TLN-LAB:

Annotazione di *Corpora* e Sense Identification

Daniele Radicioni

Task

- The task is on Semantic Word Similarity
- Given a dataset on Multilingual and Cross-lingual Semantic Word Similarity, we focus on Semantic Similarity on the Italian language.
- The original dataset is available at
  - http://alt.qcri.org/semeval2017/task2/

Consegna 1: annotazione

- La prima operazione consiste nell'annotare con punteggio di semantic similarity 100 coppie di termini.
- Il criterio da utilizzare è il seguente (https://tinyurl.com/y6f8h2kd):
  - ***4: Very similar*** -- The two words are synonyms (e.g., midday-noon).
  - ***3: Similar*** -- The two words share many of the important ideas of their meaning but include slightly different details. They refer to similar but not identical concepts (e.g., lion-zebra).
  - ***2: Slightly similar*** -- The two words do not have a very similar meaning, but share a common topic/domain/function and ideas or concepts that are related (e.g., house-window).
  - ***1: Dissimilar*** -- The two items describe clearly dissimilar concepts, but may share some small details, a far relationship or a domain in common and might be likely to be found together in a longer document on the same topic (e.g., software-keyboard).
  - ***0: Totally dissimilar and unrelated*** -- The two items do not mean the same thing and are not on the same topic (e.g., pencil-frog).

Consegna 1: annotazione

- Annotare 100 coppie di termini del file
  `test/subtask1-monolingual/data/it.test.data.txt`
  con un punteggio di similarità fra i due elementi.
  - Le 100 coppie (sul totale di 500 coppie presenti nel file) sono da individuare sulla base del cognome, tramite il programma `sem_eval_mapper.py`.
- L'output di questa prima fase è un file di 100 linee, ciascuna contenente un numero in [0,4].

Consegna 1: annotazione

```
Joule astronave
Terra Promessa Baku
macchina bicicletta
poliedro attore
sclerosi multipla sclerosi a placche
faglia sistema
arma elmetto
sceneggiatore televisione
Nazioni Unite Ban Ki-moon
Si-o-se Polponte matematico
basilica mosaico
acquerello pennello
democrazia monarchia
Gauss scienziato
tubercolosiled
auto senza conducente auto autonoma
apocalisse fuoco
velocità posto
PlayStationWii
[...]
```

```
Joule astronave 1.5
Terra Promessa Baku 2
macchina bicicletta 2.8
poliedro attore 2.5
[...]
```

Consegna 1: annotazione

- Nel caso l'esercitazione sia svolta in gruppo, tutti i componenti del gruppo devono annotare tutte le coppie con il punteggio di similarity, e devono essere riportati tutti i valori forniti dagli annotatori, il valore medio, e l'inter-rater agreement fra gli annotatori
  - inter-rater agreement deve essere calcolato utilizzando come misura gli indici di correlazione di Pearson e Spearman.

Consegna 2: sense identification

- Il secondo compito consiste nell'individuare i sensi selezionati nel giudizio di similarità.
  - La domanda che ci poniamo è la seguente: nell'attribuire un valore di similarità a una coppia di termini (per esempio, *società* e *cultura*) quali sensi vengono effettivamente selezionati?
- Per risolvere questo compito partiamo dall'assunzione che i due termini funzionino come contesto di disambiguazione l'uno per l'altro.

Consegna 2: sense identification

- Come già visto nella prima esercitazione (sulle misure di similarity calcolate tramite WordNet) è possibile utilizzare la massimizzazione della similarity per selezionare i sensi, cioè

$$\mathrm{sim}(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [\mathrm{sim}(c_1, c_2)]$$

  - In questo caso possiamo utilizzare come metrica la cosine-similarity fra i vettori NASARI dei sensi associati ai vari termini w_i.

- A differenza di quanto fatto nella prima esercitazione, in questo caso non siamo però interessati a calcolare il punteggio di similarità fra 2 termini, ma a individuare i sensi che massimizzano tale punteggio.

Consegna 2: sense identification

- A differenza di quanto fatto nella prima esercitazione, in questo caso non siamo però interessati a calcolare il punteggio di similarità fra 2 termini, ma a individuare i sensi che massimizzano tale punteggio.
- Si tratta quindi di eseguire questa operazione:

$$c_1, c_2 \leftarrow \underset{c_1 \in s(w_1), c_2 \in s(w_2)}{\arg\max} [sim(c_1, c_2)]$$

Consegna 2: sense identification

- L'output di questa parte dell'esercitazione consiste in 2 BabelNet synset ID e dalla relativa glossa.
- Valutiamo il risultato ottenuto (cioè cioè la coppia dei sensi identificati, e la relativa appropriatezza) direttamente, stabilendo se i sensi in questione sono quelli idealmente selezionati da noi stessi nel momento dell'annotazione (Consegna 1).
  - Misuriamo in questo caso l'accuratezza sia sui singoli elementi, sia sulle coppie.
  - NB: per questa valutazione non è necessario conoscere il BabelNet synset ID corretto, ma è sufficiente valutare sulla base della glossa, se il senso individuato è appropriato.