ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA

Scuola di Scienze
Corso di Laurea in Ingegneria e Scienze Informatiche

TRADUZIONE AUTOMATICA DEI DOCUMENTI SOCIAL CONDIVISI DA MALATI RARI

Elaborato in
Programmazione

Relatore
Prof.ssa. Antonella Carbonaro

Co-relatore
Dott. Giacomo Frisoni

Presentata da
Anna Fabris

Seconda Sessione di Laurea
Anno Accademico 2019 – 2020

PAROLE CHIAVE

Machine Translation

Named-entity Recognition

Natural Language Processing

Machine Translation Evaluation Metric

Rare Diseases

Normalmente si suppone che qualcosa
venga sempre perso nella traduzione;
mi aggrappo, ostinatamente, all'idea
che qualcosa può anche essere guadagnato.
- Salman Rushdie

Sommario

Circa il 5% della popolazione mondiale è affetta da una delle oltre 6000 malattie rare oggi registrate. Il termine "raro" si riferisce quindi, non tanto alla percentuale di individui colpiti complessivamente, quanto, secondo la definizione dell'Unione Europea, a un'incidenza inferiore a un caso ogni 2000 abitanti. Le limitate conoscenze disponibili per ciascuna malattia portano i pazienti a dover ricercare informazioni in maniera autonoma, spesso mediante l'utilizzo dei *social media*. Questo ha dato vita a un'enorme quantità di contenuti testuali oggi in continua crescita, che è possibile analizzare efficacemente con soluzioni NLP (Natural Language Processing). La stragrande maggioranza dei tool attualmente disponibili è pensata per l'elaborazione di documenti in lingua inglese, giustificando l'importanza sia dello sviluppo di nuovi modelli multilingua che della traduzione di dataset esistenti per un maggior supporto sul piano implementativo. L'utilizzo dei traduttori automatici presenti in commercio è frequentemente non sufficiente, in quanto le traduzioni delle entità di dominio risultano imprecise, mentre sul piano del *preprocessing* ne è richiesta una gestione corretta. Questa tesi si pone come obbiettivo il confronto di molteplici tecniche di traduzione automatica e la proposta di un approccio che sappia gestire consapevolmente le entità di dominio tramite l'adozione di sistemi NER (Named Entity Recognition), prendendo come caso di studio un corpus di post e commenti condivisi su un gruppo Facebook di pazienti affetti da Acalasia Esofagea.

Introduzione

Motivazioni e contesto

Circa il 5% della popolazione mondiale possiede una malattia rara [1]. Un paziente con una tale diagnosi difficilmente può ricevere indicazioni utili da un medico non specializzato. Per molte malattie rare, infatti, le conoscenze di base, come la causa della malattia, il suo decorso e i trattamenti terapeutici, sono limitate o non disponibili [2].

Questo porta i pazienti a dover spesso ricercare informazioni in maniera autonoma. Nell'ultimo decennio, in particolare, si è osservata la nascita di varie comunità sui *social media* che permettono di scambiarsi consigli, opinioni ed esperienze.

Tale sviluppo ha rivoluzionato la quantità e il tipo di dati disponibili. Con soluzioni di NLP (Natural Language Processing) è possibile estrarre efficacemente conoscenza a partire dai soli contenuti testuali. I modelli NLP necessitano sempre più di un'enorme quantità di dati per il training, e la maggior diffusione della lingua inglese sul Web[1] ha un conseguenziale impatto sui dataset (spesso costruiti con *scraping*) [3], raramente in italiano. Per questa ragione il numero di tool NLP per la lingua italiana è sensibilmente inferiore rispetto all'ammontare di risorse disponibile per la lingua inglese.

La traduzione automatica dei dati in lingua inglese riveste pertanto un ruolo fondamentale, consentendo un potenziale miglioramento delle prestazioni e l'apertura verso un maggior numero di software applicabili per il task considerato.

I traduttori automatici presenti in commercio hanno però difficoltà nella traduzione di entità di dominio (ad esempio, "policlinico gemelli" sarebbe erroneamente tradotto in "twins polyclinic"), la cui corretta gestione è però richiesta sul piano del *preprocessing*.

[1] `https://www.internetworldstats.com/stats7.htm`

Contributo

Questa tesi nasce dalla necessità di tradurre correttamente i post e commenti *social* condivisi da una comunità italiana di pazienti e di gestire con cura le entità menzionate al loro interno. In tal senso, il lavoro si sviluppa a partire dal dataset costruito da Giacomo Frisoni in "A new unsupervised methodology of descriptive textmining for knowledge graph learning" [4] contenente i dati provenienti da un gruppo Facebook[2] sull'Acalasia Esofagea, un raro disturbo dell'esofago [4].

L'indagine proposta analizza alcuni programmi di traduzione automatica e ne valuta la qualità dei risultati, proponendo anche una nuova soluzione atta a evitare una traduzione erronea di termini di dominio ambigui.

Organizzazione della tesi

- **Capitolo 1 - Stato dell'arte**. Espone una descrizione delle principali metodologie di traduzione automatica, i principali software presenti sul mercato e le metriche che possono essere utilizzate per valutarli. Mostra inoltre l'applicazione di tali metriche su un documento d'esempio appartenente al dominio considerato.
- **Capitolo 2 - Motivazioni e caso di studio**. Illustra le motivazioni della tesi soffermandosi sull'importanza del NLP in campo medico e di come l'utilizzo della lingua italiana possa costituire un limite sul piano implementativo.
- **Capitolo 3 - Traduzione automatica di documenti social**. Implementazione e valutazione automatica/manuale delle traduzioni generate dai software selezionati. Considerazioni su quale sia il software più efficace e osservazioni sulla correlazione tra i risultati delle metriche automatiche e il giudizio umano.
- **Capitolo 4 - Gestione di entità durante la traduzione**. Proposta di una nuova tecnica di gestione per impedire la traduzione erronea di entità di dominio.

[2] `https://www.facebook.com/groups/36705181245/`

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Stato dell'arte

1.1 Principali metodologie di traduzione

Lo sviluppo della traduzione automatica (in inglese Machine Translation, MT) risale al secolo scorso [5]. A partire dagli anni '50, ci sono stati numerosi tentativi per realizzare programmi di traduzione automatica. Solamente negli anni '70 ci furono però significativi successi [6].
Nel corso degli anni, sono emersi tre approcci principali: traduzione *rule-based*, statistica e basata su reti neurali. La Figura 1.1 mostra l'evoluzione dei tre metodi nel tempo.

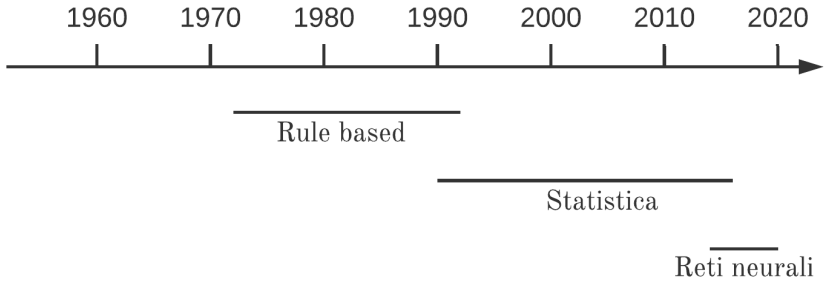


Figura 1.1: Evoluzione dei principali metodi di traduzione automatica

1.1.1 Traduzione Rule-based

I sistemi *rule-based* richiedono la conoscenza di esperti nelle lingue di partenza e di destinazione per sviluppare regole sintattiche, semantiche e morfologiche

utili ad ottenere la traduzione.
Per tradurre una frase dell'italiano all'inglese è necessario:

- un dizionario che associ ogni parola italiana con la parola inglese appropriata;
- regole che descrivano la normale struttura delle frasi in italiano;
- regole che descrivano la normale struttura delle frasi in inglese;
- regole che mettano in relazione queste due strutture.

I vantaggi di un sistema automatico di traduzioni *rule-based* consistono nella possibilità di avere controllo totale sul risultato di una traduzione (è possibile scrivere una nuova regola per ogni situazione), e nel non necessitare di un corpus bilingue. Tra gli svantaggi c'è la necessità di avere buoni dizionari e regole stabilite manualmente, inoltre con un numero elevato di regole il sistema diventa complicato da gestire.

PROMPT[1] e SYSTRAN[2] sono stati gli esempi più famosi di sistemi *rule-based*. Nel corso dell'ultimo decennio entrambe le aziende hanno però deciso di implementare tecniche di traduzione statistica e successivamente neurale.

Inizialmente sviluppato per tradurre tra lingue romanze della Spagna (spagnolo, catalano e galiziano) Apertium[3] è un software *open-source* ancora in fase di sviluppo che si basa tuttora sulla traduzione *rule-based* [7].

1.1.2 Traduzione Statistica

All'inizio del 1990, presso l'IBM Research Center[4], è stato mostrato per la prima volta un sistema di traduzione automatica estraneo ai tradizionali approcci incentrati su regole e linguistica.

Analizzando sintatticamente testi tradotti in due lingue, si cerca di trovare delle corrispondenze tra le parole della frase da tradurre e le parola della frase tradotta. Per esempio all'inizio, data la parola "tavolo" in una frase da tradurre si presuppone che sia correlata in uguale misura con qualsiasi parola della frase tradotta. Con l'apparizione di "tavolo" in altre frasi, il numero di correlazioni con "table" aumenta [8].

L'idea alla base della MT è assegnare ad ogni coppia di frasi (S, T) una probabilità $Pr(T|S)$, da interpretare come la probabilità che un traduttore produca T nella lingua di destinazione quando viene presentato con S nella

[1] `https://www.promt.com/`
[2] `https://www.systransoft.com/`
[3] `https://www.apertium.org/`
[4] `https://www.research.ibm.com/labs/`

lingua di origine. La Figura 1.2 mostra due coppie di frasi (S, T) e il valore atteso per $Pr(T|S)$.

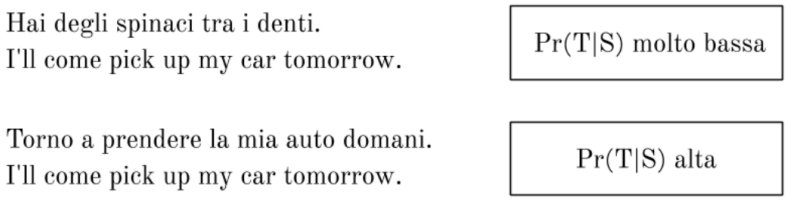


Figura 1.2: Esempio del valore $Pr(T|S)$ per una coppia di frasi (S, T)

Viene considerato un problema equivalente più semplice da risolvere. Data una frase T nella lingua di destinazione, cerchiamo la frase S dalla quale il traduttore ha prodotto la T. Sappiamo che la nostra possibilità di errore è ridotta al minimo scegliendo la frase S che è più probabile data T. Pertanto, vogliamo scegliere S in modo da massimizzare $Pr(S|T)$. Utilizzando il teorema di Bayes, possiamo scrivere l'Equazione 1.1.

$$\Pr(S|T) = \frac{\Pr(S)\Pr(T|S)}{\Pr(T)} \tag{1.1}$$

$Pr(T)$ non dipende da S, e quindi è sufficiente scegliere la S che massimizza il prodotto $Pr(S)Pr(T|S)$.

Quindi un sistema di traduzione statistica richiede un metodo per calcolare le probabilità del modello linguistico $Pr(S)$, un metodo per calcolare le probabilità di traduzione di T dato S, $Pr(T|S)$, infine, un metodo per cercare tra le possibili frasi di origine S quella che massimizza il valore $Pr(S)Pr(T|S)$ [9].

Nei paragrafi precedenti abbiamo usato la frase come unità di traduzione. Il modello più usato si chiama *phrase-based* e utilizza n-grammi (di solito digrammi o trigrammi) per il processo discusso in precedenza [10]. Per esempio, l'n-gramma inglese "I had" è tradotto in italiano "avevo".

Il sistema ha però bisogno di milioni e milioni di frasi in due lingue per raccogliere le statistiche rilevanti per ogni n-gramma. Questo corpus di testi viene utilizzato per dedurre automaticamente un modello statistico di traduzione. Il modello viene poi applicato ai testi non tradotti per fare una corrispondenza probabilistica che suggerisce una traduzione.

Alcuni inconvenienti della traduzione statistica sono l'impossibilità di correggere errori specifici e l'inabilità di considerare informazioni, se distanti dalla parole da tradurre, causando quindi incoerenze nei risultati della traduzione, come per esempio di genere.

Un altro problema causato dalla gestione di sole piccole sequenze di parole, è che l'ordine delle parole stesse deve essere pensato dal progettista del programma. Alcune classificazioni possono essere fatte in base all'ordine tipico del soggetto, del verbo e dell'oggetto in una frase, mentre altre includono modelli di riordino più complessi. In ogni caso la qualità della traduzione cala se l'ordine delle parole nelle due lingue è molto differente.

I vantaggi sono però molteplici: meno lavoro manuale da parte di esperti linguistici, possibile adattabilità a più coppie di linguaggi e con il giusto modello la traduzione è decisamente migliore rispetto a quella *rule-based*.

1.1.3 Reti neurali

I recenti progressi in ambito *deep learning* hanno determinato lo sviluppo di modelli NLP (Natural Language Processing) capaci di evolvere significativamente le soluzioni allo stato dell'arte per molteplici *task*. Come diretta conseguenza di tale evoluzione, accompagnata da una crescita esponenziale di dati testuali non strutturati senza precedenti, la MT è oggi principalmente affrontata con architetture basate su reti neurali.

Descrizione dei diversi tipi di reti neurali

Le reti neurali si ispirano ai primi modelli di elaborazione del cervello, sono composte da nodi interconnessi, ognuno dei quali riceve un certo numero di ingressi e fornisce un'uscita, vedi Figura 1.3. Applicando algoritmi che imitano i processi dei neuroni reali, possiamo far sì che la rete "impari" a risolvere molti tipi di problemi. Un neurone riceve input da una serie di altre unità o fonti esterne, pesa ciascuno di essi e realizza la loro somma. Se l'ingresso totale è al di sopra di una soglia, l'uscita dell'unità è uno, altrimenti è zero. Pertanto, l'uscita cambia da 0 a 1 quando la somma totale ponderata degli ingressi è uguale alla soglia.

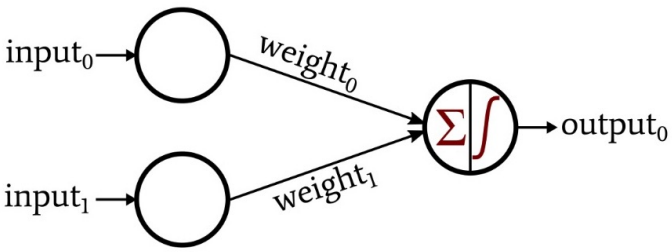


Figura 1.3: Modello di percettrone a un livello

Ottenuta da: https://www.allaboutcircuits.com/technical-articles/how-to-perform-classification-using-a-neural-network-a-simple-perceptron-example/

I fattori principali che distinguono i diversi tipi di reti tra loro è il modo in cui i nodi sono collegati e il numero di *layers*. Generalmente le reti con più *layers* nascosti sono considerate DNN (Deep neural network). In seguito vengono presentati i principali tipi di reti neurali usati per la traduzione automatica [11].

- **FFNN** (Feedforward neural network) sono una tipologia di rete basilare in cui tutti i nodi sono organizzati in strati sequenziali, con ogni nodo che riceve valori solo dai nodi degli strati precedenti.

  - **CNN** (Convolutional Neural Network) sono un sottotipo di FFNN, che impiega un'operazione matematica chiamata convoluzione. Una rete neurale convoluzionale è costituita da uno strato di ingresso e uno di uscita, oltre che da più strati nascosti, gli strati intermedi hanno neuroni disposti in 3 dimensioni.

    Alcune delle caratteristiche principali di questa tipologia di rete sono la connessione locale dei neuroni e la condivisione dei pesi. I neuroni di uno strato sono collegati solamente ad una piccola regione dello strato precedente, questo permette una grande riduzione del numero di connessioni. I pesi sono condivisi a gruppi, i neuroni di uno stesso livello eseguono lo stesso tipo di elaborazione su porzioni diverse dell'input.

- **RNN** (Recursive Neural Network) sono un tipo di rete neurale in cui i valori di ingresso di uno strato possono essere le uscite di uno strato di livello successivo. Questo permette l'uso della memoria interna per elaborare le sequenze degli ingressi. L'idea alla base delle unità delle RNN è il mantenimento di uno stato interno, che rappresenta tutto ciò che è stato visto precedentemente.

