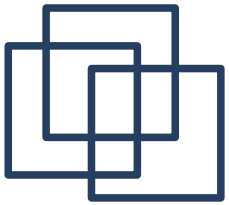
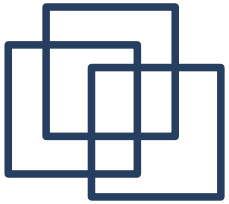


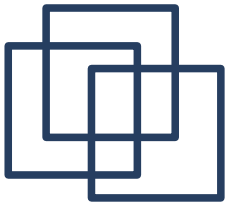
Summarzation

Elaborazione del Linguaggio Naturale
A.A. 2005 – 2006
Università di Pisa

di Alessandro Arzilli

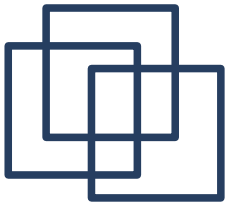


Indice



1. Introduzione

[2][3][6]

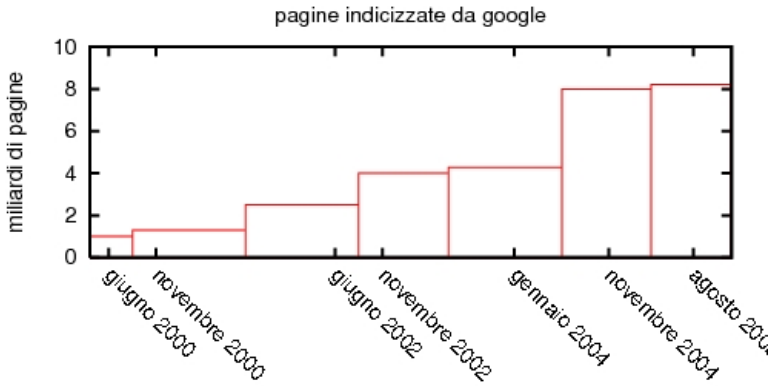


cos'è?

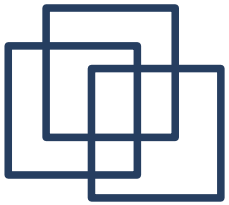
- Trasformare uno o più testi in un testo più breve che li riassuma
- Due scenari:
  - Query-driven:
    L'utente è alla ricerca di particolari informazioni che dovrebbero essere contenute nel testo, il riassunto verrà generato usando una query specificata dall'utente
  - Text-driven:
    L'utente è interessato a sapere di cosa parla un testo, il riassunto verrà generato usando criteri generali di importanza

il problema (1)

- Sovraccarico di informazioni
  - 550 miliardi di pagine nel web (stima)
    - 11 miliardi, circa, indicizzate dai motori di ricerca [4][5]

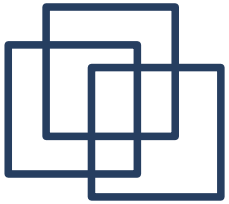


  - ... a cui vanno aggiunte le informazioni prodotte dagli altri media
- Come trovare le informazioni che ci servono?
- Come capire se un documento è interessante?



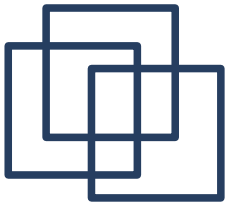
il problema (2)

- Dal punto di vista dell'intelligenza artificiale:
  - Riuscire a comprendere testi in linguaggio naturale automaticamente è un obiettivo centrale dell'IA
  - Il problema del text summarization, in linea di principio, richiede:
    - estrazione della conoscenza contenuta in un testo
    - la trasformazione di questa conoscenza in un nuovo testo (più breve)
  - Quindi, risolvere in questo modo il problema della text summarization è importante per l'IA in generale
- Inoltre:
  - idealmente un sistema per la text summarization (implementato come sopra) sarebbe quasi identico ad uno per la traduzione automatica



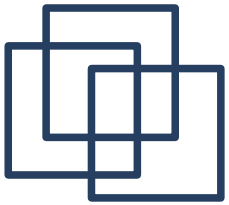
il problema (3)

- Traduzione Automatica:
  - il sistema per text summarization *ideale descritto nella slide precedente sarebbe quasi identico ad un sistema per traduzioni automatiche*
- *Sfortunatamente l'implementazione ideale è attualmente fuori portata*
  - *Si spera che diventi fattibile implementando soluzioni pragmatiche al problema sempre più sofisticate*



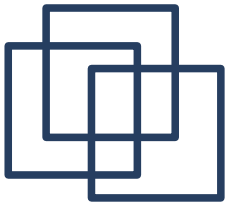
il problema (4)

- Riassunti prodotti “manualmente” vengono già usati tutti i giorni
  - abstract degli articoli tecnici, guide tv, news...
- Si vorrebbe avere la possibilità di produrre automaticamente riassunti:
  - Su qualunque insieme di documenti scelto dall'utente
  - Con dei criteri di rilevanza delle informazioni specificati dall'utente



come

- Idealmente:
  - estrazione della conoscenza contenuta in un testo
  - la trasformazione di questa conoscenza in un nuovo testo (più breve)
- Attualmente:
  1. Estrazione di passaggi “rilevanti” dal testo sorgente
  2. Fusione dei passaggi estratti in un unico testo
- Oppure:
  - Riempimento di un modulo predefinito con informazioni estratte dal testo



2. Valutazione

[2][3][6]

