Valutazione umana di DeepL a livello di frase per le traduzioni di testi specialistici dall'inglese verso l'italiano

Sirio Papa - Mirko Tavosanis
Dipartimento di Filologia, letteratura e linguistica
Università di Pisa
Via Santa Maria 36 – 56126 Pisa PI
s.papa4@studenti.unipi.it -
mirko.tavosanis@unipi.it

Riassunto[1]

Il contributo presenta una valutazione delle prestazioni di DeepL nella traduzione di testi specialistici dall'inglese all'italiano. La valutazione è stata condotta a livello di frase, su un campione di 108 frasi tratte da testi relativi ad ambiente, energia, biomedicina e discipline del farmaco, e le traduzioni prodotte sono state valutate da traduttori in formazione dotati di competenze disciplinari. La traduzione di DeepL ha ottenuto una valutazione statisticamente pari a quella della traduzione umana per quanto riguarda l'adeguatezza e leggermente inferiore per quanto riguarda la scorrevolezza. La traduzione automatica dei testi ha inoltre ricevuto un punteggio superiore a quello ottenuto, con modalità simili, dalla traduzione automatica di testi giornalistici.

Abstract

The paper presents an evaluation of the performance of DeepL in the translation of specialized texts from English to Italian. The evaluation was carried out at sentence level, on a sample of 108 sentences taken from texts relating to the environment, energy, bio-medicine and drug science, and the translations produced were evaluated by translators in training, with disciplinary skills. The translation by DeepL was statistically rated at the same level of human translation in terms of adequacy and slightly lower in terms of fluency. Machine translation of the texts also received a higher score than that obtained in another analysis, carried out in a similar way, by machine translation of journalistic texts.

1 Introduzione

La valutazione delle effettive prestazioni dei sistemi di traduzione automatica continua a essere un problema complesso sia dal punto di vista teorico sia dal punto di vista pratico.

Dal punto di vista pratico, è oggi evidente che le metriche di valutazione più usate dopo il Duemila, e in particolare BLEU, non sono in realtà in grado di descrivere adeguatamente le differenze e i miglioramenti di prestazioni dei sistemi oggi in uso, e in particolare di quelli basati su reti neurali (Bentivogli e altri 2018a; Shterionov e altri 2018; Tavosanis 2019). Metriche proposte più di recente, come BERTScore, devono ancora essere valutate a fondo e sembrano comunque fornire risultati molto simili a quelli di BLEU (Zhang e altri 2020). Si è quindi ritenuto metodologicamente

Il testo è stato concepito unitariamente dagli autori, ma ai fini della ripartizione del lavoro si dichiara che sono opera di Sirio Papa i paragrafi 4, 7 e 8 e di Mirko Tavosanis i restanti paragrafi.

Per la collaborazione generosamente prestata, si ringraziano le professoresse Silvia Barra e Isabella Blum e gli studenti del Master online in Traduzione specialistica inglese > italiano realizzato dalle università di Genova e Pisa ed erogato dal Consorzio ICoN.

opportuno non usarle per questa valutazione, nemmeno come termine di confronto.

L'unico tipo di valutazione soddisfacente, a oggi, resta quindi quello condotto da valutatori umani. Non tutti i tipi di valutazione umana sono ugualmente soddisfacenti e affidabili. Le valutazioni condotte attraverso crowdsourcing da individui di cui non sono note le competenze assegnano per esempio alla traduzione automatica, sistematicamente, punteggi più alti rispetto a quelli assegnati da persone con provata competenza nella valutazione di traduzioni (Castilho e altri 2017a; Läubli e altri 2020: 658).

In questo contesto non mancano dichiarazioni in cui si rivendica il raggiungimento della "parità" tra traduzione automatica e traduzione umana per alcuni sistemi commerciali (Hassan e altri 2018). Le verifiche indipendenti in alcuni casi hanno confermato questi risultati, ma in altri hanno evidenziato differenze significative (Toral e altri 2018). Questa incertezza è poi in buona parte riconducibile alle circostanze della valutazione, che possono essere molto diverse tra di loro.

