Marco Ottina 795130
Davide Sgaramella 938632

A.A. 2019/2020

Università degli Studi di Torino – polo Scienze della Natura
Laurea Magistrale in Informatica
Corso: Tecnologie del linguaggio naturale
Parte di: Luigi Di Caro

Esercitazione 3

L'esercitazione proposta prevede l'implementazione della teoria di P. Hanks. Tale teoria afferma che il verbo rappresenta la "***radice del significato***".

Si è deciso di usare una valenza pari a due in modo da considerare i due argomenti del verbo, in questo caso **transitivo**. Essenzialmente sono stati utilizzati due corpus di frasi generati da **Sketch Engine** relativi ad altrettanti verbi:

- To build;
- To cook.

Al fine di estrarre il *filler* di ogni frase, viene prevista una fase di ***preprocessing***: dapprima si è pulito il testo da eventuale punteggiatura, poi è stato effettuato il *parsing* delle frasi in modo da trovare i filler relativi alle frasi: ***soggetto*** e ***oggetto***.

Una volta trovati i possibili filler si è passati alla fase di *disambiguazione* degli stessi: per questo passaggio si è usato il metodo lesk() messo a disposizione dalla libreria Python "***nltk***". È opportuna una precisazione in questo caso: ci sono alcuni lessemi che non possono esseri disambiguati, in quanto non hanno un synset associato in WordNet, vale la stessa considerazione per i nomi propri. Per questo motivo ai pronomi, che in inglese vengono usati come soggetti (I, You, He/She/It, We, They), è stato associato un particolare synset: la scelta è ricaduta su due parole:

- *People*;
- *Person*.

La stessa cosa è stata effettuata sugli oggetti, in questo caso la scelta era limitata al corpus che si è scelto di analizzare, infatti, si sono scelte due parole molto esplicative:

- *Thing;*
- *Food*.

Alla luce di tutto ciò, è stata effettuato un ulteriore controllo per suddividere i pronomi sopracitati in singolari e plurali. Infatti, ai pronomi singolari è stato assegnato il synset ‘person’, mentre a quelli plurali il synset ‘people’.

Dopo aver effettuato la fase di disambiguazione si è passati alla fase di ***ricerca dei super-sensi***. In questo caso si è fatto ricorso al metodo lexname() che restituisce il super-senso della parola a cui viene applicato.

In questo modo sono state trovate le coppie di ***semantic types***, successivamente fornite in output con le relative frequenze.

Risultati

Di seguito sono mostrati i risultati ottenuti dalla sperimentazione (non tutti, solo i principali). Inoltre, si è anche previsto la generazione delle Word Cloud relative agli slot per ogni verbo analizzato.

Una premessa è dovuta in quanto ci sono delle differenze nei risultati in base al synset utilizzato nella fase di disambiguazione. Perciò sono state effettuate inizialmente 2 fasi di testing, una per “People” e una per “Person”.

Come si può notare dai risultati sottostanti, i super-sensi trovati sono differenti:

- *People → noun.group;*
- *Person → noun.tops.*

La differenza chiaramente non è solo a livello di categoria, ma di ciò che quella categoria di super-sensi rappresenta.

Il noun.group contiene tutte quelle parole la cui semantica è legata ad un insieme di persone e/o oggetti (ad esempio: pubblico, assemblea). Quindi si avrà che il super-senso in questione è quello predominante.

Il noun.tops, invece, si riferisce a circa 40 synset più generali come ad esempio fenomeno, entità, oggetto.

Quindi nel contesto del verbo to build ritengo sia più utile dare un maggior peso al primo synset, ovvero quello di People, in quanto meglio rappresenta il corpus in questione a differenza del secondo.

Invece per il secondo corpus, può essere più adatto il synset ‘person’ per il significato del verbo to cook, in quanto la maggior parte delle volte chi compie l’atto di cucinare qualcosa è sempre una sola persona.

È importante notare anche come nei risultati del secondo corpus cresca la presenza del super-senso ‘noun.Tops’ oltre che nei soggetti, anche negli oggetti. Questo è

dovuto soprattutto alla generalità del super-senso stesso. Questa cosa non accade nel corpus ‘to build’ in quanto gli oggetti di tale verbo sono essenzialmente prodotti dall’uomo (noun.artifacts).

