Modulo B

Valutare i sistemi automatici di interazione linguistica

5. Valutazione con BLEU

15 marzo 2017

Linguistica italiana II
Mirko Tavosanis
A. a. 2016-2017

BLEU

- Punto di partenza intuitivo: più una traduzione automatica assomiglia a una traduzione umana, meglio è
- Si potrebbe usare il WER, ma ci sono dei limiti
- Un limite è dato dal fatto che il WER favorisce molto le parole più probabili, ma alcune parole sono molto più probabili di altre in qualunque tipo di testo
  - Traduzione di riferimento: *Il gatto e il cane si odiano*
  - Traduzione a, 3 sostituzioni: *il il e il il si il*
  - Traduzione b, 3 omissioni o aggiunte: *il gatto il cane e si si odiano*
- BLEU permette di limitare questi fenomeni e di approssimarsi ai giudizi di valore forniti da esseri umani esaminando gli n-grammi

Calcolo

- Si usa una misura di precisione modificata p_n sugli n-grammi
  - Precisione: quante parole della traduzione automatica si ritrovano nella traduzione di riferimento? Se ripetiamo continuamente ”il”, la risposta è: tutte
- Come punto di partenza per la modifica, si calcola il numero di occorrenze dell’n-gramma nella traduzione di riferimento
- Questo numero di occorrenze viene usato per mettere un limite (*clipping*) alle occorrenze prodotte dalla traduzione automatica: non ci devono essere più di due ”il” $Count_{clip}(n\text{-}gram)$
- Il calcolo degli n-grammi nella traduzione da valutare viene fatto frase per frase, con il limite, e poi viene eseguita la sommatoria
- La sommatoria degli n-grammi ”limitati” viene divisa per la sommatoria degli n-grammi totali

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

- La misura di precisione viene poi modificata con una penalità per la brevità eccessiva (BP)

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \qquad \mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Valutazione

1. Preparazione di un corpus (in questo caso, un solo articolo di giornale, estratto dal *New York Times*)
2. Traduzione umana
3. Traduzione automatica
4. Valutazione umana delle traduzioni automatiche
5. Confronto dei risultati con mteval (BLEU-4)
6. Individuazione dei problemi delle traduzioni

Dettaglio: scale ARPA

- «Adeguatezza» (*adequacy*): quanto l'informazione si è conservata, indipendentemente dall'espressione?
  - Il valutatore riceve frasi isolate e ha a fronte il testo originale
- «Fluenza» (*fluency*): quanto è valida l'espressione nella lingua di destinazione, indipendentemente dal contenuto?
  - Il valutatore riceve frasi isolate *senza* il testo originale
- «Informatività» (*informativeness*): misura la capacità di trasmettere contenuto operativo, poi misurato attraverso test a scelta multipla
  - altamente correlata all'adeguatezza, e quindi abbandonata
- Per adeguatezza e fluenza, la valutazione viene espressa con un punteggio da 1 a 5 e la media della valutazione viene convertita in scala 0-1
- Fonte: White, J. (1995) "Approaches to Black Box MT Evaluation", in: *Proceedings of MT Summit V*

Punteggi BLEU con mteval

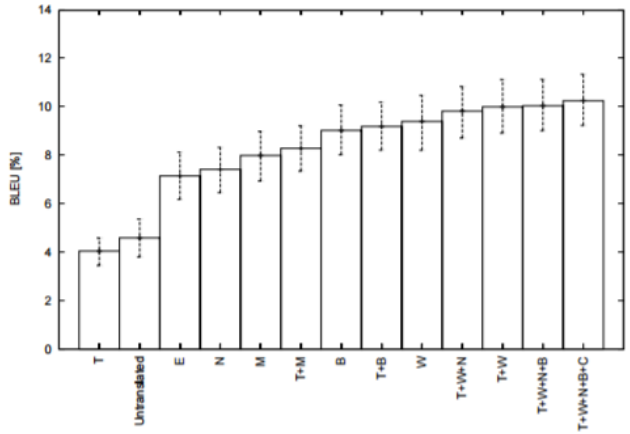
- Traduzione di Google Traduttore (sito web):
  NIST score = 4.1089 BLEU score = 0.2662
- Traduzione di Microsoft Translator (app):
  NIST score = 4.1295 BLEU score = 0.2415

- Usando come riferimento la traduzione di Google, quella di Microsoft ottiene:
  NIST score = 6.2760 BLEU score = 0.5783
- Evidentemente, i due sistemi parlano un po' la stessa lingua!

- Usando un testo identico si ottiene BLEU = 1, ma modificando una parola si ottiene
  NIST score = 7.9324 BLEU score = 0.9925

I punteggi hanno senso?

- I numeri riportati in bibliografia sono di regola più alti
- I sistemi automatici valutati da Papinieni e altri (2002) avevano punteggi inferiori a 0,2, quelli umani arrivavano a 0,6
- Non mi è ancora chiaro quale credibilità linguistica abbiano questi conti



- Soprattutto, non esistono soglie di usabilità che dicano che una traduzione è utile al di sopra di un certo livello

Quali sono i problemi delle traduzioni?

- Tutti da studiare!
- Colpisce l'assenza di controlli sintattici interni alla frase
- Per esempio, controlli che, dato un verbo di modo finito, verificano se c'è un soggetto plausibile:

Google: Due settimane fa, un dipendente di Amazon **è entrato** un insieme errato di comandi su un computer, accidentalmente **buttare giù** una buona parte dei server in un data center di Amazon in Virginia.

Microsoft Translator: Due settimane fa, un dipendente di Amazon **inserito** un set di comandi non corretto su un computer, accidentalmente **buttando giù** una buona fetta dei server presso un centro di dati di Amazon in Virginia.

Ipotesi ovvie sui risultati

- La correttezza dei risultati varierà in base al genere testuale
- I testi rappresentati nei corpora di riferimento e addestramento per la traduzione automatica saranno tradotti meglio
- I testi più rappresentati in questi corpora saranno quelli formali (i testi non formali non vengono neanche tradotti)
- Sull'entità delle variazioni non riesco a trovare dati pubblicati: l'interesse è andato finora nel miglioramento dei sistemi di traduzione su corpora limitati di riferimento