Studio di approcci statistici al problema del Question Answering in Information Retrieval.

Università degli Studi di Roma Tor Vergata
Facoltà di Scienze Matematiche Fisiche e Naturali
Corso di Laurea Specialistica in Informatica.

Relatore
Prof. Giorgio Gambosi
Correlatore
Dr. Giambattista Amati
Correlatore
Dott. Luca Ferri

Candidato
Valerio Capozio

Anno Accademico 2007/2008

Sommario

Ambito

- Information Retrieval: Question Answering.

Obiettivi

- Studio di soluzioni alternative a quelle semantiche.
- Studio di soluzioni volte ad aumentare la early-precision del sistema.

Approcci Valutati

- Query Expansion tramite risorse esterne.
- Utilizzo dei bigrammi significativi.

Strumenti di IR

- Terrier.

Sommario

Ambito

- Information Retrieval: Question Answering.

Obiettivi

- Studio di soluzioni alternative a quelle semantiche.
- Studio di soluzioni volte ad aumentare la early-precision del sistema.

Approcci Valutati

- Query Expansion tramite risorse esterne.
- Utilizzo dei bigrammi significativi.

Strumenti di IR

- Terrier.

Sommario

Ambito

- Information Retrieval: Question Answering.

Obiettivi

- Studio di soluzioni alternative a quelle semantiche.
- Studio di soluzioni volte ad aumentare la early-precision del sistema.

Approcci Valutati

- Query Expansion tramite risorse esterne.
- Utilizzo dei bigrammi significativi.

Strumenti di IR

- Terrier.

Sommario

Ambito

- Information Retrieval: Question Answering.

Obiettivi

- Studio di soluzioni alternative a quelle semantiche.
- Studio di soluzioni volte ad aumentare la early-precision del sistema.

Approcci Valutati

- Query Expansion tramite risorse esterne.
- Utilizzo dei bigrammi significativi.

Strumenti di IR

- Terrier.

Il Question Answering

Il Question Answering è uno dei tanti campi di applicazione delle tecniche di Information Retrieval.

I sistemi sviluppati in tale ambito tentano di rispondere automaticamente a domande poste dall'utente in linguaggio naturale.

In letteratura si è soliti suddividere i sistemi di Question Answering in due categorie:

- *Closed Domain*
- *Open Domain*

Scopo di questa tesi è studiare possibili approcci statistici applicabili al problema del Question Answering

Vicky, il sistema di QA dell' INPS

Vicky è un avatar virtuale sviluppato da AImavivA per conto dell'INPS.

Vicky ha il compito di accogliere gli utenti all'interno del sito dell'INPS e di rispondere alle domande che questi possono porre.

Le domande cui Vicky può tentare di fornire risposta sono inerenti ai temi della previdenza complementare e dei lavoratori domestici.

Per queste ragioni Vicky può essere classificato come un sistema Closed Domain di Question Answering.

Per funzionare Vicky sfrutta una base di conoscenza composta di FAQ, ovvero coppie $<$*domanda*, *risposta*$>$.

Metodologia di test

Per valutare l'impatto avuto dalle sperimentazioni sulle prestazioni del sistema è stato definito un ambiente di test costituito da:

TOPIC Un insieme di query da sottomettere al sistema ad ogni test.

QREL Un insieme di valutazioni di rilevanza dei documenti rispetto alle TOPIC.

BASELINE Risultati ottenuti valutando le TOPIC sottomesse al sistema senza applicare tecniche o soluzioni atte a migliorare il recupero.

Preparazione al calcolo della baseline

Le TOPIC e le QREL sono state realizzate sfruttando le informazioni contenute all'interno delle FAQ e dei LOG.

Ottenute TOPIC e QREL sono state calcolate 3 diverse baseline, basate su configurazioni distinte delle query sottoposte.

L'indice utilizzato nel corso della sperimentazione è stato ottenuto attraverso l'indicizzazione dei campi TITLE e BODY delle FAQ con l'applicazione di tecniche di stemming e stopword list. Tale indice è identificato dalla sigla TITLE+BODY-1.1.

