Elaborazione del linguaggio naturale

Question Answering Systems

Michele Guiduzzi
guiduzzi@cli.di.unipi.it

aa 2003/2004

Introduzione (1)

- Scopo dell'elaborazione del linguaggio naturale (ELN):

espressioni in linguaggio naturale
(ambiguo e impreciso)

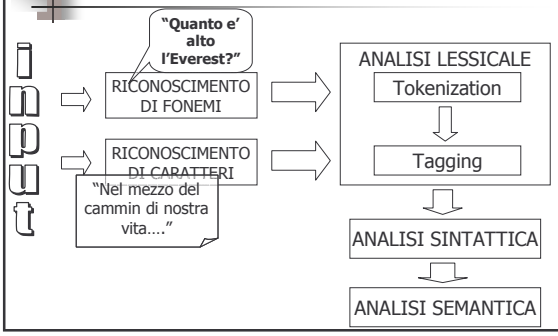
⇩

Rappresentazione interna
(non ambigua)

ELN: fasi

- Dato un input in linguaggio naturale si distinguono 3 fasi:
  - **L'analisi lessicale** che comprende:
    - Tokenizzen, cioè il riconoscimento di forme (morfologia=declinazioni e coniugazioni; catalogazione in: nomi, articoli, agg., verbi, etc.)
    - Tagging, categorizzare le forme riconosciute
  - **L'analisi sintattica**
  - **L'analisi semantica**

ELN: schema



Access Information

- Un tema di ELN e' l'Accesso all'Informazione (AI). Problemi:
  - Grande dimensione della collezione di dati
  - Duplicazione di informazioni
  - Informazioni non veritiere
- Gli approcci tipici dell AI sono:
  - Information Retrival (IR)
  - Information Extraction (IE)
  - Question Answering (Q/A)

Information Retrival (IR)

- I sistemi di IR sono caratterizzati da:
  - Uso di metodi statistici
  - Basati sulla frequenza delle parole nella query, nei documenti, nella collezione
  - Recupera documenti "interi"
- Limitazioni:
  - Non cerca di capire il significato della query dell'utente

Information Extraction (IE)

- I sistemi di IE sono caratterizzati da:
  - templates = domande predefinite
  - identifica messaggi facenti parte di argomenti specifici
  - estrae informazioni basandosi sull'uso di templates
  - restituisce "risposte"
- Limitazioni:
  - i templates sono costruiti personale esperto
  - I templates dipendono dal dominio e sono difficilmente portabili

Question Answering (1)

- Un sistema Question Answering (QA): software di recupero di informazioni
- Spesso usa database (lessicali) che si occupano di disambiguare le parole e la loro trasformazione in forma semantica
- Una prima tassonomia dei sistemi QA:
  - Open domain: integrano tecniche IR e IE a tecniche per il trattamento di fenomeni linguistici
  - Closed domain: operano su basi di dati piuttosto piccole,

Question Answering (2)

- Le caratteristiche del Q/A sono:
  - Domande poste in linguaggio naturale, non query
  - Domande specifiche per risposte "precise"
  - La risposta e' una porzione di testo, più o meno grande
- Limitazioni:
  - Risposte più lente
  - Sistemi più sofisticati

QA: un po' di storia (1)

- I primi sistemi che sono considerati di QA nascono negli anni '60
  - Architettura semplice
  - corpus di documenti limitato ⇨ closed domain
- Due tipologie di sistemi:
  - Sistemi di raccolta dati (natural language database systems): Baseball, Lunar
  - Sistemi di dialogo (dialogue systems):
    - sistemi non completi e non applicati a domini specifici: Eliza

QA: un po' di storia (2)

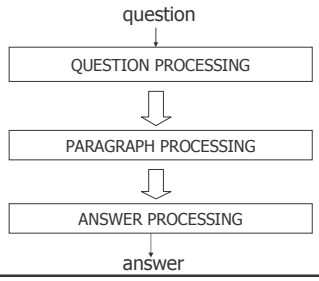
- **Baseball (1961):**
  - Risponde a domande poste in inglese riguardanti il campionato di baseball americano. Livello sintattico e semantico.
- **Lunar (1970):**
  - Costruito per aiutare i geologi ad ottenere informazioni sul suolo lunare, senza aver conoscenza tecnica. Livello sintattico e semantico.
- **Eliza (1967):**
  - Riproduce la conversazione tra paziente e psichiatra. Costruzione delle risposte basata su schemi fissi e/o parole chiave individuate nella domanda dell'utente

QA: architettura generale (1)

- I QA systems sfruttano una base di conoscenza lessicale che ha le caratteristiche di una ontologia
  - **Ontologia(=concettualizzazione di un dominio): ogni elemento di un dominio viene espresso come un concetto e l'insieme di concetti viene organizzato in un insieme di relazioni**
- Molti sistemi usano WordNet: rete semantica di concetti

QA: architettura generale (2)

- Architettura comune dei sistemi di QA:



QA: architettura generale (3)

§ **Question Processing**: è il modulo per l'analisi della domanda; è costituito da:
- Analizzatore morfologico
- Analizzatore sintattico
- Analizzatore semantico

- **Paragraph Processing**: ricerca gli elementi richiesti dalla query all'interno dei documenti
- **Answer Processing**: stabilisce la risposta migliore da riportare

QA: Wordnet (1)

- Database lessicale che vuole essere un modello della memoria lessicale umana in cui le parole organizzate su base lessicale e non alfabetica
- Nomi, verbi, aggettivi organizzati in insiemi di sinonimi (synsets), ognuno dei quali rappresenta un concetto. (es: [terra, globo, sfera])
- Gli insiemi di sinonimi organizzati in una rete tramite relazioni
- Significati – significanti
- Separazione di nomi verbi, aggettivi: categorie sintattiche diverse non possono essere sinonimi
- Relazioni semantiche (tra sysnsets) e lessicali (tra parole dei sysnsets)

QA: Wordnet (2)

Categ	Relazione	Tipo	Esempio
Nomi	Ipo/iperonimia Meronimia	Sem Sem	Dog is a kind of animal Arm is a part of body
Verbi	Implicazione Causa Opposizione Troponimia…..	Sem	The kill causes die
Aggett.	Antonimia	Sem	Hot antonym cold
Avverbi	Agg. Da cui deriva Antonimo	Less Sem	Slowly derived from slow Slow antonym quickly

