Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecniche LSA per la composizione automatica di risposte di chat-bot

S.Gaglio, G.Lupo, G.Pilato, G.Vassallo

RT-ICAR-PA-05-10 **ottobre 2005**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) – Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sede di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.na.icar.cnr.it*
– Sede di Palermo, Viale delle Scienze, 90128 Palermo, URL: *www.pa.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecniche LSA per la composizione automatica di risposte di chat-bot

S.Gaglio[1,2], G.Lupo[2], G.Pilato[1] , G.Vassallo[2]

Rapporto Tecnico : 10 ***Data:***
RT-ICAR-PA-05-10 **ottobre 2005**

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sede di Palermo Viale delle Scienze edificio 11 90128 Palermo
[2] Università degli Studi di Palermo Dipartimento di Ingegneria Informatica Viale delle Scienze 90128 Palermo

Sommario

In questo lavoro è stata presentata una tecnica sub-simbolica per la generazione automatica di frasi che abbiano un senso compiuto. Tale tecnica può essere adoperata in un sistema di dialogo come il chat-bot per la generazione automatica di risposte relative a domande poste dagli utenti su di uno specifico argomento. La generazione automatica di tali frasi sarà effettuata mediante la predizione di parole rare che hanno generalmente una semantica rilevante all'interno di queste. La tecnica proposta prende vantaggi sia dal tradizionale approccio LSA che dalla combinazione della nuova applicazione di metrica dello spazio di probabilità conosciuta come la distanza di Hellinger con la probabilità condizionata calcolata sui digrammi dei corpus d'addestramento. Sono inoltre presenti degli esperimenti che validano l'approccio proposto.

1. Introduzione

I Chat-bot, dove "Chat" sta per conversazione elettronica e "Bot" è l'abbreviazione di robot, sono la rappresentazione virtuale di un interlocutore-computer, ovvero sono un insieme di 'agenti' software che si adeguano alle esigenze dell'interlocutore. Tali agenti, dotati di interfacce audio-video, sono in grado di sostenere una conversazione, quanto più appagante possibile, in linguaggio naturale con un interlocutore umano. Sistemi di questo tipo furono creati nel tentativo di superare un test proposto in un articolo del 1950 da A.Turing [1].

Il chat-bot possiede generalmente una grande base di conoscenza costituita da migliaia di "moduli domanda-risposta". L'utente del chat-bot effettua la domanda mediante un terminale ed il chat-bot , dopo aver analizzato tutti i moduli domanda-risposta che sono all'interno della sua base di conoscenza ,

restituisce la risposta più adeguata alla domanda presa in considerazione, ed è quindi la risposta che ha la migliore corrispondenza con la domanda dell'utente.

L'architettura del software di un chat-bot deve:

- rappresentare la base di conoscenza in maniera adeguata;
- possedere un metodo che possa restituire la risposta più appropriata alla domanda che l'utente effettua;
- prevedere delle risposte predefinite (di default) in caso il metodo succitato non restituisce niente.

I moduli "domanda-risposta" si strutturano in un albero semantico, dove i rami corrispondono alle parole che costituiscono la conoscenza del chat-bot mentre i percorsi che partono dalla radice fino ad un nodo foglia rappresentano la domanda dell'utente e puntano alla risposta relativa.

Il sistema deve prevedere dei meccanismi che consentano di individuare dei percorsi precisi all'interno dell'albero in modo da fornire la risposta più idonea ad una determinata domanda, prevedendo anche dei percorsi predefiniti (di *default*) che possano consentire al chat-bot di proseguire la conversazione anche quando non è in grado di rispondere.

Il chat-bot è anche un particolare sistema di recupero automatico dell'informazione. I sistemi di recupero dell'informazione si occupano della rappresentazione, memorizzazione, organizzazione ed accesso ai contenuti informativi di una collezione di documenti[2]. Tali sistemi di information retrieval(IR) permettono all'utente di esprimere una *domanda* e di vedere soddisfatto il proprio fabbisogno informativo. I motori di ricerca (google, altavista, yahoo per citare i più famosi) sono il principale esempio di applicazione che utilizza fortemente delle tecniche di IR e ne mette anche in luce uno dei principali limiti: l'utente deve esplicitamente descrivere le caratteristiche dell'informazione desiderata, ovvero deve essere a conoscenza dei propri interessi [3].

Per valutare un sistema di recupero dell'informazione vengono utilizzate due misure: *Richiamo* e *Precisione*. Il *Richiamo* è la frazione dei documenti rilevanti che è stata recuperata mentre la *Precisione* è la frazione dei documenti reperiti che è rilevante.

I sistemi classici dell'IR riassumono il contenuto informativo dei documenti e delle richieste d'informazione dell'utente in un insieme di parole chiave e recuperano quei documenti che effettuano una corrispondenza lessicale con la richiesta d'informazione, ossia quei documenti che contengono le parole chiave contenute nella richiesta d'informazione. Tale approccio provoca molte problematiche; la maggior parte della semantica di un documento o di una richiesta d'informazione, infatti, viene persa nel momento stesso in cui il testo che esprime tale semantica viene sostituito da un insieme di termini. La causa principale di questo problema è la varietà di parole che generalmente si utilizzano per esprimere lo stesso concetto. Tra i problemi che un sistema di recupero d'informazione deve affrontare nell'analizzare testi in linguaggio naturale, risultano evidenti quelli che derivano dalla *sinonimia*, ossia la possibilità di esprimere lo stesso concetto con diversi termini, e dalla *polisemia*, ossia la possibilità di utilizzare uno stesso termine con differenti significati in contesti differenti.