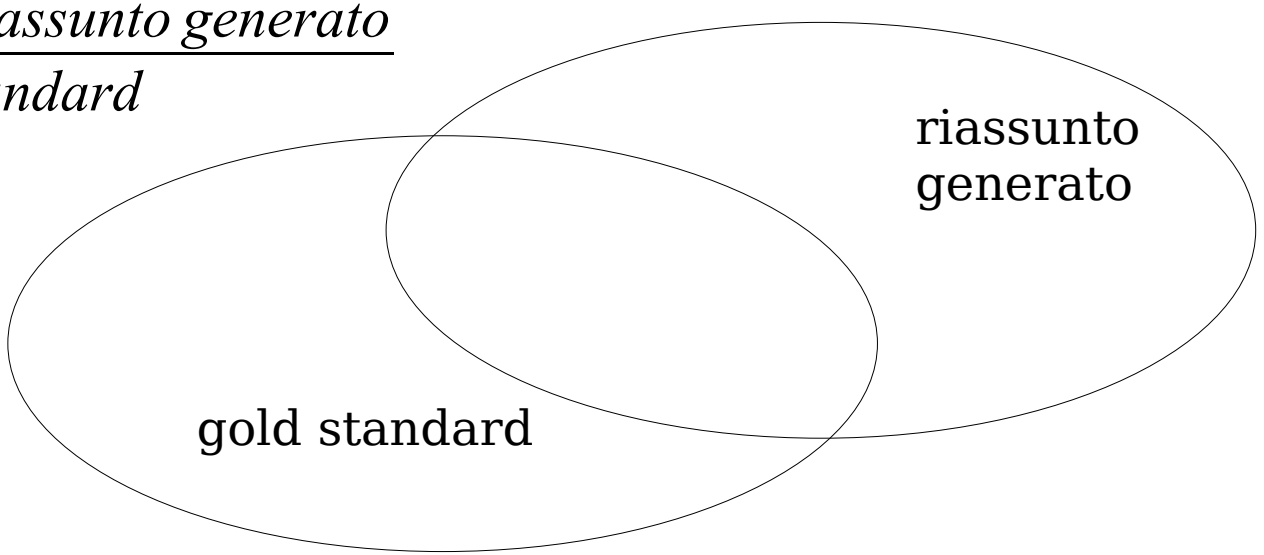
con gold standard

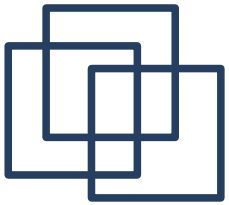
- Confronto il riassunto generato con il gold standard e misuro:
  - Precision

$$\frac{gold\ standard \cap riassunto\ generato}{riassunto\ generato}$$

  - Recall

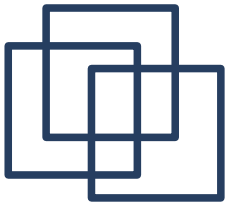
$$\frac{gold\ standard \cap riassunto\ generato}{gold\ standard}$$





con gold standard (2)

- Nel contesto dei sistemi di text summarization che riempiono dei moduli:
  - Precision:
    Numero di campi riempiti correttamente
  - Recall:
    Numero di campi riempiti (sul numero totale di campi)



senza gold standard

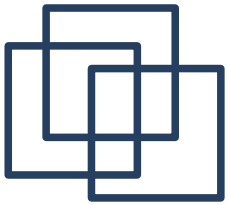
- Si confronta il riassunto generato con il testo sorgente
  - Compression Ratio:

$$\frac{lunghezza\ riassunto}{lunghezza\ sorgente}$$

  - Retention Ratio:

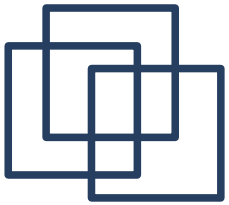
$$\frac{informazioni\ nel\ riassunto}{informazioni\ nella\ sorgente}$$

- La CR è (abbastanza) oggettiva. La RR è più difficile da misurare:
  - Q&A games...



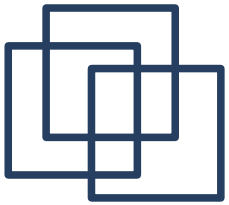
3. Metodi semplici

[2][3][6][7]



titoli

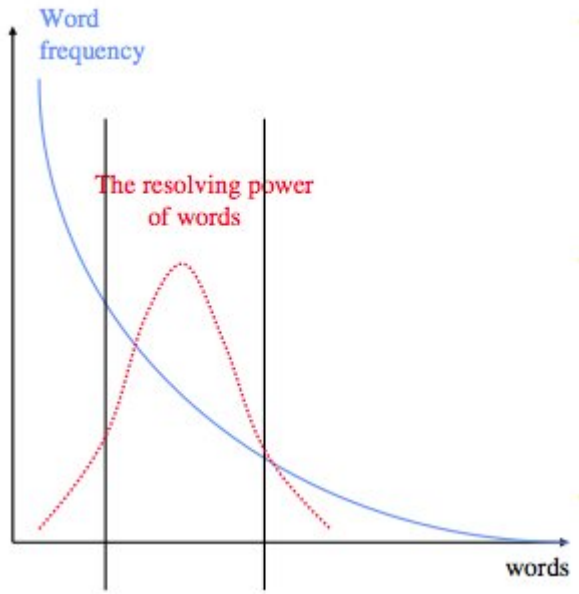
- Ipotesi:
  - L'autore sceglie dei titoli che descrivano il contenuto del testo
- Metodo:
  - Assegnare un punteggio positivo in base a quante parole del titolo contiene ogni frase del testo
  - Estrarre le frasi con il punteggio più alto



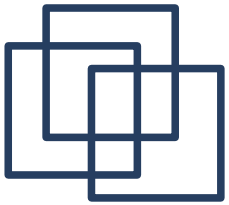
posizione

- Ipotesi
  - Le frasi importanti sono all'inizio o alla fine del testo
- Lead Method
  - Si prende la prima frase del testo (semplicemente)
  - La sua efficacia sembra diminuita col passare del tempo:
    - 52% recall & precision nel '68, 33% nel '95 (perché?)
- Quando è utilizzato assieme ad altri metodi:
  - si da un punteggio positivo a frasi che stanno all'inizio e alla fine del testo e dei paragrafi

frequenza delle parole

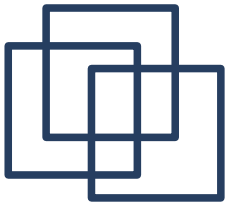


- Ipotesi:
  - Le parole che vengono utilizzate frequentemente (nel testo sorgente) sono importanti
    - Eliminando le parole con frequenza troppo alta (che saranno congiunzioni, articoli, etc)
- Metodo:
  - Calcolare la frequenza delle parole nel documento
  - Assegnare un punteggio alto alle frasi che contengono parole con frequenza compresa tra i due threshold
  - Estrarre le frasi con il punteggio più alto



cue phrases

- Ipotesi:
  - Le frasi importanti sono introdotte da locuzioni “bonus”, del tipo: “In questo articolo...”, “in conclusione...”, “significativamente...”
- Metodo:
  - Costruire un dizionario di bonus phrases e uno di stigma phrases (locuzioni che raramente compaiono in frasi importanti)
    - tramite criteri statistici e linguistici
  - Assegnare ad ogni frase del testo sorgente un punteggio basato sul numero di bonus phrases e stigma phrases che contiene
  - Estrarre le frasi con il maggior punteggio

