Elaborazione del linguaggio naturale

Professore: Amedeo Cappelli

Algoritmi per agreement

Anno accademico 2005-2006

Studente: Andrea Bartolomei

1

ELN

- L'ELN si configura come punto di incontro tra l'Intelligenza Artificiale (IA) e la linguistica:
  - dall'IA eredita i metodi di valutazione, le metodologie e la vocazione applicativa della ricerca
  - costituisce un utile banco di validazione empirica delle teorie della linguistica

2

Le parole sono importanti (1)

- Nel campo dell'ELN molti problemi basilari legati all'analisi del testo sono stati risolti adeguatamente, come l'analisi morfologica, il POS tagging …
- ma la maggior parte delle applicazioni che eseguono compiti riguardanti un'analisi a livello semantico presentano ancora seri problemi

3

Le parole sono importanti (2)

- La diffusione del web e le tecniche di videoscrittura mettono a disposizione un'enorme quantità di testo
- Per poter utilizzare al meglio l'enorme quantita di informazione si rende necessaria la disambiguazione automatica di senso, cioè assegnare automaticamente un senso corretto ad ogni parola ambigua nei testi

4

Le parole sono importanti (3)

- Un buon algoritmo di disambiguazione permetterebbe di migliorare enormemente le prestazioni di vari sistemi come ad esempio i motori di ricerca di internet, l'Information Retrieval e la traduzione automatica

5

Campo semantico

- Il campo semantico è un'area del lessico i cui costituenti sono fortemente legati tra loro, infatti il significato dei termini di una lingua è determinato dalle relazioni che hanno con gli altri termini del campo a cui appartengono
- I vari campi hanno relazioni tra loro e l'insieme di tutti i campi semantici relazionati costituisce la struttura del lessico di una lingua

6

Dominio semantico (1)

- Il fulcro di interesse per affrontare il problema semantico è la nozione di dominio che costituisce una specificazione del campo semantico
- Il dominio semantico è anch'esso una struttura formata da termini fortemente relazionati, ma a differenza del campo semantico, la relazione che determina l'appartenenza di un termine a un determinato dominio è legata a considerazioni di tipo testuale

7

Dominio semantico (2)

- Il dominio semantico è il punto d'incontro tra rappresentazione del significato lessicale e quella del significato del discorso nel suo complesso
- L'appartenenza al dominio è una proprietà che può essere posseduta sia dai costituenti lessicali che dai testi

8

Rapporti fra termini

- Distinguiamo due tipi di relazioni che determinano il significato di ogni termine:
  - relazioni sintagmatiche
  - relazioni paradigmatiche

9

Relazioni sintagmatiche

- Sono relazioni tra termini dovute alla semplice concatenazione e presenza frequente di termini all'interno di uno stesso sintagma. Non si tratta però di semplici relazioni sintattiche, infatti riguardano anche la semantica

10

Relazioni paradigmatiche

- Fanno riferimento esclusivamente a domini semantici
- Sono le relazioni di sinonimia e omonimia ovvero relazioni di tipo concettuale per la cui determinazione è necessario ricorrere al giudizio dei parlanti che decidono il significato delle espressioni

11

Sinonimia (1)

- La sinonimia è il fenomeno per cui due parole differenti hanno lo stesso significato, di conseguenza:
  - è una relazione paradigmatica in quanto ciò che lega le parole è il loro significato (esprimono lo stesso concetto)
  - può essere estesa anche alle espressioni o agli enunciati

12

Sinonimia (2)

- La sinonimia può essere di due tipi:
  - sinonimia totale (o assoluta) caratterizzata dalla totale sostituibilità di termini in qualunque contesto. Si basa sul concetto di identità
  - sinonimia semplice, caratterizzata dalla sostituibilità di termini in determinati contesti (condivisione di uno dei possibili significati che ogni termine può assumere)

13

Significato delle parole (1)

- Il significato delle parole è costituito da due aspetti principali, distinti in base al tipo di informazione che veicolano:
  - Il significato cognitivo svolge la funzione comunicativa del linguaggio come mezzo di trasmissione del pensiero per descrivere uno stato di cose (fatti)
  - Il significato espressivo riguarda invece l'espressione di contenuti valutativi relativi al significato cognitivo

14

Significato delle parole (2)

- Il significato cognitivo contribuisce a determinare una condizione di verità relativa ad un'asserzione a differenza del significato espressivo che non interviene affatto nella determinazione della condizione di verità
- Esempio:
  - Quel cane ha morso qualcuno
  - Quel maledetto cane ha morso qualcuno
- Le due frasi hanno la stessa condizione di verità, perché "maledetto" ha significato espressivo

15

Concetto formale di sinonimia

- Poiché il significato espressivo non influisce sulla veridicità di una frase, la sinonimia viene intesa come identità solo a livello cognitivo, per cui due termini sono sinonimi se sono intercambiabili all'interno di una frase senza che mutino le condizioni di verità della frase stessa
- Più formalmente due termini sono sinonimi quando possono essere sostituiti vicendevolmente all'interno di una frase senza che questa cambi di significato. Quindi la sinonimia è relativa ad un particolare uso (significato) dei termini

