UNIVERSITÀ
CA' FOSCARI
VENEZIA

DIPARTIMENTO
DI SCIENZE
DEL LINGUAGGIO

STILVEN 2008:
Un Traduttore INGLESE-VENETO e viceversa
RELAZIONE CONCLUSIVA – Luglio 2008

1. Risultati in nuce

Il progetto preliminare finanziato con 20 mila euro ha prodotto i primi risultati, che sono i seguenti:
- creazione di lessici di multiwords inglese-italiano e viceversa estratti da materiali disponibili online e messi a disposizione dal sottoscritto. Questi materiali dovranno poi essere tradotti in veneto;
- creazione di un lessico di base veneto-italiano (circa 1800 voci), con entrate completamente specificate sulla base di un lessico preesistente per la lingua italiana, da utilizzare nel sottoprogetto XLE e in quello del traduttore per interlingua;
- raccolta di materiali testuali in lingua veneta dai siti più affidabili e loro elaborazione per analisi quantitative e ortografiche – questi sono in forma grezza e necessitano un lungo lavoro di normalizzazione della ortografia, e quindi non vengono inclusi nel CD;
- ricerca di tutti i siti web utili per monitorare l'uso di veneto scritto e per verificarne l'affidabilità;
- inizio di attività del sottoprogetto XLE – presentiamo un breve resoconto;
- creazione delle infrastrutture di base per il prototipo di traduttore automatico fondato sui sistemi MOSES e GIZA – presentiamo i primi risultati.

2. Problemi affrontati

Il problema fondamentale che abbiamo dovuto affrontare è stato quello rappresentato dalla variabilità dell'ortografia. Sia per quanto riguarda i testi tradotti e pervenutici dal responsabile di Veneto.org, il dott. Pizzati, sia ancora per quanto riguarda i testi che sono disponibili online su siti web, l'ortografia è sempre soggetta a variazioni cospicue che rendono i testi inutilizzabili per il traduttore e in genere per qualsiasi attività scientifica in forma digitale.
Perché si possa utilizzare un testo digitale scritto in veneto, è indispensabile che una qualsiasi parola venga scritta sempre nello stesso modo. Non ci possono essere ortografie variabili in quanto il computer non sarà più in grado di associare il significato voluto alla parola, né tanto meno sarà in grado di associare alla parola la coppia di traduzione che permetterà di trasformare un testo in inglese nel corrispondente testo veneto e viceversa.
Quindi i testi paralleli – testi inglesi e corrispondenti testi veneti – sono stati verificati manualmente parola per parola, in modo da assicurarsi che non ci fossero variazioni ortografiche indesiderate. Ovviamente, non è sempre stato possibile eliminare le variazioni. In alcuni casi, la variazione ortografica è stata valutata in maniera diversa in quanto la presenza di certe forme morfologiche è compatibile con e quindi siamo giunti alla conclusione che il traduttore ha adottato uno stile leggermente diverso, o semplicemente, quel testo è stato tradotto da un altro traduttore.

Un secondo problema riguarda la necessità di disporre di dati linguistici di partenza indispensabili per migliorare la qualità finale della traduzione. A questo scopo abbiamo fatto una serie di studi di base del corpus veneto a nostra disposizione che ci permetteranno di capire quale sarà il livello di ambiguità dei testi. Fondamentalmente ci serve capire quali e quanto sono le parole omografe ma con significati diversi che quindi contribuiranno a creare ambiguità per il traduttore automatico. A questo scopo abbiamo creato liste di "coppie minime" per il veneto che accludiamo al CD, sulla base dei testi studiati.

SEGRETERIA DIDATTICA
tel. 041 234 5704-5705
fax 041 234 5706
filnia@unive.it
vanessa@unive.it

SEGRETERIA AMMINISTRATIVA
tel. 041 234 5780
fax 041 234 5703
mcduse@unive.it

CA' BEMBO
Dorsoduro, 1075
30123 VENEZIA

COD. FISC. 80007720271

UNIVERSITÀ
CA' FOSCARI
VENEZIA

DIPARTIMENTO
DI SCIENZE
DEL LINGUAGGIO

3. Il problema dell'ambiguità

Per fronteggiare questo problema abbiamo agito su due fronti:

- da un lato abbiamo creato le condizioni per ridurre l'ambiguità complessiva dei testi attraverso la ricerca, la classificazione e la traduzione – questo in un prossimo futuro – delle parole collocate, dette anche polirematiche e in inglese "nultiwords" termine questo che preferiamo in quanto neutrale. Si tratta di sequenze di parole che hanno significato congelato nel lessico, sono quindi di natura idiomatica o ricevono un significato non composizionale. In altri casi si tratta di collocati, cioè di associazioni preferenziali tra parole che vengono usate frequentemente. Abbiamo quindi proceduto in un primo tempo per approssimazione facendo riferimento a un'altra lingua, l'italiano, della quale esistono grandi quantità di testi paralleli con la lingua inglese. In nostro possesso c'era una lista di multiword dell'italiano di 10mila entrate e una lista di multiword dell'inglese, quest'ultima di 120mila entrate. Abbiamo acquisito un corpus parallelo messo a disposizione dalla Commissione Europea, di 80 milioni di parole contenente tutti i documenti prodotti e tradotti dalla commissione negli ultimi venti anni. Attraverso l'uso di strumenti automatici e di una rete di computer per fare girare i programmi molto pesanti – non funzionanti su un solo computer – abbiamo estratto le coppie di traduzione dall'inglese all'italiano e viceversa. Queste coppie verranno in futuro tradotte in veneto, da un parlante veneto che ne verifica la possibile appartenenza alla lingua.
- Da un altro lato, abbiamo studiato in maniera esaustiva il corpus in nostro possesso allo scopo di determinare quale sia la cosiddetta lista di frequenza dei tipi, che ci ha permesso da un lato di individuare gli hapax legomena che normalmente includono errori di ortografia, da un altro lato ci ha dato il cosiddetto lessico di frequenza.

Ovviamente, l'omografia viene in parte eliminata grazie alla presenza degli accenti all'interno di parola che servono due funzioni: individuare l'accento di parola, e differenziare le "e" e le "o" aperte e chiuse. Questa operazione però non è sempre sufficiente a garantire una disambiguazione completa. A questo scopo abbiamo creato la lista di "coppie minime" cioè parole che si differenziano per una sola lettera o fonema.

4. Il lavoro per la creazione del software di traduzione

Anche qui abbiamo operato su più fronti:
4.1 Da un lato abbiamo utilizzato i testi paralleli per iniziare il training. Questo passo però è avvenuto solo alla fine del processo di preparazione dei testi e non ci sono quindi allo stato attuale risultati concreti di traduzione effettiva. Il training serve per creare il modello statistico-matematico dedotto dai testi paralleli dopo che questi sono stati accuratamente allineati. E' attraverso il modello che poi avverrà la traduzione automatica. Per poter allineare i testi è necessario che non ci siano sbavature per quanto riguarda la corrispondenza tra frasi. In altre parole, si richiede minimamente che ad ogni frasi del corpus veneto corrisponda una frase del corpus inglese.

SEGRETERIA DIDATTICA
tel. 041 234 5704-5705
fax 041 234 5706
filnia@unive.it
vanessa@unive.it

SEGRETERIA AMMINISTRATIVA
tel. 041 234 5780
fax 041 234 5703
mcduse@unive.it

CA' BEMBO
Dorsoduro, 1075
30123 VENEZIA

COD. FISC. 80007720271

UNIVERSITÀ
CA' FOSCARI
VENEZIA

DIPARTIMENTO
DI SCIENZE
DEL LINGUAGGIO

Questa corrispondenza va verificata accuratamente e quindi è stata fatta a mano. A quel punto, si può far partire un programma che tenta di stabilire una corrispondenza parola per parola, automaticamente, sulla base delle più frequenti coincidenze. Dal momento che il corpus è molto piccolo, questa feature – caratteristica – fondamentale è difficilmente verificabile. Quindi è stato adottato un algoritmo genetico che ha lo scopo di approssimare questa corrispondenza in tutti i casi in cui i dati sono carenti. Riporto qui in basso il funzionamento di questo algoritmo in inglese, come ci è stato fornito dal ricercatore:

"The aim is to combine many features to judge the quality of aligning two sentences. The algorithm can be applied for two files of text. Remarkably, it can be executed from up to down or down to up, in both cases from source to target or from target to source. Four possibilities then. In the ideal case they should give the same result. One could augment his confidence about an alignment by taking the intersection of all possibilities. As for now, I am executing it just once. The advantage of using a genetic algorithm is that it is quite suitable for a parallel implementation, which would strongly improve the execution time (while giving excellent solutions). For a GA to be implemented we need:
- A representation scheme (in our case to represent a possible alignment)
- An evaluation function (fitness)
- A reproduction strategy (selection, crossing over, mutation)

4.1.1. The representation scheme
A chromosome in our case is meant to express which sentences from source are to be aligned to which sentences from the target. I tried many representations (lists) whose implementation was difficult and time-consuming. The best representation I could come up with is to represent a possible alignment as a sequence of movements between source-target sentences. we can imagine for example that the source sentences are the columns of a matrix and the target sentences are the lines. Consequently, a chromosome is a sequence of movements between cells in this matrix. We have three possible movements, right, down and diagonal. Some constraints are imposed on the chromosomes. For example, consecutive perpendicular moves are not allowed, as they will result in aligning a sentence more than once. It is clear that the chromosomes are of variable length (which will be a bit difficult for crossing and mutating). This schema allows us to represent one to one, many to one, one to many, and many to many alignments.