Il presente contributo punta a fornire ulteriori risultati inserendosi nel percorso di ricerca descritto in Tavosanis (2019), rispetto al quale rinforza il metodo di indagine e varia la tipologia testuale di riferimento. In Tavosanis (2019) le valutazioni sono state eseguite su testi giornalistici; nel presente contributo sono stati invece scelti testi specialistici. La valutazione punta in primo luogo a valutare la qualità delle traduzioni specialistiche in sé e in secondo luogo a vedere se i punteggi assegnati alle traduzioni specialistiche sono superiori o inferiori a quelli assegnati alle traduzioni di testi giornalistici. La traduzione automatica viene infatti normalmente usata su testi appartenenti a generi molto diversi, e valutare un unico genere è senz'altro molto limitante. (Burchardt e altri 2017: 159-160).

In particolare, date le sensibili differenze linguistiche tra i testi specialistici e i testi non specialistici, sembra verosimile che la stessa tecnologia di traduzione possa produrre risultati molto diversi nei due casi. Assicurare la qualità di traduzioni di testi provenienti da domini diversi è stato quindi considerato un problema fin dalla prima diffusione dei sistemi basati su reti neurali. Koehn e Knowles (2017: 29), per esempio, notando che "in different domains, words have different translations and meaning is expressed in different styles", presentano il *domain mismatch* come prima "sfida" per questi sistemi: nelle loro valutazioni, in questi contesti la NMT otteneva risultati inferiori a quelli dei sistemi SMT. Sembra inoltre diffusa l'idea che un sistema di traduzione a reti neurali generalista produca risultati di bassa qualità se applicato a testi specialistici (Chu e Wang 2020: 413). La *domain adaptation* è quindi un importante settore di sviluppo per la NMT (Chu e Wang 2018). Sono però rare, e quasi assenti per l'italiano, le valutazioni condotte con valutatori umani.

2 Il sistema valutato

Le verifiche descritte di seguito sono state compiute usando le traduzioni generate dal sistema DeepL, che è frequentemente segnalato come uno dei migliori prodotti della sua categoria. In particolare, nelle valutazioni comparative DeepL ha ottenuto negli ultimi anni punteggi spesso superiori a quelli di Google Traduttore (Heiss e Soffritti 2018; Tavosanis 2018; Tavosanis 2019).

Realizzato dall'azienda tedesca DeepL GmbH, DeepL è stato reso disponibile al pubblico nell'agosto del 2017 (sito: https://www.deepl.com/). Nell'ottobre 2020, il sistema copre un numero relativamente ridotto di lingue: undici in tutto, principalmente indoeuropee (italiano, inglese, tedesco, francese, spagnolo, portoghese, olandese, polacco e russo), con l'aggiunta di giapponese e cinese. Dal punto di vista tecnico, l'azienda ha dichiarato che il sistema di traduzione si basa su reti neurali, ma non ha fornito informazioni più specifiche.

Per quanto riguarda il rapporto con i domini, l'azienda non fornisce nessuna indicazione specifica. Si può quindi ipotizzare che il sistema sia generalista e non specializzato.

3 Composizione del corpus

Per la valutazione del lavoro è stato usato un corpus di testi specialistici di vario genere, composto da testi selezionati casualmente da due docenti del Master online in Traduzione specialistica inglese > italiano erogato congiuntamente dalle Università di Genova e Pisa e gestito dal Consorzio ICoN (http://www.traduzione.icon-master.it/).

I testi sono stati scelti dalle docenti di due dei domini trattati dal Master: Ambiente ed energia (professoressa Silvia Barra) e Biomedicina e discipline del farmaco (professoressa Isabella Blum). In tutti i casi dovevano essere disponibili sia il testo originale sia una traduzione professionale in lingua italiana realizzata da esseri umani. Le tipologie testuali sono state selezionate in modo da renderle rappresentative dell'ampia gamma di testi specialistici effettivamente trattati

nel Master: manuali, articoli scientifici, brevetti, schede di sicurezza. La definizione di "testo specialistico" è naturalmente piuttosto arbitraria, e comprende diverse tipologie testuali e diversi generi testuali. Tuttavia, è sembrato perfettamente adeguato agli scopi della valutazione riprendere i tipi di testo usati nella formazione dei traduttori umani professionali, senza distinzioni ulteriori.

