Seminario di “Intelligenza Artificiale: Trattamento Automatico del Linguaggio Naturale”

Titolo : Machine Translation

Studente: Bertocchi Ulisse

Corso di Laurea in Informatica

1

CAPITOLO 1: Introduzione

“Machine Translation”: Il settore che tenta di automatizzare nel suo complesso, o in parte, il processo di traduzione da una lingua umana ad un altra.

Notazione: Nel seguito il termine “Machine Translation” sarà più volte abbreviato con la sigla MT.

2

1.1 - Perché la “Machine Translation” è importante

Suddividiamo le motivazioni per campi di appartenenza;

Socio Politica: L’importanza socio-politica si evidenzia soprattutto in quelle comunità dove si parla più di una lingua. In questo caso l’unica alternativa ad un uso molto ampio della traduzione è l’adozione di una singola lingua ufficiale. Essa, però, non è una soluzione molto attraente in quanto comporta la scomparsa graduale delle altre lingue e, ancor più grave, la perdita di culture distintive e modi di pensare. E’ chiaro che in un contesto simile la mole di testi da tradurre è talmente alta che non sarebbe possibile affidarne il compito a traduttori umani e l’unica soluzione è l’uso dei traduttori automatici.
Esempi di realtà politiche all’interno delle quali convivono civiltà che utilizzano diverse lingue sono il Canada, la Svizzera, la Comunità Europea.

3

Commerciale: 1) Al fine di allargare i confini delle esportazioni di aziende commerciali è necessario fornire informazioni sui prodotti in svariate lingue diverse e molto difficilmente un traduttore umano riesce ad avere una conoscenza linguistica appropriata.
2) La traduzione è costosa. I traduttori umani devono essere molto esperti e i loro salari sono molto alti.

OSS: E’ stato stimato che circa il 40-45% dei costi di funzionamento delle istituzioni della Comunità Europea sono costi legati al linguaggio, dei quali traduzione ed interpretazione sono i principali elementi.

4

Scientifico: MT è una ovvia applicazione ed un terreno di test per molte idee in informatica, intelligenza artificiale e linguistica, e diversi dei più importanti sviluppi in questi campi sono cominciati nel campo della MT.

Filosofico: In quanto rappresenta un tentativo di automatizzare una attività che può richiedere l'utilizzo dell'intero campo della conoscenza umana, cioè, per qualsiasi porzione di conoscenza umana è possibile pensare ad una frase o testo per la cui traduzione tale conoscenza è richiesta. In questo senso, l'efficienza con la quale si può automatizzare la traduzione è una indicazione dell'efficienza con la quale si può automatizzare il pensiero.

5

1.2 – Un po' di storia

E' possibile rintracciare idee riguardanti l'automatizzazione dei processi di traduzione già nel diciassettesimo secolo, ma possibilità realistiche si presentarono solo nel ventesimo secolo.

-A metà degli anni trenta, un franco-armeno Georges Artsrouni e un russo Petr Troyanskii, si applicarono per brevettare macchine traduttrici. Dei due, il lavoro di Troyanskii fu il più significativo, proponendo non soltanto un metodo per un dizionario bilingue automatico, ma anche uno schema per codificare regole grammaticali interlingue (basate sull'Esperanto) e una rappresentazione di come le fasi di analisi e di sintesi avrebbero dovuto funzionare.

6

-***I pionieri (1947-1954):*** Poco dopo la comparsa dei primi calcolatori elettronici, la ricerca cominciò ad utilizzare i computer come supporto per la traduzione di linguaggi naturali. Entro pochi anni la ricerca sulla MT cominciò in molte università degli US, e nel 1954 fu data la prima dimostrazione della fattibilità della traduzione automatica. Sebbene si basasse su di un vocabolario ed una grammatica molto ristretti, essa fu sufficientemente impressionante da stimolare massicci contributi economici alla MT negli Stati Uniti e da provocare la nascita di progetti in tutto il mondo.

-***La decade dell'ottimismo(1954-1966):*** I primi sistemi consistevano primariamente di grossi dizionari bilingue i quali, date parole espresse nel 'source language', restituivano parole equivalenti nel 'target language' e delle regole per produrre l'ordinamento corretto dell'uscita. Ci si rese presto conto che regole specifiche guidate dal dizionario per l'ordinamento sintattico erano troppo complesse e si fece evidente la

7

necessità di utilizzare metodi più sistematici di analisi sintattica. Diversi progetti furono ispirati dagli sviluppi contemporanei della 'linguistica' ed in particolare dai modelli della grammatica formale, ed essi sembrarono offrire la prospettiva di capacità di traduzione molto più forti.

-***La disillusione(1966):***L'ottimismo rimase ad alti livelli per la prima decade di ricerca, con molte predizioni di una possibile imminente soluzione. La disillusione crebbe nel momento in cui i ricercatori incontrarono barriere semantiche per le quali essi non riuscivano a vedere soluzioni immediate. Il supporto del governo degli Stati Uniti cominciò a venire meno nel momento in cui ci si accorse della mancanza di progressi. Venne istituito il 'Automatic Language Processing Advisory Committee' (ALPAC), il quale in un famoso documento del 1966 concluse che la MT era più lenta, meno accurata e due volte più costosa della traduzione umana, escludendo anche la possibilità di sviluppi importanti immediati.

8

- ***Le conseguenze del documento ALPAC (1966-1980s):*** Sebbene il documento ALPAC fu da molti considerato come parziale e poco lungimirante, esso provocò una fine virtuale alla ricerca nel campo della MT negli Stati Uniti per più di una decade ed ebbe una grossa influenza anche sulle ricerche nell'Unione Sovietica ed in Europa. Ad ogni modo le ricerche continuarono in Canada, in Francia ed in Germania. Negli anni successivi il sistema 'Systran' fu installato dalla USAF (1970), e poco più tardi dalla Commissione delle Comunità Europee (1976) per tradurre i suoi volumi di documentazione che stavano rapidamente crescendo in quantità. Negli stessi anni, apparve in Canada un altro sistema operazionale di successo, il sistema Meteo per tradurre previsioni meteorologiche sviluppato all'università di Montreal.

9

-***1980s:*** Attraverso gli anni ottanta continuarono le ricerche su metodi e tecniche più avanzati. Per la stragrande maggioranza della decade, la strategia dominante fu quella della traduzione 'indiretta' utilizzando rappresentazioni intermedie frutto di analisi sintattiche, semantiche e morfologiche, con l'utilizzo qualche volta di basi di conoscenza non strettamente linguistiche. Sempre in questi anni ci fu un forte incremento della richiesta di traduttori automatici.

-***1990s:*** Questi anni segnarono un punto di svolta nell'approccio alla ricerca nel campo della MT. Un gruppo dell'IBM pubblicò i risultati di esperimenti su di un sistema basato puramente su metodi statistici. Inoltre, diversi gruppi giapponesi iniziarono ad utilizzare metodi basati sull'utilizzo di esempi di traduzione, utilizzando un approccio che viene oggi indicato come traduzione 'example based'. In entrambi gli approcci la caratteristica distintiva fu quella che non venivano utilizzate regole sintattiche o semantiche nell'analisi del testo o nella selezione di termini equivalenti.

10

Una terza innovazione che comparve negli anni novanta fu l'inizio della ricerca sulla traduzione del parlato, cioè sistemi che integrano moduli di riconoscimento del parlato, di sintesi del parlato e di traduzione.
Questo periodo segnò anche la nascita di altri obiettivi della MT che portò alla costruzione di sistemi basati su 'linguaggi controllati' e su domini ristretti.
Crebbe notevolmente la vendita di software per MT per personal computer, ed ancor più evidente fu la crescita della disponibilità di traduttori automatici forniti on-line.

11

CAPITOLO 2: La 'Machine Translation' in pratica

I vari passi che vengono eseguiti durante l'utilizzo di un traduttore automatico sono:

- Preparazione del documento

- Il processo di traduzione

- Revisione del documento

12

2.1 – Preparazione del documento

Questa fase ha lo scopo di organizzare il testo da sottoporre al sistema MT nella sua struttura e nella scelta del lessico al fine di facilitare il compito del sistema nel tentativo di restituire la migliore risposta possibile.

OSS: Un traduttore umano è spesso capace di rielaborare un testo scritto in una maniera confusa in una sua traduzione chiara e lineare; sicuramente noi non ci possiamo aspettare che ciò avvenga nel caso dei traduttori automatici. Nel momento in cui forniamo ad un sistema MT un testo scritto male noi sappiamo già a priori che la qualità della risposta sarà scadente.

13

La definizione di 'buon' input non è chiara e cambia da sistema a sistema. Ad ogni modo è facile identificare qualche semplice regola di scrittura e strategia che possono incrementare la performance della maggior parte dei sistemi MT.

Regole di scrittura di base

- Costruire frasi corte (i sistemi sono sempre incerti nella scelta del giusto modo di analizzare una frase; per frasi lunghe il grado di incertezza aumenta drammaticamente)
- Assicurarsi della correttezza grammaticale delle frasi
- Evitare strutture grammaticali particolarmente complesse
- Evitare (per quanto possibile) l'uso di parole che hanno molti significati
- In documenti tecnici, utilizzare soltanto parole tecniche e termini che sono ben stabiliti, ben definiti e conosciuti dal sistema

14

OSS: Realizzare una restrizione sull'insieme dei possibili input al sistema in accordo a semplici regole come quelle appena viste può innalzare fortemente la performance di un sistema MT. Ma questo non è l'unico vantaggio: ciò può anche incrementare la comprensibilità del testo da parte di un lettore umano.

Come conseguenza di tali considerazioni, diverse grosse compagnie hanno sviluppato ed esteso l'idea delle regole di scrittura, includendo vocabolari limitati, al fine di produrre forme ristrette di linguaggio usufruibili per testi tecnici. Queste forme ristrette sono conosciute come 'Controlled Languages'.

15

2.2 – Il processo di traduzione

Il passo di traduzione può consistere di funzionalità più o meno evolute. Un punto da tenere ben chiaro in mente è che un supporto alla traduzione può essere fornito anche senza realizzare una traduzione automatica completa. Di seguito riportiamo due possibili situazioni:

Strumenti di supporto alla traduzione 'dictionary based':
Tali dizionari elettronici possono essere di immenso aiuto anche nel caso in cui questi vengano utilizzati senza la traduzione automatica del testo. Un possibile scenario è il seguente: tu stai traducendo un testo a mano. Utilizzando un mouse o una tastiera, tu clicchi su una parola nel testo sorgente e una lista delle sue possibili traduzioni viene mostrata sullo schermo. Tu clicchi sulla traduzione possibile che ti sembra essere più appropriata nel contesto considerato ed essa viene inserita direttamente nel testo del linguaggio target.

16

Interazione nella traduzione: I sistemi MT analizzano il testo e devono decidere qual' è la sua struttura. Nel caso in cui ci sono dubbi o incertezze riguardo la struttura, o riguardo la scelta corretta di una parola per la traduzione, essi possono interagire in maniera utile con il traduttore umano per porre semplici domande riguardo i problemi della traduzione.

2.3 – Revisione del documento

Il principale fattore che decide la quantità di 'post-editing' che è necessario venga fatta su una traduzione prodotta automaticamente è sicuramente la qualità richiesta dell'output. Ciò a sua volta dipende dallo scopo della traduzione e dal tempo disponibile.

Ovviamente la difficoltà del 'post-editing' e il tempo da esso richiesto sono strettamente legati alla qualità della risposta del sistema MT: tanto peggiore è l'output, tanto più grande è lo sforzo da compiere per il 'post-editing'.

