UNIVERSITÀ DEGLI STUDI DI ROMA
"TOR VERGATA"

Facoltà di Ingegneria

Corso di Laurea Specialistica in Ingegneria Informatica

Progetto per il corso di Intelligenza Artificiale

SISTEMA DI QUESTION ANSWERING

Professore:
Prof. MARIA TERESA PAZIENZA
Correlatore:
Prof. ZANZOTTO - STELLATO

Studente:
STEFANO PERNA

ANNO ACCADEMICO 2008-2009

Prefazione

La seguente relazione è stata prodotta come parte integrante del progetto svolto nell'ambito del corso di Intelligenza Artificiale.

Lo scopo del progetto è quello di realizzare un sistema di Quetion Answering basato su Passage Retrival in grado di integrarsi con gli altri moduli del sistema per completarne le funzionalità.

Nel modulo presentato non sono presenti le funzionalità necessarie all'interazione con l'utente (il BOT), ma solo tutti i meccanismi necessari per processare le doamande dall'utente del sistema e fornire delle risposte.

Indice

Capitolo 1

Introduzione

L'accesso alle informazioni sta diventando un problema sempre più critico con il passare del tempo.
Le collezioni di dati disponibli sono sempre più grandi, le informazioni sono spesso duplicate, inesatte e non ben accessibili direttamente dall'utente.

I tradizionali sistemi di *Information Retrieval* aiutano l'utente nell'accesso all'informazione, ma spesso sono confusionali e poco pratici per la ricerca mirata e l'estrazione dell'informazione cercata.

Gli attuali motori di ricerca ne sono un esempio: ricercare delle informazioni sul web diventa spesso uno sforzo duplice da parte dell'utente, che dovrà prima capire come formulare la propria richiesta (query) nel modo più consono al reperimento delle informazioni, ed a seguito dei risultati proposti dovrà effettuare nuove ricerche manuali all'interno dei documenti trovati per estrarre l'informazione cercata.

Con il passare del tempo si sta verificando un fenomeno sempre più dilagante: non sono più i sistemi ad adattarsi alle nostre esigenze, ma sono gli utenti ad adattarsi alle esigenze del sistema.
In un certo qual modo l'utente viene addestrato dal sistema.

Come automatizzare il meccanismo e semplificare il reperimento dell'infor-

mazione cercata?

I sistemi di *Question Answering* (QA) risolvono in parte questi problemi, ponendosi come approccio alternativo ai tradizionali sistemi di *Information Retrieval* (IR) e di *Information Extraction* (IE).

1.1 Quali motivazioni ci spingono verso il QA?

Il crescente accesso delle informazioni presenti sul web da parte di un pubblico più ampio e non omogeneo ha creato nuove necessità di studio su come avvengono gli accessi alle informazioni disponibili.

I comuni sistemi di IR cercano di reperire i documeti interessanti basandosi sulle query formulate dagli utenti senza una profonda analisi del significato delle query stesse, ma usando principalmente approcci statistici.

Possiamo immaginare come l'efficacia del sistema aumenti scegliendo query composte dalle sole parole chiave necessarie.

I risultati sono forniti inoltre come una lista di documenti tra i quali l'utente dovrà cercare a sua volta (manualmente) l'informazione desiderata.

Molti utenti preferiscono però formulare le proprie query sotto forma di domande, e non come parole chiave, desiderando ricevere inoltre un'informazione coincisa alla propria domanda, senza necessità di dover leggere l'intera lista di documenti proposti dai sistemi di IR.

Oggi circa il 15% delle query formulate nei motori di ricerca sono sotto forma di domanda. Tale percentuale crescerà con il tempo[1].

Ma come reagiscono i motori di ricerca a query formulate come domande?

Nella figura 1.1 è riportato un esempio di come un mototre di ricerca spesso reagisce a domande formulate in linguaggio naturale.

L'interesse nel creare dei sistemi di QA funzionali è molto elevato, e la necessità di questi sistemi è in forte crescita.

[1]Studi effetuati da Yahoo Inc.