16

Polisemia (1)

- La polisemia consiste nell'attribuzione di più significati ad un termine (o espressione) in vari contesti
- I termini, anche se polisemi, devono assumere un determinato senso all'interno di un determinato contesto. Quindi il concetto di sinonimia utilizzato è molto meno restrittivo rispetto a quello della sinonimia assoluta in cui due termini devono avere lo stesso significato in qualunque contesto

17

Polisemia (2)

- In particolare è da notare che i termini coinvolti nella sinonimia potrebbero essere tutti polisemi e quindi assumere significati differenti in funzione del contesto
- Per tale motivo il giudizio sull'identità dei significati delle espressioni contenenti termini da sostituire diventa un compito difficile e in quanto è lasciato all'intuizione del parlante madrelingua che assume un ruolo centrale di fondamentale importanza

18

Polisemia: esempio

- I termini macchina ed automobile sono sostituibili nell'enunciato:
  - in autostrada si è fuso il motore della macchina

  ma non nell'enunciato:
  - questo software gira troppo lentamente su questa macchina
- Il termine macchina è polisemo ed è sinonimo di automobile nel senso di veicolo, ma non lo è quando assume il significato di computer

19

Problema della sinonimia (1)

- La completa identità di significato è solo teorica poiché in pratica esiste sempre qualche sfumatura tale per cui il significato delle parole sia diverso
- Una sinonimia intesa come identità non dovrebbe avere assolutamente alcun effetto sulla frase, ma una sinonimia assoluta è impensabile in quanto la maggior parte dei termini sono polisemi

20

Problema della sinonimia (2)

- Una sinonimia assoluta implicherebbe totale identità di significato tra due termini rendendone uno ridondante e perciò questo non apporterebbe alcuna informazione aggiuntiva e sarebbe un inutile sovraccarico cognitivo superfluo alla comunicazione

21

Giochi linguistici

- Per definire il significato delle parole si fa uso della nozione di gioco linguistico
- Un gioco linguistico è l'insieme costituito dal linguaggio e dalle attività in cui è intessuto e permette di stabilire una relazione tra il significato delle espressioni linguistiche ed il loro uso in determinati contesti

22

Linguaggi primitivi

- Sono costituiti da poche espressioni ognuna utilizzata per funzioni ben precise
- Il significato delle espressioni è facilmente identificabile con il loro utilizzo poiché gli usi in questione sono osservabili e descrivibili
- Forniscono una prospettiva nitida per osservare fenomeni linguistici in quanto si astrae dall'ambiguità tipica delle lingue naturali

23

Giochi linguistici e linguaggi primitivi

- Un gioco linguistico non è una lingua naturale ma solo un frammento della realtà in cui il linguaggio primitivo assume un significato ben determinato e osservabile
- La struttura grammaticale delle espressioni formulate non è una componente essenziale. Esempio:
  - nel gioco linguistico dei muratori l'espressione "lastra" è equivalente a "portami una lastra"

24

Espressioni di giochi linguistici

- Se il linguaggio fosse limitato ad una serie di espressioni relative a diversi giochi linguistici, il significato sarebbe esattamente definito con il solo inserimento delle espressioni nell'opportuno gioco linguistico, ma non è cosi…

25

Mancanza di identità

- Nella semantica dei termini l'identità non è un concetto valido
- I confronti tra significati vengono effettuati in base ad analogie, per cui è opportuno parlare di somiglianze e non di identità tra significati

26

Omonimia - Polisemia

- Termini ed espressioni assumono significati differenti in base al grado di parentela tra i giochi linguistici
- La differenza tra giochi linguistici porta al caso estremo dell'omonimia in cui un termine è utilizzato per riferirsi a significati completamente differenti
- Lievi sfumature di significato rientrano invece nei casi di polisemia. Un termine polisemo utilizzato in contesti (giochi linguistici) diversi assume significati differenti che però hanno una matrice comune

27

Omonimia - Polisemia: esempio

- Un caso di omonimia è costituito dal termine "calcio", inteso come gioco del calcio o elemento chimico. In tal caso è difficile notare delle somiglianze tra i due significati
- Il termine italiano "casa" assume i significati di domicilio domestico e costruzione ad uso abitativo, che in inglese sono rappresentati dai termini lessicali home e house: i due sensi hanno ovviamente una componente comune

28

Contesto delle espressioni

- Riconoscere il significato corretto di una espressione implica perciò il riconoscimento del contesto in cui è usata
- Quindi si rende necessaria una fase preliminare di disambiguazione rispetto al contesto (gioco linguistico) in modo da rintracciare il significato corretto dei termini che la compongono
- Per distinguere il gioco linguistico è necessario procedere attraverso una rete complessa di somiglianze tra famiglie di giochi. Tale capacità di discernimento deve essere una delle competenze semantiche di un parlante

