L'annotazione per il trattamento automatico del linguaggio: quando, come e perché

Irene Russo - ILC CNR

Lingua e Computer - Esperienze dalla Ricerca in Elaborazione del Linguaggio Naturale
Università degli Studi della Basilicata

Di cosa parlerò

- importanza dell’annotazione per il trattamento automatico del linguaggio
- ruolo dell’annotazione nell’analisi del testo
  - operazionalizzazione delle ipotesi in ambito umanistico
  - esercizio: prova di operazionalizzazione delle ipotesi sul populismo nei discorsi politici
- conclusioni

Annotazione nel TAL

- molti strumenti di base per il TAL (POS taggers, parser sintattici, analizzatori morfologici) hanno bisogno di un *training set* di dati manualmente annotati in una fase precedente allo sviluppo del software dedicato al compito in questione
- questo tipo di annotazione va fatta prima dello sviluppo del software
- se il software già esiste e va migliorato, si può pensare ad un'annotazione mirata che è una correzione degli errori del software

Annotazione nel TAL

La grandezza del *training set* non segue regole precise:

- in genere vale il principio *the bigger, the better*
- più un fenomeno linguistico è variabile e complesso, più dati di *training* saranno necessari
- più un fenomeno linguistico è variabile e complesso, più l'annotazione sarà onerosa
- è necessario avere più annotatori (almeno 2 o 3) che si sovrappongono nell'annotazione di almeno una parte del *training set*

Chi è l’annotatore ideale per il TAL?

- un linguista madrelingua della lingua da annotare
- un linguista che ha studiato una determinata teoria sulla quale si basano le linee guida dell’annotazione
- un linguista in grado di proporre delle correzioni/integrazioni nelle linee guida

Accordo tra annotatori

Ogni insieme di dati rispetto a un fenomeno linguistico dovrebbe sempre essere annotato da almeno due annotatori per valutare l'accordo tra di essi

Se vi è un buon accordo:

- il fenomeno analizzato è consistente e sistematico e si può addestrare un sistema automatico che svolgerà lo stesso compito dell'annotatore
- il fenomeno analizzato è consistente e sistematico e quindi rimanda ad una categorizzazione cognitivamente plausibile (in caso di fenomeni semantici e pragmatici)

Quando c’è un buon accordo?

Dipende

Inter-annotator agreement:

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

- Nel caso dell’annotazione di categorie grammaticali a livello di parola, l’accordo deve essere altissimo
- Nel case dell’annotazione della polarità delle parole, può essere più basso
- Nel caso di un’annotazione graduata (es. quanto la parola x denota un evento), soglie di accorpamento per calcolare l’accordo

Cosa si intende per annotazione nel TAL

- Annotazione di fenomeni linguistici a granularità variabile (morfemi, lemmi, collocazioni, sintagmi etc.) ma comunque definiti e inquadrati da una teoria linguistica
- Annotatori esperti e in accordo (compito di annotazione oneroso, possibilità di avere un giudice)
- Copre tutti i livelli dell'analisi linguistica: fonetica, morfologia, sintassi, semantica e pragmatica
- Con l'annotazione di fenomeni semantici e pragmatici si verifica una transizione verso l'annotazione di fenomeni testuali

Dall'annotazione linguistica all'annotazione testuale

Esempio di annotazione semantica: TimeML

```
The
<EVENT  eid="e1" class="OCCURRENCE" tense="NONE" aspect="NONE">
rains
</EVENT>
<MAKEINSTANCE  eiid="ei1" eventID="e1"/>
<EVENT  eid="e2" class="OCCURRENCE" tense="PAST" aspect="NONE">
caused
</EVENT>
<MAKEINSTANCE  eiid="ei2" eventID="e2"/>
the
<EVENT  eid="e3" class="OCCURRENCE" tense="NONE" aspect="NONE">
flooding
</EVENT>
<MAKEINSTANCE  eiid="ei3" eventID="e3"/>
<TLINK  eventInstanceID="ei1" relatedToEvent="ei2" relType="IDENTITY"/>
<TLINK  eventInstanceID="ei2" relatedToEvent="ei3" relType="BEFORE"/>
```

Le inferenze fatta su base semantica/pragmatica sono sistematiche e generalizzabili

Dall'annotazione linguistica all'annotazione testuale

- Dal 2000 il TAL inizia ad interessarsi dell'espressione delle opinioni
- Classificazione dei documenti (in questo caso recensioni di film) sulla base della loro polarità (Pang et al. 2002)
- Chi posta una recensione online associa un *rating* che costituisce un' annotazione implicita
- Attraverso metodologie machine learning si è cercato di capire quali elementi testuali fossero associati al *rating*
- Prospettiva bottom-up: MPQA Opinion Corpus nel quale opinioni e stati privati (credenze, emozioni etc.) sono annotati manualmente nelle notizie (Wiebe et al. 2005)