17

Esistono vari casi: uno nel quale è necessario fare un completo 'post-editing' e uno nel quale nessun tipo di 'post-editing' è richiesto. Un'altra opzione potrebbe essere realizzare il 'post-editing' su una traduzione al fine di rendere più facile la lettura e la comprensione del testo senza mirare alla perfezione tipica di un testo scritto pubblicato.

OSS: I sistemi MT fanno i soliti tipi di errori di traduzione ripetuti nel tempo. Qualche volta tali errori possono essere eliminati modificando le informazioni nel vocabolario.

18

CAPITOLO 3: Rappresentazione della conoscenza sintattica

In questo capitolo si introdurranno alcune delle tecniche che possono essere utilizzate per rappresentare la conoscenza sintattica necessaria per la traduzione, in modo tale che essa possa essere elaborata automaticamente

19

In generale, lo studio sintattico riguarda due tipi leggermente diversi di analisi. La prima è l'analisi della "struttura costituente", cioè la divisione delle frasi in sintagmi e la categorizzazione di questi come parte nominale, verbale etc. La seconda riguarda le "relazioni grammaticali" e quindi il riconoscimento all'interno delle frasi di soggetto, oggetto e altre relazioni.

3.1 – Grammatiche e struttura costituente

Le frasi sono formate da parole, tradizionalmente appartenenti a categorie tra le quali nomi(N), verbi(V), aggettivi(A), avverbi(ADV) e preposizioni(P).
Una ***grammatica*** di un linguaggio è un insieme di regole le quali dicono come queste categorie possono essere combinate per creare frasi corrette o 'well-formed'.

20

Per la lingua inglese tali regole possono indicare che la frase (1a) è corretta grammaticalmente, mentre la frase (1b) non lo è.

(1) a. Put some paper in the printer.
b. Printer some put the in paper.

Una semplice regola per la lingua inglese potrebbe essere: una frase consiste di un sintagma nominale (es. the user), seguito da un verbo modale o ausiliario (es. should), seguito da un sintagma verbale (es. clean the printer).

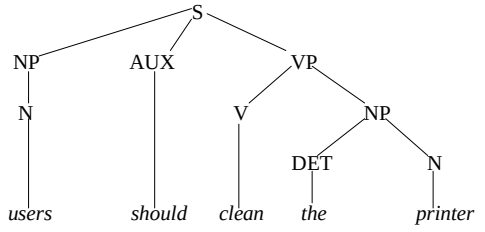
(2) The user should clean the printer

A sua volta un sintagma nominale può consistere di un articolo o determinante come *the* o *a*, ed un nome come *printer*. In alcune circostanze l'articolo può essere omesso.

21

NOTAZIONE: Le frasi sono spesso abbreviate con S, i sintagmi nominali con NP, i sintagmi verbali con VP, gli ausiliari con AUX ed i determinanti con DET.

Tali informazioni possono essere facilmente visualizzate utilizzando un albero.



22

Per convenienza i linguisti spesso utilizzano una notazione speciale per esprimere le regole grammaticali. Un esempio di grammatica che riesce a generare ed a riconoscere la frase appena utilizzata come esempio è la seguente:

S -> NP (AUX) VP
VP -> V (NP) PP*
NP -> (DET) (ADJ) N PP*
PP -> P NP
N -> user
N -> users
N -> printer
N -> printers
V -> clean
V -> cleans
AUX -> should

DET -> the
DET -> a
P -> with

Notazione: P rappresenta una preposizione e PP un sintagma preposizionale.

23

La prima regola della grammatica precedente dice che una frase (S) può essere riscritta come un sintagma nominale (NP) seguito da un ausiliario (AUX) opzionale (l'opzionalità si indica con le parentesi tonde), seguito da un sintagma verbale.
Gli argomenti marcati con il simbolo '*' possono apparire un qualsiasi numero di volte (perfino zero volte).
Le regole con parole reali come *users* nella loro parte destra realizzano una sorta di dizionario primitivo.
Ritornando alla rappresentazione ad albero precedente, ogni nodo nell'albero corrisponde alla parte sinistra di una particolare regola, mentre i figli di ogni nodo corrispondono alla parte destra della stessa regola.
OSS: La piccola grammatica che abbiamo utilizzato non è l'unica grammatica possibile per il trattamento del frammento di inglese da noi considerato. Non ci sono criteri particolari per capire quale sia la migliore. Per la valutazione della loro qualità potremmo comunque domandarci se riescono a generare tutte le frasi possibili del linguaggio e se generano solo frasi grammaticalmente corrette.

24

3.1.1 – Parsing

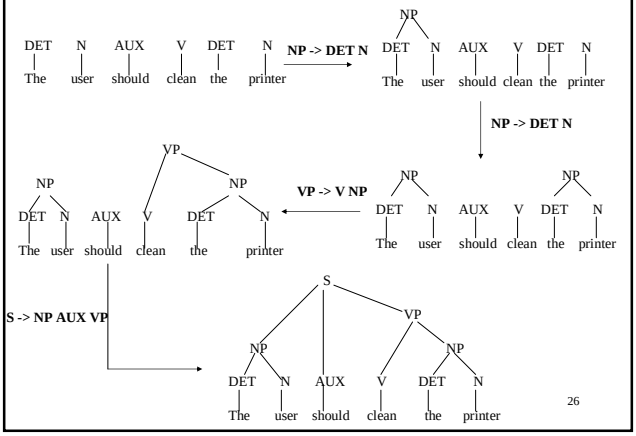
Il compito di un parser automatico è quello di prendere una grammatica formale e una frase ed applicare le regole della grammatica alla frase al fine di (a) controllare che essa sia effettivamente grammaticalmente corretta e (b) nel caso essa sia grammaticale, mostrare come le parole sono combinate all'interno dei sintagmi e come i sintagmi sono uniti per formare sintagmi più grandi (incluso le frasi).

In effetti, ciò restituisce le solite informazioni della struttura ad albero introdotta precedentemente. Così si può pensare che un parser prenda una frase e produca tale albero come rappresentazione.

Ci sono vari modi per applicare le regole all'input e produrre un albero in uscita. Nel seguito proponiamo un esempio di applicazione dell'algoritmo 'bottom-up' per la realizzazione del parsing.

25

Esempio di esecuzione dell'algoritmo 'bottom-up';



26

3.2 – Analisi delle relazioni grammaticali

Oltre alla conoscenza grammaticale espressa in termini di albero della struttura costituente, ci sono altri tipi di informazione che è utile rendere esplicito. In particolare è utile sapere quale funzione grammaticale è rappresentata da un dato elemento della frase, dove tra le varie funzioni ci sono 'SUBJECT', 'OBJECT', 'SENTENTIAL COMPLEMENT' e altre ancora.

OSS: Per capire quanto ampio e complesso è il lavoro che sta dietro la MT osserviamo che: in inglese i soggetti sono normalmente sintagmi nominali che stanno prima del verbo, e gli oggetti (o complementi oggetto) normalmente stanno immediatamente dopo il verbo. In giapponese l'ordinamento normale delle parole è 'soggetto oggetto verbo', in irlandese è 'verbo soggetto oggetto'. In molti linguaggi, come il russo, il verbo, il soggetto e l'oggetto possono apparire essenzialmente in qualsiasi ordine.

27

Sintagmi che svolgono il ruolo di SUBJECT, OBJECT, etc. dovrebbero anche essere distinti da quelli che svolgono ruoli di MODIFIERs, o ADJUNCTs, di vario tipo. Per esempio nella frase sotto *You* è il SUBJECT del verbo *clean*, *the printer casing* è l'OBJECT e i sintagmi preposizionali *with a non-abrasive compound* ed *at any time* sono ADJUNCTs.

You can clean the printer casing with a non-abrasive compound at any time.

Diversamente dai SUBJECTs le ADJUNCTs sono opzionali. Per esempio una frase che omette le ADJUNCTs è ancora perfettamente 'well-formed';

You can clean the printer casing.

Omettere il SUBJECT produce invece un risultato sgrammaticato come nel seguente caso;

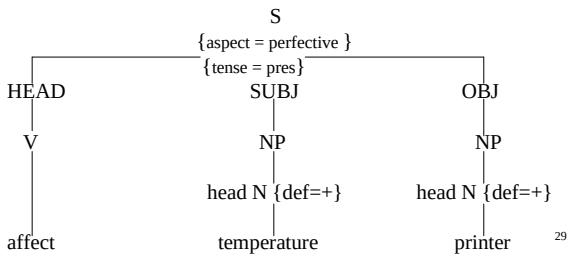
Can clean the printer casing.

28

Ci sono vari modi di rappresentare le frasi in termini delle relazioni grammaticali, ma ciò è essenzialmente poco diverso dalla rappresentazione ad albero della struttura costituente che abbiamo già incontrato.
Per esempio alla frase

The temperature has affected the printer

può essere associata la seguente rappresentazione;



29

L'elemento HEAD è, intuitivamente, l'elemento più importante dal punto di vista grammaticale dell'intero sintagma, l'elemento che guida il significato. In un sintagma nominale l'head è dato dal nome, in un sintagma verbale dal verbo e in un sintagma preposizionale dalla preposizione.

OSS: Diversamente dall'albero della struttura costituente, l'ordine dei rami in questo caso non è importante. Ciò perché sono state indicate le relazioni grammaticali e queste individuano già implicitamente un ordinamento delle parole.

Si noti che alcune parole che comparivano nella frase originale non compaiono nella rappresentazione ad albero. Queste sono state rimpiazzate da attributi come 'def', 'tense' e 'aspect'. Le specifiche 'aspect=perfective' e 'tense=pres' indicano che la frase è interamente nel present perfect tense. La specifica 'def=+' sui sintagmi nominali indica che ci si riferisce ad un particolare oggetto e non ad una categoria di oggetti.

30

IMPORTANTE: La rappresentazione delle relazioni grammaticali appena proposta ha lo scopo principale di astrarre dalla maniera particolare in cui la frase è presentata pur mantenendo tutti gli aspetti in essa espressi. Si può notare che le rappresentazioni astratte di frasi in linguaggi diversi sono spesso molto più simili tra di loro che non le frasi stesse. Tutto ciò assume un significato molto importante nella MT in quanto la chiave del successo sta proprio nella ricerca di similitudini tra le rappresentazioni astratte della frase in questione nel linguaggio sorgente e della sua traduzione nel linguaggio obiettivo.

Per descrivere la relazione tra la struttura costituente e le strutture relazionali, ci sono sostanzialmente due approcci;

I Approccio: semplicemente si aggiungono informazioni riguardanti le relazioni grammaticali direttamente nelle regole della grammatica.

31

Esempio:

S -> NP{SUBJECT} AUX VP{HEAD}

VP -> V{HEAD} NP{OBJECT} PP{ADJUNCT}*

AUX -> has{aspect=perfective, tense=pres}

L'idea è che queste annotazioni possono essere interpretate in una maniera tale che strutture ad albero delle relazioni grammaticali possono essere costruite in parallelo all'albero della struttura costituente.

II Approccio: si prevede l'utilizzo di regole speciali che relazionano la rappresentazione della struttura costituente con la rappresentazione delle relazioni grammaticali.