Tali inconvenienti sono stati superati dalle tecniche che estraggono le caratteristiche latenti del testo basate sull'Analisi Semantica Latente (LSA, Latent Semantic Analisis)[4], [5]). L'LSA è un paradigma per estrarre e rappresentare il significato calcolato tramite computazioni statistiche applicate su dei grandi corpi testuali. L'LSA è basato sul metodo dello spazio vettoriale: dato un corpo testuale di M documenti ed N parole, il paradigma LSA definisce una corrispondenza (mappatura) tra gli M documenti e le N parole in uno spazio vettoriale continuo S, dove ogni parola w (i) è associata ad un vettore U (i) in S, ed ogni documento d (j) è associato ad un vettore V (j) in

S[6]. In questo modo può essere assegnata una probabilità diversa da zero se si è "vicino" ad una parola del corpo d'addestramento. La "vicinanza" è definita dalla metrica nello spazio vettoriale S.
In questo lavoro proponiamo una tecnica per la generazione automatica di risposte tramite chat-bot, dove la riposta più rilevante per la domanda specifica viene parametrizzata seguendo la legge di Zipf[7] che individua le parole più rare le quali hanno generalmente una maggiore semantica nel contesto in cui si trovano(tali parole rare saranno quelle che il nostro sistema dovrà predire). Tale tecnica permette di non considerare gli schemi rigidi che caratterizzano generalmente un chat-bot e quindi non si basa sull'esatta corrispondenza lessicale con la richiesta di informazione. La tecnica proposta non sfrutta la tradizionale LSA basata sull'approccio parola-documento ma usufruisce del nuovo approccio parola-parola. Il nostro progetto si basa sulla codifica vettoriale delle parole, dove ognuna di queste viene codificata tramite due vettori uno che codifica il contesto destro e l'altro quello sinistro della stessa parola. Questo progetto comprende il paradigma LSA basato sull'approccio parola-parola, il calcolo della probabilità condizionata sui digrammi dei corpus d'addestramento, la distanza di Hellinger e la tecnologia dei chat-bot. La distanza di Hellinger è definita nello spazio vettoriale che viene convertito in uno spazio metrico. Questa distanza è adatta per misurare la perdita d'informazione quando si considera un'approssimazione di una distribuzione di probabilità [8], [9], [10], cosicché la nostra tecnica da al sistema di dialogo la possibilità d'estrarre sia le caratteristiche generali che quelle nascoste che provengono dall'LSA ed inoltre, utilizziamo la distanza di Hellinger, che taglia fuori solo l'informazione relativa ai dettagli strutturali del testo d'addestramento, preservando l'informazione relativa a strutture sintattiche e semantiche nascoste all'interno di questo.
Il software del sistema di dialogo risultante permette la generazione automatica di risposte coerenti alle domande effettuate dagli utenti, catturando sia le principali strutture sintattiche che le strutture semantiche nascoste all'interno del testo d'addestramento.
L'articolo è organizzato come segue: è presentata una descrizione del sistema di dialogo "chat-bot" come sistema di recupero d'informazione. Poi viene descritta la struttura matematica usata dalla nostra tecnica e quindi presentata la stessa tecnica con le sue caratteristiche. Infine sono presenti alcune prove sperimentali per la convalida del nostro progetto.

2. L'approccio proposto

2.1 Motivazioni

Lo scopo di questo lavoro è la progettazione e l'implementazione di un sistema di dialogo chat-bot che generi automaticamente delle risposte a delle domande effettuate dagli utenti. Attualmente questi sistemi si basano soltanto sui classici sistemi di recupero di informazione con tutti quei

svantaggi citati in precedenza che non permettono alcuna predizione e generalizzazione del sistema chat-bot. Il chat-bot basato sulla nostra tecnica non presenta più questi svantaggi. Per la realizzazione di questo sistema di dialogo abbiamo fatto due scelte di base:

1. l'utilizzo dell'approccio LSA basato sulla Decomposizione del Valore Singolare (SVD, Singular Value Decomposition), la quale è una tecnica numerica che fornisce un fattorizzazione di una matrice che può essere troncata facilmente per ottenere un'approssimazione a basso rango della matrice di partenza. Tale tecnica ha presentato buoni risultati sperimentali [5] ed è relativa a modelli cognitivi umani. L'SVD troncato fornisce una mappa di parole su uno spazio semantico i cui elementi mostrano una relazione stretta con le proprietà statistiche del testo di addestramento.
2. l'utilizzo della metrica di Hellinger nello spazio semantico in combinazione con il calcolo della probabilità condizionata sui digrammi del corpus dell'addestramento. Questa particolare metrica è stata scelta perché è stato dimostrato [11] [8], che misura adeguatamente la perdita di informazione relativa ad un'approssimazione di distribuzione della probabilità, considerando così l'interpretazione statistica dei vettori nello spazio semantico.

Prima di descrivere la nostra tecnica in dettaglio, verrà presentato il background matematico che servirà ad ottenere una struttura matematica tale da permettere di superare dei limiti e dei svantaggi relativi agli approcci tradizionali che diminuiscono le prestazioni e soprattutto l'efficacia dei sistemi di dialogo. Verranno introdotte le nozioni su: la distanza di Hellinger, la legge di Zipf ed infine il calcolo del coefficiente di incertezza.

2.2 La distanza di Hellinger

Sia *Z* un insieme arbitrariamente finito (denominato "spazio campionario") e *P* lo spazio di tutte le distribuzioni di probabilità su *Z*, cioè l'insieme di tutte le funzioni $p : Z \to [0,1]$ tale che $\sum_{z \in Z} p(z) = 1$. Sia inoltre $w \in R^S$ un vettore di parametri reali, che possa variare all'interno di un sottoinsieme $W \subset R^S$, e si supponga definita una corrispondenza biunivoca che associ in maniera unica una distribuzione di probabilità $p \in P$ a ciascuna $w \in R^S$; per ogni $w \in R^S$ sia p_w la distribuzione di probabilità associata su *Z*. Il sottoinsieme di *P*:

$$Q = \{ q \in P : (\exists w \in W) \Rightarrow q = p_w \}$$

è chiamato "modello computazionale". Per definizione un modello statistico è invece una distribuzione di probabilità π sull'insieme *P* e uno stimatore la funzione $\tau : Z \to Q$, se *z* è un elemento di *Z*, allora l'oggetto $\tau(z)$ rappresenta la distribuzione di probabilità su *Z*, ossia la stima.

Poiché ciascun elemento di Q può essere scritto come $q = p_w$ per una data parola $w \in W$, risulta che lo stimatore è in corrispondenza biunivoca tra lo spazio campionario e l'insieme W dei parametri consentiti (anche se, dato un $q \in Q$, possono esistere più $w \in W$ affinché si abbia $p_w = q$). Il problema nell'utilizzo di tali stimatori è la determinazione di quello ottimale.

