



UNIVERSITÀ
CA' FOSCARI
VENEZIA

DIPARTIMENTO
DI SCIENZE
DEL LINGUAGGIO

SCHEDA TECNICA PROTOTIPO “STILVEN”

1. Il sistema **STILVEN – Sistema di Traduzione dall’Inglese alla Lingua Veneta** verrà realizzato in forma di prototipo con l’aiuto delle risorse sulla lingua veneta messe a disposizione da collaboratori della Regione Veneto – nella persona del sig. Raffaele Serafini e del dott. Lodovico Pizzati e altri. L’attività implementativa verrà svolta sotto la direzione e la coordinazione del Prof. Rodolfo Delmonte.
 - 1.1. Il progetto di costruzione di un traduttore per la lingua veneta ha come scopo primario quello di creare tutte le infrastrutture linguistiche di base e le risorse perché il traduttore stesso possa funzionare e nel tempo migliorare. Un periodo ragionevole per poter avere uno strumento automatico che traduca dalla lingua inglese senza vincoli di sorta – a prescindere dal genere, dal dominio e dalle caratteristiche linguistiche del testo - si può stimare in una decina di anni di attività di ricerca e di raccolta dati. Per questo motivo, il lavoro si svilupperà in maniera incrementale in modo da ottenere ad ogni scadenza - ogni sei mesi - un prototipo da dimostrare. Il primo prototipo sarà pronto dopo il primo anno di attività.
 - 1.2. Il traduttore si svilupperà in maniera parallela considerando le due principali possibili implementazioni, sulla base della tecnologia disponibile attualmente. Da un lato creerà un prototipo che funziona solamente su basi statistiche e matematiche; dall’altro creerà un prototipo che invece utilizza un approccio linguistico di tipo simbolico. Nel secondo anno di lavoro si tenterà di unire i due approcci in un sistema di traduzione ibrido.
 - 1.3 Il primo prototipo del traduttore avrà un dominio limitato - di tipo descrittivo (storia) o prescrittivo (norme) - e una capacità di traduzione limitata a frasi brevi di lunghezza non superiore a dieci parole. Inoltre sarà soltanto funzionante nella direzione Inglese-Veneto. Possibili testi di riferimento potranno essere quelli ricavati dalla Wikipedia, o da libri per bambini disponibili in formato digitale.

2. Architettura del prototipo STILVEN

- 2.1 Il sistema di traduzione su basi statistiche utilizzerà il software liberamente scaricabile chiamato MOSES all’indirizzo,
<http://www.statmt.org/moses/>
sviluppato con fondi della Commissione Europea che utilizza sequenze di parole classificate come “phrases” per tradurre scelte come migliori candidati. Perché il sistema possa funzionare si dovrà fornire una quantità di testi paralleli – traduzioni inglese/veneto - in misura sufficiente dal punto di vista statistico. Per cui una prima attività da prendere in considerazione e da

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

affiancare a quella implementativa è la traduzione di quantità sempre maggiori di testi inglesi in veneto da parte di traduttori qualificati.

Il prototipo che verrà creato in una prima implementazione sarà molto limitato nel suo funzionamento dalla quantità di dati utilizzati per il suo addestramento (training). I successivi prototipi potranno usufruire dei materiali tradotti nel frattempo e così ampliare le proprie capacità.

Verso la fine del secondo anno si cercherà di implementare una versione ibrida nella quale si prenderanno in considerazione le versioni taggate e annotate sintatticamente dei corpora paralleli necessari per l'apprendimento automatico (Machine Learning).

2.2 Il sistema di traduzione per regole simboliche verrà realizzato sullo schema dei traduttori più utilizzati attualmente, cioè a modalità mista, Interlingua e Transfer. La parte di transfer del sistema riguarderà tutte le forme polirematiche e idiomatiche che nella lingua inglese hanno significato fisso e indivisibile e non sottostanno quindi a processi di decomposizione semantica e lessicale. La componente a Interlingua utilizzerà invece una rappresentazione semantica intermedia prodotta dagli analizzatori in lingua inglese. Questa rappresentazione intermedia o INTERLINGUA servirà poi alla sintesi o generazione della frase corrispondente in lingua veneta.

Dal momento che dai dati a nostra conoscenza, la struttura sintattica e la sua corrispondente conversione a dipendenze e semantica fornisce la base più solida per una traduzione sensata, i database su cui si farà attività di apprendimento automatico saranno in ultimo livello di quel tipo. Perché si possa procedere ad una traduzione dall'interlingua si dovrà quindi disporre di informazioni quantitative e statistiche su tutte le possibili traduzioni di una particolare struttura sintattico-semantica. In mancanza di questa informazione si procederà alla traduzione parola per parola sempre sulla base delle informazioni sulle relazioni e dipendenze create dall'Interlingua.