Il modello di pesatura

Per il calcolo della baseline, e nel resto della sperimentazione è stato utilizzato il modello Divergence From Randomness.

I modelli di DFR basano la loro analisi sull'idea che maggiore è la divergenza della frequenza di un termine in un documento, rispetto a quella nella collezione, maggiore sarà l'informazione che quel termine porterà all'interno del documento stesso.

Il peso di un termine risulta inversamente proporzionale alla probabilità che la sua frequenza nel documento sia quella derivata da un modello di casualità:

$$weigth(t \mid d) \propto -\log\left(Prob_M(t \in d \mid Collection)\right)$$

dove M indica il modello generativo utilizzato per calcolare la probabilità.

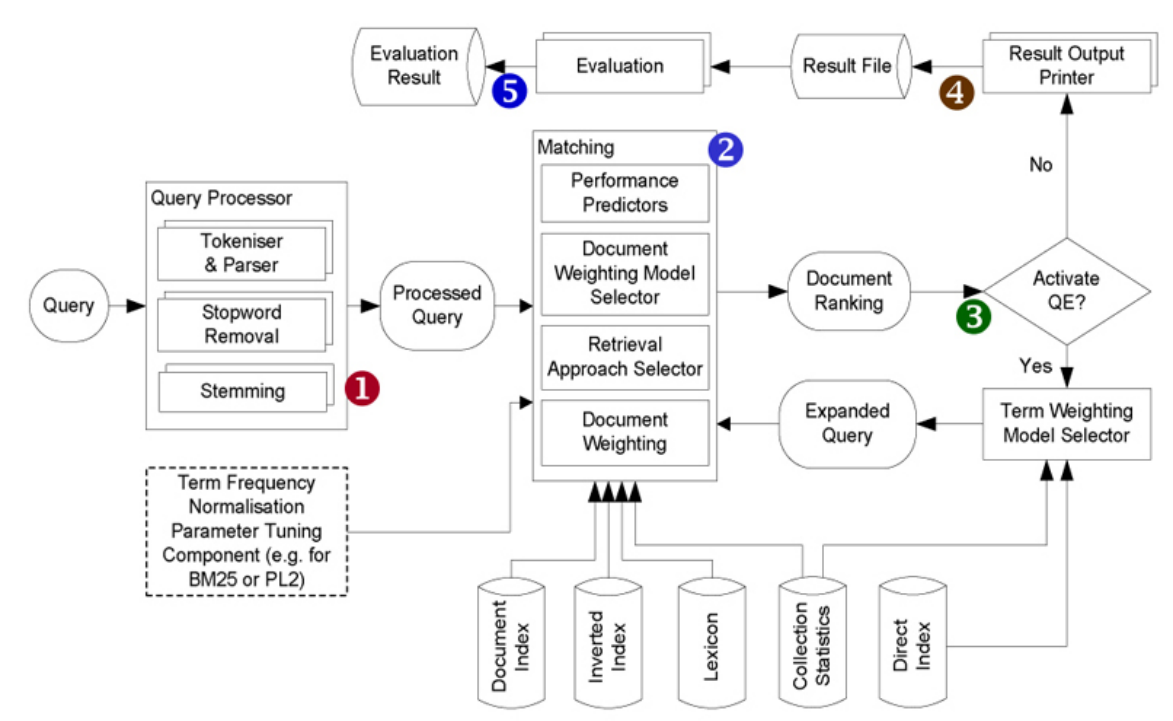
Le baseline

Nella tabella seguente sono mostrate le baseline ottenute applicando all'indice TITLE+BODY-1.1 le query estratte dalle sole FAQ, le query estratte dai soli LOG e le query ottenute sia dai LOG che dalle FAQ.

	FAQ	LOG	FAQ+LOG
Number of query	110	77	187
Retrieved	327	330	657
Relevant	110	91	201
Relevant retrieved	106	22	128
Average Precision	**0.8939**	**0.1797**	**0.5998**
R-Precision	0.8364	0.1342	0.5472
Precision at 1	0.8364	0.1558	0.5561
Precision at 2	0.4636	0.1169	0.3209
Precision at 3	0.3212	0.0952	0.2282
Precision at 4	0.2409	0.0714	0.1711

Cos'è e come si usa di solito

La Query Expansion è un processo atto a riformulare la query sottomessa al fine di migliorare le prestazioni del recupero. La tecnica di QE più utilizzata è la **pseudo-relevance feedback**.