29

Uomo e IA a confronto (1)

- Nel caso dell'uomo, il contesto è dato dall'insieme delle percezioni relative al mondo e dalla sua conoscenza
- L'uomo ha una grande capacità di focalizzare l'attenzione solo sulle caratteristiche che egli ritiene rilevanti, tralasciando molte altre caratteristiche ambientali
- In tal modo riesce ad isolare delle situazioni caratteristiche rapportabili alla sua conoscenza appresa in passate esperienze

30

Uomo e IA a confronto (2)

- I sistemi di IA operano, in domini ben più ristretti rispetto a quelli dell'uomo
- Anche se dotati di una capacità discrezionale nell'azione da eseguire, hanno bisogno che una serie di configurazioni finali (obiettivi) vengano esplicitate

31

Intelligenza Artificiale: limiti

- La totalità dei sistemi realizzati nell'ambito dell'IA in generale e dell'ELN in particolare, presentano il limite della dipendenza da un dominio specifico
- Le applicazioni manifestano una "intelligenza" solo in determinati settori della conoscenza dell'uomo

32

IA: definire il contesto (1)

- Il problema è legato alla difficoltà di fornire il sistema di un comportamento flessibile, che sia in grado di giudicare l'azione opportuna nel contesto opportuno
- Un sistema non riesce a farsi guidare dal contesto, infatti è capace di eseguire delle azioni solo dopo che gli siano stati forniti degli scopi

33

IA: definire il contesto (2)

- Il sistema dovrebbe avere due tipi di competenza:
  - la capacità di riconoscere il contesto
  - la capacità di operare in modo opportuno in relazione al contesto riconosciuto

34

Contesto: ruolo fondamentale (1)

- La definizione del contesto è un problema di fondamentale importanza ai fini della risoluzione dell'ambiguità di termini ed espressioni
- I sistemi di ELN che realizzano le prestazioni migliori sono dipendenti dal dominio perché all'interno di uno stesso dominio la polisemia dei termini tende a scomparire e quindi ogni parola assume un unico significato

35

Contesto: ruolo fondamentale (2)

- La parola “calcio” può assumere il senso di:
  - elemento chimico
  - sport
- In contesti sportivi il suo significato sarà fortemente vincolato e la polisemia non sarà visibile
- Questo problema assume una grande rilevanza in molte applicazioni:
  - traduzione automatica
  - motori di ricerca (una disambiguazione delle query renderebbe la ricerca dei documenti molto più precisa)

36

Contesto: ruolo fondamentale (3)

- Un traduttore automatico dovrebbe essere in grado di determinare esattamente il senso corretto di ogni parola all'interno del testo. Questo perché due sensi diversi di una stessa parola in una lingua si traducono spesso con due parole differenti in un'altra lingua
- In un sistema di Information Retrieval che non distingua i sensi diversi di una stessa parola può succedere che una richiesta richiami documenti non voluti in quanto rilevanti per le parole della query ma in accezioni differenti

37

Contesto: ruolo fondamentale (4)

- Una disambiguazione poco accurata sarebbe causa di grossi problemi: un sistema di IR a cui fosse applicato un modulo di disambiguazione della query malfunzionante restituirebbe solamente documenti inappropriati a causa della errata disambiguazione
- Al fine di migliorare le prestazioni è necessario realizzare sistemi di disambiguazione altamente precisi per non peggiorare le prestazioni globali

38

Studio sui domini

- Uno studio sui domini è necessario per la realizzazione di un sistema dominio indipendente, cioè di un sistema che sia in grado di processare testi di qualunque tipo in modo da risolvere l'ambiguità intrinseca del lessico

39

Sistemi di disambiguazione (1)

- Le risorse dei sistemi di disambiguazione sono:
  - dizionari
  - corpora
- I corpora sono utilizzati dai sistemi per definire (learning) caratteristiche sull'uso dei termini da disambiguare

40

Sistemi di disambiguazione (2)

- L'apprendimento può essere eseguito a partire da
  - corpora annotati semanticamente (supervised): forniscono esempi d'uso di alcuni termini indicando il senso corretto in particolari contesti
  - corpora non annotati (unsupervised): collezioni di testo senza alcuna annotazione

41

Algoritmi in base ai corpora

- Algoritmi supervised: consentono di categorizzare un nuovo esempio sulla base delle categorie apprese nel learning
- Algoritmi unsupervised: creano categorie raggruppando insiemi di esempi ritenuti simili in modo del tutto statistico (clustering)

42

Metodi di valutazione (1)

- Per valutare un sistema di ELN tipicamente vengono utilizzati due parametri:
  - Precision
  - Recall
- Il loro impiego, prima limitato alla valutazione di sistemi di IR, si è esteso a molti altri sistemi in quanto hanno permesso l'incremento delle loro prestazioni

43

Metodi di valutazione (2)

