Valutazione umana della traduzione automatica

Mirko Tavosanis

I migliori servizi disponibili oggi

- Google: sicuramente il più noto
  https://translate.google.it/?hl=it
- Microsoft Translator: integrato in prodotti come Office, ma con prestazioni inferiori rispetto a Google
  https://translator.microsoft.com/
- DeepL: apparentemente il migliore!
  https://www.deepl.com/translator

Google e le reti neurali

- Il 15 novembre 2016 Google ha annunciato il passaggio di una parte dei servizi di Google Traduttore a un sistema basato sull'apprendimento automatico: Google Neural Machine Translation (GNMT)
- Il sistema GNMT traduce "whole sentences at a time, rather than just piece by piece. It uses this broader context to help it figure out the most relevant translation, which it then rearranges and adjusts to be more like a human speaking with proper grammar"
- Il sistema GNMT è stato reso disponibile per un numero sempre più ampio di lingue: inglese, cinese, francese, tedesco, giapponese, coreano, portoghese, spagnolo e turco… e poi l'italiano, dall'aprile 2017
- Il sistema non usa una lingua ponte ma traduce direttamente da una lingua all'altra

Fonte: Barak Turovsky, *Found in translation: More accurate, fluent sentences in Google Translate*, Google Blog, 15 novembre 2016
https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

Più in dettaglio

- Partendo dal 2016, i principali sistemi di traduzione online hanno abbandonato la SMT e sono passati a sistemi basati su reti neurali (NMT)
- Il passaggio è stato rapidissimo e oggi **tutti** i principali sistemi per l'italiano sembrano basati su reti neurali

    - Apparentemente, anche nella traduzione Google usa le RNN – Recurrent Neural Networks
    - Microsoft Translator dichiara di usare«deep neural networks» in generale
    - DeepL usa le reti neurali convoluzionali (*Convolutional neural networks*, *CNN* o *ConvNet*)

Importanza delle valutazioni

- “One very important aspect of research in computational linguistics, as opposed to many other areas in humanities, is its ability to be *evaluated*…” (Schreibman e altri, *A Companion to Digital Humanities*, p. 84)
- Tuttavia, rispetto agli altri compiti della linguistica computazionale, “Machine translation (or any system that produces free, plain text) is much more difficult (and expensive) to evaluate” (*idem*, p. 85)
- “One of the major issues with translation in general, and speech-to-speech translation in particular, is the absence of a clear criterion for evaluation” (Pieraccini 2012, p. 275)

Sistemi formalizzati

- Per le traduzioni ne sono disponibili diversi
- Vediamo una rassegna generale, per approfondire poi BLEU
- Il punto chiave: un testo deve servire a esseri umani e la qualità della traduzione deve essere quindi misurata in base **all'utilità per gli esseri umani**
- Seguiremo una voce (molto utile) di Wikipedia in lingua inglese:

  https://en.wikipedia.org/wiki/Evaluation_of_machine_translation

Valutazione umana

- La componente umana è ineliminabile: anche BLEU, lo standard di riferimento, richiede come confronto parti di traduzione realizzate da esseri umani
- Alternative:
  - Vedere se i lettori riescono a capire il testo tradotto e a **usarlo** per fare qualche cosa – anche solo rispondere a domande a scelta multipla (valutazione umana **indiretta**)
  - Far valutare direttamente da esseri umani il risultato della traduzione automatica (valutazione umana **diretta**)
  - Usare come riferimento una traduzione realizzata da esseri umani e assegnare con un calcolo un punteggio alla traduzione automatica (valutazione **semiautomatica**)
- Tutto ciò che viene fatto da esseri umani ha un costo molto alto… ma i fattori soggettivi qui sono più importanti che nella trascrizione del parlato