Tipicamente la QE viene utilizzata per aumentare la **recall** del sistema.

Come vorremmo utilizzarla noi

Nel nostro caso tenteremo di utilizzare la QE al fine di migliorare non tanto la recall, come vorrebbe la teoria, quanto la **precision** del sistema.

Questo obiettivo può rivelarsi piuttosto ambizioso visto che spesso recall e precision risultano inversamente proporzionali.

Per tentare di conseguire questo obiettivo utilizzeremo delle risorse esterne per effettuare l'espansione.

Motivazioni della scelta

La collezione di riferimento risulta estremamente piccola e presenta un quantitativo di termini molto ridotto.

Questa caratteristica può creare diversi problemi nel momento in cui si decida, come nel nostro caso, di utilizzare approcci di tipo statistico.

La scelta di utilizzare risorse esterne al corpus delle FAQ è quindi mirata a sopperire alle carenze della collezione target.

Il modello di QE utilizzato

Il modello di QE utilizzato nell'arco della sperimentazione è il Bo1.

Il modello è basato sulla statistica di Bose-Einstein:

Bo1

$$w(t) = tf_x * \log_2 \left(\frac{1 + P_n}{P_n} \right) + \log(1 + P_n)$$

dove tf_x è la frequenza del termine della query nei primi x documenti della lista e P_n è dato da $\frac{F}{N}$ dove F è la frequenza del termine della query nell'intera collezione e N è il numero di documenti della collezione.

La collezione dell'INPS

La collezione esterna, da utilizzare nel processo di QE, è stata ottenuta attraverso il crawling del sito dell'INPS.

Effettuando il crawling del sito dell'INPS siamo stati in grado di costruire una collezione di 14523 documenti con 31615 termini unici.

I risultati ottenuti a seguito di tale sperimentazione sono stati confrontati con la baseline FAQ+LOG e con i dati della medesima baseline a seguito dell'applicazione della QE classica.

QE sulla collezione dell'INPS: analisi dei risultati

	BaselineFL	QE-d3-t10	I-d10-t2
Number of query	187	187	187
Retrieved	657	661	**667**
Relevant	201	201	201
Relevant retrieved	128	124	**129**
Average Precision	0.5998	0.5446	0.6052
R-Precision	0.5472	0.4750	0.5526
Precision at 1	0.5561	0.4813	0.5615
Precision at 2	0.3209	0.2834	0.3235
Precision at 3	0.2282	0.2210	0.2299

- MAP $+10\%$, P@1 $+16\%$ rispetto alla baseline con QE.
- MAP $+0.9\%$ P@1 $+1\%$ rispetto alla baseline QT-FL.
- QE su risorse esterne è percorribile, ma la collezione INPS non va bene.

La collezione di Google

Dovendo recuperare un buon numero di documenti inerenti ai temi trattati nelle FAQ abbiamo pensato di sfruttare Google.

La collezione ottenuta in questo modo è costituita di 527 documenti e ben 24450 termini unici.

In questo caso i risultati sono stati confrontati con quelli utilizzati ed ottenuti nella precedente sperimentazione.

QE sulla collezione di Google: analisi dei risultati

	Baseline	QEd3-t10	Id10-t2	Gd7-t5	IGd3/7-t3/2
Number of query	187	187	187	187	187
Retrieved	657	661	667	**660**	667
Relevant	201	201	201	201	201
Relevant retrieved	128	124	129	**135**	125
Average Precision	0.5998	0.5446	0.6052	**0.6185**	0.5882
R-Precision	0.5472	0.4750	0.5526	0.5633	0.5365
Precision at 1	0.5561	0.4813	0.5615	**0.5668**	0.5455
Precision at 2	0.3209	0.2834	0.3235	0.3316	0.3182
Precision at 3	0.2282	0.2210	0.2299	0.2406	0.2228

Rispetto alla baseline:

- 7 documenti rilevanti in più tra i restituiti
- MAP + 3.1%, P@1 + 2%

Gli n-grammi

Un **n-gramma** è una sotto-sequenza di n elementi estratti da una sequenza di partenza.