  Le RNN soffrono di due problemi *exploding gradients* e *vanishing gradients*. L'*exploding gradients* si presenta quando durante il *training* di una rete neurale l'accumulo di errore si traduce in ingiustificati grandi aggiornamenti dei pesi. Il *vanishing gradient* è il problema che sorge durante il *training* di una rete neurale che causa la diminuzione esponenziale del valore del gradiente.

  Entrambi i problemi sono causati dalla natura iterativa dell'RNN, il cui gradiente è essenzialmente uguale alla matrice dei pesi portata ad una potenza elevata. Questo fa sì che il gradiente cresca o si riduca ad un tasso esponenziale nel numero di passi.

  A differenza del problema dell'*exploding gradients* che si è rivelato relativamente facile da risolvere, il problema del *vanishing gradient* com-

porta l'impossibilità di istruire la rete in base alle *long-term dependency* [12, 13].

Questo non è tollerabile perché uno dei principali vantaggi delle reti neurali dovrebbe essere la capacità di utilizzare le informazioni precedenti per il compito attuale. Ad esempio, per determinare il genere di un aggettivo in italiano, possono dovere essere necessarie anche parole non vicine.

- **LSTM** (Long Short-Term Memory) sono una versione modificata delle RNN inventata per essere addestrata con successo utilizzando l'algoritmo BPTT (Backpropagation Through Time), evitando il problema del *vanishing gradient*.

  Le unità LSTM includono una cella di memoria, che gli permette di ricordare i dati per un periodo più lungo di tempo e *gate units* che hanno la capacità di rimuovere o aggiungere informazioni allo stato della cella, ma anche decidere di lasciar scorrere le informazioni senza modifiche [14].
- **GRU** (Gated Recurrent Unit) sono simili all'unità LSTM e non hanno il problema del *vanishing gradient*. Includono unità che modulano il flusso delle informazioni al loro interno, senza però avere una memoria separata.

  I GRU sono molto più semplici, richiedono meno potenza di calcolo e sono addestrabili più velocemente, possono quindi essere usati per formare reti molto profonde. Tuttavia le reti LSTM sono più potenti, possono garantire maggiore espressività e accuratezza, ma richiedono molta potenza di calcolo [15].

1.1.4 Architettura dei modelli di traduzione neurale

Modello encoder-decoder

L'architettura dei modelli di traduzione neurale è di tipo *end-to-end* (il modello elabora i dati di origine e genera i dati di destinazione direttamente, senza spiegazioni o risultati intermedi). Questo processo può essere diviso in due fasi: codifica e decodifica, e quindi si può separare funzionalmente l'intera architettura come *encoder* e *decoder*, vedi Figura 1.4.

L'*encoder* è responsabile della codifica della frase di partenza in un vettore di lunghezza fissa. Ogni parola della frase di ingresso viene inserita separatamente nel modello in più fasi temporali consecutive. Il *decoder* è responsabile della lettura dal vettore e della sua trasformazione nella lingua di destinazione [16].

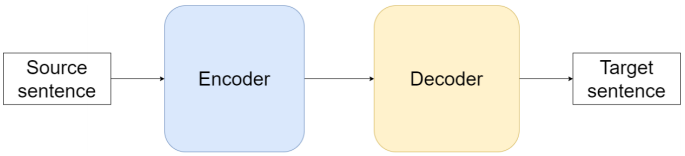


Figura 1.4: Esempio della struttura di un modello di traduzione automatica basato su reti neurali

Ottenuta da [17]

Nelle architetture più comunemente utilizzate l'*encoder* e il *decoder* sono entrambi reti neurali ricorrenti (RNN). Sotto questo punto di vista, LSTM e GRU sono molto diffuse per la loro capacità di gestire il *vanishing* e l'*exploding gradient* [17, 18, 19].

Un potenziale problema di questo approccio *encoder-decoder* è legato al fatto che una rete neurale debba essere in grado di comprimere tutte le informazioni necessarie in un vettore a lunghezza fissa. Questo può rendere difficile la gestione di frasi lunghe.

Modello Attention

Il modello *encoder-decoder* viene modificato aggiungendo un componente intermedio, chiamato *"attention"*, come metodo per allineare e tradurre, vedi Figura 1.5.

Nella traduzione automatica, come anticipato, l'allineamento ha l'obiettivo di identificare quali parti della frase di partenza siano rilevanti per ogni *token* in uscita. In tale contesto, la traduzione è interpretata come il processo che consiste nell'impiego delle informazioni rilevanti per la selezione dell'output appropriato [20]. Non si cerca più di codificare l'intera frase sorgente in un vettore a lunghezza fissa, ma si utilizzano tutte le informazioni del vettore locale, collettivamente, per decidere la sequenza successiva durante la decodifica della frase di destinazione. Questo porta il modello a prestare più attenzione alla parola corrente e da qui il nome *attention*.

In Figura 1.6 è rappresentato l'allineamento tra le parole in due diverse lingue relativamente a una frase di esempio. È fondamentale osservare come la predizione dell'i-esimo *token* in output possa necessitare anche di parti dell'input ulteriori rispetto al *token* in medesima posizione. Infatti, si può vedere come esso abbia prestato la giusta attenzione a "*European Economic Area*". In francese, l'ordine di queste parole (*zone économique européenne*) è invertito rispetto all'inglese.

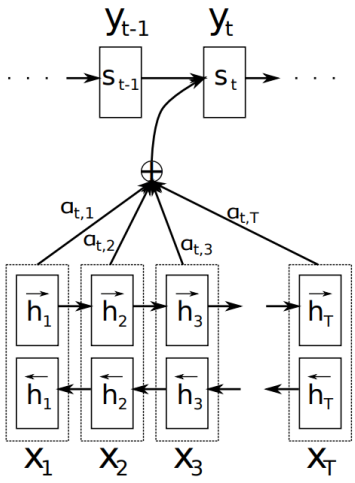


Figura 1.5: Illustrazione della generazione della t-esima parola data una frase di partenza del modello *attention*

Ottenuta da [20]

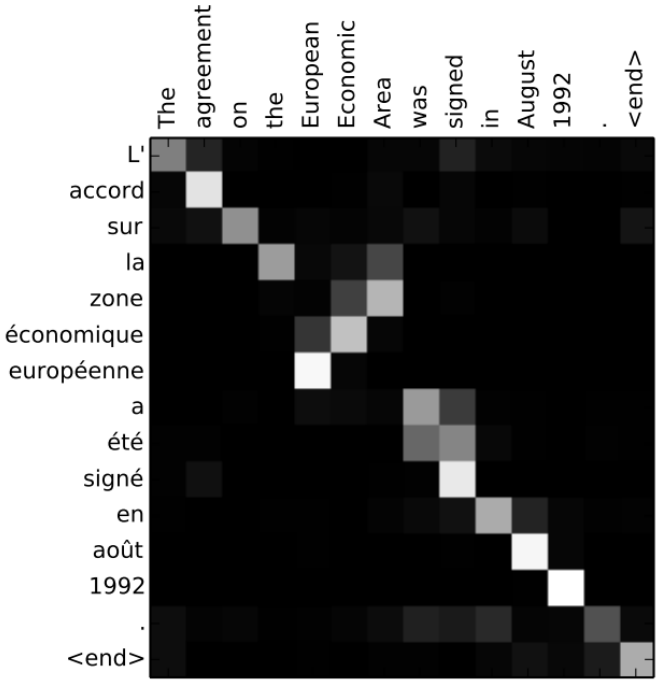


Figura 1.6: Interpretazione di una traduzione con *attention* dal francese all'inglese

Ottenuta da [20]

Modello Transformer

Il modello precedentemente illustrato non risolve completamente tutti i problemi: la natura sequenziale delle RNN comporta un ostacolo verso la parallelizzazione del processo. Nonostante la recente introduzione delle architetture LSTM e GRU, il vanishing gradient e la memoria a breve termine possono ancora costituire un problema per le RNN.

Ci sono tre tipi di dipendenze nello svolgimento di una traduzione: le dipendenze tra le parole di ingresso e di uscita, tra le parole di ingresso stesse, tra le parole di uscita stesse. Il modello *attention* (Sezione 1.1.4) ha risolto in gran parte la prima dipendenza dando al decoder l'accesso all'intera sequenza di ingresso, ma non le altre due.

Nel 2017 viene quindi presentato un nuovo modello, denominato *transformer*. Esso prende in considerazione anche i legami in ingresso e in uscita tra le parole stesse e ne permette l'elaborazione in parallelo. Il componente chiave del *transformer* è il blocco *Multi-Head Attention*. Il blocco di *multi-head attention* non fa altro che applicare il processo di "*attention*" a più blocchi in parallelo, concatenare le loro uscite e applicare una singola trasformazione lineare, come si vede in Figura 1.7.

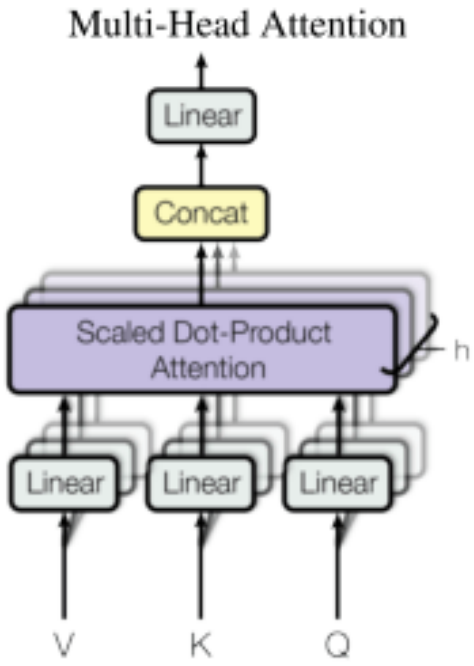


Figura 1.7: Struttura del processo di *multi-head attention*
Ottenuta da [21]

Per quanto riguarda il processo di *attention*, il *transformer* utilizza una particolare modalità chiamata *scaled dot-product attention*. A differenza della funzione di *attention* additiva comunemente usata, questa architettura utilizza una funzione di *attention* moltiplicativa. Anche se entrambe hanno la stessa complessità teorica, il *scaled dot-product attention* è stato scelto perché è molto più veloce ed efficiente [21].

Questa strategia consente di affrontare efficacemente il problema della *co-reference resolution* dove, ad esempio, si vuole individuare, in funzione del contesto, a quale sostantivo un'espressione (es. pronome) faccia riferimento all'interno di una frase. In Figura 1.8 si può vedere che viene correttamente identificato quando la parola "it" corrisponde a "animal" o a "street".

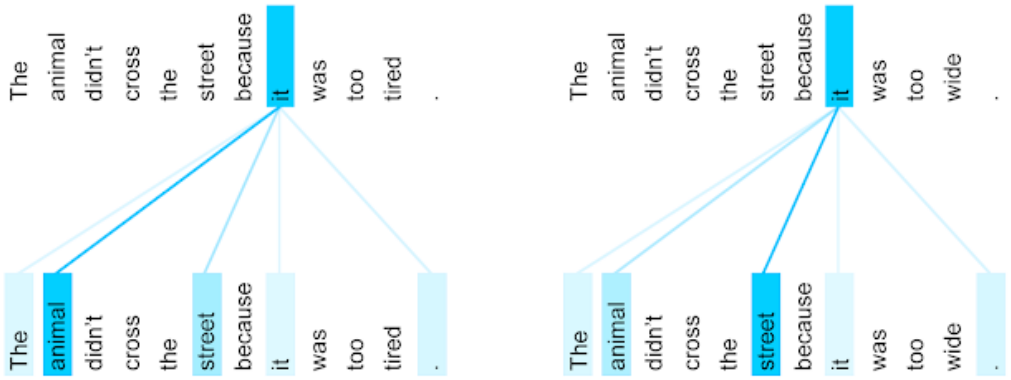


Figura 1.8: Esempio del uso di *attention* per identificare a quale sostantivo la parola "it" è riferita

Ottenuta da https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Evoluzione dell'uso delle reti neurali per la traduzione automatica

Prima dell'avvento dei *transformer*, una delle svolte nell'utilizzo delle reti neurali per la MT avviene nel 2014. Viene pubblicato un articolo [16] che dichiara un miglioramento delle prestazioni di un sistema di traduzione statistica utilizzando anche una RNN.

Successivamente nel 2016 Google[5] annuncia il progetto GNMT (Google Neural Machine Translation). Il modello di MT viene implementato utilizzando LSTM 8 *layer*, e il modello *attention*.

Uno dei più grandi problemi delle traduzioni basate su reti neurali è la gestione delle parole rare. Per risolverlo Google decide di suddividere le parole in un insieme limitato di unità formate da sotto-parole comuni. Questo metodo

[5] https://translate.google.it/

fornisce un buon equilibrio tra la flessibilità dei modelli basati sui "caratteri" e l'efficienza dei modelli basati sulle "parole", gestisce naturalmente la traduzione delle parole rare e migliora la precisione complessiva del sistema. Il progetto GNMT mostra fin da subito un'accuratezza equivalente ai sistemi di traduzione *phrase-based* esistenti, anche se allenato su corpora di dati di dimensioni ridotte [18].

Nel maggio del 2017, Facebook AI Research[6] annuncia il suo modo di implementare la traduzione neurale con CNN. Il sistema raggiunge risultati simili a quelli con RNN in tempi significativamente minori [22].

Un mese dopo Google rilasca un modello basato esclusivamente sul meccanismo *attention*, senza l'utilizzo di RNN o CNN.

A luglio dello stesso anno OpenNMT[7], un sistema di traduzione automatica neurale *open source* viene implementato [23].

Successivamente alcune grande aziende come Amazon, Microsoft, Baidu e molte altre iniziano a utilizzare metodi di traduzione neurale.

Vantaggi e svantaggi dei modelli neurali

La traduzione con reti neurali ha portato a grandi progressi negli ultimi anni. Dal punto di vista del punteggio di qualità della traduzione grezza le traduzioni neurali ottengono risultati migliori e sono generalmente più fluide. La ragione principale di questa fluidità è l'utilizzo del contesto completo di una frase e non solo del contesto immediato di poche parole prima e dopo la parola da tradurre. A volte non sono però corrette [24, 25, 26]. I modelli neurali non hanno quindi raggiunto un'indiscutibile superiorità rispetto ai modelli statistici e,in alcuni casi (lingue particolari, frasi troppo brevi, frasi troppo lunghe, varietà morfologica...), le traduzioni statistiche possono risultare migliori [27, 28, 29, 30].

1.2 Software di traduzione automatica

Sul mercato sono presenti molti sistemi di traduzione automatica. La maggior parte utilizzano almeno parzialmente tecniche di traduzione neurale. Alcuni dei software più diffusi sono: Google Translate, Bing Translator, Yandex, Amazon Translate, IBM Watson Language Translator, Baidu e Naver.

Nei paragrafi successivi alcuni software rilevanti sono analizzati. Tutti i programmi che seguono implementano almeno la traduzioni tra italiano e inglese.

[6] `https://ai.facebook.com/`
[7] `https://opennmt.net/`

1.2.1 Google Translate

Google Translate[8], o Google Traduttore in italiano, venne lanciato nel 2006. Originariamente incentrato sulla traduzione statistica, nel 2016 è stata annunciata l'introduzione di tecniche di reti neurali. Google translate supporta 108 lingue (09-2020) con tecnologia basata su reti neurali e il latino che è supportato dalla traduzione statistica *phrase-based*.

Oltre alla traduzione del testo offre anche la traduzione vocale, di immagini e di testo scritto a mano libera o con la tastiera virtuale [31]. AutoML Translation è un nuovo servizio di Google che permette di allenare un modello basato su reti neurali con coppie di parole particolari. Il costo risulta però tra i 45 e 300 dollari a seconda del numero di dati.

Tramite l'API i primi 500˙000 caratteri possono essere tradotti in modo gratuito, da 500˙000 a 1 miliardo di caratteri il costo è di 20 dollari per milione di caratteri. Superato il miliardo di caratteri è consigliato contattare un rappresentante Google.

Un esempio di traduzione effettuata con Google Translate si può vedere in Figura 1.9.