Nel caso specifico la costruzione di un modello statistico del linguaggio richiede la stima di una distribuzione della probabilità sull'insieme di tutte le possibili parole di un corpus effettuando un inferenza da un particolare insieme campione. A tale scopo, vorremmo utilizzare uno "stimatore sufficiente", cioè uno stimatore che è in grado di ricostruire solo le informazioni nascoste relative alla semantica, eliminando caratteristiche relative al particolare insieme campione utilizzato per l'inferenza. È stato mostrato [11] che adattando i parametri del modello per minimizzare la distanza di Hellinger:

$$d_H(p,q) = \sum \left(\sqrt{p} - \sqrt{q}\right)^2$$

tra la distribuzione della probabilità stimata q e la distribuzione della probabilità "vera" p.

Una distanza fra matrici può essere associata alla distanza di Hellinger fra vettori di distribuzione di probabilità allo stesso modo in cui la distanza di Frobenius è associata alla distanza Euclidea [articolo]definita:

$$d_H(X,Y) = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\sqrt{X_{ij}} - \sqrt{Y_{ij}}\right)^2}$$

per ciascuna coppia di matrici X,Y di dimensione $M \times N$ tale che $0 \le X_{ij} \le 1$, $0 \le Y_{ij} \le 1$, con $i, j = 1,2,\mathrm{K}, n$ e $\sum_{i=1}^{M}\sum_{j=1}^{M} X_{ij} = \sum_{i=1}^{M}\sum_{j=1}^{M} Y_{ij} = 1$.

Si noti che la distanza di Frobenius che si calcola come $d_F(X,Y) = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{ij} - Y_{ij})^2}$ tra le matrici X e Y è la distanza euclidea senza il calcolo della radice quadrata di ciascun elemento.

2.3 La legge di Zipf

La legge di Gorge Kingsley Zipf, stabilisce che la frequenza di utilizzo della parola *n–esima* utilizzata maggiormente in un qualsiasi linguaggio naturale è approssimativamente inversamente proporzionale a *n*. Infatti l'utilizzo classico della legge di Zipf è l'applicazione della funzione "$1/f$": dato un insieme di frequenze distribuite, ordinate dalla più alta alla più bassa, la seconda frequenza avrà un'occorrenza pari alla metà della prima, la terza pari ad un terzo della prima e così via fino alla parola *n–esima*, la cui frequenza sarà pari a $1/n$ rispetto alla prima. Inoltre la legge è di

tipo sperimentale e le distribuzioni studiate da Zipf sono comunemente osservate in diverse tipologie di fenomeni, come riportato in [12].

La legge di Zipf può essere espressa matematicamente come segue:

$$p_k(s,N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$$

dove *n* è il numero di elementi, *k* è il rango, mentre *s* è l'esponente che caratterizza la distribuzione. Riferendosi alla lingua inglese, *N* è il numero di parole presenti nel linguaggio e *s* sarà pari a 1; p_k sarà pertanto la frazione della probabilità di occorrenza della parola *n–ma* più frequente. Si noti che la distribuzione è normalizzata, ossia:

$$\sum_{k=1}^{N} p_k(s,N) = 1$$

e pertanto la legge può essere riscritta come:

$$p_k(s,N) = \frac{1/k^s}{H_{N,s}}$$

dove $H_{N,s}$ è il numero di ordine *n* della serie armonica generalizzata.

2.4 Il coefficiente di incertezza

Il paragrafo tratterà il calcolo del coefficiente di incertezza mediante la misura dell'associazione tra variabili nominali. Per ogni coppia di variabili nominali, i dati possono essere mostrati come una tabella di contingenza[13]. L'analisi dell'associazione tra variabili nominali è chiamata "analisi della tabella di contingenza".

Supponiamo di voler assegnare ad una domanda una risposta adeguata; ogni risposta ha probabilità p(un numero positivo p<1) ed a questa può essere assegnato un valore –ln p. Se ci sono I possibili risposte alla domanda (i=1,…,I) e la frazione di possibilità corrispondente all'i-esima riposta è p_i (con la somma delle probabilità p_i uguale ad 1), allora il valore della domanda è il valore aspettato della risposta , chiamato H,

$$H = -\sum_{i=1}^{I} p_i \ln p_i$$

valutando la precedente si nota che:

$$\lim_{p \to 0} p \ln p = 0$$

tramite la quale si intende che il valore di H è compreso tra 0 ed ln*I*. Il valore H è convenzionalmente definito l'*entropia* della distribuzione data dalle p_i .

Supponiamo che una domanda d1, ha I possibili risposte, etichettate con i, ed un'altra domanda d2, ha J possibili risposte, etichettate con la j.
Le entropie di d1 e d2 sono rispettivamente:

$$H(d1) = -\sum_i p_i \ln p_i \qquad H(d2) = -\sum_j p_j \ln p_j$$

L'entropia dei d1 e d2 prese in considerazione contemporaneamente è:

$$H(d1, d2) = -\sum_{ij} p_{ij} \ln p_{ij}$$

Adesso si definisce cos'è l'entropia della domanda d2 data la d1. È il valore dell'aspettativa sulle risposte a d1 dell'entropia della distribuzione ristretta a d2 che giace in una singola colonna della tabella di contingenza (corrispondendo a d1):

$$H(d2 \mid d1) = -\sum_{ij} p_{ij} \ln \frac{p_{ij}}{p_i}$$

ed in maniera analoga, l'entropia di d1 dato d2 si calcola come:

$$H(d1 \mid d2) = -\sum_{ij} p_{ij} \ln \frac{p_{ij}}{p_j}.$$

Adesso possiamo definire una misura della "dipendenza" di d2 da d1, vale a dire una misura di associazione. Questa misura è chiamata *coefficiente di incertezza* di d2. Lo indicheremo come U (d2 | d1):

$$U(d2 \mid d1) = \frac{H(d2) - H(d2 \mid d1)}{H(d2)}$$

Questa misura è compresa tra zero e uno, con il valore 0 si indica che d1 e d2 non hanno alcuna associazione, con il valore 1,invece, si indica che la conoscenza di d1 predice completamente d2.
Allo stesso modo se si vuole conoscere la dipendenza tra d1 e d2 considerando questa volta la d1 come la variabile dipendente e la d2 come quella indipendente, il coefficiente di incertezza si calcola cambiando la d1 con la d2 e viceversa:

$$U(d1 \mid d2) = \frac{H(d1) - H(d1 \mid d2)}{H(d1)}.$$

2.5 La tecnica proposta

Sia definita una matrice A delle co-occorrenze dei corpus d'addestramento. La soluzione che proponiamo è un approccio basato sul digramma parola-parola che sfrutta la TSVD(Truncated Singular Value Decomposition)[14], in modo che la matrice ricostruita è l'approssimazione migliore della matrice originale A in riferimento alla distanza di Hellinger.

