Nuovi Testi

Wiki e Blog
fonti dinamiche di testo

Caterina Lascaro
Elaborazione del Linguaggio Naturale
2005/2006

Nuovi Tipi di Testo

- Nuovi sistemi di divulgazione e soglia di pubblicazione drammaticamente *bassa* hanno dato origine a nuovi testi
  - Dinamici
  - Reattivi
  - Multilingue
  - Con numerosi autori che cooperano o perfino che avversano
  - Con controlli editoriali esigui o nulli

- I nuovi testi hanno caratteristiche di cui i testi tradizionali difettano.
- Sono interconnessi in una *rete*, resa esplicita dagli autori e dai lettori in una complessa interazione di riferimenti testuali espliciti.
- Si collocano in un contesto di altri testi molto più esplicitamente di quanto abbiano fatto precedentemente i testi.

Evoluzione del genere nel Web

- I generi già esistenti in altre forme di media sono stati convertiti in forma digitale (*giornale in notiziario elettronico*).
- Al contrario, sono sorti nuovi generi interamente dipendenti dal medium nuovo. (*homepage, search engine, webgame*)

- Quando un genere tradizionale *migra* verso il digitale, all'inizio è replicato fedelmente.
- In uno stadio successivo dell'evoluzione, si creano varianti di genere.
- Questo processo è guidato dalle capacità tecniche dei nuovi media.

5

La Credibilità dei nuovi testi

- Mentre i testi *tradizionali* sono pubblicati in forma cartacea e un numero di interventi successivi coinvolge redattori o editori per assicurare
  - accuratezza
  - attinenza
  - qualità
  - effetto del testo
- I *nuovi* testi difettano della garanzia di essere passati per molti occhi dall'autore al lettore

6

BLOG

- E' un genere evoluto *da*:
  - diari
  - logbooks / giornali di bordo
  - telecronaca,
  - rubriche ed editoriali
- *in* un misto di testi
  - dalle molte sfaccettature
  - con punti di vista e prospettive molto diversi
  - con applicazioni e ambizioni che variano da parte del creatore.

7

WIKI

- è uno spazio di lavoro condiviso da molti partecipanti
- i testi di wiki sono scritti e preparati da team aperti di autori
- Applicazioni:
  - *Wikipedia* - modellata su un genere di testo classico, quello dell'*enciclopedia*
  - gestione di progetti
  - per autori di testo

8

- È stata focalizzata l'attenzione su questo fenomeno in due importanti Workshop:
  - 11th Conference of the
    - European Chapter of the Association for Computational Linguistics, Trento, Italy Aprile 2006
  - Spring Symposium, of the
    - American Association for Artificial Intelligence, Stanford, California Marzo 2006

9

Argomenti della discussione

- Interpretare l'evoluzione del genere nel Web
- BLOG
  - Le caratteristiche linguistiche
  - Imparare a riconoscere i Blog
- Wiki
  - Descrizione
    - Analisi di scrittura digitale di Wikipedia
  - Errori in Wiki
  - Trovare frasi simili tra lingue multiple in Wikipedia
  - Cenni sull'utilizzo di Wiki
    - Costruire dizionari per il NER con Wiki

10

Interpretare l'Evoluzione del Genere nel Web

- Esplorazione dello stato corrente dell'evoluzione del *genere* nel web attraverso la *percezione degli utenti* web.
- Conferma da parte di questo studio che un numero di generi web attuali, mai visto nel mondo cartaceo, può essere *riconosciuto dai soggetti.*
- Altri generi non sono emersi del tutto e molti utenti web non conoscono le loro etichette .
- Alcune pagine di testo mostrano un alto livello di *ambiguità* per cui gli utenti web non convengono sull'assegnamento delle loro etichette.

11

Il Genere nel Web

- I generi possono essere visti come "*artefatti*", cioè oggetti culturali creati per soddisfare o semplificare le necessità comunicative.
- Questi oggetti culturali rappresentano il *ruolo* che un certo tipo di documento gioca in un ambiente.

12

Il Genere nel Web

- Ogni genere mostra delle *caratteristiche* standard o convenzionali che lo rendono riconoscibile fra gli altri, e questa identità solleva specifiche *aspettative* nei riceventi, malgrado la vaghezza delle etichette dei generi.
- I generi,però, non sono mutuamente esclusivi e generi diversi possono fondersi in un singolo documento, generando forme *ibride*.

13

Il Genere nel Web

- Da una parte, aspetti standardizzati e ricorrenti inducono ad aspettative prevedibili nei destinatari.
- Dall'altra, la libertà concessa dalla *creatività* permette ai generi di cambiare, di *evolversi*, di essere creati per soddisfare nuovi fabbisogni, specialmente sotto l'impulso di un nuovo strumento di comunicazione.

