Opinion Mining

Stefano Baccianella

Università di Pisa - A.A. 2006/2007

Sommario

- Introduzione
- Task: Determinare l'orientamento dei termini
- Task: Determinare la soggettività dei termini
- Proposte di soluzione al problema

Introduzione

Cosa vuol dire Opinion Mining?

- Opinion Mining (OM) è una disciplina molto recente, un incrocio tra l'information retrieval e la linguistica computazionale.
- Si concentra non sull'argomento di cui parla un documento ma sull'opinione che il documento stesso esprime.
- Sentiment Analysis, Sentiment Classification, Opinion Extraction sono altri nomi usati per identificare questa disciplina in letteratura.

Introduzione (cont'd)

Cosa è un'opinione?

- Un'opinione è un'affermazione che non può essere verificata in modo oggettivo. (Quirk et al., 1985)

Esempi di problemi di OM

- Qual è l'opinione generale sulla riforma delle tasse?
- I nostri clienti sono soddisfati del prodotto? Perché?

Componenti base di un'opinione

- Detentore dell'opinione: la persona o l'organizzazione che possiede un'opinione specifica su di un particolare oggetto.
- Oggetto: oggetto dell'opinione
- Opinione: un'affermazione, frase o attitudine su di un oggetto da parte del detentore dell'opinione.

Task

- Determinare la soggettività di un documento:

  consiste nel determinare se un testo descrive un fatto in modo oggettivo (senza opinioni) oppure esprime un'opinione sull'argomento.

  Questo si riduce ad una categorizzazione binaria tra Oggettivo e Soggettivo (Pang e Lee, 2004; Yu e Hatzivassiloglou, 2003);

Task (cont'd)

- Determinare l'orientamento (o polarità):

  consiste nel determinare se un testo *soggettivo* esprime un'opinione positiva o negativa sull'argomento.
- Determinare la forza dell'orientamento:

  consiste nel determinare quanto l'opinione *positiva o negativa* sia forte (Debolmente Positiva, Mediamente, Fortemente …)

Sommario

- Introduzione
- Task: Determinare l'orientamento dei termini
- Task: Determinare la soggettività dei termini
- Proposte di soluzione al problema

Task: Determinare l'orientamento dei termini

Si tratta del task principale:

- è abbastanza facile scrivere una risorsa con termini taggati positivi o negativi, non è altrettanto facile determinare se essi siano oggettivi.

Hatzivassiloglou, McKeown (1997)

- Tentarono di predire l'orientamento analizzando coppie di aggettivi (unite da e,o,oppure, ecc...)
- L'intuizione di base sta nel fatto che l'unione di aggettivi è soggetta a vincoli sull'orientamento:

  "e" unisce aggettivi dello stesso orientamento, "ma" orientamenti opposti

Hatzivassiloglou, McKeown (1997)

- Partendo da questo, si genera un grafo dove i nodo sono i termini connessi da due tipi di archi: "uguali", "opposti"
- Si applica un algoritmo di clustering che partiziona il grafo in un gruppo positivo e uno negativo.

Turney, Littman (2003)

- Partono da dei piccoli set di termini soggettivi (seed), uno con termini positivi (buono, carino, …) e uno con termini negativi (cattivo, brutto, …)
- I termini si analizzano calcolando la *pointwise mutual information (PMI),* che altro non è che l'associazione semantica del termine al seed

Turney, Littman (2003)

- Dato un termine *t* il suo orientamento *O(t)* (la positività è determinata dal segno, la forza dal valore), è calcolato dalla somma delle distanze con i componenti positivi del seed meno la somma con le componenti negative.

$$O(t) = \sum(t - t_p) - \sum(t - t_n)$$

Turney, Littman (2003)

- Il calcolo della distanza semantica viene realizzato eseguendo delle query al motore AltaVista su un set di pagine noto
- Il valore della distanza è calcolato a partire dal numero di risultati ritornati dalle query: "*t*", "t_i", "*t* NEAR t_i"

Kamps (2004)

- Kamps utilizza invece il grafo definito sugli aggettivi utilizzando la relazione di sinonimia all'interno di WordNet.
- Determina l'orientamento di un aggettivo *t* comparando:
  - La lunghezza del cammino minimo tra *t* e *buono*
  - La lunghezza del cammino minimo tra *t* e cattivo

