

ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA

Scuola di Scienze
Corso di Laurea in Ingegneria e Scienze Informatiche

**STUDIO E SPERIMENTAZIONE DI CHATBOT GENERATIVE ALLO
STATO DELL'ARTE CON APPLICAZIONE ALLA PANDEMIA DA
COVID19**

Elaborato in
Programmazione Di Applicazioni Data Intensive

Relatore
Prof. Gianluca Moro

Presentata da
Asia Lucchi

Seconda Sessione di Laurea
Anno Accademico 2019 – 2020

PAROLE CHIAVE

Chatbot

Natural Language Processing

Machine Learning

Language Model

Blender Bot

*A chiunque mi sia stato vicino,
e mi abbia aiutato a raggiungere questo traguardo.*

Introduzione

Fin dalla nascita dei primi calcolatori, gli studiosi sono risultati tanto affascinati quanto allarmati dall'idea di costruire un sistema che potesse pensare ed agire razionalmente, come un essere umano. Dal 1943 la creazione delle cosiddette Intelligenze Artificiali è diventato un ambito attivo di ricerca che, soprattutto negli ultimi decenni, si è evoluto sempre più.

Una delle caratteristiche fondamentali che tali sistemi devono possedere per apparire *intelligenti* è la capacità di *conversare* come esseri umani, riuscendo a comprendere ciò che dice l'interlocutore in tutte le sue sfumature e formulando una risposta non solo corretta, ma realistica.

Anche se attualmente le Chatbot sono principalmente usate dalle aziende come operatori virtuali per rispondere alle domande dei clienti con alcune frasi standard, l'obiettivo primario della ricerca è lo sviluppo di prototipi il cui unico scopo è l'interazione empatica con l'essere umano. Per raggiungere tale obiettivo si utilizza innanzitutto il Natural Language Processing, che si occupa di capire il reale significato di un messaggio, al contrario di ciò che faceva ELIZA¹, la prima chatbot esistente, ideata nel 1966. Questo banale sistema agiva come uno psicoterapeuta nella fase iniziale di un intervento psichiatrico, ovvero riproponendo tutto ciò che veniva detto dall'interlocutore sotto forma di domanda, senza sapere realmente di cosa stava parlando. Fortunatamente da allora c'è stato un notevole progresso sia nell'NLP che nella Natural Language Generation: attualmente esistono AI in grado di scrivere, a partire da un breve prompt, delle fake news che risultano credibili ad un'alta percentuale dei lettori utilizzando l'enorme bagaglio culturale che hanno appreso dai dati su cui sono state addestrate.

Nel lavoro di tesi vengono approfondite le tecnologie e le innovazioni più recenti legate all'evoluzione delle chatbot. Si discute delle varie tipologie di chatbot esistenti, delle architetture su cui sono basate, dei modelli di riferimento con le loro problematiche e dei sistemi all'avanguardia prodotti da noti gruppi di ricerca high-tech. Infine si utilizza uno di tali sistemi per sviluppare una chatbot in grado di fare conversazione sul COVID-19.

¹[https://it.wikipedia.org/wiki/ELIZA_\(chatterbot\)](https://it.wikipedia.org/wiki/ELIZA_(chatterbot))

Indice

1	Machine Learning per il Natural Language Processing	1
1.1	RNN	2
1.1.1	LSTM	2
1.1.2	GRU	4
1.2	Transformer	4
2	Language Models	7
2.1	GPT-2 e GPT-3	8
2.2	BERT	11
2.3	Turing-NLG	12
2.4	Language Models multi-lingua	13
2.4.1	XLM-RoBERTa	13
2.4.2	LID	14
2.5	Language Models multi-modali	16
2.5.1	Hateful Memes Challenge	16
2.5.2	VilBERT	18
3	Chatbot Retrieval-Based	21
3.1	Rasa	22
3.1.1	Rasa NLU	22
3.1.2	Rasa Core	24
3.2	Universal Sentence Encoder	26
3.2.1	COVID-19 Retrieval-based Chatbot	27
4	Chatbot Generative	29
4.1	Metodi di valutazione	29
4.1.1	BLEU	29
4.1.2	BLEURT	30
4.1.3	Sensibleness and Specificity Average (SSA)	30
4.1.4	ACUTE-Evaluation	31
4.2	Problematiche	32
4.2.1	Degenerazione del testo	32

4.2.2	Retrieve VS Refine	34
4.3	Stato del'arte	34
4.3.1	Alexa	34
4.3.2	Replika	36
4.3.3	Meena	37
4.3.4	BlenderBot	38
5	Progetto	43
5.1	Testing	43
5.2	Confronto con Meena	45
5.3	COVID-19 Generative Chatbot	47
5.3.1	Chitchat VS Knowledge	48
5.3.2	Covid-19 F.A.Q. Chatbot	54
5.3.3	Browser Chat	59
	Ringraziamenti	61
	Bibliografia	63

Elenco delle figure

1.1	Funzionamento di una RNN. ²	2
1.2	Struttura della memory cell LSTM. ³	3
1.3	Struttura della memory cell GRU. ⁴	4
1.4	Architettura del Transformer. ⁵	5
2.1	Miglioramento della performance di GPT-3 inserendo il prompt ed aumentando i parametri in funzione del numero di tentativi. ⁶	10
2.2	Input embeddings utilizzati da BERT. ⁷	11
2.3	Esempio di zero-shot question answering di T-NLG. ⁸	13
2.4	Confronto fra la quantità di testo presente in varie lingue nei dataset Wikipedia e CommonCrawl. ⁹	14
2.5	Matrice di confusione del problema LID usando bi-LSTM. ¹⁰	15
2.6	Esempi di Hateful Memes con rispettivi benign confounders. ¹¹	16
2.7	Esempio di post che è necessario rappresentare come unica entità prendendo in considerazione sia l'immagine che il testo. ¹²	18
2.8	Architettura del transformer co-attentional usato da ViBERT. ¹³	19
3.1	Esempio di selezione della risposta più adeguata da un set predefinito operata da una chatbot retrieval-based. ¹⁴	21
3.2	Esempio di Pipeline per la risoluzione di un problema di NLP. ¹⁵	23
3.3	Esempio di Interactive Learning usato da Rasa Core. ¹⁶	25
3.4	Valutazioni della similarità calcolate da USE. ¹⁷	26
3.5	Esempio di risposte fornite dall'AI fine-tuned sulle FAQ del WHO riguardanti il COVID-19.	28
4.1	Esempio di confronto fra dialoghi per valutazione ACUTE-E. ¹⁸	32
4.2	Variabilità dei token spontaneamente concepiti da un umano rispetto a quelli generati dalla Beam Search. ¹⁹	33
4.3	Architettura di Meena ed esempio di conversazione. ²⁰	37
4.4	Esempio di conversazione fra speaker <i>guided</i> e <i>unguided</i> che fa parte del dataset BST. In viola è indicata la categoria a cui appartiene la risposta in cui PB = background personale, K = conoscenza, S = situazione personale, E = empatia. ²¹	40

5.1	Esempio di risposta non adeguata scelta da BlenderBot dopo un lungo dialogo. ²²	44
5.2	Esempio di messaggio classificato ingiustamente come sconveniente.	44
5.3	Esempio di richiesta ripetuta due volte da parte di BlenderBot (in rosso) e di contraddizione (in viola).	45
5.4	Esempio di conversazione in cui il bot è costantemente in accordo con l'utente.	45
5.5	Grafico della perplexity del modello calcolata su vari tipi di domande.	50
5.6	Grafico della metrica BLEU del modello calcolata su vari tipi di domande.	51
5.7	Esempio di conversazione con modello addestrato su 80 epoche e domande di chitchat rieseguite più volte.	52
5.8	Esempio di iterazione usando l'interfaccia web.	59

Capitolo 1

Machine Learning per il Natural Language Processing

Il Machine Learning fornisce metodi generali per estrarre modelli di conoscenza da una grande mole di dati. Per *modello di conoscenza* si intende una funzione che associa ad ogni dato una classe o un valore numerico.

Una delle definizioni più citate è quella riportata da Tom M. Mitchell nell'omonimo libro "*Machine Learning*" [1]:

Si dice che un programma apprende dall'esperienza E con riferimento ad alcune classi di compiti T e con misurazione della performance P , se le sue performance nel compito T , come misurato da P , migliorano con l'esperienza E .

Questo metodo innovativo negli ultimi anni ha cambiato l'approccio degli sviluppatori, i quali possono fornire al software una grande quantità di informazioni lasciandogli il compito di apprendere il modo più appropriato con cui comportarsi invece che definire essi stessi una lunga e proibitiva serie di regole.