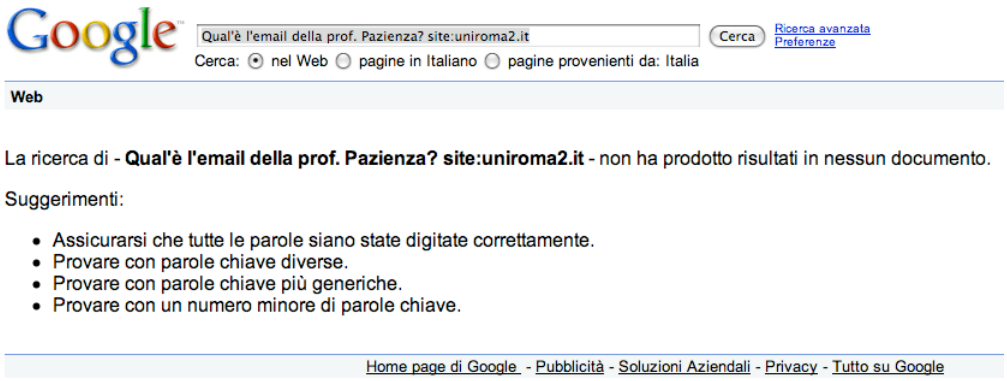

Figura 1.1: Query in linguaggio naturale su Google.

Con il passare del tempo sarà un punto critico poter presentare agli utenti delle informazioni precise. Gli interessi industriali e commerciali verso queste tecnologie sono molto forti.

1.2 Dai sistemi di IR al QA

I motori di ricerca più popolari (come Google, Yahoo, Altavista) utilizzano approcci puramente statistici per il retrival dei documenti, non fornendo una vera e propria informazione diretta poichè non si tenta di interpreatre il senso della query o dei documenti stessi.

L'obiettivo di questo approccio non è infatti quello di fornire una vera e propria informazione, ma viene sostituito tale concetto con quello di documento.

Il risultato di una query in un motore di ricerca è spesso una lista di documenti, ordinati per rilevanza, in base a metriche statistiche come la frequenze delle parole nella query, nel documento o nella collezione.

I sistemi di QA possono fornire l'informazione effettivamente cercata, senza dover presentare all'utente liste di documenti spesso irrilevanti ai fini del reperimento dell'informazione stessa.

Figura 1.2: Query in linguaggio naturale su Google.

Un tecnica largamente adottata per realizzare sistemi di QA consiste nello sfruttare le possibilità offerte dall'IR per il reperimento dei documenti rilevanti ai fini dell'estrazione dell'informazione combianata con tecniche di IE per estrarre l'informazione cercata.

Come visto, nei sistemi di IR le query sono considerate semplicemnte a fini statistici senza nessuna comprensione del significato della stessa o dei concetti in essa espressi.

Nei sistemi di QA la query è invece analizzata per cercare di capirne il significato e per supportare l'estrazione dell'informazione cercata.

Non è da nascondere il fatto che comunque anche i sistemi di QA (come i sistemi di IE) non sono in grado di comprendere il testo nella sua interezza. La comprensione è infatti spesso limitata all'organizzazione delle informazione e all'estrazione delle stesse tramite template o metodi statistici.

Capitolo 2

Passage Retrival Question Answering

Il Question Answering (QA) si prefigge l'obiettivo di trovare la risposta ad una domanda posta in linguaggio naturale, partendo da un insieme di documenti.

I classici sistemi di QA utilizzano varie tecniche per poter fornire le risposte. Tra le più comuni troviamo l'uso di sistemi di document retrieval supportati da tecniche di analisi delle domande e tecniche per l'estrazione delle risposte, consentendo di selezionare un insieme di risposte candidate, tra le quali sarà reperita la risposta (o le risposte) alla domanda.

La suddivisione del tipo di domande in categorie distinte ci consente l'uso di strategie mirate per affrontare il problema.

Le principali categorie di domande possono essere riassunte in: *factoid questions* e *complex questions*.