Valutazione umana delle traduzioni umane

- Si è sempre fatta, ma in modo poco strutturato
- Ancora oggi si usano criteri di senso comune nati in età prescientifica, spesso per la traduzione letteraria
  - «brutte e fedeli o belle e infedeli?»
  - «devo fare una traduzione letterale o libera?»
- Nel Novecento si è molto dibattuto sulla teoria della traduzione, e il consenso è che la traduzione sia sempre un'opera di reinterpretazione, non un'attività meccanica
- In fin dei conti, di regola un testo:
  - è espressione parziale di un significato più ampio (mentale e non linguistico)
  - lega strettamente forma e contenuto

Traduzione

- La traduzione da una lingua all'altra («traduzione interlinguistica») è un caso particolare di traduzione
- Un testo può essere
  - riscritto **nella stessa lingua**, ma con parole diverse (traduzione intralinguistica)
  - riscritto **in un'altra lingua** (traduzione interlinguistica)
  - riscritto **in un altro sistema di segni** (traduzione intersemiotica)
- La valutazione della traduzione da una lingua all'altra è un caso particolare di un caso particolare di rielaborazione
- Come riferimento sintetico: Raffaella Bertazzoli, *La traduzione: teorie e metodi*, seconda edizione, Roma, Carocci, 2015
- Nella prossima diapositiva: un esempio di tabella di valutazione proposto da Bruno Osimo

SIGLE	SPIEGAZIONE	TIPO DI CAMBIAMENTO	ESEMPI
C	**Cadenza,** punteggiatura, rima, metrica, capoversi	è stato alterato uno di questi elementi, modificando il ritmo del testo	
CAC	**CACofonia**	allitterazioni, assonanze involontarie	le ostiche ostriche
D	**Deittici,** rimandi interpersonali, punto di vista	migliore/peggiore riproduzione del punto di vista del narratore o del personaggio, ideologia personale	questo→quello ora→allora qui→là
DEN-A	**Aggiunte**	una singola parola è aggiunta	il gatto→il gatto bianco
DEN-CS	**Calchi Semantici** e **Sintattici**	calco di parola che determina senso diverso e incomprensibile	il tuo comportamento è morbido
DEN-M	cambiamento radicale di senso **Mistranslation**	l'errore è tale da compromettere il senso generale della frase	the triumph of spirit over circumstance→il trionfo della spiritualità sul caso
DEN- MOD	**MODulazione:** specificazione-generalizzazione, parole-termini, ambiguazione-disambiguazione	una parola è resa più specifica o più generica. un termine è diventato parola comune o viceversa. ridondanza semantica. modifica del livello di ambiguità di un'espressione in entrambi i sensi	non mi dà fastidio, lo sopporto
DEN- OM	**OMissioni**	una singola parola è omessa	il gatto bianco→il gatto
DEN-W	errori lessicali riguardanti una sola parola **Word**	una singola parola è fraintesa in modo netto (altro campo semantico)	il gatto→il cane
DT	destinatario – **Dominante** del **Testo –** leggibilità	migliore/peggiore coglimento del lettore modello e della dominante del testo	
E	**Enciclopedia** - precisione fattuale – conoscenza del mondo	la dotazione enciclopedica della traduttrice è insufficiente a colmare l'implicito culturale	blue helmets→elmetti celesti
ENF	**ENFasi,** ordine delle parole	dislocazioni, frase scisse, ordine anomalo delle parole che determina diversa accentuazione della frase	
G-S	errori **Grammaticali** e **Sintattici**	errori di grammatica o sintassi nella cultura ricevente	sebbene è; inerente il; in stazione
I	rimandi **Intertestuali,** realia	migliore/peggiore coglimento dei rimandi esterni ad altri testi o altre culture	
INTRA	uso di **SINonimi,** ripetizioni, rimandi intratestuali	sinonimizzazione e desinonimizzazione. coglimento di rimandi interni da un capo all'altro del testo. ridondanza lessicale	domandare una domanda
L	**Logica**	la logica della traduttrice è insufficiente a colmare l'implicito culturale	sapeva che non sarebbe sopravvissuta alla propria morte
O	**Ortografia**	errori d'ortografia nella cultura ricevente	un po'; qual è; ti dò
P	**Presentazione** - forma grafica – layout – impaginazione	migliore/peggiore riproduzione degli aspetti grafici rispetto alle norme suggerite dal committente	
R	**Registro,** tipo di testo	uso di parole di registro uguale a/diverso da quello desiderato. migliore/peggiore	
S	**Stile** complessivo dell'autore	migliore/peggiore rendimento dello stile	
U	**Uso:** locuzioni, collocazioni, calchi non semanticamente sbagliati, resa inefficace	una singola parola, sebbene non semanticamente sbagliata, è collocata in modo involontariamente marcato	l'ho mandato in quella città (anziché "a quel paese");SEP è supposto saperlo

