Università degli Studi di Genova

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Informatica e Scienze dell'Informazione

Tesi

Sistemi di Question Answerng Multilingue: L'importante Ruolo della Traduzione

Anno Accademico
2004/2005

Relatore:
Stefano Rovetta

Correlatori:
Paolo Rosso (Università Politécnica Valencia, Spagna)
Manuel Montes-y-Gómez (INAOE Puebla, Messico)

Candidato:
Sabatino Larosa

Matricola n° 2455796

Sommario

Abstract

Attualmente nel Web è presente quasi ogni tipo di informazione. Ma senza adeguati strumenti, come *Information Retrieval* (IR), *Information Extraction* (IE) e recentemente *Question Answering* (QA), che aiutano l'utente a trovare ciò di cui ha bisogno, tutte queste informazioni sarebbero inutili. Un passo molto importante per un sistema di Question Answering multilingue è la traduzione di una domanda da una lingua sorgente ad una di destinazione. Attualmente per questo compito la maggior parte dei sistemi usa un singolo traduttore. Una cattiva traduzione può penalizzare in maniera significativa le possibilità finali di successo del sistema. Lo scopo finale di questo lavoro è quello di fornire al sistema di QA la migliore traduzione possibile. Questa viene scelta estraendola da un insieme ottenuto usando più di un traduttore. La scelta viene effettuata attraverso l'uso di tecniche implementate con basi statistiche e l'approccio usato è indipendente dal linguaggio.

1. Svolgimento del lavoro

Questa relazione descrive il lavoro che abbiamo svolto per lo sviluppo di un modulo da associare ad un sistema di QA multilingue. Un aspetto molto importante per un QA multilingue è la traduzione di una domanda da una lingua sorgente ad una di destinazione. Attualmente la maggior parte di questi sistemi usa un singolo traduttore e spesso la scelta ricade su quelli disponibili sul Web. Tuttavia la qualità dei traduttori non è molto alta e questo ha un impatto decisamente negativo sull'efficienza dei QA multilingue. Tramite questo studio si è cercato di sopperire in parte, alla carenza di qualità delle traduzioni fornite ai sistemi di QA multilingue. Il compito del modulo sviluppato è quello di scegliere la migliore traduzione prendendola da un insieme e fornendola ad un sistema di QA multilingue. L'insieme è ottenuto traducendo una domanda con un numero di traduttori differenti. Nel nostro caso, prendiamo in esame traduzioni dall'Italiano allo Spagnolo, in quanto si vogliono studiare le potenzialità offerte da un sistema di QA multilingue basato su Web. Poiché i documenti scritti in Spagnolo presenti nella rete sono sensibilmente maggiori rispetto a quelli scritti in Italiano

(2,658,631,000 pagine web in Spagnolo contro 1,845,026,000 in Italiano) [1], il sistema di QA sfrutterà la maggiore ridondanza dei primi, aumentando le possibilità di successo. Data l'idea originale, il primo passo è stato quello di studiare dei metodi che potessero fornire buoni risultati nell'estrazione della migliore traduzione in combinazione ad un numero adeguato di traduttori. I metodi sviluppati sono due - *Word-Count e Double Translation* - e sfruttano tecniche statistiche. L'idea di partenza del primo metodo è quella di confrontare le traduzioni appartenenti allo stesso insieme e scegliere quella che ha il più alto numero di parole in comune con le altre. Il secondo, invece, effettua una doppia traduzione dall'Italiano allo Spagnolo e successivamente dallo Spagnolo all'Italiano. Vengono confrontate le traduzioni dell'insieme con la domanda in italiano. Quella con la maggior attinenza verso l'originale sarà la prescelta.

Per effettuare il confronto fra le traduzioni e stabilire una similarità, sono state adottate due formule - *Dice e Coseno* - consigliate dai correlatori che hanno visionato la prima parte del lavoro. La formula di *Dice* è comunemente adottata per ottenere una similarità tra due oggetti. Viene usata in vari campi da quello genetico a quello biologico e adoperata anche in quello informatico. [2]. La formula del *Coseno* trova larga applicazione in *Information Retrieval e Information Extraction*, dove si confronta un vasto numero di documenti per stabilire qual è quello più rilevante [3]. Proprio per questa particolarità abbiamo adattato la formula originale ai nostri usi, non essendo in possesso di un grosso quantitativo di documenti, ma solo di un insieme composto da quattro traduzioni.

