Metodi di valutazione per sistemi di traduzione automatica

Corso di

Elaborazione del linguaggio naturale

di francesca bonin

Anno accademico 2005-2006

Sommario

- Cenni agli MT system
- Problematiche legate alla valutazione
- Algoritmi per valutazione automatica
- Tassonomia di proprietà per una valutazione standard
- Valutazione NIST

Sistemi di machine translation- 1/2

Tradurre:

"dire quasi la stessa cosa"

(Umberto Eco)

Si potrebbe discutere a lungo sul significato della parola "quasi" della definizione di traduzione tradizionale, ma i sistemi di traduzione automatica, alla fine degli anni '80, hanno aperto nuove problematiche.

Tradurre automaticamente, infatti, significa

Programmare un computer perché capisca il testo e lo rielabori in un nuovo codice

Sistemi di machine translation- 2/2

Processo di traduzione:

- decodificare il significato del testo sorgente
- ri-codificare il significato in un nuovo codice
- strutturare questo significato in un testo

L'uomo, dietro a questa operazione, nasconde un complesso processo cognitivo che coinvolge competenze linguistiche, conoscenza del mondo e della situazione

Come può un computer raggiungere gli stessi risultati?

Può usare vari approcci…

Metodi e progetti di MT

- **Approccio dictionary based:**
  traduzione parola per parola senza preoccuparsi del contesto
- **Approccio statistico**:
  recupero di regolarità statistiche dallo studio di corpora bilingue (come il Canadian Hansard Corpus). Algoritmo che si può applicare a qualsiasi coppia di lingue purché rappresentato da ricche collezioni di corpora paralleli.
- **Approccio basato su interlingua**:

la traduzione target è generata a partire da un'interlingua language independent

Valutazione dei sistemi di MT

Valutare l'output di un' applicazione di traduzione è sicuramente complesso, perché, a differenza di quanto accade in matematica o fisica, il risultato linguistico mantiene un forte grado di soggettività

E' necessario creare parametri standard e universalmente riconosciuti che rendano confrontabili e quantificabili traduzioni diverse, tenendo conto del fatto che :

non esiste una traduzione perfetta, ma tante traduzioni perfette a seconda del contesto e dell'uso che si farà della traduzione.

Importante organizzare le metriche di valutazione per avere un metro di giudizio universale

Perché valutare?-1/2

1. Per i ricercatori che così hanno un feedback sui propri sistemi e un metro di paragone con gli altri
2. Per gli sviluppatori commerciali che possono sfruttare i risultati della valutazione nelle strategie di marketing

Perché valutare-2/2

Le caratteristiche da valutare le possiamo distinguere in due gruppi, dai confini abbastanza sfumati:

Linguistiche:

- **qualità:** della traduzione (testo lessicalmente e sintatticamente ben formato)
- **fedeltà**: adeguatezza semantica fra l'informazione di partenza e l'output di arrivo (richiede competenze bilingue)

Commerciali

- prezzo
- coverage: capacità del sistema di offrire un output coerente con il dominio di riferimento
- estendibilità: possibilità dell'utente finale di aggiungere al sistema nuove entrate, parole,regole.

Valutazioni umane vs valutazioni automatiche

Le traduzioni di sistemi di MT possono essere valutate da valutatori umani o da sistemi automatici.

Valutazione umana: facile da implementare, ma soggettiva, ed estremamente costosa in termini di tempo e denaro

Valutazione automatica: algoritmi complessi, ma oggettiva, economica, ripetibile, language independent. Ideale per gli sviluppatori che hanno bisogno di feedback di valutazione continui.

