di.unito.it

DIPARTIMENTO
DI INFORMATICA

TLN-LAB:

Annotazione di *Corpora* e Sense Identification

Daniele Radicioni

Consegna 1: semantic similarity

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 50 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 50 coppie di termini del file

  `it.test.data.txt`

  con il punteggio di similarità fra i due elementi della coppia.

  - Le 50 coppie (sul totale di 500 coppie presenti nel file) sono da individuare sulla base del cognome, tramite la funzione definita nel notebook **`semeval_mapper.ipynb`**.

- L'output della prima consegna è un file (in formato tsv) di 50 linee, ciascuna contenente un numero in [0,4].

Consegna 1: annotazione

```
Joule   astronave
Terra Promessa Baku
macchina bicicletta
poliedro attore
sclerosi multiplasclerosi a placche
faglia sistema
Si-o-se Pol  ponte matematico
democrazia monarchia
Gauss  scienziato
auto senza conducente auto autonoma
apocalisse fuoco
velocità posto
PlayStation  Wii
[...]
```

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

Consegna 1: annotazione in gruppo

- Se l'esercitazione è svolta in gruppo, è necessario preliminarmente analizzare la coerenza dell'annotazione prodotta dal gruppo.
- In questo caso tutti i componenti del gruppo devono annotare le stesse coppie con il punteggio di similarity: quelle attribuite al primo —in base all'ordinamento alfabetico— cognome del gruppo.
  - per ogni coppia di termini devono essere riportati tutti i valori forniti dagli annotatori, e il valore medio;
  - è inoltre necessario calcolare l'agreement fra gli annotatori (inter-rater agreement), utilizzando gli indici di correlazione di Pearson e Spearman (in caso il gruppo sia costituito da 3 componenti, riportare la media fra le 2 coppie).

Consegna 1: valutazione

- La valutazione dei punteggi annotati dovrà essere condotta in rapporto alla similarità ottenuta utilizzando i vettori NASARI (versione embedded; file *mini_NASARI.tsv*, nel materiale della lezione).
  - Si tratta di vettori distribuzionali (saranno trattati diffusamente nella terza parte del corso); li utilizziamo massimizzando la cosine similarity al posto della generica funzione sim(c_1, c_2)

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

Consegna 1: valutazione

- Massimizzazione della cosine similarity al posto della generica funzione sim(c_1, c_2)

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

la *cos-sim* fra V_1 e V_2 (rappresentazioni vettoriali corrispondenti ai sensi c_1 e c_2) è

$$\text{cos-sim}(\overrightarrow{V_1}, \overrightarrow{V_2}) = \frac{\overrightarrow{V_1} \cdot \overrightarrow{V_2}}{\| \overrightarrow{V_1} \| \| \overrightarrow{V_2} \|}.$$

- Dove al numeratore abbiamo il prodotto interno dei due vettori, e al denominatore il prodotto delle lunghezze Euclidee di V_1 e V_2 entrambi di lunghezza N; la lunghezza Euclidea (o norma 2) di un generico vettore v è definita come

$$\| v \| = \left(\sum_{i=1}^{N} |v_i|^2 \right)^{1/2}.$$

- Per alcuni termini i vettori non sono presenti; in questi casi si esclude la coppia dalla valutazione.

di.unito.it
Dipartimento di Informatica

Consegna 1: valutazione

- La valutazione della nostra annotazione è condotta calcolando i coefficienti di Pearsons e Separman fra (la media dei) i punteggi annotati a mano e quelli calcolati con la versione embedded di NASARI.

- Questa valutazione è sostanzialmente diversa dalla verifica di agreement nell'annotazione:
  - nell'inter-rater agreement verifichiamo se gli annotatori umani hanno annotato correttamente i vari elementi, e se il task è chiaramente formulato/risolvibile univocamente da esseri umani;
  - La valutazione consiste invece nel calcolare il livello di correlazione fra giudizio umano e punteggi calcolati algoritmicamente (massimizzazione *cos-sim* dei vettori).

Consegna 2: individuazione dei sensi alla base del giudizio

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: quali sensi abbiamo effettivamente utilizzato quando abbiamo assegnato un valore di similarità a una coppia di termini (per esempio, *società* e *cultura*)?
  - NB: questa annotazione, sebbene svolta successivamente a quella della prima consegna, deve essere coerente con l'annotazione dei punteggi di similarità.
- 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

- L'output di questa parte dell'esercitazione consiste in 2 Babel synset ID e dai termini dei synset
  - il formato di output è quindi costituito da 6 campi (separatore fra campi la tabulazione, mentre usiamo la virgola ',' come separatore all'interno dello stesso campo):

```
#Term1 Term2 BS1 BS2 Terms_in_BS1 Terms_in_BS2

macchina bicicletta bn:00007309n bn:00010248n
auto,automobile,macchina bicicletta,bici,bike
```

Agreement nell'annotazione

- Calcoliamo nuovamente il livello di agreement nelle annotazioni, questa volta utilizzando il punteggio kappa di Cohen
  - Chi usa Python può utilizzare il `cohen_kappa_score` della libreria sklearn.metrics.
  - Se il gruppo di annotatori è formato da 3 componenti, calcolare la kappa di Cohen per ogni coppia e riportare la media risultante, che sarà il valore sintetico di agreement sulle annotazioni prodotte.

Cohen's kappa

From Wikipedia, the free encyclopedia

Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance. There is controversy surrounding Cohen's kappa due to the difficulty in interpreting indices of agreement. Some researchers have suggested that it is conceptually simpler to evaluate disagreement between items.[1] See the Limitations section for more detail.

Calculation [edit]

Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The first mention of a kappa-like statistic is attributed to Galton (1892);[2] see Smeeton (1985).[3]

The definition of κ is:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where p_o is the relative observed agreement among raters (identical to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreement then $\kappa = 1$. If there is no

Valutazione dell'annotazione

- Valutiamo il risultato ottenuto (cioè la coppia dei sensi identificati, e la relativa appropriatezza) in rapporto all'output di un semplice sistema realizzato come segue
  - Utilizziamo nuovamente i vettori NASARI (versione densa, embedded) presenti nel file *mini_NASARI.tsv*, disponibile all'interno del materiale della lezione.
    - NB: il file contiene solo i vettori per i synset associati ai termini delle coppie; NON tutti i termini delle coppie hanno un vettore…
  - Con tali vettori calcoliamo la coppia di sensi che massimizzano lo score di similarità

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

Valutazione dell'annotazione

- Misuriamo in questo caso l'accuratezza sia sui singoli elementi, sia sulle coppie.