La reperibilità di un numero sufficiente di traduttori che potesse svolgere il nostro compito è stata ardua. Infatti, sono pochi i traduttori disponibili che forniscono una traduzione diretta per la coppia di lingue da noi scelta e la maggior parte sono a pagamento. Perciò, abbiamo attuato una scelta intermedia fra traduttori che provvedono o meno ad una traduzione diretta. Si è quindi optato per un numero minimo, di quattro traduttori, che potesse essere in grado di offrire delle prime valutazioni sulle effettive potenzialità dei metodi. Ci siamo affidati ai seguenti: Power Translation Pro7[1], IdiomaX[2], Google[3], FreeTranslation[4].

[1] Power Translation Pro: www.lec.com
[2] idioma: www.IdiomaX.com
[3] Google: www.google.it/language_tools
[4] FreeTranslation: www.freetranslation.com

I primi due sono applicazioni software, mentre gli ultimi due sono disponibili on-line. Google e FreeTranslation non consentono la traduzione in maniera diretta dall'Italiano allo Spagnolo, di conseguenza, ho dovuto effettuare una traduzione intermedia in Inglese.

Per effettuare lo sviluppo del software ci siamo basati sul linguaggio Java e sul tool di sviluppo Ecplise[5]. Ho iniziato questa esperienza durante la fine del mio soggiorno Erasmus presso l'Universidad Politécnica de Valencia (Spagna) nell'ottobre del 2004. Ho collaborando inizialmente con i professori Paolo Rosso (docente presso l'Universidad Politécnica de Valencia, Spagna) e Manuel Montes-y-Gómez (docente presso il Laboratorio di Tecnologie del Linguaggio dell'Instituto Nacional de Astrofísica, Optica y Electrónica, Messico) che mi hanno fatto partecipe dell'iniziativa.

Il lavoro è proseguito anche dopo il mio ritorno in Italia, grazie al supporto del professor Stefano Rovetta (docente presso il Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova). Questi studi hanno portato come risultato alla redazione di due scritti, presentati in conferenze internazionali. Nel giugno del 2005, a Tetuan, in Marocco (iCtiS'05), è stato presentato l'articolo intitolato: "Best Translation for an Italian-Spanish Question Answering System" (vedasi Appendice A) e prodotto in collaborazione con i docenti. Sempre nel giugno del 2005 ho partecipato ad un workshop (CLIP2005) presso il DISI, Università degli Studi di Genova, con una relazione dal titolo: "Cross-Language Question Answering by Multiple Automatic Translations". Attualmente, è stato accettato un ulteriore lavoro, frutto della collaborazione con i medesimi docenti, intitolato: "Cross-language Question Answering: The Key Role of Translation" (vedasi Appendice B) che sarà presentato alla fine di settembre 2005, a Puebla, Messico (ENC'2005).

2. Introduzione al Question Answering

Attualmente, nel Web, è presente quasi ogni tipo di informazione in più di 1,500 lingue e in formato elettronico. Ma senza adeguati strumenti che aiutino l'utente a trovare le informazioni come *Information Retrieval* , *Information*

[5] Eclipse: www.eclipse.org

Extraction e recentemente *Question Answering* , tutte queste informazioni sarebbero inutili. Lo scopo di un sistema di QA è quello di aiutare utenti poco esperti a trovare risposte precise, formulando le domande in linguaggio naturale. Un attuale motore di ricerca permette all'utente di trovare documenti rilevanti per i suoi bisogni, ma è incapace di fornire una risposta precisa ad una specifica richiesta di informazione. L'alternativa ai motori di ricerca per risolvere una specifica richiesta di informazione sono i sistemi di QA. Se ad esempio un utente formulasse la seguente domanda: *"Dove si trova il Colosseo?"* il sistema risponderebbe: *"Roma"*, anziché fornire una lunga serie di documenti correlati al Colosseo. Esistono sistemi di QA che fanno uso di numerose risorse linguistiche per comprendere la domanda e fare la ricerca della risposta. Il principali problemi derivano dalla complessità dell'approccio e dallo stretto legame con uno specifico linguaggio. Quindi l'idea principale per superare questi ostacoli è un approccio che sfrutti la vastità di documenti presenti nel Web. Il sistema di QA basato sul Web sfrutterà quindi la ridondanza dei documenti in modo statistico e indipendente dal linguaggio.