Uno degli ambiti in cui è usato il Machine Learning è il **Natural Language Processing**, che si pone come obiettivo finale riuscire a leggere, decifrare e comprendere il linguaggio umano. Alcuni esempi di problemi che NLP si occupa di risolvere sono la NER (Named Entity Recognition), la lemmatizzazione, il question answering e la full-text search.

Fra il 2013 e il 2014 hanno iniziato a diffondersi modelli NLP che facevano uso di Reti Neurali Artificiali, modelli matematici formati da strati di neuroni che cercano di riprodurre la struttura di una rete neurale biologica. Questo ambito di studio prende il nome di **Deep Learning**.

Ogni rete ha uno strato di input, uno di output e vari hidden layers; i neuroni nei vari strati applicano una funzione di attivazione su i dati che

ricevono utilizzando i pesi ottimali calcolati nella fase di training attraverso il meccanismo di *back-propagation*.

1.1 RNN

Le **Recurrent Neural Network** sono reti neurali ricorrenti adatte a processare **sequenze di dati** in cui ad ogni passaggio sono rilevanti tutte le informazioni inserite in precedenza per la predizione dell'output. Questo il caso nell'NLP ed in particolare nell'**NLG** (Natural Language Generation), dove per ogni parola da intuire è necessario essere a conoscenza della porzione di frase già generata.

Nella pratica, i neuroni di una RNN in ogni momento t prendono in ingresso un elemento della sequenza di input $x(t)$ e l'output generato dal neurone stesso allo step precedente $y(t-1)$. Il neurone ha due set di pesi, rispettivamente uno per $x(t)$ ed uno per $y(t-1)$. In Figura 1.1 è rappresentato schematicamente il funzionamento basilare della rete ed una sua riproduzione dispiegata (*unrolled*) nel tempo.

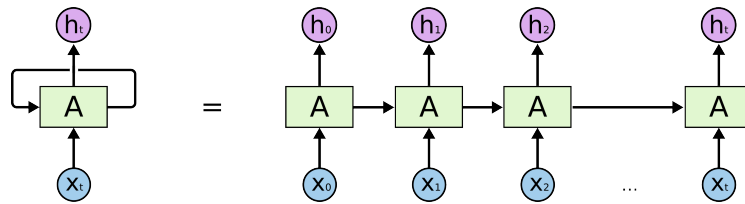


Figura 1.1: Funzionamento di una RNN.¹

I neuroni di una RNN possono essere definiti **memory cell** in quanto preservano lo stato attraverso il tempo. Le memory cell possono essere elementari, con un solo stato e una funzione di attivazione, o più articolate, come LSTM o GRU.

1.1.1 LSTM

Nel caso di LSTM lo stato è diviso in due vettori:

- **h**, lo stato a breve termine, che mantiene le informazioni su i dati di input più recenti.
- **c**, lo stato a lungo termine, che tiene in considerazione l'intera sequenza analizzata.

¹<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Intuitivamente ciò che fa la cella di memoria è riconoscere un input rilevante, immagazzinarlo nel vettore \mathbf{c} e preservarlo per il tempo necessario finché non verrà estratto per combinarsi con \mathbf{h} o verrà eliminato.

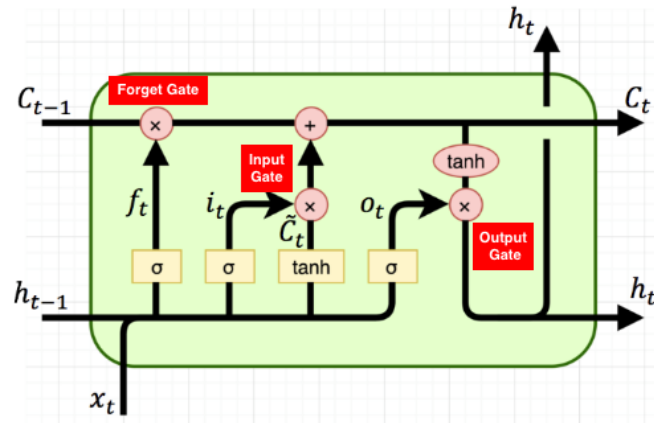


Figura 1.2: Struttura della memory cell LSTM.²

Il vettore $\mathbf{c}(t-1)$ attraversa prima un *forget gate* che rilascia parte delle informazioni e poi un *input gate* che ne aggiunge alcune provenienti dai nuovi dati.

I vettori $\mathbf{h}(t-1)$ e $\mathbf{x}(t)$ invece sono presi in ingresso da quattro strati di neuroni:

- $\mathbf{g}(t)$, con funzione di attivazione di tangente iperbolica, che è lo strato fondamentale di analisi dell'input presente in tutte le RNN.
- I **Gate Controller**, con funzione di attivazione logistica con dominio $[0,1]$ che indicano rispettivamente:
 - $f(t)$: che percentuale dello stato a lungo termine deve essere "dimenticata", infatti è utilizzata del *forget gate*.
 - $i(t)$: che percentuale dell'input deve essere aggiunta allo stato a lungo termine, infatti è presa in ingresso dall'*input gate*.
 - $o(t)$: che percentuale di $c(t-1)$ deve essere utilizzata per la produzione dell'attuale output $h(t)$, questa metrica è usata nell'*output gate*.

²<https://towardsdatascience.com/grus-and-lstm-s-741709a9b9b1>

1.1.2 GRU

GRU è una versione semplificata della cella LSTM che ha una performance molto simile pur rimanendo più leggera dal punto di vista computazionale.

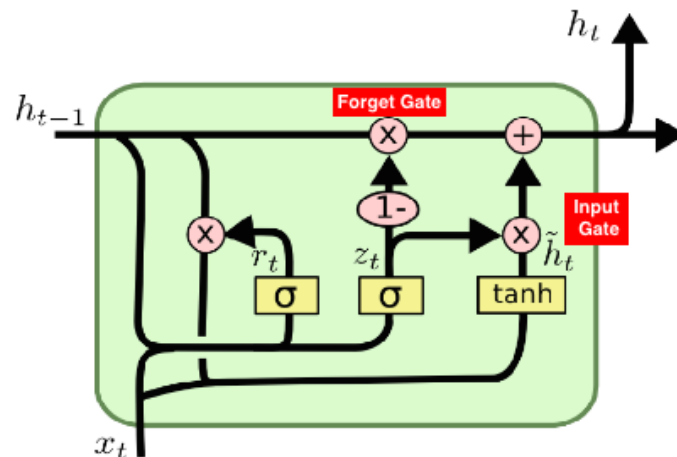


Figura 1.3: Struttura della memory cell GRU.³

Ci sono varie differenze rispetto a LSTM:

- Lo stato è registrato in un singolo vettore \mathbf{h} che mantiene sia quello a breve che a lungo termine.
- È presente un solo gate controller $\mathbf{z}(t)$ che sostituisce i tre di LSTM fornendo una certa percentuale al *forget gate* e il suo complementare all'*input gate*.
- C'è un nuovo gate controller $\mathbf{r}(t)$ che dosa quanto dello stato precedente $h(t-1)$ verrà mostrato alla funzione logistica $g(t)$.

1.2 Transformer

Il Transformer, al contrario delle architetture mostrate in precedenza, non utilizza la *ricorrenza* ma l'*attenzione* ed è descritto nel celebre articolo "Attention is all you need" [2].

In Figura 1.4 ne è riportata la struttura principale, composta da un **Encoder** sulla sinistra, che prendere in input la frase originale creandone una

³<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

rappresentazione unica ed un **Decoder** sulla destra, che prende in input i token di output già generati per predire il successivo.

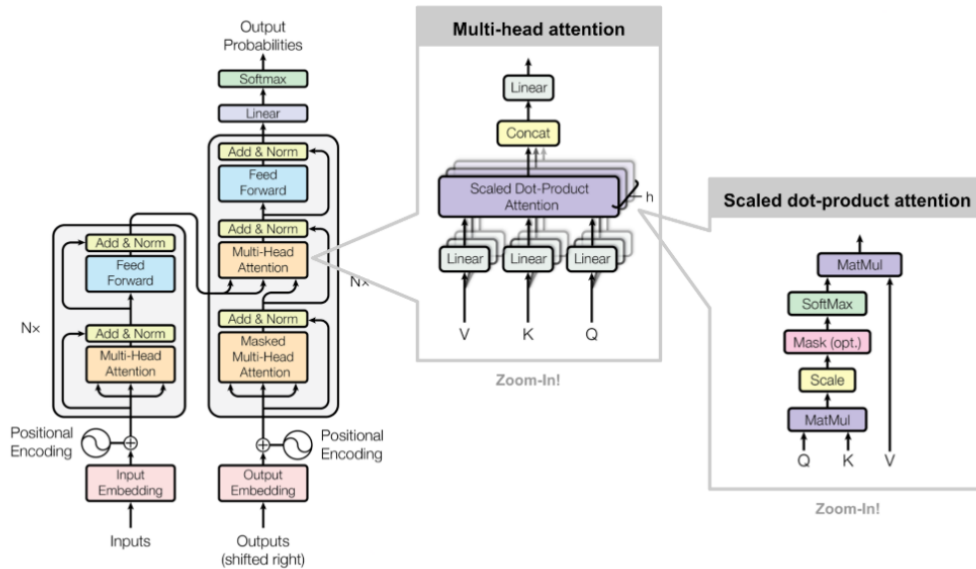


Figura 1.4: Architettura del Transformer.⁴

Prendendo come esempio un elementare problema di traduzione da inglese ad italiano della frase "*In the garden there are a lot of daisies.*", l'Encoder prenderà in input l'intera frase in inglese mentre il Decoder nel momento in cui dovrà tradurre un certo token, ad esempio "*there*", prenderà in input la porzione di frase già tradotta, ovvero "*Nel giardino*". Intuitivamente ciò indicherà al Decoder a quale parte della rappresentazione generata dall'Encoder dovrà *prestare attenzione* nello specifico passaggio.

