QUESTION-ANSWERING SYSTEMS

Un esempio concreto:

FALCON

- Un QUESTION-ANSWERING SYSTEM è un sistema di recupero automatico delle informazioni, destinato a rispondere alle domande che gli sono poste nel linguaggio naturale.
- Diversamente dagli attuali motori di ricerca, i sistemi Q/A non ricercano interi documenti ma forniscono risposte specifiche situate in piccoli frammenti di testo. Questa loro funzionalità rende la ricerca molto più lenta, ma anche più precisa e puntuale.

I sistemi Q/A vengono vengono classificati in base al loro livello di precisione e raffinatezza in:

1. Slot-filling: Il sistema risponde a domande molto semplici che possono essere computate anche da un sistema IE (information extraction); Ciò che cambia è il modo di operare: mentre i sistemi IE usano metodi logici di estrazione, i sistemi Q/A usano queries ad hoc.
2. Limited-domain: Il sistema risponde a domande di media difficoltà legate a sfere specifiche di conoscenza.

3. Open-domain: sono i sistemi più complessi ed articolati che integrano assieme le tecniche di IE, IR (information retrivial) ed NLP.

Esempi di possibili domande alla portata di un sistema Q/A:

- Qual è l'altezza del monte Everest?
- Quand'è nata la televisione?
- Da quanti giocatori è composta una squadra di calcio?

Falcon: conoscenza d'amplificazione per i motori di risposta

- Falcon è un motore di risposta open-domain elaborato presso la Southern Methodist University di Dallas che unisce assieme diverse forme di conoscenza (sintattica, semantica e pragmatica) ed integra i metodi del NLP ad una complessa tecnologia d'amplificazione che si basa su nuovi approcci legati alla conoscenza pragmatica.

Principali caratteristiche di Falcon:

- Invece di operare a livello di ogni singola parola, Falcon opera al livello di dipendenza tra le parole, attraverso unificazioni libere delle loro forme semantiche.
- Presenza di un nuovo meccanismo di recupero del paragrafo che è in grado di modificare le parole chiave usate nella ricerca della risposta.
- Uso del metodo del "question reformulation".

QR QS EAT QK AP

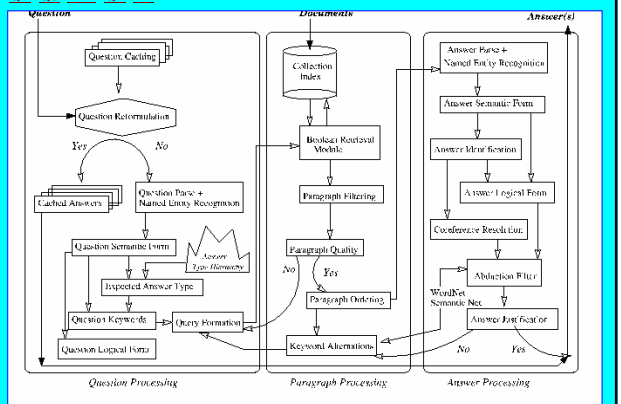


Figure 1: Question, Paragraph and Answer Processing in FALCON

Meccanismo di funzionamento

In generale la ricerca di una risposta è basata sulla supposizione che questa si trova in un paragrafo del testo che contiene gli elementi della domanda più rappresentativi. Viene utilizzato un modello di *recupero Booleano* dotato di un *meccanismo di filtro* che trattiene solo quei passaggi del testo che contengono il modello atteso di risposta.

- Question reformulation: La domanda posta al sistema viene inizialmente riformulata attraverso **l'algoritmo della somiglianza,** in maniera tale da renderla meglio analizzabile. La riformulazione della domanda può avvenire in tre modi:

a) Può essere trasformata la radice morfologica di una parola così che dalla domanda "WHO IS THE OWNER OF CNN?" otteniamo "WHO OWNS CNN?".

b) Una parola può essere sostituita con un suo sinonimo contenuto nella rete semantica di WordNet; così "HOW LONG IS HUMAN PREGNANCY?" può divenire "HOW LONG IS HUMAN GESTATION?"

c) Una parola può essere sostituita con un'iperonimo, così che "WHEN WAS BERLIN'S WALL ERECTED?" diviene "WHEN WAS BERLIN'S WALL BUILT?"