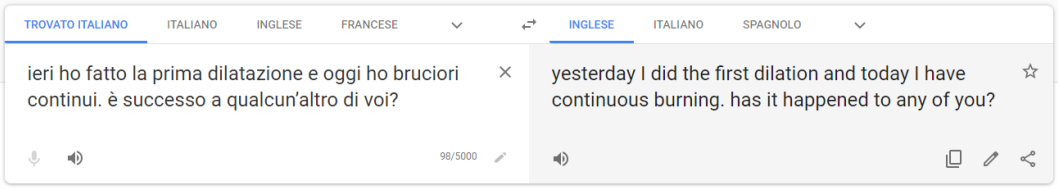


Figura 1.9: Esempio traduzione con Google Translate

1.2.2 Yandex Translate

Yandex[9] è una società russa che offre diversi servizi, tra i quali un motore di ricerca (il più popolare in Russia), un sistema di pagamento elettronico e un programma di MT.

Yandex Translate nasce nel 2011 con un sistema di traduzione automatica statistico, dal 2017 è stato implementato un nuovo sistema che utilizza sia un modello statistico che un modello neurale. Supporta circa 90 linguaggi, e il costo tramite l'API è di circa 6 dollari per milione di caratteri.

Un esempio di traduzione effettuata con Yandex Translate si può vedere in Figura 1.10.

[8] `https://translate.google.it/`
[9] `https://yandex.com/company/`

Figura 1.10: Esempio traduzione con Yandex Translate

1.2.3 IBM Watson Language Translator

IBM Watson Language Translator[10] è inizialmente progettato per rispondere alle domande del popolare programma televisivo Jeopardy, ora è anche un traduttore che supporta 52 linguaggi (08-2020) [32]. Watson Language Translator sfrutta la traduzione automatica neurale.

Un milione di caratteri possono essere tradotti gratuitamente, superato il limite il costo è di 2 dollari per milione di carattere. La maggior parte dei modelli di traduzione forniti possono essere estesi per imparare termini e frasi personalizzate o uno stile derivante dai dati. Il costo della traduzione personalizzata è pari a 8 dollari per milione di carattere.

Un esempio di traduzione effettuata con IBM Watson Language Translator si può vedere in Figura 1.11.

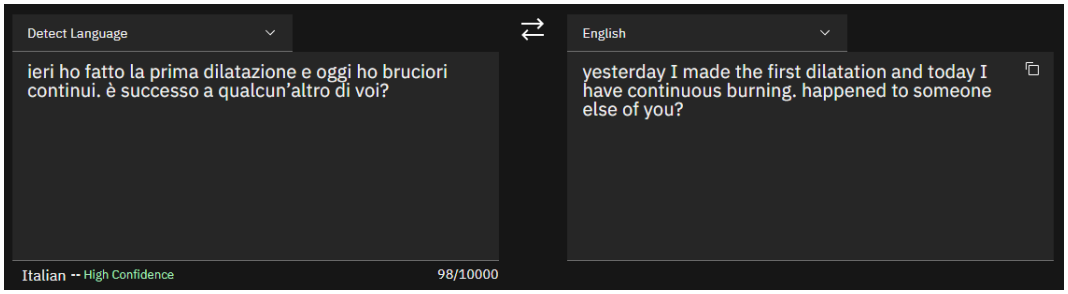


Figura 1.11: Esempio traduzione con IBM Watson Language Translator

1.2.4 Bing Translator

Bing Translator[11], anche noto come Microsoft Translator è un programma di MT dell'azienda Microsoft. Il traduttore è stato utilizzato dallazienda dal 2007 ed è stato reso disponibile tramite API per i clienti dal 2011. Dal 2018 Bing Translator implementa la traduzione basata sulla rete neurale LSTM con l'algoritmo *attention*. Il traduttore supporta più di 70 lingue per la traduzione

[10] `https://www.ibm.com/watson/services/language-translator/`
[11] `https://www.bing.com/translator`

di testi (09-2020) ed è usato in molti dei programmi Microsoft, alcuni degli esempi più noti sono Office, Skype e Cortana.

Il servizio *Custom Translator* consente agli utenti di personalizzare la traduzione per le lingue in cui la traduzione neurale è supportata. Questo è ottenuto tramite l'addestramento di un sistema di traduzione neurale che comprende la terminologia specifica di settore. Un'altra funzione interessante è la traduzione dal vivo di conversazioni anche in diverse lingue fino a un massimo di 500 persone.

Tramite l'API i primi 2 milioni di caratteri possono essere tradotti in modo gratuito, sorpassato il limite il costo è di 10 dollari per milione di caratteri. Sono anche disponibili degli sconti se si dovesse tradurre un volume elevato di dati.

Un esempio di traduzione effettuata con Bing Translator si può vedere in Figura 1.12.

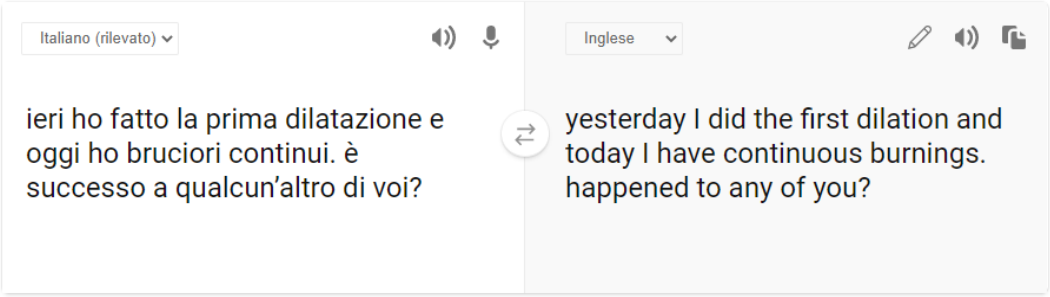


Figura 1.12: Esempio traduzione con Bing Translator

1.2.5 Marian

Marian, è un modello basato su reti neurali che può essere addestrato con una miscela di testi in una qualsiasi coppia di linguaggi [33]. Il sistema è stato creato nel 2018 dall'Università di Edinburgh e dall'Università Poznań, ed è attualmente utilizzato per i servizi di traduzione automatica neurale di Microsoft Translator.
Hugging Face[12] è un'azienda del 2016 che gestisce un elenco di modelli *NLP*. Uno di questi è **MarianMT** che offre il sistema Mariam già addestrato per molte coppie di linguaggi.

Un esempio di traduzione effettuata con MarianMT si può vedere in Figura 1.13.

[12] `https://www.huggingface.co/transformers/`

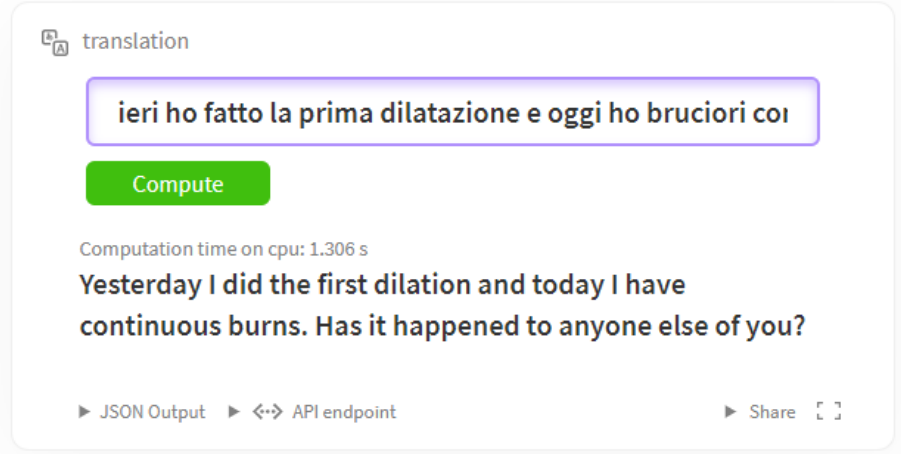


Figura 1.13: Esempio traduzione con MarianMT

1.2.6 DeepL

Il sistema di traduzione DeepL[13] è nato in Germania nel 2017 come startup innovativa partner di Linguee. L'azienda dichiara di aver prodotto il miglior sistema di MT al momento disponibile. Copre un numero relativamente ridotto di lingue, solo 11, ma tra queste ci sono l'italiano e l'inglese. Il sistema di traduzione si basa su reti neurali, ma non sono disponibili ulteriori informazioni [31].

Al costo di 6 euro al mese è possibile iscriversi all'abbonamento base di DeepL. E quindi possibile tradurre 5 file e anche variare la forma della traduzione. Per le lingue con termini di indirizzo formali e informali, questa funzione è particolarmente utile.

Un esempio di traduzione effettuata con DeepL si può vedere in Figura 1.14.

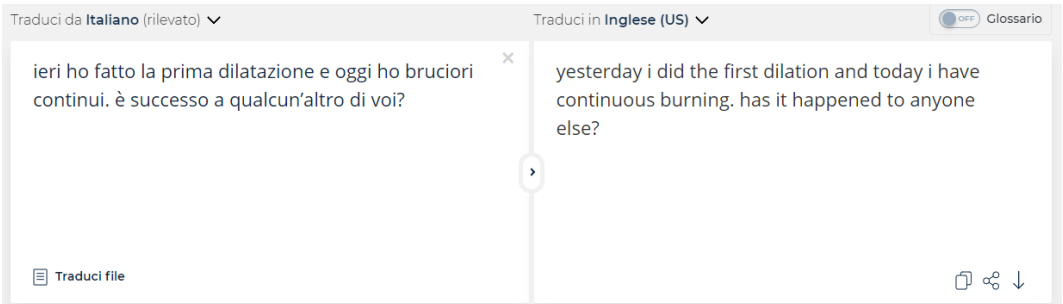


Figura 1.14: Esempio traduzione con DeepL

[13] `https://www.deepl.com/translator`

1.2.7 Amazon Translate

AWS[14] (Amazon Web Services) è la piattaforma di *cloud computing* (la distribuzione di servizi di calcolo, come server, database risorse di archiviazione, tramite Internet) più utilizzata del mondo.

Uno dei principali motivi della sua popolarità è l'opportunità per le aziende di eliminare gli investimenti iniziali in infrastrutture e l'elasticità che il servizio offre. Il costo finale dipende dall'utilizzo effettivo ed possibile aumentare o diminuire le risorse in base ai carichi di lavoro e alla domanda. Fondata nel 2006, al momento AWS offre oltre 175 prodotti, uno dei quali è Amazon Translate.

Amazon Translate è un servizio di traduzione automatica neurale che supporta 55 lingue. Registrandosi al sito di AWS si accede alla possibilità di tradurre 2 milioni di caratteri gratuitamente per 12 mesi. Il costo è altrimenti pari a 15 dollari per milione di carattere.

Amazon Translate offre anche la possibilità di creare un dizionario con terminologia personalizzata per assicurarsi che certi contenuti vengano tradotti esattamente come desiderato, a prescindere dal contesto e dalla decisione dell'algoritmo Amazon Translate.

Un esempio di traduzione effettuata con Amazon Translate si può vedere in Figura 1.15.

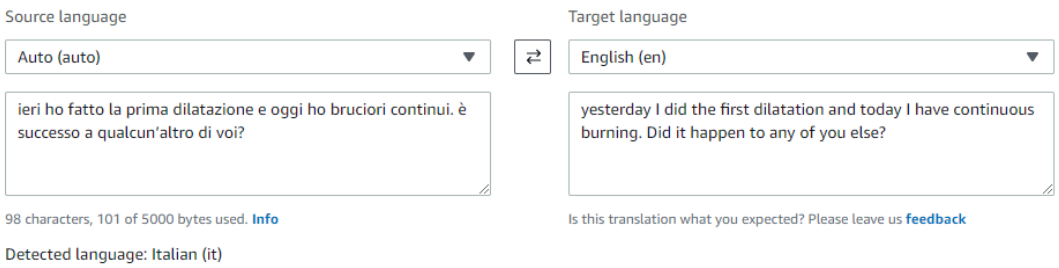


Figura 1.15: Esempio traduzione con Amazon Translate

1.3 Metodi di valutazione delle traduzioni automatiche

Il ruolo sempre più importante dei sistemi di traduzione automatica, utilizzati spesso come principale modalità di traduzione, ha reso necessaria l'abilità di giudicare la qualità delle traduzioni. Sono presenti due alternative: la valuta-

[14] `https://aws.amazon.com/`

zione umana, e l'utilizzo di un algoritmo che calcoli un punteggio indicativo della qualità della traduzione.

Gli algoritmi che valutano la qualità della traduzione si distinguono generalmente in *glass-box* e *black-box*. La valutazione *glass-box* misura la qualità di un sistema sulla base delle sue proprietà interne. Data l'impossibilità di procedere in questo modo con la maggior parte dei software da analizzare, tutte le metriche più popolari utilizzano un algoritmo di tipo *black-box*, richiedente la sola traduzione per il calcolo del punteggio [34]. Per essere utile, una metrica per la valutazione delle traduzioni automatiche deve soddisfare diversi criteri. Il requisito primario è che la metrica deve assegnare punteggi di qualità correlati con il giudizio umano. Inoltre, una buona metrica dovrebbe essere il più possibile sensibile alle differenze di qualità tra i diversi sistemi e tra le diverse versioni di uno stesso sistema. La metrica dovrebbe essere coerente (lo stesso sistema di traduzione automatica su testi simili dovrebbe produrre punteggi simili), affidabile e generale (applicabile a diversi compiti di MT in una vasta gamma di domini e scenari). Soddisfare tutti i criteri è estremamente difficile. Tuttavia, tali requisiti possono stabilire uno standard complessivo che consenta di confrontare diverse metriche di valutazione [35].

Alcune delle metriche di fatto utilizzate sono BLEU, NIST, METEOR, ROUGE, WER e TER. Tutte hanno bisogno di dati etichettati per confrontare il risultato della traduzione e fornire i punteggi di confronto [36]. NIST e METEOR si basano su BLEU, mentre WER e TER vengono utilizzate per misurare quanto contributo umano sia necessario a seguito della traduzione al fine di rilasciare un risultato che gli esseri umani possano considerare di buon livello.

La cronologia della divulgazione delle principali metriche in Figura 1.16.

1.3.1 BLEU

Il compito principale dell'algoritmo BLEU è quello di confrontare quante sequenze di parole (tecnicamente, n-grammi) della traduzione automatica si ritrovano nella traduzione umana di riferimento dello stesso testo. In Figura 1.17 un esempio di come diversi n-grammi vengono considerati. In rosso gli n-grammi non trovati nella traduzione di riferimento, in viola le singole parole presenti, in azzurro le coppie di parole e in verde i gruppi in cui almeno 4 parole sono uguali a quelle della frase di riferimento.

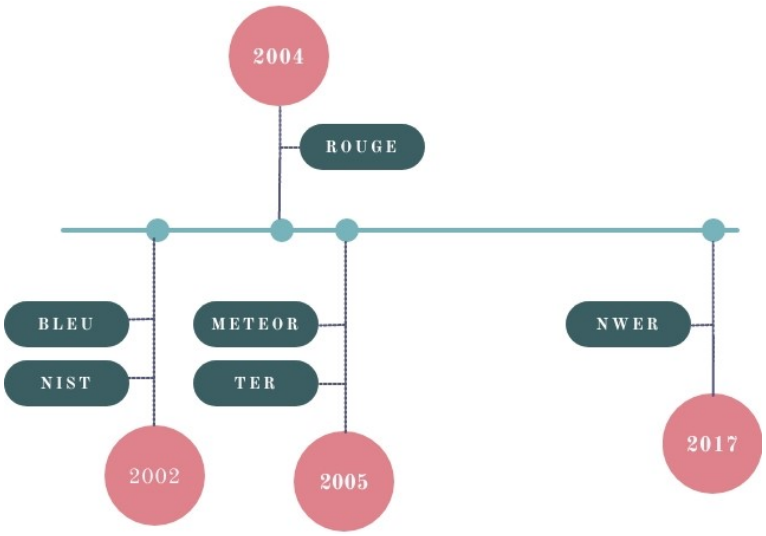


Figura 1.16: Cronologia della divulgazione delle principali metriche per la valutazione della traduzione automatica

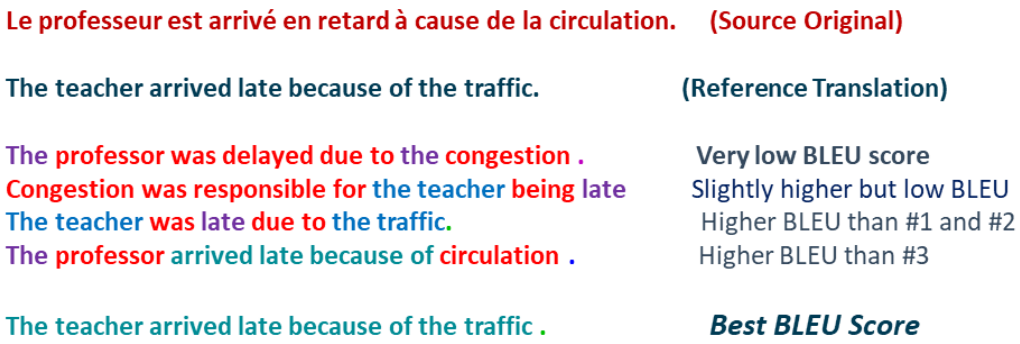


Figura 1.17: Esempio BLEU
Modificata da https://blog.sdl.com/blog/understanding-mt-quality-bleu-scores.html

Più il testo è simile a quello umano, migliore è il punteggio assegnato. Il punteggio assegnato da BLEU va da 0 a 1. Poche traduzioni raggiungono un punteggio di 1, incluse quelle effettuate da esperti, che di solito si sono comprese nel range $[0.3, 0.6]$ [37, 18]. È importante notare che incrementare il numero delle traduzioni di riferimento porta un aumento anche nel punteggio assegnato.