3. COMPONENTI DEL PROGETTO

3.1 Il progetto si compone delle seguenti parti:

- Risorse
- Algoritmi di Analisi e di Traduzione
- Interfaccia Utente su Web

3.2 Le Risorse comprendono i seguenti items:-

- materiali linguistici digitalizzati in lingua veneta raccolti a cura dei collaboratori della Regione Veneto e manipolati presso il LLC per ricavare bigrammi e trigrammi frequenti da utilizzare nel processo di traduzione;
- un lessico di frequenza della lingua veneta derivato dai testi digitalizzati che verrà utilizzato per produrre il lessico di base completamente specificato con features morfologiche e categorie sintattiche e semantiche;
- tabelle di flessione grammaticale per le categorie lessicali che lo richiedono – verbi, aggettivi, nomi da utilizzare nella fase di realizzazione della frase veneta corrispondente a quella 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

- consulenze linguistiche di parlanti veneti nativi per le problematiche relative alle varianti regionali e alla qualità grammaticale dell'output del traduttore;
- le "translation-pairs" o coppie di traduzione, parola inglese-parola veneta: queste in parte coincidono con il dizionario messo a disposizione dei collaboratori della Regione Veneto e in parte dovranno essere prodotti ex novo;
- le "translation-pairs" o coppie di traduzione, per multiwords inglese-multiwords veneta: queste sono inesistenti e devono essere prodotte ex-novo.

Gli Algoritmi di Analisi e di Traduzione verranno forniti dall'LLC. Comprendono i seguenti elementi:

- un sistema di tagging (etichettare con categorie grammaticali) della lingua veneta
- un sistema di disambiguazione dei tag ambigui prodotti dal tagger
- una grammatica context-free della lingua veneta
- il programma di analisi e sintesi della lingua inglese che è stato elaborato dal LCL ed è proprietà intellettuale del suo direttore prof. Rodolfo Delmonte
- una versione del traduttore statistico MOSES che contenga il modello da utilizzare per la creazione del prototipo
- una versione in veneto della risorsa semantica più utilizzata in campo computazionale, cioè WordNet, adattandola sulla base di quella esistente per l'italiano;
- lo sviluppo della grammatica LFG sul sistema XLE

L'interfaccia Utente su Web scritta in linguaggio JAVA verrà prodotta da un programmatore di fiducia della Regione Veneto e non viene inclusa in questo documento.

4. MILESTONES

Le fasi principali dell'attività o milestones di implementazione di STILVEN sono le seguenti:

4.1 Primo Anno - 2008

- nei primi quattro mesi si raccoglieranno i materiali linguistici e le tabelle di flessione grammaticale. Inoltre si inizierà il lavoro di messa a punto dell'algoritmo di analisi dell'inglese;
- nei primi quattro mesi si realizzerà il primo prototipo di traduttore automatico sulla base di MOSES e dei testi paralleli inglese-veneto che verrà dimostrato a luglio 2008;
- nel secondo semestre si realizzerà un treebank del veneto che prevederà il tagging, la disambiguazione, l'analisi sintattica per chunks e infine la strutturazione dei chunks per pseudo-frasi sulla base della grammatica context-free messa a punto fino ad allora. Il treebank, o banca di frasi strutturate sintatticamente, verrà corretto e annotato manualmente se necessario;
- nel primo anno si implementerà il primo prototipo di analizzatore simbolico utilizzando le informazioni ricavate dallo studio dei corpora di lingua veneta, la realizzazione della grammatica del veneto, la costruzione del lessico di base del veneto;

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

Alla fine del primo anno di attività si farà una demo del prototipo.

4.2 Secondo Anno - 2009

- il primo semestre del secondo anno di attività servirà ad ampliare le risorse esistenti e a perfezionare gli algoritmi di analisi. In particolare si amplierà il lessico a features, la grammatica e le regole sintattiche di strutture non canoniche. Inoltre si correggerà il treebank del veneto e lo si porterà a una forma standard;
- si inizierà l'attività di codifica della risorsa semantica chiamata WordNet;
- il secondo semestre sarà dedicato a implementare l'algoritmo di Realizzazione superficiale delle strutture della lingua veneta, cioè alla generazione della traduzione di una qualsiasi struttura semantica inglese derivante dall'Interlingua prodotta dall'analisi;
- l'ultimo trimestre verrà anche utilizzato per sperimentare una versione ibrida del traduttore MOSES in cui oltre alle informazioni sulle "phrases" o sequenze di parole, si aggiungeranno le informazioni sulle strutture sintattiche a chunks prodotte automaticamente dall'analisi e depositate nel treebank del veneto.

5. Costi

Il costo complessivo del progetto è di € 115.000. La Regione verserà al Dipartimento di Scienze del Linguaggio la somma in due tranches separate: € 45.000 nel 2008, e € 70.000 nel 2009.

I pagamenti verranno disposti con accreditamento alla Banca Popolare di Verona e Novara SCRL, filiale Fondamenta S.Chiera, via Fondamenta S.Chiera, conto tesoreria n. 14015,

SWIFT: VRBPIT2V,

IBAN: IT 08 I 05188 02002 000000014015

intestato al Dipartimento di Scienze del Linguaggio.

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