4.1.2. The evaluation function
A chromosome is evaluated in two steps. First, every pair of aligned sentences is evaluated. The criteria I used are:
- Difference in the number of words (Gale & Church), to be minimized
- Number of words lexically related (using the small dictionary we have), to be maximized
- Number of similar named entities (they should be written in a similar way), to be maximized
- Number of similar special punctuation, to be maximized
Second, an overall evaluation of the chromosome is evoked.
- Minimize many to many alignments as much as possible (which means keep close to diagonal)

SEGRETERIA DIDATTICA
tel. 041 234 5704-5705
fax 041 234 5706
filnia@unive.it
vanessa@unive.it

SEGRETERIA AMMINISTRATIVA
tel. 041 234 5780
fax 041 234 5703
mcduse@unive.it

CA' BEMBO
Dorsoduro, 1075
30123 VENEZIA

COD. FISC. 80007720271

UNIVERSITÀ
CA' FOSCARI
VENEZIA

DIPARTIMENTO
DI SCIENZE
DEL LINGUAGGIO

- Maximize the number of moves (in order to cover all sentences) .
Once those features are evaluated for a given chromosome, each one is normalized, so that the combination will not be dominated by the feature with large values. The normalization I used here was by dividing by the sum of all values. Finally, a weighted sum of the normalized values is considered as the fitness of the corresponding chromosome. The weights are set subject to personal preference (no optimization was done to select the weights)

4.1.3. The reproduction
According to my experience the tournament selection gives better results than the roulette wheel. The selection, crossover, and mutation were implemented canonically. However, it should be mentioned that:
- The crossing point is chosen randomly according to the chromosome of the min length of the two parents
- We keep trying a number of times to cross or to mutate. Since the chosen point could lead to incorrect chromosome
-After crossing or mutating, a revision of the resulting child/children is performed, as the moves could go beyond the sentences."

Si potrà verificare successivamente il risultato di questo algoritmo attraverso il training. Sul CD saranno presenti i primi risultati dell'allineamento e il corpus allineato.

4.2 Dall'altro lato abbiamo posto le condizioni per la creazione del traduttore per regole che utilizzerà i dati ricavati dal sistema XLE. Questo sistema è reso disponibile gratuitamente per fini di ricerca a istituzioni pubbliche, dalla XEROX di Palo Alto, e per il quale abbiamo firmato una regolare licenza di uso. Il sistema funziona utilizzando la teoria LFG – ideata dalla prof. Joan Bresnan dell'Università di Stanford - e gli algoritmi prodotti da Martin Kay e Ron Kaplan appunto della XEROX. A questo sistema verranno forniti i dati lessicali di basi e le regole grammaticali del veneto. Un ricercatore ha iniziato, purtroppo in ritardo per abbandono del precedente incaricato, a lavorare al sistema. Il test set di frasi che utilizzeremo è quello utilizzato dall'iniziativa ASIS e presente sul sito di Dialect Syntax, che accludiamo nel CD. Si tratta di frasi grammaticalmente ricche che servono a mettere in luce le caratteristiche peculiari del dialetto veneto.
Contemporaneamente abbiamo creato il primo lessico di base completamente specificato del veneto. Anche questo verrà incluso nel CD. Questo lessico di base ha un migliaio di forme di parola a cui sono associate le caratteristiche, morfologiche, sintattiche e semantiche e il lemma di riferimento. Queste parole sono quelle più frequenti del veneto sulla base appunto dei lessici di frequenza prodotti.
Il traduttore per regole utilizzerà il lessico di base e le regole che verranno messe a punto sul sistema XLE.

SEGRETERIA DIDATTICA
tel. 041 234 5704-5705
fax 041 234 5706
filnia@unive.it
vanessa@unive.it

SEGRETERIA AMMINISTRATIVA
tel. 041 234 5780
fax 041 234 5703
mcduse@unive.it

CA' BEMBO
Dorsoduro, 1075
30123 VENEZIA

COD. FISC. 80007720271