Un’altra considerazione da fare riguarda la bassa percentuale del super-senso “*noun.person*” nello slot 1 relativo ai soggetti di entrambi i corpus: questo è dovuto al fatto che nei dataset utilizzati non sono presenti molte frasi che riconducono a quel super-senso, ovvero non hanno un nome che identifica una persona come soggetto, **ad esempio** architetto, cuoco, progettista, padre ecc.

Inoltre, è stata effettuata una terza fase di sperimentazione per ricavare i risultati dopo la suddivisione dei pronomi tra singolari e plurali.

Nei risultati relativi al corpus ‘**to build**’ è possibile quindi notare come in questa fase ci sia un maggior equilibrio tra i due synset e di conseguenza i relativi super-sensi noun.group per i plurali e noun.Tops per i singolari.

Questo invece non accade nel corpus ‘**to cook’** in quanto vi è una maggiore presenza del synset person e quindi del suo super-senso ‘noun.Tops’. Si nota una riduzione del suo impiego rispetto a ‘noun.group’, ma ‘noun.Tops’ rimane predominante in questo caso.

To build corpus

Utilizzo di People.

Top 10 semantic types:

```
Semantic Type:   ('noun.group', 'noun.state') 9.29  %
Semantic Type:   ('noun.group', 'noun.artifact') 6.09  %
Semantic Type:   ('noun.group', 'noun.group') 5.49  %
Semantic Type:   ('noun.group', 'noun.communication') 4.9  %
Semantic Type:   ('noun.group', 'noun.act') 4.8  %
Semantic Type:   ('noun.group', 'noun.cognition') 4.0  %
Semantic Type:   ('noun.group', 'noun.person') 2.9  %
Semantic Type:   ('noun.group', 'noun.attribute') 2.3  %
Semantic Type:   ('noun.group', 'noun.location') 2.1  %
Semantic Type:   ('noun.group', 'adj.all') 1.8  %
```

Top 10 Filler slot 1:

```
Filler slot 1:   group 59.64  %
Filler slot 1:   artifact 7.19  %
Filler slot 1:   person 4.4  %
Filler slot 1:   communication 4.2  %
Filler slot 1:   act 3.8  %
Filler slot 1:   cognition 3.1  %
Filler slot 1:   all 2.0  %
Filler slot 1:   location 1.8  %
Filler slot 1:   attribute 1.5  %
Filler slot 1:   possession 1.0  %
```

Top 10 Filler slot 2:

```
Filler slot 2:   state 14.79  %
Filler slot 2:   artifact 10.09  %
Filler slot 2:   communication 9.29  %
Filler slot 2:   group 8.39  %
Filler slot 2:   act 8.19  %
Filler slot 2:   cognition 7.69  %
Filler slot 2:   person 5.19  %
Filler slot 2:   location 4.0  %
Filler slot 2:   attribute 4.0  %
Filler slot 2:   all 3.1  %
```

Utilizzo di Person.

Top 10 semantic types:

```
Semantic Type:   ('noun.Tops', 'noun.state') 7.89  %
Semantic Type:   ('noun.Tops', 'noun.artifact') 5.69  %
Semantic Type:   ('noun.Tops', 'noun.group') 4.8  %
Semantic Type:   ('noun.Tops', 'noun.communication') 4.5  %
Semantic Type:   ('noun.Tops', 'noun.act') 4.1  %
Semantic Type:   ('noun.Tops', 'noun.cognition') 3.5  %
Semantic Type:   ('noun.Tops', 'noun.person') 2.8  %
Semantic Type:   ('noun.Tops', 'noun.attribute') 2.2  %
Semantic Type:   ('noun.Tops', 'adj.all') 1.8  %
Semantic Type:   ('noun.Tops', 'noun.location') 1.8  %
```

