# Modulo B

# Valutare i sistemi automatici di interazione linguistica

4. BLEU

9 marzo 2017

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

# Valutazioni automatiche

- Le valutazioni umane sono sempre costose
- Alcune valutazioni automatiche hanno mostrato una buona correlazione con valutazioni umane
- L'uso di valutazioni automatiche non serve a migliorare la **qualità** della valutazione: rende la valutazione più efficiente dal punto di vista **economico**

# Ritraduzione

- Se il sistema lo consente, si può tradurre un testo in una lingua e poi ritradurlo da lì nell'originale
- Si può immaginare che una traduzione perfetta porti a ricostruire l'originale (e che una traduzione imperfetta sia misurabile usando semplicemente il WER)
- Nella pratica, tuttavia, le traduzioni sono sempre piuttosto lontane
- Inoltre, a volte la ricomposizione può essere indizio di una traduzione meccanica, inutile nella lingua di arrivo
- Il metodo della ritraduzione oggi non viene usato, anche se può essere utile per farsi un'idea rapida delle prestazioni di un sistema in una lingua che non si conosce

# Esempio

«Prima di tutto, la simpatia e la semplicità del disegno, che rendono immediatamente riconoscibili tutti i personaggi della serie»

Vediamolo con Google Traduttore

- Passando dall'inglese: Prima di tutto, la cordialità e la semplicità del **design**, che rendono immediatamente **riconoscibile** tutti i personaggi della serie (2 variazioni)
- Passando dal vietnamita: Prima di tutto, la cordialità e la semplicità del disegno, che rende immediatamente **riconoscere** tutti i personaggi **del film** (3 variazioni)