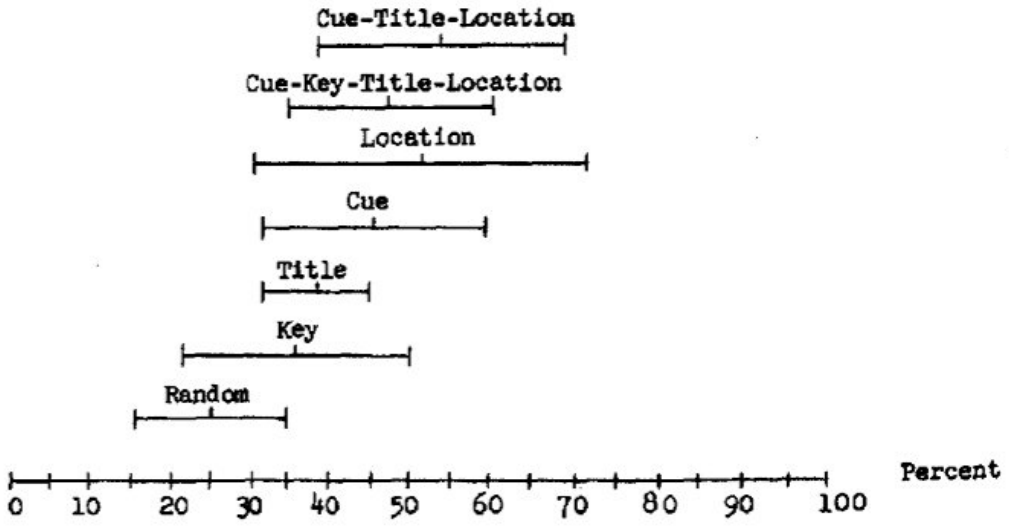


Edmundson '69

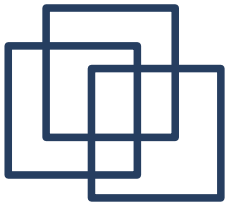
- Metodo che combina i metodi visti fino qui
- Queste differenze:
  - Il metodo delle frequenze viene chiamato Key
  - Nelle cue phrases vengono utilizzate soltanto locuzioni composte da una sola parola
  - La frequenza viene calcolata solo per le parole non comprese tra le cue phrases
  - Il punteggio finale di una frase è calcolato come:

$$a_1 C + a_2 K + a_3 T + a_4 L$$

Edmundson '69 (2)

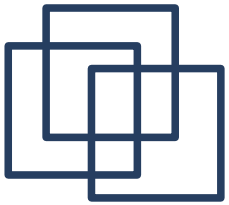


- Recall o Precision?
- Il metodo Key (delle frequenze) è stato eliminato in base ai risultati sperimentali, ottenendo un miglioramento delle prestazioni



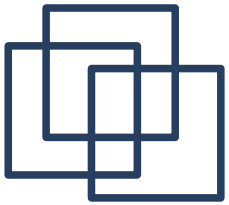
optimum position policy (1) [8]

- Ipotesi
  - Le frasi importanti si trovano in posizioni particolari all'interno del testo
  - Queste posizioni possono essere imparate automaticamente grazie ad un training set di testi annotati
- Apprendimento
  - Il training set consiste in testi accompagnati da un abstract e una lista di keyword
  - Ogni frase viene annotata con la coppia:
    (no. paragrafo, no. frase)
  - Ad ogni frase viene assegnato un punteggio in base alla similarita con l'abstract e la lista di keyword

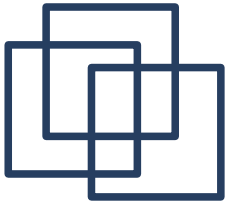


optimum position policy (2)

- Apprendimento (continua)
  - In questo modo si ottiene una lista di coppie (no. paragrafo, no. frase) ordinate per importanza
- Metodo
  - Essenzialmente uguale al metodo della posizione:
  - Si utilizza la lista di coppie generata dall'apprendimento per estrarre le frasi più importanti o per assegnare i punteggi
- Valutazione:
  - Recall 35% Precision 38%
  - Estratti lunghi il 10% del testo sorgente coprono il 91% delle parole importanti

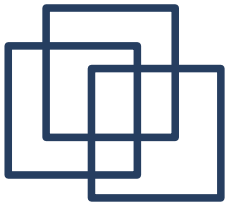


4. Metodi più complessi



coerenza del testo

- I testi non sono sequenze di frasi scollegate tra loro
- esistono relazioni tra frasi e gruppi di frasi
  - inserite deliberatamente dall'autore
  - al fine di esporre un argomento
- Individuare la struttura delle relazioni tra queste frasi permetterebbe di riconoscere le frasi più importanti
  - e quindi di estrarle per creare un riassunto



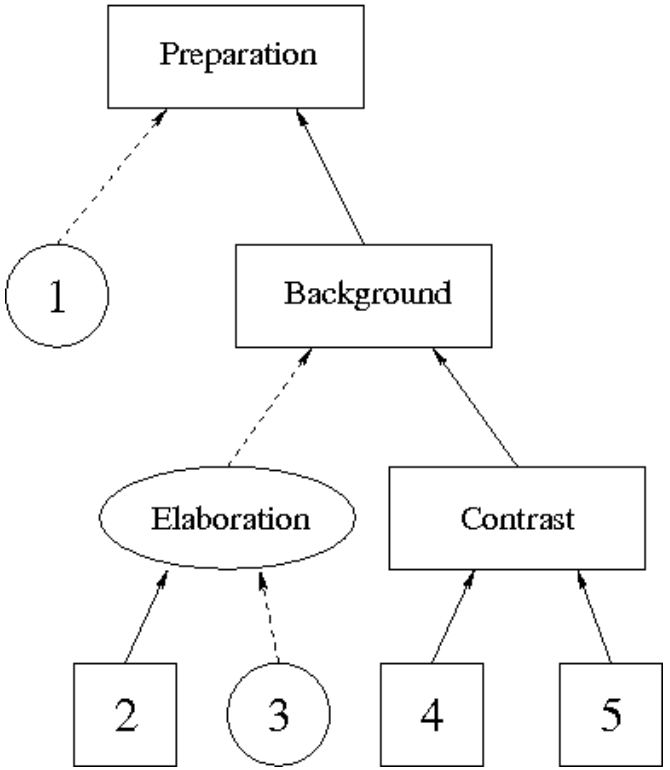
rhetorical structure trees (1)