- Un sistema di elaborazione del linguaggio dovrà eseguire una serie di scelte in risposta a problemi di un certo tipo
- Tali scelte potranno essere selezionate a partire da un insieme di possibili risposte (*answer space*), alcune delle quali sono ritenute corrette (*target*)
- La correttezza delle risposte è valutata mediante la formulazione di giudizi forniti da esperti

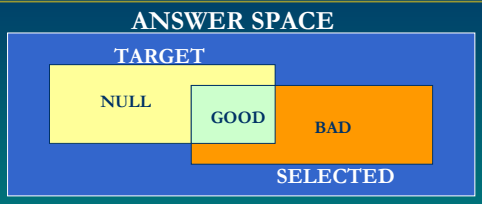
44

Metodi di valutazione (3)

- Ad esempio in un sistema di IR l'obiettivo è quello di rintracciare un insieme di documenti ritenuti pertinenti rispetto alla query formulata. In tal caso:
  - Answer space = corpus di testi
  - Target = testi ritenuti pertinenti rispetto alla query da un comitato di parlanti competenti
  - Selected = insieme di risposte del sistema
  - Good = insieme di risposte corrette fornite dal sistema (intersezione tra Target e Selected)
  - Bad = insieme di risposte sbagliate fornite dal sistema (Selected – Good)
  - Null = insieme di risposte corrette non fornite dal sistema (Target – Good)

45

Parametri fondamentali (1)



- Sulla base della cardinalità degli insiemi appena descritti si possono calcolare i due parametri fondamentali per la valutazione dei sistemi:
  - Precision = Good / Selected
  - Recall = Good / Target
- Il range di valori che possono assumere è compreso tra 0 ed 1

46

Parametri fondamentali (2)

- Precision: indica la percentuale di correttezza fornita nelle risposte
- Recall: indica la percentuale di risposte corrette fornite dal sistema rispetto al numero totale delle possibili risposte corrette (target)
- La valutazione di un sistema avviene tenendo conto di entrambi i parametri, ma il peso da attribuire ad ognuno di essi nella fase di progettazione sarà determinato dagli obiettivi per i quali il sistema è realizzato

47

Sistemi di disambiguazione

- L'input di un sistema di disambiguazione è un testo presentato come una sequenza di caratteri all'interno del quale una o più parole sono evidenziate
- Il sistema dovrà fornire in output il senso opportuno per ogni parola sulla base di un repertorio di sensi tratti dal dizionario utilizzato

48

Word Sense Disambiguation (WSD)

- La valutazione della bontà di un sistema di WSD (Word Sense Disambiguation) si esegue determinando Precision e Recall
- Answer space (per una singola parola da disambiguare) coincide con il numero dei sensi presenti nel dizionario per quella data parola
- Selected è l'insieme dei sensi (uno o più) fornito in output dal sistema
- Target è l'insieme dei sensi (uno o più) ritenuti corretti da un comitato di annotatori esperti

49

Valutazione del sistema WSD

- Good: la risposta fornita dal sistema coincide con quella fornita dagli esperti (Target = Selected)
- Bad: risposta non corretta fornita dal sistema
- Null: il sistema non fornisce alcuna risposta
- La precision e la recall di un sistema sono calcolate assegnando ad ognuno dei casi sopra esposti valore unitario
- Alcuni sistemi forniscono risposte multiple ad ognuna delle quali è assegnata una probabilità: precision e recall sono calcolate assegnando un peso proporzionale alla probabilità di ogni risposta

50

Condizione necessaria al sistema

- Per affrontare un qualunque problema, un sistema di IA necessita del requisito della Turing-esprimibililità
- Un comportamento “intelligente” non può essere simulato se non è possibile stabilire un criterio di valutazione “oggettivo” del comportamento stesso
- Perciò nel problema non può essere presente una marcata componente di discrezionalità nelle scelte eseguite dagli esperti tale da non permettere loro di trovare un accordo (agreement)

51

Gold standard (1)

- Un gold standard è un corpus annotato (supervised) utilizzato sia per l'addestramento che per la valutazione di un sistema di WSD
- Per realizzare un gold standard è necessario costruire un corpus annotato da almeno due parlanti madrelingua che, indipendentemente, eseguano le opportune distinzioni di senso

52

Gold standard (2)

- Il problema è dovuto al fatto che gli esperti parlanti madrelingua potrebbero trovarsi in disaccordo sulle scelte di senso effettuate in contesti specifici
- Il disaccordo è legato alle differenziazioni di significato molto sottili presenti sui dizionari a causa della polisemia dei termini

53

Gold standard (3)

- Quindi per realizzare un gold standard è importante la scelta di un dizionario che renda il più possibile univoca l'attribuzione di senso nelle varie occorrenze dei termini nel testo
- Ciò non è sufficiente ad evitare casi di disaccordo. In tali situazioni come è possibile eseguire il raffronto?
  - assumere l'esistenza di un esperto di livello superiore che faccia da giudice?
  - considerare una "media" tra i vari giudizi formulati?