32

Esempio:

$[_S$ NP:\$1, AUX:\$2, $[_{VP}$ V:\$3, NP:\$4]]

↔

$[_S$ HEAD:\$3, SUBJ:\$1, OBJ:\$4]

Nella regola presentata, \$1, \$2, etc. sono variabili, o nomi temporanei di parti della struttura. La regola è molto semplificata dal momento che non vengono nemmeno menzionate le informazioni riguardo gli attributi 'aspect', 'def' e 'tense', ma ad ogni modo essa dovrebbe essere in grado di dare un'idea del concetto.

OSS: Si noti come la freccia usata nella regola sia bidirezionale. Ciò suggerisce che la regola descrive una corrispondenza tra la rappresentazione della struttura costituente e quella delle relazioni grammaticali, senza dire quale delle due ha priorità sull'altra. In tal modo la regola può essere utilizzata per trasformare una rappresentazione della struttura costituente in una delle relazioni grammaticali e vice versa.

33

3.2.1 – Forme attive e forme passive

Molti verbi hanno una forma passiva ed una forma attiva, come nell'esempio seguente:

(1) a. *Temperature affects printers.* (attiva)
b. *Printers are affected by temperature.* (passiva)

Notiamo che l'oggetto nella frase attiva corrisponde al soggetto in quella passiva. Ciò crea la domanda di cosa significano le relazioni grammaticali SUBJECT e OBJECT. In particolare, *temperature* sarebbe il soggetto di (15a), e *printers* sarebbe il soggetto di (15b). L'alternativa è adottare una notazione che restituisce il solito elemento sia nella forma passiva che in quella attiva. Noi diremo che il D-OBJECT (deep object) corrisponde al sintagma nominale dopo il verbo (in inglese) nelle frasi attive e al sintagma nominale prima del verbo nella corrispondente forma passiva. OSS: Interpretare SUBJECT come 'deep subject' è consistente con l'idea generale di astrarre dalle caratteristiche di superficie della frase, proprietà tipica della rappresentazione delle relazioni grammaticali.

34

CAPITOLO 4: I Motori per la Machine Translation

E' arrivato il momento di guardare dentro il componente non umano più importante nella MT, il componente che attualmente realizza la traduzione automatica.

35

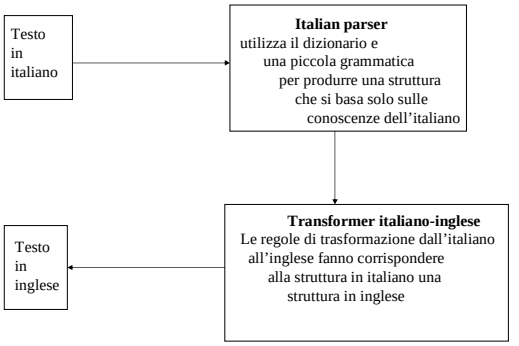
Tradizionalmente, la MT si è basata su motori con architettura '**transformer**', e questa è ancora l'architettura che si trova in molti dei più affermati sistemi commerciali. All'interno del capitolo ci occuperemo anche di una architettura più recente, l'architettura '**linguistic knowledge',** la quale sta cominciando ad essere disponibile in forme commerciali dopo un periodo in cui ha dominato nel campo della ricerca.

4.1 – Architetture transformer

L'idea base di questi motori è che le frasi in ingresso (espresse nel source language) possono essere trasformate in frasi di uscita (espresse nel target language) realizzando il più semplice 'parsing' possibile, rimpiazzando le parole del linguaggio sorgente con il loro equivalente nel linguaggio obiettivo come specificato in un dizionario bilingue, e poi riordinando le parole ottenute al fine di soddisfare le regole grammaticali del linguaggio obiettivo.

36

Di seguito riportiamo uno schema complessivo del funzionamento di un motore ad architettura transformer. Nel caso particolare si tratta di un traduttore dall'italiano all'inglese.



37

Il primo passo di elaborazione include il parser, il quale realizza qualche analisi preliminare della frase nel linguaggio sorgente. Non è necessario che il parser restituisca una rappresentazione completa come quella di cui si è parlato nel capitolo 3, ma può restituire anche una semplice lista di parole. Tutto ciò è passato ad un pacchetto di regole le quali trasformano la frase in ingresso in una frase espressa nel linguaggio target. Le regole di trasformazione includono le regole incluse nel dizionario bilingue e quelle per riordinare le parole. Esse possono anche includere regole per cambiare la forma delle parole target, per esempio, quelle che assicurano la correttezza della persona e del numero del verbo.

Cerchiamo ora di evidenziare in punti le caratteristiche di un generico motore con architettura transformer;

- Alta **robustezza**. Cioè, il motore non si blocca in condizioni di errore quando incontra input che contengono parole o strutture grammaticali sconosciute. Ciò perché raramente il sistema avrà una conoscenza della grammatica del linguaggio sorgente sufficiente a riconoscere frasi sgrammaticate.

38

- Nel caso peggiore può funzionare in maniera insoddisfacente in quanto può produrre uscite del tutto inaccettabili nel linguaggio obiettivo. Ciò è dovuto alla poco dettagliata conoscenza grammaticale da parte del sistema della grammatica del 'target language'.
- Il processo di traduzione include molte regole differenti che interagiscono in molti modi diversi. Ciò rende i sistemi 'transformer' piuttosto difficili da comprendere e ciò a sua volta rende difficile una sua eventuale espansione o modifica.
- L'approccio dei sistemi 'transformer' è quello di essere progettati per la traduzione in un'unica direzione tra una coppia di linguaggi, e ciò li rende poco adatti alla costruzione di sistemi per la traduzione multi-lingua.

39

4.2 – Architetture 'Linguistic Knowledge'

NOTAZIONE: Nel seguito le architetture 'linguistic knowledge' saranno più volte abbreviate con LK.

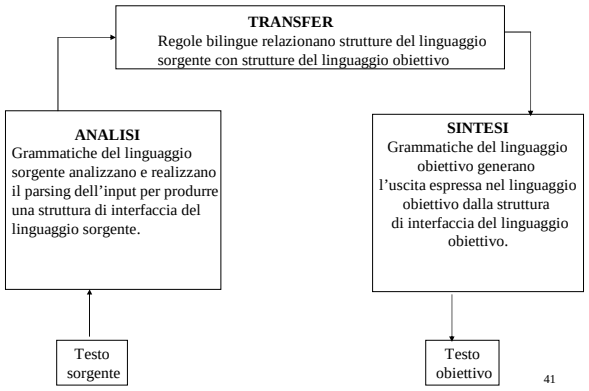
L'idea che sta dietro i motori LK è sostanzialmente la seguente:

Una MT di alta qualità richiede conoscenza linguistica sia del linguaggio sorgente che del linguaggio obiettivo, ma anche conoscenza riguardo le differenze tra i due linguaggi.

OSS: In questo contesto il termine conoscenza linguistica si riferisce alle grammatiche formali che permettono analisi abbastanza profonde e astratte come quelle viste nel capitolo 3.

40

Di seguito proponiamo lo schema generale di un tipico sistema per MT con motore ad architettura LK;



41

Come si può dedurre dallo schema precedente, le architetture LK richiedono due cose:

1- Una dettagliata grammatica sia del linguaggio sorgente che del linguaggio obiettivo. Queste grammatiche sono usate dai parser per analizzare le frasi al fine di produrre rappresentazioni che mostrino la loro struttura sottostante e dai generatori (fase di sintesi) per produrre frasi in uscita che corrispondano ad una particolare rappresentazione.

2- Una grammatica comparativa addizionale la quale è usata per relazionare ogni rappresentazione della frase sorgente a qualche rappresentazione corrispondente nel linguaggio target. Quest'ultima costituisce la base per generare una traduzione nel linguaggio target.

42

I motori LK hanno una grammatica per ogni linguaggio con il quale devono funzionare: in un sistema che traduce dall'italiano all'inglese, ci dovrebbero essere una grammatica per l'italiano ed una per l'inglese. Ognuna di queste grammatiche è una entità indipendente. In effetti la separazione fisica e concettuale tra le due grammatiche è tale che nella fase iniziale di sviluppo del motore LK, un gruppo di specialisti inglesi potrebbe scrivere la grammatica per la lingua inglese interamente indipendentemente da un altro gruppo di specialisti italiani che stanno scrivendo la grammatica per l'italiano del sistema.

OSS: In tal caso, entrambi i gruppi dovrebbero mirare ad una simile profondità di rappresentazione dei loro linguaggi, altrimenti si possono creare discrepanze strutturali che richiederebbero l'uso di regole extra nella fase di transfer per far si che queste diverse strutture tornino ad avere livelli di astrazione simili.

43

IMPORTANTE: Il fatto che venga utilizzata una grammatica propria del linguaggio obiettivo significa che l'uscita del sistema è con molta più probabilità corretta grammaticalmente rispetto a quella di un sistema 'Transformer' come quello del paragrafo 4.1 (ricordiamo che questi ultimi non avevano una grammatica esplicita del linguaggio obiettivo che li guidasse). Infatti, se noi avessimo (per assurdo) un sistema LK con una grammatica 'perfetta' del linguaggio obiettivo, l'unico tipo di errore che esso potrebbe fare sull'uscita sarebbe quello sulla accuratezza della traduzione. Cioè, il sistema produrrebbe sempre frasi perfettamente 'well-formed' anche quando non produce la traduzione corretta.

OSS: In linea teorica il sistema dovrebbe essere reversibile, cioè dovrebbe essere in grado di tradurre tra due lingue diverse a prescindere da quale delle due è la lingua sorgente e quale la lingua obiettivo.

44

Un altro vantaggio dei sistemi LK è che, siccome i linguaggi sono gestiti in moduli separati (una grammatica per ogni linguaggio e una grammatica comparativa per ogni coppia di linguaggi), è relativamente facile in principio aggiungere nuovi linguaggi al sistema.

SVANTAGGIO: Siccome le grammatiche che i linguisti computazionali sono in grado di scrivere sono molto meno complete della grammatica complessiva ideale di ogni linguaggio, ci saranno delle frasi in ingresso grammaticalmente complicate che il sistema fallisce a riconoscere anche se corrette. Sotto questo aspetto i sistemi con architettura 'transformer' hanno il vantaggio di accettare qualsiasi cosa che venga dato loro.

45

4.2.1 – La fase di 'transfer' e le grammatiche comparative

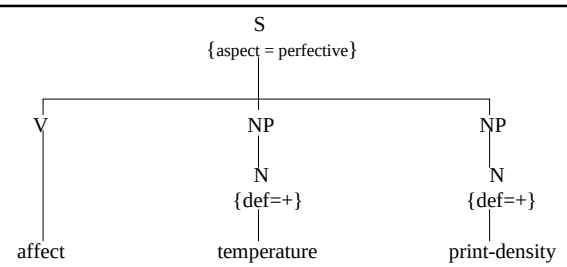
Abbiamo già detto che i parser nei motori LK tipicamente analizzano la frase per generare rappresentazioni astratte. Sicuramente ogni sistema individuale differisce dagli altri per la particolare forma di rappresentazione che utilizza, ma in questo contesto noi supponiamo che il nostro motore produca una rappresentazione sintattica come quella vista nel capitolo 3, anche se questa è ben lontana dall'essere la rappresentazione più astratta possibile.

Ora evidenziamo il significato della fase di transfer attraverso un esempio. Supponiamo di voler tradurre la frase sotto dall'inglese al tedesco;

The temperature has affected the print density.

La fase di analisi potrebbe aver prodotto un risultato simile allo schema seguente, il quale rappresenta così l'ingresso alla fase di transfer.