La nostra tecnica consta di cinque fasi:

1- Pre-elaborazione;
2- Eliminazione stop-word;
3- Generazione della matrice A;
4- Applicazione della SVD;
5- Predizione di una parola rara mediante LSA con l'ausilio del calcolo della probabilità condizionata sul digramma contenente la stessa.

La prima fase di pre-elaborazione serve per effettuare un primo filtraggio del corpus testuale in esame. In questa fase si eliminano: punteggiatura, numeri e caratteri speciali come lo spazio, gli apici etc.

La seconda fase esegue un secondo filtraggio del corpus eliminando le stop-word o le cosiddette parole comuni, le quali non hanno una semantica rilevante nel contesto. Tale fase sottolinea il fatto che la nostra tecnica si utilizzerà per sistemi di tipo semantico e non di tipo sintattico.

La terza fase genera la matrice A delle co-occorrenze ovvero dei digrammi individuati nel corpus che ha subito già la pre-elaborazione ed è privo delle parole comuni. Le righe della matrice A rappresentano i vettori relativi al contesto sinistro mentre le colonne rappresentano i vettori relativi al contesto destro. I digrammi vengono individuati contando le occorrenze di ognuno e riordinandoli per valori di occorrenza decrescenti. Questi sono ricercati all'interno di una finestra di parole di grandezza 10(individuata sperimentalmente), quindi si considerano tutte le combinazioni dei digrammi del corpus all'interno delle 10 parole. Si noti che tutte le parole rare del corpus ricercate mediante la legge di Zipf che stabilisce che la frequenza di utilizzo della parola *n–esima* utilizzata maggiormente in un qualsiasi linguaggio naturale è approssimativamente inversamente proporzionale a *n*, sono sostituite con dei rappresentanti.

La quarta fase applica la tecnica della decomposizione ai valori singolari ridotta alla matrice A. Se la matrice A è di ordine t x d questa può essere scomposta nel prodotto matriciale di tre matrici come nella figura 1.

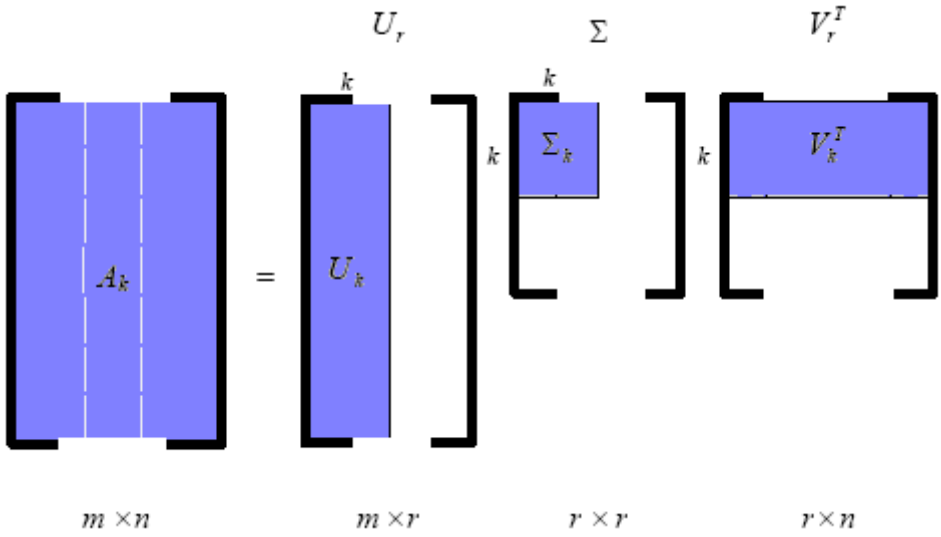

Figura 1: Decomposizione ridotta

Nella figura, K rappresenta il numero della dimensione del modello ridotto. L'evidenza matematica assicura che è possibile decomporre perfettamente la matrice utilizzando un numero di fattori non maggiore della dimensione più piccola della matrice originale. La matrice ricostruita $\tilde{A}$ risulta la migliore approssimazione di *A* secondo la norma di Frobenius, ma poiché gli elementi di *A* sono delle frequenze normalizzate, quest'ultima può essere interpretata come una funzione di probabilità discreta sullo spazio delle configurazioni generata a partire da tutti i digrammi possibili. Ricordando che la distanza di Hellinger tra due matrici è la distanza di Frobenius ottenuta a partire delle matrici originali estraendo la radice quadrata di ciascun elemento, è possibile porre:

$$B_{ij} = \sqrt{A_{ij}}$$

per $i, j = 1 : N$. Applicando la decomposizione ai valori singolari ridotta(o troncata,TSVD) sulla matrice *B* trattenendo soltanto gli *R* valori singolari più grandi e se $\tilde{B} = \tilde{U}\tilde{S}\tilde{V}^T$ è la matrice ricostruita ottenuta, $\tilde{A}$ risulta pertanto la matrice costituita da tutti gli elementi di $\tilde{B}$ al quadrato e la migliore approssimazione di *A* di rango *K* secondo la distanza di Hellinger. Per i vari significati delle matrici U e V e le loro approssimazioni vedere in [14].

A differenza dell'approccio parola-documento nel quale ogni parola è rappresentata da un vettore, si associa a ciascuna parola due vettori, uno relativo al contesto sinistro e l'altro al destro. L'efficienza di tale tecnica è fondata sull'assunzione della distanza di Hellinger tra le probabilità della matrice di occorrenza del digramma originale e la matrice ricostruita di rango *K*, come funzione obiettivo da minimizzare mediante la TSVD: la naturalezza di tale funzione è evidente quando la perdita d'informazione nell'approssimazione delle distribuzioni di probabilità deve essere misurata.

La quinta ed ultima fase della tecnica proposta effettua la predizione delle parole rare di un certo corpus analizzando questo tramite l'analisi della semantica latente basata sull'approccio parola-parola con l'ausilio del calcolo della probabilità condizionata sui digrammi contenenti la stessa.