Le nuove domande ottenute vengono racchiuse in uno schema, chiamato **classe di riformulazione**, che può contenere da 2 a 8 domande.

Q397: *When was the Brandenburg Gate in Berlin built?*	
Q814: *When was Berlin's Brandenburg gate erected?*	
Q-411: *What tourist attractions are there in Reims?*	
Q-711: *What are the names of the tourist attractions in Reims?*	
Q-712: *What do most tourists visit in Reims?*	
Q-713: *What attracts tourists to Reims?*	
Q-714: *What are tourist attractions in Reims?*	
Q-715: *What could I see in Reims?*	
Q-716: *What is worth seeing in Reims?*	
Q-717: *What can one see in Reims?*	

Table 1: Two classes of TREC-9 reformulations.

Fig 1: Esempi di classi di riformulazione

- Question parse, named entity recognition: In questo passaggio la frase viene riconosciuta come una sequenza di tante entità, ad ognuna delle quali viene assegnata una certa struttura sintattica. La sola analisi della sintassi di superficie della frase non è però sufficiente per trovare il modello di risposta atteso, dato che la radice della domanda (how, what, who, when..) è associata a modelli di risposta differenti.

- Question semantic form: La risposta attesa può essere trovata solo conoscendo il contenuto semantico della domanda.

  Questo viene calcolato derivando tutte le dipendenze tra le parole, così da creare un diagramma di relazioni anonime che toccano ogni termine della domanda.

  Questo tipo di informazione è molto più importante di quella sintattica prodotta dagli analizzatori di frase.

Vantaggi nel processare forme semantiche:

- Permettono abduzioni basate sulle dipendenze tra parole piuttosto che su ogni singolo elemento.
- Facilitano la scoperta della risposta attesa poiché il nodo che nella rappresentazione è più connesso agli altri viene mappato nella gerarchia delle possibili risposte contenute in WordNet.
- Permettono al sistema di connettere nuovi concetti alle parole chiave utilizzate per il recupero della risposta.

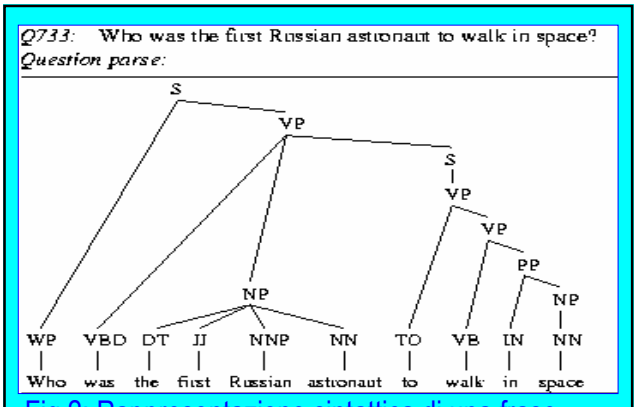
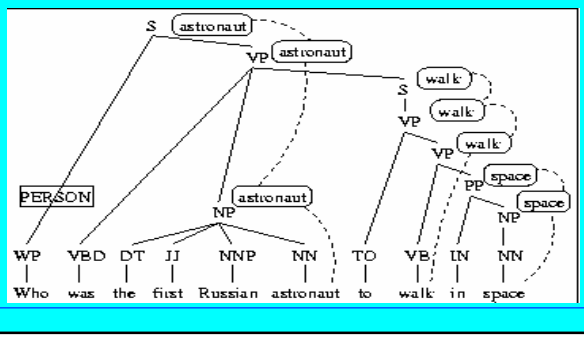