Un sistema di questo tipo è stato sviluppato presso il Laboratorio di Tecnologie del Linguaggio, INAOE, Messico [4]. Fornita una domanda in input il sistema crea una combinazione di parole che si ottiene manipolando l'ordine delle parole nella domanda. Queste combinazioni, saranno passate ad un motore di ricerca (es. Google) per fare delle ricerche in rete. Per ogni combinazione il sistema colleziona un certo numero di snippet (la parte di un documento rilevante, che contiene quasi tutte le parole della query). Infine, le possibili risposte sono estratte tramite basi statistiche, e viene fatta un classifica delle risposte candidate. Quindi le fasi principali di questo sistema sono: la riformulazione della query, la collezione degli snippet, l'estrazione della risposta.

Inoltre, esistono sistemi di QA multilingue basati su Web, che permettono all'utente di formulare una domanda in una lingua differente da quella dei documenti che il sistema userà per fornire la risposta. Cercare la risposta in documenti scritti in un'altra lingua da quella della domanda permette di sfruttare la ridondanza dei documenti presenti nel Web. Cercando la stessa informazione in un insieme di documenti scritti in una lingua più comune, si aumentano le possibilità di ottenere la risposta esatta. [5-6-8-9-10] Inoltre, può capitare di dover cercare la risposta in documenti che non sono stati tradotti nella nostra lingua di

appartenenza. Un aspetto molto importante per un QA multilingue è la traduzione di una domanda da una lingua ad un'altra. Attualmente la maggioranza dei sistemi usa traduttori disponibili on-line. Sono in corso alcuni studi su Sistemi di Qa multilingue con varie coppie di lingue (Catalano-Spagnolo, Arabo-Inglese) [13] dove si vuole sottolineare l'importanza della traduzione.

3. Problematiche della Traduzione

Chi ha provato almeno una volta nella sua vita ad usare dei software di traduzione automatica, sa bene che la speranza di ottenere traduzioni di documenti vicine al linguaggio naturale è abbastanza remota. A volte i testi prodotti mostrano ambiguità comiche ed errori madornali che sarebbe assai pericoloso lasciare in documenti di lavoro. Questo accade perché i processi linguistici sono ancora oggetto di studio e di dibattito tra diverse scuole di pensiero, come gran parte delle attività psichiche e cerebrali umane.

Le cose si complicano se si tenta di "insegnare" a una macchina a riconoscere il linguaggio e a riprodurlo. Dietro a un software di traduzione automatica, che spesso viene scartato con giudizi impietosi, c'è un bagaglio di ricerca immenso, che coinvolge discipline scientifiche e umanistiche quali la linguistica, la linguistica computazionale, l'informatica, l'intelligenza artificiale.

Eppure alle origini dell'Intelligenza Artificiale e dell'Elaborazione del Linguaggio Naturale, il settore dedicato alla Traduzione Automatica era animato dall'ambiziosa speranza che "offrendo traduzioni parola per parola sarebbe stato possibile comprende ciò che si diceva senza l'ausilio di un traduttore umano"[7].

Questa speranza non si realizzo mai e oggi lo scopo della Traduzione Automatica non è quello di tradurre il testo così come farebbe un traduttore umano, ma di offrire come risultato un testo in una diversa lingua che faccia capire almeno il succo di quanto è scritto nel testo originale. Questa evoluzione ha certamente reso più fattibile l'impresa della Traduzione Automatica, ma riuscire a 'far capire il succo' di un testo implica che la traduzione non sia un processo di codifica e decodifica che richieda il semplice passaggio da un codice a un altro. Al contrario richiede un processo di comprensione e di riformulazione del messaggio. L'ostacolo principale è costituito dalla semantica, ovvero la capacità di

caricare i segni di contenuti, o viceversa astrarre i contenuti dai segni, elaborare concetti e relazioni tra concetti, comprensione dei contesti[11].