L'LSA permette di analizzare il contesto sinistro e destro mediante i due vettori che codificano la parola. Per contesto destro e sinistro della parola rara intendiamo rispettivamente le successive tre parole e le precedenti tre parole semanticamente significative, in quanto il corpus è stato privato delle stop-word nella seconda fase della nostra tecnica. Ogni parola totalizzerà un cero score che viene calcolato come:

$$LSA = \alpha \cdot \text{contesto sinistro} + \beta \cdot \text{contesto destro}$$

dove $0 \leq \alpha, \beta \leq 1$. Le parole che verranno prese in considerazione saranno ordinate in ordine decrescente rispetto al punteggio totalizzato che terrà conto sia del contesto sinistro che di quello destro.

Da questa prima classificazione si prenderanno in considerazione le prime venti parole che sono le potenziali sostitute della parola rara parametrizzata, ed agli score di queste totalizzati con la sola LSA si aggiunge la probabilità condizionata calcolata sui digrammi del corpus come nella seguente formula:

$$scoreTot = \alpha * LSA + \beta * P(A \mid B)$$

dove LSA rappresenta lo score relativo alla sola analisi della semantica latente, P(A|B) la probabilità condizionata calcolata sul digramma AB dove A è a parola da predire e B è quella successiva semanticamente significativa ed infine α e β sono due pesi tali che $0 \leq \alpha, \beta \leq 1$.

Il risultato finale della nostra tecnica permette di ricercare la parola prima classificata. Tale parola sostituirà la parola rara individuata con una la parola fittizia che non ha alcun significato come ad esempio "PARAM". Tali sostituzioni permetteranno al sistema di dialogo "chat-bot" di predire le parole rare del corpus d'addestramento e quindi nel caso specifico delle risposte generandole automaticamente in modo tale che siano semanticamente correlate alla domanda dell'utente e che contemporaneamente siano corrette sia da un punto di vista semantico che sintattico.

3. Prove sperimentali

Dagli esperimenti effettuati, relativi al solo punteggio realizzato dalla tecnica LSA,

$$LSA = \alpha \cdot \text{contesto sinistro} + \beta \cdot \text{contesto destro}$$

si è verificato che i migliori risultati si sono avuti considerando i due pesi alfa e beta:

$$\alpha = \beta = 0{,}5.$$

Mentre i risultati migliori, relativi alla nostra tecnica,

$$scoreTot = \alpha * LSA + \beta * P(A \mid B)$$

si sono ottenuti con α=0,4 e β=1. Si noti che l'utilizzo della probabilità condizionata P(A|B) rispetto alla P(B|A) è giustificato dalle prove sperimentali fatte per il calcolo del coefficiente di incertezza, dalle quali si è riscontrato che la probabilità condizionata di A rispetto a B ha un maggiore potere predittivo che la probabilità condizionata di B rispetto ad A.

3.1 Valutazione del sistema di recupero dell'informazione

Le prove sperimentali sono state fatte in riferimento a tre corpus d'addestramento per un totale di circa 4500 parole differenti semanticamente rilevanti (ossia non consideriamo le stop-word). I tre corpus[16] sono:

- MED, relativo alla medicina;
- CRAN, relativo alla fisica;
- CISI, relativo agli articoli scientifici.

Abbiamo effettuato un confronto calcolando Richiamo e Precisione del nostro sistema chat-bot tra l'approccio parola-parola e quello parola-documento ed anche con un classico sistema di recupero d'informazione di tipo sintattico Lucene[15]. In seguito saranno mostrate delle tabelle relative a due dei corpus presi in considerazione MED e CISI ed assieme a queste saranno mostrati dei grafici riguardanti le tabelle(si noti che nelle tabelle i risultati sono stati approssimati alla seconda cifra decimale).

Dati relativi agli esperimenti effettuati sul corpus MED:

MED	Parola-Parola	Parola-Documento	Lucene
Richiamo	0,38	0,38	0,38
Precisione	0,68	0,63	0,45

Tabella 1: Richiamo e precisione di MED

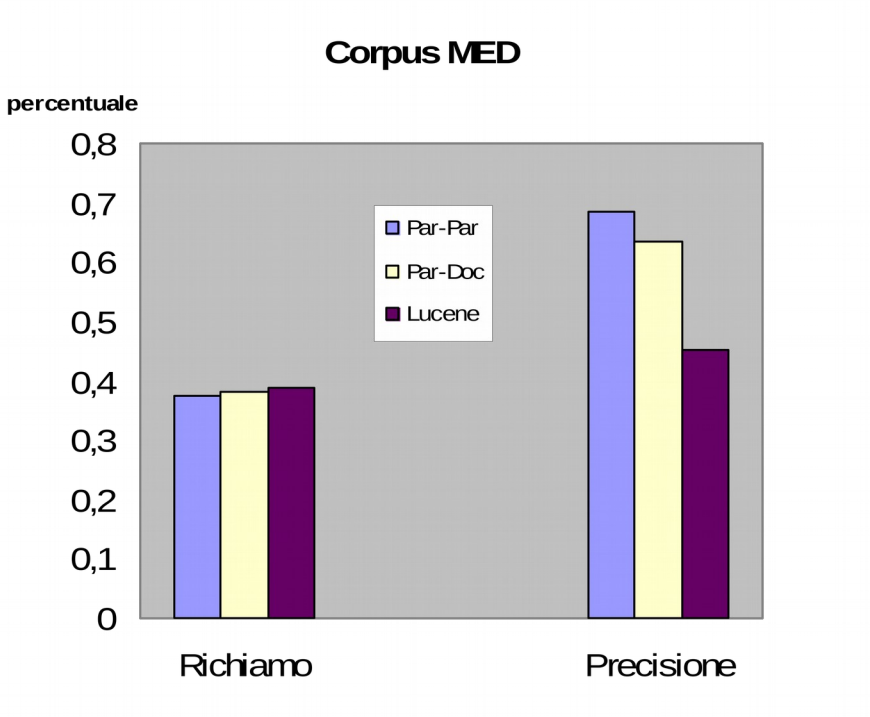

Figura 2: Grafico relativo al corpus MED

Per quanto riguarda il corpus MED relativo alla medicina si può notare sia dalla tabella che dal grafico che a parità di richiamo di circa 0,38 la precisione migliore è quella relativa al sistema basato sull'approccio parola-parola che è 0,68 dopo si classifica l'approccio parola-documento con una precisione 0,63 ed infine Lucene con una precisione 0,45.