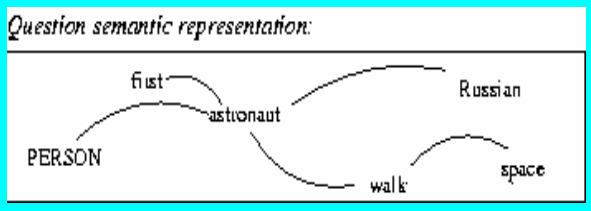
Fig 2: Rappresentazione sintattica di una frase presa come esempio

La rappresentazione semantica dell'albero della Fig 2 è ottenuta attraverso un'analisi chiamata **Parse tree traversal**. Fig 3



La Fig 4 mostra il risultato dell'analisi dell'albero. La rappresentazione semantica della domanda

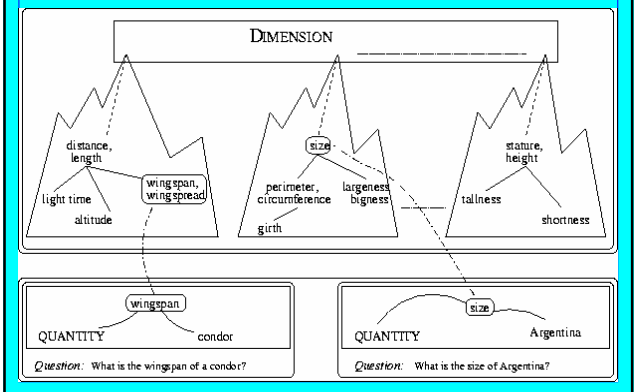
1. Comprende tutte le teste della frase.
2. Cattura le interrelazioni tra le teste con collegamenti anonimi. Fig 4



- Expected answer type: E' il passaggio in cui viene individuato il modello di risposta attesa. Il sistema è dotato di un riconoscitore di nomi (**named entity recognizer**) che copre 18 categorie, chiamate **Top Categories**, ognuna delle quali è collegata a numerose classi di parole contenute in WordNet. Fig 5

DATE	*TIME*	*ORGANIZATION*
REASON	*MANNER*	*NATIONALITY*
PRODUCT	*MONEY*	*LANGUAGE*
MAMMAL	*GAME*	*DOG BREED*
LOCATION	*REPTILE*	*NUMERICAL VALUE*
QUOTATION	*ALPHABET*	*PERCENTAGE*

Fig 6: Esempio di concetti contenuti in WordNet

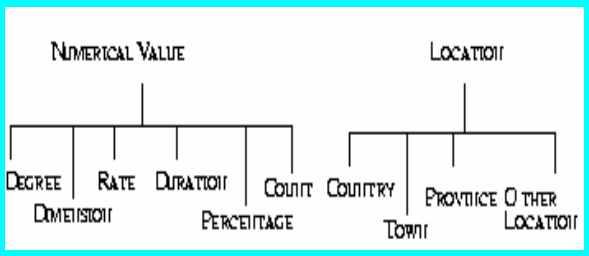


- **WordNet** è una rete semantica di concetti connessi gli uni agli altri e contenuti in una lista organizzata sottoforma di grafi di relazioni.

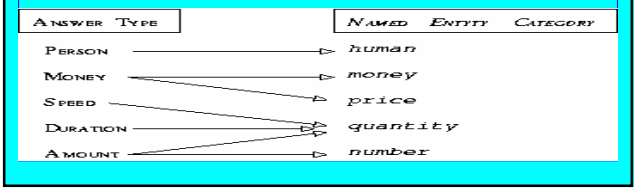
La Fig 6 mostra un esempio di come avviene l'individuazione del modello di risposta atteso relativo alla domanda " WHAT IS THE WINGSPAN OF A CONDOR?" (Quant'è l'apertura alare di un condor?).