Le domande centrate sui fatti (*factoid questions*) richiedono un semplice fatto come risposta (come un numero di telefono, il nome di una persona, o un'email), e sono relativamente più semplici da trattare.

Le domande complesse (*complex questions*) richiedono invece risposte complesse e spesso più lunghe, che possono rappresentare sia fatti che relazioni o processi (come per esempio chiedere *cosa* è una mela, e non semplicemente

quanto pesa).

Molti sistemi di QA sono in grado di rispondere alle *factoid questions* usando risorse ontologiche.

Il sistema presentato in questo progetto utilizzarà un approccio più statistico, basato sull'accesso ad una larga collezione di documenti, e mediante l'ausilio di diverse tecniche di IR ed IE combinate per estrarre l'informazione cercata.

2.1 Obiettivi del sistema

L'obiettivo del sistema è quello di fornire una risposta ad una determinata domanda.

Il sistema appoggierà il sistema esistente di QA nel caso in cui non fosse possibile reperire la risposta dall'ontologia utilizzata.

Il BOT del sistema, nel caso di fallimento del reperimento della risposta dall'ontologia, interrogherà il sistema di Passage Retrival Question Answering (PRQA) passandogli la domanda posta dall'utente.

Come visto le domande possono essere divise in due categorie: *factoid questions* e *complex questions*.

Le due categorie di domande saranno trattate in modo diverso, fornendo tipologie di risposte differenti.

Per ogni domanda sottoposta al sistema si procederà all'analisi del tipo di domanda cercando di capire la natura della stessa e se è possibile rispondere con un fatto.

Nel caso non sia possibile rispondere semplicemente con un fatto, il sistema tenterà di rispondere presentando la lista dei documenti rilevanti che possono contenere la risposta complessa cercata.

In un esempio più pratico, alla domanda "*Qual'è l'email della Prof. Pazienza?*" il BOT analizzerà che la domanda richiede come risposta un fatto: l'**email** della Prof. Pazienza.

Dalla domanda verranno estratti i due termini da utilizzare per l'interrogazione dell'ontologia (**email**, **Pazienza**) e la loro relazione.

Cosa succede se l'interrogazione dell'ontologia non porta a nessun risultato o se il Bot non è in grado di estrarre i termini necessari per interrogare l'ontologia?

Il fallimento dell'interrogazione dell'ontologia comporterebbe il fallimento dell'intero sistema, in quanto non saprebbe come rispondere all'utente.

Il sistema PRQA interviene in questi casi: il BOT passa la domanda dell'utente al sistema PRQA, il quale si farà carico di dare una risposta.

La domanda posta dall'utente verrà analizzata nuovamente mediante metodi statistici per valutare se è possibile collocarla in una categoria di fatti conosciuti dal sistema. Individuata la categoria della domanda si potrà procedere al reperimento della risposta.

In entrambi i casi il sistema avrà a disposizione tutti i documenti del web dell'ateneo per poter cercare di estrarre l'informazione e nel caso fallisse nel trovarla può presentare una lista di documenti ritenuti rilevanti ai fini della risposta.

2.2 Architettura del sistema

Il sistema presentato si occupa di svolgere molti compiti del QA: a partire dall'analisi della domanda, al reperimento dei documenti interessanti, alla loro analisi ed all'estrazione della risposta.

L'architettura del sistema presentato è suddivisibile in diversi processi:

1. *Question Analysis* - L'analisi della domanda.

2. *Document Retrieval* - Il reperimento dei documenti.

3. *Document Processing* - L'analisi dei documenti.

4. *Answer Extraction* - L'estrazione delle risposte.