In genere, ci sono molti possibili traduzioni corrette a partire da una data frase di partenza, queste traduzioni possono variare nelle parole utilizzate o nell'ordine in cui vengono espresse. Il confronto effettuato da BLEU avviene quindi indipendentemente dalla posizione e utilizzando, se possibile, anche più

traduzioni di riferimento.

Uno dei maggiori problemi di BLEU è che non tiene conto di parafrasi o sinonimi, i punteggi possono quindi essere fuorvianti. Ad esempio, "rapido" non ottiene credito parziale per "veloce". I sistemi che utilizzano tecnologie basate su reti neurali, possono generare eccellenti traduzioni che sono diverse dal riferimento e quindi hanno un basso punteggio [36, 37].

BLEU è stata la prima metrica automatica a dimostrare una correlazione generale con i giudizi umani sulla qualità della traduzione. In quanto tale, BLEU è stata utilizzata come base per molte altre valutazioni di traduzione automatica. Con la crescita della popolarità delle metriche di valutazione, BLEU è diventata rapidamente quella più utilizzata [38].

Per calcolare il punteggio BLEU di una frase è necessario determinare BP (*Brevity Penalty*). Data la lunghezza della traduzione di riferimento r e la lunghezza della traduzione da valutare c, BP è definita come riportato in Equazione 1.2.

$$BP = \begin{cases} 1 & \text{if c} > \text{r} \\ e^{1-r/c} & \text{if c} \leq \text{r} \end{cases} \tag{1.2}$$

Successivamente è necessario determinare p_n, la media geometrica delle precisioni modificate degli n-grammi e w_n i pesi positivi che sommano a uno, infine il punteggio BLEU viene ottenuto come mostrato in Equazione 1.3 [37].

$$BLEU = BP * exp\left(\sum_{n=1}^{N} w_n log \rho_n\right) \tag{1.3}$$

1.3.2 NIST

La metrica NIST (prende il nome dal *National Institute of Standards and Technology* che l'ha inventata) è una metrica di traduzione automatica progettata per migliorare BLEU premiando la traduzione di parole più informative. Il punteggio assegnato da NIST va da 0 a 10.

L'obiettivo è quello di prevenire l'inflazione dei punteggi di valutazione a causa di parole comuni o traduzioni facili. Di conseguenza, la metrica NIST assegna un peso maggiore alle parole più rare. Ad esempio, la sequenza di termini "*of the*" riceverà un peso inferiore rispetto a "*particular branches*", in quanto è meno probabile che ciò si verifichi.

L'affidabilità e la qualità della metrica NIST si è dimostrata superiore alla metrica BLEU in molti casi [39].

1.3.3 METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) è una metrica di valutazione basata su BLEU che cerca di migliorarne le prestazioni. Il punteggio assegnato da METEOR va da 0 a 1.

BLEU non tiene conto direttamente della *recall* (la proporzione degli n-grammi giusti sul numero totale di n-grammi nella traduzione di riferimento) La *recall* è estremamente importante per valutare la qualità della traduzione automatica, poiché riflette in che misura la traduzione rispecchia il contenuto della frase tradotta. Un altro problema è la mancanza di correlazione esplicita tra i termini presenti nell'output di un modello di MT e quelli contenuti nella rispettiva traduzione di riferimento. Questo può portare a conteggiare corrispondenze errate, in particolare per le parole di uso comune.

METEOR valuta quindi una traduzione calcolando un punteggio in base alle corrispondenze esplicite di ogni parola tra l'output effettivo e quello desiderato. Se sono disponibili più traduzioni di riferimento, la traduzione viene valutata indipendentemente con ogni riferimento e viene riportato il punteggio migliore.

METEOR non si limita a supportare la corrispondenza tra le parole che sono identiche nelle due stringhe confrontate, ma abbina anche i sinonimi e le parole che sono semplici varianti morfologiche l'una dell'altra (cioè hanno la prima parte identica).

Questo permette a METEOR di raggiungere mediamente una correlazione più alta con il giudizio umano [35].

1.3.4 WER

WER (Word Error Rate) cerca la sequenza più breve di operazioni di modifica necessarie per trasformare la traduzione candidata in una delle traduzioni di riferimento. WER è semplice da calcolare, efficiente e riproducibile. Al contrario della maggior parte delle altre metriche, che misurano la qualità delle traduzioni, WER misura gli errori da risolvere e una traduzione viene ritenuta migliore se ha un punteggio minore.

1.3.5 TER

TER (Translation Error Rate) è un'evoluzione di WER. Alle modifiche consentite (inserimento, cancellazione e sostituzione di singole parole) si aggiunge lo scambio di una sequenza di parole adiacenti. Questa nuova operazione evita che le frasi tradotte correttamente, ma da riordinare, vengano penalizzate eccessivamente [40].

1.3.6 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) è un insieme di metriche utilizzate per la valutazione della MT e altri task NLP:

- ROUGE-N: misura il numero di n-grammi (es. 1 o 2) che corrispondono tra la traduzione e il riferimento;
- ROUGE-L (Longest Common Subsequence): misura la sequenza più lunga di parole corrispondenti;
- ROUGE-W (Weighted Longest Common Subsequence): misura il numero maggiore di parole consecutive con una corrispondenza nella frase di riferimento;
- ROUGE-S (Skip-Bigram Co-Occurrence Statistics): misura il numero di *skip-bigram* corrispondenti. Uno *skip-bigram* è una qualsiasi coppia di parole in ordine in una frase, non considerando gli spazi. Per esempio "non domani" nella frase "non tornare domani";
- ROUGE-SU (Skip-bigram plus Unigram-based Co-Occurrence statistics) è uguale a ROUGE-S, ma considera anche le singole parole.

Il punteggio assegnato va da 0 a 1 [41].

1.3.7 MEWR

MEWR (Machine translation Evaluation Without Reference text) è una delle uniche metriche che non ha bisogno di una traduzione di riferimento. Questo è un grande vantaggio perché spesso traduzioni di riferimento non sono disponibili o sono molto costose da ottenere o produrre.

MEWR ha uno stretto legame sia con BLEU che con la valutazione umana. La valutazione di un testo si basa sulla fedeltà, sulla fluidità e sulla qualità generale del testo. La fedeltà è la misura di quanto una traduzione sia conforme alla sua fonte. La fluidità valuta invece la naturalezza del testo.

Uno dei problemi da affrontare è che il tempo impiegato per valutare i testi è piuttosto lungo [42], inoltre dalla sua nascita nel 2017 non sembra essere più stata utilizzata.

1.3.8 Test delle metriche

Questa sezione riporta alcuni esempi pratici legati al caso di studio affrontato nella tesi (vedi Capitolo 2), con l'intento di chiarire il funzionamento e l'applicazione di alcune delle metriche descritte. Le frasi utilizzate per i test sono in Tabella 1.1.

Traduttore	BLEU
Frase in italiano	ieri ho fatto la prima dilatazione e oggi ho bruciori continui. è successo a qualcun'altro di voi?
Primo Riferimento	I had my first dilation yesterday and today I have a continuous burning sensation. did it happen to any of you?
Secondo Riferimento	yesterday I did the first dilation and today I have continuous burnings. has it ever happened to someone else?
Traduzione Google	yesterday I did the first dilation and today I have continuous burning. has it happened to any of you?

Tabella 1.1: Frasi per la sperimentazione delle metriche di traduzione automatica

BLEU

Per prima cosa viene calcolato BP, una penalità per evitare che le traduzioni troppo brevi ricevano punteggi esageratamente alti, considerando la prima frase di riferimento (con 21 parole) e la traduzione Google (con 19 parole) $BP = e^{1-r/c} = e^{1-21/19} = 0.90$.

Per procedere bisogna scegliere un valore N, solitamente N = 4 in modo da prendere in considerazione le corrispondenze tra gli unigramm, i bigrammi, i trigrammi e i quadrigrammi.
Calcolare ρ_n, è significativamente più complicato ed indicato in Equazione 1.4.

$$\rho_n = \frac{\text{numero di n-grammi corrispondenti nelle frasi di riferimento e da valutare}}{\text{numero di n-grammi nella traduzione da valutare}} \tag{1.4}$$

In questo caso $\rho_4 = \frac{2}{16} = 0.125$ (i 2 quadrigrammi trovati sono "and today I have", "to any of you" sul numero totale di quadrigrammi, 16), $\rho_1 = \frac{15}{19} = 0.79$, $\rho_2 = \frac{8}{18} = 0.44$, $\rho_3 = \frac{4}{17} = 0.23$.
Infine $w_n = N^{-1} = N^{-4} = \frac{1}{4} = 0.25$ e si può calcolare il risultato (Equazione 1.5) [37].

$$\begin{aligned} BLEU = BP * exp\left(\sum_{n=1}^{N} w_n log \rho_n\right) = \\ = 0.9 * exp\left(\frac{log 0.79 + log 0.44 + log 0.23 + log 0.13}{4}\right) = \mathbf{0.55} \end{aligned} \tag{1.5}$$

NIST

Il valore NIST viene calcolato con lo stesso processo di BLEU aggiungendo però un peso per ogni parola in base alla sua rarità e cambiando la formula di BP. Ad esempio, la parola "I" appare due volte nella traduzione da valutare, gli verrà quindi assegnato un peso minore. Per calcolare ρ_n non si somma il numero di parole corrispondenti (come in BLEU), ma ogni parola può avere un peso diverso, la parola "I" in questo esempio avrà importanza minore. Il nuovo BP viene calcolato come illustrato in Equazione 1.6.

$$BP = exp\left(0.5 * log\left(min\left(\frac{\text{parole traduzione da valutare}}{\text{parole traduzione di riferimento}}, 1.0\right)\right)\right) \quad (1.6)$$

Il punteggio NIST finale per le frasi in analisi è **3.46**, ma non molto significativo dato che è necessario usare un insieme di dati più grande per individuare le parole più e meno rare [43].

METEOR

Utilizzando la prima traduzione di riferimento le corrispondenze trovate sono: "i had my first dilatation" con "I did my first dilatation", "and today I have" con "and today I have", "to any of you" e "to any of you". Da notare come METEOR non abbia bisogno di una corrispondenza perfetta.
Per calcolare METEOR è necessario calcolare precision (Equazione 1.7), recall (Equazione 1.8) e Fmean (Equazione 1.9).

$$precision = \frac{\text{numero parole corrispondenti}}{\text{numero parole traduzione}} = \frac{13}{19} = 0.68 \quad (1.7)$$

$$recall = \frac{\text{numero parole corrispondenti}}{\text{numero parole riferimento}} = \frac{13}{21} = 0.62 \quad (1.8)$$

$$Fmean = \frac{10 * p * r}{9 * p + r} = 0.626 \quad (1.9)$$

Per favorire le sequenze più lunghe si calcola una penalità come segue in Equazione 1.10 (sono state accoppiate 3 sequenze con un totale di 13 parole).

$$Penalty = 0.5 * (\frac{3}{13})^3 = 0.0061 \quad (1.10)$$

Concludendo $METEOR = Fmean * (1 - DF) = 0.626 * (1 - 0.0061) = \mathbf{0.62}$.

I numeri che appaiono nei calcoli senza spiegazione sono parametri che possono essere cambiati per massimizzare la correlazione con diverse nozioni di giudizio umano: 10 e 9 nell'Equazione 1.9 specificano il peso che si vuole dare al valore *precision* rispetto alla *recall*, 0.5 nell'Equazione 1.10 rappresenta il massimo valore che la penalità può assumere [35].

WER

Per ottenere la seconda traduzione di riferimento è necessario sostituire la parola "*burning*" con "*burnings*", aggiungere la parola "*ever*", trasformare "*someone*" in "*any*" e eliminare la parola "*of*".

Dato il numero totale di parole nel riferimento (19) e il numero di trasformazioni (4), $WER = \frac{4}{19} = 0.21$. Prendendo la prima frase di riferimento $WER = \frac{8}{21} = \mathbf{0.38}$.

TER

Il punteggio TER per la seconda frase di riferimento è uguale a quello WER. Prendendo la prima frase di riferimento si può però spostare la parola "*yesterday*", evitando una modifica e avendo una misura più corretta degli errori presenti nella traduzione $TER = \frac{7}{21} = \mathbf{0.33}$.

ROUGE-L

Per prima cosa si calcola LCS (longest matching sequence), prendendo in considerazione la prima frase di riferimento. Le parole uguali nelle due frasi sono state segnalate in arancione.
Primo rifetimento: I had my first dilation yesterday and today I have a continuous burning sensation. did it happen to any of you?
Traduzione Google: yesterday I did the first dilation and today I have continuous burning. has it happened to any of you?

LCS risulta quindi essere 13. ROUGE-L calcola due parametri: *precision* (Equazione 1.11) e *recall* (Equazione 1.12).

$$precision = \frac{\text{LCS}}{\text{numero parole traduzione}} = \frac{13}{19} = 0.68 \tag{1.11}$$

$$recall = \frac{\text{LCS}}{\text{numero parole riferimento}} = \frac{13}{21} = 0.62 \tag{1.12}$$

L'output che di solito viene utilizzato si chiama *F-score* e viene calcolato come in Equazione 1.13.

$$Fscore = 2 * \frac{precision * recall}{precision + recall} = \mathbf{0.65} \tag{1.13}$$

1.4 Gestione di entità di dominio

Una *named entity* (NE) è un oggetto del mondo reale corrispondente ad un identificazione specifica di persone, luoghi, organizzazioni, ecc. Le entità di

dominio e i nomi propri sono un serio problema per i sistemi di traduzione automatica, ad esempio, "policlinico gemelli" è erroneamente tradotto in "twins polyclinic".

La traduzione di queste entità è scomoda a causa delle loro caratteristiche peculiari. Il numero di parole utilizzate nella costruzione di NE è potenzialmente infinita e non possono essere trattate come le altre parole. Le entità devono essere individuate e poi deve essere deciso se possono essere solamente trascritte (es. Giulia Rossi) o se richiedono un mix di traslitterazione e traduzione (es. ospedale pediatrico Bambino Gesù). Inoltre, non possono essere memorizzate dai modelli di MT a causa della loro unicità. Di conseguenza, la loro traduzione richiede approcci e metodi diversi [44].

L'individuazione delle entità è difficile, in quanto i nomi non hanno confini esatti. Spesso possono essere semanticamente ambigui, per esempio la parola "gemelli" in "Politecnico Gemelli" non significa fratelli nati dallo stesso parto, né indica il segno zodiacale, ma è una NE [45]. Vengono utilizzati diversi metodi per individuare le entità. I primi sistemi NER (Named entity recognition) erano basati su regole manuali, glossari e ontologie. Questi sistemi sono stati seguiti da quelli basati su *feature-engineering* e *machine learning*. Successivamente sono nate le prime architetture di reti neurali, costruite a partire da caratteristiche ortografiche (ad esempio, capitalizzazione del primo carattere), dizionari e glossari. Infine gli ultimi modelli sono basati su RNN che utilizzano architetture *word level*, *character level* o combinate [46, 47]. Alcuni programmi che eseguono il processo di NER sono: Spacy[15], TextRazor[16] e GATE[17].