4 Formazione del campione

Il campione da esaminare è stato costruito innanzitutto sottoponendo a DeepL, nella loro interezza, i testi selezionati; la traduzione è stata eseguita nel giugno del 2020. Dagli stessi testi sono poi state selezionate casualmente 108 frasi, 40 provenienti dal dominio Ambiente ed energia e 68 dal dominio Biomedicina e discipline del farmaco; la distribuzione per dominio è proporzionale alla consistenza del rispettivo corpus. Si è ritenuto che non fosse possibile indicare a priori uno dei due domini come più difficile da tradurre rispetto all'altro e che quindi non fosse necessario bilanciare la composizione. Nella selezione sono state evitate le frasi ripetute e quelle nominali o disposte in tabella o in elenco.

La dimensione del campione è ridotta rispetto a quello di campagne di valutazione recenti come Intento, che ha preso in esame 500 "segmenti" per numerose coppie di lingue e numerosi domini (Intento 2020). Tuttavia, Intento ha valutato le frasi usando il sistema automatico BERTScore, menzionato al § 1, senza ricorrere a valutatori umani. La dimensione del campione usato qui è invece simile a quelle dei campioni usati in altre esperienze con valutatori umani, condotte per esempio con 150 frasi (Hassan e altri 2018), 299 frasi (Läubli e altri 2020: 657), 104 frasi (Läubli e altri 2020: 658-659), e così via.

Le dimensioni complessive del campione sono state di 2.826 token (1664 per Biomedicina e 1162 per Ambiente). La lunghezza media è quindi di poco superiore ai 26 token per frase e la mediana si attesta a 22 token. La frase più breve è lunga 9 token, mentre quella più lunga 91, ma rappresenta chiaramente un outlier dato che il 75% delle frasi ha una lunghezza entro i 33 token.

Per ogni frase selezionata sono stati raccolti:

1. La frase originale in inglese
2. La corrispondente traduzione in italiano realizzata da un traduttore umano
3. La corrispondente traduzione in italiano realizzata da DeepL

Le 108 frasi tradotte da DeepL sono state divise in due gruppi di 54, denominati gruppo A e gruppo B. All'interno di ogni gruppo sono state poi inserite le altre 54 frasi nella versione realizzata da traduttori umani. Ognuna delle 108 frasi tradotte da DeepL e da esseri umani è stata poi valutata sia per l'adeguatezza (*adequacy*) sia per la scorrevolezza (*fluency*) da ogni valutatore del gruppo a cui era stata associata. Questo ha permesso di ottenere una valutazione di tutte le 108 frasi tradotte da esseri umani e di tutte le 108 frasi tradotte da DeepL.

Durante la valutazione, le frasi state sottoposte ai valutatori umani in ordine casuale e senza indicazioni sulla loro origine: i valutatori non avevano quindi elementi esterni per decidere se l'origine di una singola frase era un traduttore umano o DeepL. Nella valutazione per adeguatezza le frasi erano accompagnate dal testo originale in lingua inglese, secondo l'orientamento *DA-src* (Bentivogli e altri 2018b: 62), mentre nella valutazione per scorrevolezza era disponibile solo il testo italiano. La valutazione è stata eseguita online, usando il sistema KantanLQR[2], per un tempo medio di un'ora per ogni campione.

5 Criteri di valutazione

Anche se i risultati delle verifiche sulla traduzione automatica condotte in rapporto ai convegni WMT hanno confermato la maggior rilevanza dell'adeguatezza rispetto alla fluenza (Bentivogli e altri 2018b: 62), le due diverse valutazioni sono state conservate per verificare l'esistenza di differenze tra di loro. Va comunque notato che, nonostante sia teoricamente possibile che una frase tradotta con sistemi a reti neurali si allontani molto dal senso testo di partenza, nella pratica non si è prodotto nessun caso di questo genere.