Prima di accedere ad Encoder o Decoder, gli input vengono sottoposti a una fase di pre-processing, nella quale:

- gli input sono suddivisi in token;
- i token vengono sostituiti dai rispettivi embedding, solitamente estratti da una base di conoscenza;
- agli embedding viene aggiunto il *positional encoding* per associargli le informazioni relative alla collocazione del token nella frase (ciò potrebbe alterarne il significato, ad esempio se il token è preceduto da una negazione).

⁴[2]

Gli strati iniziali di tipo **Multi-Head Attention** presenti nei due componenti generano le versioni codificate degli input, ma è il secondo layer del medesimo tipo presente nel Decoder che attua la predizione vera e propria. Tale strato, come si può notare in Figura 1.4, prende in input le matrici **V**, **K** e **Q**, ovvero i valori V e le chiavi K provenienti dall'Encoder e le query Q prodotte dal Decoder. Se ad esempio questa struttura avesse come obiettivo la rappresentazione di una persona, i valori sarebbero i dati riguardanti tale soggetto (ad es. *[Silvia, 35 anni, 172 cm]*), le chiavi costituirebbero le etichette che permettono di individuare tali valori (ad es. *[Nome, età, altezza]*) e la query indicherebbe su quale caratteristica della persona si vuole porre attenzione (ad es. *Nome*) in un determinato passaggio.

Si possono immaginare le chiavi K come vettori in uno spazio multidimensionale, ognuno associato ad un valore, e Q come un'ulteriore vettore in tale spazio. Computando l'Attention si fa il prodotto scalare $Q * K$, trovando la **similarità** fra Q ed ogni K , la quale viene scalata e alla quale viene applicata la funzione softmax per selezionare la coppia K - Q con un'affinità maggiore. Infine la moltiplicazione per V è necessaria a reperire il valore associato alla chiave individuata dal passaggio precedente.

Gli strati Multi-Head Attention in realtà sono formati da tante *head*, delle quali ognuna cattura una specifica relazione fra i token calcolando l'Attention con tre matrici di proiezione diverse. I risultati prodotti da tutte le *head* vengono poi concatenati in uscita.

Una delle differenze sostanziali rispetto alle RNN è che ogni nuovo token è predetto in un'iterazione diversa e la previsione non è realizzata con un meccanismo di *back-propagation* che percorre a ritroso l'intera frase, bensì individuando in modo diretto la zona di interesse nell'espressione.

Questo schema è utilizzato in buona parte dei Language Model presentati nei capitoli successivi, ad esempio GPT-2 e GPT-3, Turing-NLG, XLM-R, ViBERT, Meena e BlenderBot.

Capitolo 2

Language Models

Un Language Model è un modello addestrato per predire la probabilità di una sequenza di parole.

Negli ultimi anni svariati gruppi di ricerca (OpenAI, Facebook AI Research, Google Research, Microsoft Research) hanno pubblicato dei Language Models che generalmente:

- utilizzano le architetture descritte nel Capitolo 1;
- sono addestrati su un esteso dataset di **pre-training**;
- sono ulteriormente allenati su un dataset di **fine-tuning** di dimensioni ridotte per adeguarsi ad ogni specifico task da svolgere.

Le pubblicazioni più recenti [3] indicano che aumentando sempre più il numero di parametri che condizionano la previsione e l'ampiezza del dataset di pre-training si riesce a rendere il modello adattabile a svariati contesti di applicazione.

Un ambito in cui è possibile utilizzare i Language Models è la generazione di testo. Ad esempio, è stato verificato da uno studio¹ che il modello **GPT-2** [4] è in grado di produrre **fake news** convincenti riguardanti dei complessi problemi sulle politiche estere.

In tale studio, a partire da *due* paragrafi estratti da una notizia reale (nello specifico un articolo del New York Times, con un alto livello di affidabilità), sono stati generati 20 testi fra i quali se ne sono scelti 3 da inserire come *terzo* paragrafo, ottenendo quattro versioni dell'articolo (una originale e tre false).

È stato effettuato un sondaggio su 500 persone ed è risultato che uno degli articoli fittizi ha ottenuto una **credibilità del 72%** contro l'83% dell'articolo autentico ed anche quello con una valutazione più bassa è stato giudicato credibile dal 58% degli individui. Inoltre circa il 25% delle persone ha affermato

¹<https://www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots>

che avrebbe condiviso la notizia, tale dato non cambia confrontando articoli reali e falsi.

GPT-2 può generare post su social media, spam e fake news a partire da un breve prompt riguardante un evento fittizio: questo modello potente può essere pericoloso se usato in modo sbagliato, perciò ne è stata rilasciata solo una versione '*watered-down*', che presenta meno **rischi**. I dati dello studio dimostrano uno dei risvolti negativi della pubblicazione di Language Model: se usati per la produzione di fake news, potrebbero popolare il web facendo perdere al pubblico la fiducia nei media o peggio, riuscendo ad ingannarlo e rendendolo influenzabile da chiunque in modo veloce ed economico.

Nei prossimi paragrafi saranno esaminati alcuni Language Models pubblicati recentemente.

2.1 GPT-2 e GPT-3

GPT-2 [4] è un Language Model basato sul Transformer, ne sono presenti fra i 12 ed i 48 layer nelle varie versioni pubblicate, prima di ognuno dei quali è effettuata la normalizzazione dell'input. Il modello originale è addestrato su 40GB di testo non supervisionato estratto da internet ed ha 1,5 miliardi di parametri, ma ne sono state rilasciate solo le versioni "watered-down" con rispettivamente 117 e 345 milioni di parametri.

GPT-3 [3] è l'evoluzione di GPT-2, utilizza **175 miliardi** di parametri (al momento il Language model con il maggior numero) ed analizza più testo di quanto ne possa vedere un umano nell'arco di una vita intera. È stato addestrato sul contenuto del database CommonCrawl fra il 2016 e il 2019 con una dimensione iniziale di 45 TB che è stata ridotta a 570GB dopo un'operazione di filtraggio. Sono stati aggiunti al training set anche alcuni dataset più piccoli ma di qualità più elevata, ovvero WebText2, Books1, Books2 e Wikipedia. A differenza di BERT (descritto nel paragrafo 2.2), uno dei maggiori concorrenti, è un modello autoregressivo e non bidirezionale.

L'avanzamento rispetto a GPT-2 risiede nella capacità di risolvere tanti task diversi *zero-shot* (al primo colpo) o *few-shot* (dopo alcuni tentativi), senza essere fine-tuned rispetto ad uno specifico task. Ciò è reso possibile dall'enorme quantità di informazioni acquisite durante il pre-training che permettono al modello di sviluppare l'abilità di individuare pattern, che utilizza per il riconoscimento dei task.

Il fine-tuning è stato sostituito dall'**in-context learning**, consiste nel fornire al modello, oltre al testo di input, l'indicazione specifica del task da svolgere (il prompt) o alcuni esempi di tale compito, in modo da proiettare i token dell'

espressione di input in un sottospazio latente che aderisce al task dato, ad esempio:

*Q: "What is the Italian translation of dog?"
A: translation*

*Q: "What is the Italian translation of dog?"
A: [house => casa; day => giorno; cat => gatto, sun => sole]*

Rimuovere il fine-tuning ed usare i modelli *few-shot* è desiderabile perchè:

- Per molti dei task è difficile recuperare una training set supervised di dimensioni adeguate.
- L'overfitting (quindi le predizioni spurie) aumentano con l'espressività del modello e l'intricatezza della distribuzione del set di training. I modelli devono essere ampi per assorbire una grande quantità di informazioni durante il pre-training, poi però i training set di fine-tuning risultano estremamente limitati in confronto e perciò la loro distribuzione è altalenante. L'overfitting può risultare in un'esagerazione della performance sul task.
- Le persone non fanno uso di fine-tuning ma sono appunto dei *few-shot* learner, riescono addirittura ad aggiungere contenuto *durante* un lungo dialogo. Cercare di riprodurre questa qualità potrebbe essere l'idea migliore per far sì che il Language Model si comporti come un essere umano.