Un **bigramma** è quindi un n-gramma di lunghezza due.

Lo studio dei bigrammi consiste nell'analisi della probabilità condizionata di poter trovare una determinata parola conoscendo la precedente. La relazione di probabilità condizionata è data dalla seguente formula:

$$P(W_n|W_{n-1}) = \frac{P(W_{n-1}, W_n)}{P(W_{n-1})}$$

L'analisi dei bigrammi può essere vista come un caso specifico di studio della proximity.

La proximity

La **proximity** è una forma di dipendenza tra termini, basata sulla distanza dei termini stessi all'interno documento.

È stato dimostrato come un documento, in cui compaiono tutti i termini della query, abbia una più alta probabilità di risultare maggiormente rilevante rispetto ad un documento contenente solo alcuni dei termini della query. Allo stesso modo, un ulteriore indice di rilevanza è costituito dalla vicinanza dei termini di una query all'interno di un documento.

Utilizzeremo lo studio condotto sui bigrammi significativi per superare alcuni dei limiti imposti dal modello **bag of words**.

Tavola di contingenza

Al fine di identificare i bigrammi significativi si è pensato di studiare la correlazione esistente tra i termini utilizzando alcuni coefficienti. Per calcolare questi coefficienti è stato necessario definire correttamente lo spazio degli eventi. Siano quindi:

- N il numero totale di bigrammi.
- n_{t_i} la frequenza dell'unigramma t_i.
- $n_{t_{ij}}$ la frequenza del bigramma t_{ij}.
- n_{t_j} la frequenza dell'unigramma t_j.

	$t_i = 1$	$t_i = 0$	
$t_j = 1$	$n_{t_{ij}}$	$n_{t_j} - n_{t_{ij}}$	n_{t_j}
$t_j = 0$	$n_{t_i} - n_{t_{ij}}$	$N - n_{t_i} - n_{t_j} + n_{t_{ij}}$	$N - n_{t_j}$
	n_{t_i}	$N - n_{t_i}$	N

Tabella: Tavola di contingenza.

Coefficienti di correlazione

Indicando con $[X]$ la cardinalità dell'insieme riportato nella tavola di contingenza precedente otteniamo

	$t_i = 1$	$t_i = 0$	
$t_j = 1$	$[1]$	$[2]$	$[7]$
$t_j = 0$	$[3]$	$[4]$	$[8]$
	$[5]$	$[6]$	$[9]$

Da cui possiamo calcolare vari coefficienti attraverso le seguenti formule:

Coefficienti

$$\rho(t_i, t_j) = \frac{[1][9]-[5][7]}{\sqrt{[5][7][6][8]}}.$$

$$EMIM(t_i, t_j) = \frac{[1]}{[9]} \log \frac{[1][9]}{[5][7]} + \frac{[2]}{[9]} \log \frac{[2][9]}{[6][7]} + \frac{[3]}{[9]} \log \frac{[3][9]}{[5][8]} + \frac{[4]}{[9]} \log \frac{[4][9]}{[6][8]}.$$

$$MI(t_i, t_j) = \frac{[1]}{[9]} \log \frac{[1][9]}{[5][7]}.$$

$$CMI(t_i, t_j) = \log \frac{[1]}{[7]}.$$

SVM

Le frequenze dei bigrammi e degli unigrammi che li compongono sono state utilizzate, congiuntamente ai coefficienti di correlazione, per addestrare un classificatore SVM.

L'obiettivo da raggiungere, grazie all'ausilio del classificatore, consiste nel riconoscere i bigrammi significativi tra tutti i bigrammi presenti nella collezione.