54

Gold standard (4)

- È difficile postulare l'esistenza di un parlante più esperto di altri perché spesso i compiti sono oggettivamente ambigui o non formulati nella maniera più congeniale in contesti insufficienti alla determinazione esatta del senso opportuno per ogni parola
- Il disaccordo è fisiologico nelle questioni che riguardano il significato, perciò è necessario determinare una stima dell'accordo al fine di valutare la validità del corpus per il sistema

55

Stime dell'accordo

- Esistono diverse stime dell'accordo tra osservatori, ma le due maggiormente utilizzate sono:
  - Pa (Percentuale di agreement)
  - Statistica kappa

56

Percentuale di agreement (Pa)

- Esistono diverse formulazioni per il calcolo di Pa che tengono conto di giudizi “accoppiati” forniti dai valutatori
- Nel caso in cui i valutatori siano due:
  Pa = C / T
  C = n° di scelte concomitanti; T = totale delle scelte
- Se il numero di valutatori è maggiore di due, la percentuale è uguale alla media aritmetica delle Pa di tutte le possibili coppie di esperti

$$Pa = 2 * \Sigma C_{Pa} / [N * (N-1)]$$

C_{Pa} = Pa di una coppia di valutatori; N = n° di valutatori

57

Pa: un esempio

- Supponiamo che a due valutatori venga chiesto di assegnare uno tra due possibili sensi a 100 occorrenze di una parola con il seguente risultato:

	senso 1	senso 2	valutatore 2
senso 1	51	9	60
senso 2	4	36	40
valutatore 1	55	45	100

l’accordo osservato risulta: Pa = (51+36) / 100 = 87%

58

Statistica kappa (1)

- Ideata da Scott per il caso di due valutatori, fu poi riveduta da Cohen e Light che introdussero la possibilità di associare dei pesi alle scelte. Infine Fleiss sviluppò la generalizzazione al caso di più osservatori
- Ha sostituito la Pa fornendo una stima più accurata in quanto prende in considerazione il fatto che una parte di accordo può derivare da scelte casuali

59

Statistica kappa (2)

- Può accadere che i valutatori forniscano le stesse scelte esprimendo giudizi a caso e determinando cosi una sovrastima della Pa soprattutto nei casi in cui:
  - il numero di categorie è notevolmente ridotto
  - una delle possibili categorie di scelta occorre molto più di frequente rispetto alle altre (la probabilità di ottenere un accordo casuale su tale categoria è molto alta)

60

Statistica kappa: un esempio (1)

- Riprendiamo l'esempio precedente sulla Pa ma questa volta facciamo valutare le 100 occorrenze a due persone che non conoscono la lingua italiana. Queste decidono di assegnare i sensi lanciando una moneta. Il risultato:

	senso 1	senso 2	valutatore 2
senso 1	25	25	50
senso 2	25	25	50
valutatore 1	50	50	100

Pa = (25+25) / 100 = 50%

- È corretto affermare che concordano al 50%?

61

Statistica kappa: un esempio (2)

- Esiste chiaramente la possibilità che una quota di concordanza sia dovuta al caso
- La concordanza dovuta al caso si calcola dai marginali
- Riprendiamo il primo esempio:

	senso 1	senso 2	valutatore 2
senso 1	51	9	60
senso 2	4	36	40
valutatore 1	55	45	100

l'accordo osservato è: Pa = (51+36) / 100 = 87%

62

Statistica kappa: un esempio (3)

- I valori attesi dovuti al caso si calcolano in ogni cella:

	senso 1	senso 2	valutatore 2
senso 1	55*60/100	45*60/100	60
senso 2	55*40/100	45*40/100	40
valutatore 1	55	45	100

- Otteniamo:

	senso 1	senso 2	valutatore 2
senso 1	33	27	60
senso 2	22	18	40
valutatore 1	55	45	100

- Perciò l'accordo dovuto al caso è: (33+18) / 100 = 51%

63

Statistica kappa: un esempio (4)

- Riassumendo:
  - Accordo osservato: Pa = 87%
  - Accordo per caso: Pc = 51%
  - Accordo dovuto a concordanza vera = 87% - 51% = 36%
  - Accordo residuo non casuale = 100% - 51% = 49%
- k = concordanza vera / concordanza residua non casuale
- k = (Pa – Pc) / (1 – Pc)
- k = 36% / 49% ≈ 0,734

64

Statistica kappa generalizzata

- K = (Pa – Pc) / (1 – Pc)

  Pc = Σ [Cj / (2T)]^2

  T = totale delle scelte per il termine in esame

  Cj = n° di scelte concomitanti per il significato j, con 1≤ j ≤ n° di possibili sensi del termine in esame

- Quindi i suoi valori possono variare tra 0 e 1:
  - completo disaccordo (0)
  - completo accordo (1)

65

Valutazione dell'accordo

- Una percentuale di agreement non ragionevolmente alta sarà interpretabile come sintomo di una cattiva formulazione del compito stesso
- Infatti se un comitato di esperti differisse nel 50% delle risposte fornite, che senso avrebbe per un sistema ottenere prestazioni maggiori?