Top 10 Filler slot 1:

```
Filler slot 1:   Tops 53.95  %
Filler slot 1:   artifact 7.19  %
Filler slot 1:   group 5.89  %
Filler slot 1:   person 4.4  %
Filler slot 1:   communication 4.2  %
Filler slot 1:   act 3.8  %
Filler slot 1:   cognition 3.1  %
Filler slot 1:   all 2.0  %
Filler slot 1:   location 1.8  %
Filler slot 1:   attribute 1.5  %
```

Top 10 Filler slot 2:

```
Filler slot 2:   state 14.79  %
Filler slot 2:   artifact 10.09  %
Filler slot 2:   communication 9.29  %
Filler slot 2:   group 8.39  %
Filler slot 2:   act 8.19  %
Filler slot 2:   cognition 7.69  %
Filler slot 2:   person 5.19  %
Filler slot 2:   location 4.0  %
Filler slot 2:   attribute 4.0  %
Filler slot 2:   all 3.1  %
```

Suddivisione dei pronomi singolari e plurali.

Top 10 semantic types:

```
Semantic Type:   ('noun.group', 'noun.state') 4.9  %
Semantic Type:   ('noun.Tops', 'noun.state') 4.5  %
Semantic Type:   ('noun.Tops', 'noun.artifact') 3.2  %
Semantic Type:   ('noun.group', 'noun.act') 3.1  %
Semantic Type:   ('noun.group', 'noun.group') 3.1  %
Semantic Type:   ('noun.group', 'noun.artifact') 3.0  %
Semantic Type:   ('noun.Tops', 'noun.communication') 2.7  %
Semantic Type:   ('noun.Tops', 'noun.group') 2.4  %
Semantic Type:   ('noun.group', 'noun.cognition') 2.3  %
Semantic Type:   ('noun.group', 'noun.communication') 2.2  %
```

Top 10 Filler slot 1:

```
Filler slot 1:   group 31.37  %
Filler slot 1:   Tops 28.47  %
Filler slot 1:   artifact 7.19  %
Filler slot 1:   person 4.4  %
Filler slot 1:   communication 4.2  %
Filler slot 1:   act 3.8  %
Filler slot 1:   cognition 3.1  %
Filler slot 1:   all 2.0  %
Filler slot 1:   location 1.8  %
Filler slot 1:   attribute 1.5  %
```

Top 10 Filler slot 2:

```
Filler slot 2:   state 14.79  %
Filler slot 2:   artifact 10.09  %
Filler slot 2:   communication 9.29  %
Filler slot 2:   group 8.39  %
Filler slot 2:   act 8.19  %
Filler slot 2:   cognition 7.69  %
Filler slot 2:   person 5.19  %
Filler slot 2:   location 4.0  %
Filler slot 2:   attribute 4.0  %
Filler slot 2:   all 3.1  %
```

To cook corpus

Utilizzo di People.

Top 10 semantic types:

```
Semantic Type:   ('noun.group', 'noun.Tops') 10.11  %
Semantic Type:   ('noun.group', 'noun.artifact') 6.25  %
Semantic Type:   ('noun.group', 'noun.food') 5.05  %
Semantic Type:   ('noun.group', 'noun.person') 3.8  %
Semantic Type:   ('noun.group', 'noun.time') 3.25  %
Semantic Type:   ('noun.group', 'noun.cognition') 3.2  %
Semantic Type:   ('noun.group', 'noun.group') 2.75  %
Semantic Type:   ('noun.group', 'noun.act') 2.75  %
Semantic Type:   ('noun.group', 'noun.communication') 2.0  %
Semantic Type:   ('noun.group', 'noun.attribute') 2.0  %
```

Top 10 Filler slot 1:

```
Filler slot 1:   group 61.28  %
Filler slot 1:   person 7.1  %
Filler slot 1:   food 4.1  %
Filler slot 1:   artifact 3.0  %
Filler slot 1:   all 2.5  %
Filler slot 1:   act 2.2  %
Filler slot 1:   communication 2.05  %
Filler slot 1:   cognition 1.95  %
Filler slot 1:   contact 1.4  %
Filler slot 1:   plant 1.35  %
```

Top 10 Filler slot 2:

```
Filler slot 2:   Tops 15.26  %
Filler slot 2:   artifact 10.46  %
Filler slot 2:   food 8.5  %
Filler slot 2:   person 5.8  %
Filler slot 2:   cognition 5.6  %
Filler slot 2:   time 5.45  %
Filler slot 2:   communication 4.35  %
Filler slot 2:   act 4.25  %
Filler slot 2:   all 4.05  %
Filler slot 2:   group 4.0  %
```

Utilizzo di Person.