46



Possiamo vedere ora come la grammatica comparativa relazioni una tale rappresentazione con le corrispondenti rappresentazioni per le frasi nel linguaggio target.

Proprio come ogni grammatica monolingua ha un dizionario di regole (es. N -> temperature), così anche la grammatica comparativa ha regole che realizzano il dizionario bilingue;

47

Nella versione più semplice queste regole possono associare termini lessicali sorgente a termini lessicali obiettivo:

temperature <-> *temperatur*
print-density <-> *druckdichte*
affect <-> *beeinflußen*

OSS: Queste regole del dizionario possono essere viste come relazionanti foglie (i nodi parola) dell'albero del linguaggio sorgente con foglie dell'albero del linguaggio obiettivo.

La grammatica comparativa contiene anche regole strutturali le quali mettono in relazione altre parti dei due alberi. Una tale regola potrebbe essere data da:

[$_S$ HEAD:$HEAD, D-SUBJ:$SUBJECT, D-OBJ:$OBJECT]
$\longleftrightarrow$
[$_S$ HEAD:$H, D-SUBJ:$S, D-OBJ:$O]

48

Nella regola precedente, la parte sinistra descrive una struttura inglese e la parte destra una struttura tedesca. Al suo interno $H, $S e $O sono variabili interpretate come rappresentanti elementi della struttura inglese nella parte sinistra della regola, e come loro traduzione nella parte destra.

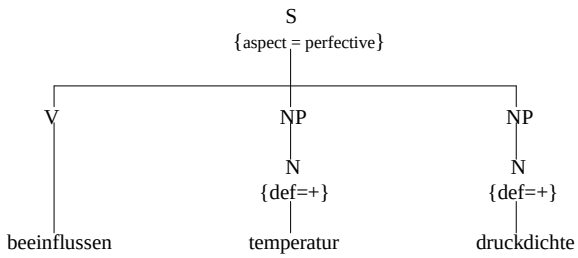
Devono essere tradotte anche le annotazioni sui nodi. Nel nostro caso le regole che realizzano tale traduzione sono immediate e potrebbero essere scritte nel seguente modo:

{def=+} <-> *{def=+}*

{aspect = perfective} <-> *{aspect = perfective}*

Applicando queste regole alla rappresentazione inglese precedente otteniamo la corrispondente rappresentazione tedesca che riportiamo di seguito.

49



La rappresentazione sopra serve come input per il modulo di sintesi del tedesco, il quale applica le regole della grammatica tedesca per produrre una frase in tedesco. In questo caso la frase restituita potrebbe essere:

Die temperatur hat die druckdichte beeinflußt

50

OSS: Sebbene l'esempio qui riportato consiste di regole immediate, ed infatti le strutture che si sono ottenute per i due linguaggi sono molto simili, in genere ciò non è valido. Le regole necessarie sono di solito ben più complesse e le strutture ottenute per i linguaggi in questione sono quindi molto diverse tra di loro.

CONCLUSIONI: Dovrebbe essere chiaro che le architetture LK e quelle Transformer manipolano il problema dell'ordinamento delle parole in maniera diversa. Un motore Transformer generalmente preserva l'ordine del linguaggio sorgente e direttamente lo riusa, con modifiche appropriate per ordinare le parole del linguaggio target. Un motore LK, invece, estrae tutte le informazioni possibili dall'ordinamento delle parole sorgenti e rielabora tali informazioni in una rappresentazione più o meno astratta. Il generatore utilizza le informazioni in tale rappresentazione e nella grammatica del linguaggio target per costruire una frase nel linguaggio target che ha un ordinamento delle parole grammaticalmente appropriato per quel linguaggio.

51

4.2.2 – Interlingua

Da una osservazione generale si è dedotto che le grammatiche comparative della fase di Transfer nelle architetture LK diventano molto più semplici quando l'analisi linguistica riesce ad andare più in profondità e cioè quando la rappresentazione diventa più astratta.

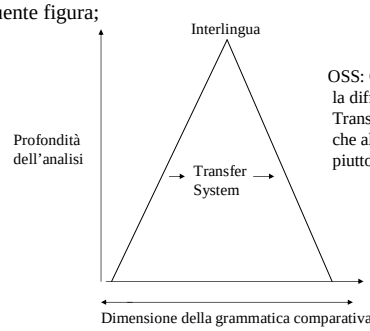
In effetti, uno dei maggiori obiettivi della ricerca nel campo della MT è definire un livello di analisi che sia così profondo ed accurato da far sì che il componente della grammatica comparativa scompare completamente. Dato un tale livello di rappresentazione, l'uscita della fase di analisi potrebbe essere direttamente l'entrata alla fase di sintesi.

OSS: Rappresentazioni di un tale livello dovrebbero catturare qualsiasi cosa in comune tra le frasi e la loro traduzione, cioè, in un certo senso, dovrebbero essere capaci di rappresentare il significato. Esse dovrebbero essere, quindi, anche completamente indipendenti dal linguaggio utilizzato per esprimere la frase.

52

Per tutte le ragioni viste sopra, un tale livello di rappresentazione è normalmente chiamato un '**Interlingua**', e i sistemi che lo utilizzano sono chiamati '**Interlingual**'.

La relazione tra i sistemi Transfer e Interlingual può essere descritta dalla seguente figura;



OSS: Come la figura suggerisce, la differenza tra la rappresentazione Transfer e quella Interlingual è più che altro una distinzione di grado piuttosto che concettuale.

Ci sono vari motivi che rendono molto attraenti i sistemi interlingual.

1 - Da un punto di vista puramente scientifico ed intellettuale, l'idea di tali sistemi è interessante ed eccitante.

2 – Da un punto di vista più pratico, un sistema interlingual promette di essere molto più facile da estendere, aggiungendo nuove coppie di linguaggi, rispetto ad un sistema transfer. Ciò perché dovrebbe essere possibile aggiungere un nuovo linguaggio ad un sistema semplicemente inserendo solo le specifiche componenti di analisi e di sintesi, mentre in un sistema transfer è richiesto l'inserimento anche di tutte le grammatiche comparative tra il linguaggio inserito e tutti i linguaggi già presenti nel sistema. Dal momento che esiste un transfer per ogni coppia di linguaggi, N linguaggi richiedono N x (N-1) componenti transfer (non c'è bisogno di un transfer tra un linguaggio e se stesso). Per esempio, estendere un sistema per 3 linguaggi in uno da 5 significa scrivere 14 nuovi componenti transfer (si passa da 6 a 20 componenti transfer).

CAPITOLO 5: I Dizionari

Questo capitolo tratta il ruolo svolto dai dizionari nella MT. Ad essi viene dedicato un intero capitolo in quanto rappresentano una delle parti più importanti in un sistema per la traduzione automatica.

I motivi per i quali i dizionari rappresentano una parte importantissima all'interno di un sistema per la MT sono i seguenti:

- I dizionari sono le componenti più grandi di un sistema per la MT in termini di quantità di informazione in essi contenuta. Nel caso essi siano qualcosa di più di semplici liste di parole (e lo dovrebbero essere per avere buone prestazioni), allora possono essere anche la componente più costosa da costruire.
- Più di qualsiasi altro componente, la dimensione e la qualità del dizionario limita gli obiettivi del sistema e la qualità della traduzione che ci si può aspettare.
- I dizionari sono la parte dove l'utente finale si aspetta di poter contribuire maggiormente al funzionamento del sistema, in quanto l'utente si aspetta di dover fare delle aggiunte ai dizionari per rendere il sistema realmente utile.

5.1 – Tipi di informazione sulle parole

In questa sezione introdurremo le varie parti di informazione riguardanti le parole che un buon sistema per la MT deve contenere.

E' utile fare una distinzione tra le caratteristiche intrinseche di una parola (le sue proprietà inerenti) e le restrizioni che essa impone sulle altre parole del suo ambiente grammaticale.

L'informazione riguardante l'ambiente grammaticale nel quale una parola può apparire è normalmente divisa in due tipi: l'informazione di '**subcategorization**', che indica gli ambienti sintattici all'interno dei quali una parola può occorrere, e le '**selectional restrictions**' le quali descrivono le proprietà semantiche dell'ambiente.

57

La tipica informazione riguardo la 'subcategorization' è l'indicazione che *button* è un verbo transitivo. Più precisamente, ciò indica che è un verbo che compare come 'HEAD' di frasi con un (sintagma nominale) SUBJECT e un (sintagma nominale) OBJECT.

Di seguito riportiamo alcuni esempi con relative informazioni sulla 'subcategorization' dei verbi che vi compaiono;

a – The president died. [I]
b – The Romans destroyed the city. [Tn]
c – Sam gave roses to Kim. [Dn.pr]
d – Sam gave Kim roses. [Dn.n]
e – Sam persuaded Kim to stay at home. [Cn.t]
f – Kim believed that the library was closed. [Tf]
g – The quality is low. [La]
h – Sam appeared the best man for the job. [Ln]

58

Negli esempi precedenti abbiamo introdotto alcune sigle di cui diamo qui la spiegazione;

[I] - verbo intransitivo
[Tn] - verbo transitivo
[Dn.pr] - verbo ditransitivo il quale prende un soggetto e due oggetti, dove il secondo è introdotto dalla preposizione '*to*'
[Dn.n] - verbo ditransitivo che prende un soggetto e due oggetti sostantivo
[Cn.t] - verbo transitivo complesso che richiedono un soggetto, un oggetto e una clausola infinitivale (non coniugata) introdotta dal '*to*'
[Tf] - verbo transitivo che prende un soggetto, un oggetto e una frase coniugata introdotta da '*that*'
[La] - verbo che collega un sintagma aggettivale (che descrive il soggetto) al soggetto
[Ln] - verbo che collega un sintagma nominale al soggetto

59

I verbi non sono la sola categoria di parole che subcategorizzano per certi elementi nel loro ambiente grammaticale. I sostantivi esibiscono lo stesso fenomeno, come quei sostantivi che sono stati derivati dai verbi.

a – *The death of the president shocked everybody.*

b – *The destruction of the city by the Romans was thorough*

Similmente, ci sono degli aggettivi che subcategorizzano per certi complementi.

60

Analizziamo ora le ‘selectional restrictions’;
Riguardo al verbo *button* noi sappiamo molte più cose rispetto a ciò che abbiamo appena detto, cioè che esso compare con un OBJECT costituito da un sintagma nominale. Sappiamo infatti che l’OBJECT appena menzionato, o in termini di ruoli semantici il PATIENT del verbo, deve essere una cosa abbottonabile, come pezzi di tessuto, e che il SUBJECT (o AGENT in termini semantici) del verbo è normalmente animato.

OSS:Questa informazione è implicita nei dizionari di carta. Al loro interno non troviamo espresso che il soggetto del verbo deve essere una entità animata (probabilmente umana) in quanto è giustamente assunto che il lettore umano può dedurre tutte queste cose da solo. Al contrario, questa informazione deve essere resa esplicita nei dizionari utilizzati per la MT in quanto necessari per una corretta realizzazione delle fasi di analisi, sintesi e trasferimento all’interno dei sistemi per la MT.

61

Le informazioni inerenti e le informazioni riguardo la ‘subcategorization’ e le ‘selectional restrictions’ possono essere rappresentate in una maniera immediata per scopi di MT. Essenzialmente, le entrate in un dizionario per MT sono equivalenti a collezioni di attributi e relativo valore. Per esempio, per il nome *button* potremmo avere una struttura come la seguente la quale, tra le altre cose, indica la forma base del nome stesso, il fatto che si tratta di un nome comune e che è concreto (piuttosto che astratto come ‘felicità’ o ‘sincerità’).

lex = button
cat = n
ntype = common
number =
human = no
concrete = yes

OSS: Il campo ‘number’ è senza valore in quanto un valore per l’attributo è possibile ma non è inerente alla parola stessa la quale può avere diversi valori in situazioni diverse (al contrario *trousers* è solo plurale).