Valutazione umana

E' stata la prima forma di valutazione dei sistemi di MT, ma oggi rischia di essere un ostacolo e di rallentare la ricerca in questo campo, a causa degli alti costi di performance e della sua soggettività

Ancora oggi, però, ci sono misure molto interessanti che non possono fare a meno dell'uomo:

1. Tempo di lettura: *wpm (word per minute)= #parole/tempo di lettura*
2. Closed reading times: tempo di comprensione del testo
3. Tempo di correzione: inversamente proporzionale alla qualità della traduzione

Valutazione umana 2-2

Test:

1. **Cloze test**: test di leggibilità, alcune parole della traduzione vengono cancellate per vedere se il lettore ricostruisce il senso (con la parola esatta o con un sinonimo).
   Limite: dipende dalle competenze lessicali del valutatore.
2. **Test di adeguatezza**: tutte le informazioni della lingua sorgente sono espresse nella traduzione (non si perde informazione nel passaggio da un codice all'altro)
3. **Test di comprensione** : domande sul testo per valutare il livello di comprensibilità della traduzione

Machine Evaluation

Fattori positivi:

- oggettività
- ripetibilità
- economicità
- indipendenza dalla lingua (cambia la lingua dei testi che si passano al sistema ma non l'algoritmo)

I sistemi di valutazioni automatica confrontano la traduzione candidata con traduzioni referenti (gold standard) usando algoritmi diversi.

Algoritmo **BLEU** (Bilingual Evaluation Understudy)

Filosofia di base:

Una traduzione è tanto migliore quanto è più vicina alla traduzione di un traduttore professionista umano

Sviluppato dalla IBM, presso il Watson Research Center (2001)

Sviluppatori: **Papineni, Roukos, Ward e Zhu**

Obiettivo: dare risulti il più vicini possibili a quelli che si avrebbero con una valutazione umana.

Algoritmo BLEU - 2/7

Ingredienti base:

1. Un corpus di traduzioni referenti umane che faccia da gold standard
2. Una metrica numerica di "vicinanza"

La traduzione con score più alto è quella più "vicina" alla traduzione umana, e la vicinanza si calcola in termini di **precision**.

Dato un set di traduzioni con:

- t1, t2, t3: traduzioni candidate
- r1: traduzione referente

Algoritmo BLEU – 3/7

Sarà assegnato lo score più alto alla traduzione t(i) che condivide il maggior numero di parole con la traduzione referente (vicinanza vettoriale tra il vettore traduzione candidata e il vettore traduzione referente in uno spazio vettoriale n-dimensionale).

Per traduzione candidata viene calcolata la **precision,** come:

P= #n-grams condivisi/# n-grams candidato

n- gram: il pattern che viene confrontato tra le traduzioni con *n* numero dei tokens presi in considerazione (n=1, un token alla volta, n=2 due token per volta). Indica l'ampiezza della finestra.

Algoritmo BLEU- 4/7

Es: voglio andare a Londra

referente*: I want to go to London*

candidato 1:*I want to go London* **P(c1)= 5/5** *vince!!!!*

candidato 2: *wants going London* **P(c2)= 1/3**

Problema over-generazione:

candidato 3: go go go **P(c3)= 3/3**

Modified n-grams precision = 1/3

(si termina il conto della parola una volta raggiunto il numero max di volte che la stessa occorre nel referente)

Algoritmo BLEU - 5/7

Il confronto avviene per n-grams che possono essere 1-gram, 2-grams, 3-grams….

Problema:

- quando Bleu analizza 1-grams (finestra ad ampiezza 1), non tiene conto dell'ordine delle parole, quindi possono risultare molto vicine tra loro traduzioni semanticamente diverse.
  es. : *riso freddo - freddo riso*

- quando Bleu analizza 2-grams possono sembrare diverse traduzioni simili, perché non riconosce le perifrasi
  es. : *politics of south Wales- the south Wales politcs*

Algoritmo BLEU –6/7

Confronto fra la valutazione di BLEU e i risultati di human evaluation

Papineni confronta i risultati della valutazione automatica di Bleu con quelli di valutatori monolingui e bilingui umani.

Dati dell'esperimento:

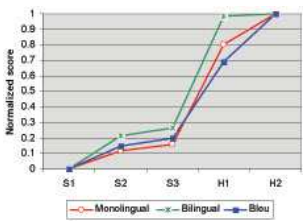
- 5 sets di traduzioni di 5 MT systems diversi (cinese-inglese)
- 1 gruppo di monolingui inglesi (per valutare la qualità)
- 1 gruppo di bilingui cino-americani (per valutare la fedeltà).