Principali relazioni tra categorie di parole in Wordnet

QA: question processing (1)

- La prima cosa che fa ogni sistema è individuare le informazioni presenti nella domanda che permettono di giungere alla risposta
- Le domande poste in linguaggio naturale sono ambigue; ciò è causato da:
  - Sinonimia
  - Polisemia
  - Anafora
  - Metafora
  - Variabilità nella costruzione delle frasi

QA: question processing (2)

- Il contesto di una frase aiuta a disambiguare
- Per automatizzare il processo di disambiguazione è necessario bisogna conoscere le relazioni tra le parole
- **Approccio statistico:** assegna il significato ad una parola in base alla Prob che ha di essere inserita insieme alle altre del contesto
- **Approccio basato sulla distanza semantica**: si usano reti semantiche (WordNet) per calcolare la distanza tra due concetti

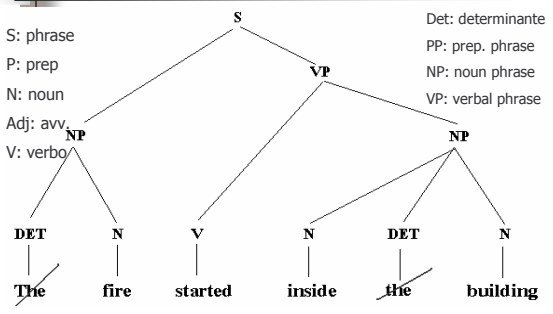
QA: question processing (3)

- **Question parse e Named Entity Tagged:**
  - La domanda viene scomposta in tante entità in base alle categoria lessicale di appartenenza
  - Si vuole una rappresentazione interna della query concetti e dipendenze binarie tra concetti
  - Le "stop words" vengono eliminate

"How much could you rent a Volkswagen bug for in 1966?"

la sua rappresentazione interna cattura la relazione binaria tra il concetto "rent" e "1966"

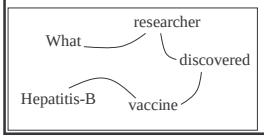
QA: question processing (4)



QA: question processing (5)

- Question semantic form & Expeted answer:
  - Un Diagramma delle relazioni tra parole esprime le dipendenza tra le stesse. Esso viene usato per ottenere la il concetto semantico della risposta, ovvero il tipo di risposta atteso

What researcher discovered the vaccine against Hepatitis-B?



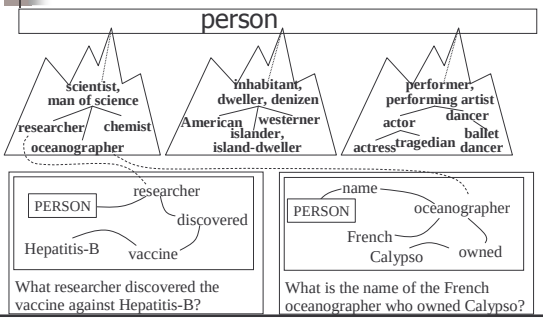
Categorie interne al sistema:

Vaccine -> MEDICINE

Hepatitis-B -> MEDICINE

Researcher -> PERSON

QA: question processing (6) (dentro a Wordnet)



QA: question processing (7)

- § Un set di concetti contenuti nella domanda vengono selezionati come "keywords"
- § **Question reformulation:**
  - § Generare nuove domande semanticamente analoghe alla domanda originale.
    - Avviene attraverso l'espansione delle "keywords":
      - Trasformazione della radice morfologica di una parola
      - Sostituzione di una parola con sinonimi
      - Sostituzione di una parola con iperonimo
    - Aiuta ad individuare il contesto
- ü Derivazioni morfologiche delle parole chiave. casa -> caseggiato
- ü Derivazioni semantiche: casa -> abitazione

QA: Paragraph & Answer Processing

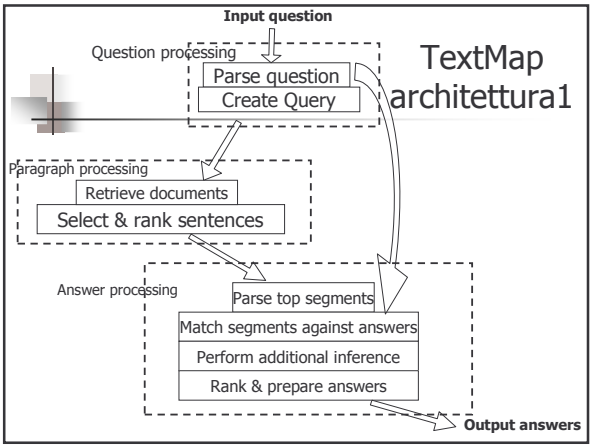
- Le query vengono trasformate in forma booleana
- Si comincia a cercare i documenti che contengono tutte le keywords
- Si cerca i passaggi di testo che contengonon più keywords per identificare le frasi migliori
- Frasi candidate ⇨ forma booleana
- Boolean Query Vs Boolean Answer
- Answer ranking
- answer

TextMap: introduzione

- Evoluzione di Webclopedia del 2002
- Sistema di QA sviluppato dall'Information Sciences Institute (ISI) - University of Southern California
- Usa BBN's Identifinder, un modulo che isola nomi propri in un testo e li classifica in persone, organizzazioni o luoghi
- Presente al TREC 2003* concorso mondiale che valuta i QA systems; ha risposto a 3 tipi di domande:
  - Factoid questions
  - List questions
  - Definition questions

** Vedi dopo*

TextMap architettura1



TextMap: architettura (1)

- I moduli del sistema che intervengono quando devono rispondere ad una factoid question sono:
  - Question analizer, che identifica il tipo di risposta attesa
  - Query generator, che produce specifiche TREC query e Web Query
  - Le Web query sono sottoposte a Google e le TREC query al motore di IR "Inquery". Lo scopo è recuperare 100 Web documenti e 100 TREC documenti