Per l'adeguatezza è stata usata una scala di valori basata su criteri relativi:

1. Il contenuto informativo dell'originale è stato completamente alterato
2. È stata trasmessa una parte del contenuto informativo, ma non la più importante
3. Circa metà del contenuto informativo è stata trasmessa
4. La parte più importante del contenuto informativo originale è stata trasmessa

[2] KantanLQR è un sistema che fornisce strumenti automatizzati per valutazione e l'analisi di segmenti linguistici. Il sistema è implementato sulla piattaforma di KantanMT (<https://kantanmt.com/>), ma può essere utilizzato indipendentemente da essa, su qualsiasi corpus organizzato e diviso in singole frasi.

5. Il contenuto informativo è stato tradotto completamente

Per la scorrevolezza, sulla base del livello medio di traduzione visto in altre verifiche, la scala è invece stata basata su criteri in parte assoluti:

1. Impossibile da ricondurre alla norma
2. Con più di due errori morfosintattici
3. Con non più di due errori morfosintattici e/o molti usi insoliti di collocazioni
4. Con non più di un errore morfosintattico e/o un uso insolito di collocazioni
5. Del tutto corretta

6 Composizione del gruppo dei valutatori

Il gruppo dei valutatori è stato interamente composto da studenti del Master online in Traduzione specialistica inglese > italiano citato al § 3. La maggior parte dei valutatori, all'interno del Master, aveva approfondito l'uno o l'altro dei domini presi in esame, o entrambi. Tutti avevano comunque l'italiano come lingua madre e disponevano di una conoscenza della lingua inglese valutabile tra C1 e C2. Nessuno di loro è stato coinvolto nella fase di scelta e preparazione degli articoli.

La scelta di valutatori specializzati è conseguenza di due idee di base: innanzitutto, solo le persone dotate di conoscenze disciplinari sono i destinatari normali di testi specialistici; inoltre, solo le persone dotate di conoscenze disciplinari possono valutare con cognizione di causa un testo specialistico. Per esempio, valutare anche solo la correttezza grammaticale di frasi come questa sembra possibile solo a chi sa se nell'italiano specialistico sono o no accettabili sintagmi come *in aperto* e parole come *farmacocinetica*:

"È stato condotto uno studio a dose singola in aperto per valutare la farmacocinetica di una dose ridotta di sitagliptin (50 mg) in pazienti con vari gradi di compromissione renale cronica rispetto a soggetti sani di controllo."

Per migliorare l'omogeneità del risultato, alcuni mesi prima della valutazione vera e propria è stata fatta una sessione di addestramento con i valutatori interessati. In questa sessione sono state valutate numerose frasi (diverse da quelle esaminate in seguito), e i punteggi assegnati sono stati discussi collettivamente, cercando di arrivare a parametri di valutazione quanto più possibile condivisi

I valutatori sono stati complessivamente 15: 7 per il gruppo A, 8 per il gruppo B. Il numero è quindi superiore a quello usato in valutazioni umane simili, come quelle descritte in Hassan e altri (2018) e Läubli e altri (2020).

7 Esito generale della valutazione

I risultati della valutazione sono riportati in Tabella 1.

Traduttore e sottocorpus	Media adeguatezza	Media scorrevolezza	σ adeguatezza	σ scorrevolezza
Umano complessivo	4,29	4,17	0,43	0,60
Biomedicina	4,38	4,41	0,36	0,38
Ambiente	4,15	3,78	0,49	0,68
DeepL complessivo	4,31	4,09	0,45	0,56
Biomedicina	4,36	4,06	0,48	0,59
Ambiente	4,24	4,14	0,39	0,51

Tabella 1: Risultati della valutazione.

La variazione nei giudizi è, in generale, piuttosto limitata. Per quanto riguarda l'adeguatezza della traduzione umana, la deviazione standard è stata di 0,43 e 39 frasi su 108 hanno ottenuto un punteggio maggiore di 4,50. Solamente 2 frasi hanno ottenuto un punteggio minore o uguale a 3. Le traduzioni di DeepL hanno ottenuto una deviazione standard di 0,45; 40 frasi hanno ottenuto un punteggio maggiore di 4,50, e solo una un punteggio minore o uguale a 3.