Detti S1, S2, S3 le traduzioni dei sistemi automatici e H1 e H2 le traduzioni umane, troviamo…

Algoritmo BLEU- 7/7

Confronto fra la valutazione di BLEU e i risultati di human evaluation

…che le valutazioni di Bleu sono molto vicine alle valutazioni umane.



Altri algoritmi di valutazione

NIST

Sviluppato dal National Institute of Standards and Technology.

Stessa filosofia di BLEU.

Qualche leggera modifica:

- Case sensitive
- **Tiene conto del peso di ogni n-grams nel calcolo della precision: le parole più importanti per la traduzione hanno un peso maggiore nel confronto con il gold standard.** L'importanza della parola si calcola sulla base della sua frequenza nella singola traduzione in rapporto alla rarità nell'intera collezione (tf-idf).

Altri algoritmi di valutazione

F-Measure

Sviluppata al Computer Science Department della New York University

Calcolo di precision e recall del massimo match, cioè della più lunga corrispondenza di n-grams.

Precision: rapporto fra numero n-grams condivisi tra candidato e referente e numero n-grams candidati

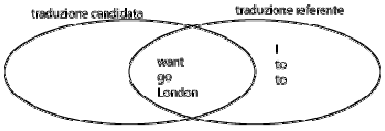
Recall: rapporto fra numero n-grams condivisi tra candidato e referente e numero n-grams referente

F-M= 2*(p*r)/p+r

Altri algoritmi di valutazione

Referente: I want to go to London

Candidato: want go London



Precision=1 **Recall=0.5**

F-M= 0.66

Altri algoritmi di valutazione

METEOR

Sviluppato alla Carnegie Mellon University di Pittsburgh.

Anche questo calcola le sovrapposizioni di n-grams, ma sfrutta due tipi di matching (in Perl)

1. Match esatto
2. Stemmed match

- Procede per passi: in una prima fase individua gli n-grams identici, in una seconda fase riprende le forme escluse inizialmente, le porta alla radice e cerca le corrispondenze per stemmed matching.
- Non si limita a contare gli n-grams perfettamente corrispondenti ma anche i sinonimi con l'aiuto di WordNet.

Standard in MT Evaluation

Non è facile confrontare vari sistemi di MT, perché sia gli output che le caratteristiche, hanno un valore intrinseco diverso a seconda dello scopo per cui sono stati pensati

Ma è importante creare un modello sistematico e universale delle proprietà da prendere in considerazione per una valutazione, in modo da avere un metro di paragone unico fra i vari sistemi

ISO-IEC 9126-1

ISO-IEC 9126-1 [1991]

L'International Organization of Standardization e la International Electrotechical Commission hanno elaborato nel 1991 delle linee guida per standardizzare i sistemi di valutazione dei software (ISO-IEC 9126-1).

Definiscono il concetto di:

Qualità*: misura della validità di un sistema in un particolare contesto dato. L'insieme delle caratteristiche e delle proprietà di un prodotto che portano a soddisfare i bisogni per cui quel prodotto era stato pensato.*

ISO-IEC 9126-1 [1991]

E i concetti di:

- **Requisiti di qualità interni:** legati alla performance della traduzione.
- **Requisiti di qualità esterni**: caratteristiche dell'applicazione.
- **Requisiti d'uso:** dipendenti dall'uso che si decide di fare dell'applicazione, come:
  1. sicurezza
  2. produttività
  3. efficacia
  4. soddisfazione

Tassonomia di Popescu-Belis,Hovy e King-1/3

All'interno del modello ISO-IEC, Popescu-Belis, King e Hovy sentono l'esigenza di creare una **tassonomia specifica per la valutazione dei sistemi di traduzione automatica**. Per fare ordine nel mare della valutazione e sintetizzare insieme le caratteristiche legate al contesto d'uso e i requisiti di qualità intrinseci del sistema, si sforzano di classificare le caratteristiche, gli attributi e le metriche per l'MT evaluation in questa tassonomia:

- **specifiche del contesto d'uso:**
  - caratteristiche del task di traduzione
  - caratteristiche legate all'uso
  - caratteristiche dell'input…

Tassonomia di Popescu Belis,Hovy King-2/3

- **caratteristiche di qualità**

-interne:
- modello di traduzione (statistico, rule based...),
- risorse linguistiche,
- utility del sistema sul testo: ricerca nel vocabolario, possibilità di editing;

- esterne:
- funzionalità,
- affidabilità,
- usabilità,
- efficienza (tempo di generazione della traduzione, tempo di correzione, tempo di lettura...)
- portabilità,
- costi,
- facilità di mantenimento.