Dati relativi agli esperimenti effettuati sul corpus CISI:

CISI	Parola-Parola	Parola-Documento	Lucene
Richiamo	0,05	0,05	0,05
Precisione	0,30	0,24	0,20

Tabella 2: Richiamo e precisione di CISI

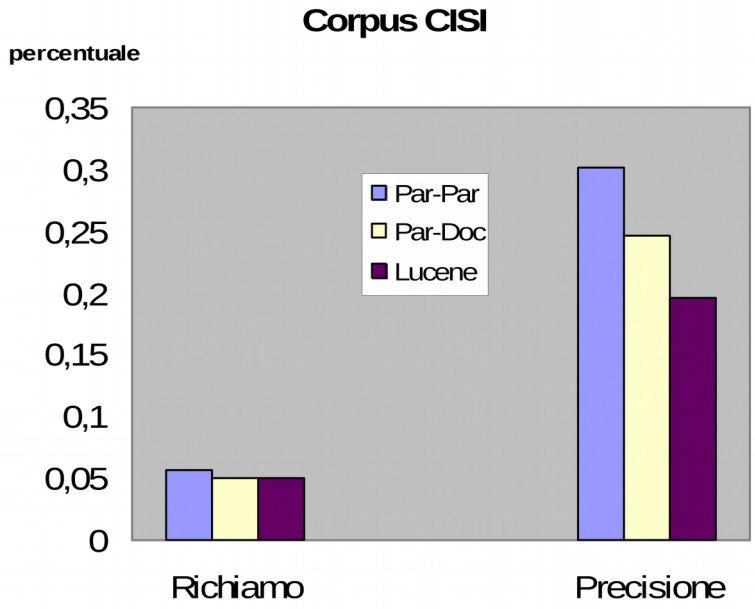

Figura 3: Grafico relativo al corpus CISI

Per quanto riguarda il corpus CISI relativo agli articoli scientifici si può notare sia dalla tabella che dal grafico che a parità di richiamo di circa 0,05 la precisione migliore è quella relativa al sistema basato sull'approccio parola-parola che è poco superiore a 0,30 dopo si classifica l'approccio parola-documento con una precisione di 0,25 ed infine Lucene con una precisione 0,20.
Tali risultati sperimentali validano l'approccio dell'LSA parola-parola.

3.2 Risultati relativi al coefficiente di incertezza

La verifica, che serviva per calcolare sperimentalmente quale delle probabilità condizionate portava ad una maggiore predizione P(A|B) oppure P(B|A), è stata effettuata su di una campione di cinquanta digrammi AB(dove A rappresenta la prima parola e quindi quella da predire e B la seconda). Tali campioni sono stati scelti casualmente tra le parole dei corpus d'addestramento quali:

- MED , relativo alla medicina;
- CRAN, relativo alla fisica;
- CISI, relativo agli articoli scientifici.

I risultati ottenuti, come si può vedere dalla tabella sottostante, indicano che la probabilità P(A|B) ha una maggiore predizione rispetto alla P(B|A) in quanto ha mediamente un coefficiente di incertezza più elevato. Infatti calcolando la media il coefficiente di incertezza relativo a P(A|B) è U(A|B)=0,93 mentre la media del coefficiente di incertezza relativo a P(B|A) è U(B|A)=0,91.

A	B		U(A|B)		U(B|A)

intervention	authors		0,9395		0,9025
abdominal	weight		0,9655		0,9431
abortion	response		0,9447		0,8991
bacterial	species		0,9784		0,9732
controls	autistic		0,977		0,975
dogs	transformation		0,9789		0,9825
esterified	glucose		0,9736		0,9593
foetal	post		0,9517		0,9429
gamma	protected		0,8393		0,8135
hysterical	tics		0,6477		0,6477
lipids	adults		0,9607		0,9557
magna	oxygen		0,9549		0,9238
nasal	allergic		0,9292		0,9354
objects	paranoid		0,8815		0,8998
prengnancy	acids		0,985		0,9822
quantity	blood		0,9774		0,9558
radioactive	ratio		0,9709		0,9657
spontaneous	resistance		0,9712		0,9679
unions	eye		0,8064		0,7487
ventricle	pressure		0,9875		0,9837
zinc	retention		0,8907		0,8349
animals	science		0,8923		0,811
behavior	result		0,9887		0,9821
competition	technology		0,9286		0,8812

Tabella 3: Parte dei risultati relativi al campione dei digrammi

3.3 Capacità di predizione

Abbiamo eseguito i nostri esperimenti sui tre corpus d'addestramento precedentemente menzionati mediante il sistema di dialogo "LUPO-BOT" da noi realizzato che sfrutta la tecnica proposta in questo articolo. Dopo la prima fase di pre-elaborazione, la successiva fase dell'eliminazione delle stop-word e quindi la fase in cui si costruisce la matrice A parola-parola, abbiamo eseguito la TSVD

con K=100. Gli esperimenti relativi alla predizione consistono nel effettuare una domanda per la quale il nostro sistema ricerca la risposta più rilevante, questa viene parametrizzata secondo la legge di Zipf. Le parole da predire sono quindi quelle rare per il corpus d'addestramento. Il sistema individua le parole da predire con la tecnica da noi proposta e precedentemente menzionata.
In seguito saranno mostrati degli esempi di predizione di parole rare mediante la quale è possibile generare automaticamente le risposte di un chat-bot su di uno specifico argomento.
Esempio relativo a MED:

- **domanda utente**: "*tell me the electron microscopy of lung or bronch*i".
- **risposta generata dal sistema**: "*acute experimental pneumococcal(type i) pneumonia in the mouse.. the migration of leucocytes from the pulmonary capillaries into the alveolar spaces as revealed by the electron microscope. in this preliminary study of experimental pneumococcal pulmonary pneumonia in the mouse the leucocytes were observed to pass from the capillaries into the interstitial tissue and eventually into the alveolar spaces through the intercellular* ***JUNCTIONS*** *of the endothelial and epithelial cell membranes*".