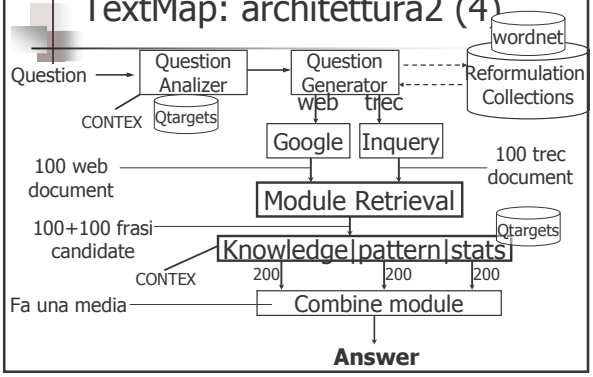
TextMap: architettura (2)

- Un modulo recupera 100 frasi da documenti Web e 100 dai documenti del TREC, scegliendo quelle che sembrano contenere una risposta migliore
- Tre moduli distinti (**knowledge-, pattern-, statistical-based**) individuano in modo indipendente, le risposte corrette tra le 200 e assegnano loro un punteggio
- Un modulo combina i risultati dei tre moduli in una singola classifica

TextMap: architettura (3)

- List questions, ritorna una lista di frasi che hanno ottenuto un certo punteggio
- Per le definition questions l'architettura vengono sfruttate risorse addizionali per eventuali espansioni:
  - WordNet
  - Una collezione di 14,414 biografie (biography.com)
  - Corpus di Mike Fleschman, formato da 966,557 descrittori di persone proprie
  - Un set di relazioni soggetto-verbo, oggetto-verbo, soggetto-copula-oggetto

TextMap: architettura2 (4)



TextMap: query analizer (1)

- Usando BBN's Identifinder, CONTEX (un parser) analizza la domanda e determina il tipo semantico della risposta aspettata (Qtarget); Gli sviluppatori hanno costruito 185 tipi, organizzati in molte classi:
  - **Semantic (concept) Qtargets:** rappresenta la classe più vasta e limita la ricerca a frasi che soddisfano un particolare tipo semantico, estratte per lo più da Wordnet; include C-PROPER-ORGANIZATION, C-BODY-PART, C-COLOR, C-PROPER-ANIMAL

TextMap: query analizer (2)

- **Abstract Qtargets:** è la classe che comprende il tipo di domanda tipici del QA. Per esempio la domanda "who was Mother Teresa?" è equivalente a "Why is the individua known as Mother Teresa famous?". Il suo tipo è A-WHY-FAMOUS. Altri Qtargets: A-YES-NO-QUESTION, A-ABBREVIATION-EXPANSION
- **Syntatic Qtargets:** comprende frasi di cui il sistema non è riuscito a comprendere il tipo semantico (però ha individuato il tipo sintattico). I tipi sintattici sono deboli e spesso non restringono il campo di ricerca.

TextMap: query analizer (3)

S-NP è il Qtargets di default Altri sono S-NP, S-NOUN ("What does Pegeout manifacture?"); S-VP ("That did John Hinckley do to impress Jodie Foster?");
S-PROPER-NAME.

- **Role Qtargets:** questo Qtargets specifica gli elementi del parse tree della domanda e della risposta candidata; ROLE-REASON ("Why did David Koresh ask the FBI for a word processor?"); ROLE-MANNER ("How did David Koresh die?")

TextMap: query analizer (4)

esempio di parse-tree semplificato: "The tournament was cancelled due to bad weather"

((SUBJ LOG-OBJ) The tournament
(PRED) was cancelled
(REASON) due to bad weather
)

La frase "due to bad weather" soddisfa il ROLE Qtargerts

TextMap: query analizer (5)

§ **Slots Qtargets:** riguarda informazioni non sintattiche associate alle frasi. Gli slots possono essere riempiti prima o dopo il parsing.
SLOT TITLE-P TRUE ("Name a novel written by Proust");
SLOT QUOTE-P TRUE ("What did Richard Feynman say upon hearing he would receive the Nobel Prize in Physics?";
SLOT POSSIBLE-REASON-P TRUE

TextMap: query analizer (6)

- **Relations Qtargets:** esprime relazioni tra due tipi semantici come *Person* e *Date* per esprimere il Qtargets R-BIRTHDAY o *Person* e *Noun* per esprimere R-INVENTION

- I Qtargets possono essere combinati con forza variabile:
Question: Where is the Getty Museum?
Qtarget: ((C-PROPER-CITY 1.0)
(C-AT-LOCATION 0.7)
(C-PROPER-PLACE 0.7 ….)

TextMap: query generation (1)

- CONTEX restituisce in output una rappresentazione semantica delle domande
- Vengono indentificati noun phrases, noun, verb phrases, adjective….
- Viene assegnato un punteggio alle parole/frasi della domanda in base (in ordine di rilevanza):
  - alla frequenza del loro tipo in un corpus di domande (27,000+)
  - Alla loro lunghezza
  - Alla frequenza delle parole nel corpus

TextMap: query generation (2)

- Per ridurre il gap tra le parole contenute nella query e nell'answer da recuperare, TextMap genera riformulazioni della query, per aumentare la probabilità di recupero. *E*sempio:

  question: *"How did Mahatma Gandhi die?"*
  Reformulation patterns:
  1) Mahatma Gandhi died <how>?
  2) Mahatma Gandhi died of <what>?
  3) Mahatma Gandhi lost his life in < what >?
  4) Mahatma Gandhi was assassinated ?

….fino a 40 riformulazioni

TextMap: query generation (3)

- Q: "Mahatma Gandhi was assassinated ?"
  A1: "Mahatma Gandhi was assassinated by a young Hindu extremist"
  A2: "Mahatma Gandhi died in 1948"
  A1 è considerata migliore di A2
- La collezione di riformulazioni in TextMap contiene 550 asserzioni raggruppate in circa 105 blocchi di equivalenza
- In TREC-2003 5.03 riformulazioni medie per query

TextMap: answer selection (1)

- Il modulo di risposta **knowledge-based**, usa "Context" che facilita il riconoscimento di Qtargets, arricchito di:
  - set di 1.200 domande
  - Named entity tagging
- La selezione della risposta è guidata:
  - da il grado di "matching" a livello semantico/sintattico tra i parse tree della domanda e della risposta
  - Dall'uso di Wordnet

TextMap: answer selection (2)

- Da euristiche che penalizzano la risposta per alcuni motivi:
  - Qtarget match factor
    Q: How long did the Manson trial last?
    Semantic mismatch: 20 miles
  - Vagueness penalty:
    Q: Where is Luxor? Too vague: on other side
  - Negation penalty:
    Q: Who invented the electric guitar?
    Negation: Fender did **not** invent the electric guitar