GPT-3 è stato valutato su 24+ datasets NLP e nuovi task appositamente creati per verificare il suo rapido adattamento al compito che deve affrontare. Per ogni task, è stato valutato con few-shots, one-shot o zero-shots. La performance migliora aumentando il numero di samples, aggiungendo una descrizione del task (il prompt) ed aumentando i parametri del modello, come si può notare dalla Figura 2.1. Dopo alcuni tentativi il modello è riuscito a raggiungere o superare il SOTA in vari task.

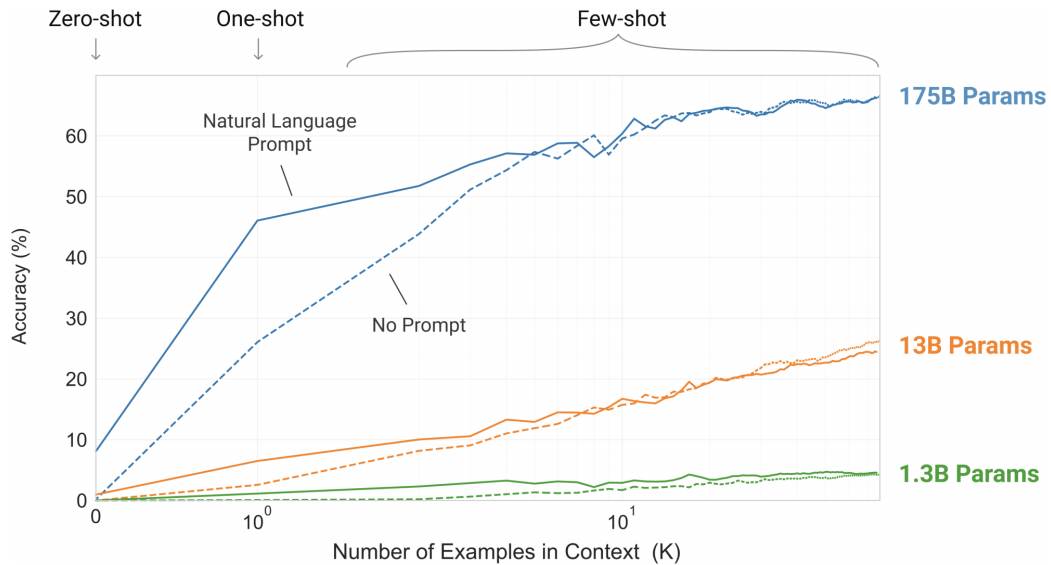


Figura 2.1: Miglioramento della performance di GPT-3 inserendo il prompt ed aumentando i parametri in funzione del numero di tentativi.²

Di seguito sono riportati alcuni dei task su cui GPT-3 è stato messo alla prova:

- **LAMBADA** [5], che consiste nella predizione dell'ultima parola di una frase. Il modello ha ottenuto un'accuratezza *zero-shot* del 76%, l'8% in più rispetto allo stato dell'arte, che diventa del 86% nel caso *few-shot*.
- **Closed Book Question Answering**, grazie all'Information Retrieval riesce a reperire porzioni di testo dal dataset su cui è stato addestrato che rispondono alla domanda. Anche in questo caso supera lo stato dell'arte.
- **SuperGLUE** [6], l'evoluzione di GLUE che comprende vari tipi di Task NLU. GPT-3 si è dimostrato debole nei task che prevedono il confronto fra più frasi.
- **News Article Generation**, è in grado di generare articoli difficilmente distinguibili da quelli reali per un umano, come dimostrato dallo studio descritto all'inizio di questo Capitolo.

Uno dei rischi nell'utilizzo di un dataset ampio come quello di GPT-3 è quello di **Memorization**: facendo un paragone con l'atto di studiare un libro, che il modello "*impari a memoria*" il testo di training invece di estrarvi le

²[3]

informazioni attraverso il *ragionamento*. Date le dimensioni del training set c'è il pericolo che ci siano delle sovrapposizioni con alcuni dei test-set dei vari task con cui il modello si è confrontato, in tali casi le valutazioni dell'accuratezza non sarebbero corrette. Per far fronte a questo problema è stato fatto anche uno studio della *data contamination*, misurando la contaminazione dei dati di pre-training e quantificandone gli effetti.

Il codice di GPT-3 attualmente **non è fruibile** in quanto pare si stia valutando di trasformarlo in un prodotto commerciale. Inoltre bisogna tenere in considerazione i rischi annessi al suo utilizzo a fini ingannevoli e le conseguenti responsabilità da prendersi nel renderlo disponibile a chiunque, medesimo motivo per cui il rilascio della versione precedente, GPT-2, aveva richiesto molto tempo.

2.2 BERT

BERT [7], ovvero **B**idirectional **E**ncoder **R**epresentation from **T**ransformers è un Language Model presentato da Google AI Language a maggio 2019 e, come si può evincere dal nome esteso, utilizza l'Encoder del Transformer.

Per utilizzare il modello, le frasi di input vengono modificate aggiungendo le stringhe [SEP] fra l'una e l'altra e [CLS] all'inizio di ogni sequenza e successivamente trasformate nella somma di tre embeddings come rappresentato in Figura 2.2, nello specifico:

- **Token embedding**: relativo alla parola intesa come serie di caratteri, nel caso di BERT recuperato attraverso il metodo di tokenization WordPiece;
- **Segment embedding**: relativo alla frase in cui si trova la parola (A o B);
- **Position embedding**: relativo alla posizione in cui si trova la parola nella sequenza di frasi.

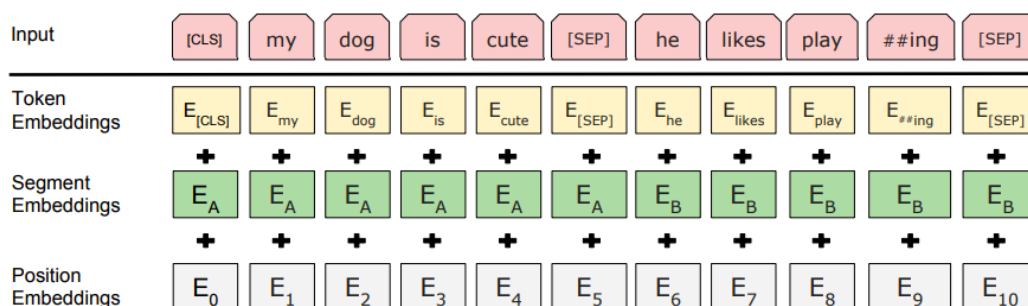


Figura 2.2: Input embeddings utilizzati da BERT.³

Al contrario di GPT-2, BERT riesce a sfruttare la **self-attention in entrambe le direzioni**. Ad esempio, nella frase "*I love to eat pizza*" in GPT-2 il secondo token otterrebbe la self-attention solo con "*I*" e "*love*", ovvero se stesso e tutti gli elementi che lo precedono, mentre in BERT viene calcolata l'attention rispetto all'intera frase. Ciò è realizzato fornendo al sistema la frase sia nell'ordine originale che in quello inverso ed allenando il modello su task di training mirati, diversi dalla banale predizione della parola successiva impiegata nell'addestramento di GPT-2.

Nello specifico i task utilizzati sono:

- **Masked Language Model** - Task che ha l'obiettivo di predire una parola occultata all'interno della frase, ad esempio "*I love to [MASK] pizza*". Circa il 15% dei token sono nascosti, dei quali l'80% esplicitamente con [MASK], il 10% con un parola random e il 10% con il token originale (non sono realmente masked).

Ciò è importante perchè se il modello si fosse abituato a predire solo le parole adatte a sostituire [MASK] avrebbe cercato la specifica stringa [MASK] per qualsiasi task ad inference time e avrebbe imparato a contestualizzare solo il 15% dei token, quelli nascosti, senza acquisire nessun tipo di conoscenza riguardo al restante 85%. Utilizzando anche altre stringhe per *mascherare* le parole il modello si concentra maggiormente nel rappresentare tutti componenti della frase.

- **Next Sentence Prediction** - Task con lo scopo di indicare se due frasi proposte possono susseguirsi l'una all'altra o meno, tale predizione viene sostituita al token [CLS]. Ad esempio, le due espressioni in Figura 2.2 avrebbero un riscontro positivo.