I problemi legati alla traduzione come ambiguità, omonimia, polisemia etc. spesso richiedono il ricorso al contesto immediato o ad informazioni extralinguistiche e pertanto la loro soluzione non dipende solo dalla competenza del parlante, ma dall'uso effettivo di ciascuna lingua. Una parola polisemica che può essere presente in una comune frase è ad esempio: "calcio". Nella lingua Italiana questa parola può essere interpretata sia come sinonimo di uno sport, oppure di un elemento chimico. La sua disambiguazione è insita nel contesto della frase che si vuole tradurre, ma questo è solo uno dei tanti problemi in cui si può imbattere un traduttore automatico. Non bisogna dimenticare possibili errori di ortografia e di grammatica presenti nel testo da tradurre. Nel primo caso la soluzione è quella di confrontare ogni singola parola con il relativo archivio disponibile , se la parola appartiene al database allora verrà considerata valida, in caso contrario verrà segnalato un errore. Errori di grammatica sono di più difficile individuazione poiché anche in questo caso ci imbattiamo nell'analisi del contesto[12].

4. Metodo Word-Count

Questo metodo, sfruttando la ridondanza delle parole, sceglie come migliore traduzione quella che ha il più alto numero di termini in comune rispetto alle altre. Più la frequenza di un termine è maggiore e molto più probabilmente questo indica che sarà stato tradotto correttamente. Per stabilire il numero di termini in comune e calcolare la similarità tra le traduzioni sono state scelte due formule: la formula di *Dice* e la formula del *Coseno.*
Dato un insieme di traduzioni in Spagnolo queste vengono intersecate tra loro per trovare il numero di parole in comune.

Esempio di domanda tradotta con quattro traduttori differenti:

Che cosa significa la sigla CEE?

1. ¿Qué significa la sigla CEE?
2. ¿Qué cosa significa siglas el EEC?
3. ¿Qué significa la CEE de la abreviación?
4. ¿Qué cosa significa la pone la sigla CEE?

Tabella 1. Risultati dell'intersezione con gli esempi precedenti. "No" significa che l'intersezione di una traduzione con se stessa non è considerata.

	1 Tran.	*2 Tran*	*3 Tran.*	*4 Tran.*
1 Tran.	No	2	4	5
2 Tran.	2	No	2	3
3 Tran.	4	2	No	5
4 Tran.	5	3	5	No

La formula di Dice viene usata per stabilire il grado di similitudine tra le traduzioni e creare una gerarchia, sfruttando l'informazione che hanno in comune, ovvero le parole.

$$sim(t_i, t_j) = \frac{2 \times len(t_i \cap t_j)}{len(t_i) + len(t_j)} \quad (1)$$

- t1 e t2 sono le traduzioni che prendiamo in considerazione.
- (t1 ∩ t2) rappresenta l'intersezione (numero di parole in comune).
- len1 e len2 rappresentano il numero di parole di cui è composta ogni traduzione.

Ad ogni traduzione viene associato un grado di similitudine, che si ottiene svolgendo il calcolo descritto dalla formula precedente, tra la traduzione di

riferimento e le altre dell'insieme. Infine sommando tra loro i risultati parziali si ottiene il grado di similitudine desiderato. Ad esempio per ottenere il valore della traduzione numero uno dovremmo fare: ($Sim_{trad1\text{-}trad2}$ + $Sim_{trad1\text{-}trad3}$ + $Sim_{trad1\text{-}trad4}$). Infine si sceglierà la traduzione che ha ottenuto il maggior grado di similitudine. Per aumentare la precisione nella scelta della traduzione migliore, sono stati usati gli n-grammi (sequenze di n parole). L'uso degli n-grammi è stato sino ai trigrammi.

Esempio di bigramma della frase *"Qué significa la sigla CEE":*
"Qué significa" "significa la" "la sigla" "sigla CEE"

Gli n-grammi sono molto utili in casi in cui si debbano confrontare traduzioni formate da stesse identiche parole, ma poste in ordine diverso. Se per esempio ci sono due traduzioni che hanno le stesse identiche parole ma in ordine differente, grazie agli n-grammi possiamo migliorare la precisione del calcolo per ottenere il grado di similarità.

Il metodo word-count è stato implementato anche con la formula del coseno per calcolare il grado di similarità. In questo modello le traduzioni sono rappresentate come vettori in uno spazio t-dimensionale (t è il numero complessivo di termini indice o keywords). Per calcolare i pesi delle keywords si utilizza lo schema di pesatura TermFrequency-InverseDocumentFrequency (tf-idf).
Tutte le parole facenti parte dell'insieme di traduzioni vengono considerate come keywords (sono prese una sola volta senza ripetizioni).

Esempio di domanda tradotta con quattro traduttori differenti:

Qual è la capitale della Repubblica del Sud Africa?