66

Limite massimo di precisione

- Una volta determinata la misura di accordo tra esperti, il sistema è valutato sulla base del confronto delle sue risposte con le informazioni contenute nel gold standard
- Consideriamo il caso in cui:
  - la percentuale di accordo sul corpus è x
  - il sistema di WSD risponde con una precision y

67

Valutazione del sistema

- Come deve essere valutato il sistema se y > x?

  Distinguiamo due casi:

  - il sistema risponde solo ad alcune domande
  - il sistema risponde a tutte le domande

68

Risposte solo ad alcune domande

- Il sistema potrebbe fornire delle risposte corrette solo nei casi in cui sia riscontrato accordo tra i valutatori accordandosi con ognuno degli esperti in misura maggiore del loro reciproco agreement
- In tal caso il sistema sarebbe in grado di selezionare solo le risposte “possibili” tralasciando quelle dubbie. Una situazione del genere mostrerebbe una cattiva formulazione del compito perché non fornirebbe risposte ad una tipologia di domande

69

Risposte a tutte le domande (1)

- Il sistema è concorde con un determinato valutatore in misura maggiore di quanto questi lo sia con un altro ipotetico esperto (la precision è una stima della percentuale di agreement tra i valutatori ed il sistema)
- Quindi il sistema si accorda maggiormente con un determinato valutatore piuttosto che un altro perché l’agreement tra valutatori non è totale

70

Risposte a tutte le domande (2)

- Il sistema possiede un modello della competenza lessicale che si approssima ad un determinato annotatore piuttosto che ad un altro
- Questo non è auspicabile se si vuole valutare una competenza linguistica generica, cioè una stima della competenza del sistema rispetto alla totalità di una lingua
- Perciò un tale sistema non ha ragione di esistere

71

Rischi nella suddivisione del corpus

- Spesso gli sviluppatori di sistemi suddividono il corpus annotato in due sottoinsiemi disgiunti per:
  - Learning supervised
  - Valutazione
- In tal caso il sistema apprende la competenza di un determinato esperto e in fase di valutazione potrebbe fornire una precisione maggiore rispetto all'agreement tra due valutatori
- Questa precisione non sarebbe rappresentativa della competenza linguistica generale del sistema

72

Limite minimo di precisione

- Se il livello massimo è delimitato dalla percentuale di agreement, un altro parametro da tenere in considerazione nella realizzazione di un sistema di WSD è la baseline, ovvero il valore di precision e recall ottenuti da un algoritmo particolarmente semplice adottato come base di partenza per la ricerca
- Esempi classici di algoritmi baseline sono:
  - Random
  - Most Frequent

73

Algoritmi baseline

- Random assegna ad ogni parola da disambiguare un senso selezionato a caso dal repertorio a disposizione
- Most Frequent assegna il senso più frequente per ogni parola, in cui la frequenza è determinata sulla base di una stima eseguita su un corpus annotato
- Ovviamente prestazioni al di sotto del livello della baseline sono un chiaro segno di una direzione sbagliata nella ricerca

74

Livello di agreement tra esperti

- Con un accordo tra gli esperti non ragionevolmente alto il sistema non sarebbe di alcuna utilità in quanto non vi sarebbe un raffronto valido con quello dei valutatori
- Infatti tale agreement limiterebbe le prestazioni del sistema e sarebbe il sintomo di una cattiva formulazione del task che sarebbe escluso dal campo di indagine dell'IA

75

Precisione richiesta

- La maggior parte delle applicazioni che trarrebbero benefici dall'avvalersi di un modulo di disambiguazione, richiedono per quest'ultimo una precisione non inferiore al 70-80%
- In caso contrario le applicazioni potrebbero avere prestazioni peggiori di quelle ottenute senza l'impiego di moduli di disambiguazione

76

Gold standard fondamentale

- La determinazione di un gold standard (e dunque di un opportuno vocabolario) è di fondamentale importanza ai fini di una effettiva risoluzione del problema della WSD
- Il gold standard deve consentire la realizzazione di sistemi che, valutati su di esso, possano raggiungere un elevato grado di precisione

77

Un caso di studio (1)

- È stato stimato l'accordo di annotatori su un corpus di sensi etichettati relativo a più di 30.000 istanze delle più frequenti occorrenze di sostantivi e verbi inglesi
- Questo corpus è l'intersezione del corpus WordNet Semcor e del corpus DSO ed è stato etichettato con i sensi di WordNet da 2 gruppi separati di valutatori
- Il corpus Semcor è un sottoinsieme del corpus Brown etichettato con i sensi di WordNet e consiste di più di 670.000 parole ottenute da 352 files di testo

78

Un caso di studio (2)

- Il corpus DSO consiste di scelte fatte sul corpus Brown e sul giornale Wall Street: un totale di circa 192.800 scelte in cui ogni occorrenza di parola è stata etichettata con un senso
- Quindi l'intersezione tra il corpus Semcor e il corpus DSO consiste di scelte fatte sul corpus Brown

