ed iperonimi. Le informazioni sui contesti sono state pre-processate eseguendo rimozione di simboli di punteggiatura e di spazi multipli, lemmatizzazione e rimozione delle stopwords. Infine viene prodotto un set eliminando i duplicati. Il file "**DataPreprocessing.py**" contiene la definizione della funzione "**bag_of_words_mapping**" che per l'appunto svolge le operazioni appena dette.

Nei documenti in allegato all'esercizio è presente un file "**main.py**". Seguendo il flusso di esecuzione del programma verranno illustrati grossomodo i dettagli implementativi. Il programma istanzia un oggetto della classe **FrameNetDisambiguation** che riceve come costruttore il **path** al file json contenente le annotazioni dei termini. In seguito vengono prelevate queste informazioni ed aggregate in oggetti di tipo **StudentFrames** in modo da associare ad ogni studente le corrette liste di frame, frame element e lexical unit. In questa fase si fa attenzione a discriminare un

termine semplice da uno composto, ricordando che nel file json i synset per questi casi particolari sono stati annotati manualmente. Ogni parola nel file viene considerata come generico elemento lessicale, dunque una volta recuperato il contesto da frameNET viene creato un oggetto della classe **DisambiguateTerm**, che riceve come costruttore il termine da disambiguare, il suo synset e il bag of words mapping fatto sulla sua definizione. L'oggetto, all'atto della creazione disambigua il termine recuperando da wordNET la lista dei synset associati e procedendo nella costruzione del contesto come spiegato in precedenza. Infine il metodo evaluation della classe FrameNetSynset si occupa di valutare il processo di disambiguazione per ogni frameSet di ogni studente.

La valutazione consiste nel confrontare i synset restituiti in output dal sistema con quelli annotati nel file "annotation", in particolare viene calcolata l'accuratezza del sistema come rapporto degli elementi corretti sul totale degli elementi annotati. A termine del processo di valutazione, il sistema restituisce i seguenti risultati:

- Termini presenti nel file "annotation": 189
- Termini annotati correttamente: 118
- Accuratezza: 62.0%

Le informazioni sulle singole annotazioni sono reperibili in due file di testo, uno per ogni studente, nella cartella output. L'accuratezza restituita va da riferirsi ai soli termini annotati che, come spiegato in precedenza, rappresentano un sottoinsieme degli elementi lessicali presenti su FrameNET per i frameSet considerati.