Per tradurre una NE sono disponibili vari metodi. Uno si basa sugli anchor text, il testo cliccabile in un link. Le entità vengono tradotte recuperando un elenco di documenti web nella lingua di destinazione, estraendo gli *anchor text* e trovando la migliore traduzione, partendo dal presupposto che gli *anchor text* in diverse lingue che si ripetono frequentemente e che si verificano raramente separatamente sono probabilmente traduzioni l'uno dell'altro. Il vantaggio degli *anchor text* è il fatto che sono spesso una descrizione succinta della pagina web di destinazione. Avere una grande quantità di testo introduce molti altri candidati non corretti, il che rende più difficile recuperare la traduzione corretta [48]. Un altro popolare metodo consiste nella consultazione della versione inglese dell'articolo di Wikipedia italiano sull'entità. Il suo titolo è considerato la traduzione di base. Altre potenziali traduzioni possono venir estratte dal testo dell'articolo in inglese, se necessario. Successivamente si deciderà la traduzione migliore.

[15] `https://spacy.io/`
[16] `https://www.textrazor.com/`
[17] `https://gate.ac.uk/`

Capitolo 2

Motivazioni e Caso di studio

Questo capitolo analizza i problemi che sorgono dall'utilizzo esclusivo della lingua italiana nelle soluzioni NLP e l'importanza che la traduzione automatica di testi può avere se applicata in campo medico.

Il NLP è vitale nell'estrazione di informazioni da testi. Una delle tante applicazioni di queste tecnologie è in campo medico. Queste informazioni possono essere utilizzate per aiutare il processo decisionale dei fornitori di assistenza sanitaria e dei pazienti. fornendo informazioni facilmente accessibili, nel momento in cui sono necessarie [49].

Questa possibilità risulta particolarmente utile nell'ambito delle malattie rare. In molti casi un singolo medico può non aver mai visto un paziente con tale disturbo e non poterlo aiutare. La variabilità di queste malattie rende particolarmente difficile il riconoscimento e la diagnosi. Essere in grado di accedere a un insieme organizzato di informazioni può portare a grandi miglioramenti nel campo.

Il caso di studio affrontato in questa tesi prende in analisi l'Acalasia Esofagea, un raro disturbo dell'esofago, sviluppandosi a partire dal dataset costruito da Giacomo Frisoni in "A new unsupervised methodology of descriptive text mining for knowledge graph learning" [4].

2.1 NLP e modelli multilingua

NLP (Natural Language Processing) è una branca dell'intelligenza artificiale che permette ai computer di interpretare e manipolare il linguaggio naturale (umano). Per i computer, è un lavoro estremamente difficile, in quanto il linguaggio umano è incredibilmente complesso e vario.

Ci esprimiamo in infiniti modi, sia verbalmente che per iscritto. Oltre al numero di lingue disponibili, ogni lingua è un insieme unico di regole grammati-

cali e di sintassi. Un testo scritto può contenere errori ortografici, abbreviazioni o punteggiatura inesatta.

L'origine della materia non è ben definita, ma è stata influenzata da altre discipline. Le principali contribuenti sono:

- Linguistica — lo studio scientifico del linguaggio, compresa la grammatica, la semantica e la fonetica;
- Linguistica computazionale — lo studio moderno della linguistica utilizzando gli strumenti dell'informatica;
- Informatica — si occupa della rappresentazione e dell'elaborazione delle informazioni rilevanti;
- Psicologia cognitiva — lo studio dell'uso del linguaggio in base ai processi cognitivi umani [50].

Con lo sviluppo del *deepl learnig*, varie reti neurali, principalmente RNN, CNN e graph-based neural networks (GNN) sono state ampiamente utilizzate per risolvere compiti di NLP [51].

2.1.1 Task NLP

Alcune dei principali compiti trattati sono spiegati in seguito. Il processo di traduzione automatica e un task NLP ma non viene incluso in questa lista essendo stato ampiamente trattato nel Capitolo 1. Per i motivi descritti nella successiva Sezione 4.1, NER e NEL sono due task centrali in questa tesi e vengono trattati più approfonditamente.

Part of speech tagging

Part of speech tagging consiste nella caratterizzazione di una parola in un testo con un tag identificativo di una particolare parte del discorso (es. nome), in base sia alla sua definizione che al suo contesto. La principale difficoltà nel processo è l'individuazione corretta delle parole ambigue [52].

Per esempio, la parola "marca" è un sostantivo nella frase "ho comprato un profumo di marca", ma è un verbo nella frase "il difensore marca l'attaccante".

Lemmatization

Lemmatization, o lemmatizzazione in italiano si occupa di ridurre ogni parola alla sua forma di base, nota anche come "lemma", in modo che possa essere analizzata come un unico elemento. Una tipica trasformazione coinvolge il

passaggio dal tempo passato alla forma base del verbo (es. camminavo in camminare). Un'altra modifica consiste nel riportare tutti gli aggettivi nella loro forma positiva al maschile singolare (es. alte in alto).

Tokenization

Il compito di *tokenization* consiste nel separare un testo in unità linguistiche minime chiamate *token*, che verranno elaborate successivamente. Un *token* deve essere linguisticamente significativo e utile.

Mentre in lingue come l'italiano e l'inglese (dove più comunemente la separazione viene fatta in corrispondenza di spazi bianchi) non è un compito particolarmente difficile, in lingue come il cinese è molto più complicato [53]. Da notare però che anche in inglese e italiano due parole sparate possono formare un unico *token* (es. New York).

Coreference resolution

Coreference resolution consiste nel trovare tutte le espressioni che si riferiscono alla stessa entità in un testo. È un passo importante per molti compiti NLP che implicano la comprensione del linguaggio naturale, come la *summarization* dei documenti, la risposta alle domande e l'estrazione di informazioni [54]. Un esempio di *coreference resolution* nella frase "Emma ha ricevuto il suo regalo" è individuare che la parola "suo" si riferisce a "Emma".

Sentiment analysis

Sentiment analysis, chiamato anche *opinion mining*, è il campo di studio che analizza le opinioni e i sentimenti di una persona sull'argomento in discussione. Un orientamento semantico positivo implica desiderabilità (es. "onesto", "bello") e un orientamento semantico negativo implica indesiderabilità (es. "inquietante", "superfluo") [55].

Named entity recognition

NER (*Named entity recognition*) è una tecnica di identificazione e categorizzazione delle informazioni chiave (entità) presenti in un testo. Un'entità può essere qualsiasi parola o serie di parole che si riferiscono costantemente alla stessa cosa. Ogni entità rilevata è classificata in una categoria predeterminata come per esempio nomi di persone, località e organizzazioni [56]. Nell'esempio in Figura 2.1 è stato rilevato il nome di un'ospedale composto da due parole, il nome di una città e di un alloggio.

Figura 2.1: Esempio processo NER

Named entity linking

NEL (*Named entity linking*) è l'assegnazione di un'identità univoca alle entità menzionate nel testo. Il processo di NEL si distingue dalla classificazione delle entità (parte del procedimento di NER) in quanto solamente il primo riesce ad identificare quale entità specifica è descritta. Uno dei modi più diffusi di ottenere questo risultato si basa sull'utilizzo di conoscenza derivata da Wikipedia[1] e l'associazione all'entità della pagina corrispondente tramite l'utilizzo di DBpedia[2]. Nell'esempio in Figura 2.2 le entità rilevate tramite i processi di NER sono state collegate ad un identificatore univoco.

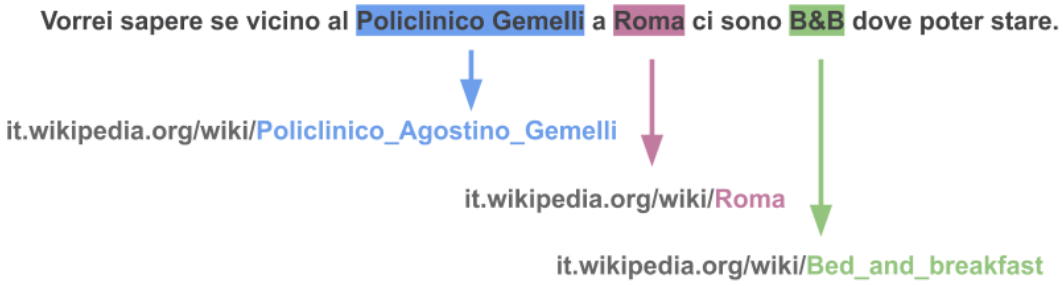


Figura 2.2: Esempio processo NEL

2.2 Problemi della lingua italiana

In seguito si propone una panoramica di alcuni dei principali strumenti NLP, ponendo particolare attenzione alla loro disponibilità in italiano e inglese.

Hugging Face

Hugging Face è un'azienda del 2016 che gestisce un elenco di modelli NLP estremamente popolari in vari task tra cui traduzione, risposta a domande, riassunto testi. Sul sito, al momento della scrittura (09-2020) sono disponibili

[1] `https://it.wikipedia.org/wiki/`
[2] `https://wiki.dbpedia.org/`

547 modelli in inglese, ma solo 52 in italiano, di questi 39 si occupano di traduzione[3].

GPT-3

GPT-3 (Generative Pre-trained Transformer 3) dell'azienda OpenAI[4] GPT-3 è stato addestrato con un modello *autoregressive* con 175 miliardi di parametri, questo lo rende il più grande modello del suo genere [57]. Le potenzialità di GPT-3 sono molto vaste, forniti i giusti esempi si può chiedere a GPT-3 eseguire traduzioni, programmare, scrivere poesie e altro ancora.

Dato che uno dei punti forza di GPT-3 è la quantità dei dati utilizzati per addestrarlo, viene fatta un'analisi dei siti indicati in Figura 2.3 utilizzati appunto per il *training*.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Figura 2.3: Siti dal quale sono stati presi i dataset utilizzati nel processo di training di GPT-3 e le loro percentuali di utilizzo
Ottenuta da [57]

Common Crawl[5] è un'organizzazione che si dedica a fornire una copia di Internet a ricercatori, aziende e singoli individui senza alcun costo a scopo di ricerca e analisi. Il sito contiene pagine web in 160 diverse lingue, in 43% dei casi il linguaggio principale è l'inglese, mentre nel 2% è l'italiano (09-2020)[6].

WebText2 è un raccolta di pagine web come Common Crawl, ma essendo stato creato da OpenAI (stessa azienda propietaria di GPT-3) i testi o le informazioni sul linguaggio dei dati non sono disponibili [3].

Books1 e Books2 sono semplicemente dei gruppi di libri disponibili al pubblico. Mentre risulta difficile analizzare il numero di libri disponibili in un certo linguaggio, si può notare come il numero di libri pubblicati in Italia nel 2018

[3] `https://huggingface.co/languages`
[4] `https://openai.com/`
[5] `https://commoncrawl.org/`
[6] `https://commoncrawl.github.io/cc-crawl-statistics/plots/languages`

sia pari a 75˙000 [58], mentre il numero di libri pubblicati negli Stati Uniti è stato circa 700˙000 nel 2016[7].

Wikipedia[8] è un'enciclopedia online gratuita, estremamente popolare e curata da volontari in tutto il mondo. il numero di pagine in italiano è 1˙635˙060 mentre quello di pagine in inglese è 6˙158˙519 (09-2020)[9].

Si può quindi notare come la quantità di dati disponibili in inglese sia decisamente superiore a quella in italiano. Molti dei principali tool NLP sono stati creati sulla base della lingua inglese e, mentre alcune soluzioni si possono trovare in varie lingue, la quantità di risorse disponibili nel linguaggio inglese non è paragonabile a quella per l'italiano.

Lo sviluppo dei metodi NLP si basa sempre più su approcci basati sui dati che aiutano a costruire modelli più potenti e robusti. Le reti neurali profonde di solito hanno un gran numero di parametri, con piccoli insiemi di dati il processo di *training* risulta difficile.

2.3 Il contesto delle malattie rare

Una malattia è definita *rara* in Europa se ha una prevalenza inferiore a 5 persone su 10˙000, secondo i dati riportati nel database Orphanet[10], progetto europeo nato nel 2000 che fornisce informazioni su farmaci orfani e sulle malattie rare sia per medici che per pazienti. Al momento della scrittura sono registrati 6172 disturbi rari, ma questo numero è in costante crescita e se ne stimano sino a 8000. Mentre la parte della popolazione affetta da ogni singola malattia è ridotta, 400 milioni di persone (5% della popolazione) sono affette da una di queste malattie [1].

Tali disturbi sono scarsamente rappresentati nelle classificazioni internazionali e conseguentemente ignorati dai sistemi informativi sanitari, divenendo frequentemente invisibili agli occhi della società [59].

A causa della scarsa disponibilità di informazioni e della loro disgregazione, negli ultimi anni si sta assistendo a una forte crescita di comunità di pazienti sul web. I contesti social (quali i gruppi Facebook) divengono pertanto il territorio attraverso cui si condividono esperienze, si richiedono pareri e si scambiano informazioni di indubbia rilevanza durante tutto il percorso di un malato raro [4].

[7] `http://www.bowker.com/`
[8] `https://it.wikipedia.org/`
[9] `https://meta.wikimedia.org/wiki/List_of_Wikipedias`
[10] `https://www.orpha.net`

2.4 Importanza del NLP in ambito medico

L'avvento dei *social media* ha fornito nuovi servizi per una condivisione efficiente ed efficace delle proprie conoscenze, idee e opinioni [60]. Tutto questo ha reso il Web una grande fonte di dati, per lo più espressi in linguaggio naturale.

Per aver accesso alla conoscenza significativa racchiusa in questa enorme quantità di dati, oggi in crescita esponenziale, è necessario avere delle soluzioni software e architetturali capaci di gestir ed elaborare quest'ultima.

Il campo NLP è particolarmente in espansione nel settore sanitario. Questa tecnologia sta migliorando l'erogazione delle cure e le diagnosi delle malattie. Alcune organizzazioni sanitarie adottano cartelle cliniche elettroniche. Molte *chatbot* vengono utilizzate, una delle più comuni è Ada Health[11], un'app in grado di valutare lo stato di salute dell'utente sulla base dei sintomi indicati. Un altro strumento molto popolare è Babylon Health[12] che collabora con il Servizio Sanitario Nazionale[13] del Regno Unito per offrire informazioni basate sul percorso clinico personale e una consulenza video dal vivo con un vero medico, se necessario.

In Figura 2.4 è indicato il numero totale di pubblicazioni contenenti "natural language processing" dall'anno 1977 all'anno 2020 (sono presenti solamente gli articoli che precedono la data di scrittura 09-2020) su PubMed[14], comprendente più di 30 milioni di articoli e studi pertinenti alla letteratura biomedica. Come si può osservare l'interesse accademico nei confronti dell'utilizzo del NLP in campo medico è in decisivo aumento.

Per i malati rari internet offre la preziosa opportunità di connettersi e fornire supporto ad altri che hanno esperienze simili, sono spesso anche l'unico modo per i pazienti per descrivere i sintomi, i trattamenti terapeutici, i medici di riferimento e le caratteristiche delle possibili cure [61]. I *social media* permettono ai pazienti di sollevare problemi che in precedenza non erano stati considerati o apprezzati dai professionisti del settore medico [62].

Data la potenzialità dei dati raccolti dai singoli pazienti con malattie rare e l'interesse in crescita in campo medico per le tecnologie NLP, è incredibilmente utile unire il contesto reale al mondo della ricerca scientifica. Questo permetterebbe di rispondere agli interrogativi dei pazienti e fornire loro un ulteriore strumento attraverso il quale operare decisioni che richiedano approfondimenti e molteplici punti di vista [4].