La deviazione standard collegata alla scorrevolezza è stata più alta, ma comunque contenuta: 0,60 per la traduzione umana, 0,56 per la traduzione di DeepL. Per la scorrevolezza, va notato, inoltre, che il punteggio 5 è stato assegnato all'unanimità solo a pochissime frasi e il punteg-

gio minimo ottenuto (2,00 in entrambe le traduzioni) è più basso di quello dell'adeguatezza (2,75 per entrambe le traduzioni). Tuttavia, 26 traduzioni di DeepL hanno ottenuto un punteggio medio superiore a 4,5, contro 32 traduzioni umane.

Complessivamente, la traduzione automatica ha ricevuto un punteggio migliore della traduzione umana per quanto riguarda l'adeguatezza, e inferiore per quanto riguarda la scorrevolezza. I dati sono stati, inoltre, sottoposti ad un t-test per verificare la significatività delle differenze. I risultati presentano un p value di 0,762 per l'adeguatezza e un p value di 0,313 per la scorrevolezza. I valori dei p value fanno concludere che, con il 95% di confidenza statistica, non è possibile affermare che i risultati dell'adeguatezza ottenuti da DeepL siano effettivamente migliori dei risultati ottenuti dalla traduzione umana, o viceversa (p value > 0,05). Al contrario, i risultati della scorrevolezza ottenuti dalla traduzione umana possono dirsi significativamente migliori rispetto ai risultati ottenuti dalla traduzione automatica (p value < 0,05).

8 Valutazioni particolari

Per l'adeguatezza, solo una frase tradotta da DeepL ha ottenuto un risultato minore o uguale a 3:

Originale: "After discontinuation of short-term and long-term treatment with pregabalin withdrawal symptoms have been observed in some patients".
Traduzione: "Dopo l'interruzione del trattamento a breve e a lungo termine con sintomi di astinenza da pregabalin sono stati osservati in alcuni pazienti".

Lo stesso è avvenuto per due frasi tradotte da traduttori umani:

Originale: "Ampersand's leadership knew that to keep the product from being cost prohibitive, it'd have to create a model that was sustainable for the people who needed the electric mototaxis the most: the motars".
Traduzione: "I dirigenti di Ampersand sapevano che evitare che il prodotto avesse un costo proibitivo avrebbe creato un modello sostenibile per coloro che avevano bisogno più degli altri del mototaxi elettrico: i motars".

Originale: "This information is based on our current knowledge and is intended to describe the product for the purposes of health, safety and environmental requirements only."
"Queste informazioni sono basate sulle nostre conoscenze attuali e sono intese descrivere il prodotto per il solo scopo dei requisiti di salute, sicurezza e ambientali."

Il punteggio pieno è stato assegnato a 3 frasi tradotte da DeepL:

Originale: "This medicinal product does not require any special storage conditions".
Traduzione: "Questo medicinale non richiede particolari condizioni di conservazione".

Originale: "The other ingredients are: lactose monohydrate, maize starch, talc, gelatine, titanium dioxide (E171), sodium laurilsulphate, anhydrous colloidal silica, black ink, (which contains shellac, black iron oxide (E172), propylene glycol, potassium hydroxide) and water".
Traduzione: "Gli altri ingredienti sono: lattosio monoidrato, amido di mais, talco, gelatina, biossido di titanio (E171), laurilsolfato di sodio, silice colloidale anidra, inchiostro nero, (che contiene gommalacca, ossido di ferro nero (E172), glicole propilenico, idrossido di potassio) e acqua".

Originale: "Animal data do not suggest an effect of treatment with sitagliptin on male and female fertility".
Traduzione: "I dati relativi agli animali non suggeriscono un effetto del trattamento con sitagliptina sulla fertilità maschile e femminile".

