NETS, quando la traduzione assistita dal calcolatore incontra la linguistica

Adriano Allora
Università degli Studi di Torino

Abstract
The word *serendipity* means "to make discoveries, by accident and sagacity, of things not in quest" (wikipedia, 2006). There's some serendipity in the NETS history.
NETS (Network Enhanced Translators' Software) was implemented as a tool to aid and speed up the translators' job: it can split each text in segments and record every segment and its translation.
The very first time I saw the 6406 segments (dirty result) in Source Language – Target Language pairs I thought of parallel corpora.
In this presentation I'll describe NETS (in user's and programmer's perspectives) and its potentiality in linguistic (and lexicographic) perspective.

1 Introduzione

Il presente testo è articolato in due parti; nella prima parte verrà descritto il programma NETS (Network Enhanced Translation Software) sia nella prospettiva del funzionamento da parte dell'utente che da parte del programmatore. Una – pur breve – discesa negli inferi del dettaglio informatico è, come si vedrà, resa necessaria dalla natura open source e di prodotto *in fieri* del programma.

Nella seconda parte verrà invece descritto il corpus di segmenti tradotti e, con un riepilogo dei problemi posti dal trattamento di corpora paralleli, verranno prospettate alcune eventualità di sviluppo di NETS in chiave linguistica.

2 Un nùovo programma di assistenza alla traduzione

La scarsa bibliografia sull'argomento[1] distingue innanzitutto due differenti tipi di procedure di traduzione assistita dal calcolatore: la Human Aided Machine Translation (HAMT),[2]

[1] In effetti la bibliografia sulla connessione tra calcolatori e traduzione si concentra sul ben più stimolante nesso della traduzione automatizzata, o Machine Translation, che vanta una bibliografia assai ricca (valga per tutti il fondamentale Hutchins / Somers 1992; per riferimenti espliciti alla traduzione assistita si rimanda a Macklovitch 2001). I testi disponibili sono soprattutto manuali d'uso (che però raramente spiegano almeno a grandi linee la logica della traduzione assistita, anche se Prior 2004 lo fa), tesi di laurea o documenti descrittivo-comparativi per traduttori interessati ad avvicinarsi alla traduzione assistita (come dmoz.org e, naturalmente, wikipedia).

[2] Si tratta di traduttori automatici che richiedono l'aiuto di operatori umani nei casi ambigui. Ambigua è anche la presenza di HAMT nell'ambito della CAT (accade sia in bibliografia – wikipedia 2006 – che nella presentazione di software come Masterin Pro della Master's Innovation).

che rientra nell'ambito della traduzione automatica, e Machine Aided Human Translation (MAHT), identificata a ragione con la Traduzione Assistita dal Calcolatore (d'ora in poi anche CAT, dall'inglese: *Computer Aided Translation*). E tuttavia CAT è un'etichetta che riunisce diversi tipi di agenti software:

- supervisori allo spelling o grammaticali (sia come programmi autonomi che integrati in altri programmi);
- gestori di (basi di dati di) terminologia, anche disponibili on-line, e dizionari elettronici;
- motori di indicizzazione e/o ricerca su testi già tradotti ed estrattori di concordanze;
- gestori di memorie di traduzione.

E anche se la maggior parte dei programmi di CAT raramente assolve a più di una delle funzioni menzionate, quella certo più rappresentativa è la gestione di memorie di traduzione.

Una Memoria di Traduzione (d'ora in poi MT) è una coppia di segmenti in due lingue diverse, un segmento da tradurre (che serve al programma per individuare segmenti già tradotti simili al segmento che di volta in volta il traduttore deve affrontare) ed un segmento tradotto (che viene proposto dal programma al traduttore come traduzione o base per una traduzione).

Qualsiasi gestore di MT è tanto più efficace quanto maggiore è il suo bagaglio di segmenti tradotti ai quali attingere, anche se – come un ingiustificato analogo della esperienza personale del traduttore – me MT rimangono generalmente appannaggio esclusivo degli individui o delle aziende che finanziano una traduzione. Per questo motivo – e per colmare una lacuna nell'attuale panorama della traduzione assistita – è stato creato NETS: un gestore di MT in cui archivio è condiviso in rete.

1.1 Usare NETS

Come anticipato, NETS è sostanzialmente un gestore di Memorie di Traduzione accessibile da Internet e condiviso.[3] La stessa povertà di mezzi impiegata nella sua implementazione che l'ha reso essenziale ha anche notevolmente contribuito alla sua semplicità d'uso.

La prima schermata contiene una breve ed informale descrizione del programma e le opzioni essenziali per iniziare a lavorare: la creazione di un nuovo gruppo di lavoro (indispensabile per isolare i tipi di utenti e, quindi, la qualità delle traduzioni) o di una nuova memoria di traduzione (che individua invece il tipo di traduzione su un ipotetico asse diatemico e in base alle lingue dalla quale e nella quale si traduce); la scelta della MT nella quale scrivere e delle MT da consultare in cerca di suggerimenti; i moduli di selezione del testo da tradurre.[4]

Iniziata la traduzione NETS propone segmento per segmento il testo, isolando sulla sinistra il segmento da tradurre e il suo immediato cotesto (il segmento precedente e quello successivo); la prima proposta è quindi quella di una navigazione sequenziale dei segmenti, ma è anche disponibile l'intero testo, segmentato e navigabile.