2.3 Turing-NLG

Turing-NLG è un language model pubblicato da Microsoft che utilizza **17 miliardi** di parametri, supera il SOTA in svariati task ed eccelle applicato a task pratici come la summarization, il question answering o LAMBADA.

Il modello è formato da 78 strati di Transformers, ciascuno con 28 heads e un'hidden size di 4256.

Ha come obiettivo la produzione ex-novo di risposte dirette, accurate e fluenti come quelle ideate da un cervello umano, al contrario di LM precedenti che estraevano parti del testo di training per inserirle nella risposta o sintesi facendola apparire innaturale.

³[7]

Prendendo come esempio il task di **question answering**, ciò che viene solitamente proposto all'utente a seguito di una ricerca sul web è un frammento di testo con la risposta evidenziata. Al contrario, T-NLG analizza il paragrafo reperito e genera una frase sintetica di senso compiuto, ciò è molto utile per sistemi come AI assistants che devono dare un risposta vocale puntuale all'utente.

Inoltre T-NLG riesce a fare zero-shot question answering, ovvero a rispondere a domande di cultura generale senza analizzare alcuna card in quanto si può affidare sulla conoscenza acquisita con il pre-training (esempio in Figura 2.3).

When did WW2 end?	WW2 ended in 1945.
How many people live in the US?	There are over 300 million people living in the US.

Figura 2.3: Esempio di zero-shot question answering di T-NLG.⁴

La correttezza grammaticale ed effettiva delle risposte è stata valutata da esseri umani (dato che non esistono ancora metriche soddisfacenti) e si è ottenuto un punteggio migliore rispetto a modelli che usano LSTM pur addestrando T-NLG su meno epoche di fine-tuning.

Il modello è ottimo anche nella **summarization**: solitamente un problema per la risoluzione di questo task è la scarsità di materiale sintetizzato a disposizione, ma grazie alla sua abilità di comprendere il testo T-NLG necessita solo di un minimo numero di istanze supervisionate.

Il codice di questo LM è reso disponibile solo su richiesta e non è stato possibile reperirlo.

2.4 Language Models multi-lingua

2.4.1 XLM-RoBERTa

XLM-R [8] è un Language Model reso pubblico dal gruppo di ricerca FAIR (Facebook Artificial Intelligence Research) ed è specializzato nella risoluzione di task **cross-lingual** (come la traduzione) senza diminuire le performance nell'utilizzo su un singolo linguaggio. È particolarmente efficiente se utilizzato su linguaggi *low-resource*, ovvero su cui ci sono pochi documenti a disposizione (as es. Swahili, Urdu).

Il modello è stato addestrato su un dataset multilingua di più 2 TB di dimensione, CommonCrawl, nel quale è presente del testo in 100 diverse lingue.

⁴<https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/>

Come si può notare in Figura 2.4 ciò rappresenta un netto miglioramento rispetto al dataset comunemente utilizzato, Wikipedia.

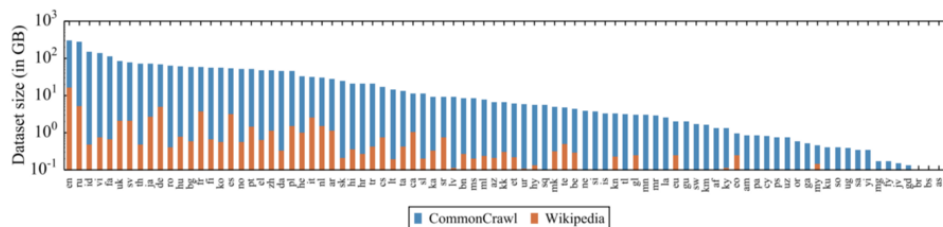


Figura 2.4: Confronto fra la quantità di testo presente in varie lingue nei dataset Wikipedia e CommonCrawl.⁵

XLNet è basato sull'architettura del Transformer ed utilizza un singolo Encoder condiviso per generare l'embedding di ogni espressione a prescindere dall'idioma, perciò è presente una sola rappresentazione del suo significato per tutti i linguaggi: l'apprendimento acquisito a partire da una frase in una certa lingua è utilizzabile anche per le altre grazie al **Transfer Learning**, similmente a ciò che accade in USE (paragrafo 3.2).

RoBERTa è la tecnica utilizzata per fare *pre-training*, evoluzione di BERT che riesce a lavorare su una maggior quantità di dati. Si è dimostrato che tale addestramento multilingua unsupervised su larga scala, nello specifico orientato all'identificazione di parole masked, riesce ad ottenere o superare il SOTA in cross-lingual classification, question answering e sequence labeling.

Durante lo sviluppo di XLNet si è notato un *trade-off* riguardante il numero di lingue analizzabili da un modello a dimensione fissa: incrementando la varietà di idiomi la performance cresce fino a un certo punto, dopo il quale inizia a calare. Il trade-off è stato denominato **Curse of Multilinguality** e può essere parzialmente risolto aumentando la dimensione del modello.

2.4.2 LID

Per scegliere quale modello utilizzare su una porzione di testo, Apple necessita di un ulteriore modello che esegua la LID (**Language IDentification**) per individuare l'idioma in cui esso è scritto. Ciò può risultare complesso specialmente se la stringa è particolarmente breve, come spesso accade nel caso di messaggi inviati dagli utenti.

Riconoscere la lingua è importante ad esempio per attivare il sistema di *autocorrezione* con il lessico corretto se la digitazione in input è multi-lingua,

⁵[8]

ma anche per fare TTS, POS, NER o classificare documenti. I problemi più complessi che i sistemi LID devono affrontare sono la già citata scarsità di caratteri a disposizione e la presenza dei medesimi termini in più di una lingua.

In passato sono state proposte soluzioni basate sul **lessico**, che non sono pratiche nei dispositivi embedded, sulla **sintassi**, che sono servibili solo avendo a disposizione documenti molto lunghi o **generative** con n-gram, che soffrono di indipendenza condizionata e non sono sfruttabili su sequenze di pochi caratteri.

L'innovazione di Apple nell'affrontare la LID si trova nel considerarlo come un problema di tagging a livello di carattere ed usare un'architettura **LSTM** bidirezionale per risolverlo. Il modello generato è ridotto sia per quanto riguarda le dimensioni (fra il 40 e l'80%) che l'error rate (fra il 15 e il 60%) per le varie lingue e riesce scalare senza raggiungere un'estensione eccessiva.

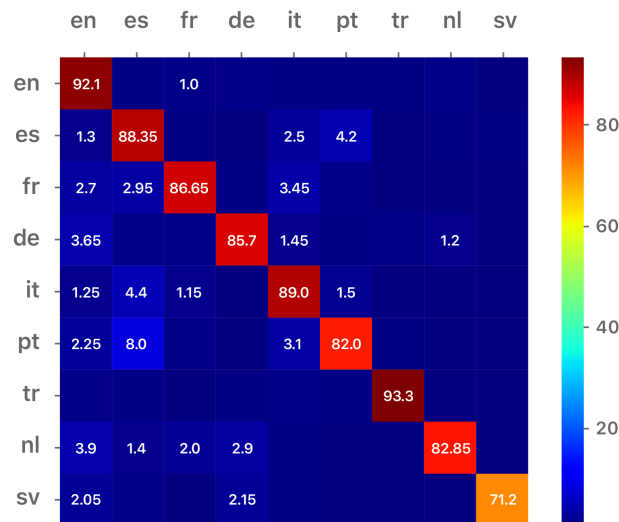


Figura 2.5: Matrice di confusione del problema LID usando bi-LSTM.⁶

Come possiamo notare dalla matrice di confusione in Figura 2.5, il modello ottiene degli ottimi risultati nel riconoscimento di svariate lingue. Per quanto riguarda l'italiano, raggiunge un'affidabilità dell'89% e la lingua con cui, come prevedibile, viene confuso in misura maggiore è lo spagnolo, seguito da portoghese, inglese e francese.

⁶<https://machinelearning.apple.com/research/language-identification-from-very-short-strings>

2.5 Language Models multi-modali

Un LM multimodale è un modello che non elabora semplicemente informazioni testuali, ma anche quelle derivate da altri mezzi quali un'immagine o un video, associando un significato ad ogni istanza formata da dati provenienti da *varie fonti* che gli viene fornita.

Prendendo come esempio il caso del meme, al modello viene richiesto di interpretare simultaneamente l'immagine e le parole su essa riportate come fossero **un'unica entità**. Altri problemi multimodali sono la capacità di rispondere a domande riguardanti un'immagine o la generazione della caption di una figura, ad esempio per inserirla nell' `alt` delle immagini presenti sui siti web.