62

E’ chiaro che a parole di diverse categorie grammaticali corrisponde una diversa collezione di attributi. Per esempio, i verbi avranno un attributo *vtype* piuttosto che *ntype*, e mentre i verbi potrebbero avere campi per l’indicazione del numero, della persona e della coniugazione, noi non ci aspettiamo che tali campi siano replicati nel caso di preposizioni.

lex = button
cat = v
vtype =main
finite =
person =
number =
subcat = subj_obj
sem_agent = human
sem_patient = clothing

63

Riguardo le informazioni da inserire nel dizionario concernenti la fase di traduzione, una possibilità è tentare di rappresentare tutte le informazioni rilevanti per mezzo di attributi e valori. Così, come aggiunta alle entrate del dizionario per il termine *button* visto sopra, un sistema ‘transformer’ potrebbe specificare la traduzione aggiungendo l’attributo *trans* al quale si fa corrispondere come valore la traduzione nella lingua target. Se la lingua target è l’italiano ciò significherebbe aggiungere *trans = bottone.*

Osserviamo però che tale soluzione non è particolarmente attraente. Essa è chiaramente orientata in una direzione, e sarà difficile o almeno poco immediato inserire entrate che si riferiscono all’altra direzione di traduzione (cioè dall’italiano all’inglese).

Ciò suggerisce l’utilizzo di regole di traduzione bidirezionali che relazionano ‘head word’ del linguaggio sorgente con quelle del linguaggio obiettivo. Per esempio ciò significherebbe l’introduzione di regole del tipo *temperature* <-> *temperatura.*

64

5.2 – Dizionari e Morfologia

La morfologia riguarda la struttura interna delle parole, e come le parole possono essere formate. Di solito si distinguono tre differenti processi di formazione.

1 – **Inflection**: processo per mezzo del quale una parola è derivata dalla forma di un'altra parola, acquisendo certe caratteristiche grammaticali ma mantenendo la solita parte di parola o categoria (es. *walk, walks*);
2 – **Derivation**: processo nel quale una parola di una categoria diversa è derivata da un'altra parola o radice di parola attraverso l'applicazione di qualche processo (es. *grammar -> grammatical, grammatical ->grammaticality*);
3 – **Compounding**: processo nel quale parole indipendenti si uniscono in qualche modo per ottenere una nuova unità (es. *buttonhole*).

65

5.2.1 - Inflection

Di regola, i dizionari di carta astraggono dall' **inflection**. Ci sono varie ragioni per giustificare tale scelta;

1 – Il processo di **inflection** è relativamente regolare e, una volta che si sono isolate le eccezioni, tale processo si applica a tutti i membri di una data categoria. Per esempio (in inglese) , per formare la terza persona singolare del 'present tense' dei verbi semplicemente si aggiunge una *s* o una *es* alla forma base del verbo. Ci sono molte poche eccezioni a tale regola, ed esse devono essere descritte esplicitamente.
2 – Ciò risparmia spazio, tempo e sforzo nel costruire le entrate del dizionario. Dal momento che l'inglese ha dei processi di inflection piuttosto poveri, tale risparmio non è enorme. Ma in italiano o in spagnolo esistono sei diverse forme verbali solo per il presente e ciò evidenzia l'enorme risparmio che si ha nel costruire il dizionario se si trascura il processo di inflection.

66

Nel contesto della MT è chiaramente desiderabile utilizzare un approccio simile, dove il dizionario monolingue e quello della fase di transfer contengono solo le HEADs e non 'inflected words'.
Per realizzare ciò un sistema deve essere capace di catturare gli schemi regolari del processo di inflection. Ciò può essere fatto aggiungendo al sistema un **componente morfologico** che descrive tali processi in termini di regole, con regole esplicite addizionali per i casi irregolari. Tale componente dovrà riuscire ad associare alle parole 'inflected' la corrispondente 'head word' ed ad estrapolare il significato che il processo di inflection ha aggiunto alla parola base.

Esempio: Consideriamo sempre il verbo *affects* nella semplice frase *Temperature affects printer density*. Prima di tutto vogliamo che il nostro componente morfologico riconosca *affects* come una forma 'inflected' di *affect*. Secondariamente, non vogliamo perdere le informazioni aggiunte dal suffisso in modo tale che esse possano essere utilizzate nel generare la frase di uscita.

67

Ci sono vari modi di descrivere tali informazioni, ma probabilmente la più semplice è la seguente:

(lex=V, cat=v, +finite, person=3rd, number=sing, tense=pres) <-> V+s

Abbiamo introdotto una regola la quale dice che i verbi finiti in terza persona singolare coniugati in 'present tense' possono essere formati aggiungendo una *s* alla forma base rappresentata dal valore dell'attributo 'lex'. Tale regola può essere letta anche nella direzione opposta: se una parola può essere divisa in una stringa di caratteri e una *s*, allora essa può essere un verbo finito coniugato alla terza persona singolare del present tense.

Altre regole dovrebbero essere date per indicare che la *s* finale può essere aggiunta a tutti i verbi, tranne che a quelli che terminano in *s, ch, sh, o, x* e *z* ai quali si aggiunge *es*.

68

La ricerca del termine che rappresenta la forma base del verbo può essere fatta nel dizionario monolingue. Così, se l'analizzatore morfologico incontra una parola come *affects*, controllerà se all'interno del dizionario monolingue esiste una entrata con le caratteristiche *cat = v, lex = affect.* Dal momento che tale entrata esisterà sicuramente, *affects* può essere rappresentato per mezzo delle informazioni contenute nella rispettiva entrata del dizionario e di quelle fornite dalla regola applicata del componente morfologico. Il risultato delle analisi morfologiche è quindi una rappresentazione che consiste sia delle informazioni fornite dal dizionario che delle informazioni fornite dal suffisso.

lex = affect	*sem_patient = ?*
cat = v	*vform = finite*
vtype = main	*person = 3rdSing*
subcat = subj_obj	*tense = pres*
sem_agent = ?	

69

Al fine di riconoscere le forme irregolari il componente morfologico deve contenere regole esplicite. Per esempio potremmo descrivere tale eccezioni nel seguente modo:

(lex=be,cat=v,+finite,person=3rd,number=sing,tense=pres) <-> is

(lex=have,cat=v,+finite,person=3rd,number=sing,tense=pres) <-> has

Per essere sicuri che le regole delle forme regolari non producano *bes* e *haves*, potremmo dividere le regole in due insiemi; un gruppo di regole eccezione e uno di regole di default. Dovremmo poi assicurarci che nessuna regola di default venga utilizzata nel caso in cui una regola eccezione può essere applicata.

70

5.2.2 - Derivation

Il processo di derivazione forma nuove parole (generalmente di una categoria diversa) da parole esistenti. Per esempio, *industrialization*, e *destruction* possono essere viste come derivate nella maniera illustrata sotto.

a. $[_N\ [_V\ [_{ADJ}\ [_N$ *industry]* + *ial]* + *ize]* + *ation]*

b. $[_N\ [_V$ *destroy]* + *ion]*

OSS: Come si può vedere dall'esempio di *destruction*, non appare necessariamente la forma di citazione della parola nella derivazione, e per questa ragione è comune parlare di processi di derivazione che utilizzano la radice della parola (o 'stem').

71

Alcuni dei processi di derivazione sono piuttosto regolari e possono essere descritti per mezzo di una grammatica. Ciò significa:

1 – inserire i vari prefissi e suffissi nel dizionario;

2 – permettere loro di subcategorizzare per ciò con cui essi possono combinarsi (es. *–able* si combina con verbi transitivi come read -> readable).

3 – assicurarsi che le regole che combinano parole con suffissi e prefissi diano alla parola derivata le caratteristiche giuste per il risultato, e gestiscano qualsiasi possibile cambiamento di scrittura della parola e della parte aggiunta.

4 – trovare un modo di specificare il significato in termini dei significati della parola e della parte aggiunta.

72

Un approccio per gestire la morfologia derivazionale nel campo della MT è semplicemente elencare tutte le parole derivate, e per alcune di esse tale approccio è sicuramente il più giusto in quanto il loro significato è impredicibile.

Esempio: Consideriamo il suffisso '*ing*'.

a. *Painting*: può rappresentare un prodotto (il dipinto)
b. *Covering*: può rappresentare una cosa (la copertina) che realizza l'azione di coprire.
c. *Cutting*: può rappresentare una cosa (il ritaglio) che subisce l'azione di tagliare.
d. *Crossing*: può rappresentare un posto (l'incrocio).

Inoltre i termini del tipo X-ing hanno di solito anche come significato ' l'azione di Xing'.

OSS: Ciò evidenzia come ci sia quasi sempre un problema di ambiguità con le parole derivate.

73

Riguardo alla traduzione, ci sono casi in cui si può tradurre le parole derivate traducendo la radice (stem) e il particolare prefisso o suffisso.

Esempio: La traduzione in italiano degli avverbi inglesi formati da un aggettivo più *–ly* è spesso realizzata traducendo l'aggettivo e aggiungendogli *–mente* (es. *quick+ly -> rapido+mente, easy+ly -> facile+mente*).

Ma ciò non è possibile per tutti i prefissi e suffissi. Le difficoltà nel tradurre le parole derivate traducendo separatamente lo 'stem' e l'aggiunta possono essere viste dalla traduzione dei termini seguenti in tedesco;

a. *Killing -> doden*
b. *driving off -> wegrijden*
c. *painting (the act) -> schilderen*

74

Dagli esempi precedenti si nota una relazione tra le parole inglesi terminanti in *ing* e quelle tedesche terminanti in *en*. I successivi esempi fanno però crollare la nostra ipotesi;

d. *painting (the product) <> schilderen*, ma *-> schilderij*
e. *covering <> bedekken*, ma *-> bedekking*
f. *cutting <> knippen*, ma *-> knipsel*
g. *crossing <> kruisen*, ma *-> kruispunt*

Quindi, sebbene l'idea di fornire regole per tradurre le parole derivate può sembrare attraente, essa solleva troppi problemi e così attualmente è più un obiettivo della ricerca sulla MT che una possibilità pratica.

75

5.2.3 - Compounds

Un compound è la combinazione di due o più parole che funge da parola singola. In inglese, il tipo più comune di compound è probabilmente quello composto di due nomi, come quelli nell'entrata del dizionario per *button*:

a. **buttonhole**: $[_N [_N$ *button* $] [_N$ *hole* $]]$

b. **buttonhook**: $[_N [_N$ *button* $] [_N$ *hook* $]]$

c. **button mushroom**: $[_N [_N$ *button* $] [_N$ *mushroom* $]]$

OSS: Ortograficamente, linguaggi diversi seguono diverse convenzioni. Per esempio, in italiano i *compound* sono generalmente scritti come una singola parola, ma in inglese alcuni sono scritti come parola singola (es. *buttonhole*), altri come parole affiancate (es. *small-scale*) e altri come parole giustapposte (es. *button mushroom*).

76

Come per la derivazione, è possibile descrivere i possibili **compounds** per mezzo di una grammatica su parole, e come per la derivazione la possibilità di poter realizzare la traduzione traducendo le parti componenti è molto attraente, specialmente perchè non è possibile elencare tutti i compounds in inglese in quanto teoricamente si possono ottenere parole di lunghezza arbitraria.