79

Un caso di studio (3)

- Per ogni parola di una lista di 191 frequenti parole inglesi (121 sostantivi e 70 verbi) sono state selezionate occorrenze (al singolare o plurale oppure in forma verbale)
- Queste occorrenze, a cui è stato assegnato un senso da due gruppi indipendenti di valutatori, costituiscono l'insieme di dati con cui stimare l'accordo degli stessi valutatori

80

Definizione del matching (1)

- Per determinare la misura dell'accordo dei valutatori, il primo passo è quello di confrontare ogni scelta in Semcor con la corrispondente nel corpus DSO
- Una scelta in Semcor è considerata equivalente a quella in DSO se le scelte sono identiche o se differiscono solo per la presenza o assenza del carattere "." o del carattere "–"
- Per ogni rimanente scelta in Semcor, se il 75% o più delle parole collima con quelle nel corpus DSO, allora un match potenziale viene registrato

81

Definizione del matching (2)

- Questi match potenziali sono verificati manualmente per avere la certezza della corrispondenza
- In tal modo 13.188 coppie di scelte contenenti sostantivi e 17.127 coppie di scelte contenenti verbi hanno trovato il match con entrambi i corpora, producendo 30.315 scelte che costituiscono il corpus di intersezione usato nel presente caso di studio

82

Calcolo dell'agreement

- Un k ≥ 0,8 indica un buon agreement
- Accordo dei valutatori sul corpus intersezione:

Tipo	n° di parole	A	T	Pa	Media k
Sostantivi	121	7.676	13.188	0,582	0,300
Verbi	70	9.520	17.127	0,556	0,347
Tutti	191	17.196	30.315	0,567	0,317

  - La media k è la media aritmetica dei valori delle k individuali di ogni parola
  - La stima dell'accordo su 30.315 scelte valutata con Pa risulta essere circa il 57%
- L'accordo ottenuto dai valutatori sul corpus di intersezione non è alto (Media k = 0,317)

83

Greedy search algorithm (1)

- Sarebbe interessante scoprire come varia l'accordo in base alle classi di senso utilizzate
- L'algoritmo Greedy Search può ottenere automaticamente generiche classi di senso basate sui sensi assegnati dagli annotatori
- Ciò permette di ottenere un più alto livello di accordo mantenendo il più possibile le classi originali

84

Greedy search algorithm (2)

- loop: let C1,…,CM [le attuali M classi di senso]
  k* = – ∞
  for all i, j such that $1 \leq i < j \leq M$
    let $C'_1,\ldots,C'_{M-1}$ [M – 1 classi per unione di C_i e C_j]
    calcola $k(C'_1,\ldots,C'_{M-1})$
    if $k(C'_1,\ldots,C'_{M-1}) > k^*$ then
      $k^* = k(C'_1,\ldots,C'_{M-1})$
      $i^* = i$
      $j^* = j$
  end for
  unisce le classi di senso C_{i^*} e C_{j^*}
  M = M – 1
  if $k^* < k_{min}$ goto loop

85

Greedy search algorithm (3)

- L'algoritmo opera su un insieme di scelte effettuate dagli annotatori
- Ad ogni iterazione l'algoritmo unisce una coppia di classi di senso C_i e C_j in una sola se tale unione produce un più alto valore di k per il gruppo di classi risultante
- Tale processo è ripetuto fino a che il valore di k supera la soglia di $k_{min} = 0,8$

86

Greedy search algorithm (4)

- Nei casi in cui l'agreement tra annotatori non è sufficientemente alto, può essere usato questo algoritmo per ottenere un ridotto insieme di classi di senso al fine di migliorare l'accordo tra i valutatori
- Quindi l'algoritmo permette di ottenere un corpus con un livello di accordo sufficiente da poterlo impiegare come gold standard in sistemi di disambiguazione

87

Risultati (1)

- Per ogni parola della lista di 121 sostantivi e 70 verbi, applichiamo l'algoritmo Greedy Search ad ogni insieme di scelte nel corpus di intersezione
- Per un sottoinsieme di 95 parole (53 sostantivi e 42 verbi), l'algoritmo ha ottenuto un insieme più generico di 2 o più sensi per ognuna di queste 95 parole tale che $k \geq 0,8$
- Per le altre 96 parole, ottenendo sempre $k \geq 0,8$, l'algoritmo ha riunito tutti i sensi di ogni parola su una singola classe

88

Risultati (2)

- Risultato dei 53 sostantivi:

	n° medio di sensi	A	T	Pa	Media k
Prima	7,6	3.387	5.339	0,634	0,463
Dopo	4,0	5.033	5.339	0,943	0,862

- Risultato dei 42 verbi:

	n° medio di sensi	A	T	Pa	Media k
Prima	12,8	5.115	8.602	0,595	0,441
Dopo	5,6	8.042	8.602	0,935	0,852

89

Risultati (3)