14

Lo studio del Web

- Uno studio è stato fatto da ricercatori, e basato su partecipanti volontari all'interno di varie università (in Usa, UK, Canada, Europa)
- Il *riconoscimento* e *l'accettazione* di un genere è basato su elementi come l'educazione, la cultura, la comunità e la società.

15

Pagine Web e Generi Web

- La percezione degli utenti può essere divisa in 3 gruppi.
  - Generi web facili
  - Generi web ambigui
  - Generi web difficili
- I partecipanti dovevano assegnare a 25 screenshot di pagine web 23 etichette di generi web.

16

Interpretazioni dei dati: tre metodi

- Conteggi approssimativi e percentuali
  - Tre gradi di concordanza identificati:
    1. pagine con una concordanza sopra l'80%; (*generi web stabili*)
    2. pagine con una concordanza tra il 79% e il 50% (*generi web emergenti*)
    3. pagine con una concordanza tra il 49% e il 20%. (*generi web confusi* dagli utenti)
- Test esatto di Fisher
- Residui aggiustati

17

Conteggi approssimativi e percentuali

- I partecipanti mostrano la più alta concordanza sui *"generi web facili"*, ad eccezione delle seguenti meno accettate:
  - front page
  - net ad
  - splashscreen
- Il gruppo *medio* include per la maggior parte i generi web ambigui insieme al *ezine*, ritenuto difficile da parte dell'autore.
- L'*ultimo* gruppo contiene il resto dei generi ambigui insieme alle altre pagine web difficili e tre pagine web del primo gruppo.

18

TIPOLOGIE	GENERI WEB, suddivisi dall'autore		
Generi Web Facili	Eshop, frontpage, Searchpage	personal_hp, FAQ, Net ad Splashscreen, corporate_hp	
Generi Web Ambigui	Email, sitemap, hotlist, newsletter	academic_hp aboutpage, organiz._hp, howto, tutorial	Blog, clog, searchM, onlineForm
Generi Web Difficili	Ezine, dontknow (x3)		

19

RANGES	GENERI WEB		
Top: oltre 80%	Eshop, personal_hp, search_page,	Corportate_hp, FAQ,	
Middle: tra 79% e 50%	Blog, academic_hp, online_form,	About_page, ezine, dontknow, organiz_hp,	Clog, howto, tutorial,
Bottom: Tra 49% e 20%	Email, spalshscreen sitemap,	Hotlist, dontknow(x3) newsletter	frontpage net ad

20

Test esatto di Fisher

- Le percentuali non chiarificano se l'etichette e i tipi di pagina web sono associate nella maniera suggerita.
- Il test esatto di *Fisher* può esserci d'aiuto.
- Il valore ritornato per questo test dalla SPSS è 9898.275, ed è abbastanza grande per rifiutare l'ipotesi che le etichette e le pagine web sono indipendenti

21

Residui aggiustati

- Un test statistico non indica quante e quali celle si allontanano molto da questa ipotesi.
- I *residui*, cioè le differenze tra le frequenze delle celle previste e osservate, possono esserci d'aiuto.
- Le analisi dei residui aggiustati sostengono che c'è un'associazione significativa tra le 25 pagine web e le 23 etichette

22

Conclusioni

- I tre gruppi di percezione sono venuti fuori chiaramente dalle percentuali, ma la distribuzione di pagine web nei 3 gruppi è leggermente differente da ciò che ci si aspettava.
- La visione generale dei risultati (test di Fisher) rivela che c'è un'associazione significativa tra le 25 pagine web e le 23 etichette.
- Le analisi dei residui aggiustati sostengono questa interpretazione.

23

BLOG

- I Blog *sembrano* essere principalmente
  - personali, spesso scritti da una prospettiva personale, ed esprimono le opinioni o i sentimenti dell' autore.
  - spesso mal editati e messi insieme frettolosamente in un linguaggio che ricorda le brevi note, parole sussurrate o lettere corte, piuttosto che saggi o articoli di giornali.

24

- In realtà i blog possono presentare vari tipi di linguaggio,
  - da quello più *informale*, *disinvolto* e *personalizzato*,
  - a quello *forbito* e a volte squisitamente *letterario*.
- In genere si crede che il primo tipo di linguaggio sia il più utilizzato.

25

Caratteristiche Linguistiche dei Blog – il linguaggio letterario

- Molti sondaggi hanno messo in evidenza, invece, che lo standard di questo medium sia molto più alto di quanto si creda, in quanto molti blogs puntano ad avere una *dignità letteraria*.
- Uno studio ha provato a dimostrare ciò.