  Se *t* è più vicino a *buono* sarà classificato come *positivo*, *negativo* altrimenti

Kim, Hovy (2004)

- Cambiano approccio, dando un punteggio di positività *e* uno di negatività ai termini
- In questo modo si evidenzia come i termini possano avere orientamenti sia positivi che negativi, e con quanta forza portano quell'orientamento

Kim, Hovy (2004)

- Il sistema parte, come molti altri, da un set di termini positivi e negativi (seed) e lo espande aggiungendo sinonimi e antinomi
- Il sistema classifica un termine *t* basandosi sulla probabilità di apparire nel set espanso
- Uno dei limiti di questo metodo è che possono essere classificati solo termini che condividono sinonimi o antinomi col seed

Sommario

- Introduzione
- Task: Determinare l'orientamento dei termini
- Task: Determinare la soggettività dei termini
- Proposte di soluzione al problema

Task: Determinare la soggettività dei termini

- E' un task chiave, ancora molto difficile da ottenere.
- Permette di scremare i termini interessanti (Soggettivi) da quelli non interessanti (Oggettivi)

Riloff (2003)

- Utilizza un algoritmo *bootstrap* per determinare i *nomi* soggettivi
- Usa per il bootstrap un set di 20 termini giudicati dall'autore come *fortemente soggettivi* e *molto frequenti* nel testo da cui il *nome* proviene

Baroni, Vegnaduzzo (2004)

- Applicano il già visto algoritmo PMI di Turney e Littman per determinare non l'orientamento dei termini ma la loro soggettività
- Il metodo utilizza un set S_s di 35 aggettivi marcati come soggettivi da un valutatore umano, per dare un punteggio di soggettività

Sommario

- Introduzione
- Task: Determinare l'orientamento dei termini
- Task: Determinare la soggettività dei termini
- Proposte di soluzione al problema

SentiWordNet

- Sviluppato da Andrea Esuli (ISTI-CNR, Pisa) e Fabrizio Sebastiani (Dip. Matematica, Università di Padova)
- E' un'estensione di WordNet, assegna ad ogni synset (un insieme di sinonimi) tre punteggi: *Obj(s), Pos(s) e Neg(s),* che stanno a indicare quanto oggettivi, positivi o negativi siano quei termini.

SentiWordNet

- L'assunzione di fondo che permette di assegnare i punteggi ai synset e non ai termini è che differenti significati di un termine portano opinioni differenti.
- Ogni punteggio varia tra 0.0 e 1.0, e la loro somma è sempre 1.0.
- Es.: Synset [estimable(3)], corrispondente a: "may be computed or estimated", dell'aggettivo estimable, ha punteggio: Obj 1.0. Mentre il synset [estimable(1)] corrispondente a "deserving of respect or high regard" ha un punteggio: Pos 0.75, e Obj 0.25.

SentiWordNet

- La costruzione di SentiWordNet parte dalla classificazione dei synset.
- Allo scopo sono stati predisposti vari classificatori ternari, differenti per il training set adottato per il loro allenamento, e si è fatto classificare ogni synset da tutti i classificatori, i punteggi ottenuti sono proporzionali ai risultati dei classificatori.
  (es. se tutti i classificatori ritengono positivo un synset, Pos sarà 1.0)

SentiWordNet

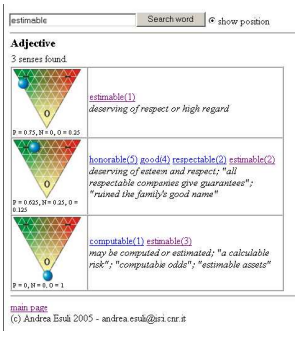
- Classificazione dei synset, in media:

Tipo	Positivo	Negativo	Oggettivo
Aggettivi	0,106	0,151	0,734
Sostantivi	0,022	0,034	0,944
Verbi	0,026	0,034	0,940
Avverbi	0,235	0,067	0,698
Tutti	0,043	0,054	0,903

SentiWordNet

- Valutare la bontà e l'affidabilità della rete in modo sperimentale è impossibile, sarebbe necessario taggare manualmente WordNet
- E' al vaglio la costruzione di un corpus di 1000 synset da usare per la valutazione, ma bisogna comunque notare che 1000 synset sono solo l'1% di WordNet