Dall’annotazione linguistica all’annotazione testuale

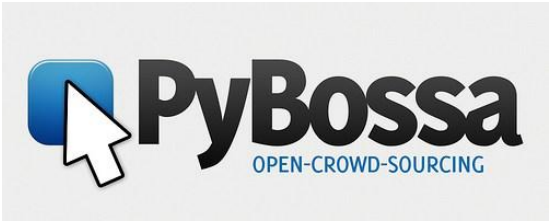
- L’ annotazione testuale può incentrarsi su una classificazione globale (recensioni positive o negative di uno smartphone) o su una classificazione puntuale all’interno di una classificazione globale

> *Telefono perfetto, originale samsung italia, al prezzo a cui si trova adesso è un ottimo acquisto. non risulta essere scomodo in tasca nonostante le dimensioni, funzionalità al top, nessun lag dopo mesi di utilizzo e la batteria ha ottime performance, quasi due giorni usandolo normalmente e disattivando l'always on display, con quest'ultimo attivo l'autonomia scende di un 20%.*

- Il primo tipo di annotazione viene ritenuto alla portata di annotatori non esperti e si può far ricorso a metodologie di analisi crowdsourcing

Cosa è il crowdsourcing

Delegare l'esecuzione di un compito a più persone esterne al gruppo di ricerca o all'azienda



Funziona? Dipende dalla motivazione degli annotatori...

L’annotazione testuale complessa in ambito umanistico

- L’analisi del testo è più complicata ed è influenzata da un maggior numero di variabili rispetto all’analisi dei fenomeni linguistici
- Per analizzare La Divina Commedia bisogna essere esperti di Dante

Chi è l’annotatore ideale per le *digital humanities*?

- un linguista (?) madrelingua rispetto alla lingua da annotare
- una persona che ha assimilato studi critici sui quale si basano le linee guida dell’annotazione OPPURE una persona che conosce in maniera ottimale la produzione testuale di riferimento, tanto da aver elaborato delle ipotesi di partenza su cosa si può ritrovare in un testo a partire dall’annotazione

Operazionalizzare le ipotesi

- Per passare dagli studi critici sui testi di un autore alla definizione delle linee guida per l'annotazione è necessario estrapolare dagli studi critici un insieme di ipotesi che possono essere operazionalizzate a livello di annotazione
- Operazionalizzare un'ipotesi significa dargli una traduzione a livello linguistico
- Solo con l'ancoraggio all'elemento linguistico è possibile elaborare uno schema di annotazione per testare le ipotesi di partenza

Alcuni esempi di ipotesi operazionalizzabili

- Uno scrittore, nell'ambito di una produzione decennale, tende a semplificare la struttura narrativa (per aumentare le vendite o perché la sua affiliazione a un filone sperimentale si interrompe)
- I romanzi di uno scrittore, nella versione data alle stampe, sono stati riscritti da uno o più *editors* (Raymond Carver)
- Non si conosce l'identità di un autore che si cela dietro uno pseudonimo ma ci sono delle ipotesi a partire da altri scritti di autori noti (Elena Ferrante)

Annotare Grazie Deledda

- Collaborazione con un linguistica sardo (Simone Pisano)
- Definizione di un insieme di fenomeni ritenuti rilevanti nella produzione letteraria di Grazia Deledda:
  - teoria del doppio codice
  - paesaggio emotivo
  - caratterizzazione dei personaggi sulla base del loro genere (rapporto con gli stereotipi di genere dell'epoca?)

Annotare Grazie Deledda, *Marianna Sirca* (1915)

<<Marianna>>, egli disse, prendendo la tazza con l'altra mano, senza lasciarle il polso; <<sì, quando mi dissero: "Marianna è alla Serra", pensai subito: "voglio andare a trovarla." Contenta sei, di vedermi?>>

Vide il cielo schiarirsi, e fra i cespugli brillare lo specchio d'una conca dove si raccoglieva l'acqua di una sorgente, che dopo essere scesa con impeto dai macigni sopra la grotta pareva fermarsi in mezzo a una ghirlanda di giunchi fioriti, per riposarsi, come faceva lui, prima di correre per la sua via.

Annotare per le scienze sociali: il discorso politico populista

populismo

po|pu|lì|ṣmo
s.m.
1921; dall'ingl. populism, der. del lat. populus "[3]popolo", cfr. russo narodničestvo.

1. **TS** stor. movimento politico e culturale sorto in Russia nel XIX sec., precedente al diffondersi del marxismo, che teorizzava il dovere degli intellettuali di porsi al servizio del popolo attraverso l'attività di propaganda rivoluzionaria volta a ottenere un miglioramento delle condizioni delle classi più povere
2. **TS** polit. estens., spreg., atteggiamento politico di esaltazione velleitaria e demagogica dei ceti più poveri
3. **TS** arte, lett. rappresentazione idealizzata del popolo in quanto considerato come depositario di valori etici e sociali

populismo [po-pu-lì-smo] s.m.