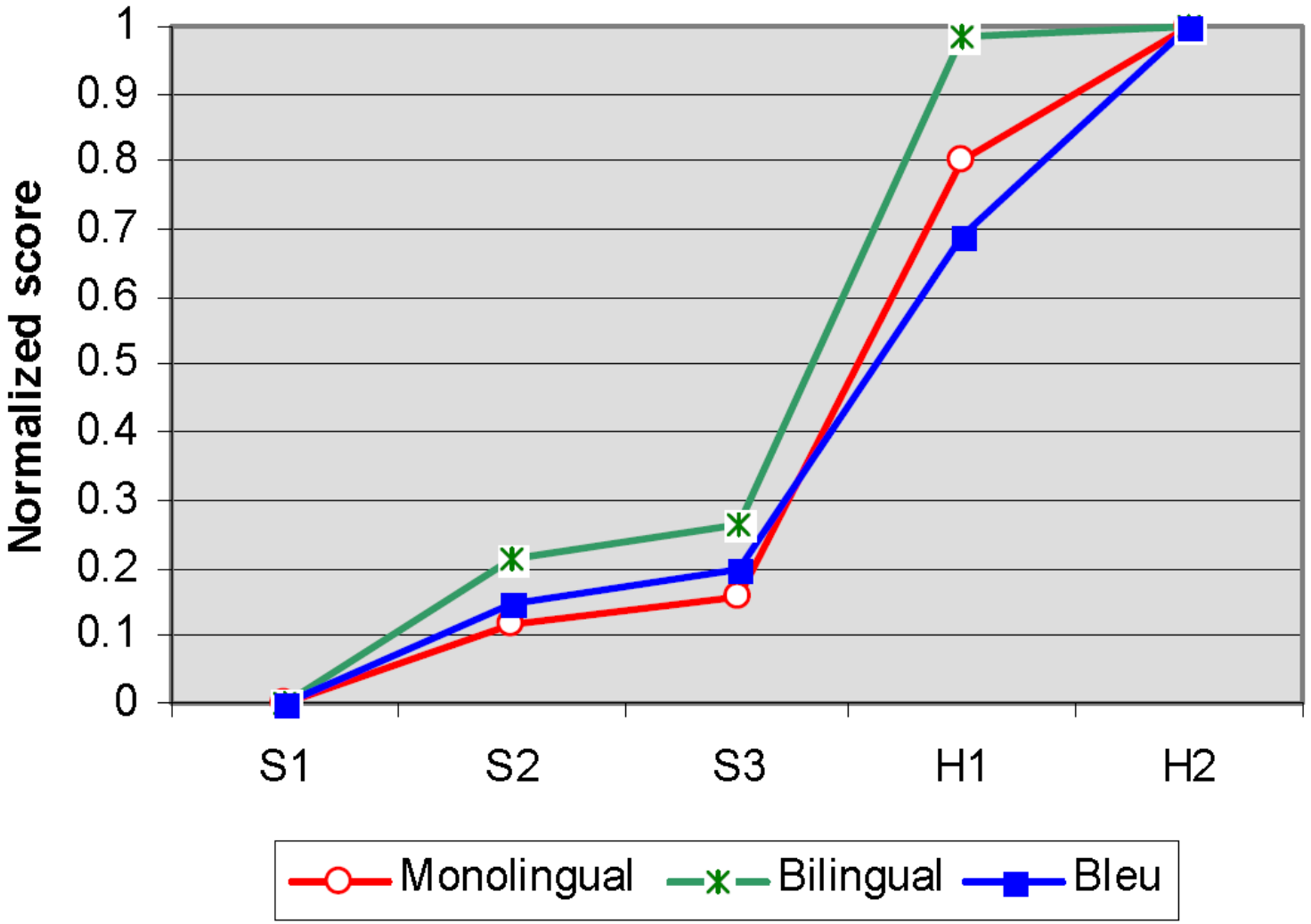
# BLEU

- «Bilingual Evaluation Understudy»: come il WER nel suo settore, è lo standard più diffuso
- Non sembra che esistano metriche in grado di avere una correlazione migliore con le valutazioni umane (Graham e Baldwin 2014)
- Alla base: un algoritmo per il calcolo della precisione – cioè la frazione di «parole» generate dal traduttore che si ritrovano nel corpus di riferimento
  - Nel caso di BLEU, in pratica, non si usano parole singole ma n-grammi di lunghezza 4 (in pratica, sequenze di 4 parole)
- Il confronto viene fatto poi su una traduzione di riferimento fatta di brevi sezioni (frasi) del testo originale
- Fonte: Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation*. In: *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, pp. 311–318.

# Presupposto di BLEU

- Anche se le traduzioni possibili sono diverse, una buona traduzione avrà molti punti di contatto con altre buone traduzioni
- Maggiori sono i punti di contatto con le traduzioni di riferimento, migliore sarà la traduzione
  - **Testo originale:** The murder defendant, James Bates, agreed late Monday to allow Amazon to forward his Echo's data to Arkansas prosecutors.
  - **Traduzione umana di riferimento: L'accusato, James Bates,** nella serata di lunedì ha dato ad Amazon il permesso di consegnare i dati del suo Echo ai pubblici ministeri dell'Arkansas.
  - **Traduzione Microsoft Translator:** L'**accusato** di omicidio, **James Bates**, concordato late Lunedi consentire **Amazon** inoltrare **i dati** di sua Echo ai **pubblici ministeri Arkansas**.
- Naturalmente, è una soluzione imperfetta perché ci sono molti modi per tradurre la stessa frase
- Anche per questo, i risultati di BLEU hanno senso su un corpus e non su singole frasi

# Correlazione



# Uso di mteval per BLEU

- Anche in questo caso usiamo un pacchetto messo a disposizione dal NIST: mteval https://www.nist.gov/itl/iad/mig/tools
- Il programma richiede che sul computer sia installato Perl
  - nel mio caso ho usato ActivePerl 5 su Windows 10
  - oltre all'installazione standard è richiesta l'installazione del pacchetto XML-Twig (potete gestire l'installazione usando il sistema di gestione dei pacchetti di ActivePerl, PerlPackageManager)
- Il lavoro si svolge usando tre file:
  - ref.xml – traduzione di riferimento, realizzata da un essere umano
  - src.xml – testo nella lingua di partenza
  - tst.xml – traduzione automatica
- Comando: mteval-v13a.pl -r ref.xml -s src.xml -t tst.xml

# Valori

- Il prodotto del calcolo è un valore compreso tra 0 e 1
- Il valore massimo è 1 e indica la perfetta corrispondenza con il prodotto della traduzione umana
- Un risultato possibile
  NIST score = 8.3006  BLEU score = 0.4929 for system "sample_system”
- NIST è quasi la stessa cosa, ma assegna un peso maggiore agli n-grammi più rari

# Codifica

- <?xml version="1.0" encoding="UTF-8"?>
  <!DOCTYPE mteval SYSTEM
  "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
  **<mteval>**
  <**refset** setid="example_set" srclang="Arabic" trglang="English"
  refid="ref1">
  <**doc** docid="doc1" genre="nw">

  <**seg** id="1"> Roed-Larsen: Middle East a "Powder Keg with Lit Fuse" </seg>

  <seg id="2"> Oslo 2-6 (AFP) - Terje Roed-Larsen, the former United Nations Middle East envoy, considered the situation in the region as having never been as dangerous as it is today and compared the region to a "powder keg with a lit fuse". </seg>

- Le singole frasi sono incluse in un <seg> all'interno di 

# Operativo

Possibile soggetto di relazione: valutare l'output di una traduzione automatica…

- scegliendo un testo
- traducendolo in italiano
  - a mano
  - con Google Translate
  - con Microsoft Translator
- calcolando il punteggio BLEU
- commentando gli aspetti linguistici più vistosi

Il lavoro può essere gestito in modo simile a quanto fatto per il modulo A