26

Analisi preliminare - un campione testuale

- Il campione di riferimento è rappresentato da 10 blogs di *Splinder.com* nella lista degli "ultimi aggiornati". La selezione è stata fatta in tempi diversi in uno stesso giorno

27

Analisi di un blog meno formale del campione:

"di ritorno da…" Il post descrive un viaggio di vacanza in Scozia per la fine dell'anno e sentimenti personali. Il post mostra scelte grafiche non ortodosse:

- mancanza delle maiuscole anche per i nomi propri
- parole con accenti inappropriati

28

- uso frequente ed esagerato dei puntini di sospensione.
- Questa ultima caratteristica è molto frequente nei blogs perché da la sensazione del linguaggio parlato.

29

Analisi di un blog letterario:

"*SoleLuna*" Una donna ricorda il suo perduto amore.

Il linguaggio di questo post ha pretese letterarie, vestendosi di liricità e utilizzando costruzioni retoriche a volte verbose. Presenta, però alcune cadute di tono nelle forme grafiche come i puntini di sospensione, la "d" eufonica e così via.

30

Analisi di un blog con linguaggio intermedio:

"*Incontrista*" il testo di questo post parla della differenza fra le splinderine e le meetiche, privilegiando le prime. Il linguaggio è simil giornalistico, con una prosa brillante. Significativamente, non ci sono puntini di sospensione.

31

Nessuno di questi 3 tipi di linguaggio sembra essere il modello principale per scrivere un blog.

- Questo è dovuto alla piccola dimensione del campo di ricerca.
- Se si aumenta quest'ultimo, è possibile trovare un significativo modello di linguaggio che mediamente costituisca la base per scrivere i blogs.
- L'aumento del campo di studio è stato reso possibile parzialmente utilizzando moderni motori di ricerca.

32

Analisi quantitativa di grandi corpora usando il motore di ricerca

- Facendo una ricerca preliminare, si è visto che da un punto di vista *ortografico*, i blogs italiani sono corretti *qualitativamente* almeno come i giornali online.
- Anche altri indicatori correlati all'uso di forme italiane "*neostandard*" nel campo dei pronomi e degli aggettivi dimostrativi suggeriscono una parentela fra blogs e giornali.

33

- Secondo queste ricerche, le differenze principali tra i post dei blogs e gli articoli di giornali non erano collegati all'*accuratezza* dello scritto o alle scelte *morfologiche* diverse.
- Quindi possiamo supporre come ipotesi di lavoro che le differenze principali fra i blogs e i giornali si riferiscono, infatti, al *lessico* e alla *sintassi*.

34

- Lo status *sintattico* di molti blog è probabilmente ben rappresentato dai campioni testuali scelti precedentemente (l'uso diffuso di punti di sospensione che sono la caratteristica più cospicua)
- Tuttavia un sondaggio preciso di questo livello può essere probabilmente ottenuto solo attraverso la *codifica* di un largo corpus con etichette sintattiche.

35

Analisi quantitativa di grandi corpora usando motori di ricerca

- Le caratteristiche lessiche dei blogs possono essere studiate attraverso una semplice analisi con il motore di ricerca*.*

36

- Due corpora web sono stati, quindi, selezionati:
  - il complesso dei blog indicizzati nella versione beta del *blogsearch.google.com* .
  - il sito web del giornale *LaRepubblica* , indicizzato e interrogato attraverso l'interfaccia Google
- I due corpora sembrano avere la stessa grandezza.

37

Dobbiamo fare alcune considerazioni:

- Non analizzare forme basse di linguaggio, perché è chiaro che i giornali raramente le usano, mentre si trovano nell'uso comune dei blogs.
- Non analizzare parole assai comuni come "*questo*" e "*quello*".

38

- L'analisi comparativa si è basata sulla più alta *frequenza del linguaggio letterario*, che nella tradizione italiana ha un lessico vasto e svariato.
- Abbiamo, quindi focalizzato l'attenzione su gruppi di parole "deboli".
- Precisamente, ci siamo basati su liste di verbi "letterari" che iniziavano con la ***b***, la ***e*** e la ***v***, considerati all'infinito, prelevati dal dizionario DeMauro.
- Non abbiamo considerato quelli che presentavano omografie

39

Alcuni verbi sono stati esclusi, in quanto

- non presenti oppure
- parimenti bilanciati, o ancora
- risultanti come forme spezzettate di verbi diversi.

Dopo questa selezione, le forme rappresentate nel corpus avveniva come descritte in Tavola 2.