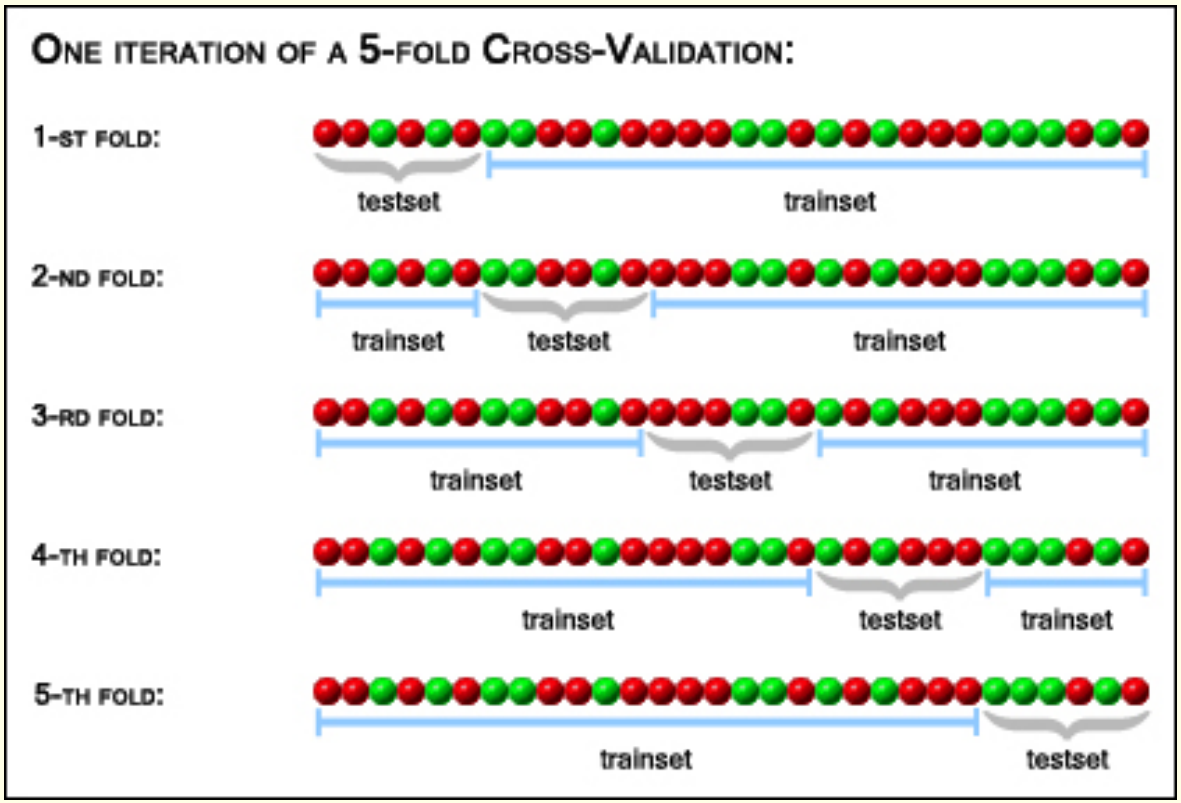
Il vettore contente le feature rappresentanti il bigramma è stato così strutturato:

$$< label \quad 1 : tf_{t_i,t_j} \quad 2 : tf_{t_i} \quad 3 : tf_{t_j} \quad 4 : \rho \quad 5 : MI \quad 6 : EMIM \quad 7 : CMI >$$

Il classificatore SVM è stato addestrato attraverso raffinamenti successivi del training set, effettuati tramite la tecnica del **n-fold cross validation**.

N-Fold Cross Validation

La tecnica del **N-Fold Cross-Validation** viene utilizzata per valutare l'accuratezza e il grado di apprendimento raggiunto dalla SVM.



Nel nostro caso abbiamo utilizzato la N-Fold Cross Validation con $N = 5$

Il Document Score Modifier

I bigrammi significativi sono stati utilizzati per alterare lo score dei documenti attraverso l'utilizzo di un Document Score Modifier che ridefinisce il peso assegnato ad un documento d nel seguente modo:

$$score(d) = score_U(q_U) + score_B(q_B) \times p_B \times \lambda$$

dove $score_U(q_U)$ è lo score assegnato al documento d recuperato tramite la query q_U sull'indice degli unigrammi, mentre $score_B(q_B)$ è lo score del documento d recuperato tramite la query q_B sull'indice dei bigrammi. p_B è il valore restituito dalla SVM e λ è un parametro di **smoothing**. Teoricamente $0 < \lambda < \infty$, nel nostro caso $\lambda = 0.63$.