SentiWordNet



Blog Mining

- Sviluppato da Giuseppe Attardi e Maria Simi (Dip. Informatica, Università di Pisa)
- Cercare di estrarre opinioni dai blog, in particolare dalla piattaforma Blogger (attualmente di Google) e WordPress, con un approccio sia di NLP che di Information Retrieval

Blog Mining

- Quando si effettuano ricerche su internet capita spesso di cercare opinioni sulla parola chiave (recensioni, soluzioni a problemi, …) e non la parola chiave stessa.
- I motori di ricerca invece mostrano i risultati basandosi sul rank della parola chiave
- L'idea è quella di aggiungere al momento del rank anche la presenza di opinioni, o sinonimi della parola chiave

Blog Mining

- La prima difficoltà è il mining dei testi, allo scopo è stata usata una collezione di feed RSS di 100.000 elementi raccolti in 3 mesi
- I feed sono stati altresì filtrati dallo spam utilizzando una blacklist di *splogs* (spam blogs)
- Sono stati presi in esame solo blog in lingua inglese

Blog Mining

- Dopo la fase di crawling dei testi e la loro indicizzazione, si è passati ad arricchire gli indici con le annotazioni
- Per ogni parola dell'indice è stato aggiunto, se necessario, il tag OPINIONATED che significa che la parola porta con sé un'opinione rilevante
- Allo scopo è stato usato un subset di SentiWordNet, di 8.427 synset tutti con un orientamento superiore a 0.4

Blog Mining

- Questo sistema permette ad esempio di eseguire query di questo tipo:

  content matches proximity 6 [OPINIONATED: 'George Bush']*

  Che significa: “trova tutti i documenti che contengono 6 parole soggettive dopo la frase ‘George Bush’”

Blog Mining

- La valutazione delle prestazioni è stata fatta su una base di 11,530 documenti:

run	topic		opinion	
	relevant	p@5	relevant	p@5
title	6150	56.80	3566	33.60
title + opinionated	4287	54.40	2500	32.80
title + description	5874	61.60	3293	36.00
title + opinionated + description	4290	69.60	2469	47.60

Blog Mining

- Il sistema ha partecipato al TREC 2006 classificandosi al 3°posto (47.60) dietro l’Università di Amsterdam (48.80), e la vincitrice l’Università di Chicago (52.60)
- Ma per l’elaborazione ha impiegato solo 6,28 sec contro le svariate ore del sistema statunitense

News Mining

- Sviluppato da Soo-Min Kim e Eduard Hovy (USC Information Sciences Institute, Marina del rey, CA)
- Pone l’attenzione all’identificazione non solo dell’opinione ma ad identificare anche gli *opinion holders* e gli *opinion topic.* La loro identificazione è fondamentale nelle news, che solitamente racchiudono diverse opinioni di diversi soggetti.

News Mining

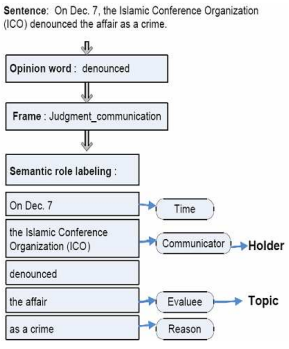
- Il lavoro è decomposto in vari task:
  - Identificare le opinioni
  - Etichettare i ruoli semantici legati alle opinioni
  - Identificare gli *holders* e i *topic* dei ruoli
  - Salvare in un DB la tripletta < opinione, holder, topic >

ØCi occuperemo solo dei primi tre, in quanto il quarto non è affatto significativo per i nostri scopi

News Mining

- Tutta la procedura è basata su FrameNet (un corpus di annotazioni semantiche assegnate manualmente)
- L'algoritmo in sostanza si preoccupa di identificare la parola *opinionata*, da questa identificarne il contesto (frame) e a partire dal frame assegnare significati alle parole o multi-word presenti nella frase

News Mining



- Ecco l'algoritmo come si comporterebbe con un titolo di una notizia

News Mining

- Come abbiamo visto in precedenza identificare una parola *opinionata*, ovvero portatrice di un'opinione, è uguale a classificarla come soggettiva (Positiva o Negativa)
- In questo caso è stato utilizzato un classificatore triplice che a partire da 1860 aggettivi e 2011 verbi ha classificato come soggettivi solo 220 verbi e 503 aggettivi