TextMap: answer selection (3)

§ Il modulo di risposta **pattern-based** usa un set di pattern imparati automaticamente, sfruttando altavista (che recupera 1000 documenti per query)

- L'apprendimento consiste di due passi:
  - Dato un Qtarget contenuto nell'ontologia di TextMap (BIRTHYEAR) e qualche istanza di coppie <question; answer>, come <NAME_OF_PERSON, BIRTHYEAR>, vengono estratte dal web tutti i diversi pattern (TEMPLATEs) che contengono queste coppie
  - Calcolare la precisione di ogni pattern e mantenere il pattern con più alta precisione

TextMap: answer selection (4)

Pattern appresi dal sistema

Prec.	#correct	#found	pattern
1	122	122	<NAME> (<BD>-<DD>
1	15	15	<NAME> (<BD>-<DD>),
1	13	13	,<NAME> (<BD>-<DD>)
0.9166	11	12	<NAME> was born on<BD> in
0.9090	10	11	<NAME> : <BD>-<TIME>
0.6944	25	36	<NAME> was born on <BD>

TextMap: answer selection (5)

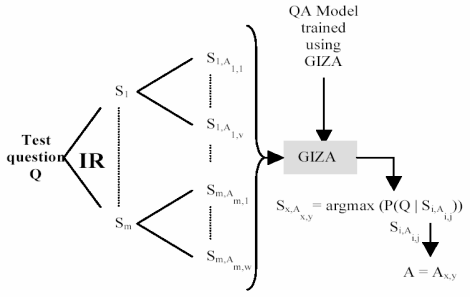
- Dopo aver appreso i pattern, il modulo li usa per identificare le risposte ad una domanda
- Il set di potenziali risposte viene ordinato
- Il pattern-based module viene "allenato" usando 1192 domande del TREC 9 e TREC 10 e 500 questions del TREC 11

TextMap: answer selection (6)

Il **modulo statistico**:

- Sia Sa una frase che contiene al suo interno la risposta A alla domanda Q
- Dato un corpo di coppie (Q, Fa) si può istruire un modello probabilistico in modo che stimi la P(Q|Sa)
- Una volta appresi i parametri del modello, data una domanda Q e un set Σ di frasi (ottenuto da un motore di IR), si trova la frase $S_i \in \Sigma$ e la risposta in essa contenuta $A_{i,j}$, cercando $S_{i,A_{ij}}$ che massimizza la $P(Q|S_{i,A_{ij}})$
- Per stimare i parametri del modello, è stato usato GIZA, un software pubblico di machine-traslation (http://www.clsp.jhu.edu/ws99/projects/mt/)

TextMap: answer selection (7)



TextMap: combining output of multiple answer-selection modules

- Webclopedia riportava alcuni errori. TEXTmap usa un framework basato sulla "massima entropia", per scartare le risposte errate e ri-ordinare quelle corrette:
  - Il modulo-pattern rispondeva bene a domande relative a Qtarget ben definiti (NAMES, ORGANIZATIONS, LOCATIONS) ma non Np Qtargets
  - Il modulo statistico non restringeva il tipo di risposta in base al Qtarget: ciò causava inesattezze
  - Tutti i moduli commettevano altre imprecisioni: ad esempio i moduli statistic e pattern selezionavo come risposte migliori, frasi che iniziavano per "he", "she" e "it"; queste rispste non sono certo buone per le domande factoids

TextMap: special modules resources for aswering definition questions (1)

- La sfida consiste nel estrarre frasi "rilevanti" da un corpo di risposte ottenute da un modulo IR
- Per fare ciò TextMap sfrutta alcune risorse:
  - 14,414 biografie ottenute da http://www.biography.com
  - Inoltre sono state identificate 6,640 parole che occorrono almeno cinque volte nelle biografie
- Collezione di descrittori di Proper People
- Wordnet
- La lista delle parole permette di giudicare in modo positivo una risposta che contiene un termine "in alto nella lista delle parole"

TextMap: special modules resources for aswering definition questions (2)

- Top 20 terms:

494.0 Nobel	251.4 studied	188.3 edited
467.5 Oxford	247.0 travelled	187.5 Painter
406.0 Poems	209.0 poem	183.0 Angeles
384.0 knighted	206.0 Labour	181.7 Physicist
290.0 Info	204.0 Composer	171.9 War
278.0 Ballet	194.5 St	169.2 commanded
257.0 Broadway	188.7 poetry	

QA: differenze tra i sistemi (1)

- Le differenze dei QA system si evidenziano in termini di:
  - **Natura dell'interrogazione** cioè le caratteristiche che influenzano la forma delle domande, il tipo della domanda (chi, cosa, come, quando) e lo scopo della domanda (elencare, ordinare, informare)
  - **Grado di precisione ed esattezza della risposta** (singola parola, frase o frammento di documento)

QA: differenze tra i sistemi (2)

- **Tipologia dei dati utilizzati**
- **Tipo di dominio** su cui opera (open o close domain)
- **Performance** ottenute del sistema
- **Supporto di risorse di conoscenza**
- **Tipo di interazione** (qualità della domanda e risposta) e caratteristiche qualitative dei documenti

QA: parametri di qualità

- I parametri di qualità di un sistema di QA sono:
  - **Tempestività**: minore è il tempo di risposta del sistema, migliore sarà la qualità
  - **Accuratezza**: è il rapporto tra prestazioni offerte e risorse impiegate dal sistema
  - **Utilizzabilità**: riguarda la difficoltà con cui l'utente si interfaccia al sistema
  - **Affidabilità**: capacità del sistema di garantire determinate prestazioni sotto certe condizioni
  - **Rilevanza**: riguarda la risposta. Essa è rilevante se è precisa e libera da contesto

TREC (1)

- **TREC (Text Retrieval Conference)** : è un concorso annuale con lo scopo di incoraggiare e promuovere la ricerca nel campo di recupero di informazioni
- Nato nel 1992
- Il TREC è finanziato e patrocinato dal **NIST** (National Institute of Standards and Tecnology), dal **DARPA/IAO** (Information Awareness Office of Defence Advanced Research Projects Agency) e dall'**ARDA** (US Department of Defence Advanced Research and Development Activity)

TREC (2)