- un modello per rappresentare le relazioni tra le frasi di un discorso
- Due tipi di relazione:
  - **Nucleo-satelliti**: la frase **nucleo** è centrale allo scopo dell'autore mentre le frasi **satelliti** chiarificano o aggiungono informazioni riguardo al nucleo, esempi:
    - **Background:** i satelliti facilitano la comprensione del nucleo
    - **Elaboration:** il nucleo fornisce l'informazione di base, i satelliti informazioni aggiuntive
    - **Evidence:** i satelliti forniscono delle prove a favore dell'affermazione fatta nel nucleo
  - **Multinucleare:** tutte le frasi che compongono la relazione hanno la stessa importanza, esempio:
    - **Contrast:** due alternative

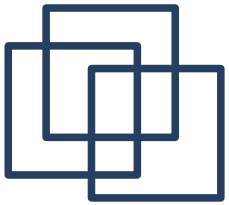
[9][10][11]

RST: esempio

- Il risultato è una struttura ad albero
- Esempio:
  [Lactose and Lactase. 1][Lactose is milk sugar; 2][the enzyme lactase breaks it down. 3][For want of lactase most adults cannot digest milk. 4][In populations that drink milk the adults have more lactase, perhaps through natural selection 5]

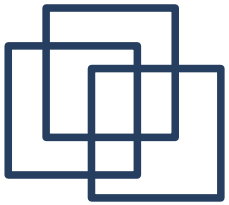


- I satelliti sono in cerchi/ellissi collegati da freccie tratteggiate
- I nuclei sono in quadrati/rettangoli collegati da freccie continue



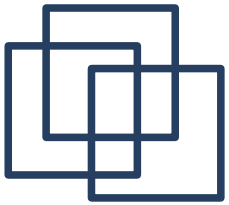
RST: generazione automatica (1)

- Gli RST non sono stati pensati inizialmente per essere generati automaticamente
  - La differenza tra due tipi di relazione retorica spesso è sottile
  - In generale le relazioni scelte non permettono un interpretazione non ambigua del testo
    - D'altra parte però i testi in linguaggio naturale hanno sempre un certo grado di ambiguità...
  - Esistono almeno 300 tipi di relazioni retoriche
    - Quelle più usate sono circa 30...
    - e possono essere raggruppate in 16 tipi principali



RST: generazione automatica (2)

- Algoritmo proposto da Marcu nel 1997
  - utilizza i “marcatori di discorso” per ipotizzare le possibili relazioni tra le frasi del testo
    - Le ipotesi possibili per ogni marcatore sono state identificate analizzando un corpus
    - i “marcatori di discorso” sono simili alle locuzioni bonus del cue method
  - Vengono eliminate le ipotesi che non soddisfano i vincoli di validità per RST
  - Tra gli alberi RST sopravvissuti viene scelto quello più sbilanciato a destra
    - Anche questa euriistca è stata derivata empiricamente dall'analisi del corpus

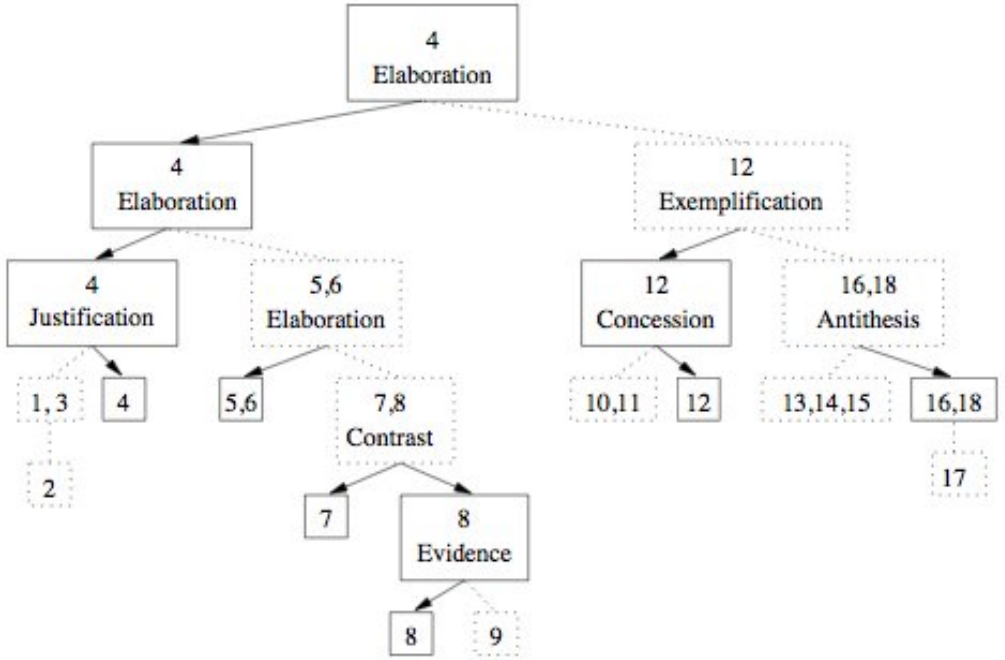


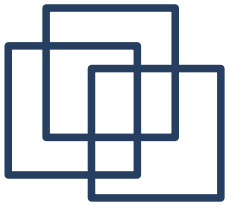
RST: summarization (1)

- Una volta generato l'RST determino la frase più importante della relazione radice
- La frase più importante di una relazione è:
  - le frasi più importanti dei nuclei, se la relazione è multinucleare
  - la frase più importante del nucleo, se la relazione è nucleo-satelliti
- Si visita l'albero in ampiezza:
  - emettendo le frasi più importanti ad ogni livello
  - finché non si raggiunge la lunghezza desiderata per il riassunto