[3] Per ogni riferimento al sito, si rinvia all'URL: www.corpora.unito.it//cgi-bin/lingue/tac/tac_index.pl?var=NETS

[4] In teoria NETS è in grado di leggere un testo dalla rete, scevrarlo di tutto ciò che non è testo piano e permettere al traduttore di lavorare sulla traduzione. In pratica la soluzione del testo piano letto dal client dell'utente è la soluzione migliore.

Sotto il campo di testo nel quale inserire la traduzione compaiono i bottoni con le proposte di traduzione: cliccandoli il loro testo viene copiato nel campo che deve ospitare la traduzione. Se nessuna traduzione è adeguata al segmento da tradurre, l'utente può inserire una nuova traduzione e memorizzarla, passando al segmento successivo.

Esistono alcuni dettagli forniti dal programma, come la distanza tra il segmento da tradurre e quello la cui traduzione viene proposta o il promemoria delle MT consultate, ma si tratta appunto di dettagli non essenziali in questo contesto.

Più interessante è invece la funzionalità di dizionario, che permette di effettuare ricerche con espressioni regolari POSIX – presto anche Perl –. Il risultato della ricerca è una lista di segmenti che contengono il termine ricercato; l'utente può scegliere: la lingua (sorgente o di destinazione) nella quale fare ricerche, la o le memorie di traduzione da interrogare e se visualizzare, per ogni segmento trovato, anche la sua traduzione.

1.1 Programmare NETS

NETS è stato sviluppato con Perl e il modulo CGI.pm,[5] ma al momento alcune funzionalità poggiano ancora su programmi nativi dei sistemi operativi X-like (come il programma *grep*), compromettendone la compatibilità.

Come nel caso di altri programmi ospitati sul sito www.corpora.unito.it, alla base di NETS esiste la volontà di provare a scrivere un programma accessibile e modificabile anche per quanti non siano avvezzi alla programmazione (seppure una certa dose di coraggio e pazienza rimangano requisiti fondamentali).

La programmazione è partita dall'interfaccia: scritte in HTML la prima pagina e la pagina dei risultati, si è proceduto alla scomposizione dei processi di trasformazione ed elaborazione del testo, dal caricamento alla segmentazione, dalla proposizione dei segmenti in lingua sorgente alla ricerca e proposizione di possibili traduzioni alla rielaborazione dei segmenti tradotti in un unico testo. In ognuno di questi frammenti di processo sono stati cercati sotto-frammenti comuni agli altri frammenti, come la pulizia del testo e la conversione dei caratteri speciali in stringhe di *escape* compatibili con tutti i browser.

I singoli frammenti, o moduli, e sotto-moduli sono poi stati sviluppati individualmente, in un processo di continuo miglioramento: il modulo di ricerca e riconoscimento di segmenti simili, ad esempio, verrà presto integrato con alcune procedure già ampiamente testate con il motore di ricerca EnTeR (Allora 2006), al fine di migliorarne precisione, velocità e mantenibilità.[6]

Entrando un po' più nel dettaglio, il modo in cui NETS gestisce i segmenti è quantomeno originale: i gruppi di lavoro sono cartelle che contengono altre cartelle, le MT. I nomi delle MT sono sottoposti ad una rigida normazione, devono infatti specificare la Lingua Sorgente (LS), la Lingua Traducente (LT) e devono fornire qualche informazione sul tema dei testi; la sintassi di tali nomi è quindi LS+LT_informazioni.

[5] Per una introduzione al linguaggio Perl si rimanda a Allora 2006.

[6] Probabili migliorie su questo terzo fronte saranno poi riversate sugli altri progetti di www.corpora.unito.it.

All'interno delle singole cartelle/MT sono allineati i segmenti, distinti per lingua e numerati (il segmento e la sua traduzione saranno chiamati, ad esempio, IT1.txt e DE1.txt) in maniera da permettere al programma di accedere velocemente, dato una qualsiasi segmento, anche alla sua traduzione.

Questa struttura permette non solo alla macchina di accedere in maniera molto trasparente ai dati,[7] ma anche agli utenti umani che intendano, ad esempio, trattare i dati automaticamente con altri strumenti: estrarre con uno script shell o con un programma in Perl una lista di frequenza per lingua, o passare i dati ad un parser ancora facendo distinzioni in base alla lingua o alla MT.

3 NETS al servizio della linguistica

Perché un programma come NETS può rivestire qualche interesse per la linguistica?