La parola scritta in maiuscolo, corsivo e sottolineato nel caso specifico JUNCTIONS è la parola rara che il sistema individua come prima classificata. Vediamo nel dettaglio i vari score relativi alla sola LSA(Score LSA), alla sola probabilità condizionata(Score P(A|B)) ed alla combinazione dei due score(Score LSA+P(A|B)):

Parole	Score LSA	Parole	Score P(A|B)	Parole	Score LSA+P(A|B)
alveolar	1,0263205	***junctions***	0,25	***junctions***	0,53982866
endothelial	1,0071642	intercellular	0,03125	alveolar	0,41662574
lamellar	0,9424033	leucocytes	0,027777778	endothelial	0,4028657
epithelial	0,90259314	lattice	0,01369863	lamellar	0,38819727
spaces	0,90154886	lamellar	0,011235955	lattice	0,3733704
lattice	0,89917946	spaces	0,009259259	epithelial	0,37000588
cytoplasmic	0,8469275	epithelial	0,00896861	spaces	0,3698788
intercellular	0,82481486	granules	0,006289308	intercellular	0,36117595
amorphous	0,8214915	alveolar	0,006097561	cytoplasmic	0,33877102
junctions	0,7245716	bodies	0,005797102	amorphous	0,3285966

Tabella 4: Score relativi alle parole da individuare(Esempio MED)

Tale esempio è molto significativo perché giustifica la nostra scelta di utilizzare in combinazione con l'analisi della semantica latente anche la probabilità condizionata calcolata sui digrammi. Con questa combinazione infatti il sistema raggiunge una maggiore percentuale di successi.
Esempio relativo a CRAN:

- **domanda utente**: "*what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft* ".

- **risposta generata dal sistema**: "*similarity laws for aerothermoelastic testing. the similarity laws for aerothermoelastic testing are presented in the range. these are obtained by making nondimensional the appropriate governing equations of the individual external aerodynamic flow, heat conduction to the interior, and stress- deflection problems which make up the combined aerothermoelastic problem. for the general aerothermoelastic model, where the model is placed in a high- stagnation- temperature wind tunnel, similitude is shown to be very difficult to achieve for a scale ratio other than unity. the primary conflict occurs between the free- stream mach number reynolds number aeroelastic parameter heat conduction parameter and thermal expansion parameter. means of dealing with this basic conflict are presented. these include(1) looking at more specialized situations, such as the behavior of wing structures and of thin solid plate lifting surfaces, and panel flutter, where the aerothermoelastic similarity parameters assume less restrictive forms,(2) the use of/ incomplete aerothermoelastic/ testing in which the pressure and/ or heating rates are estimated in advance and applied artificially to the model, and(3) the use of/ restricted purpose/ models investigating separately one or another* ***FACET*** *of the complete aerothermoelastic problem. some numerical examples of* ***MODELING*** *for the general aerothermoelastic case as well as for the specialized situations mentioned in(1) above are given. finally, extension of the aerothermoelastic similarity laws to higher speeds and temperatures is discussed*".

In questo caso le parole rare individuate sono due: FACET e MODELING. Vediamo nel dettaglio i vari score delle parole del corpus relativo alla fisica mediante la seguente tabella:

Parole	Score LSA	Parole	Score P(A|B)	Parole	Score LSA+P(A|B)
facet	0,5558948	***facet***	0,33333334	***facet***	0,8892281
parameter	0,4291584	modeling	0,333333333	modeling	0,6059382
similarity	0,373224	similarity	0,005524862	parameter	0,4291584
equations	0,3526761	conduction	0,003759399	similarity	0,37874886
model	0,3519324	models	0,003456221	model	0,35349894
heating	0,3300821	model	0,00156658	equations	0,3531775
conduction	0,329044	numerical	0,001184834	conduction	0,3328034
models	0,3287734	heating	0,000941619	models	0,33222967
aerodynamic	0,3282214	aerodynamic	0,000855432	heating	0,33102375
numerical	0,3228719	equations	0,000501379	aerodynamic	0,3290768
modeling	0,7809622	***modeling***	0,333333334	***modeling***	1,1142955
facet	0,4858418	facet	0,333333333	facet	0,8191751
similarity	0,4059961	specialized	0,045454544	similarity	0,411521
laws	0,3721761	laws	0,02238806	laws	0,39456412
equations	0,3609248	similarity	0,005524862	equations	0,36142614
aeroelastic	0,2945298	conduction	0,003759399	aeroelastic	0,29452986
numerical	0,2744703	numerical	0,001184834	numerical	0,2756551

conduction	0,2599984	equations	0,000501378	specialized	0,27487248
puzzled	0,2391053	heat	0,000373902	conduction	0,26375782
approximate	0,2315538	approximate	0	puzzled	0,23910528

Tabella 5: Score relativi alle parole da individuare(Esempio CRAN)

Esempio relativo a CISI:

- **domanda utente**: "*How can actually pertinent data, as opposed to reference or entire articles themselves, be retrieved automatically in response to information request?*".
- **risposta generate dal sistema**: "*the cost- performance evaluation of the supars system is reported.. supars was an on- line, free- text bibliographic retrieval system; cost- effectiveness data of such systems are not readily available.. in our evaluation, two measures of cost were employed: a computer processing charge expressed in dollars, and the number of documents retrieved(a measure of work that must be expended to review the retrieved items).. the measure of performance was an estimate of the recall ratio.. to obtain the requisite measures an experimental plan was developed in which experts searched the data base of psychological abstracts forming their queries from written statements of information needs.. these statements(along with the list of documents relevant to them) were produced by people with information problems.. **TALLIES** were kept of the number of documents retrieved before each of the designed relevant items were found.. the major findings are noted below..(1) queries to the system employing simple boolean operators(AND, OR) have better cost- performance characteristics than queries using more elegant searching operators..(2) On- demand access to the index or dictionary contributes **SIZEABLY** to improving the cost- performance of the system..(3) the argument is raised that human factors, such as the differences among users of a system, probably should be a major factor in the design, operation and evaluation of retrieval systems.. it appears that consideration of these factors will improve system cost- performance...*".