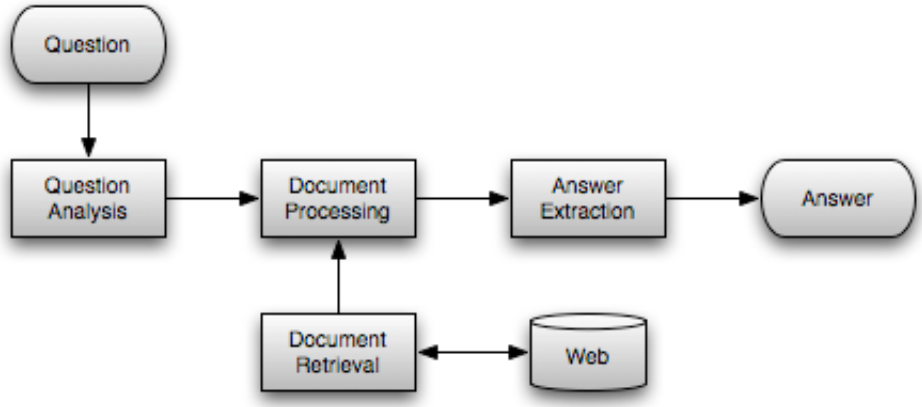

Figura 2.1: Architettura del sistema.

Ogni elemento svolge delle funzione specifiche all'interno del sistema e coopera con gli altri elementi per raggiungere gli obiettivi preposti.

Il modulo di *Question Analysis* si occupa di analizzare la domanda per cercarne di capirne il contesto ed estrarre i termini necessari alla formulazione della query utilizzata per il reperimento dei documenti rilevanti.

I moduli di *Document Retrieval* e *Document Processing* si occupano rispettivamente di reperire l'insieme dei documenti rilevanti e la loro analisi.

Il modulo di *Answer Extraction* si occupa di estrarre l'informazione dai documenti recuperati, a partire dal tipo di domanda presentata.

Nelle sezioni successivi sono riportati i processi nel dettaglio.

2.2.1 Question Analysis

L'analisi della domanda è uno dei punti chiavi del sistema. La buona riuscita di questo processo concorrerà al raggiungimento degli obiettivi preposti.

Il processo può essere diviso in due fasi: il *Question Classification* ed il *Query Extraction*.

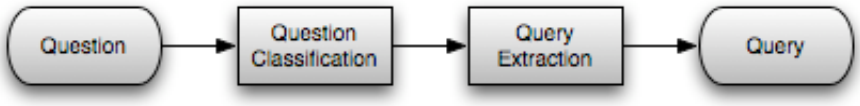

Figura 2.2: Processo d'analisi della domanda.

La classificazione della domanda consente di identificare la categoria semantica della domanda ed identificare il tipo di risposta richiesta.

Le categorie semantiche conosciute dal sistema sono raccolte in una collezione di di domande pre-impostate con il quale il sistema verrà addestrato al riconoscimento.

Il tipo di risposta sarà quindi relazionato alla categoria semantica della domanda.

In generale da un insieme di domande possiamo aspettarci delle risposte ben precise (dei fatti o delle risposte complesse).

In tabella sono riportai alcuni esempi:

Question	**Stem**	**Answer Type**
Chi è il direttore del dipartimento?	Chi	PERSON
Qual'è il numero di telefono della segreteria?	Quale	PHONE
Dove si trova il dipartimento di matematica?	Dove	ADDRESS

Per ogni domanda è possibile individuare degli stem (Chi, Quale, Quanto, Dove...) che possono dirci molto sul tipo di risposta attesa.

Per ogni classe di domanda sarà quindi possbile identificare un tipo di risposta ben specifico.

In tabella è possibile per esempio vedere come allo stem "**Chi**" è possibile associare la risposta del tipo **PERSON**.

Spesso però gli stem delle domande possono essere ambigui.

Consideriamo due domande: "*Qual'è l'email della segreteria?*" e "*Qual'è il nome del rettore?*".

In entrambe le domande lo stem è lo stesso ("**Quale**"), ma la risposta attesa è ben differente.
Nel primo caso si cerca un indirizzo email (il fatto **EMAIL**), nel secondo il nome di una persona (**PERSON**).

La semplice individuazione dello stem non è quindi sufficiente, bisogna anche disambiguare il tipo di risposta richiesto.