- La tabella dei sostantivi indica che prima della fusione delle classi di senso, questi 53 sostantivi avevano una media di 7,6 sensi per sostantivo
- 5.339 scelte nel corpus intersezione si riferiscono a questi sostantivi e 3.387 di queste scelte sono state equivalenti nei due gruppi di annotatori
- La media aritmetica della kappa statistica calcolata sulle k di ogni sostantivo individuale è 0,463

90

Risultati (4)

- Dopo la fusione delle classi il numero medio di sensi per sostantivo è sceso a 4,0
- Il numero di scelte che sono state fatte sullo stesso senso generico dagli annotatori è cresciuto a 5.033
- Quindi circa il 94,3% delle scelte è stato fatto sullo stesso senso generico e la k statistica è cresciuta a 0,862 mostrando un alto accordo tra gli annotatori

91

Discussione (1)

- Nonostante il buon esito ottenuto dall'algoritmo, questo studio mostra che per un linguaggio medio è molto difficile ottenere un alto valore di accordo se agli annotatori è chiesto di assegnare sensi ben definiti come quelli di WordNet
- Nondimeno osserviamo che un linguaggio medio degli utenti permette di processare testi senza una disambiguazione di senso delle parole ad un livello "fine" tipico dei dizionari tradizionali

92

Discussione (2)

- In contrasto, gli esperti etichettano le occorrenze delle parole facendo riferimento ad esempi che mostrano l'uso di ogni senso di parola in dizionari molto completi (ad esempio HECTOR)
- Questi sono i fattori che hanno contribuito in maniera determinante alle differenze osservate

93

Discussione (3)

- Accorpamento di sensi ottenuto dall'algoritmo Greedy Search:

Sostantivo	Classe di senso generico
air	wind / gas vs aura / atmosphere
board	committee vs plank
body	physical / natural object vs group / collection
change	modification vs coins
country	nation vs region / countryside
course	class vs action vs direction
field	land vs subject
foot	human body part vs unit vs lower part / support
force	strength vs personnel
light	illumination vs perspective
matter	concern / issue vs substance
party	political party vs social gathering vs group

94

Discussione (4)

- L'algoritmo può fornire interessanti raggruppamenti di sensi di parola che corrispondono all'intuitivo giudizio di granulosità di senso
- Alcuni dei dissensi tra i due gruppi di annotatori possono essere attribuiti solamente alla raffinata distinzione di senso di WordNet

95

Discussione (5)

- Ad esempio, nel corpus intersezione c'è un totale di 111 scelte che contengono occorrenze della parola "change" a cui sono stati assegnati 7 sensi dai due gruppi di annotatori. Il risultato:
  - Pa = 0,38
  - k = -0,09 (disaccordo sistematico)
- L'algoritmo Greedy Search ha riunito cinque sensi in una classe e i due rimanenti in un'altra. Il risultato:
  - Pa = k = 1 (accordo perfetto)
- Questo mostra che esiste una distinzione troppo sfumata tra i sensi assegnati alle occorrenze delle parole

96

Discussione (6)

- Similmente alcune delle 96 parole che sono state riunite in un unico senso, hanno sensi sono troppo simili per essere considerarli distinti
- In breve, le classi di senso generico derivate da Greedy Search possono contribuire ad una migliore classificazione dei sensi per algoritmi di valutazione automatica della disambiguazione di senso delle parole

97

Conclusioni

- L'algoritmo Greedy Search costruisce classi di senso generiche basate su etichette di senso assegnate da annotatori
- Ci sono interessanti raggruppamenti di sensi di parola che corrispondono al giudizio intuitivo di granulosità di senso

98

Bibliografia (1)

- Cuppari, G. (2000). I campi semantici: la teoria di Trier. PhD thesis, Università degli studi di Messina
- Magnini, B., Strapparava, C., Pezzulo, G., and Gliozzo, A. (2002). The Role of Domain Information in Word Sense Disambiguation. Natural Language Engineering. Special issue on Senseval-2, in press
- Berruto, G. (1977). La semantica. Zanichelli
- Gale, W., Church, K., and Yarowsky, D. (1992). One Sense per Discourse. In Proc. of the ARPA Workshop on Speech and Natural Language Processing

99

Bibliografia (2)

- Véronis, J. (1988). A study of polisemy judgements and inter-annotator agreement. Technical report
- Dreyfus, H. L. and Dreyfus, S. E. (1990). Ricostruire la mente o progettare modelli del cervello? L'intelligenza artificiale torna al bivio. In Negrotti, M., editor, Capire l'artificiale. Bollati Boringhieri
- Cohen, J. (1968). Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. Psychological Bullettin, 70(4):213-220
- Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational linguistics, 22(2):249-254

100

Bibliografia (3)

- Fleiss, J. J. (1971). Measuring Nominal Scale Agreement among Many Raters. Psychological Bullettin, 76(5):378-382
- Ng, H. T., Lim, C. Y., and Foo, S. K. (1997). A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. In Proceedings of the SIGLEX workshop Tagging text with lexical semantics: why, what and how?

101