Valutazione umana diretta

- Per alcuni tipi di traduzione – per esempio, tecniche – alcuni aspetti della tabella possono essere tralasciati
- Le valutazioni umane rimangono ineliminabilmente soggettive, ma la variabilità può essere limitata usando tecniche di buon senso; per esempio:
  - Ai valutatori si può fornire una serie di istruzioni dettagliate, tipo scala Mercalli (soprattutto se si varia da «perfetto» a «incomprensibile»)
  - Si possono usare più valutatori e deve essere calcolata una media dei loro giudizi
- Esempi:
  - I programmi ALPAC (1964)
  - Le valutazioni ARPA (dal 1991)

Dettaglio: valutazioni ARPA

- Negli anni Novanta l'ARPA ha provato diversi sistemi di valutazione per la traduzione automatica
- Molto valida (ma costosa): la valutazione diretta dei testi da parte di esperti che si servono dei criteri usati per valutare le traduzioni fatte da esseri umani
- Simile nei risultati, ma più economica: la valutazione di **segmenti** di testo da parte di parlanti L1
  - La valutazione viene fatta usando due o tre scale diverse e ha mostrato **una buona correlazione** con la valutazione dei testi interi da parte di esperti
- Fonte: White, J. (1995) "Approaches to Black Box MT Evaluation", in: *Proceedings of MT Summit V*

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*

Valutazioni automatiche

- Le valutazioni umane sono sempre costose
- Alcune valutazioni automatiche (o meglio, semiautomatiche) hanno mostrato una buona **correlazione** con valutazioni umane
- L'uso di valutazioni automatiche **non** serve a migliorare la qualità della valutazione: rende solo 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 Word Error Rate, la differenza tra il testo voluto e quello giusto)
- 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
- Passando dal vietnamita: Prima di tutto, la cordialità e la semplicità del disegno, che rende immediatamente **riconoscere** tutti i personaggi **del film**

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.

Perché BLEU

- Punto di partenza intuitivo: più una traduzione automatica **assomiglia** a una buona traduzione umana, meglio è
- Si potrebbe usare il semplice Word Error Rate, 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