40

TAVOLA 2 a

FORMA	Num. in Blog	Num. in *La Repubblica*
Basire	1	0
Bastarsi	12	1
Beare	9	2
Biasmare	0	1
Biondeggiare	0	2
Biscazzare	1	0
Bruire	1	0
Bruttare	0	1
Bugiare	1	0
Elicere	0	1
Ergere	24	9
Esondare	6	4

41

TAVOLA 2 b

FORMA	Num. in Blog	Num. in *La Repubblica*
Esperire	56	21
Esplicare	79	15
Estimare	2	0
Evoluire	2	0
Vacare	4	0
Vagolare	7	1
Vanire	1	0
Vaticinare	17	6
Ventare	2	0
Vigoreggiare	2	0
Villaneggiare	1	0
Volvere	2	0
Volversi	0	1

Totale delle occorrenze

Blog: 230

LaRepubblica: 65

42

Conclusione

Le analisi preliminari dei blog italiani sembrano confutare la semplice equivalenza "**Blog = testo informale**".

43

- Chiaramente sia *mezzi statistici* che un *software speciale* da monitoraggio, sono necessari per dare a questo tipo di ricerca più focus e più profondità.
- Questa ricerca può essere migliorata con
  - una migliore copertura dei motori di ricerca
  - altri *indicatori di ricerca della qualità* linguistica di un testo.

44

Imparare a riconoscere i Blogs

- Esperimenti con l'applicazione del *machine learning* su una classificazione binaria dei blog,
  - cioè determinare se una data pagina web è una pagina di un blog.
- *Qual è la performance degli algoritmi della machine learning di base su questo compito?*
- *Può la performance di questi metodi essere migliorata usando metodi di ricampionamento come Bootstrapping e Co-Training?*

45

Applicazione del machine learning

- Uso di un piccolo *dataset* annotato manualmente e di una grande varietà di algoritmi.
- Selezione di una gran quantità di attributi caratterizzanti, tra cui
  - numero di post
  - lunghezza media/max/min dei post
  - i vari host

46

Prima questione

- È stata istruita una vasta gamma di *learners* ,usando i dati manualmente notati e testati usando 10 volte una convalida incrociata.
- Comparazione dei risultati alla baseline.
- Il miglior algoritmo è risultato SMO basato su un vettore di supporto. (94,75%)

47

Ricampionamento

- Il dataset precedente è suddiviso in due dataset
  - Dataset di training (100 casi)
  - Dataset di test (101 casi)
- Un nuovo dataset è stato creato con un crawler, filtrato e poi suddiviso in sottoset di 1000 casi, per ogni iterazione.

48

Bootstrapping

- ***Inizializzazione***: con il set di training (100 casi annotati manualmente) predire l'etichette o nomi dei primi subset di 1000 casi non etichettati.
- ***Iterazioni***: etichettare i casi non etichettati secondo la previsione dell'*algoritmo* e aggiungere questi casi al precedente set di training per formare un nuovo set di training. Costruire un nuovo modello basato sul nuovo set di training.
- Ogni iterazione viene testata con il *set di test*.

49

Risultato del Bootstrapping

- Dopo 36 iterazioni, era chiaro che si era giunti al massimo delle prestazioni.
- Nonostante il campione così grande, il bootstrapping non migliora la performance del machine learning, probabilmente a causa degli outliers, mancati a causa della particolarità del blog (MSN Space…).

50

Co-Training

- L'obiettivo è di prendere le predizioni unanimi dai migliori 3 algoritmi del machine learning e usarle per fare il *Bootstrap* del set di training.
  - *SMO* con vettori di supporto,
  - *J48* (con albero di decisione e un'implementazione C4,5)
  - *Jrip* (basato sulle regole).
- Poi, verrà testato se offre un miglioramento sull'algoritmo SMO.

51

- Il procedimento ha inizio con il set di *training* e, con l'aiuto delle predizioni dei 3 algoritmi, si etichettano vari casi, aggiunti poi al set di training.
- Usando quest'ultimo, vengono etichettati i casi del subset di 1000 casi.
- Ancora una volta quei casi venivano aggiunti al training set e così via per quante più iterazioni possibili.
- Dopo ogni iterazione, viene testato il set con l'algoritmo SMO.

52

Risultato del Co-Training

- Dopo 30 iterazioni, l'esperimento fu terminato, in quanto era finita la memoria.
- Anche se ad ogni iterazione, la percentuale non è sempre migliore di quella del Bootstrapping, questo classificatore è più accurato

53

WIKI

- I Wiki tendono
  - ad avere alte ambizioni riguardo la correttezza dei fatti, persistenza, qualità editoriale e affidabilità.
- E dove non riescono gli autori, i vari Wiki sono a cura di altri autori e lettori.
- Wikipedia può essere considerato come esempio dell'*evoluzione del genere tradizionale* già esistente, evoluzione preservata nelle forme superficiali degli articoli, ma non nei processi di scrittura e lettura

54

Un'analisi di scrittura digitale di Wikipedia

- L'analisi del contrasto linguistico fra Wikipedia - nuovo tipo di testo - e l'Enciclopedia britannica online.
- *Enciclopedia Britannica* – la più grande enciclopedia del mondo di lingua inglese.
- Contiene oltre 120.000 articoli, scritti in maniera accurata e affidabili

55

- *Wikipedia* – uno dei siti più popolari, progettato con lo scopo di creare un'enciclopedia *gratuita*, contenente informazioni su *tutti* gli argomenti, scritti da volontari *cooperanti* tra di loro.
- Consiste di 200 edizioni indipendenti di lingue diverse, tra cui quella inglese è la più fornita.