- **Obiettivi:**
  - Incoraggiare ricerca del recupero di informazione, premiando sistemi software migliori
  - Incentivare la comunicazione tra industrie, università e istituzioni governative, attraverso scambi di idee e tecnologie
  - Spingere lo sviluppo di software più efficienti
- Il concorso è diviso in sezioni differenti (tracks) il cui numero varia ogni anno; ogni sistema può partecipare a più tracks

TREC (3)

TRACKS	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Ah hoc	18	24	26	23	28	31	42	41	-	-	-
Routine	16	25	25	15	16	21	-	-	-	-	-
Interactive	-	-	3	11	2	9	8	7	6	6	6
Spanish	-	-	4	10	7	-	-	-	-	-	-
Confusion	-	-	-	4	5	-	-	-	-	-	-
Database Merging	-	-	-	3	3	-	-	-	-	-	-
Filtering	-	-	-	4	7	10	12	14	15	19	21
Chinese	-	-	-	-	9	12	-	-	-	-	-
NLP	-	-	-	-	4	2	-	-	-	-	-
Speech	-	-	-	-	-	13	10	10	3	-	-
Cross-language	-	-	-	-	-	13	9	13	16	10	9
High precision	-	-	-	-	-	5	4	-	-	-	-
Very large Corpus	-	-	-	-	-	-	7	6	-	-	-
Query	-	-	-	-	-	-	2	5	6	-	-
Q/A	-	-	-	-	-	-	-	20	28	36	34
Web	-	-	-	-	-	-	-	17	23	30	23
Video										12	19
Novelty											13
Partecipanti totali	22	31	33	36	38	51	56	66	69	87	93

TREC (4)

- Il Corpus può essere costituito da:
  - Un insieme di documenti
  - Un insieme di informazioni connesse a ciascun documento (chiamate topics)
  - Giudizi di pertinenza o rilevanza (Relevance Judgments)
- TREC-1: documenti costituiti da articoli vari e quotidiani; qualche documento governativo

TREC (5)

- Topics:
  - un'unità di informazione che evidenzia le caratteristiche principali di ogni documento
  - Elementi:
    - <id>,<title>,<description>,<corpus>
  - I topics vengono costruiti dagli esaminatori (assessors)
- QA track: nasce nel 1999 (TREC-8)

TREC-8*(1999)* Vs. TREC-9*(2000)*

- Ogni sistema partecipante viene fornito di un corpus di documenti e di un set di 200 domande, generalmente brevi
- Ogni domanda ha risposta in almeno un documento
- Per ogni domanda i sistemi devono riportare una lista di 5 frammenti di testo, definiti come [*riferimento al documento*, *stringa della risposta*]
- Caratteristiche frammenti:
  - Lunghezza variabile di 50 o 250 caratteri
  - recuperati dal documento di origine o generati da più documenti
  - gli assessor assegnano il punteggio alle risposte

TREC-8*(1999)* Vs. TREC-9*(2000)*

- **TREC-8:** Raccolta di 200 domande
  - risposta valutata da tre esaminatori
    - Esaminatore super partes: assegna la valutazione in caso di discordanza
  - Domande costruite dagli esaminatori NIST:
    - ⇨ domande ricavate dal corpus dei documenti
    - ⇨ risposte più facili (condividono i vocaboli della domanda)

528.000 articoli presi dal Los Angeles Times, Financial Times, Foreign Broadcast Information Service (FBIS), Federal Register

TREC-8*(1999)* Vs. TREC-9*(2000)*

- **TREC-9:**
  - risposta valutata da un esaminatore
  - Raccolta di 500 domande + 193 domande che sono variazioni sintattiche delle prime
  - Domande estratte dal Encarta ed Exicite Log
    maggiore difficoltà nel recupero della risposta
  - concetto di risposta "non supportata"
    - Una risposta non è supportata se inserita in un contesto sbagliato
  - 979.000 documenti

TREC-8*(1999)* Vs. TREC-9*(2000)*

- Criterio generale di valutazione: la risposta (la coppia [doc-id, answ-string]) è giudicata corretta se:
  - La coppia contiene la stringa di risposta
  - La risposta è corretta e pertinente
  - Il documento supporta la risposta
- Una risposta è non supportata se inserita in un contesto sbagliato

TREC-8 Vs. TREC-9*(2000) main task*

- La valutazione di una risposta i è data in base al "Mean Reciprocal Rank" (MRRi)
  - I 5 frammenti di risposta vengono ordinati dal sistema in base alla correttezza
  - Gli assessor stabiliscono il frammento più corretto p
  - Per ogni risposta i:
    - MRRi= 1/posizione del frammento p riportata dal sistema
  - MRR**totale** = $\sum_{i=1}^{n} MRR_i$ (per ogni risposta corretta)

TREC-8 Vs. TREC-9*(2000) main task*

Frammento1	Frammento2	Frammento3	Frammento4	Frammento5
1	0.5	0.33	0.25	0.2

(procedura di calcolo MRR per ogni risposta)

Risposte date dal sistema:

- <doc-id, string>
- **<doc-id, string>**
- <doc-id, string>
- <doc-id, string>
- <doc-id, string>

Risposta giudicata corretta dagli assessors:

MRR = 0.5

Normalmente le risposte giudicate corrette sono tra i primi frammenti

TREC-10 (2001)

- Vengono aggiunti al main task altri due moduli: **list task** e **context task**
- Corpus di documenti attinto da: AP newswire, Wall Street Journal, San Jose Mercury News, Financial Times, Los Angeles Times, Foreign Broadcast Information Service
- 979.000 articoli
- 3,033 Mbytes ti testo

TREC-10 (2001): main task

- Raccolta di 500 domande
- 5 frammenti riportati <= 50 caratteri
- Selezionate solo domande che contengono pronomi e congiunzioni interrogative (what, when,…), verbo essere o verbi modali e frasi interrogative
- Filtraggio degli esaminatori del NIST; vengono eliminate:
  - Domande che richiedono una lista di elementi
  - Le "Yes/no questions"
  - Domande procedurali
  - Domande troppo attuali

TREC-10 (2001): main task

- Le domande non sono necessariamente collegate ad almeno un documento:
  - Possibile risposta: <doc-id, NIL>

    dove NIL significa che nessun documento contiene la risposta alla domanda. La risposta NIL può essere corretta o scorretta a seconda che il sistema si sia effettivamente sbagliato o meno