Presupposto di BLEU

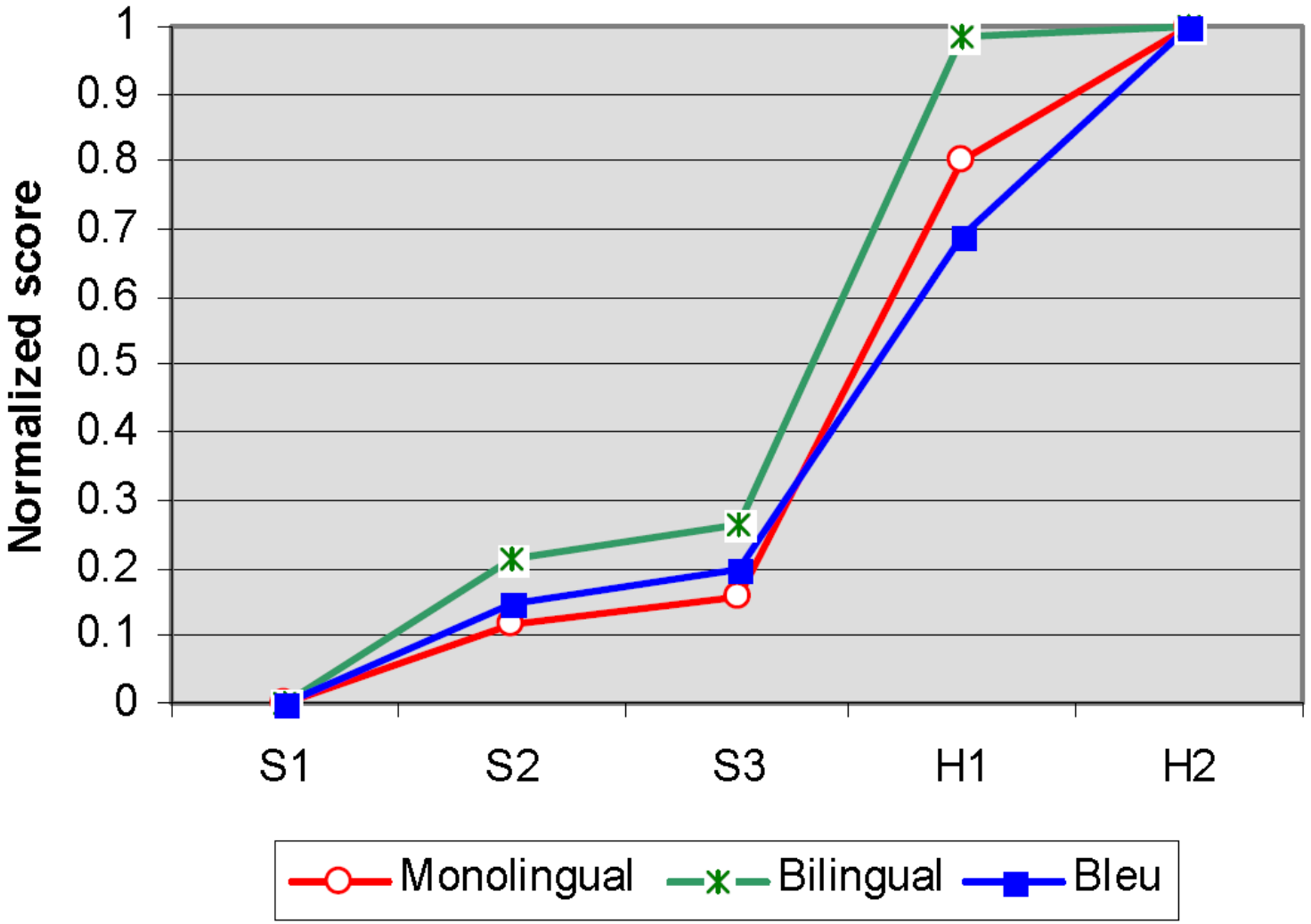
- Anche se le traduzioni possibili sono diverse, una buona traduzione avrà molti punti di contatto con altre buone traduzioni
- Maggiori sono i punti di contatto con le traduzioni di riferimento, migliore sarà la traduzione
  - **Testo originale:** The murder defendant, James Bates, agreed late Monday 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

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)

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

Correlazione



Misurazione della qualità

- Si è visto rapidamente che i sistemi NMT ottenevano risultati migliori di quelli SMT
- Un contributo di Google: Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv preprint arXiv:1609.08144
- L'esperimento era basato su due corpora WMT'14:
  English-to-French (WMT En→Fr)
  English-to-German (WMT En→De)
- I risultati arrivavano vicini alle prestazioni di un traduttore umano, anche se gli stessi ricercatori ammettevano che il campione di riferimento non era molto adatto a un esame di dettaglio (e comunque, il punto del loro lavoro era un altro)

Miglioramento rispetto a SMT

Wu e altri, 2016: valutazione umana delle traduzioni su scala 1 (incomprensibile) – 6 (perfetto):

