Marco Ottina 795130
Sgaramella Davide 938632 A.A. 2019/2020

Università degli Studi di Torino – polo Scienze della Natura
Laurea Magistrale in Informatica
Corso: Tecnologie del Linguaggio Naturale
Parte terza - professor Luigi Di Caro

Esercitazione 3

L'esercitazione proposta prevedeva l'implementazione della teoria di P. Hanks in cui si ha che il verbo è 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 **pre-processing**: 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**".

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.

To build corpus

Top 10 semantic types:

```
('noun.group', 'noun.artifact') 5.47  %
('noun.group', 'noun.communication') 5.0  %
('noun.group', 'noun.group') 4.84  %
('noun.group', 'noun.act') 4.06  %
('noun.group', 'noun.cognition') 3.75  %
('noun.group', 'noun.person') 2.34  %
('noun.group', 'noun.location') 1.41  %
('noun.group', 'noun.attribute') 1.41  %
('noun.artifact', 'noun.artifact') 1.25  %
('noun.group', 'noun.state') 1.25  %
```

Top 10 Filler slot 1:

```
group 42.5  %
artifact 10.16  %
person 6.09  %
communication 5.94  %
act 5.62  %
cognition 4.53  %
all 2.66  %
location 2.5  %
attribute 2.19  %
possession 1.56  %
```

Top 10 Filler slot 2:

```
artifact 11.56  %
communication 11.41  %
act 9.53  %
group 9.53  %
cognition 9.06  %
person 6.09  %
location 4.69  %
attribute 4.22  %
time 3.59  %
state 2.97  %
```

To cook corpus

Top 10 semantic types:

```
('noun.group', 'noun.artifact') 6.94  %
('noun.group', 'noun.food') 4.15  %
('noun.group', 'noun.person') 3.7  %
('noun.group', 'noun.cognition') 3.47  %
('noun.group', 'noun.time') 3.17  %
('noun.group', 'noun.group') 2.57  %
('noun.group', 'noun.act') 2.49  %
('noun.group', 'noun.communication') 1.81  %
('noun.group', 'adj.all') 1.74  %
('noun.group', 'noun.attribute') 1.66  %
```

Top 10 Filler slot 1:

```
group 48.75  %
person 9.43  %
food 5.13  %
artifact 4.3  %
all 3.4  %
communication 2.79  %
act 2.72  %
cognition 2.42  %
contact 1.96  %
plant 1.74  %
```

Top 10 Filler slot 2:

```
artifact 13.28  %
food 9.36  %
cognition 6.87  %
person 6.72  %
time 6.49  %
communication 4.91  %
act 4.75  %
all 4.68  %
group 4.45  %
attribute 3.47  %
```