[11] `https://ada.com/`
[12] `https://www.babylonhealth.com/`
[13] `https://www.nhs.uk/`
[14] `https://pubmed.ncbi.nlm.nih.gov/`

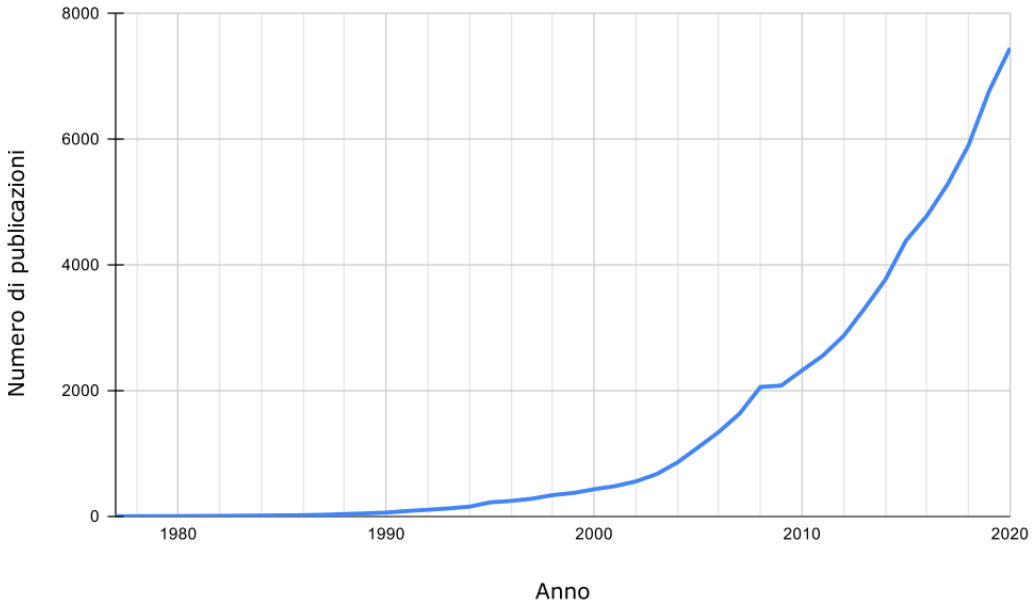


Figura 2.4: Numero totale di pubblicazioni su PubMed dal 1977

2.5 Dataset sull'Acalasia Esofagea

Il *dataset* scelto per come già anticipato all'inizio di Capitolo 2, proviene dallo studio di Giacomo Frisoni in "A new unsupervised methodology of descriptive textmining for knowledge graph learning" [4]. Il lavoro presentato nella tesi sopraindicata intende estrarre la conoscenza racchiusa da un gruppo Facebook per pazienti di Acalasia Esofagea, al fine di rappresentarla in una forma organizzata e abilitando un ragionamento logico deduttivo al di sopra di essa [4].

2.5.1 Acalasia Esofagea

L'acalasia (ORPHA:930) è un raro disturbo dell'esofago, l'organo che trasporta il cibo dalla gola allo stomaco, ed è caratterizzata da una ridotta capacità di spingere il cibo verso lo stomaco (peristalsi). La malattia colpisce circa una persona ogni 10˙000 [63, 64].

I sintomi presentati possono essere disfagia (arresto del bolo alimentare nell'esofago), rigurgito di cibo non digerito, problemi respiratori (tosse notturna, aspirazione ricorrente e polmonite), dolore toracico e perdita di peso [65].

La causa esatta dell'acalasia non è nota e al momento non c'è cura. Una volta che l'esofago è paralizzato, il muscolo non può più funzionare corretta-

mente. I sintomi possono di solito essere trattati in modo appropriato, ma si possono ripresentare e il paziente potrebbe aver bisogno di un trattamento intermittente. I pazienti hanno una normale aspettativa di vita [64]. I trattamenti per l'acalasia includono opzioni chirurgiche e non chirurgiche. Le opzioni non chirurgiche includono:

- dilatazione pneumatica — un palloncino viene inserito nello sfintere esofageo inferiore e gonfiato per allargare l'apertura;
- botox — viene iniettato del Botox direttamente nello sfintere esofageo, è raccomandato solo per le persone che non sono candidate alla dilatazione pneumatica o all'intervento chirurgico;
- prescrizione di farmaci — i farmaci prescritti sono miorilassanti come la nitroglicerina o la nifedipina prima di mangiare. Questo tipo di terapia è raramente indicato in quanto ha risultati scarsi ed effetti collaterali gravi.

Le opzioni chirurgiche per il trattamento dell'acalasia includono:

- LHM (laparoscopic heller myotomy) — il chirurgo taglia il muscolo all'estremità inferiore dello sfintere esofageo per permettere al cibo di passare più facilmente nello stomaco;
- POEM (Peroral endoscopic myotomy) — il chirurgo utilizza un endoscopio inserito attraverso la bocca per creare un'incisione nel rivestimento interno dell'esofago e poi tagliare il muscolo [66].

2.5.2 Gruppo Facebook

Il dataset proviene dal gruppo Facebook *Acalasia esofagea... i malati "rari" non sono soli...!*[15]. Il gruppo — gestito dall'associazione AMAE Onlus (Associazione Malati Acalasia Esofagea) — è stato creato il 16 Ottobre 2008, e alla data di scrittura (09-2020) ha circa 2100 utenti.

Sono stati considerati i post e i commenti trovati nella pagina a partire da febbraio 2009 fino ad agosto 2019. Il numero totale di post presi in considerazione risulta essere 6˙917, quello dei commenti è 61˙694.

I dati sono stati ottenuti tramite l'API Graph di Facebook[16] da Giacomo Frisoni.

[15] `https://www.facebook.com/groups/36705181245/`

[16] `https://developers.facebook.com/docs/graph-api/`

2.5.3 TextRazor

TextRazor[17] è il software scelto per effettuare i processi di NER e NEL. TextRazor dispone di un'enorme base di conoscenza dei dettagli delle entità estratte da varie fonti web, tra cui Wikipedia, DBPedia e Wikidata. Al momento (versione: 2020-03) le entità presenti sono 36˙584˙406. Vengono utilizzati anche metodi statistici per identificare persone, luoghi e aziende che non sono mai stati individuati precedentemente.

Per ogni specifica entità, textRaxor individua una moltitudine di parametri, tra i più importanti e utili per gli scopi di questa tesi:

- L'ID Freebase e Wikidata;
- Link alla pagina di Wikipedia dell'entità;
- Una lista di categorie in cui l'entità può essere classificata;
- Un punteggio da 0.5 a 10 (confidenceScore) rappresentativo della correttezza con cui l'entità è stata individuata;
- Un punteggio tra 0 e 1 (relevanceScore) rappresentativo dell'importanza dell'entità nel testo;
- La traduzione in inglese dell'entità effettuata tramite Wikipedia (se possibile).

Particolarmente importante come TextRazor oltre alle operazioni di NER e NEL offra anche una traduzione inglese, se necessario, delle entità.

2.5.4 Struttura dataset

I file utilizzati contengono dati elaborati da Giacomo Frisoni [4] a partire dai contenuti della pagina Facebook descritta precedentemente.

- **post_ita.cvs**
  Contiene l'elenco completo dei post ottenuti.
  I campi più rilevanti presenti in questo file sono: *post_id* (un numero univoco generato da Facebook per ogni nuovo post), *message* (il testo contenuto nel post), *created_time* (il giorno e l'ora in cui il post è stato pubblicato), *updated_time* (la data e l'ora in cui il post è stato aggiornato, uguale a created_time se non è mai stato modificato).

[17] `https://www.textrazor.com/`

- **comments_ita.cvs**
  Contiene l'elenco completo dei commenti ottenuti.
  I campi più rilevanti presenti in questo file sono: *parent_post_id* (l'id del post commentato), *id* (un numero univoco generato da Facebook per ogni commento, diverso da quello dei post), *message* (il testo contenuto nel commento), *created_time* (il giorno e l'ora in cui il post è stato pubblicato).

- **textrazor_ner_posts_ita.cvs**
  Contiene le 15˙688 entità individuate da TextRazor a partire dalla lista dei 6˙918 post.
  I campi utilizzati di questo file sono l'*id* del post di provenienza dell'entità, i punteggi di *confidencescore* dell'entità, *matched_texts* (il testo trovato corrispondente all'entità), *entity_id* (nome dell'entità), *entity_english_id* (la traduzione in inglese dell'entità).

- **textrazor_ner_comments.cvs**
  Contiene le 73˙096 entità individuate da TextRazor a partire dalla lista dei 61˙693 commenti.
  I campi utilizzati di questo file sono: *id*, *confidencescore*, *matched_texts*, *entity_id*, *entity_english_id*.

- **documents_quality_nerid.rds**
  Contiene tutti i post e commenti della pagina Facebook prodotti dal processo di elaborazione di qualità effettuato.
  I campi utilizzati sono il testo del messaggio o post e l'id corrispondente.
  Tra le operazioni più rilevanti realizzate sono presenti la separazione di parole attaccate, la rimozione di spazi e lettere in eccesso e la correzione dell'ortografia.

Capitolo 3

Traduzione automatica di documenti social

In questo capitolo viene utilizzato un campione di post proveniente dalla pagina Facebook indicata nella Sezione 2.5.2 per analizzare le prestazioni dei traduttori automatici scelti. In particolare i software vengono analizzati manualmente e con alcune delle metriche di valutazione descritte nella Sezione 1.3.

Le diverse metodologie utilizzate per giudicare i traduttori permettono di fornire riflessioni sulla correlazione delle diverse metriche con il giudizio umano.

Viene infine individuato il traduttore automatico migliore per tradurre il dataset dei post e commenti dell'Acalasia Esofagea e vengono confrontati i risultati con altre valutazioni già effettuate.

3.1 Implementazione

3.1.1 Colab

Google Colaboratory[1] (più comunemente noto come Colab) è una piattaforma online gratuita che offre un servizio basato sui blocchi note Jupyter.

Jupyter[2] è un'applicazione web open-source e accessibile attraverso il browser che permette di creare e condividere documenti che contengano codice, equazioni e grafici. Il codice è organizzato in celle, che possono essere modificate ed eseguite individualmente. L'output di ogni cella appare direttamente sotto di essa e viene memorizzato come parte del documento [67]. L'utilizzo di

[1] `ttps://colab.research.google.com/`
[2] `https://jupyter.org/`

questa tecnologia rende più facile la condivisione e la replica di lavori scientifici poiché gli esperimenti e i risultati sono presentati in modo autonomo [68].

Colab permette quindi a chiunque di scrivere ed eseguire celle di codice Python tramite il browser e successivamente condividerlo come un qualunque oggetto Google Docs. É particolarmente adatto per machine learning e analisi dei dati grazie alle numerose librerie già configurate e al supporto di accelerazione GPU [69]. Colab viene utilizzato in questa tesi per valutare le traduzioni, analizzare e gestire le entità e modificare i post.

3.1.2 Librerie utilizzate

NLTK

Il NLTK[3] (Natural Language Toolkit) è una raccolta di librerie e programmi pubblicata nel 2002 per analisi nel settore NLP in inglese, scritto in Python. Contiene esempi, corpora di dati e librerie per molti task NLP, per esempio *tokenization*, *part of speech tagging*, classificazione, *named entity recognition*, *tagging* [70].
Successivamente è riportato il codice utilizzato per valutare, con BLEU e NIST, la traduzione fatta con Google con la libreria NLTK utilizzando due traduzioni di riferimento.

```
from nltk.translate.bleu_score import corpus_bleu
references = [[open("/Reference.txt", "r").read().split(),
    open("/Reference2.txt", "r").read().split()]]
candidate = [open("/Google.txt", "r").read().split()]
corpus_nist(references, candidate)
corpus_bleu(references, candidate)
```

ROUGE

Per la valutazione ROUGE è stata invece utilizzata la libreria Python rouge[4]. Successivamente è riportato il codice utilizzato per valutare con la libreria rouge e due traduzioni di riferimento la traduzione fatta da Google.

```
!pip install rouge
from rouge import Rouge
rouge = Rouge()
reference1 = [open("/Reference.txt", "r").read()]
reference2 = [open("/Reference2.txt", "r").read()]
```

[3] https://www.nltk.org/
[4] https://pypi.org/project/rouge/

```
candidate = [open("/Google.txt", "r").read()]
rouge.get_scores(candidate, reference1)
rouge.get_scores(candidate, reference2)
```

3.2 Valutazione

Per la valutazione delle traduzioni è stato utilizzato un campione di post estratto dal dataset originario. I programmi di traduzione che si è deciso di analizzare sono: Google Translate, IBM Watson Language Translator, DeepL e MarianMT. La metriche utilizzate sono BLEU, NIST e ROUGE.

Si è ritenuto opportuno realizzare la valutazione sia con metriche automatiche che manuali (delegando la traduzione a individui qualificati), proponendo conseguentemente uno studio il più comprensivo possibile.

3.2.1 Valutazione Automatica

La valutazione è stata eseguita con 16 post contenenti 30 frasi, riportati in Appendice.

Per ogni post sono state effettuate due traduzioni di riferimento in inglese, da parte di quattro individui con conoscenza della lingua di livello C1 o superiore. A tali traduzioni si sono affiancati i risultati dei software sopra menzionati. Di seguito si riportano gli esiti delle traduzioni manuali e automatiche per un post appartenente al campione considerato per la valutazione.

Originale: *ieri ho fatto la prima dilatazione e oggi ho bruciori continui. è successo a qualcun'altro di voi?*
Prima Traduzione di riferimento: *I had my first dilation yesterday and today I have a continuous burning sensation. did it happen to any of you?*
Seconda Traduzione di riferimento: *yesterday i did the first dilation and today I have continuous burnings. has it ever happened to someone else?*
Google Traduttore: *yesterday I did the first dilation and today I have continuous burning. has it happened to any of you?*
IBM Watson Language Translator: *yesterday I made the first dilatation and today I have continuous burning. happened to someone else of you?*
DeepL: *yesterday i did the first dilation and today i have continuous burning. has it happened to anyone else?*
MarianMT: *Yesterday I did the first dilation and today I have continuous burns. Has it happened to anyone else of you?*

L'esempio riportato determina un punteggio BLEU di 0.51 per Google Traduttore, 0.25 per Watson Language Translator, 0.42 per DeepL e 0.53 per MarianMT.

Watson Language Translator viene correttamente penalizzato per l'errata traduzione in campo medico *dilatation* e per la parola *burns* che non appare nelle frasi di riferimento. Si può però anche osservare che DeepL ottiene un punteggio minore di Google Translator nonostante l'unica differenza tra le due frasi sia "anyone else" al posto di "any of you", che è ugualmente corretto.

Esito Valutazione

In Tabella 3.1 si possono osservare i risultati della valutazione su tutte le frasi, con le metriche BLEU,NIST e ROUGE-L.

Traduttore	**BLEU**	**NIST**	**ROUGE-L**
Google Traduttore	0.37	5.1	0.59
IBM Watson Language Translator	0.32	3.8	0.55
DeepL	0.32	4.9	0.55
MarianMT	0.36	5.1	0.57

Tabella 3.1: Valutazione dei software di traduzione automatica con BLEU, NIST e ROUGE-L

Come è possibile constatare, Google Translate ha raggiunto risultati superiori a quelli di MarianMT, che è a sua volta leggermente migliore di IBM Watson Language Translator e DeepL. Secondo la valutazione NIST però, Google Translate e MarianMT risultano invece avere prestazioni uguali.

3.2.2 Valutazione umana

Per quanto concerne la valutazione manuale, ogni post tradotto da una soluzione automatica di riferimento è stato annotato da esperti con un numero reale $s \in [0, 1]$, dove 1 rappresenta il punteggio per una traduzione ritenuta perfetta. Nell'analisi è stata inoltre calcolata anche la varianza, per evidenziare anche la variabilità della qualità delle traduzioni dei singoli post. I risultati sono in Tabella 3.2.

Traduttore	Valutazione	Varianza
Google Traduttore	0.72	0.30
IBM Watson Language Translator	0.60	0.36
DeepL	0.67	0.50
MarianMT	0.65	0.32

Tabella 3.2: Valutazione manuale dei software di traduzione automatica

Google Translate risulta essere il traduttore con la valutazione migliore e più consistente. DeepL risulta essere considerevolmente più inconsistente degli altri sistemi di traduzione automatica, ma viene comunque valutato positivamente.

3.3 Analisi dei risultati

Per riuscire a confrontare i risultati delle valutazioni automatiche con quelli invece assegnati manualmente, si è deciso di aggregare i primi in un solo punteggio, calcolato come indicato nell'Equazione 3.1.

$$PunteggioValutazioneAutomatica = \frac{BLEU + \frac{NIST}{10} + ROUGE}{3} \quad (3.1)$$

In Tabella 3.3 vengono riportati i valori delle valutazioni ottenuti con le metriche e il loro punteggio medio appena descritto.

Traduttore	BLEU	NIST	ROUGE	Val. Automatica
Google Traduttore	0.37	5.1	0.59	0.49
Watson Translator	0.32	3.8	0.55	0.42
DeepL	0.32	4.9	0.55	0.45
MarianMT	0.36	5.1	0.57	0.48

Tabella 3.3: Valutazione dei software di traduzione automatica con metriche e punteggio aggregato

Si può osservare come le valutazioni BLEU e ROUGE-L siano molto simili, mentre NIST ha valori leggermente diversi. Watson Translator viene valutato

negativamente da NIST rispetto agli altri traduttori, lasciando intendere che molte parole meno comuni non vengano tradotte correttamente. Questo implica un punteggio di valutazione automatica totale inferiore a quello degli altri traduttori analizzati. Google Translate risulta avere il punteggio massimo con tutte le metriche utilizzate (anche se pari con MarianMT per NIST).