TREC-10 (2001): main task

- Parametri di valutazione:
  - Valutazione strict: risposte non supportate considerate scorette
  - Valutazione lenient: risposte non supportate considerate corrette
  - NIL ritornate: numero di nil restituite dal sistema
  - NIL corrette: nil "reali"
  - Final sure: percentuale della sicurezza del sistema sulle risposte date
  - Sure correct: percentuale delle risposte effettivamente "indovinate" dal sistema

TREC-10 (2001): main task

RUN	Strict no correct MRR	#qs	%	*Lenient* no correct MRR	#qs	%	#qs NIL Returned	#qs NIL Correct	Final Sure Sure	Sure correct
Insight	0.68	152	30.9	0.69	147	29.9	120	38	75%	77%
LCC1	0.57	171	34.8	0.59	159	32.3	41	31	100%	51%
Orcl1	0.48	193	39.2	0.49	184	37.4	82	35	100%	40%
Isi1a50	0.43	205	41.7	0.45	196	39.8	407	33	80%	38%
Uwmtal	0.43	212	43.1	0.46	200	40.7	492	49	100%	35%
Mtsuna0	0.41	220	44.7	0.42	213	43.3	492	49	100%	32%

ID dei Sistemi in gara

TREC-10 (2001): list task

- 25 domande costruite dagli assessori NIST
  ⇒ Difficoltà di estrazione della risposta < main task
- Le risposte sono formulate attingendo da documenti differenti
- Ogni assessore crea una domanda breve (numero risposte<=5), due medie (tra 5 e 20) e una grande(tra 30 e 40)
- Lunghezza massima della risposta 50 caratteri
- 1 lista = 1 risposta

TREC-10 (2001): list task

- A tutte le risposte della lista vengono assegnati i giudizi "corretta", "incorretta" e "non supportata"
- "accuracy": parametro che si riferisce ad ogni risposta calcolato come:
  - Accuracy = rc/rr dove rc=risposte corrette riportate dal sistema
    rr=risposte da recuperare

§ Accuracy totale: calcolata come media delle accuracy di ogni singola domanda

TREC-10 (2001): context task

- Domande divise in sezioni differenti
- Il sistema deve rispondere a più domande della stessa serie
- I sistemi non sono in grado di rispondere con la stessa abilità a tutte le domande ⇒ insuccesso del task

TREC-11 (2002)

- Main task e list task
- Documenti estratti dal "AQUAINT Corpus of English News Text". Provenienti da 3 fonti:
  - New York Times newswire 1998-2000
  - AP newswire 1998-2000
  - Xinhua News Agency 1996-2000
- 3 Gbyte di documenti
- Gli assessori correggono gli errori nel corpus di documenti

TREC-11 (2002): main task

- Il TREC-11 possiede alcune nuove caratteristiche rispetto all'edizione 2001:
  - Ogni sistema deve riportare una risposta per domanda
  - Nuovo parametro di valutazione: Confidence-Weighted Score (CWS)

$$1/Q \sum_{i=1}^{Q} num\,correct\ first\,i\,ranks/i$$

dove Q è il numero totale di domande sottoposte al sistema

TREC-11 (2002): main task

- Parametro Precision: rapporto tra numero di informazioni pertinenti estratte e il numero totale di informazioni estratte
- Parametro Recall: rapporto tra #informazioni pertinenti estratte e #totali di informazioni pertinenti da estrarre La risposta deve essere "una stringa esatta" e non più un frammento di documento
- Nel punteggio finale concorrono solo le risposte corrette e le stringhe NIL (che sono vere)

TREC-11 (2002): main task

- Possibili valutazioni della risposta [doc-id, answ-string]:
  - Scorretta: la risposta non è quella desiderata
  - NIL: il sistema non ha trovato la risposta; ciò può essere "vero o falso", a seconda che la risposta sia o meno contenuta nel corpus dei documenti
  - Non supportata: la risposta è inserita in un contesto non opportuno
  - Inesatta: la stringa è corretta e supportata ma contiene informazioni ridondanti
  - Corretta: la stringa è supportata ed esaustiva

TREC-12 (2003)

- passages task e il main task

§ 3 Gbyte per 1.033.000 documenti

- 413 domande ricavate da AOL e MSNSearch Log.
- 30 domande non trovano risposta nel corpus

TREC-12: *passage task*

- 1 sola Risposta (<= 250 caratteri da 1 solo documento) per domanda:
  - giudicata da 2 assessori: corretta, scorretta, non supportata; può essere:
    <offset-doc-char, lenght snippet> o NIL
  - **non supportata**: contiene la risposta giusta ed esaustiva ma il documento non è pertinente
  - **corretta** se:
    - Contiene la giusta risposta
    - Il frammento risponde comunque alla domanda
    - Il documento individuato è quello giusto

TREC-12: *passage task*

- **E' giudicata scorretta:**
  - contiene entità multiple della stessa categoria semantica, senza indicare quale entità sia la risposta, non risponde alla domanda
  - risposta che non includono unità di misura corrette
  - riferimenti errati a copie di entità famose

  Q:"Dove si trova il Taj Mahal?" R:"Il casinò ~~Taj Mahal è.....~~"
- Parametri di valutazione:
  - Accuracy
  - NIL Recall: #NIL riportati/30
  - NIL Precision: #NIL riportati/#NIL effettivi

TREC-12: *passage task*

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
LCCpass03	Language Computer Corp.	0.685	0.381	0.800
nuslamp03a	National University of Singapore (Lee)	0.419	0.156	0.333
uwmtCQ2	University of Waterloo (MultiText)	0.351	—	—
umassql	University of Massachusetts	0.201	—	—
answfind1	Macquarie University	0.191	—	—
Saarland	Saarland University	0.169	0.097	0.367
IITBQA1	Indian Institute of Technology Bombay	0.133	0.045	0.100
clr03p2	CL Research	0.119	0.109	0.233
UAmsT03P1	University of Amsterdam	0.111	0.128	0.333
pircsqa3	Queens College, CUNY	0.097	0.000	0.000
NSIR	University of Michigan	0.085	0.075	0.100

TREC-12: *main task*

- Comprende tre tipo di domande:
  - Factoids
  - Lists
  - Definitions
- 54 "corse" per 25 sistemi differenti
- CL Research, Language Computer Corp e University of Amsterdam hanno partecipato ad entrambi i tasks