Table 10: Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

Sottostima

- Nel corso del 2018 la comunità dei ricercatori si è accorta che per i sistemi a reti neurali le valutazioni fatte con BLEU non erano più correlate alla valutazione umana
- Shterionov e altri (2018) hanno eseguito una misurazione dettagliata delle differenze
- Ipotesi di partenza: che tutti i sistemi basati su n-grammi, incluso BLEU, sottovalutino la qualità dei risultati della traduzione automatica
- Riferimento: Shterionov, Dimitar, et al. “Human versus automatic quality evaluation of NMT and PBSMT.” *Machine Translation* 32.3 (2018): 217-235 https://www.researchgate.net/publication/325027945_Human_versus_automatic_quality_evaluation_of_NMT_and_PBSMT

Sistemi basati su n-grammi

- Abbiamo visto BLEU
- Esistono anche altri sistemi, meno usati
- Tra questi, Shterionov e altri hanno provato:
  - F-measure
  - TER
- I risultati comunque sono stati simili a quelli di BLEU, o peggiori

F-measure

- Il sistema matematicamente più semplice
- Fa la media armonica di precisione e richiamo
- Il livello usato è quello delle singole parole (1-grammi)
- Non viene quindi preso in considerazione l'ordine delle parole all'interno del testo
- Non a caso, è stato superato da BLEU

$$F_1 = \frac{2}{\frac{1}{\mathrm{r}} + \frac{1}{\mathrm{p}}} = 2 \cdot \frac{\mathrm{p} \cdot \mathrm{r}}{\mathrm{p} + \mathrm{r}}$$

TER

- Translation Edit Rate
- Concettualmente simile al WER:
  - opera a livello di parola
  - data una traduzione umana di riferimento, calcola il numero di trasformazioni necessario per passare dalla traduzione automatica alla traduzione umana
- A differenza del WER, calcola anche lo spostamento di blocchi di parole (problema frequente nel caso delle traduzioni)
- Fornisce risultati interessanti… e una traduzione corretta!

Motivazioni per la sottostima

- In sostanza, le traduzioni fatte con NMT sono più libere di quelle fatte con SMT, e in particolare con il sistema più usato per le traduzioni automatiche: PBSMT (Phrase-Based Statistical Machine Translation)
- Alla base c'è proprio il modo in cui funzionano questi meccanismi

PBSMT

- PBSMT opera a livello di segmenti e blocchi limitati…
- … però, attenzione! Questo non significa «a livello di frase» nel senso in uso nella linguistica italiana
- «Phrase» in inglese ha spesso il significato di «sintagma»: può essere una frase, ma anche un suo componente
- In questo caso addirittura «phrase» indica un blocco di parole di lunghezza determinata, indipendentemente dai confini dei sintagmi e dalla lunghezza di frase
- Il sistema di traduzione cerca corrispondenze in n-grammi, e questo apparentemente produce risultati che ottengono un buon punteggio con i sistemi di valutazione basati su n-grammi

NMT

- Il sistema prende una **frase** (in senso italiano! *Sentence* in inglese) e la traduce nel suo complesso
- Alla base c'è la sequenza completa dei token della frase di partenza, non una sua sezione
- Questo permette traduzioni più libere, indipendentemente da qualunque considerazione di n-grammi
  - io credo che renda anche molto difficili gli interventi che richiederebbero di cambiare i confini tra le frasi
- Inoltre, a volte i sistemi non usano come blocco fisso le parole, ma anche unità più piccole delle parole, e perfino i singoli caratteri

Esiti possibili

Come per le traduzioni umane, con NMT si possono avere traduzioni corrette ma che in pratica non condividono nessuna delle parole del testo di riferimento

Example 1 An NMT translation with low BLEU score that is better (judged by human evaluators) than a PBSMT one with a higher BLEU score.

Source (EN): All dossiers must be individually analysed by the ministry responsible for the economy and scientific policy.

Reference (DE): Jeder Antrag wird von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik individuell geprüft.

PBSMT: Alle Unterlagen müssen einzeln analysiert werden von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik. **BLEU:** 55.82%

NMT: Alle Unterlagen müssen von dem für die Volkswirtschaft und die wissenschaftliche Politik zuständigen Ministerium einzeln analysiert werden. **BLEU:** 3.21% △

Quindi?