Lo stesso punteggio è stato assegnato a una sola frase tradotta da un essere umano:

Originale: "Pregabalin should be discontinued immediately if symptoms of angioedema, such as facial, perioral, or upper airway swelling occur".
Traduzione: "Il trattamento con pregabalin deve essere immediatamente interrotto in presenza di sintomi di angioedema come gonfiore del viso, gonfiore periorale o gonfiore delle vie respiratorie superiori".

Per la scorrevolezza, nessuna frase ha ottenuto un punteggio pieno, né per la traduzione umana né per quella automatica. Sono state più frequenti, invece, le frasi che hanno ottenuto un punteggio minore o uguale a 3. Nel caso delle traduzioni di

DeepL sono state quattro, tra cui per esempio questa:

“L'analisi del ricovero in ospedale per insufficienza cardiaca è stata adattata per una storia di insufficienza cardiaca al basale”.

Le traduzioni umane ad avere ottenuto un punteggio di scorrevolezza minore o uguale a 3 sono state invece cinque, tra cui per esempio questa:

“Modulo di cella solare comprendente un insieme di pre-laminazione per cella solare, in cui l’insieme è come elencato in qualsiasi rivendicazione da 1 a 11.”

9 Confronto con il testo giornalistico

In Tavosanis (2019) la valutazione umana delle traduzioni di testi giornalistici, condotta con gli stessi criteri di valutazione e con un numero di valutatori comparabile, aveva fornito i risultati riportati nella Tabella 2.

Traduttore	N. frasi	Media adeguatezza	Media scorrevolezza
Google	37	4,15	3,90
DeepL	39	4,30	3,94
Umano	24	4,60	4,46

Tabella 2: Valutazione complessiva delle traduzioni di testi giornalistici in Tavosanis (2019).

Confrontando questi risultati con quelli presentati nella Tabella 1, la differenza principale consiste nel peggioramento del punteggio assegnato alla traduzione umana. Se si presuppone che la qualità della traduzione umana sia stabile da una rilevazione all’altra e da un tipo di testo all’altro, questo peggioramento potrebbe essere attribuito a una maggiore severità dei revisori in quanto esperti di dominio (possibilità anticipata nel § 5). Intuitivamente, esistono però numerose altre spiegazioni possibili, in isolamento o in combinazione: per esempio, che il testo specialistico sia più adatto a questo tipo di traduzione automatica rispetto al testo giornalistico, o che sia più difficile da gestire per i traduttori umani. Allo stato attuale delle conoscenze non ci sono fattori che spingano a preferire una spiegazione rispetto a un’altra.

10 Conclusioni e sviluppi futuri

I risultati ottenuti con questa prova supportano l’ipotesi che anche per l’italiano, perlomeno per alcune tipologie testuali e a livello di frase, la traduzione automatica abbia raggiunto un livello qualitativo statisticamente pari a quello della traduzione umana per quanto riguarda l’adeguatezza e leggermente inferiore per quanto riguarda la scorrevolezza. Sono quindi coerenti con diversi altri risultati recenti, presentati per altre lingue (Läubli e altri 2020: 660); va però ricordato che l’italiano non è stato incluso negli importanti task di WMT 2019 (Barrault e altri 2019).

Inoltre, i risultati ottenuti non supportano l’ipotesi che la NMT di sistemi di uso generale ottenga risultati inferiori quando viene applicata a testi specialistici rispetto a quando viene applicata a testi non specialistici.

L’analisi ha naturalmente diversi limiti: per esempio, il campione valutato è relativamente ristretto, le oscillazioni nel giudizio dei valutatori non possono essere confrontate con una media professionale sperimentata e i domini specialistici presi in considerazione sono solo due. Tuttavia, l’estensione e il miglioramento di queste pratiche sembrano a oggi l’unico modo per valutare correttamente le capacità della traduzione automatica in italiano.

Per quanto riguarda gli sviluppi futuri, la necessità di una valutazione realistica sembra rendere indispensabile il passaggio dalla valutazione di singole frasi a quella di testi interi. La qualità della traduzione automatica a livello di testo risulta infatti, in diversi casi, sensibilmente peggiore rispetto a quella a livello di frase (Läubli e altri 2020: 660). La mancanza di sistemi strutturali per garantire la coerenza a livello di testo nella traduzione a reti neurali fa pensare che il fenomeno sia strutturale; per verificare queste ipotesi sono però necessarie valutazioni dedicate.