La Tabella 3.4 riporta i risultati complessivi delle valutazioni da parte degli utenti, il punteggio di valutazione automatica e la media dei due valori.

Traduttore	**Val. Umana**	**Val. Automatica**	**Media**
Google Traduttore	0.72	0.49	0.61
Watson Translator	0.60	0.42	0.51
DeepL	0.67	0.45	0.56
MarianMT	0.65	0.48	0.57

Tabella 3.4: Valutazione manuale e automatica dei risultati prodotti dai software di traduzione automatica

Google Translate è il sistema di traduzione automatica più performante e accurato, sia secondo i valutatori che secondo gli algoritmi di traduzione automatica.

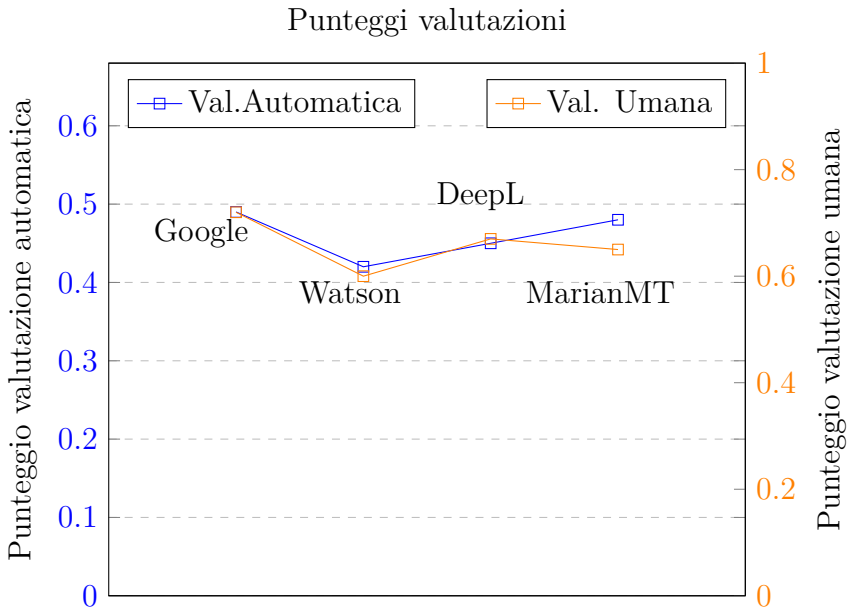
È stato inoltre calcolato il coefficiente di correlazione tra le tre metriche e la valutazione umana. Il coefficiente di correlazione misura il grado di dipendenza o la correlazione lineare tra due campioni casuali. Al momento del suo rilascio, BLEU era associato a un coefficiente di correlazione compreso tra 0.96 e 0.99 [37]. Con l'avvento di nuove tecnologie di traduzione (in particolare le reti neurali) la metrica BLEU è stata molto criticata [71, 72]. In questa tesi i coefficienti di correlazione BLEU sono prevedibilmente risultati più bassi e sono riportati in Tabella 3.5.

	BLEU	**NIST**	**ROUGE**	**Val. Automatica**
Coefficiente di correlazione	0.66	0.81	0.77	0.96

Tabella 3.5: Coefficienti di correlazione

È importante osservare come il punteggio di valutazione automatica risulti avere un coefficiente di correlazione con la valutazione umana significativamente migliore di quello ottenibile con ogni singola metrica presa in considerazione.

Come constatabile dai dati riportati in Tabella 3.4, i punteggi assegnati manualmente dagli esperti e quelli ottenuti per mezzo di metriche applicate alle traduzioni automatiche, si distribuiscono diversamente. Per confrontare graficamente i risultati delle due modalità, si è pertanto scelto di riportare quest'ultimi su scale diverse.



I due metodi di valutazione risultano essere consistenti nell'analisi di Google Translate, Watson Language Translator e DeepL ma non nella valutazione di MarianMT.

Assumendo che la valutazione umana sia più indicativa delle vere prestazioni di MarianMT, è possibile desumere che le metriche abbiano erroneamente valutato alcune delle frasi a causa dei problemi illustrati nella Sezione 1.3.

Un altro possibile problema è il punteggio assegnato a DeepL. L'azienda ha fatto scegliere a traduttori professionisti la traduzione migliore tra quelle prodotte in automatico, da diversi sistemi, di identità non nota. Il campione utilizzato a tal scopo è composto da 119 brani, tratti da un vasto numero di argomenti. Come diretta conseguenza dei risultati ottenuti, gli autori di DeepL ritengono di aver superato gli altri tool di traduzione automatica [73].

Una delle ragioni per questo contrasto risiede nelle modalità di valutazione utilizzate. A differenza di DeepL, in questa tesi si è assegnato un punteggio ad ogni traduzione. Dato l'alto valore di varianza in Tabella 3.2, è ragionevole pensare che il calo di valutazione di DeepL sia dovuto a un piccolo insieme di traduzioni di bassa qualità.

Un altro motivo potrebbe essere il registro del linguaggio utilizzato. Il linguaggio informale tipico dei contesti social rappresenta un'ulteriore sfida

per gli strumenti di NLP, la maggior parte dei quali si è sviluppata sulla base di testo formale e dipende in modo significativo dalla qualità del testo scritto. I post e commenti di Facebook contengono spesso errori di ortografia, parole che non sono parte di gergo tradizionale (es. slang) o inventate. I concetti descritti possono essere espressi tramite l'utilizzo di grammatica incorretta e scarsa punteggiatura. La scarsa qualità dei testi di partenza ha probabilmente influito significativamente nella traduzione.

Un altro fattore da considerare è l'ampio cambiamento della qualità di traduzione che si ottiene con testo minuscolo. In questa analisi è stato utilizzato testo completamente in lower-case. Nella Figura 3.1 si può osservare come la traduzione passi da incomprensibile a perfetta col il solo posizionamento corretto dei caratteri maiuscoli.

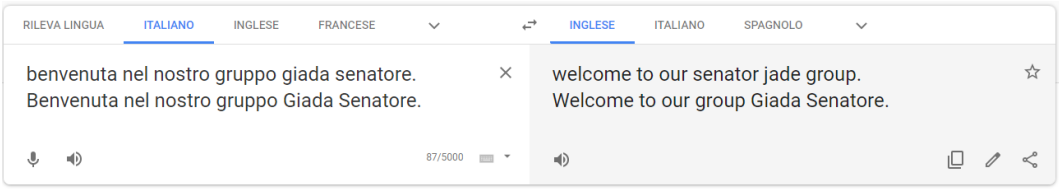


Figura 3.1: Esempio della stessa frase tradotta con Google Translate considerando testo esclusivamente minuscolo o con le maiuscole in corretta posizione

Un altro elemento capace di contribuire pesantemente sulla qualità di una traduzione è la punteggiatura. Nella Figura 3.2 si può osservare come la qualità della traduzione migliori sensibilmente con l'aggiunta di un punto interrogativo che segnali la presenza di una domanda, e come l'ultima frase con la punteggiatura corretta sia migliore delle altre.

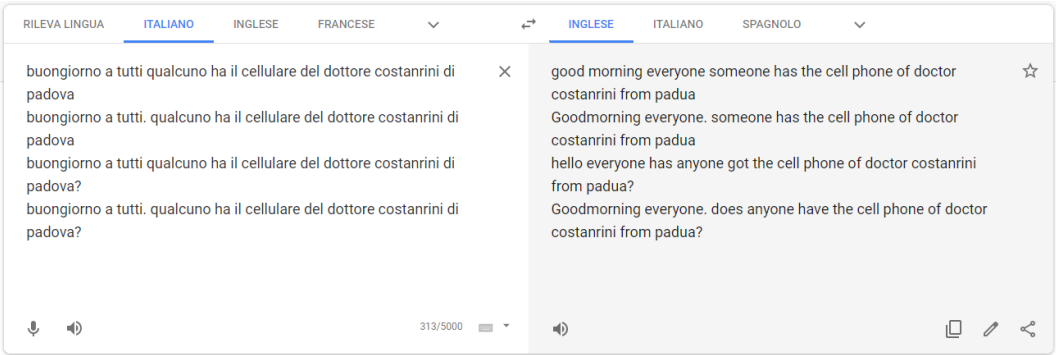


Figura 3.2: Esempio della stessa frase tradotta con Google Translate senza punteggiatura e aggiungendo una virgola, punto indicativo o entrambi

Capitolo 4

Gestione di entità durante la traduzione

4.1 Introduzione

I traduttori automatici hanno spesso difficoltà nella traduzione di termini specifici di dominio. All'interno del caso di studio considerato, tale problema emerge principalmente con nomi di medici e centri specializzati che sarebbe necessario non venissero tradotti. In Figura 4.1 sono riportati alcuni esempi dei suddetti errori, commessi da parte di diversi software di traduzione automatica.

Dal momento che l'interpretazione di questi dati con tool NLP risulterebbe inaccurata, in questo capitolo vengono quindi trattatati i metodi utilizzati per l'individuazione delle parole da non tradurre tramite il riconoscimento di entità.

4.2 Eliminazione entità superflue

Le entità riconosciute nei post e nei commenti per mezzo di TextRazor sono etichettate sia secondo la tassonomia gerarchica di Freebase che quella di Wikidata. Tuttavia, le due tipizzazioni sono unite e trasformate negli elementi di una nuova tassonomia (Figura 4.2) capace di rappresentare tutti i concetti di interesse nel dominio delle malattie rare, e ignorando di conseguenza quelli ritenuti non sufficientemente rilevanti ai fini dell'analisi.

Per garantire una migliore traduzione non è però necessario trattare tutte le entità appartenenti a questa classificazione. Da queste categorie di partenza sono state quindi eliminate le entità facilmente traducibili da Google Translate

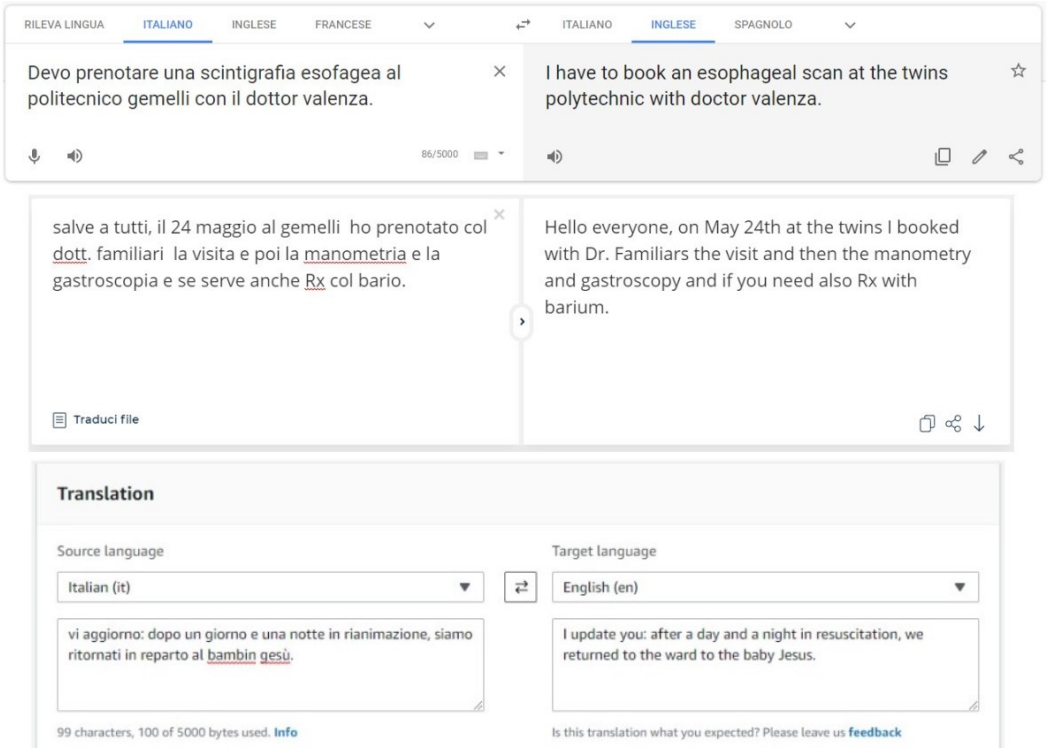


Figura 4.1: Traduzioni con gestione non appropriata di entità di dominio, da parte di Google Translate, DeepL e Amazon Translator

(es. sport, cibi) o comunque trattate tipicamente in maniera corretta (es. le città vengono tradotte solo se possibile e lasciate inalterate nel caso contrario).

Le entità rimanenti sono quelle che non devono essere tradotte o devono essere tradotte solo parzialmente (es. "Policlinico Gemelli" → "Gemelli Polyclinic"), e non propriamente gestite dai software di MT. Le entità trattate appartengono ad almeno una delle seguenti categorie: persone, organizzazioni, ospedali e università.

Seguendo le decisioni prese da Giacomo Frisoni sono state eliminate anche le entità possibilmente individuate erroneamente da parte di TextRazor, mantenendo quindi le sole entità con *confidenceScore* superiore a 0.5. Il numero totale di entità rimanenti è 10˙075.

4.3 Preprocessing e traduzione

Per tutte le operazioni successive viene utilizzato Colab, lo stesso strumento descritto nella Sezione 3.1.1.

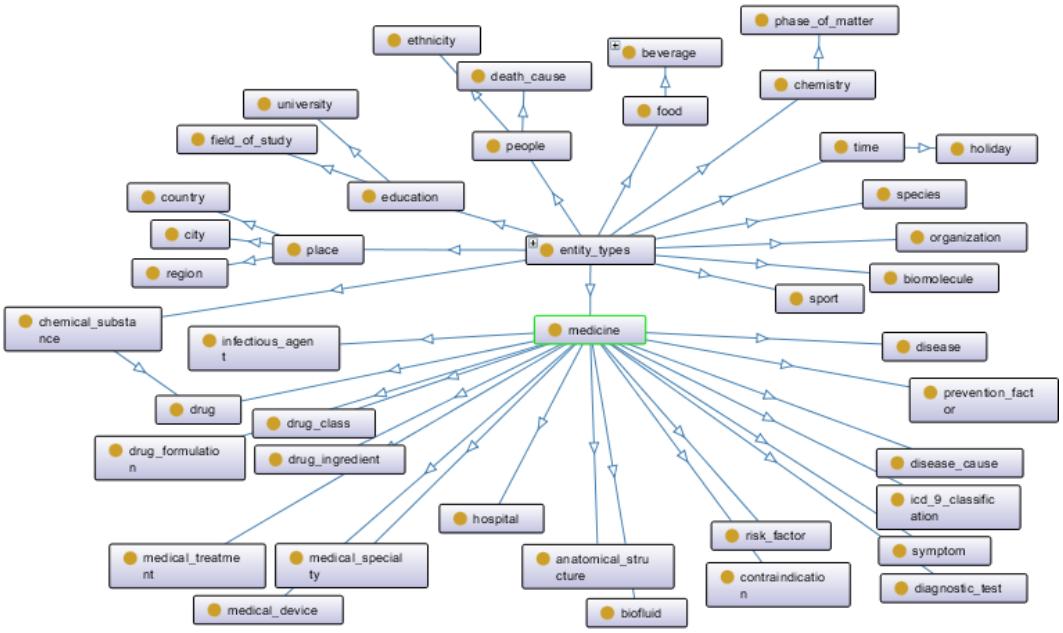


Figura 4.2: Tassonomia delle entità riconosciute da post e commenti
Ottenuta da [4]

Date le entità presenti nei file *textrazor_ner_posts_ita.cvs* e *textrazor_ner_comments_ita.cvs* sono state eliminate le entità non rilevanti. Tutte le entità rimanenti sono state trasformate nel seguente modo: entità→e-n-t-i-t-à. Questo assicura che l'entità non venga tradotta.

Le entità con il campo *entity_english_id* non disponibile sono state sostituite nel file *documents_quality_nerid.rds*, le altre sono state sostituite nel file con la corretta traduzione.

Il file creato è **massage_to_translate.xlsx** e contiene l'elenco dei post e commenti pronti per essere tradotti e il loro corrispondente id.

Per la traduzione è stato utilizzato Google Translate per i motivi già discussi nel Capitolo 3. Google Translate elimina i "-" presenti tra le lettere di una parola in automatico.

In seguito viene riportato lo pseudo-codice rappresentativo del processo di preparazione alla traduzione.