- L’unica cosa da fare è usare come punto di riferimento le valutazioni umane
- In fin dei conti, il metro per misurare la qualità delle traduzioni automatiche è questo – e i vari sistemi semiautomatici sono utili solo finché hanno buona correlazione con questo

Nella pratica

Nel caso di Shterionov e altri 2018 è stata fatta una valutazione (descritta da p. 10 in poi); i componenti sono stati:

- 5 coppie di lingue, cioè traduzioni dall'inglese in…
  - Tedesco
  - Spagnolo
  - Italiano
  - Giapponese
  - Cinese mandarino
- 3 valutatori umani per ogni lingua (totale: 15 valutatori): «**Translation studies students** recruited from five different universities in Europe, holding certificates of English proficiency or attending courses taught in English» (p. 11), tutti **madrelingua** nella lingua target

Oggetto della prova

- I testi da tradurre sono stati presi da raccolte della Commissione Europea e da Opus, per un totale di 350 milioni di parole
- I domini selezionati sono stati:
  - Tecnico
  - Medico
  - Legale
- Attenzione! Le traduzioni non sono state fatte usando strumenti commerciali, ma strumenti sviluppati dal gruppo di ricerca (sia per PBSMT sia per NMT) attraverso la piattaforma cloud KantanMT: https://www.kantanmt.com/

Esecuzione della valutazione

- La prova è stata gestita usando KantanLQR, un sistema online per fare questo tipo di test
- Per ogni lingua sono stati isolati 200 frasi, da dividere tra i valutatori
- In pratica, ogni valutatore, da valutare consecutivamente
  - senza pause (a occhio, ci sarà voluto un paio d'ore)
  - senza scambi di pareri con gli altri valutatori
- Il confronto è stato fatto con il sistema A-B:
  - Il valutatore vede l'originale inglese
  - Vede allo stesso tempo anche due traduzioni, una fatta con PBSMT e una fatta con NMT (ma l'ordine con cui le due traduzioni sono presentate su schermo è casuale)

Giudizio semplice

- Il valutatore doveva decidere tra tre possibilità, indicando il rapporto di qualità tra le due traduzioni:
  - Uguale qualità
  - PBSMT > NMT
  - NMT > PBSMT

Risultati test A-B

Per la maggior parte delle coppie, i valutatori hanno preferito NMT (tabella a p. 12):

	Same				PBSMT				NMT				κ
Lang. Pair	H1	H2	H3	Avg.	H1	H2	H3	Avg.	H1	H2	H3	Avg.	coef.
EN–DE	19%	14%	6%	13%	27%	35%	40%	34%	54%	51%	54%	53%	33.14%
EN–ES	12%	10%	7%	10%	28%	26%	31%	28%	60%	64%	62%	62%	36.82%
EN–IT	25%	29%	19%	24%	19%	14%	25%	19%	56%	57%	56%	56%	54.69%
EN–JA	21%	14%	27%	21%	19%	28%	16%	21%	60%	58%	57%	58%	62.65%
EN–ZH_CN	41%	34%	37%	37%	20%	26%	25%	24%	39%	40%	38%	39%	62.68%

Table 3 Side-by-side evaluation of PBSMT and NMT output performed by human evaluators. H1, H2 and H3 denote the different human evaluators.

Ovviamente, l'accordo tra valutatori umani non è perfetto; un'indicazione di quanto i valutatori siano d'accordo è data dal coefficiente k

Coefficiente *k*

- Il coefficiente *k* (o «*k* di Cohen») indica quanto l'accordo presente in una classificazione fatta da più giudici è più forte di quello che si ottiene distribuendo a caso i giudizi
  - con valori inferiori a 0, la distribuzione dei giudizi è **casuale**
  - Con valori compresi tra 0 e 40%, l'accordo è **scarso**
  - Con valori compresi tra 40 e 60%, l'accordo è **discreto**
  - Con valori compresi tra 60 e 80%, l'accordo è **buono**
  - Con valori compresi tra 80 e 100%, l'accordo è **ottimo**
- Nel caso che abbiamo visto, l'accordo dei valutatori va da **scarso** a **buono**, a seconda delle coppie di lingue (per l'italiano: **discreto**)
- Ovviamente, questo è un punto debole