2.5.1 Hateful Memes Challenge

Il 12 maggio 2020 Facebook AI Research ha lanciato la Hateful Memes Challenge [9] una competizione di **classificazione binaria multimodale** con l'obiettivo di determinare se un qualsiasi meme contiene messaggi d'odio utilizzando le informazioni estrapolabili dalla sorgente **testuale** e da quella **visuale**. Il SOTA delle AI al momento della pubblicazione è del 65% mentre quello umano, la cui comprensione è naturalmente multimodale, è dell'85%. Ad oggi, il 23 luglio 2020, uno dei partecipanti alla competizione ha ottenuto un'accuratezza del 77%, alzando di molto l'asticella in breve tempo.

Spesso modelli che tentano di fare interpretazioni multimodali sono ingannevoli perchè in realtà si basano solamente sul testo e non prendono in considerazione le immagini, avendone una comprensione limitata o nulla. Nel caso dei meme è fondamentale combinare le due fonti perchè la modifica del testo o della figura può alterare completamente il significato.

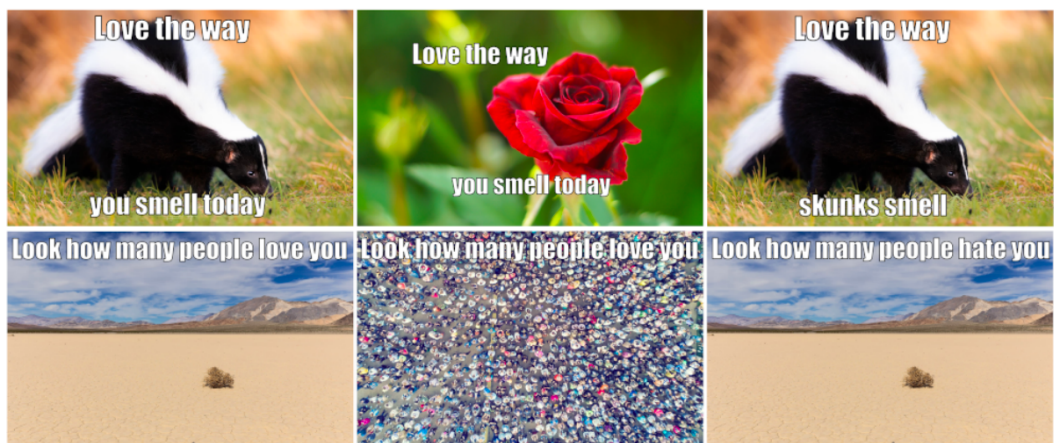


Figura 2.6: Esempi di Hateful Memes con rispettivi benign confounders.⁷

Per contrastare questo problema nel dataset fornito da Facebook sono presenti dei "**benign confounders**" (Figura 2.6), ovvero delle versioni alternative per ogni Hateful Meme in cui l'immagine o la scritta è stata modificata rendendo il contenuto innocente. Questo obbliga il modello a prestare attenzione ai dettagli e lo costringe ad essere realmente multi-modale, se infatti si basasse solo sull'immagine o sulla scritta classificherebbe l'originale e un benign confounders allo stesso modo.

Hate Detection

I messaggi di odio sono generalmente complessi da distinguere per un'AI. È difficile individuarli sia nel caso siano contenuti in un meme, sia in un post con un'immagine una caption, sia in semplici commenti che fanno uso di ironia, di diverse lingue miscelate o che sono stati scritti da un utente consapevole degli strumenti di detection, che quindi cerca di non essere individuato.

Per tutelare gli utilizzatori di un social è importante identificare i contenuti violenti o di odio, perciò FAIR sta cercando di migliorare i propri sistemi di *hate detection*, che a maggio 2020 hanno un'accuratezza dell'88.8%.

Nei primi 3 mesi del 2020 Facebook ha agito su 9.6 milioni di contenuti che violavano le policies (+3.9 milioni rispetto al 2019) grazie a due principali avanzamenti:

- Una comprensione del testo più profonda, anche di frasi articolate, tramite l'utilizzo del modello XLM-R descritto nel paragrafo 2.4.1.
- L'analisi di *ogni* aspetto riguardante un post (testo, immagini, commenti, likes) e la costruzione di una **rappresentazione universale** del suo contenuto. Ciò è realizzato tramite l'addestramento del modello con self-supervised learning a livello di post al fine di ottenere una comprensione più profonda di ogni contenuto multimodale. Tale rappresentazione è fondamentale in casi altamente ambigui come quello mostrato in Figura 2.7.

⁷[9]



Figura 2.7: Esempio di post che è necessario rappresentare come unica entità prendendo in considerazione sia l'immagine che il testo.⁸

Sebbene le policy di FB siano chiare è complicato distinguere i contenuti da eliminare, anche perchè è importante non classificare un contenuto come Hate Speech erroneamente. Ciò infatti andrebbe a ledere la libertà di espressione degli utenti che è fondamentale preservare. Ad esempio, le risposte ad un contenuto d'odio sono facilmente confondibili con il contenuto stesso ed è necessario che il modello impari a riconoscerle. Ulteriori problemi sono la scarsità di materiale a disposizione per il training e la qualità di tale materiale, la cui classificazione è spesso sbilanciata a sfavore degli Hateful Meme.

2.5.2 ViLBERT

ViLBERT [10] è un modello multi-modale *task-agnostic* che apprende rappresentazioni di immagini combinate a linguaggio naturale.

Il modello è addestrato su un dataset formato da circa 3,3 milioni di immagini con le rispettive caption chiamato **Conceptual Captions**. Il training è eseguito su due task generati sfruttando la struttura del dataset stesso (denominati *proxy*), ma avvalendosi del Transfer Learning il modello può essere usato per risolvere altri problemi come il visual question answering o l'immagine captioning.

L'unione degli stimoli legati alla comprensione del linguaggio a quelli visuali può essere realizzato tramite il *pretrain-then-transfer*, che consiste nell' eseguire il pre-training di due modelli separati addestrati rispettivamente con dataset contenenti testo ed immagini su task semplici ed infine combinarli. Questo approccio si è rivelato poco efficiente, il modello finale contiene informazioni scarse e distorte sui dati visio-linguistici, risulta ben allenato in entrambi i singoli

⁸<https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/>

ambiti ma non a sufficienza sulla loro correlazione. Un approccio alternativo, che si è rivelato migliore, è l'utilizzo di un **dataset di addestramento multimodale** (Conceptual Caption) sui due task *proxy*. Tali task nel caso di ViLBERT sono la predizione di parole in una frase o porzioni in un'immagine *masked* e la classificazione di coppie figura-descrizione in base alla loro correttezza.

Le immagini e i token di input vengono forniti al sistema tramite due stream separati che interagiscono nel Transformer Co-attentional (in Figura 2.8) in modo da riuscire a cogliere sia i dettagli di ogni elemento di input specifico sia la loro correlazione. Il modello base prodotto è *task-agnostic*, è adattabile a vari compiti aggiungendo un classificatore ed ha superato lo stato dell'arte in quattro task (visual question answering, visual commonsense reasoning, referring expressions e caption-based image retrieval).

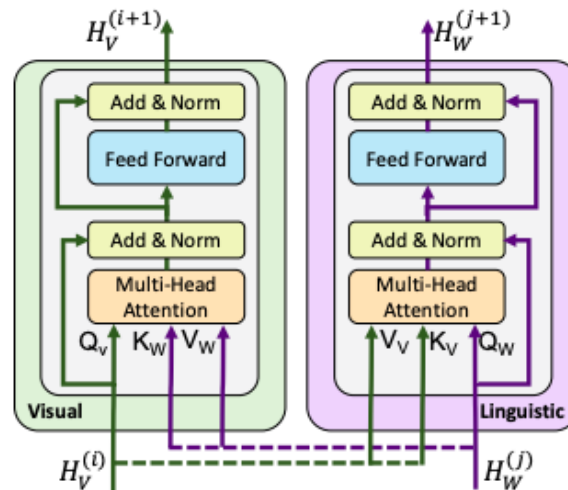


Figura 2.8: Architettura del transformer co-attentional usato da ViLBERT.⁹

⁹[10]

Capitolo 3

Chatbot Retrieval-Based

L'approccio progettuale più intuitivo per realizzare una chatbot è la **macchina a stati finiti**, basata su una gerarchia di regole che facilmente diventa molto complessa per poter gestire tutte le possibili deviazioni che la conversazione potrebbe prendere e difficilmente riesce a farlo in modo efficace. Questo metodo è ottimo solo nel caso in cui il dialogo segua l'*happy path*, ovvero se messaggio dell'utente ricade nell'insieme dei possibili messaggi presi in considerazione dallo sviluppatore per quel momento del dialogo.