Al momento della stesura del presente contributo, il maggior pregio di NETS consiste nella sua natura di raccolta di corpora paralleli liberamente interrogabili (e incrementabili!) in rete, che, grazie alle distinzioni per gruppi e MT, permette di fare ricerche non solo per argomento, eventualmente valutando differenze e somiglianze d'uso in settori differenti, ma anche per gruppi ed eventualmente tipi di traduttori, in prospettiva di analisi della traduzione e dell'apprendimento.

L'avere come informatori linguistici dei traduttori, di vari livelli, impegnati nel proprio compito comporta anche però un prezzo: i corpora vanno costantemente tenuti sotto osservazione e periodicamente "ripuliti" da errori dovuti ad uno scorretto uso del programma.[8]

La raccolta, ripulita, comprende 15 corpora, 3 dall'italiano al tedesco, 4 dall'italiano all'inglese, 3 dall'italiano allo spagnolo e 4 dall'italiano al francese. I tipi di testi raccolti in questi corpora sono segmenti di tratti da pagine web di: *curricula*, un catalogo di tastiere industriali e commutatori elettrici, un sito linguistica (una lezione on-line sulla deissi) e alcune pagine di www.corpora.unito.it.

I segmenti, in tutto, sono 3636 (il confronto con gli iniziali 6406 segmenti fornisce un'idea abbastanza chiara di quanti segmenti possono essere stati memorizzati per errore o con errori, non di traduzione ma di redazione dei segmenti stessi), sui quali sono spalmate 84950 parole, con una media di 26 parole per segmento.

In una prospettiva futura, possono essere valorizzati alcuni aspetti di NETS che promettono una maggiore utilità per quanti vi si avvicinino come linguisti invece che come traduttori o studenti di traduzione:

- piena accessibilità ai dati: il corpus può essere liberamente fruito ed arricchito, ed anche se la questione del bilanciamento dei corpora non può che rimanere irrisolta, l'etichettatura per gruppo e MT e la prospettiva di crescita potrebbero rappresentare una parziale compensazione;

[7] Un ulteriore vantaggio di tale trasparenza consiste nel fatto che, in fase di riscrittura di singoli moduli o sotto-moduli, i dati rimangono inalterati.

[8] Ad esempio, includenti, oltre alla traduzione, anche il segmento in LS; oppure totalmente privi di traduzione. Tutti i segmenti malformati sono stati spartanamente eliminati (a vantaggio dei traduttori più che dei linguisti).

• buona accessibilità del codice: per ovvie ragioni di sicurezza non è possibile manipolare il codice direttamente sul server che lo ospita, ma come detto il codice è abbastanza leggibile e chiunque può contattare l'autore del programma, farsene inviare una copia, installarlo e modificarlo sulla propria macchina e, successivamente, anche sul server di riferimento. Qualsiasi proposta di modifica è naturalmente ben accetta;

• continuo sviluppo: sono in previsione le seguenti migliorie: implementazione del sistema di indicizzazione già usato per EnTeR; implementazione di un dispositivo per l'aggiornamento dinamico degli indici; progettazione e implementazione di un modulo di *tracking* per memorizzare, osservare e quantificare i comportamenti dei traduttori in termini di riutilizzo e/o modifica delle MT e di progressive trasformazioni di singoli segmenti; sviluppo di uno strumento del modulo dizionario che parallellizza più finemente i segmenti,[9] individuando su base statistica anche i probabili traducenti della parola cercata nei testi in LT allo scopo di fornire un vero dizionario multilingue che solo in seconda istanza, e su richiesta dell'utente, mostri anche i contesti d'uso.[10]

References

A. Literature

Allora, A. (2006), *PERLinguisti, manuale di programmazione in Perl per umanisti*, Roma, Aracne.

Ahrenberg, L., Andersson, M., Merkel, M. (1998), 'A simple hybrid aligner for generating lexical correspondences in parallel texts' in *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montréal, Canada, 10-14 August 1998, pp. 29-35.

Hutchins, W.J., Somers H.L. (1992), *An Introduction to Machine Translation*, Roma-San Diego, Academic Press.

Pianta E., Bentivogli, L. (2003), 'Translation as Annotation' in *Proceedings of the AI*IA 2003 Workshop "Topics and Perspectives Processing in Italy"*, Pisa, Italy, September 2003, pp. 40-48.

B. Web References

Allora, A. (2005), *EnTeR, Engine for Textual Researchers*. Available at: http://www.corpora.unito.it/cgi-bin/lingue/enter/enter_index.pl?corpus=VALICO

Macklovitch, E. (2002), *The New Paradigm in NLP and its impact on Translation Automation*. Available at: www.onterm.gov.on.ca

Prior, M. (2004), *The ASAD Manual for OmegaT*. Avaliable at: http://www.omegat.org

Boitet, C. (1995), *Machine-aided Human Translation*. Avaliable at: http://cslu.cse.ogi.edu:HLTsurvey:ch8node6.html

[9] A partire da Ahrenberg/Andersson/Merkel 1998.

[10] La speranza è quella di realizzare la prospettiva di traduzione come etichettatura, di ricorso all'elaborazione delle differenze tra segmenti paralleli per l'analisi del linguaggio (Pianta, Bentivogli 2003).