Il sistema proposto cerca di risolvere il problema mediante un approccio di Machine Learning.

Nel sistema è integrata una macchina ad apprendimento supervisionato addestrata con un collezione di domande di cui si conosce il tipo di risposta richiesta.

Per ogni nuova domanda presentata al sistema la macchina tenterà di predire la categora della risposta richiesta usando dei Decision Trees[1].

Se la categoria d'appartenenza della domanda non fosse predicibile, la macchina assegnerà una categoria vuota.

La seconda fase del processo consiste nell'estrazione dei termini necessari per la formulazione della query.

L'estrazione dei termini dalla domanda avviene analizzando le caratteristiche semantiche della domanda.

La domanda viene processata mediante un analizzatore morfo-sintattico in grado di individuare le part-of-speech (POS) delle parole.

Le parole che non contraddistingueranno la domanda (come per esempio articoli, avverbi, etc...) saranno scartate. Le parole distintive (come nomi o particolari verbi) saranno invece individuate per la formulazione della query.

[1] Vedi Appendice A

2.2.2 Document Retrieval e Processing

Il processo di Document Retrieval e Processing si colloca nella seconda fase dell'intero processo. L'obiettivo è quello di recuperare tutti i documenti interessanti partendo dalla query formulata.

Possiamo identificare 3 fasi principali in questo processo:

1. *Query Formulation* - Formulazione della query a partire dalle parole chiave.
2. *Document Retrieval* - Il reperimento dei documenti.
3. *Document Processing* - L'analisi dei documenti.

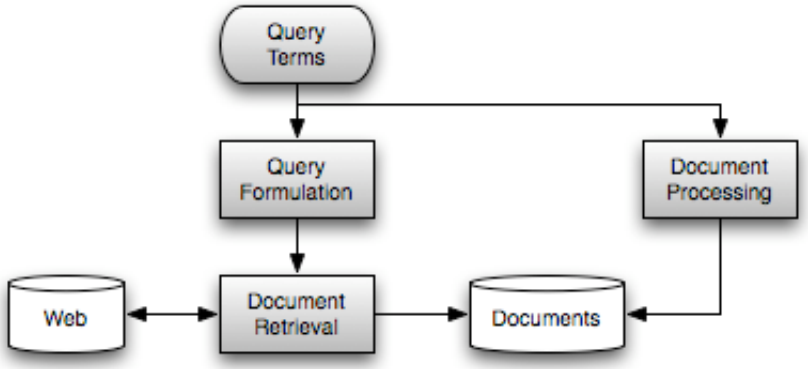

Figura 2.3: Processo per il reperimento dei documenti.

Le query sono formulate a partire dalle parole chiave a disposizione.

Le parole chiave verranno combinate con sequenze di AND e di OR in base alle necessità.

Stabilite delle soglie per il numero di documenti recuperati, possiamo aggiustare le query se il numero di risultati fosse troppo elevato o troppo basso.

Recuperare un numero di risultati troppo alto è probabilmente conseguenza di una query non troppo selettiva, e questo comporterebbe una maggiore difficoltà nel reperire l'informazione cercata. Possiamo quindi rendere la query più selettiva aumentando o concatendando le parole chiave in modo diverso.

Ottenere invece un numero di documenti troppo basso è sintomo di una query troppo selettiva. Bisognerebbe quindi rilassare i vincoli della query per permettere di recuperare un numero di documenti uguale o superiore alla soglia che riteniamo minima.

La tecnica di Document Retrieval utilizzata è basata sul Passage Retrieval.

Per ogni documento trovato dal sistema vengono estratti dei *passage* (finestre di testo) rilevanti, nei quali si spera sia contenuta la risposta cercata.

Il sistema di IR ritorna tutti i documenti associati alla query, e per ogni docmento vengono estratte delle porzioni di testo interessanti.

Per ogni *passage* viene calcolato il Passage Rank, uno valore di riferimento che ci consentirà di decidere quali passage siano più rilevanti rispetto agli altri.

Le metriche per calcolare lo score (rank) possono essere diverse.