Bigrammi significativi: analisi dei risultati(1/2)

A seguito dell'introduzione dei bigrammi significativi sono dunque stati ripetuti i test di verifica delle performance ottenendo i risultati riportati in tabella.

	QE-G	QE-G+Phrasal	QE-G+Bigrammi
Number of query	187	187	187
Retrieved	660	632	**660**
Relevant	201	201	201
Relevant retrieved	135	132	**135**
Average Precision	0.6185	0.6167	**0.6266**
R-Precision	0.5633	0.5667	0.5793
Precision at 1	0.5668	0.5622	**0.5829**
Precision at 2	0.3316	0.3297	0.3316
Precision at 3	0.2406	0.2378	0.2406

- MAP $+1.3\%$ P@1 $+2.8\%$ rispetto alla QE.
- MAP $+1.6\%$ P@1 $+3.7\%$ rispetto al phrasal.

Bigrammi significativi: analisi dei risultati(2/2)

L'introduzione dei bigrammi significativi ha portato ad un buon aumento della early precision con quasi 3 punti percentuali.

Ovviamente un approccio di questo genere non ha minimamente intaccato il numero di documenti restituiti, ma questo era un risultato atteso visto il tipo di sperimentazione intrapresa.

Con la nostra sperimentazione siamo inoltre riusciti ad ottenere risultati migliori rispetto al modello del phrasal attualmente presente in Terrier.

Considerazioni finali sui risultati ottenuti(1/2)

Le analisi condotte hanno contribuito ad aumentare le prestazioni del sistema di circa un 5%. Con un incremento specifico della MAP pari a 4.5% ed un'aumento della sola precision at 1 del 5%.

In realtà tali valori potrebbero risultare fuorvianti se non si considerasse che la baseline su cui ci siamo basati finora (FAQ+LOG) è il risultato della composizioni di due baseline distinte (FAQ e LOG) che mostravano dati di partenza diametralmente opposti.

Se da una parte infatti la baseline del LOG risultava estremamente bassa e dunque altamente migliorabile, quella delle FAQ otteneva una MAP che sia attestava vicino al 90%.

Una baseline così elevata rende estremamente difficile apportare significativi miglioramenti al sistema, poiché si avvicina ai limiti fisici imposti dal modello scelto.

Considerazioni finali sui risultati ottenuti(2/2)

Analizzando le singole baseline è possibile notare come l'elevata baseline delle FAQ mascheri i reali miglioramenti compiuti dal sistema.

	BL FAQ	QE+Bigr FAQ	BL LOG	QE+Bigr LOG
Number of query	110	110	77	77
Retrieved	327	327	330	**332**
Relevant	110	110	91	91
Relevant retrieved	106	106	22	**29**
Average Precision	0.8939	**0.8985**	0.1797	**0.2468**
R-Precision	0.8364	0.8455	0.1342	0.2121
Precision at 1	0.8364	**0.8455**	0.1558	**0.2208**
Precision at 2	0.4636	0.4636	0.1169	0.1494
Precision at 3	0.3212	0.3212	0.0952	0.1255

- MAP $+37\%$, P@1 $+41\%$ sui LOG

Per concludere

Durante questo lavoro di tesi abbiamo:

- effettuato uno studio sulla possibilità di impiego di tecniche statistiche per migliorare i sistemi di QA.
- effettuato uno studio sull'efficacia di utilizzo di risorse esterne in presenza di piccole collezione tramite la QE.
- effettuato uno studio sulla proximity e la dipendenza dei termini attraverso l'analisi dei bigrammi significativi.

Possibili sviluppi potrebbero essere i seguenti:

- utilizzare i bigrammi del WEB rilasciati da Google per generare una lista di bigrammi significativi universali, così da non dover addestrare il modulo SVM su ogni collezione
- analizzare nuovi modelli di scoring per i bigrammi significativi testandoli su collezioni di grandi dimensioni.

Grazie!