RST: summarization (2)

[With its distant orbit 1][- 50 percent farther from the sun than Earth - 2][and slim atmospheric blanket, 3][Mars experiences frigid weather conditions. 4][Surface temperatures typically average about 60 degrees Celsius (76 degrees Fahrenheit) at the equator 5][and can dip to 123 degrees C near the poles. 6][Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, 7][but any liquid water formed in this way would evaporate almost instantly 8][because of the low atmospheric pressure. 9]
[Although the atmosphere holds a small amount of water,10][and water-ice clouds sometimes develop, 11][most Martian weather involves blowing dust or car-bon dioxide. 12][Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, 13][and a few meters of this dry-ice snow accumulate 14][as previously frozen carbon dioxide evaporates from the opposite polar cap. 15][Yet even on the summer pole, 16][where the sun remains in the sky all day long, 17][temperatures never warm enough to melt frozen water. 18]





RST: Valutazione (1)

- Recall: 66%
- Precision: 68%
- L'autore stima che sia possibili migliorare fino al 70% per Precision e Recall

Sentences	Random	38.4	38.4
	Microsoft Summarizer	41	39
	Our summarizer	**66**	**68**
	Analysts	67.5	78.5

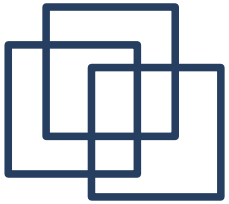
RST: Valutazione (2)

Reduction	Method	Recall	Precision	F-score
10%	Humans	83.20%	75.95%	79.41%
	Program	68.33%	84.16%	75.42%
	Lead	82.91%	63.45%	71.89%
20%	Humans	82.83%	64.93%	72.80%
	Program	59.51%	72.11%	65.21%
	Lead	70.91%	46.96%	56.50%

TREC Corpus

Level	Method	Recall	Precision	F-score
Clause	Humans	72.66%	69.63%	71.27%
	Program	67.57%	73.53%	70.42%
	Lead	39.68%	39.68%	39.68%
Sentence	Humans	78.11%	79.37%	78.73%
	Program	69.23%	64.29%	66.67%
	Lead	54.22%	54.22%	54.22%

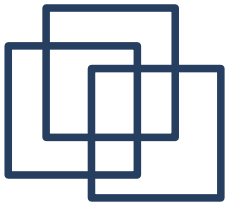
Scientific American Corpus



coesione

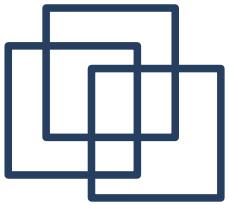
- Gli strumenti linguistici utilizzati per dare coerenza al testo
  - ripetizione di parole, sinonimia, congiunzioni, ellissi…
- È generalmente più facile da individuare della struttura del discorso
  - Ma la coesione è un segnale della presenza di strutture del discorso
- Le catene lessicali permettono di identificare la coesione di un testo

[12]



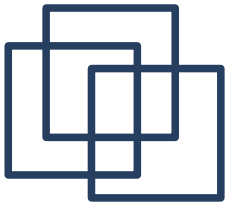
catene lessicali [12]

- Catena lessicale:
  - Gruppo di parole semanticamente vicine
  - Ovvero: sinonimi, iperonimi, meronimi, etc
- Ipotesi:
  - Le frasi importanti sono attraversate da catene lessicali “forti”



individuazione catene lessicali

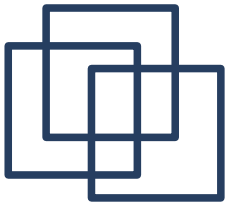
- Algoritmo generale:
  - Si selezionano le parole del testo che si intendono prendere in considerazione per le catene lessicali (solitamente solo i sostantivi)
  - per ogni parola selezionata:
    - si stabilisce se appartiene ad una catena lessicale esistente:
      - se sì, si aggiunge la parola alla catena lessicale
      - altrimenti si crea una nuova catena lessicale per la parola



metodo 1

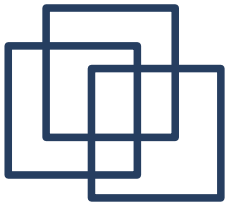
G. Hirst & D. St-Onge, Lexical chains as representation of context for the detection and correction of malapropisms

- Scelta delle parole: Tutti i sostantivi presenti in WordNet
- Tre tipi di relazione tra due parole
  - extra-strong: le due parole sono uguali
  - strong: collegate da una relazione di WordNet
  - medium-strong: collegamento tra i synset più lungo di uno (ma non tutti i collegamenti vanno bene)
- Una parola appartiene ad una catena lessicale se esiste una parola della catena lessicale:
  - in relazione extra-strong con essa
  - in relazione strong con essa e distante al più sette frasi
  - in relazione medium-strong con essa e dist. al più tre frasi



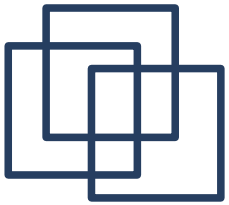
metodo 1 (2)

- la scansione delle parole selezionata viene effettuata seguendo la loro posizione nel testo
- la disambiguazione del significato della parola viene effettuata al momento in cui viene inserita in una catena lessicale
  - una volta effettuata la disambiguazione il significato della parola non cambia più



metodo 1: problema

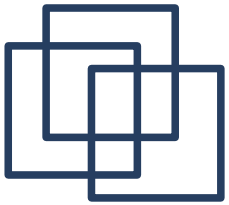
- “Mr. Kenny is the person that invented an anaesthetic machine which uses micro-computers to control the rate at which the anaesthetic is pumped into the blood. Such machines are nothing new. But his device uses two micro-computers to achieve much closer monitoring of the pump feeding the anaesthetic into the patient”
  - Parole selezionate: “Mr.”, “person”, “anaesthetic”, “machine”, “micro-computers”, “rate”, “anaesthetic”..
- person viene associato alla CL di “Mr.” disambiguato come: `[lex “person”, sense {person, individual, someone, man, mortal, human, soul}]`
- “anaesthetic” viene inserito in una catena lessicale nuova
- “machine” viene disambiguato come olonimo di “person” (“an efficient person”) e assegnato alla catena lessicale con “Mr.” e “person”
- **vogliamo che venga associato alle CL di “micro-computer”, “device”, “pump”...**



metodo 2

R. Barzilay & M. Elhadad, Using Lexical Chains for Text Summarization

- Tre tipi di relazioni:
  - sinonimia e ripetizione: peso 10
  - antonimia: peso 7
  - iperonimia/meronimia: peso 4
- Trattamento della polisemia:
  - Si considerano tutti i significati di ogni parola
    - Il numero delle interpretazioni di un testo cresce esponenzialmente
      - “person” (2 significati) -> 2 interpretazioni
      - aggiungo “machine” (5 significati) -> 2*5 = 10 interpr.
    - Quando le interpr. diventano troppe quelle più deboli vengono eliminate
    - Alla fine scelgo l'interpr. più forte (= più connessioni)