4.3.1 Pseudo-codice

```
/* eliminazione di tutte l'entità con confidenceScore
   minore o uguale a 0.5 o che non appartengono alle
   categorie precedentemente selezionate                */
for tutte le entità do
    if confidenceScore è minore uguale a 0.5 o l'entità non appartiene
     a una delle categorie selezionate then
        elimina l'entità;
    end
end

/* modifica delle entità e la loro traduzione in modo che
   dal formato "Abc def" risultino nel formato "abc_def,
   formato utilizzato nel file
   documents_quality_nerid.rds"                         */
for tutte le entità e la loro traduzione do
    trasformazione da spazi a "-"
    trasformazione da maiuscole a minuscole
end

/* Sostituzione delle entità trovate con la loro
   traduzione o dell'entità alternata con '-' per evitare
   la traduzione di Google Translate                    */
for tutti i post do
    scrittura dell'id e del messaggio completo nel file
     massage_to_translate.xlsx
    for tutti le entità do
        if entità(abc_def) appartiene al post then
            if l'entità ha una traduzione inglese then
                sostituzione dell'entità con la traduzione con i -
            else
                sostituzione dell'entità con la parola italiana con i -
            end
            riscrizione del messaggio nel file massage_to_translate.xlsx
        end
    end
end
```

4.4 PyPi

Nella comunità scientifica in generale, Python si sta affermando come uno dei linguaggi di programmazione più popolari. Una delle ragioni di questo successo è il ricco ecosistema di librerie e applicazioni. Oltre ha essere dotato di una libreria standard ben documentata, facile da usare e potente sono anche disponibili una grande varietà di pacchetti esterni [74].

Python Package Repository, o PyPI è il registro ufficiale per i pacchetti Python sviluppati da terzi. Al momento è anche uno dei più antichi e grandi depositi di software. Durante i suoi 15 anni di attività, PyPI[1] è cresciuto fino ad ospitare oltre 260˙144 pacchetti (09-2020) [75].

4.4.1 Contenuto progetto in PyPI

Una versione più generalizzata del codice illustrato nella Sezione 4.3.1 è stata utilizzata per creare una libreria PyPI. Il progetto è stato pubblicato ed è disponibile su PyPI al link: `https://pypi.org/project/strlistsmanipulator/`, si può installare con il comando `pip install strlistsmanipulator`.

In Figura 4.3 è stata indicata la struttura generale del progetto.

```
strlistsmanipulator
├── LICENSE
├── README.md
├── package
│   ├── __init__.py
│   └── list_manipulator.py
└── setup.py
```

Figura 4.3: Struttura progetto pubblicato su PyPI

La licenza scelta è la licenza MIT[2]. Il file *__init__.py* è necessario per importare la cartella *package*. Il file *setup.py* contiene informazioni sul pacchetto (come nome, versione, descrizione) e quali file di codice includere. Il file *list_manipular.py* contiene tutte le funzioni della libreria e la loro documentazione in Docstring[3].

[1] `https://pypi.org/`
[2] `https://opensource.org/licenses/MIT`
[3] `https://www.python.org/dev/peps/pep-0257/`

Capitolo 5

Conclusione

La valutazione dei programmi ha portato a trarre diverse osservazioni.

Considerando i post e commenti condivisi da una comunità di pazienti sul Gruppo Facebook *Acalasia esofagea... i malati "rari"non sono soli...!*[1], si è constatato come la traduzione automatica di più alta qualità sia quella realizzata da parte di Google Translate, adottando sia una valutazione automatica con metriche che una valutazione manuale condotta da utenti esperti.

Il calcolo del coefficiente di correlazione che le varie metriche di valutazione hanno ottenuto con il giudizio umano, ha reso evidente come la media tra più metriche produca risultati più somiglianti rispetto a quelli ottenibili considerando ciascuna di esse separatamente.

La traduzione di testo proveniente da *social media* è più difficoltosa perché il linguaggio è informale, la capitalizzazione è incoerente, le variazioni ortografiche e gli errori grammaticali sono più frequenti.

Partendo da tali considerazioni, questa tesi ha sfruttato TextRazor, per individuare le entità e decidere se possono essere solamente trascritte o se richiedono un mix di traslitterazione e traduzione. Adottando Google Translate e una propria soluzione per la gestione di entità specifiche per il dominio, è stato possibile completare la traduzione dell'intero dataset in lingua inglese.

Possibili sviluppi futuri includono la possibilità di riapplicare la metodologia di *text mining* svolta nella tesi [4] da Giacomo Frisoni sul corpus tradotto, sfruttando la possibilità di utilizzare i tool NLP disponibili in lingua inglese. Questo per verificare la presenza di un eventuale miglioramento dei risultati, per quanto concerne l'estrazione automatica di correlazioni tra concetti medici.

[1] `https://www.facebook.com/groups/36705181245/`

Appendice A

Frasi in italiano

Lista delle frasi in italiano usate per la valutazione dei programmi di traduzione automatica.

- io soffro di acalasia e la prossima settimana avro l'intervento heller dor.
- giornata brutta, l'esofago sembra strapparsi e spossatezza totale. maledetta acalasia.
- vi aggiorno: dopo un giorno e una notte in rianimazione, siamo ritornati in reparto al bambin gesù. hanno assorbito l'aria che aveva fatto collassare un polmone, un effetto collaterale della poem, ora è tutto ok. il mio giovanotto ha cominciato a bere e si sente bene. grazie a tutti voi per la vicinanza.
- oggi è programmato l'intervento poem di mio figlio, ma qualcosa non è andata stanotte. ha vomitato fino alle 4, anche se a digiuno. sono veramente a pezzi.
- comunicato di servizio, sto abbandonando il cucchiaio per la forchetta.
- potrei sapere quali, tra i primi esami, per diagnosticare l'acalasia. grazie.
- esattamente un anno fa venivo operato con la tecnica heller dor al sant'orsola di bologna. per me è come festeggiare la rinascita a nuova vita.
- buonasera, è capitato ad altri di voi che, ripetendo un esame a distanza di anni, vi diagnosticassero un acalasia esofagea che dapprima i precedenti referti escludevano?
- ieri ho fatto la prima dilatazione e oggi ho bruciori continui. è successo a qualcun'altro di voi?

- ciao a tutti, volevo sapere se qualcuno di voi è riuscito ad avere il pantoprazolo in esenzione con ricetta a costo zero con la nostra esenzione ri0010?
- ragazzi ieri mi sono operato heller dor oggi ho dolori atroci retrosternali come quando bevevo acqua è normale? qui mi dicono di si ma volevo un confronto.
- ciao, in marzo farò la prima dilatazione. una curiosità: la sedazione è simile a quella della gastroscopia o un po' più pesante? grazie delle eventuali risposte.
- benvenuta nel nostro gruppo giada senatore!
- buongiorno a tutti qualcuno ha il cellulare del dottore costanrini di padova?
- buongiorno a voi. una domanda: quanto dura l'intervento poem?
- presto sarò operata anch'io al gemelli il 10 vado per la preospedalizzazione e volevo ringraziare di essere entrata in questo gruppo. sono un po' spaventata ma leggere tanti commenti mi ha rincuorata non sono sola.

Bibliografia

[1] "Global genes, allies in rare disease - rare facts." `https://globalgenes.org/rare-facts/`, consultato 15 Settembre 2020.

[2] "Priority diseases and reasons for inclusion - who." `https://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf?ua=1`, consultato 15 Settembre 2020.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[4] G. Frisoni, G. Moro., and A. Carbonaro., "Learning interpretable and statistically significant knowledge from unlabeled corpora of social text messages: A novel methodology of descriptive text mining," in *Proceedings of the 9th International Conference on Data Science, Technology and Applications - Volume 1: DATA,*, pp. 121–132, INSTICC, SciTePress, 2020.

[5] I. S. Mukhin, "An experiment on the machine translation of languages carried out on the besm," *Proceedings of the IEE - Part B: Radio and Electronic Engineering*, vol. 103, no. 3, pp. 463–472, 1956.

[6] W. H. Hutchins, "Machine translation: A brief history," in *Concise history of the language sciences: from the Sumerians to the cognitivists*, 1995.

[7] M. Boyan, I. Bonev, S. Ortiz, R. Juan, A. Pérez, O. Gema, G. Ramírez-Sánchez, F. Sánchez-Martínez, M. Carme, M. Montava, and F. Tyers, "Documentation of the open-source shallow-transfer machine translation platform apertium," 04 2010.

[8] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, R. L. Mercer, and P. Roossin, "A statistical approach to language translation," in *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*, 1988.

[9] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Comput. Linguist.*, vol. 16, p. 79–85, June 1990.

[10] M. Junczys-Dowmunt and R. Grundkiewicz, "Phrase-based machine translation is state-of-the-art for automatic grammatical error correction," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (Austin, Texas), pp. 1546–1556, Association for Computational Linguistics, Nov. 2016.

[11] A. Krogh, "What are artificial neural networks?," *Nature biotechnology*, vol. 26, pp. 195–7, 03 2008.

[12] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," 2012.

[13] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 2342–2350, JMLR.org, 2015.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, p. 1735–1780, Nov. 1997.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014.

[16] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[17] S. Yang, Y. Wang, and X. Chu, "A survey of deep learning techniques for neural machine translation," 2020.

[18] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[19] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, "Massive exploration of neural machine translation architectures," 2017.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[22] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," 2017.

[23] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*, (Vancouver, Canada), pp. 67–72, Association for Computational Linguistics, July 2017.

[24] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T.-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou, "Achieving human parity on automatic chinese to english news translation," 2018.

[25] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Névéol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, "Findings of the 2016 conference on machine translation," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, (Berlin, Germany), pp. 131–198, Association for Computational Linguistics, Aug. 2016.

[26] A. Toral, M. Wieling, and A. Way, "Post-editing effort of a novel with statistical and neural machine translation," *Frontiers in Digital Humanities*, vol. 5, p. 9, 2018.

[27] M. Dowling, T. Lynn, A. Poncelas, and A. Way, "SMT versus NMT: Preliminary comparisons for Irish," in *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, (Boston, MA), pp. 12–20, Association for Machine Translation in the Americas, Mar. 2018.

[28] A. Beyer, V. Macketanz, A. Burchardt, and P. Williams, "Can out-of-the-box nmt beat a domain-trained moses on technical data," *Proceedings of EAMT User Studies and Project/Product Descriptions*, pp. 41–46, 2017.

[29] Y. Wang, L. Zhou, J. Zhang, and C. Zong, "Word, subword or character? an empirical study of granularity in chinese-english nmt," in *Machine*

Translation (D. F. Wong and D. Xiong, eds.), (Singapore), pp. 30–42, Springer Singapore, 2017.

[30] A. Toral and V. M. Sánchez-Cartagena, "A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions," 2017.

[31] M. Tavosanis, "Valutazione umana di google traduttore e deepl per le traduzioni di testi giornalistici dall'inglese verso l'italiano," 2019.

[32] D. A. Ferrucci, "Introduction to 'this is watson'," *IBM Journal of Research and Development, Volume: 56*, 2012.

[33] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, (Melbourne, Australia), 2018.

[34] M. S. Maučec and G. Donaj, *Recent Trends in Computational Intelligence*, ch. 8 Machine Translation and the Evaluation of Its Quality, Section 4. IntechOpen, 2019.

[35] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, June 2005.

[36] M. S. Maučec and G. Donaj, *Recent Trends in Computational Intelligence*, ch. 8 Machine Translation and the Evaluation of Its Quality, Section 5. IntechOpen, 2019.

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," 2002.

[38] M. Przybocki, K. Peterson, S. Bronsart, and G. Sanders, "The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results," *Machine Translation*, vol. 23, pp. 71–103, 09 2009.

[39] K. Wołk and D. Koržinek, "Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking," 10 2015.

[40] D. Cer, C. D. Manning, and D. Jurafsky, "The best lexical metric for phrase-based statistical mt system optimization," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 555–563, 2010.

[41] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[42] N. Huyen and C. Joseph, "Mewr: Machine translation evaluation without reference texts," 2017. `https://pbs.twimg.com/media/DQJBTbxV4AApFJE.jpg:large`, consultato il 5 Settembre 2020.

[43] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the second international conference on Human Language Technology Research*, pp. 138–145, 2002.

[44] T. Phan Thi Thanh and I. Thomas, "English-vietnamese machine translation of proper names," in *Error Analysis and Some Proposed Solutions*, pp. 386–393, 09 2012.

[45] R. Zarei and S. Norouzi, "Proper nouns in translation: Should they be translated?," *International Journal of Applied Linguistics and English Literature*, vol. 3, no. 6, pp. 152–161, 2014.

[46] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," 2019.

[47] B. Babych and T. Hartley, "Improving machine translation quality with automatic named entity recognition," 01 2003.

[48] W. Ling, P. Calado, B. Martins, I. Trancoso, A. Black, and L. Coheur, "Named entity translation using anchor texts," in *IWSLT*, 2011.

[49] D. Demner-Fushman, W. Chapman, and C. Mcdonald, "What can natural language processing do for clinical decision support?," *Journal of biomedical informatics*, vol. 42, pp. 760–72, 09 2009.

[50] E. D. Liddy, "Natural language processing," 2001.

[51] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2020.

[52] A. Voutilainen, *Part-of-speech tagging*, vol. 219. The Oxford handbook of computational linguistics, 2003.

[53] P. W. Jusczyk, D. M. Houston, and M. Newsome, "The beginnings of word segmentation in english-learning infants," *Cognitive psychology*, vol. 39, no. 3-4, pp. 159–207, 1999.

[54] J. Zheng, W. Chapman, R. Crowley, and G. Savova, "Coreference resolution: A review of general methodologies and applications in the clinical domain," *Journal of biomedical informatics*, vol. 44, pp. 1113–22, 08 2011.

[55] P. D. Turney and M. L. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," 2002.

[56] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," *Information Processing & Management*, vol. 51, p. 32–49, Mar 2015.

[57] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[58] "Produzione e lettura di libri in italia," 2018. `https://www.istat.it/it/archivio/236320`, consultato il 15 Settembre 2020.

[59] A. Rath, A. Olry, F. Dhombres, M. Milicic Brandt, B. Urbero, and S. Aymé, "Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users," *Human mutation*, vol. 33, pp. 803–8, 05 2012.

[60] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[61] N. Garcelon, A. Burgun, R. Salomon, and A. Neuraz, "Electronic health records for the diagnosis of rare diseases," *Kidney International*, vol. 97, 01 2020.

[62] W. Davies, "Insights into rare diseases from social media surveys," *Orphanet journal of rare diseases*, vol. 11, no. 1, pp. 1–5, 2016.

[63] "National organization for rare disorder. achalasia," 2018. https://rarediseases.org/rare-diseases/achalasia/, consultato il 15 Settembre 2020.

[64] "Orphanet. idiopathic achalasia." https://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=EN&data_id=302&disease=Idiopathic-achalasia&search=Disease_Search_Simple, consultato il 15 Settembre 2020.

[65] G. Boeckxstaens, G. Zaninotto, and J. Richter, "Achalasia," *Lancet*, vol. 383, 07 2013.

[66] "Mayo clinic. achalasia." https://www.mayoclinic.org/diseases-conditions/achalasia/diagnosis-treatment/drc-20352851, consultato il 15 Settembre 2020.

[67] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, and J. development team, "Jupyter notebooks ? a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (F. Loizides and B. Scmidt, eds.), pp. 87–90, IOS Press, 2016.

[68] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, "Using the jupyter notebook as a tool for open science: An empirical study," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–2, 2017.

[69] T. Carneiro, R. V. Medeiros Da NóBrega, T. Nepomuceno, G. Bian, V. H. C. De Albuquerque, and P. P. R. Filho, "Performance analysis of google colaboratory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61677–61685, 2018.

[70] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[71] E. Reiter, "A structured review of the validity of BLEU," *Computational Linguistics*, vol. 44, pp. 393–401, Sept. 2018.

[72] R. Ananthakrishnan, P. Bhattacharyya, M. Sasikumar, and R. M. Shah, "Some issues in automatic evaluation of english-hindi mt: more blues for bleu," *ICON*, 2007.

[73] "Nuova svolta nella qualità della traduzione basata sull'intelligenza artificiale." `https://www.deepl.com/blog/20200206.html`, consultato il 6 Febbrio 2020.

[74] T. Megies, M. Beyreuther, R. Barsch, L. Krischer, and J. Wassermann, "Obspy - what can it do for data centers and observatories?," *Annals Of Geophysics*, vol. 54, pp. 47–58, 04 2011.

[75] E. Bommarito and M. Bommarito, "An empirical analysis of the python package index (pypi)," 2019.