1. ¿Cuál es la capital de la República de la Sur África?
2. ¿Cuál es entendido ellos de la república de la África del sur?
3. ¿Cuál es la capital de la República del Sur una Africa?
4. ¿Cuál es el capital de la república del sur Africa?

Si ottiene il seguente elenco di parole chiave:

“cuál” “es” “la” “capital” “de” “república” “sur” “África” “entendido” “ellos” “del” “una” “africa” “el”

Dopo ciò si determina la freq(i,j), la frequenza di ogni keywords (k_i) per ogni traduzione.
Per calcolare il peso per ogni traduzione ci serviamo della seguente formula:

$$t_{ij} = f_{ij} \times \log(1 + \frac{n_i}{N}) \quad (2)$$

dove:
- N = numero totale di traduzioni nell’insieme
- ni = numero di documenti che contengono k_i
- f(i,j) = freq(i,j) / max(freq(l,j))

  rappresenta la frequenza della parola chiave nella traduzione, divisa per il massimo, calcolato su tutte le parole chiave di quella traduzione. La divisione è fatta per normalizzare il risultato.

La formula differisce dall’originale di Salton [3] per la presenza del termine 1 nel log, questo perchè se ci fossero casi in cui N=ni il log non sarebbe uguale a 0. Si fa presente che la formula originale viene usata con una grande collezione di documenti e le probabilità di avere lo stesso termine in ogni documento sono quasi nulle. Nel nostro caso si ha a disposizione un piccolo gruppo di traduzioni ed è molto frequente avere un termine in tutte le traduzioni.
Al termine di questo passaggio, otteniamo per ogni traduzione, il vettore contenente i pesi associati ad ogni parola chiave:

t1: [1.33, 1.33, 4, 0.62, 2.6, 1.33, 1.33, 0.35, 0, 0, 0, 0, 0, 0]
t2: [2, 2, 4, 0, 4, 2, 2, 0.5, 0.3, 0.3, 0.93, 0, 0, 0]
ecc...

Ottenuti i vettori dal passo precedente, si procede con il calcolo del grado di similitudini tra traduzioni tramite la seguente formula:

$$sim(t_i, t_j) = \frac{\left(\sum_{\forall k} t_{ik} \times t_{jk}\right)}{\sqrt{\sum_{\forall k} t_{ik}^{\,2}} \times \sqrt{\sum_{\forall k} t_{jk}^{\,2}}} \quad (3)$$

Nella formula tji e tqi rappresentano due generici vettori contenenti i pesi.

Il calcolo finale si esegue in questo modo:
Trad1 = Sim(t1,t2) + Sim(t1,t3) + Sim(t1,t4)
Trad2 = Sim(t2,t1) + Sim(t2,t3) + Sim(t2,t4)
Trad3 = Sim(t3,t1) + Sim(t3,t2) + Sim(t3,t4)
Trad4 = Sim(t4,t1) + Sim(t4,t2) + Sim(t4,t3)

Viene scelta la traduzione che ha ottenuto il valore più alto.

5. Metodo Double-Translation

Ogni domanda in Italiano viene tradotta in Spagnolo, quindi ritradotta nuovamente in Italiano. Si utilizzano quattro traduttori ottienendo quindi un nuovo insieme. Verrà scelta la traduzione dell'insieme che risulta più simile alla domanda originale. Anche in questo caso ci siamo serviti della formula di *Dice* e successivamente della formula del Coseno. Gli algoritmi usati sono quelli precedentemente illustrati, vi sono semplicemente delle piccole variazioni.

Esempio di domanda originale e doppia traduzione:

Che cosa significa la sigla CEE?

1. Che cosa significa la sigla CEE?
2. Che cosa significa le abbreviazioni il EEC?

3. Che significa il CEE dell'abbreviazione?
4. Che cosa ha importanza la mette la sigla di CEE?

Anche con questo metodo vengono usati gli n-grammi sino ai trigrammi.

Per quanto riguarda l'implementazione con la formula di Dice, la differenza con il precedente algoritmo riguarda l'intersezione tra traduzioni. Infatti in questo metodo vengono fatte intersezioni tra la domanda originale e le domande ritradotte. Quindi viene usata la formula di Dice per ottenere il grado si similarità. Infine viene scelta la traduzione con il più alto grado si similarità .