Esempio:

a. student film
b. student film society
c. student film society committee
d. student film society committee scandal
e. student film society committee scandal inquiry

77

Sfortunatamente, sebbene ci siano casi in cui decomporre un **compound** e tradurre le sue parti restituisce il risultato corretto (es. in tedesco *Wassersportverein* si traduce come *water sport club*), i problemi della interpretazione e della traduzione sono perfino più grossi di quelli incontrati per la derivazione.

Ci sono problemi di ambiguità. Per esempio, *student film society* potrebbe avere entrambe le strutture indicate sotto, con differenti interpretazioni;

a. $[_N [_N$ *student film] society]*
   può rappresentare la società dei film sugli studenti
b. $[_N$ *student* $[_N$ *film society]]*
   può rappresentare la società di film composta da studenti

78

Un altro tipo di ambiguità può essere illustrato con il seguente esempio: *satellite observation* può in una occasione significare ‘osservazione da satellite’ mentre in altre occasioni può significare ‘osservazione del satellite’.

IMP: In generale esiste un’ampia varietà di relazioni possibili tra elementi di un **compound**. Così, *buttonhole* è un foro per bottoni, ma *button mushroom* è un fungo che assomiglia ad un bottone e non un fungo per bottoni. Non è chiaro come queste relazioni possono essere catturate.

La maggior parte delle volte il lettore umano riesce, basandosi sulla conoscenza del mondo o sul particolare contesto, a decifrare il particolare significato delle composizioni che si trovano nelle frasi.

Così, come per la derivazione, un approccio realmente generale per il trattamento dei **compounds** rimane un obiettivo della ricerca in MT più che una possibilità pratica.

79

CAPITOLO 6: Le Problematiche della Traduzione

In questo capitolo considereremo alcune problematiche particolari che il compito della traduzione pone al costruttore del sistema per la MT

80

Le problematiche che andremo ad evidenziare sono tra quelle che rendono il compito della MT veramente difficile.

E' utile pensare che queste problematiche siano divise in tre gruppi concettuali:

1- Problemi di **ambiguità**

2- problemi che nascono a causa delle **differenze strutturali e lessicali** tra i linguaggi

3- unità multiparola come gli **idiomi**

81

6.1 - Ambiguità

Nel migliore dei mondi possibile, ogni parola avrebbe un solo significato. Ma, come noi ben sappiamo, ciò non è la realtà.

Quando una parola ha più di un significato, allora essa è detta essere **lessicalmente ambigua.**

Quando un sintagma o una frase possono avere più di una struttura essi sono detti essere **strutturalmente ambigui.**

OSS: L'ambiguità è un fenomeno pervasivo nelle lingue umane. E' molto difficile trovare parole che non abbiano almeno ambiguità 2 (con ambiguità 2 intendiamo una parola con due possibili significati), ed è normale trovare frasi con (fuori dal contesto) parecchi gradi di ambiguità. Ciò non è problematico per il solo fatto che alcune delle possibili interpretazioni sono sbagliate, ma anche perché le ambiguità si moltiplicano (vedi esempio seguente).

82

Esempio: Nel caso peggiore, una frase contenente 2 parole, ognuna delle quali con ambiguità 2, può avere ambiguità (2 x 2), una frase con tre parole può avere ambiguità (2 x 2 x 2)=8. Secondo questa logica si possono ottenere numeri veramente elevati. Per esempio, una frase che consiste di 10 parole, ognuna delle quali ha ambiguità 2, e con 2 possibili analisi strutturali potrebbe avere $2^{9+2}=2^{11}=2048$ analisi diverse. Fortunatamente, comunque, le cose non sono sempre così pessime.

Immaginiamo di dover tradurre queste due frasi in italiano:

a. *You must not use abrasive cleaners on the printer casing.*

b. *The use of abrasive cleaners on the printer casing is not recommended.*

Nella prima frase *use* è un verbo, e nella seconda un nome, cioè abbiamo un caso di **ambiguità lessicale**.

83

Un dizionario italiano tradurrebbe la stessa parola *use* in due modi diversi a seconda che si tratti di un nome o di un verbo.

Un modo per capire quale è la giusta categoria da attribuire ad *use* è verificare se è grammaticalmente possibile avere un nome o un verbo nella posizione dove occorre. Per esempio, non esistono in inglese sequenze grammaticali di parole che consistono di *the + V + PP*, e così per esempio nella frase (b) l'unica soluzione possibile è che *use* rappresenta un nome.

Come abbiamo già visto, noi possiamo dotare i sistemi di traduzione della capacità di riconoscere frasi grammaticalmente corrette fornendogli una grammatica sotto forma di regole del linguaggio considerato. Ciò è molto utile perché permette di escludere molte analisi della frase dall'insieme di tutte le analisi possibili.

84

OSS: Ad ogni modo, il dare al nostro sistema una conoscenza riguardo la sintassi non ci permette di risolvere il problema dell'ambiguità. Ciò perché le parole possono avere diversi significati anche all'interno della solita categoria sintattica. Per esempio *button* può essere sia un nome che un verbo. Restringendoci alla categoria nome, *button* può essere sia un 'bottone' che un 'pulsante'.

IMP: Ne deduciamo che è necessario fornire alla macchine delle conoscenze riguardo il significato delle parole.

Esempio: Consideriamo la seguente frase

Cleaning fluids can be dangerous.

Una possibile analisi vede *cleaning* come verbo e un'altra come aggettivo. E' chiaro che l'interpretazione che vede *cleaning* essere un verbo è meno realistica dell'altra, ma è importante notare come tale osservazione nasce da nozioni sul significato delle parole e non sintattiche.

85

6.2 – Incongruenze lessicali e strutturali

All'inizio della sezione precedente si è detto che noi vorremmo vivere (per gli scopi della MT) in un mondo in cui ad ogni parola corrisponde un unico significato. Sarebbe sicuramente una situazione più gestibile, ma non la migliore in assoluto.

1- Alcuni dei problemi che continuerebbero ad esistere hanno a che fare con le differenze lessicali tra le lingue, differenze sulla maniera in cui le diverse lingue sembrano classificare il mondo, quali concetti esse scelgono come esprimibili in una singola parola, e quali non vengono lessicalizzati.

2- Altri problemi nascono in quanto lingue diverse utilizzano diverse strutture per lo stesso scopo, e la solita struttura per scopi diversi.

In entrambi i casi è necessario complicare notevolmente il processo di traduzione per ottenere risultati accettabili.

86

Proponiamo degli esempi di diversa classificazione del mondo esibita da diversi linguaggi:

a. *Know (V)* -> *savoir (un fatto)*
connaître (una cosa)
b. *leg (N)* -> *patte (di un animale)*
jambe (di un umano)
pied (di un tavolo)
c. *brown (A)* -> *brun*
châtain (di capelli)
marron (di scarpe/pelle)
d. *wear/put on (V)* -> *kiku*
haku (scarpe)
kakeru (occhiali)
kaburu (cappelli)
hameru (guanti)
haoru (cappotto)
shimeru (sciarpa)

87

Osserviamo come il compito di scegliere la migliore traduzione possibile per un termine che nel linguaggio obiettivo ha più associazioni può, nel caso più banale, richiedere la sola lettura delle parole che compaiono nel testo sorgente. In casi meno fortunati, ciò può avvenire attraverso deduzioni semantiche sugli oggetti che compongono l'ambiente in questione. Esiste però anche la possibilità che tale scelta per la traduzione si riveli un problema in qualche particolare istanza perfino indecidibile.

OSS: Pertinenti con i problemi lessicali sono anche tutti quei problemi che nascono dalla presenza nella associazione di termini del linguaggio sorgente con quelli del linguaggio obiettivo dei cosiddetti **lexical holes**. Con tale termine indichiamo quei casi in cui un linguaggio deve utilizzare un sintagma o una espressione linguistica per esprimere ciò che in un altro linguaggio è esprimibile da una singola parola (es. *suicidarsi* in italiano deve essere tradotto in inglese da *to commit suicide*).

88

Abbiamo anche introdotto le incongruenze strutturali tra due lingue, e si è visto come queste compaiano nel momento in cui due linguaggi utilizzano la stessa costruzione per scopi diversi, o usano costruzioni diverse per quello che sembra essere lo stesso scopo.

Forniamo ora degli esempi in cui vengono utilizzatec costruzioni diverse per ottenere lo stesso effetto;

Esempio:

(1) a. *He is called Sam.*
b. *Er heißt Sam.*
'He is named Sam'.
c. *Il s'appelle Sam* oppure *Si chiama Sam.*
'He calls himself Sam'

(2) a. *Sam has just seen Kim.*
b. *Sam vient de voir Kim.*
'Sam comes of see Kim'

Il problema fondamentale in questi casi è che la rappresentazione astratta della frase nel linguaggio sorgente e quella della rispettiva traduzione nel linguaggio obiettivo sono notevolmente diverse. Il passare dall'una all'altra richiede regole di trasformazione particolarmente complesse. Per la frase (2) tali regole devono necessariamente realizzare i seguenti punti:

1- L'avverbio *just* deve essere tradotto nel verbo *venir-de* anche se tale associazione appare del tutto innaturale.
2- *Sam*, il soggetto di *see* deve diventare il soggetto di *venir-de*.
3- Alcune informazioni riguardo la coniugazione del verbo devono essere prese dal nodo S del quale *see* è la HEAD, e portate sul nodo S la cui HEAD è *venir-de.* Ciò è una complicazione, in quanto, normalmente ci si aspetta che tale informazione vada a finire sul nodo la cui HEAD è la traduzione di *see* cioè *voir*.

6.3 – Unità Multiparola: gli Idiomi

Informalmente, gli idiomi possono essere visti come espressioni il cui significato non può essere completamente compreso dal significato delle parti componenti.

Per esempio, mentre è possibile estrarre il significato della frase (1a) sulla base della conoscenza della grammatica inglese e del significato delle parole, tali nozioni non sono sufficienti per capire il significato della frase (1b).

(1) a. *If Sam mends the bucket, her children will be rich.*
b. *If Sam kicks the bucket, her children will be rich.*

Il problema è che *kick the bucket* è un **idioma** che nel suo insieme significa 'morire'.

In molti casi, una traduzione naturale per un idioma è data da una singola parola.

OSS: I lexical holes e gli idiomi rappresentano normalmente istanze di traduzione del tipo word <-> phrase. La differenza è che con i 'lexical holes', il problema di solito si pone nel tradurre dalla lingua con la 'word' alla lingua che utilizza il 'phrase', mentre con gli idiomi, si manifestano i problemi nel tradurre dalla lingua che contiene l'idioma (phrase) alla lingua che utilizza una singola 'word'.

Un possibile approccio per la gestione degli idiomi è quello di rappresentarli come unità singole nel dizionario monolingue. Ciò significa che all'interno del dizionario si avrà un'entrata lessicale del tipo **kick_the_bucket**.

Il vero problema con gli idiomi è che questi non hanno generalmente una forma fissa, e che le variazioni della loro forma non sono limitate a variazioni di 'inflection'. Esiste, quindi, un serio problema a riconoscere gli idiomi.

Gli idiomi possono variare nella forma del verbo, in base al tempo, alla persona e al numero.