Una valutazione alternativa: la produttività

- A 15 traduttori (con caratteristiche simili a quelle già viste) è stato chiesto di intervenire sull'output dei sistemi di traduzione, incluso quello umano, migliorandolo
- La produttività è stata espressa in termini di parole per ora: più parole si riesce revisionare, migliore si suppone la qualità dell'output su cui è stato condotto il lavoro

	trans				pe-PBSMT				pe-NMT			
Lang. Pair	H1	H2	H3	Avg.	H1	H2	H3	Avg.	H1	H2	H3	Avg.
EN–DE	522	622	807	650	1641	1331	2016	1663	**1989**	**2192**	**2923**	**2368**
EN–ES	410	741	766	576	1264	2589	**1754**	**1869**	**1425**	**2849**	1097	1790
EN–IT	559	493	432	495	956	1046	560	854	**1338**	**1173**	**682**	**1064**
EN–JA	304	166	261	243	538	203	569	437	**644**	**235**	**812**	**564**
EN–ZH_CN	368	129	245	247	**1100**	**434**	550	**695**	886	302	**605**	597

Table 4 Words per hour for translating (trans), post-editing PBSMT (pe-PBSMT) or post-editing NMT (pe-NMT) output performed by human translators. H1, H2 and H3 denote the different human evaluators. In bold font are the highest rates for each translator and language pair.

Sottovalutazione

Confrontando i risultati del testo A-B con le metriche tradizionali, la sottovalutazione è risultata questa (tabella 5, p. 14):

	F-measure		BLEU		TER	
	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT
EN–DE	4%	52%	8%	52%	8%	41%
EN–ES	6%	52%	12%	53%	6%	37%
EN–IT	33%	33%	41%	28%	33%	26%
EN–JA	0%	61%	0%	65%	0%	44%
EN-ZH_CN	20%	38%	20%	35%	20%	28%
Average	13%	47%	16%	47%	13%	35%

In altri termini, BLEU **sottovaluta** clamorosamente i risultati della NMT (all'italiano va meglio che alle altre lingue)

Unica soluzione: la valutazione umana

- I sistemi di valutazione semiautomatici erano utili perché (sembrava) fornivano una buona correlazione con la valutazione umana
- Se manca la correlazione, i sistemi sono inutili
- A oggi, non ci sono sistemi che forniscano una correlazione migliore di BLEU = non ci sono sistemi utili
- Il lavoro che ho presentato mercoledì a CLiC-it 2019 è stato un primo studio di valutazione: realizzare una procedura di valutazione è piuttosto complesso

UNIVERSITÀ DI PISA
FILOLOGIA, LETTERATURA E LINGUISTICA

Valutazione umana di Google Traduttore e DeepL

Mirko Tavosanis – Università di Pisa

- Valutazione umana di 100 frasi di articoli giornalistici in inglese tradotte in italiano nel maggio 2019 da
  - Google
  - DeepL
  - Traduttori umani
- Le frasi sono state presentate ai valutatori umani senza informazioni sulla loro provenienza e sono state valutate su scala 1-5 per
  - Adeguatezza
  - Fluenza
- I valutatori erano studenti della laurea magistrale in Informatica umanistica dell'Università di Pisa (italiano come madrelingua, conoscenza dell'inglese almeno di livello B2) che avevano partecipato a un'attività di formazione dedicata

- Risultati

Traduttore	N. frasi	Media adeguatezza	Media fluenza	BLEU
Google	37	4,15	3,90	0,2538
DeepL	39	4,30	3,94	0,3254
Umano	24	4,60	4,46	n. a.

- La distanza rispetto alla traduzione umana è molto ridotta
- BLEU non ha una buona corrispondenza con la valutazione umana per i prodotti di traduzione a reti neurali