56

Wiki – un nuovo genere testuale

- I wiki sono considerati come nuove *peculiarità* aggiuntive agli attuali strumenti sincroni e asincroni della prima generazione CMC (***C****omputer* ***M****ediated* ***C****ommunity*).
- Contrariamente agli altri siti, Wikipedia invita alla scrittura di articoli, usando i link wiki e creando così una rete di pagine interconnesse.

57

Processo di interlink

- Nel redigere un articolo, l'autore può collegare una o più parole (WikiWord) ad un altro articolo, racchiudendole tra parentesi quadre.
- L'interlink è automatico e semplice.
- L'autore crea percorsi differenti per il lettore, anche se non vi è un ordine predefinito di pagine da seguire.

58

- I testi *tradizionali* creano una netta *separazione* tra lo scrittore e il lettore.
- La *tecnologia Wiki* media questo divario, in quanto i due attori hanno ruoli interscambiabili.
  - Difatti anche il lettore può apportare modifiche, commenti o creare nuovi articoli
- In questo modo la conoscenza diventa *dinamica* e *contestualizzata*.

59

Due modi di scrittura

Modalità Documento

- I contributori creano i documenti in collaborazione e possono lasciare aggiunte
- I documenti sono
  - espositivi, estesi e rifiniti
  - formali e anonimi
  - scritti a guisa di monologo e in terza persona

60

Due modi di scrittura

Modalità Thread (collaborativo)

- I contributori portano avanti discussioni "postando" messaggi firmati nelle pagine connesse all'articolo principale
- I thread sono
  - esplorativi, aperti e collettivi
  - dinamici e informali
  - scritti a guisa di dialogo e in prima persona.

61

Due tipi di conoscenze

- La *Modalità Documento* dimostra che
  - La conoscenza è collettiva
  - Le idee, *non gli scrittori*, sono il focus principale
- La *Modalità Thread* dimostra che
  - La conoscenza è il risultato della collaborazione costruttiva, e *non* una produzione *solitaria*.

62

Obiettivi della ricerca

- Investigare gli articoli di Wikipedia e analizzare il *WikiLanguage*
  - linguaggio formale, neutro e impersonale usato negli articoli ufficiali enciclopedici
- analizzare il *WikiSpeak,* considerando Wikipedia come CMC
  - linguaggio parlato-scritto dagli utenti di Wikipedia nelle loro comunità informali (dietro le quinte)

63

Primo Obiettivo
Wiki vs. Britannica

- Analisi comparativa di un campione di articoli, scelti a caso.
- Il campione include file testo di articoli (di Wiki e della Britannica) su argomenti, presi da 8 categorie di Wikipedia
  - cultura, geografia, storia, vita, matematica, scienze, società, tecnologia

64

- Un programma ha analizzato il campione e sono stati utilizzati dei *fattori* con cui misurare la formalità di Wikipedia:

1. Lunghezza dell'articolo (totale delle parole), in quanto la concisione è una caratteristica del discorso formale scritto.
2. Lunghezza media delle parole (in lettere), in quanto le parole corte sono una caratteristica del genere informale.

65

3. Un alto livello di densità lessicale è peculiare di scritti formali accademici.
4. Numero di elementi lessicali unici.
5. Frequenza dei pronomi impersonali e dei suffissi (come -age, -ment,- ance/ence, -ion), tipici del genere *formale* .
6. frequenza di abbreviazioni, acronimi, contrazioni e pronomi personali, tipici del genere *informale*, come faccia-a-faccia e conversazione telefonica

66

I Primi Risultati

- Gli articoli in Britannica sono più corti e presentano una densità lessicale superiore
- Nonostante il livello di formalità totale sia superiore in Britannica, la frequenza di parole formali e pronomi impersonali e la lunghezza media delle parole è similare in entrambe.

67

I Primi Risultati

- Di conseguenza, la differenza risiede nella densità lessicale.
- C'è da considerare, però, che il numero di parole medio nelle due enciclopedie è notevole e che i termini lessicali differenti sembrano avere una maggiore *evidenza* nei testi più corti.

68

- Si può dedurre che grazie al controllo editoriale collettivo, il WikiLanguage mostra uno stile formale e standardizzato simile a quello trovato nella Britannica.