Al tempo stesso, il confronto con la valutazione dei testi giornalistici suggerisce l’idea che i risultati possano variare in modo sensibile da un genere testuale all’altro, e che almeno in alcuni casi possano essere migliori rispetto a quelli che si ottengono con testi non specialistici. La variabilità collegata al genere non è contemplata nella peraltro dettagliatissima sintesi di Läubli e altri (2020), ma sembra indispensabile prenderla strutturalmente in considerazione per rendere più solide tutte le valutazioni future.

Bibliografia

Barrault, Loïc, e altri (2019). *Findings of the 2019 Conference on machine translation (WMT 2019*). In *Proceedings of the WMT*, Firenze, Association for computational linguistics, pp. 1-61.

Bentivogli, Luisa, e altri (2018a). *Neural versus phrase-based MT quality: an in-depth analysis on English–German and English–French*. In *Computer speech & language*, 49, pp. 52-70.

Bentivogli, Luisa, e altri (2018b). *Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment*. In *Proceedings of the 15th international workshop on spoken language translation*, Iwslt, pp. 62-69.

Burchardt, Aljoscha, e altri (2017). *A linguistic evaluation of rule-based, phrase-based, and neural MT engines*. In *The Prague bulletin of mathematical linguistics*, 108, pp. 159-70.

Castilho, Sheila, e altri (2017). *Crowdsourcing for NMT evaluation: professional translators versus the crowd.* In *Proceedings of translating and the computer*, 39, pp. 1-22.

Chu, Chenhui e Rui Wang (2018). *A survey of domain adaptation for machine translation*. In *Proceedings of the 27th international conference on computational linguistics*, Association for computational linguistics, pp. 1304-1319.

Chu, Chenhui e Rui Wang (2020). *A survey of domain adaptation for machine translation*. In *Journal of information processing*, 28, pp. 413-426.

Hassan, Hany, e altri (2018). *Achieving human parity on automatic Chinese to English news translation.* arXiv preprint arXiv:1803.05567 (2018).

Heiss, Christine e Marcello Soffritti (2018). *DeepL Traduttore e didattica della traduzione dall'italiano in tedesco–alcune valutazioni preliminari*. In *Translation and Interpreting for Language Learners (TAIL). Lessons in honour of Guy Aston, Anna Ciliberti, Daniela Zorzi*, a cura di Laurie Anderson, Laura Gavioli e Federico Zanettin, Milano, AItLA, pp. 241-258.

Intento (2020). *Independent multi-domain evaluation of commercial machine translation engines*. Intento, Inc.

Koehn, Philipp e Rebecca Knowles (2017). *Six Challenges for Neural Machine Translation*. In *First Workshop on Neural Machine Translation*, Association for Computational Linguistics, pp. 28-39.

Läubli, Samuel, e altri. (2020). *A set of recommendations for assessing human-machine parity in language translation*. In *Journal of artificial intelligence research*, 67, pp. 653-672.

Shterionov, Dimitar, e altri (2018). *Human versus automatic quality evaluation of NMT and PBSMT*. In *Machine Translation*, 32, 3, pp. 217-235.

Tavosanis, Mirko (2018). *Lingue e intelligenza artificiale.* Roma: Carocci.

Tavosanis, Mirko (2019). *Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano*. In *CLiC-it 2019 – Proceedings of the Sixth Italian Conference on Computational Linguistics*, a cura di Raffaella Bernardi, Roberto Navigli e Giovanni Semeraro, CEUR Workshop Proceedings, Aachen University, pp. 1-7.

Toral, Antonio, e altri (2018). *Attaining the unattainable? Reassessing claims of human parity in neural machine translation*. arXiv preprint arXiv:1808.10432.

Zhang, Tianyi, e altri (2020). *BERTScore: evaluating text generation with Bert*. arXiv preprint arXiv:1904.09675.