Anche nell'implementazione con la formula del coseno vi sono delle differenze rispetto al primo metodo. In questo metodo creiamo una lista di keywords includendo anche la domanda originale. Quindi il passo successivo è quello di calcolare i vettori contenenti i pesi, ma se per per le domande ritradotto usiamo la formula (2), per la domanda originale useremo la seguente formula:

$$(0.5+[0.5* f(i, j)]) * \log(1+\frac{n_i}{N}) \quad (4)$$

Questa formula viene usata in quanto la domanda originale viene confrontata ad una Information Retrieval query. Questa differisce dall'originale di Salton & Bucley per le medesime ragioni illustrate in precedenza. Ottenuti i vettori contenenti i pesi si applica la formula (3) tra il vettore contenente i pesi della domanda originale e gli altri vettori. Infine viene scelta la traduzione con il grado di similarità più elevato. Anche questo metodo fa uso degli n-grammi sino ai trigrammi.

6. Risultati Ottenuti

Il Clef[6] (Cross-Language Evaluation Forum) è un consorzio europeo che organizza una competizione a livello internazionale riguardante sistemi di

[6] Cross-Language Evaluetion Forum: www.clef-campaign.org

information retrieval, operanti su lingue europee in ambito monolingua e multilingua. Gli esperimenti sono stati effettuati utilizzando un set di 450 domande fattuali e derivanti dalla gara del Clef 2003. Per ogni domanda in Italiano, si ottiene un insieme di quattro traduzioni in Spagnolo (*Word-Count*) oppure quattro doppie traduzioni (*Double Translation*).
Le tabelle seguenti mettono a confronto i risultati ottenuti usando i differenti traduttori e applicando le tecniche spiegate precedentemente. Per ogni esperimento vengono indicate le percentuali di successo e il numero di domande.

Unigrammi	*Bigrammi*	*Trigrammi*
51,33 %	51,11 %	51,55 %
231 su 450	230 su 450	232 su 450

Word-Count e formula Dice

Unigrammi	*Bigrammi*	*Trigrammi*
46,66 %	49,11 %	50,22 %
210 su 450	221 su 450	226 su 450

Double Translation e formula Dice

Unigrammi	*Bigrammi*	*Trigrammi*
48,66 %	49,33%	50,00%
219 su 450	222 su 450	225 su 450

Word-Count e formula Coseno

Unigrammi	*Bigrammi*	*Trigrammi*
45,77 %	48,44%	49,11%
206 su 450	218 su 450	221 su 450

Double Translation e formula Coseno

Ogni tabella mostra la percentuale di successo e il numero di domande che sono state tradotte in ogni esperimento.

7. Analisi dei risultati

Da questi esperimenti emerge che alcuni traduttori producono una traduzione di bassa qualità. In particolare il riferimento va a Google e FreeTranslation, ovvero i traduttori on-line. Questo probabilmente dovuto al fatto che bisogna effettuare una traduzione intermedia in Inglese per giungere alla traduzione finale in Spagnolo. Come conseguenza, vi sono casi dove la cattiva traduzione e di conseguenza una cattiva ridondanza penalizzano l'elezione della migliore traduzione, specialmente nel secondo metodo (Doble-Translation). Il traduttore che ha ottenuto i migliori risultati è il Power Translator Pro (55.33%). Questo traduttore si è dimostrato più efficace del nostro migliore risultato (51.55%) ottenuto con il metodo word-count. Tuttavia questi risultati preliminari sembrano essere promettenti. Infatti una combinazione ottimale tra i due metodi potrebbe aumentare le percentuali di successo. Tramite una stima approssimativa si potrebbe arrivare ad un incremento del 20%. Questo è dovuto al fatto che le scelte effettuate dai due sistemi non sono le stesse. Infine è stato fatto un ulteriore esperimento per capire come combinare al meglio i due metodi. Sono state divise le domande in categorie (date, persone, organizzazioni, luoghi e misure) e si sono confrontati i risultati ottenuti dal traduttore di riferimento (Power Translation Pro7) e dai due metodi.