TREC-12: *main task - factoids*

- È simile al passage task e il sistema restituisce una sola risposta per ogni factoids question
- Il sistema restituisce "la risposta" e non il documento la contiene
- La risposta non deve essere necessariamente estratta da "un solo" documento
- La risposta è della forma:
  <query-id, run-tag, doc-id, answer-string>

TREC-12: *main task - factoids*

- Se il sistema non è in grado di recuperare l'informazione, la risposta sarà del tipo:
  <query-id, run-tag, NIL, " ">
- Sono ammessi tre giudizi alla risposta:
  - **Incorrect:** la stringa non contiene la risposta o non risponde alla domanda
  - **Not supported:** la stringa contiene una giusta risposta ma il documento non supporta la risposta
  - **Not exact:** risposta giusta, documento giusto ma la risposta contiene informazioni inutili o ne mancano alcune
  - **Correct:** la stringa contiene esattamente la giusta risposta e il documento la supporta

TREC-12: *main task - factoids*

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
LCCmainE03	Language Computer Corp.	0.700	0.381	0.800
lexiclone92	LexiClone	0.622	—	—
nusmml03r1	National University of Singapore (Yang)	0.562	0.160	0.400
isi03a	University of Southern California, ISI	0.337	0.071	0.200
IBM2003a	IBM Research (Prager)	0.298	0.082	0.233
MITCSAIL03b	Massachusetts Institute of Technology	0.295	0.258	0.267
uwbqitekat03	University of Wales, Bangor	0.259	0.092	0.967
Albany03l2	University of Albany	0.240	0.109	0.200
irstqa2003w	ITC-irst	0.235	0.121	0.267
BBN2003B	BBN	0.208	0.068	0.100
FDUT12QA1	Fudan University	0.194	0.077	0.233
ntt2003qam1	NTT Communication Science Labs	0.150	0.090	0.267
MITRE2003A	MITRE Corp.	0.148	0.000	0.000
ICTQA2003C	Chinese Academy of Sciences (CAS-ICT)	0.145	0.105	0.467
UAmsT03M2	University of Amsterdam	0.145	0.273	0.100

TREC-12: *main task - lists*

§ Risposta è una "lista "di <doc-id, answer string>
- più risposte da più documenti
- 37 list-question costruite dagli assessori NIST
- Gli assessori creano le risposte alle 37 domande usando PRISE, un motore di ricerca
- Ogni istanza della lista in risposta viene giudica incorretta, corretta, inesatta o non supportata
- Se sono rinvenute nuove istanze di risposta, la lista viene aggiornata
- Set di risposte equivalenti ⟹ gli assessori ne marcano una come "distinct" e le altre "not distinct"

TREC-12: *main task - lists*

- Solo le risposte corrette possono essere marcate come "distinct"
- Sia S la dimensione della lista finale delle risposte, D il numero di risposte "distinct" riportate dal sistema, N il numero totale di risposte riportate dal sistema.
- Sia IP (instance precision) = D/N
- Sia IR (instance recall) = D/S
- Sia F = 2 x IP x IR / (IP + IR)
- La classifica finale è fatta in base alla media del parametro F riportato dai sistemi per le 37 list questions

TREC-12: *main task - lists*

Stimorol	Big Red	Hubba Hubba
Dirol	Winterfresh	Nicorette
Doublemint	Spearmint	
Juicy Fruit	Freedent	
Orbit	Chiclets	
Trident	Double Bubble	
Dentyne	Bazooka	

Answer list for list question 1915 – name of chewing gums found within the AQUANT corpus

TREC-12: *main task - lists*

Run Tag	Submitter	F
LCCmainS03	Language Computer Corp.	0.396
nusmml03r2	National University of Singapore (Yang)	0.319
MITCSAIL03c	Massachusetts Institute of Technology	0.134
isi03a	University of Southern California, ISI	0.118
BBN2003B	BBN	0.097
Albany03I4	University of Albany	0.096
ICTQA2003A	Chinese Academy of Sciences (CAS-ICT)	0.091
FDUT12QA1	Fudan University	0.088
IBM2003c	IBM Research (Prager)	0.077
irstqa2003w	ITC-irst	0.076
MITRE2003A	Mitre Corp.	0.069
UAmsT03M1	University of Amsterdam	0.054
CMUJAV2003	Carnegie Mellon University (Javelin)	0.052
lexiclone92	LexiClone	0.048
ntt2003qam1	NTT Communication Science Labs	0.040

TREC-12: *main task - definitions*

- Una definizione è una domanda del tipo: "who is Colin Powell?"
- Si usano dei metodi per confrontare il concetto della risposta desiderata e il concetto della risposta data dal sistema
- Set di 50 domande:
  - 30 con oggetto persone fisiche
  - 10 con oggetto organizzazioni
  - 10 con oggetto altro
- Gli assessori scelgono le domande tra quelle contenute nei logs dei motori di ricerca e poi cercano i documenti che contengono la risposta

TREC-12: *main task* - *definitions*

- In questo tipo di domande è "bene" conoscere "chi" fa la domanda per conoscere il livello di dettaglio richiesto
- Per risolvere questo problema si fanno assunzioni:
  "Chi fa la domanda è adulto, di lingua Inglese, è un lettore medio del US newspaper….Ha un'idea di base della domanda che pone….Non è un esperto sul dominio della domanda e perciò non cerca dettagli esoterici…. [da "Overview of the TREC2003 Question Answering Track"]

TREC-12: *main task* - *definitions*

- Le risposta sono sempre del tipo <doc-id, anw-string>
- Non c'è limite di lunghezza alla singola risposta e al numero delle risposte;
- La valutazione avviene nel modo seguente:
  - ü Le risposte presentate sotto forma di un'unica "lunga" stringa; usando le risposte e le ricerche precedenti, effettuate per creare le domande, gli assessors creano una lista di "**information nuggets**" riguardanti l'obiettivo della domanda; un nugget è un fatto che permette all'assessore di stabilire se il nugget è contenuto (si/no) oppure no nella risposta

TREC-12: *main task* - *definitions*

- ü Ogni nugget viene poi definito "vital" o "not vital" a seconda se deve apparire in una definizione perché questa sia "buona"
- ü Gli assesori marcano le risposte del sistema che contengono i "nuggets"
- § Nella valutazione, gli assessori valutano "solo il contenuto della risposta"
- § Nugget recall : calcolato su i "vital nuggets"
- § Nugget precision: calcolato su "vital" e "non vital nuggets"