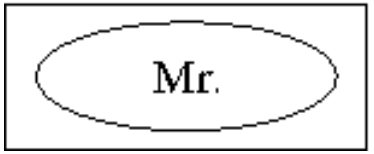


metodo 2: dettagli (1)

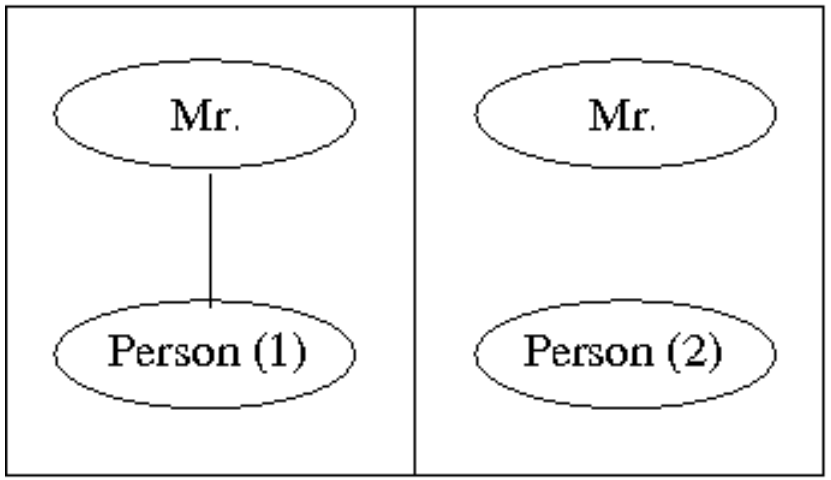
- Selezione delle parole:
  - Invece di utilizzare solo i sostantivi presenti in WordNet si considerano anche i nomi composti
    - Ad esempio “quantum computer”
    - Si utilizzano dei criteri per determinare quale è la testa di un nome composto (in “quantum computer” si utilizzerà “computer”)
- Segmentazione del testo
  - Il testo viene diviso in segmenti in base alla frequenza delle parole
  - Ogni segmento viene esaminato a parte
  - Due catene lessicali contenute in due segmenti diversi vengono unite solo se contengono almeno una parola uguale usata con lo stesso significato

metodo 2: esempio (1)

- Inserisco “Mr.”:



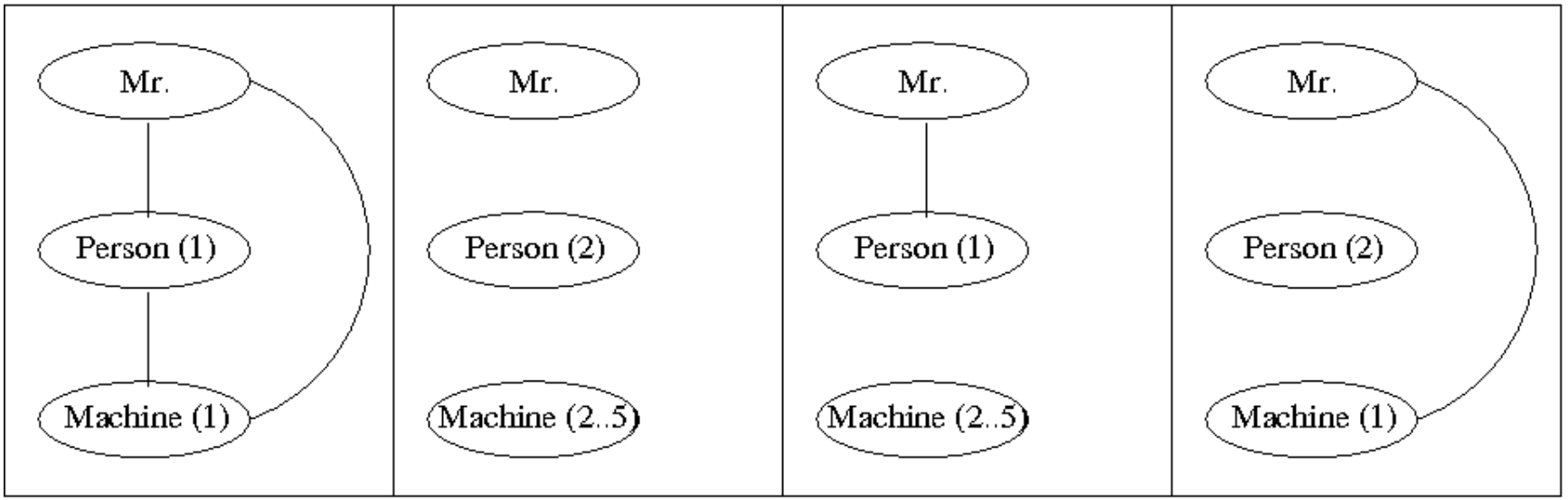
- Inserisco “person” e poi “anaesthetic”:

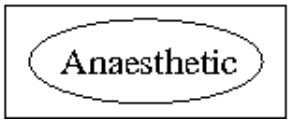


- person (1): human being, etc
- person (2): sign. grammaticale

metodo 2: esempio (2)

- Inserisco “machine”:

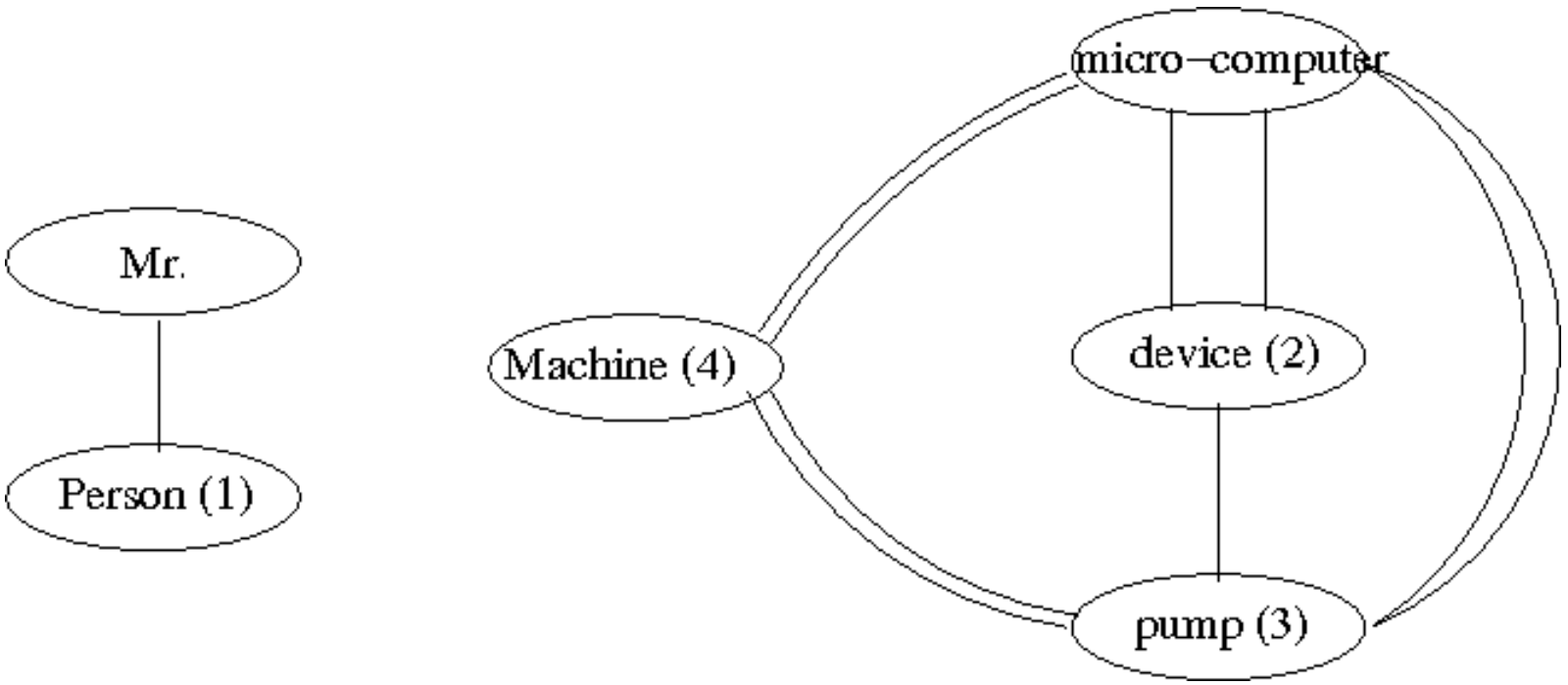


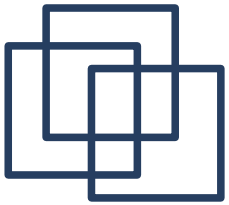


- machine (1): an efficient person
- machine (4): device, etc...

metodo 2: esempio (3)

- Alla fine viene selezionata l'interpretazione più forte (in base ai pesi connessioni):



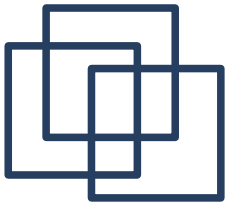


catene lessicali: summarization

- Una volta identificate le catene lessicali:
  - Scelgo le catene più importanti
    - punteggio_catena = lunghezza * omogeneità
      - lunghezza = numero di occorrenze dei membri della catena nel testo
      - omogeneità = 1 - il numero di occorrenze distinte diviso la lunghezza
    - Le catene importanti sono quelle che hanno: punteggio_catena > punteggio_medio + 2*dev_std

catene lessicali: summarization (2)

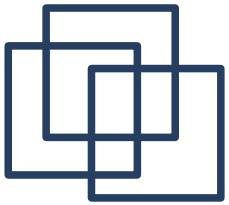
- Dopo aver identificato le catene lessicali più importanti:
  - Estraggo le frasi più importanti dal testo:
    Tre euristiche:
    - per ogni catena scelta estraggo la prima frase in cui compare un membro della catena
      - Problema: non tutti i membri di una catena sono ugualmente rappresentativi della catena
    - per ogni catena scelta estraggo la prima frase in cui compare un membro rappresentativo della catena
    - per ogni catena scelta si identifica l'unità di testo in cui la catena è l'argomento centrale:
      - in base alla densità di membri della catena nel testo
      - Da questa unità si estrae la prima frase contenente un membro significativo della catena
- NB: estraggo una sola frase per catena



catene lessicali: valutazione

- Euristiche:
  - La prima e la seconda euristica di solito producono lo stesso risultato
    - Quando producono un risultato diverso la seconda è migliore
    - In generale la seconda euristica produce i migliori risultati
    - La terza euristica, più sofisticata, produce i peggiori risultati

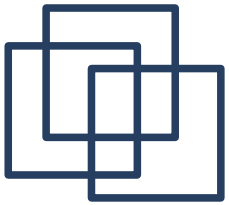
	Microsoft		Lexical Chain	
	Prec	Recall	Prec	Recall
10%	33	37	61	67
20%	32	39	47	64



information extraction (1)

- Idea:
  - Utilizzare sistemi di information extraction per produrre riassunti
  - Si deve produrre un modulo che verrà riempito dal sistema con le informazioni richieste
  - Query based
    - È necessario avere una idea precisa delle informazioni che ci interessano
      - Altrimenti non è possibile produrre il modulo da riempire
    - Nessuno dei sistemi visti fino qui è facilmente adattabile al caso di summarization query based.