Densità Lessicale Media		Nomi Formali + P.P.		Lunghezza media articoli		Lunghezza media parole		Formalità totale %	
B	**W**	**B**	**W**	**B**	**W**	**B**	**W**	**B**	**W**
44,9	31,4	5,3	5,2	1728	3510	5,3	5,2	50,2	36,6

69

WikiSpeak

- WikiSpeak è un linguaggio non ufficiale.
- La sua peculiarità è l'immediata evidenza nei *WikiLogismi* (es. stub, NPV, wikify, backlogs, FAQ, village pump, ecc) , considerati per la propria densità lessicale, una suprema sintesi del WikiSpeak.

70

WikiSpeak

- È stata fatta un'analisi per misurare l'impatto
  - del contenuto,
  - della forma
  - delle funzionalità sul lettore
- e il WikiSpeak usato nelle discussioni connesse agli articoli.
- Da ciò sono emersi una gran quantità di nuovi termini.

71

- WikiSpeak è ricco di
  - acronimi:
    - NPOV - neutral point of view
    - COTW - collaboration of the week
    - IFD - image for deletion
    - WDYS - what did you say?
    - CIO - check it out (controllalo)
  - abbreviazioni:
    - pls - please
    - bb ppls - bye bye peoples
    - b4n - bye for now
    - cyl - see you later

72

- fusione di termini:
  - infobox
  - quickpoll
  - Namespace
- traslazione di significati:
  - orphan
  - mirror
  - stub
- nuova grafologia:
  - il lower-case è di default
  - la capitalizzazione marca il concetto
  - BiCaps o CamelCase

73

- Plurale inglese:
  - sostituzione della -s con la -z
- Punteggiatura a volte assente
- Ripetizione di vocali e consonanti, punteggiatura e simboli
  - Yayyyyyyyy
  - WHAT???????
  - # (…) (---) (***)

74

Conclusioni

- Wikipedia, come una nuova espressione per un genere enciclopedico, appare simile alle *tradizionali* enciclopedie grazie alla omogeneità stilistica, al punto di vista neutrale e allo stile formale.
- La collaborazione degli utenti rispetta le norme stilistiche, l'etica del lavoro sociale condiviso, quindi *diversità* e *controversie* sono cancellate.
- Quindi le voci, anche se individuali, originarie della comunità CMC, sono fuse e omogeneizzate nella neutralità e formalità degli articoli.

75

Errori nei Wiki

- affidabilità offerta dai wiki
- tecniche dell'elaborazione di un linguaggio per aiutare a decidere *se* affidarsi ad un testo particolare.

76

Wikis e il problema della fiducia

- Un articolo in Wikipedia può essere
  - ben scritto,
  - apparire autorevole,
- il lettore è inclinato a *credervi*.
- Il lettore non sa che alcune aggiunte di altri autori sono *incorrette* !
- I wikis sono universalmente utilizzati, quindi il potenziale di disinformazione tende a crescere.

77

Esempio: storia di uomini politici

- In genere,gli articoli su determinati uomini politici vengono scritti o rivisti da personale appartenente al loro staff.
- Questi articoli tendono a fare apparire il personaggio politico sotto una luce più favorevole.
- L'informazione è, evidentemente, manipolata.

78

Le correzioni

- Si possono sempre fare, ma ci sono dei problemi:
  - È necessario un determinato tempo.
  - Gli autori, soprattutto di articoli vecchi, possono non essere interessati.
  - Lo staff preposto alla correzione non è sufficientemente numerico.

79

- Questo problema della disinformazione è meno rilevante nei siti web *non-wiki*, anche se gli argomenti sono trattati da autori sconosciuti, in quanto è il *dominio* stesso ad offrire la garanzia di affidabilità.

80

- Il problema non consiste nel numero di errori nel wiki, ma come il lettore possa decidere se l'articolo sia affidabile e quindi utilizzarlo come guida.
- C'è bisogno di strumenti automatici per fornire aiuto ai:
  - lettori, per valutare l'affidabilità
  - autori e moderatori per analizzare le modifiche
  - moderatori, inoltre, per identificare vandalismi, diffamazioni e propaganda

81

Imparare l'affidabilità

- La storia dei cambiamenti su wiki può essere utile a categorizzare gli utenti come *affidabili* o *sconosciuti*, purché noi abbiamo a disposizione criteri indipendenti.
- Infatti l'impiego del criterio basato sulle tecniche stilistiche, anche se rilevante, non è attuabile, in quanto dobbiamo categorizzare piccoli frammenti di testo.

82

Una ipotesi

Una ipotesi di lavoro è dividere il testo in *fattoidi*, la cui identificazione è davvero difficile, ma è probabile che gli stessi cambiamenti del testo possano aiutarci in questo senso.