Comunemente viene calcolato:

- *Same Word Sequence Score* - il numero di parole chiave della query trovate nello stesso ordine all'interno della finestra.
- *Unique Word Score* - il numero di parole chiave uniche (non ripetute) trovate all'interno della finestra.
- *Distance Score* - il numero di parole intercorrenti tra le due più distanti parole chiave presenti nella finestra.

Considerata la query contenente i termini "**email Einstein**". Tra i documeti recuperati possiamo trovare per esempio la finestra:

*"Albert **Einstein** tel: +123456789 **email**: albert@einstein.com. I risultati del corso di fisica..."*

In grassetto sono evidenziati i termini presenti nella query.

Calcolando gli score avremo che: il *Same Word Sequence* ha score pari a 0 (i termini della query non sono presenti nello stesso ordine all'interno della

finestra), l'*Unique Word Score* ha valore pari a 2 (entrambi i termini della query sono presenti) e il *Distance Score* ha volore pari a 2 (sono presenti 2 parole, "tel" e "+123456789", tra le parole chiave della query).

L'approccio utilizzato nel sistema presentanto combina entrambi gli score per fornire un raking più accurato.

In questo modo possiamo ottenere un'insieme ordinato di documenti rappresentati dai propri *passage*, che potranno essere utilizzati per le fasi successive di estrazione dell'informazione cercata.

2.2.3 Answer Extraction

L'ultima fase del processo è l'Answer Extraction: si cercherà di estrarre l'informazione cercata e di fornire una risposta.

Nella fase di Answer Extraction verranno analizzati tutti i *passage* recuperati precedentemente e verranno estratte le informazioni in base al tipo di risposta richiesto.

In questa fase abbiamo già a disposizione il tipo di domanda e l'insieme dei documenti.

Le fasi del processo consisteranno quindi nell'estrazione delle risposte candidate e nella selezione della risposta scelta.

1. *Answer Extraction* - Estrazione delle risposte candidate.
2. *Answer Selection* - Selezione della risposta.

Come visto il tipo di domande proposte al sistema può essere suddiviso in due categorie: domande che richiedono fatti o domande complesse.

Nelle precedenti fasi il sistema è stato in grado di comprendere il tipo di domanda proposta e il tipo di risposta attesa ed ha recuperato una lista di

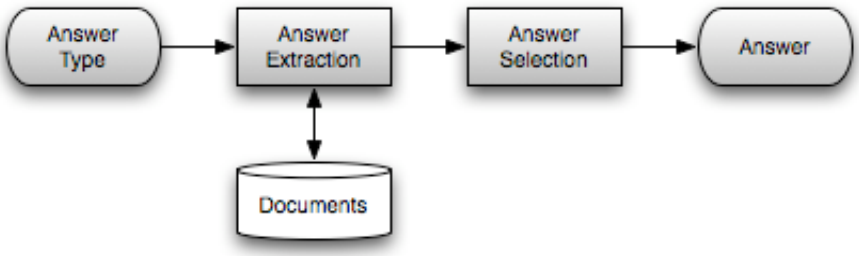

Figura 2.4: Processo di estrazione della risposta.

documenti dai quali estrarre l'informazione.

Conoscendo il tipo di risposta richiesta (l'Answer Type: PERSON, PHONE, ADDRESS...) è possibile definire dei pattern per estrarre l'informazione dai documenti.

Ogni informazione estratta rappresenterà una risposta candidata (*answer candidate*) alla domanda.
Le answer candidate saranno poi soggette ad un ranking per definire quale sarà la risposta finale da fornire.

L'operazione di ranking è basata su euristiche legate alla Candidate-Query Distance.
Per ogni risposta viene conteggiato il numero di parole che la separa dalle parole chiave della query.

Il calcolo della distanza può essere effettuato utilizzando la Mean Square Distance:

$$dist(w_k, Q, P_i) = \sqrt{\sum_{t=0}^{numterms \in Q} dist(w_k, w_{Q_t}, P_i)^2}$$