Esempio: Consideriamo l'idioma *bury the hatchet* che significa porre fine alle ostilità e riconciliarsi. Tale forma può variare notevolmente a seconda del particolare contesto.Alcune variazioni potrebbero essere:

He buries / buried / will bury the hatchet

They bury / buried / shall bury the hatchet

93

Una seconda forma di variazione comune è la forma del pronome possessivo nell'espressione.

Esempio: Consideriamo l'idioma *to burn one's bridges* che significa 'darsi la zappa sui piedi'. Tra le possibili variazioni appartenenti alla categoria considerata ci sono:

He has burned his bridges.

She has burned her bridges.

Variazioni possono anche riguardare la configurazione sintattica.

Esempio: Consideriamo sempre l'idioma *bury the hatchet*. Esso può apparire sia nella forma attiva che nella forma passiva:

He buried the hatchet

94

The hatchet seems to have been buried.

Tutti gli esempi indicati evidenziano la complessità del trattamento degli idiomi nel campo della traduzione automatica. Molti idiomi per essere riconosciuti richiedono un'analisi sintattica molto dettagliata.

Allo stato attuale i sistemi per la MT non riescono ancora a garantire risultati accettabili nel trattamento degli idiomi soprattutto di quelli che possono presentarsi in svariate forme sintattiche.

95

CAPITOLO 7: Rappresentazione del Significato

All'interno di questo capitolo parleremo dell'importanza di arricchire la conoscenza del sistema con rappresentazioni orientate al significato.

96

Le varie discussioni nei capitoli precedenti hanno mostrato che per realizzare una traduzione di qualità non è sufficiente la sola analisi sintattica. Esistono molti casi in cui il problema sembra richiedere una conoscenza più profonda, più orientata verso il significato.

E' utile pensare a questo tipo di conoscenza come suddivisa i tre tipi:

1- conoscenza linguistica indipendente dal contesto detta anche **conoscenza semantica.**

2- conoscenza linguistica legata al contesto detta anche **pragmatica**.

3- conoscenza in generale non linguistica, basata sul senso comune e sulla conoscenza del mondo detta anche **conoscenza del mondo reale**.

97

7.1 – La Semantica

La semantica riguarda il significato delle parole e come queste si combinano per costruire il significato dell'intera frase.

Ci sono vari modi di pensare e di rappresentare il significato delle parole, ma un modo per il quale è stata dimostrata l'utilità nel campo della MT è quello di associare alle parole delle caratteristiche semantiche che corrispondono alle loro componenti di senso.

Esempio:

man = (+HUMAN, +MASCULINE and +ADULT)
woman = (+HUMAN, -MASCULINE and + ADULT)
boy = (+HUMAN, +MASCULINE and –ADULT)
girl = (+HUMAN, -MASCULINE and –ADULT)

98

OSS: Associare alle parole delle caratteristiche semantiche è utile in quanto alcune di esse impongono vincoli semantici sulle parole con le quali possono comparire.

Consideriamo il verbo *eat* il quale richiede che il suo AGENT sia un essere animato e che il suo PATIENT sia commestibile, concreto (piuttosto che astratto come *sincerità* o *bellezza*), e solido (in quanto non si mangia una cosa liquida tranne poche eccezioni che non consideriamo). Noi possiamo codificare questi vincoli nella nostra grammatica associando le caratteristiche HUMAN e EDIBLE con appropriati nomi nel nostro dizionario e descrivendo la nostra entrata per *eat* come qualcosa del genere *cat=verb, AGENT=HUMAN, PATIENT=EDIBLE.*

La grammatica ora accetterà solo oggetti commestibili per il verbo *eat*, realizzando così una selezione che elimina tutte le analisi che non soddisfano i requisiti semantici descritti.

99

Esempio: Consideriamo la frase seguente:

John ate the game.

La parola inglese *game* è ambigua in quanto può avere più significati. Essa può significare, tra le altre cose, 'una gara sportiva' o 'della cacciagione'. Utilizzando i vincoli descritti sopra, possiamo escludere dalle possibili interpretazioni quella di 'gara sportiva', supponendo comunque che il sistema sia in grado di dedurre che la cacciagione è qualcosa di commestibile mentre la gara sportiva non lo è.

OSS: Esiste un stile linguistico che mette in gravi difficoltà tutte le teorie semantiche viste fino ad ora. Si tratta dell'uso all'interno del linguaggio di metàfore che rendono il testo molto figurativo. Consideriamo la frase

This car eats money.

100

La frase precedente è chiaramente utilizzata per indicare il fatto che la macchina in questione richiede molti soldi per essere mantenuta.
Con un simile stile linguistico non è proponibile cercare di soddisfare i vincoli semantici sul verbo *eat* anche perché i soldi non sono qualcosa di commestibile.

7.2 – La pragmatica

Ricordiamo che la pragmatica si riferisce al significato dipendente dal particolare contesto. Per contesto intendiamo sia il resto del testo all'interno del quale occorre la frase, sia tutte le circostanze esterne al testo stesso come chi è l'autore e qual'è la sua particolare posizione sociale.

Per introdurre le varie problematiche relative alla pragmatica vediamo degli esempi.

101

Esempio: Analizziamo la traduzione dei cosiddetti **anaphoric pronouns** che rappresentano quei pronomi che si riferiscono a oggetti precedentemente incontrati nel testo. Consideriamo la frase

Sam took the cake from the table. Then he ate it.

Supponiamo di voler tradurre tale frase dall'inglese all'italiano. Noi sappiamo che *it* deve riferirsi a qualche nome singolare nella parte di testo precedente. Il pronome *it* può potenzialmente riferirsi a tre diversi sintagmi nominali che sono *Sam*, *the cake* o *the table*. La struttura sintattica dell'inglese costringe il pronome a concordare nel numero e nel genere con i suoi antecedenti, e quindi *it* non può riferirsi a *Sam* in quanto pronome neutro. Tale osservazione ci lascia la scelta tra *the cake* e *the table*. Potremmo sperare a questo punto che in entrambi i casi la traduzione sia la stessa. Sfortunatamente però 'il tavolo' è un termine maschile e 'la cioccolata' è un termine femminile e ciò incide sulla traduzione in italiano.

102

Nell'esempio particolare che stiamo considerando potremmo risolvere i problemi di ambiguità utilizzando ancora i vincoli semantici associati al verbo *eat*. Ciò porterebbe all'esclusione tra le varie alternative di *table* in quanto rappresenta un oggetto non commestibile.

OSS: Automatizzare il procedimento di risoluzione dell'ambiguità esaminato nell'esempio precedente non è particolarmente impegnativo, ma sfortunatamente le cose non sono sempre così facili. Le cose potrebbero essere complicate dal fatto che il pronome si riferisce ad oggetti che non compaiono né nella frase corrente né in quella precedente.

Esempio:
a. A: Now insert the cartridge at the back.
b. B: Okay.
c. A: By the way, did you order more toner today?
d. B: Yes, I got some when I picked up the new paper.
e. A: OK, how far have you got?
f. A: Did you get **it** fixed?

103

It nell'ultima frase si riferisce alla *cartridge*, sebbene questa fosse stata menzionata per l'ultima volta nella prima frase.
Per affrontare questi tipi di testo è necessario vedere il dialogo precedente non come una struttura intera, o una sequenza di frasi, ma piuttosto come una serie di **segmenti**, dove un segmento è una porzione di discorso (non necessariamente continua) nella quale le frasi si riferiscono al solito argomento. Sintagmi particolari come *By the way* segnalano dove finisce un segmento e ne comincia un altro.

METODO: Noi quindi vincoleremo il anaphoric pronoun ad appartenere al solito segmento dell'oggetto riferito.

Nell'esempio precedente ci sono tre ovvi referenti per *it*: la *cartridge* (a), *toner* (c), e *paper* (d). Ad ogni modo le frasi (c) e (d) appartengono ad un altro segmento rispetto a quello cui appartiene *it* (tale segmento è una digressione che comincia con *by the way* e termina con *OK*). La *cartridge* è quindi l'unico referente possibile per *it*.

104

Consideriamo adesso il lato della pragmatica non dipendente dal testo ma dalle circostanze esterne. Non faremo particolari trattazioni teoriche ma riporteremo semplicemente un esempio per focalizzare il problema;

Come interpreteremo la frase sottostante ? Come un comando (per esempio dato dal datore di lavoro) o come un suggerimento (che potrebbe essere dato dal commesso di un negozio) ?

The front cover should be closed.

Notiamo che il fatto che la frase precedente venga interpretata come un comando o come un suggerimento ha influenza sulla eventuale traduzione per diverse lingue obiettivo.

105

7.3 – Conoscenza del Mondo Reale

Non tutta la conoscenza di cui abbiamo bisogno per estrarre il significato di frasi e per tradurle può essere trovata nei testi cui appartengono. Consideriamo i seguenti esempi:

a. *Little Johnny was very upset. He had lost his toy train. Then he found it. It was in his **pen**.*

b. *I saw the soldiers aim at the women, and I saw several of **them** fall.*

c. *The council refused the women a permit because **they** advocated violence.*

106

Nell'esempio (a) *pen* deve essere interpretata come 'box' e non come 'penna per scrivere', in quanto affinchè A sia in B deve valere che A è più piccolo di B, ed in questo caso il trenino giocattolo è più piccolo del 'box' ma non della penna da scrivere.

Nell'esempio (b) la domanda è chi cadde a terra, i soldati o le donne? In generale, sappiamo che il mirare è spesso seguito dallo sparare, e che generalmente a cadere a terra sono le persone verso cui si è mirato e non quelle che miravano.

Nell'esempio (c) non è chiaro chi sosteneva la violenza, il consiglio o le donne? Anche in questo caso la conoscenza delle situazioni del mondo reale ci fanno pensare che a favorire la violenza siano le donne in quanto ciò rappresenta una giustificazione plausibile da parte del consiglio per negare un permesso.

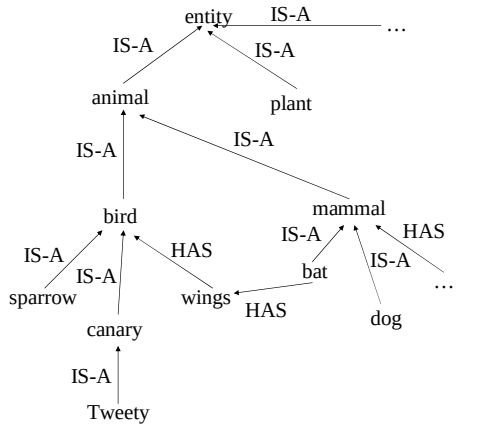
107

In tutte le spiegazioni precedenti si è utilizzato una conoscenza che non è linguistica, ma si sono seguiti dei ragionamenti dettati soprattutto dal senso comune da una conoscenza generale e da fatti riguardanti domini più ristretti.

Il rappresentare e manipolare una tale conoscenza automaticamente è uno dei più importanti campi di ricerca dei nostri tempi e probabilmente la ragione di esistere di un'intera disciplina, cioè l'intelligenza artificiale (AI).

Un modo particolarmente comodo di rappresentare tale conoscenza è dato dalle **reti semantiche** (semantic net) di cui diamo un esempio di utilizzo di seguito.

108



109

Intuitivamente, i nodi in una tale rete rappresentano cose, ed i collegamenti tra di essi sono relazioni. Ciò significa che la struttura può essere facilmente generalizzata per altri tipi di relazioni. Per esempio, aggiungendo altri oggetti, e utilizzando una relazione 'parte di', si può rappresentare il fatto che una stampante è costituita di vari componenti e che questi a loro volta hanno altri sottocomponenti. Una tale informazione potrebbe essere utile nell'interpretare frasi come la seguente;

Put the toner in the cartridge in the reservoir.