Tassonomia di Popescu Belis,Hovy King-3/3

Ogni attributo della tassonomia deve essere quantificato, e inserito con coordinate precise in uno “spazio di qualità”. Ogni foglia di questa tassonomia diventa la proprietà ultima di valutazione, con un punteggio specifico, per una comparazione di sistemi diversi.

La scala di valore che si utilizza deve essere standard e ben correlata con i giudizi umani

NIST Evaluation 2005- 1/3

Workshop organizzato dal NIST per fare il punto sui risultati raggiunti nel campo della ricerca dei sistemi di traduzione automatica.

I partecipanti dovevano testare i propri algoritmi di MT da arabo e cinese verso inglese.

Dati:

- due sets di frasi per lingua sorgente (due per arabo e due per cinese).
- quattro sets di traduzioni referenti per ogni set sorgente

NIST Evaluation 2005- 2/3

Ogni partecipante ha generato un output con il proprio sistema e lo ha inviato ai valutatori.

Il sistema di valutazione automatico, **GALE**, ha preso in input i risultati e li ha confrontati con i sets referenti usando l’algoritmo **BLEU**.

Ogni traduzione candidata ha ottenuto un punteggio in un rank fra 0-1.

Diversi risultati per l’arabo e per il cinese (e diversi a seconda del set di dati)…..

NIST Evaluation 2005- 3/3

Ecco i sistemi che si sono aggiudicati le prime tre posizioni:

Arabic-to-English Task	
Site	**Bleu score**
Google	0.51
ISI	0.46
IBM	0.468

Chinese-to-English Task	
Site	**Bleu score**
Google	0.35
ISI	0.3007
UMD	0.3000

Conclusioni

Il dibattito sul tema della valutazione è ancora aperto, per definire come debba essere valutato un MT system:

- attraverso caratteristiche intrinseche di qualità dell'output. Se sì, quanto sono realistici i metodi statistici di valutazione automatica nel confrontare l'output con le traduzioni umane?
- Attraverso qualità estrinseche dell'uso dell'applicazione (portabilità, usabilità, efficienza…)
- Attraverso una tassonomia che le racchiuda entrambe.

Bibliografia

- The NIST Machine Translation Plan (MT 2005)
- An Introduction to MT Evaluation, Hovy, Popescu-Belis, King, Marina del Rey, Ginevra
- BLEU: A method for Automatic Evaluation of Machine Translation, Papineni, Roukos, Zhu, atti del 40th meeting di ACL, Association of Computational Linguistic, Philadelphia, Luglio 2002, pp. 311-318
- Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Workshop at the Annual Meeting of the Association of Computational Linguistics, Ann Arbor, Michigan, 29 giugno 2005.
- Traduzione automatica: ricerca, sviluppi e prospettive, Valentina Felici, ForumTAL, gennaio 2006
- NIST 2005, Machine Transaltion Evaluation Official Result, agosto 2005,
- Al mondiale delle traduzioni gli italiani sono ai primi posti, Di Miceli, articolo novembre 2005, RepubblicaOnLine.

Sitografia

- Sito della University of Southern California's Information Sciences Institute (ISI): http://www.isi.edu
- Sito del Association of Computational Linguistic: http://www.aclweb.org
- Sito del NIST: http://www.nist.org
- University of Maquarie, Sydeny, Australia: http://www.ics.mq.ed.au. (visitato 1/6/2006)
- Forum di discussione sul Trattamento Automatico della Lingua Italiana: http://forumtal.fbu.it