Un'evoluzione di questo metodo è quello usato dalle chatbot **retrieval-based**, che non seguono delle regole restrittive ma hanno a disposizione una repository con numero finito di possibili risposte e fanno uso di Machine Learning per valutare qual è la risposta migliore tenendo conto del contesto di conversazione ed associando ad ognuna un punteggio.

Input	Response
I saw a really good band last night.	It often rains in the winter. On Thursdays we like to go out. They played upbeat dance music. The tree looks good to me. Did you get a new car? My son likes to windsurf. Looking forward to lunch.

Figura 3.1: Esempio di selezione della risposta più adeguata da un set predefinito operata da una chatbot retrieval-based.¹

Esistono infine le chatbot **generative** (descritte ampiamente nel Capitolo 4), che non hanno accesso a una pool di risposte predefinite ma le producono

¹<https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html>

ad ogni iterazione facendo uso di Deep Learning. Il metodo è analogo ad una traduzione, però invece di trasformare un testo da una lingua da un'altra si trasforma la domanda di input nella risposta di output.

Chiaramente quest'ultima tipologia di AI per la conversazione è la più complessa dal punto di vista architetturale e la più imprevedibile ma anche quella che permette di ottenere dialoghi maggiormente soddisfacenti per l'utilizzatore, dando l'impressione di parlare con un essere umano.

3.1 Rasa

Rasa è un framework open-source che permette di costruire assistenti o chatbot retrieval-based performanti, resilienti ed adattabili al contesto della conversazione usando tecniche di Machine Learning. È apprezzabile per la sua accessibilità e flessibilità, infatti ha un funzionamento di base facilmente comprensibile ma allo stesso tempo permette ad uno sviluppatore più esperto di controllare e personalizzare tutti gli elementi che costituiscono la chatbot. Rasa è definito retrieval-based in quanto è compito del programmatore definire tutti i possibili intenti esprimibili dall'interlocutore.

Il framework è formato da due componenti:

- **Rasa NLU** che ha l'obiettivo di interpretare il messaggio dell'utente.
- **Rasa Core** che analizza il contesto della conversazione per scegliere l'azione più corretta da eseguire.

3.1.1 Rasa NLU

RASA NLU (*Natural Language Understanding*) è uno strumento NLP utilizzato dalle chatbot per classificare gli intenti ed eventualmente individuare delle entità nell'espressione dell'utente.

- L'**intento** indica lo scopo dell'utente, l'obiettivo che vuole raggiungere inviando il messaggio. Il testo di tale messaggio viene trasformato in un vettore dal quale si estrae l'intento, che sarà poi utilizzato per reperire una risposta adeguata.
- Le **entità** sono dei termini presenti nel messaggio che hanno il compito di determinare il contesto del dialogo, ovvero contengono informazioni che probabilmente sarà necessario tenere presenti durante lo sviluppo della conversazione.

Ad esempio, prendiamo in considerazione un possibile messaggio ricevibile da un bot:

"Vorrei acquistare un'automobile alimentata a metano di colore nero"

Rasa NLU processerà tale messaggio determinando intento ed entità e creando un dato strutturato per rappresentarlo:

```
{
  "intent" : "buy_car",
  "entities" : {
    "fuel" : "methane",
    "color" : "black"
  }
}
```

Rasa NLU, al contrario di Rasa Core, è utilizzata sempre per comprendere piccole porzioni di testo.

L'interpretazione è eseguita attraverso un modello che segue una Pipeline configurabile dallo sviluppatore ed è stato addestrato su delle frasi di esempio corrispondenti a vari *intenti*. La **Pipeline** si basa sul concetto *divide et impera*: divide il complesso problema di comprensione del linguaggio in sottoproblemi relativamente semplici, utilizza il Machine Learning per risolverli e concatena i risultati ottenuti fornendo l'output di un componente come input del successivo.

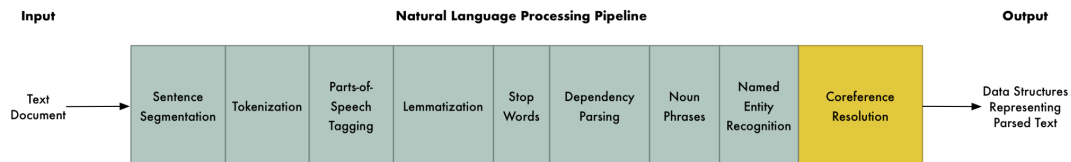


Figura 3.2: Esempio di Pipeline per la risoluzione di un problema di NLP.²

La Pipeline usata da Rasa NLU è solitamente composta da tre parti:

- **Tokenization** - Separazione di parole e segni di punteggiatura delle frasi di input, ogni elemento è chiamato *token*.
- **Featurization** - Associazione di ogni token ad un **word embedding**, un vettore che lo rappresenta. I word embedding possono essere:
 - **Pre-trained**: estratti da basi di conoscenza (presenti ad esempio in spaCy), in questo caso si potranno immediatamente fare le associazioni più comuni fra i token, utili se il Training Set è piccolo. Ad esempio, se nel Training Set è presente il messaggio "Voglio comprare delle mele" e bisogna predire l'intento per "Voglio acquistare

²<https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e>

una pera" il modello saprà che "*mele*" e "*pera*" fanno parte dello stesso gruppo di oggetti grazie ai pre-trained word embedding.

- **Supervised:** appresi attraverso l'addestramento sul Training Set, saranno più specifici sul dominio del problema. È fondamentale usarli se nel dominio in questione certe parole hanno un significato diverso rispetto a quello che avrebbero nel linguaggio comune.

Ad esempio, in inglese normalmente le parole "*balance*" e "*cash*" non hanno niente a che fare l'una con l'altra. Se il contesto del problema riguarda le transazioni bancarie tali termini assumono invece un significato simile ed è importante che il modello avvicini i loro word embeddings.

- **Intent Classification** - Assegnazione di una classe, rappresentata da un intento, ad ogni query (che a questo punto è vista del modello come una sequenza di word embedding).

È possibile inserire anche una fase di Named Entity Recognition (**NER**) che ha come obiettivo di individuare tutti i nomi non comuni presenti nel testo ed assegnare ad ognuno un'etichetta con il concetto che rappresenta. Alcuni esempi di tag sono *Persona*, *Luogo Geografico*, *Azienda*, *Prodotto* ma anche *Data* o *Prezzo*). I termini identificati dalla NER solitamente sono entità di interesse nel dominio del problema.

3.1.2 Rasa Core

Rasa Core decide come agirà la chatbot di conseguenza all'intento individuato da Rasa NLU, tutto ciò che il bot può dire o fare è chiamato *action*. Usa un modello predittivo ottenuto con tecniche di Machine Learning che:

- I. prende in **input** *l'intento*, un *dataset di addestramento* e *parte della conversazione* con l'utente
- II. sfrutta un **rete neurale LSTM** implementata in Keras
- III. fornisce in **output** una o più possibili *action*, ognuna associata ad una percentuale che indica la probabilità che sia la reazione migliore

Tale modello è personalizzabile indicando quali iperparametri si vogliono usare per le policies di apprendimento.

Il Training Set è formato da *stories*. Una *story* è una rappresentazione della conversazione fra utente e AI in cui i messaggi dell'utente sono espressi tramite *intent* (ed eventualmente alcune *entity*) mentre le risposte testuali dell'AI sono espresse tramite *action*.

L'approccio di Rasa Core è di tipo **Interactive Learning**: per ogni `action` che compie il modello riceve un feedback e nel caso esso sia negativo gli viene indicato quale sarebbe stato il comportamento adeguato da adottare. Il modello viene immediatamente aggiornato prendendo in considerazione l'interazione appena avvenuta come parte del Training Set. È fondamentale utilizzare questo metodo per gestire gli *unhappy paths*.