[13][14]

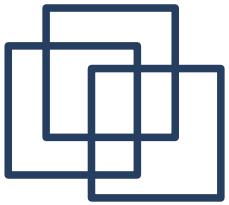


information extraction (2)

- Esempio di modulo/template riempito:

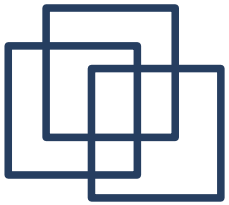
MESSAGE:ID	TSL-COL-0001
SECSOURCE:SOURCE	Reuters
SECSOURCE:DATE	26 Feb 93 Early afternoon
INCIDENT:DATE	26 Feb 93
INCIDENT:LOCATION	World Trade Center
INCIDENT:TYPE	Bombing
HUM TGT:NUMBER	AT LEAST 5

- Passaggi per il riempimento di un template:
  - identificazione dei nomi nel testo
  - risoluzione di riferimenti (alias, pronomi, sintagmi nominali...)
  - estrazione dal testo (elaborato dai passi precedenti) delle informazioni necessarie a riempire il template



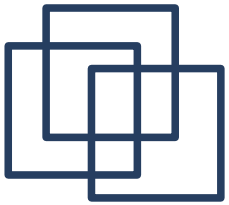
information extraction (3)

- Identificazione dei nomi
  - Approcci: Rule Based, Hidden Markov Model
  - Accuratezza vicina all'80%
- Risoluzione dei riferimenti
  - Alcuni tipi di riferimento
    - **alias**: International Business Machines, IBM, Big blue...
    - **descrizioni**: “il gigante di redmond”
    - **pronomi**
  - Accuratezza: dipende molto dal dominio (può essere anche solo 50 – 60%)



information extraction (4)

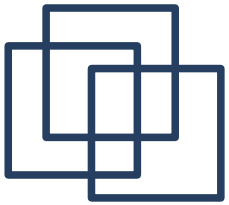
- Estrazione di relazioni ed eventi
  - Approccio “molecolare”
    - Esaminando i testi si individuano dei pattern utili per l'identificazione delle informazioni da inserire nel modulo, si codificano questi pattern in regole
  - Approccio “atomico”
    - Si costruisce una relazione/evento a partire da quasi tutti i sintagmi verbali
    - Queste parti di descrizione vengono fusi assieme cercando di ricostruire le informazioni necessarie a riempire il modulo
  - In entrambi i casi si devono usare informazioni sul dominio:
    - codificate a mano o provenienti da una KB pubblica (WordNet...)



IE: valutazione

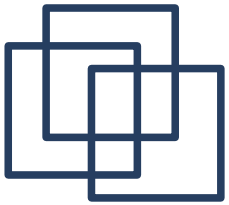
- Problemi (nel contesto della summarization):
  - I moduli/template sono specifici per un dominio
    - È necessaria molta conoscienza specifica del dominio e i moduli/template difficilmente potranno essere usati per altri tipi di testo
  - Strettamente query based
    - La query c'è sempre, implicitamente o esplicitamente
    - Soltanto le query che possono essere soddisfatte da un modulo/template presente nel sistema sono accettabili
- Conferenze MUC
  - Risultati dei migliori algoritmi di IE per anno:

	1989	1992	1996
Recall	63.9	71.5	67.1
Precision	87.4	84.2	78.3



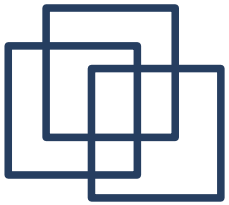
5. Generazione

[2][3][6]

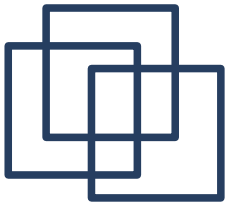


generazione

- Una volta estratte le frasi più importanti si può produrre il riassunto, possibilità
  - Nessuna elaborazione:
    Le frasi estratte vengono semplicemente concatenate:
    - il problema è la risoluzione dei riferimenti
      - soluzione: includere la/le frase/i precedenti
      - altra soluzione: risolutore di riferimenti (come per l'IE)
  - Elaborazioni semplici:
    assemblare gruppi di frasi estratte tra loro
  - Completo:
    - pianificazione della frase, linearizzazione grammaticale...



6. Bibliografia



bibliografia (1)

1. Summarization Website
   http://www.summarization.com/

2. Automated Text Summarization Tutorial, COLING/ACL '98, by E.Hovy e D. Marcu
   http://www.isi.edu/~marcu/acl-tutorial.ppt

3. Text Summarization Tutorial ACM SIGIR, by D. R. Radev
   http://www.summarization.com/sigirtutorial2004.ppt

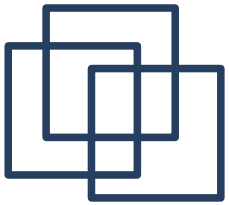
4. Google – Corporate History
   http://www.google.com/intl/en/corporate/history.html

5. World Wide Web, wikipedia article
   http://en.wikipedia.org/wiki/World_Wide_Web#Statistics

6. Automatic Text Sumarisation, by Hugo de Vries
   http://www.comp.mq.edu.au/units/slp803/students/s3141227/Pres.htm

7. New Methods in Automatic Extraction, Journal of ACM, by H.P.Edmundson
   http://portal.acm.org/citation.cfm?coll=GUIDE&dl=GUIDE&id=321519

8. Identifying Topics by Position, by C-Y Lin e E. Hovy
   http://acl.ldc.upenn.edu/A/A97/A97-1042.pdf



bibliografia (2)

9. Introduction to: Rhetorical Structure Theory, by W.C. Mann
   http://www.sfu.ca/rst/01intro/intro.html

10. From Discourse Structure to Text Summaries, by D. Marcu
    http://www.isi.edu/~marcu/papers/summary97.ps

11. The Rhetorical Parsing of Natural Language Texts, by D. Marcu
    http://acl.ldc.upenn/edu/P/P97/P97-1013.pdf

12. R. Using Lexical Chains for Text Summarization, by Barzilay & M. Elhadad
    http://www.cs.bgu.ac.il/~yaeln/papers/summary.ps.gz

13. Information Extraction: A User Guide, by H. Cunningham
    ftp://ftp.dcs.shef.ac.uk/home/hamish/auto_papers/Cun99c.ps.gz

14. Introduction to Information Extraction, by D.E.Appelt, D.J.Israel
    http://coleweb.dc.fi.udc.es/docencia/ln/biblioteca/ie.pdf