Il sapere che il serbatoio non ha una cartuccia come sua parte permetterebbe di dedurre che quella sopra rappresenta un'istruzione per mettere il toner che è nella cartuccia nel serbatoio, piuttosto che mettere il toner in una particolare cartuccia (cioè quella cartuccia che si trova nel serbatoio).

110

CONCLUSIONI: Abbiamo ora un modo di rappresentare almeno alcune delle conoscenze del mondo reale. Allo stato attuale si ha che:

1- Il problema di manipolare la conoscenza del mondo in una maniera simile a come questa viene gestita dall'uomo è un problema irrisolto e forse anche irresolubile (questione filosofica).

2- Sotto particolari circostanze restrittive, si può riuscire a fare qualcosa di utile. Per circostanze restrittive intendiamo ambienti specifici nei quali esistano pochi oggetti e con relazioni piuttosto limitate.

111

CAPITOLO 8: Le Nuove Direzioni della MT

In questo capitolo diamo uno sguardo a quelli che oggi sono argomenti di ricerca ma che con buona probabilità diventeranno parte integrante dei futuri sistemi per la MT.

112

Tra i nuovi approcci alla traduzione automatica, in questo capitolo analizzeremo solo quelli classificati come **approcci empirici**. Tali approcci utilizzano tecniche di 'pattern matching' e basate su statistiche.

Con il termine empirico si vuole evidenziare come qualsiasi conoscenza linguistica che il sistema utilizza viene derivata empiricamente, esaminando testi reali, piuttosto che esplicitata da qualche linguista.

Vedremo in particolare due di questi approcci: l'approccio **Example-Based** e quello **statistico**.

8.1 – La traduzione Example-Based

L'idea base di questo approccio è quella di avere a disposizione grosse quantità di esempi di traduzione che vengono poi riutilizzati per dirigere traduzioni future. Ciò avviene andando a ricercare il particolare sintagma da tradurre all'interno degli esempi memorizzati in modo tale che la traduzione registrata fornisca indicazioni sulla traduzione attuale.

OSS: Tale idea è riconducibile al modo in cui un traduttore umano realizza una traduzione servendosi di un dizionario bilingue: osservando gli esempi forniti all'interno del dizionario si cerca un esempio che approssimi nel migliore dei modi possibili ciò che deve essere tradotto, e successivamente si costruisce una traduzione sulla base della traduzione fornita dal dizionario per quel particolare esempio.

Esempio: Un generico dizionario bilingue (inglese-francese) potrebbe presentare le seguenti informazioni in corrispondenza dell'entrata *printer*:

a. *Print's error* faute *f* d'impression, coquille *f*;

b. *Print's reader* correcteur *m*, -trice *f* (d'épreuves).

Supponiamo di dover tradurre la frase seguente:

This seems to be a printer's mistake.

Un traduttore umano certamente sceglierebbe *faute d'impression* o *coquille* come traduzione, sulla base del fatto che un *mistake* è molto più simile ad un *error* che ad un *reader*.

Il cercare l'esempio che meglio approssima il sintagma da tradurre può richiedere il calcolo della *vicinanza* tra gli argomenti in una gerarchia di termini e concetti fornita da una specie di dizionario dei sinonimi.

Una possibile estensione di questa idea base è data dall'introduzione di coppie che relazionano espressioni del linguaggio sorgente con espressioni del linguaggio obiettivo, includendo anche esempi di traduzione scritti tra parentesi e interpretabili come descriventi le condizioni sotto le quali l'equivalenza data vale.

Esempio: Consideriamo la regola per la parola giapponese *sochira* data sotto;

sochira ⟶
this ((desu {be}),…)
you ((okuru {send}),…)
this ((miru {see}),…)

La regola precedente indica che *sochira* si traduce come *this* quando l'esempio include *desu* (che si traduce *be*), come *you* quando l'input contiene qualcosa come *okuru* (che si traduce *send*).
Se volessimo tradurre un ingresso come *sochira ni tsutaeru*, verrebbe selezionato il pronome *you* come traduzione, in quanto *tsutaeru* (convey) è più vicino ad *okuru* (send) nella gerarchia del dizionario dei sinonimi.

OSS: E' evidente che la fattibilità dell'approccio example-based dipende in maniera cruciale dalla collezione di 'buoni' dati.

VANTAGGI: 1) La qualità della traduzione aumenta incrementalmente man mano che l'insieme di esempi si rende più completo, senza il bisogno di aggiornare ed incrementare le descrizioni dettagliate del lessico e della grammatica.
2) L'approccio può essere molto efficiente in quanto nel migliore dei casi non c'è da applicare alcuna regola complessa, tutto ciò che c'è da fare è trovare l'esempio appropriato e talvolta calcolare le 'distanze'.

117

SVANTAGGI: Il principale problema di questo approccio è quello che si pone quando esistono più esempi ognuno dei quali concorda con parte della stringa in input ma non la copre nella sua interezza. In tali casi, calcolare l'esempio migliore può richiedere di considerare un gran numero di possibilità.

8.2 – MT Statistica

L'approccio può essere visto come un tentativo di applicare alla MT le tecniche che hanno avuto risultati soddisfacenti nel campo della *'speech recognition'* , e sebbene una discussione dettagliata richieda una trattazione statistica sofisticata, è possibile introdurre i concetti base di questo approccio in maniera piuttosto semplice.

118

Le due nozioni chiave di nostro interesse sono quelle di **language model** e quella di **translation model**.

Il language model ci fornisce le probabilità per le stringhe di parole (cioè le frasi), le quali possiamo denotare con Pr(S) (per una frase S espressa nel linguaggio sorgente) e con Pr(T) (per ogni frase T espressa nel linguaggio obiettivo). Intuitivamente, Pr(S) è la probabilità di presentarsi di una stringa di parole S espresse nel linguaggio sorgente. Analogo è il significato di Pr(T) rivolto invece al linguaggio obiettivo.

Il translation model ci fornisce invece le probabilità Pr(T|S), le quali rappresentano le probabilità condizionate che una frase target T occorrerà in un testo espresso nel linguaggio obiettivo che traduce un altro testo espresso nel linguaggio sorgente all'interno del quale compare la frase S.

119

Sfruttando le formule sul calcolo delle probabilità otteniamo

$$Pr(S,T) = Pr(T|S)\ Pr(S)$$

la quale rappresenta la probabilità che la coppia (T,S) compaia nei testi delle rispettive lingue.

Come viene calcolata la Pr(S) ?
Tale calcolo può essere decomposto nella probabilità della prima parola moltiplicata per la probabilità condizionale delle parole successive, nel seguente modo:

$$Pr(s1)\ x\ Pr(s2|s1)\ x\ Pr(s3|s1,s2)\ x\ \ldots\ldots\ldots$$

Intuitivamente, la probabilità condizionata Pr(s2|s1) è la probabilità che s2 si presenterà, supponendo che si è appena presentato s1;

120

Per esempio, la probabilità che *am* ed *are* compaiano in un testo potrebbe essere approssimativamente la stessa, ma la probabilità che *am* compaia dopo *I* è piuttosto alta, mentre quella di *are* è molto bassa.

STRATEGIA: Per mantenere i calcoli entro limiti di maneggevolezza, di solito nel calcolo delle precedenti probabilità condizionate si prendono in considerazione soltanto una o due parole precedenti.

OSS: Al fine di calcolare tutte queste probabilità sul linguaggio sorgente, è richiesta una grande quantità di dati monolingua, dati che incideranno in maniera forte sulla validità, l'utilità e l'accuratezza del modello ottenuto (con il termine *corpus* si indica proprio questa grande collezione di dati).

Un altro compito che richiede grosse quantità di dati è quello per la specifica dei parametri per il translation model, il quale richiede una grossa quantità di dati bilingue.

121

Consideriamo due tipi di *corpus* bilingue (inglese-francese):

A Sentence-Aligned Corpus

Often, in the textile industry, businesses close their plant in Montreal to move to the Eastern Townships.
Dans le domaine du textile souvent, dans Montreal, on ferme et on va s'installer dans les Cantons de l'Est.

There is no legislation to prevent them from doing so, for it is a matter of internal economy.
Il n'y a aucune loi pour empêcher cela, c'est de la régie interne.

But then, in the case of the Gulf refinery it is different: first of all, the Federal Government asked Petro-Canada to buy everything, except in Quebec.
Mais là, la différence entre la Gulf… c'est différent parce que la vente de la raffinerie Gulf: premiérement, le gouvernement fédéral a demandé á Petro-Canada de tout acheter, sauf le Québec.

122

Word Aligned Corpus

The Federal Government asked Petro-Canada to by everything.
Le(1) gouvernement(3) fédéral(2) a demandé(4) á Petro-Canada(5) de(6) tout(8) acheter(7).

In un *Word Aligned Corpus* vengono indicate quali parole del linguaggio target corrispondono ad ogni parola nel linguaggio sorgente. Il numero dopo le parole del linguaggio sorgente indicano la posizione della stringa della corrispondente parola o parole del linguaggio obiettivo.

DEF: La *fertilità* di una parola nella lingua sorgente è data dal numero di parole ad essa corrispondenti nella stringa obiettivo.

123

Esempio: La fertilità di *asked* è 2, in quanto essa si allinea con *a demandé.*

NOZIONE: La *distorsione* rappresenta il fatto che le parole del testo sorgente e le loro corrispondenti nel linguaggio obiettivo non necessariamente compaiono nella solita posizione all'interno della stringa (es. *tout acheter* e *buy everything*).

METODO:In base ad informazioni estratte automaticamente dal 'corpus' riguardanti le probabilità di fertilità per ogni parola del linguaggio sorgente (la probabilità che tale parola sia tradotta con una, due, tre etc. parole del linguaggio obiettivo), le possibili traduzioni e le probabilità di distorsione, vengono calcolate dal *translation model* le probabilità Pr(T|S).

124

Il problema a questo punto può essere ridotto a trovare la frase S che è la più probabile dato T. Si deve quindi scegliere la S che massimizza la quantità

$$Pr(S|T) = [Pr(S)\ Pr(T|S)] / Pr(T)$$ formula di Bayes

VANTAGGI: In un approccio come quello appena studiato il problema dell'acquisizione della conoscenza linguistica è completamente assente.

SVANTAGGI: 1) L'applicabilità generale del metodo potrebbe essere dubbia, in quanto essa è pesantemente dipendente dalla disponibilità di dati bilingue o monolingua di buona qualità ed in grosse quantità, il che è al momento mancante per molti linguaggi.
2) Parole che hanno una relazione morfologica tra di loro sono trattate come completamente separate l'una dall'altra, in maniera tale che le informazioni su *sees* non contribuiscono al calcolo dei parametri per *see* e *saw* (per risolvere tale problema si è cominciato ad inserire informazioni grammaticali di basso livello in questi sistemi).

125

BIBLIOGRAFIA

-**W.J.Hutchins and H.L.Somers –** " An Introduction to Machine Translation". *Academic Press, London, 1992.*

-**A.Spencer –** " Morphological Theory". *Basil Blackwell, Oxford, 1991.*

-**Ronnie Cann –** "Formal Semantics". *Cambridge University Press, Cambridge, 1993.*

Molte informazioni che compaiono in questo lavoro sono state prese dal sito web

www.essex.ac.uk/linguistics/clmt/MTbook

126