La parola WINGSPAN è contenuta nella sottoclasse "DISTANZA" , la quale assieme a "MISURA" e ad

"ALTEZZA" fa parte di una categoria più grande chiamata "DIMENSIONE". Dimensione è a sua volta una sottoclasse della Top Category "NUMERICAL VALUE", cioè valore numerico. Fig 7



Una volta stabilito che la risposta a tale domanda dovrà essere un valore numerico, il termine "NUMERICAL VALUE" sarà incluso tra le parole chiave per la ricerca della risposta. La Fig 8 mostra invece come diverse risposte attese possono essere associate alla stessa categoria nome o viceversa.

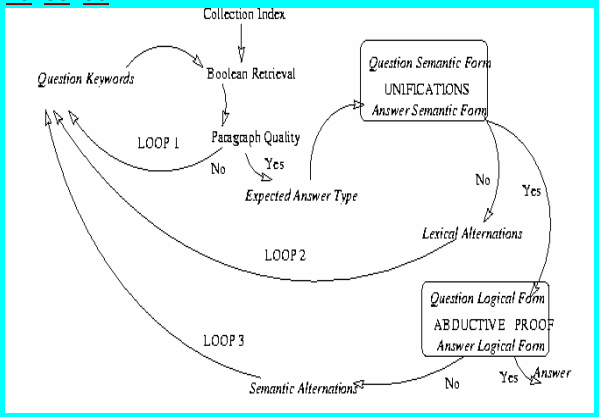


- Question Keywords: I paragrafi dei documenti che contengono le risposte attese, sono recuperati tramite l'utilizzo di parole chiave che vengono strutturate all'interno di una query ed inserite in un processo di ricerca booleana. Complessivamente la qualità dei paragrafi è misurata dal numero di volte che questi vengono nuovamente immessi nel ciclo booleano. Esistono dei limiti alla reiterazione dei paragrafi stabiliti a priori per ogni risposta attesa.

Se il recupero dei paragrafi è dentro i limiti,questi vengono preparati per essere processati nel modulo dell' Answer processing.

Se invece i paragrafi processati sono troppi, vengono aggiunte alla query nuove parole chiave in modo da affinare e migliorare la scelta; nel caso in cui siano troppo pochi, le parole chiave usate in precedenza vengono abbandonate: in entrambi i casi il ciclo viene ripetuto. Loop 1-Fig 9

29 35 36



I paragrafi che hanno superato il primo ciclo vengono ordinati ed inviati nel modulo dell' Answer processing: questo modulo identifica ed estrae la risposta dai paragrafi che contengono le parole chiave.

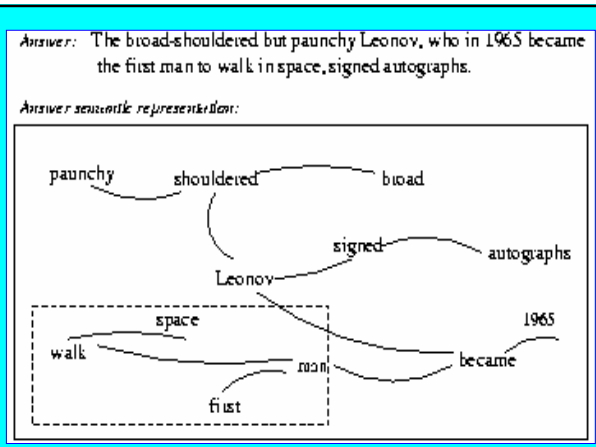
- Unification between question and answer semantic form

A questo punto il sistema unifica la forma semantica della domanda con la forma semantica della risposta attesa.

Tale unificazione avviene attraverso particolari *euristiche* che creano una corrispondenza tra le relazioni concettuali espresse nella domanda e le relazioni derivate dalla risposta.

Nella Fig 10 è rappresentata la forma semantica della risposta alla domanda "WHO WAS THE FIRST RUSSIAN ASTRONAUT TO WALK IN SPACE?" (Fig 2)

Il rettangolo tratteggiato indica il risultato dell'unificazione: nessuno di quei concetti rappresenta la risposta attesa, cioè Leonov.