Anche in questo caso le parole rare individuate sono due: TALLIES e SIZEABLY. Vediamo nel dettaglio i vari score delle parole del corpus relativo alla fisica mediante la seguente tabella:

Parole	Score LSA	Parole	Score P(A|B)	Parole	Score LSA+P(A|B)
queries	0,5733317	***tallies***	0,111111111	***tallies***	0,6049845
retrieved	0,5076553	boolean	0,04597701	queries	0,5898606
tallies	0,4938734	queries	0,016528925	boolean	0,5234206
boolean	0,4774436	technique	0,009242144	retrieved	0,50765526
system	0,4271326	file	0,005532504	system	0,42793688
retrieval	0,3825824	measures	0,003663004	technique	0,3903272
technique	0,381085	recall	0,003552398	retrieval	0,38413516

recall	0,3489161	procedure	0,003125	recall	0,35246852
measures	0,3389096	produced	0,002444989	measures	0,34257257
data	0,3369865	cost	0,002403846	produced	0,33891502
sizeably	0,8219073	minimizes	0,33333334	***sizeably***	0,9330184
contributes	0,4581551	***sizeably***	0,111111111	minimizes	0,61287194
cost	0,4397506	determinants	0,083333334	contributes	0,5322291
elegant	0,4158617	minimize	0,08163265	elegant	0,47141722
operators	0,4014196	contributes	0,074074075	cost	0,4397506
factors	0,3691475	elegant	0,055555556	operators	0,42364177
system	0,3652399	queries	0,024793388	factors	0,38902515
queries	0,3610575	operators	0,022222223	queries	0,38585085
costs	0,3608301	factors	0,019877676	system	0,37368548
raised	0,3363048	raised	0,014285714	minimize	0,36515078

Tabella 6: Score relativi alle parole da individuare(Esempio CISI)

Anche in questo caso la probabilità condizionata ha permesso al sistema di avere successo per la parola TALLIES che si classifica solamente terza nello score relativo all'LSA.

Dai molteplici esperimenti è risultato che:

- con la sola LSA la percentuale media di successi è di circa 84%;
- con la sola probabilità calcolata sui digrammi la percentuale media di successi è di circa 76%;
- con la combinazione delle due tecniche LSA e probabilità condizionata sui digrammi il sistema raggiunge una percentuale media di successi dell'87%.

Precisiamo che per successo del nostro sistema si intende l'individuazione della parola rara come prima classificata nella classifica relativa allo score LSA+P(A|B). I successi relativi alla predizione di tali parole permettono al sistema di generare automaticamente delle risposte alle domande poste dall'utente mantenendo una corretta sintassi ed anche una corretta semantica.

Da questi dati sperimentali, si mostra chiaramente che la tecnica proposta oltre a catturare le principali strutture sintattiche cattura anche quelle semantiche tra le parole.

4. Conclusioni

Abbiamo presentato una tecnica che sfruttando la combinazione tra LSA e la probabilità condizionata calcolata sui digrammi, basata su un'interpretazione statistica della matrice ricostruita dal TSVD, permette di generare automaticamente delle frasi predicendo delle parole rare mediante i loro contesti destro e sinistro. Questa tecnica può essere utilizzata efficacemente per realizzare un sistema di dialogo come il chat-bot che generi automaticamente delle riposte alle domande poste dagli utenti su di uno specifico argomento.

Possibili lavori futuri includono prove sperimentali più estese e costruzioni di matrici con gli stessi criteri descritti in [14] ma con un affinamento della fase di pre-elaborazione per la ricerca delle parole rare e la costruzione dei digrammi contenenti parole funzionali. Un altro sviluppo dell'attuale sistema potrebbe essere quello di integrare a quest'ultimo il modulo relativo alla componente acustica del sistema di riconoscimento vocale tenendo presente i vincoli nelle risorse dell'applicazione.

I. 5. Bibliografia

[1] A.Turing. Computing machinery and intelligence. *Mind*, 1950, 59, pp. 433-460.

[2] B.Ribeiro-Neto R.Baeza-Yates. *Modern Information Retrieval*. Addison Wesley, 1999, chap. 1, pp.1-18.

[3] Resources avalaible at: http://sra.itc.it/people/massa/thesis/node45.html .

[4] T.K. Landauer, P.W. Foltz, D. Laham, *An Introduction to Latent Semantic Analysis*, Discours Processes, 25, 1998, pp. 259-284.

[5] J.R. Bellegarda,*A Multispan Language Modelling Framework for Large Vocabulary Speech Recognition*, IEEE Transactions on Speech and Audio processing, vol. 6, no. 6, Semptember 1998.

[6] D. Widdows, S. Peters, *Word vectors and Quantum logic Experiments with negation and disjunction*, Mathematics of Language, 8, Bloomington, Indiana, June 2003, pages 141-154.

[7] Nobuhide Kitabayashi. Occurrence of Zipf's law of transition in the processes of learnig the language. *Iizuka*, 2000, pp.283-286.

[8] H. Zhu, *Bayesian Geometric Theory of Learning Algorithms*, ICNN '97,vol. 2, pp. 1041-1044.

[9] D.C. Brody, L.P. Hughston, *Geometric Models for Quantum Statistical Inference*, in Geometric Issues in the Foundations of Science, volume in honour of Roger Penrose, eds. Huggett, S.A., et al. (OUP, Oxford 1997).

[10] J.P. Levy, J.A. Bullinaria, *Learning Lexical Properties from Word Usage Patterns: Which Context Words Should be Used ?*, in R.F. French & J.P. Sougne (Eds), Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop, 273-282. London: Springer.

[11] H. Zhu, *On Information and Sufficiency*, Annals of Statistics, 1997.

[12] J. R. Pierce, "Introduction to Information Theory: Symbols, Signals and Noise 2nd rev. Ed.", *New York: Dover*, pp. 86-87 e 238-239, 1980.

[13] Sample page from NUMERICAL RECIPES IN C: THE ART OF SCIENTIFIC COMPUTING (ISBN 0-521-43108-5) Copyright (C) 1988-1992 by Cambridge University Press. Programs Copyright (C) 1988-1992 by Numerical Recipes Software. http://www.nr.com/ .

[14] S. Gaglio, G. Pilato, G. Vassallo and F. Agostaro, "A Subsymbolic Approach to Language Model Construction for Speech Recognition", CAMP '05, *International Workshop on Computer Architecture for Machine Perception*, Terrasini – Palermo, July 4 – 6 2005.

[15] Apache Jakarta. Resources avalaible at : http://jakarta.apache.org/ .

[16] Common IR Test Collection. Resources avalaible at: http://www.cs.utk.edu/~lsi/corpa.html.