1 Atteggiamento o movimento politico tendente a esaltare il ruolo e i valori delle classi popolari

2 spreg. Atteggiamento demagogico volto ad assecondare le aspettative del popolo, indipendentemente da ogni valutazione del loro contenuto, della loro opportunità

3 Movimento rivoluzionario russo della fine del sec. XIX, che propugnava l'emancipazione delle classi contadine e dei servi della gleba attraverso la realizzazione di una sorta di socialismo rurale

4 In ambito artistico, raffigurazione idealizzata del popolo, presentato come modello etico positivo

• a. 1921

Il discorso politico populista

An ideology which pits a virtuous and homogeneous people against a set of elites and dangerous 'others' who are together depicted as depriving (or attempting to deprive) the sovereign people of their rights, values, prosperity, identity and voice.
(Albertazzi & McDonnell, 2008)

Esempi di discorsi populisti

E’ festa qui tra il popolo sovrano della Padania, con il Parlamento ed il Governo Padani. Ed è festa là nel parlamento romano, con il potere costituito e rigido, con il potere mummificato schierato in difesa di una Costituzione che i cosiddetti Padri della Patria previdero immortale come le rovine di Roma.
In realtà è festa soltanto qui, dove si parla di libertà e di democrazia. Qui è la festa del coraggio della giovane ed eterna nazione padana. E’ il raduno della paura e dei risentimenti statalisti a Roma. 2 giugno. A Roma è festa del tradimento delle speranze di chi è morto lottando per la libertà. I morti si onorano con i fatti, non soltanto con le corone di fiori e con i bei discorsi.

Umberto Bossi, Pontida, 1996

Su questo prato c'è gente perbene e io vi chiedo di dire si o no a una proposta di legge che presenteremo quando andremo al governo: giudici eletti direttamente dal popolo, si o no? E chi sbaglia paga. [...] Quando saremo al governo daremo mano libera alle forze dell'ordine per portare pulizia e sicurezza alle nostre città. Questo impegno ce lo prendiamo si o no? [...] Quando ci mandate al governo lo reintroduciamo quello che è un limite preciso e di buon senso: entri in casa mia se chiedi permesso. Lo introduciamo il reato di immigrazione clandestina, si o no?

Matteo Salvini, Pontida, 2017

Annotatione crowdsourcing?

- Che scala di valori caratterizza le persone che simpatizzano con i discorsi populisti?

- Secondo la teoria dei valori di base di Schwartz potrebbe essere benevolence: *preserving and enhancing the welfare of those with whom one is in frequent personal contact (the 'in-group')*.

Su questo prato c'è gente perbene e io vi chiedo di dire si o no a una proposta di legge che presenteremo quando andremo al governo: giudici eletti direttamente dal popolo, si o no? E chi sbaglia paga. [...] Quando saremo al governo daremo mano libera alle forze dell'ordine per portare pulizia e sicurezza alle nostre città. Questo impegno ce lo prendiamo si o no? [...] Quando ci mandate al governo lo reintroduciamo quello che è un limite preciso e di buon senso: entri in casa mia se chiedi permesso. Lo introduciamo il reato di immigrazione clandestina, si o no?

Matteo Salvini, Pontida, 2017

L’annotazione nelle scienze umane e sociali

Può avvalersi dell’uso di strumenti sviluppati per il TAL:

- per l’annotazione dei toponimi in un testo ci si può avvalere di un POS tagger che ritrova le entità nominate
- costruzioni marcate in una lingua possono essere individuate come una sequenza di POS
- porzioni di testo dove viene narrato lo sviluppo emotivo del personaggio possono essere ritrovate partendo da una risorsa lessicale, come WordNet-Affect che contiene tutti i synsets relativi ai concetti emotivi (Strapparava & Valitutti 2004)

Può contribuire allo sviluppo di strumenti per il TAL

Conclusioni

Ma oggi con il deep learning abbiamo ancora bisogna dell'annotazione manuale?

- SI! Perché:
- chi può insegnare ad una macchina come funziona il linguaggio se non chi lo ha inventato?
- quali sono i limiti del linguaggio (soprattutto a livello di generazione) se non quelli modellati sulla comprensibilità per un utente umano?
- come può essere possibile processare contenuto testuale in base a categorie di alto livello (positivo/negativo, commento aggressivo/non aggressivo) se queste categorie non sono state predefinite da utenti umani?
- la classificazione automatica è supportata dalla frequenza di elementi ritenuti salienti rispetto ad una determinata classe, la salienza è stabilita dall'annotatore

Conclusioni

Si, ma io non ho tempo per fare annotazione!

- se il compito che ti proponi è nuovo e non esistono dati annotati, annotare un piccolo insieme di unità (insieme ad un'altra persona) è l'unico modo per rendersi conto della generalizzabilità del compito
- se il compito che ti proponi è nuovo ma riesci a trovare un insieme di dati già annotati in maniera implicita dagli utenti online, usali
- se il compito che ti proponi è nuovo e hai bisogno dell'annotazione di molte persone (o di molti elementi annotati da persone diverse) pensa alla motivazione di chi partecipa al crowdsourcing
- se il compito che ti proponi è nuovo e hai bisogno di annotatori esperti, pensa a un ciclo di sviluppo furbo che non abusi del loro lavoro (e della loro pazienza) (annotazione di singole istanze opportunamente selezionate per algoritmi di reinforcement learning)