Quando l'unificazione delle forme semantiche non porta al recupero della risposta, il sistema allora effettua delle modifiche di natura lessicale e morfologica sulle parole chiave, prima di immetterle nuovamente nel ciclo booleano.

Loop 2-Fig 9

a) variazioni Morfologiche: (es.) Per rispondere alla domanda "WHO INVENTED THE PAPER CLIP?" Falcon mappa il verbo *invented* all'interno del nome *inventor*.

Quest'ultimo si trova a sua volta nella sottogerarchia della risposta attesa *Person*. La query che viene passata nel recupero booleano sarà perciò

QUERY: [paper AND clip AND (invented OR inventor OR invent)]

b) variazioni Lessicali: Consistono nel sostituire le parole chiave con dei sinonimi. (es.) In "WHO KILLED J.F.KENNEDY?" il sistema considera, come sinonimo di killer, il verbo *assassin*.

- Abductive proof, Answer justification:

Una volta recuperata, la risposta viene estratta solo se può essere provata una giustificazione logica della sua correttezza. A questo scopo le forme semantiche della domanda e della risposta sono trasformate in **Forme Logiche**.

Attraverso l'euristica **LFT** (Logical Form Transformation) il sistema genera le **QLF** (Question Logical Formulae) e le **ALF** (Answer Logical Formulae).

- *Costruzione di una forma logica*: Fig 11

a) I verbi sono rappresentati come verbi predicati (e,x,y,z…) con la convenzione che:

e rappresenta l'eventualità che un'azione o un evento abbia luogo, mentre gli argomenti **x y z** rappresentano i predicati del verbo.

b) I nomi sono mappati nello stesso modo dei predicati a cui si riferiscono.

Answer: The broad-shouldered but paunchy Leonov, who in 1965 became the first man to walk in space, signed autographs.

Answer logic form:

paunchy(y) ^ shouldered(e1 y x) ^ broad(x) ^ Leonov(x) ^ first(z) ^
^ man(z) ^ space(t) ^ walk(e2 t z) ^ became(e3 z u x) ^
^ 1965(u) autographs(v) ^ signed(e4 v x) ^ HUMAN(x) ^ DATE(u)

Q733: Who was the first Russian astronaut to walk in space?

Question semantic representation:

Question logic form:

first(x) ^ astronaut(x) ^ Russian(x) ^ space(z) ^ walk(y z x) ^
^ HUMAN(x)

- Le trasformazioni in forma logica unite al complesso sistema di filtraggio delle risposte scorrette, permettono al sistema di essere molto veloce e preciso nel recupero della risposta attesa.
- Quando non è giustificata nessuna risposta, poiché il numero di relazioni di coreferenza tra le due forme logiche è nullo, il sistema immette nuovamente nel ciclo le parole chiave, dopo averle modificate dal punto di vista semantico.

- variazioni Semantiche: (es.) "WHERE DO LOBSTERS LIKE TO LEAVE?" Il verbo *like* viene mappato in WordNet come *prefer*, così che la query sarà:
  QUERY: [(lobster OR lobsters) AND (like OR prefer)].

  Loop 3-Fig 9

- Quando avviene la giustificazione di correttezza della risposta, questa viene estratta ed il sistema si arresta fino ad una nuova interrogazione.
  Fig 9

Conclusioni

E' stato ampiamente provato che Falcon è in grado di rispondere in maniera corretta al 79% delle domande che gli sono poste nel linguaggio naturale. Le ottime performances di Falcon sono in gran parte da attribuire al sistema di ricerca e recupero della risposta.

Questo complesso sistema utilizza tre diverse forme di conoscenza che determinano un modello di ricerca booleana basato su tre cicli.Tale tecnologia apre nuovi spazi di ricerca ad ogni reiterazione dei paragrafi, poiché ogni nuovo ciclo prevede una ricerca sempre più sofisticata e puntuale.

UNIVERSITA' DEGLI STUDI DI SIENA
a.a 2002/2003
Modulo di linguistica Computazionale
Prof. Amedeo Cappelli
di: Chiara Brinchi Giusti