News Mining

- Per tutti i vocaboli classificati come soggettivi sono stati assegnati uno o più frame
- Un frame consiste di elementi lessicali chiamati Unità Lessicali (LU) e elementi collegati.
  - Per esempio il frame *ATTACK* contiene come LU i verbi *assail, assault, attack* e come nomi *invasion raid* e *strike*

News Mining

- Alla fine del processo sono stati assegnati 49 frame per i verbi e 43 frame per gli aggettivi.
  - Per esempio il frame *Desiring* è stato assegnato a parole come *wish, want, hope*
- E sono state collegate ai frame 8256 e 11877 frasi rispettivamente per i verbi e per gli aggettivi

News Mining

- Dopo la classificazione si passa all'annotazione dei ruoli semantici, sono state annotate solo le parole portatrici di opinioni secondo il seguente schema

Feature	Description
target word	A predicate whose meaning represents the frame (a verb or an adjective in our task)
phrase type	Syntactic type of the frame element (e.g. NP, PP)
head word	Syntactic head of the frame element phrase
parse tree path	A path between the frame element and target word in the parse tree
position	Whether the element phrase occurs *before* or *after* the target word
voice	The voice of the sentence (*active* or *passive*)
frame name	one of our opinion-related frames

News Mining

- Una volta identificati i Frame Element è molto semplice decidere gli opinion topics e gli opinion holder
- E' stato deciso manualmente che gli element taggati come come *Experiencer* sono considerati *holder* e i *Focal_participant* sono considerati *topic*

News Mining

- Un po' di risultati:

	Topic			Holder		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
V	69.1	64.0	**66.5**	81.9	75.7	**78.7**
A	67.5	73.4	**70.3**	66.2	77.9	**71.6**

Il TestSet per la valutazione è composto da 2028 frasi, 834 da frame verbali, 1194 da frame di aggettivi estratti dal set utilizzato per l'apprendimento del sistema

News Mining

- Il Testset 2 è costituito invece da news collezionate online casualmente.

Table 5. Opinion-bearing sentence identification on Testset 2. (P: precision, R: recall, F: F-score, A: Accuracy, H1: Human1, H2: Human2)

	P (%)	R (%)	F (%)	A (%)
H1	56.9	67.4	61.7	64.0
H2	43.1	57.9	49.4	55.0

Conclusioni

- L'Opinion Mining è un settore di ricerca in rapido sviluppo e di grande importanza per il NLP.
- Ma ancora una volta si evince come ogni applicazione abbia necessariamente bisogno di un corpus di riferimento
- Assumono allora importanza fondamentale i progetti che mirano a costruire corpus e KB di supporto come SentiWordNet, OntoText, FrameNet ecc…

Bibliografia

- SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining - Andrea Esuli, Fabrizio Sebastiani – 2006 – http://sentiwordnet.isti.cnr.it
- Blog Mining through Opinionated Words - Giuseppe Attardi, Maria Simi – 2006
- Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text - Soo-Min Kim, Eduard Hovy – 2005 - http://www.isi.edu/natural-language/people/hovy/papers/06ACL-WS-opin-topic-holder.pdf
- Opinion Mining – Bing Liu – Chapter 11, Lessons Slides
- Predicting the semantic orientation of adjectives - Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, pages 174–181, Madrid, ES - Vasileios Hatzivassiloglou and Kathleen R. McKeown – 1997

Bibliografia

- Using WordNet to measure semantic orientation of adjectives. In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, volume IV, pages 1115–1118, Lisbon, PT - Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten DeRijke – 2004
- Determining the sentiment of opinions. In Proceedings of COLING-04, 20th International Conference on Computational Linguistics, pages 1367–1373, Geneva, CH - Soo-Min Kim and Eduard Hovy – 2004
- A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, pages 271–278, Barcelona, ES - Bo Pang and Lillian Lee - 2004

Bibliografia

- Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of CONLL-03, 7°Conference on Natural Langu age Learning, pages 25–32, Edmonton, CA - Ellen Riloff, Janyce Wiebe, and Theresa Wilson – 2003
- Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 1(4):315–346 - Peter D. Turney and Michael L. Littman -2003
- Acquisition of subjective adjectives with limited resources. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US - Stefano Vegnaduzzo - 2004