Top 10 semantic types:

```
Semantic Type:   ('noun.Tops', 'noun.Tops') 9.35  %
Semantic Type:   ('noun.Tops', 'noun.artifact') 5.65  %
Semantic Type:   ('noun.Tops', 'noun.food') 4.9  %
Semantic Type:   ('noun.Tops', 'noun.person') 3.2  %
Semantic Type:   ('noun.Tops', 'noun.cognition') 3.15  %
Semantic Type:   ('noun.Tops', 'noun.time') 2.8  %
Semantic Type:   ('noun.Tops', 'noun.act') 2.5  %
Semantic Type:   ('noun.Tops', 'noun.group') 2.3  %
Semantic Type:   ('noun.Tops', 'noun.communication') 1.9  %
Semantic Type:   ('noun.Tops', 'noun.attribute') 1.75  %
```

Top 10 Filler slot 1:

```
Filler slot 1:   Tops 56.18  %
Filler slot 1:   person 7.1  %
Filler slot 1:   group 5.65  %
Filler slot 1:   food 4.1  %
Filler slot 1:   artifact 3.0  %
Filler slot 1:   all 2.5  %
Filler slot 1:   act 2.2  %
Filler slot 1:   communication 2.05  %
Filler slot 1:   cognition 1.95  %
Filler slot 1:   contact 1.4  %
```

Top 10 Filler slot 2:

```
Filler slot 2:   Tops 15.26  %
Filler slot 2:   artifact 10.46  %
Filler slot 2:   food 8.5  %
Filler slot 2:   person 5.8  %
Filler slot 2:   cognition 5.6  %
Filler slot 2:   time 5.45  %
Filler slot 2:   communication 4.35  %
Filler slot 2:   act 4.25  %
Filler slot 2:   all 4.05  %
Filler slot 2:   group 4.0  %
```

Suddivisione dei pronomi singolari e plurali.

Top 10 semantic types:

```
Semantic Type:   ('noun.Tops', 'noun.Tops') 6.7  %
Semantic Type:   ('noun.Tops', 'noun.artifact') 4.4  %
Semantic Type:   ('noun.group', 'noun.Tops') 3.45  %
Semantic Type:   ('noun.Tops', 'noun.food') 3.35  %
Semantic Type:   ('noun.Tops', 'noun.person') 2.45  %
Semantic Type:   ('noun.Tops', 'noun.cognition') 2.3  %
Semantic Type:   ('noun.Tops', 'noun.time') 2.15  %
Semantic Type:   ('noun.group', 'noun.artifact') 2.0  %
Semantic Type:   ('noun.group', 'noun.food') 1.7  %
Semantic Type:   ('noun.Tops', 'noun.act') 1.7  %
```

Top 10 Filler slot 1:

```
Filler slot 1:   Tops 39.97  %
Filler slot 1:   group 21.86  %
Filler slot 1:   person 7.1  %
Filler slot 1:   food 4.1  %
Filler slot 1:   artifact 3.0  %
Filler slot 1:   all 2.5  %
Filler slot 1:   act 2.2  %
Filler slot 1:   communication 2.05  %
Filler slot 1:   cognition 1.95  %
Filler slot 1:   contact 1.4  %
```

Top 10 Filler slot 2:

```
Filler slot 2:   Tops 15.26  %
Filler slot 2:   artifact 10.46  %
Filler slot 2:   food 8.5  %
Filler slot 2:   person 5.8  %
Filler slot 2:   cognition 5.6  %
Filler slot 2:   time 5.45  %
Filler slot 2:   communication 4.35  %
Filler slot 2:   act 4.25  %
Filler slot 2:   all 4.05  %
Filler slot 2:   group 4.0  %
```