83

Punto della situazione:

- *possiamo automaticamente classificare i contributori wiki come affidabili o inaffidabili?*
- *i cambiamenti degli utenti affidabili forniscono buoni dati di preparazione?*
- *ci sono delle caratteristiche in frammenti di testo che permettono una classificazione di affidabilità?*
- *quali mezzi possono essere adattati da altre aree dell'elaborazione del linguaggio per affrontare queste problematiche?*

84

Una ontologia di errori?

- Vi porto ad esempio un errore rilevato nel Wikipedia inglese riguardo all'università di Cambridge:
  - *L'ammissione ai colleges di Cambridge di prima laurea dipendevano, una volta, dalla conoscenza del latino e del greco antico, materie insegnate principalmente nel UK nelle scuole a pagamento, dette "scuole pubbliche"*

85

- Questo frammento è chiaramente sbagliato, in quanto è generalmente risaputo che le scuole pubbliche, nel senso stretto della parola, sono nel UK una assai piccola proporzione, rispetto a quelle a pagamento.
- L'equiparazione dei due tipi di scuola è un errore.

86

Ipotesi

- Ammettiamo che un editor affidabile corregga questo errore.
- Ammettiamo che possiamo analizzare e immagazzinare la correzione automaticamente.
- Di conseguenza sarà possibile controllare e correggere lo stesso errore in altri testi.

87

Frasi simili attraverso linguaggio multipli

- *Il corpus Wikipedia è adattabile ad un'analisi multilingue che tende a generare corpora paralleli?*
- Le motivazioni:
  - corpora allineati alle frasi hanno un ruolo importante nei metodi d'elaborazione di linguaggi basati su corpus
  - forniscono una conoscenza profonda in Wiki come fonte di conoscenza
  - Sono un utile tipo di supporto di modifica.

88

Due approcci

1. sistema MT per ottenere una traduzione approssimativa di una data pagina da una lingua in un'altra + sovrapposizione delle parole delle frasi.
2. lessico bilingue che è generato da Wikipedia usando la struttura dei link e i titoli delle pagine collegate.

89

Approccio basato su MT

- Traduciamo la pagina wikipedia olandese in inglese, usando Babelfish di Altavista.
- Colleghiamo ogni brano di testo o frase in inglese ad ogni brano o frase in olandese.
- Calcoliamo un punteggio semplice di sovrapposizione di parole per ciascuna coppia
- Abbiamo usato la misura di similarità di *Jacard*

90

Approccio basato su MT

- Supponendo una corrispondenza 1:1, filtriamo la lista generata di brani e di frasi, ordinata in ordine decrescente di similarità:
  - Per ogni coppia della lista, si eliminano tutte le occorrenze di frasi della coppia selezionata, da tutte le altre coppie.

91

Usare un lessico bilingue

- Algoritmo usato:
  - Generare un lessico bilingue
  - Dato un argomento, prendere le pagine corrispondenti dall'inglese e dall'olandese di Wikipedia
  - Dividere le pagine in frasi e arricchire gli hyperlinks nella frase o identificare le entità con nome nelle pagine
  - Rappresentare le frasi in queste pagine usando il lessico bilingue.
  - Computare la sovrapposizione dei termini tra le frasi pertanto rappresentate.

92

Un lessico bilingue

- Per ogni pagina di Wikipedia in una lingua, le traduzioni del titolo in altre lingue, per cui ci sono entry separate, sono date come hyperlinks.
- La maggior parte di questi titoli sono frasi che contengono il nome del contenuto e sono molto utili nel computo di similarità multilingue.

93

Rappresentazione canonica di una frase

- Rappresentare le frasi in ambedue le coppie di lingue usando questo lessico.
- Ciascuna frase è rappresentata da un set di hyperlinks che essa contiene. Noi cerchiamo ciascun hyperlink nel lessico bilingue.

94

Arricchire la struttura a link

- In Wikipedia non tutte le occorrenze di entità che hanno entry in wikipedia sono in realtà anchor text di un link di ipertesto.
- Per evitare ciò, vengono identificati altri hyperlinks automaticamente, cercando nel lessico gruppi di parole (da 4 fino a 1) contenuti nell'articolo.

95

Identificare frasi simili

- Il passo finale coinvolge il computo della sovrapposizione dei termini tra le coppie di frasi e il filtraggio della lista risultante.
- I passi rimanenti sono simili a quelli descritti nell'approccio basato sul MT.