Domande separate per categorie

	Date	*Persone*	*Organizzazioni*	*Luoghi*	*Unità di Misura*
Numero di Domande	44	71	26	61	77
WordCount Dice (1-gram)	--	--	**46%**	59%	**58%**
WordCount Dice (2-gram)	--	--	--	--	**58%**
Double Trans Dice (2-gram)	61%	--	--	--	--
Double Trans Dice (3-gram)	61%	**64%**	--	--	--
Double Trans Cosine (3-gram)	61%	--	--	--	--
Traduttore di Riferimento	70%	64%	42%	72%	40%

La tabella mostra i risultati ottenuti con i due metodi e con il miglior traduttore. Per ogni metodo compare solo la migliore percentuale ottenuta tra i metodi. I numeri in neretto indicano che il metodo è stato in grado di ottenere una migliore performance rispetto al traduttore di riferimento.

8. Conclusioni e Sviluppi Futuri

Con questo lavoro si è voluto investigare sulla possibilità di aumentare la qualità di traduzione per un sistema di QA multilingue. Sono stati sviluppati due metodi totalmente statistici e indipendenti dal linguaggio.

I risultati preliminari risultano essere promettenti considerando il fatto di aver usato un ristretto numero di traduttori, e che inoltre due di essi non permettono la traduzione in maniera diretta dall'Italiano allo Spagnolo. L'ultimo esperimento effettuato con la suddivisione delle domande ha dimostrato che in alcuni casi i metodi ottengono dei risultati migliori rispetto al traduttore di riferimento. Sono necessari ulteriori esperimenti per trovare una combinazione ottimale tra i metodi con lo scopo di aumentare le percentuali di successo.

Nell'edizione del Clef-2006, molto probabilmente, il task che vedrà impegnati i sistemi di QA multilingue sarà many-to-many. Ovvero, le domande fornite saranno in più di una lingua e i sistemi dovranno essere in grado di trovare la risposta corretta in collezioni di documenti multilingue. Quindi, per il successo finale da parte dei sistemi di QA multilingue, la buona riuscita del processo di traduzione delle domande, sarà di fondamentale importanza.

9. Bibliografia

[1] Kilgarriff A. y Greffenstette G., Introduction to the Special Issue on Web as Corpus, Computational Linguistics 29(3), pp.1-15, 2003.

[2] Lin D. 1998, An information-theoretic definition of similarity. Proceedings 15th International Conf. on Machine Learning.

[3] Baeza-Yates, R., and Ribeiro-Neto, B., 1999.,Modern Information Retrieval. Addison-Wesley.

[4] M. Del Castillo, M. Montes y Gómez, and L. Villaseñor, "QA on the web: A preliminary study for Spanish language", Proc. of the 5th Mexican Int. Conf.on Computer Science (ENC), Colima, Mexico, 2004.

[5] E. Brill, J. Lin, M. Banko, and S. Dumais, "Data-intensive question answering", Proc. TREC-10, 2001.

[6] Solorio T. Perez M. Montes M. Villaseñor L. E Lopez. A 2004. "A language independent Method for Question Classification". Inc proc of the 20th Int. Conf. on Computationla Linguistic (COLING-04) Geneva, Switwerland.

[7] E. Charniak, D. McDermott, *Introduction to Artificial Intelligence,* Addison Wesley Publishing
Company, Reading, MA, 1985, p. 172.

[8] E. Hovy, L. Gerber, U. Hermajakob, M. Junk, and C. Lin, "Question answering in Webclopedia", Proc. TREC-9, 2000.

[9] C. Kwok, O. Etzioni, and D. Weld, "Scaling question answering to the Web", Proc. of the WWW Conference, 2001.

[10] J. Lin, J., "The Web as a resource for question answering: perspectives and challenges", Proc. of the 3rd Int. Conf. on Language Resources and Evaluation (LREC), 2002.

[11] N. Ruimy, E.Gola, "Traduzione Automatica e processi di comprensione", http://www.ilc.cnr.it

[12] A. De Simone, S.Ciampoli, "Do you speak pc?", Pc Magazine Italia, Aprile 2005, p. 150.

[13] P. Rosso, A. Lyhyaoui, J. Peñarrubia, M. Montes y Gómez , Y. Benajiba, and N. Raissouni, "Arabic-English Question Answering", Proc. of Information Communication Technologies Int. Symposium (ICTIS), Tetuan, Morocco, June 2005.

Un particolare ringraziamento va ai professori Paolo Rosso e Manuel Montes, che tramite il loro aiuto e il costante supporto, mi hanno permesso di svolgere questo lavoro. Un grazie anche al contributo fornito dal professor Stefano Rovetta.