TREC-12: *main task* - *definitions*

- § Per la difficoltà nell'ottenere certi parametri viene usata la "lunghezza di una risposta". La risposta "più corta" è meglio accetta
- § I sistemi sono penalizzati se:
  - § Non recuperano informazioni contenenti "vital nuggets"
  - § Non recuperano informazioni contenenti "nuggets"
- § Il risultato finale è misurato dalla metrica F che ha come parametro β=5; il valore 5 indica che recall è cinque volte più importante che precision

TREC-12: *main task - definitions*

- **r** numero di "vital nuggets" presenti nella risposta
- **a** numero di nuggets presenti nella risposta (nuggets non vital)
- **R** numero totale di nuggets presenti nella lista degli assessori
- **Len** numero di caratteri in ψna risposta (escluso spazi bianchi)

recall=r/R
allowance=100 x (r+a)

precision = 1 se len < allowance
1 – (len – allowance / len) altrimenti

F(β=5) = (26 x Precision x Recall) / (25 x Precision +recall)

TREC-12: *main task - definitions*

1 vital provides remuneration to executives who lose jobs
2 vital assures officials of rich compensation if lose job due to takeover
3 vital contract agreement between companies and their top executives

4 aids in hiring and retention
5 encourages officials not to resist a merger
6 IRS can impose taxes

Information nuggets created for question 1905 "What is a golden parachute?"

TREC-12: *main task - definitions*

Run Tag	Submitter	F($\beta = 5$)	Ave Length
BBN2003C	BBN	0.555	**2059.20**
nusmml03r2	National University of Singapore (Yang)	0.473	**1478.74**
lsi03a	University of Southern California, ISI	0.461	**1404.78**
LCCmainS03	Language Computer Corp.	0.442	**1407.82**
cuaqdef2003	Univ. of Colorado/Columbia Univ.	0.338	**1685.60**
irstqa2003d	ITC-irst	0.318	**431.26**
UAmsT03M1	University of Amsterdam	0.315	**2815.08**
MITCSAIL03a	Massachusetts Institute of Technology	0.309	**620.28**
shef12simple	University of Sheffi eld	0.236	**338.42**
UIowaQA0303	University of Iowa	0.231	**3039.26**
CMUJAV2003	Carnegie Mellon University (Javelin)	0.216	**182.34**
FDUT12QA3	Fudan University	0.192	**203.54**
piq002	University of Pisa	0.185	**89.52**
IBM2003b	IBM Research (Prager)	0.177	**223.16**
ntt2003qam1	NTT Communication Science Labs	0.169	**2219.24**

TREC-12: *main task*

Run Tag	Submitter	Component Score			Final Score
		Factoid	List	Def	
LCCmainS03	Language Computer Corp.	**0.700**	**0.396**	**0.442**	**0.559**
nusmml03r2	National University of Singapore (Yang)	**0.562**	**0.319**	**0.473**	**0.479**
lexiclone92	LexiClone	**0.622**	**0.048**	**0.159**	**0.363**
lsi03a	University of Southern California, ISI	**0.337**	**0.118**	**0.461**	**0.313**
BBN2003C	BBN	**0.206**	**0.097**	**0.555**	**0.266**
MITCSAIL03a	Massachusetts Institute of Technology	**0.293**	**0.130**	**0.309**	**0.256**
irstqa2003w	ITC-irst	**0.235**	**0.076**	**0.317**	**0.216**
IBM2003c	IBM Research (Prager)	**0.298**	**0.077**	**0.175**	**0.212**
Albany03I2	University of Albany	**0.240**	**0.085**	**0.146**	**0.178**
FDUT12QA3	Fudan University	**0.191**	**0.086**	**0.192**	**0.165**
UAmsT03M1	University of Amsterdam	**0.136**	**0.054**	**0.315**	**0.160**
shef12simple	University of Sheffi eld	**0.138**	**0.029**	**0.236**	**0.135**
CMUJAV2003	Carnegie Mellon University (Javelin)	**0.133**	**0.052**	**0.216**	**0.134**
ICTQA2003C	Chinese Academy of Sciences (CAS-ICT)	**0.145**	**0.091**	**0.149**	**0.133**
uwbqitekat03	University of Wales, Bangor	**0.259**	**0.000**	**0.000**	**0.130**

Final score =1/2FactidScore+1/4ListScore+1/4DefScore

Bibliografia (1)

- *"Multiple-Engine Question Answering in TextMap "* A. Echihabi, U. Hermjakob, E. Hovy, D. Marcu, E. Melz, D. Ravichandran – Information Sciences Institute – Univiersity of Souhtern of California
- *"Knowledge-Based Question Answering"* – U. Hemjakob, E. Hovy, Chin-Yew Lin - Information Sciences Institute – Univiersity of Souhtern of California
- *"The use of External Knowledge in Factoid QA"* – U. Hemjakob, E. Hovy, Chin-Yew Lin - Information Sciences Institute – Univiersity of Souhtern of California
- *Tesi di Laurea di Anna Brinchi Giusti* – Corso di Laurea in Scienze della Comunicazione– Università degli studi di Siena

Bibliografia (2)

- *A new discipline for Information Access: An introduction to "Question Answering"* - Simon Sweeney (http://www.cis.strath.ac.uk/research/digest/rd_slides/InformationAccessQA.ppt)
- *"Question Answering Techniques and Systems"* – M. Surdeanu (TALP), M. Paşca (Google - Research)*
  TALP Research Center Dep. Llenguatges i Sistemes Informàtics Universitat Politècnica de Catalunya

*The work by Marius Pasca (currently mars@google.com) was performed as part of his PhD work at Southern Methodist University in Dallas, Texas.

Bibliografia (3)

- *"Overview of the TREC 2003 Question Answering Track"* – Ellen M. Voorhees
- *"Performance issues and error analysis in an Open-Domain Question Answering System"*
  D. Moldovan, M. Paşca, S. Harabagiu and M. Surdeanu -Language Computer Corporation
  http://portal.acm.org/citation.cfm?id=763694

§ http://www.trec.nist.gov
- http://acl.ldc.upenn.edu (A Digital Archive of Research Papers in Computational Linguistics)