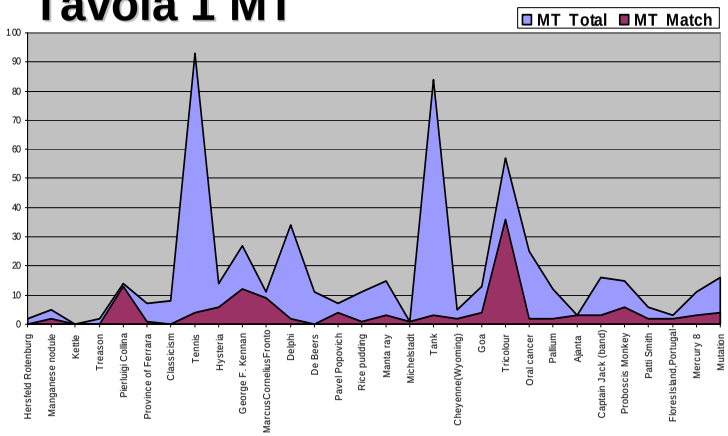
96

Conclusione

- L'approccio del lessico bilingue restituisce meno coppie incorrette dell'approccio basato su MT.
- Noi interpretiamo questo dicendo che il metodo basato sul lessico bilingue fornisce una rappresentazione più accurata sui contenuti delle frasi in Wikipedia rispetto all'approccio basato su MT.

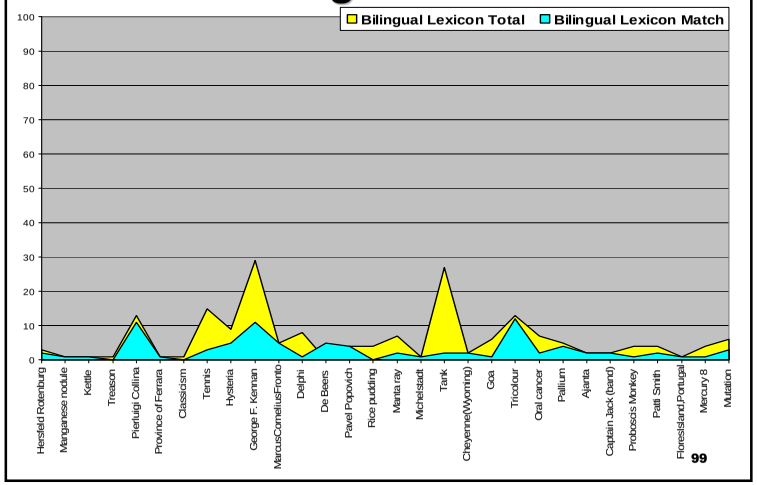
97

Tavola 1 MT



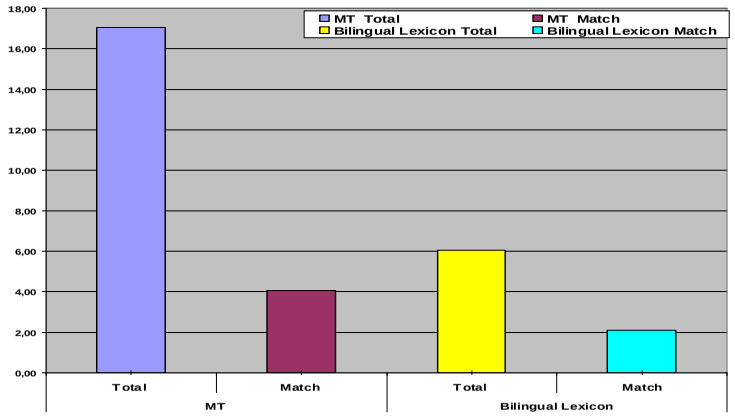
98

Tavola 1 Bilingual Lexicon



99

Tavola 2 Medie dei Risultati



100

Dizionari per il NER con Wiki

- Wikipedia è stata scelta come risorsa linguistica per la creazione e la manutenzione automatica di dizionari per il ***N**amed **E**ntity **R**ecognition* (NER), perché:
  - è una grande fonte di informazioni
  - ha la licenza gratuita
  - ha dati formali e strutturati
  - è multilingue
  - è continuamente aggiornato.

101

Bibliografia

- Tutte le informazioni sono reperibili al sito: http://www.sics.se/jussi/newtext/
- Tutti gli studi sono approfonditi e più particolareggiati nei file pdf, linkati al sito indicato.

102

References

- Ann Copestake: *Errors in Wiki*
- Mirko Tavosanis: *Linguistic features of Italian Blogs: literary language*
- Antonella Elia: *An Analysis of Wikipedia digital writing*
- Erik Elgersma, Maarten de Rijke: *Learning to recognize blogs: a preliminary exploration*
- Marina Santini: *Interpreting genre evolution on the Web*
- Sisay Fissaha Adafre, Maarten de Rijke: *Finding Similar Sentences across Multiple Languages in Wikipedia*

103