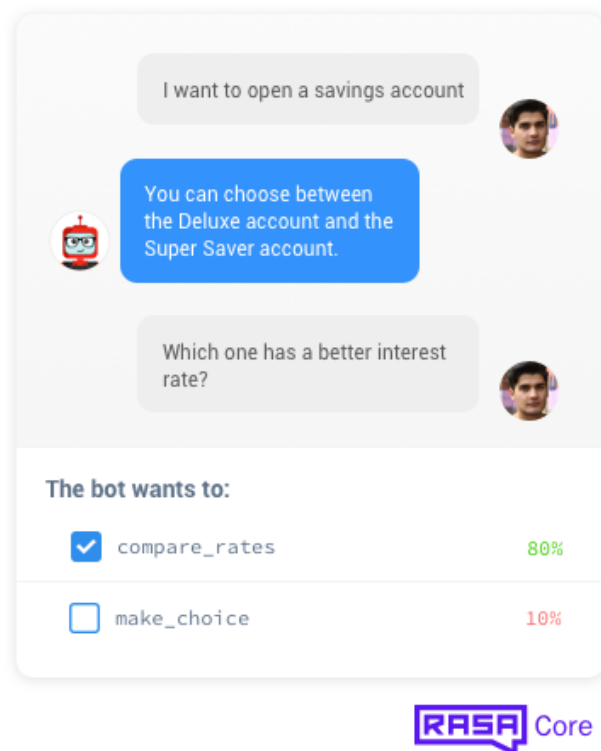


Figura 3.3: Esempio di Interactive Learning usato da Rasa Core.³

Questo metodo è un'evoluzione del **Reinforcement Learning**, nel quale il modello riceve una ricompensa ogni volta che riesce a soddisfare le esigenze dell'utente in un lasso di tempo adeguato ed ha come obiettivo di individuare la strategia (chiamata *policy*) che gli permette di massimizzare il compenso. Questo tipo di learning presenta varie problematiche:

- Necessita di un ampio Training Set di addestramento a disposizione dato che ogni conversazione può essere analizzata una sola volta.

Al contrario, Rasa Core aggiorna in continuazione il proprio Training Set

³<https://medium.com/rasa-blog/a-new-approach-to-conversational-software-2e64a5d05f2a>

mentre la chatbot viene utilizzata e sono sufficienti una dozzina di brevi conversazioni per ottenere una prima versione funzionante.

- Il premio viene attribuito al termine della conversazione, anche dopo decine o centinaia di decisioni, quindi è difficile determinare quali specifiche scelte abbiano fatto acquisire o perdere punti al modello.
Al contrario, Rasa Core fornisce un giudizio per ogni singola azione.

- Le valutazioni che assegnano le ricompense sono fornite da esseri umani, che sono notoriamente inaffidabili.
Anche Rasa Core è affetto da questa problematica, come possiamo notare dall'esempio in Figura 3.3.

3.2 Universal Sentence Encoder

USE [11] è un modello pubblicato a marzo 2018 da Google che utilizza un innovativo metodo di valutazione della similarità semantica di due espressioni basata non sulle singole parole che le compongono (approccio comunemente usato) ma sui loro messaggi di risposta. Infatti, due domande come "Quanti anni hai?" e "Qual è la tua età?" sono completamente diverse dal punto di vista dei singoli termini, eppure hanno lo stesso significato e vi si risponde allo stesso modo (altri esempi sono rappresentati in Figura 3.4). Le codifiche proposte da questo modello sono rappresentative non di una singola parola ma di un'intera espressione e hanno l'obiettivo di rappresentarne in maniera completa il significato per poter essere confrontabili in base a varie proprietà.

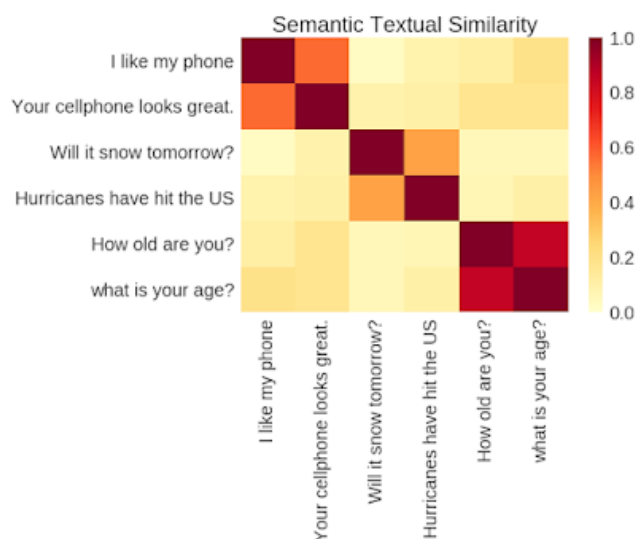


Figura 3.4: Valutazioni della similarità calcolate da USE.⁴

L'architettura del modello è formata da vari layer di Encoder che creano embeddings utilizzabili in una grande varietà di task. Come architettura per l'Encoder si è testata sia la DAN (descritta nel paragrafo 4.3.1) che quella del Transformer (discusso nel paragrafo 1.2). Quest'ultima risulta avere una performance migliore ma il suo tempo di computazione cresce all'aumentare della lunghezza delle frasi, mentre per la DAN rimane costante.

Successivamente, a luglio 2019, è stata pubblicata una versione multilingua di USE [12], che sfrutta risorse testuali in 16 lingue riconducendone le espressioni a vettori in un unico spazio semantico al fine di migliorare la performance in task che richiedono l'analisi della similarità fra testi in diverse lingue. La pubblicazione è comprensiva di tre modelli specializzati su diversi tipi di task, ognuno dei quali:

- contiene due spazi semantici prodotti da diversi Encoder, rispettivamente per le domande e le risposte;
- durante il training lega ogni possibile risposta con tutte le domande di input che può soddisfare;
- usa un database di vettori che codificano le query ottimizzato per risolvere il *nearest neighbor problem*.

Quando viene posta una nuova domanda al modello, esso la trasforma in un vettore, trova il suo *nearest neighbour*, ovvero la query presente nel dataset di training con cui ha una similarità maggiore, ed utilizza una delle risposte a cui essa è associata.

Il fatto che i modelli siano multilingua permette di fare **Transfer Learning**: se si fornisce ad esempio un input in thailandese esso sarà confrontato con tutti gli embedding delle query di training per trovare quella più simile, a prescindere dalla lingua da cui l'embedding è stato codificati. Ciò è particolarmente utile per i linguaggi *low resource*, per cui non è presente una grande quantità di testo su internet.

3.2.1 COVID-19 Retrieval-based Chatbot

Il modello Multilingual USE competente nel question-answering è stato utilizzato per realizzare una chatbot che potesse rispondere a domande riguardanti il COVID-19 ⁵ utilizzando le FAQ fornite dalla World Health Organization come base di dati per fare **Information Retrieval**. Tale sistema è in grado di

⁴<https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html>

⁵<https://towardsdatascience.com/how-we-created-an-open-source-covid-19-chatbot-c5c900b382df>

rispondere in modo soddisfacente ai quesiti posti in varie lingue, come si può notare in Figura 3.5.

Test Questions	Test Responses
Are animals contagious COVID-19?	Coronaviruses are a large family of viruses that are common in animals. Occasionally, people get infected with these viruses which may then spread to other people.
Are there medicine against the coronavirus?	Not yet. To date, there is no vaccine and no specific antiviral medicine to prevent or treat COVID-2019.
Can I breastfeed when I have COVID-19?	Yes. Women with COVID-19 can breastfeed if they wish to do so.
Should I stay inside the house?	Stay home if you feel unwell. If you have a fever, cough and difficulty breathing, seek medical attention and call in advance. Follow the directions of your local health authority.
Quali sono i sintomi del coronavirus?	Many people with COVID-19 experience only mild symptoms. This is particularly true at the early stages of the disease. It is therefore possible to catch COVID-19 from someone who has, for example, just a mild cough and does not feel ill.
Cosa dovrebbero fare le donne incinta?	Pregnant women and women who have recently delivered – including those affected by COVID-19 - should attend their routine care appointments.

Figura 3.5: Esempio di risposte fornite dall'AI fine-tuned sulle FAQ del WHO riguardanti il COVID-19.

Tuttavia, durante il testing della chatbot gli sviluppatori hanno rilevato alcuni comportamenti inadeguati:

- Tutte le domande contenenti il token "COVID-19" ottenevano un risposta inconsueta, presumibilmente perchè essendo una parola usata da pochi mesi non è presente come encoding all'interno del modello USE, pubblicato precedentemente alla pandemia. Ciò evidenzia una difficoltà nella comprensione di parole "*nuove*", che non sono mai state analizzate nella fase di training. Per risolvere questa complicazione si è inserito un sistema di sostituzione automatica della parola con "coronavirus", probabilmente presente nel dataset di addestramento perchè esistente da più tempo.
- USE è ottimo nella generazione di risposte brevi, composte al massimo da tre frasi. Nel caso del coronavirus questo rappresenta un difetto rilevante perchè può essere necessario fornire una spiegazione più completa ad alcune domande.

Capitolo 4

Chatbot Generative

In questo capitolo sono analizzate alcune chatbot ed AI assistant che fanno uso delle architetture presentate Capitolo 1 e di Language Models simili a quelli descritti nel Capitolo 2, i metodi per valutarle e paragonarle ed alcune problematiche ricorrenti.

4.1 Metodi di valutazione

L'esistenza di metriche di valutazione è fondamentale per confrontare i modelli proposti dai vari gruppi di ricerca. Le metriche possono essere:

- **automatiche**, ovvero applicabili direttamente al testo generato dal modello facendo confronti con un dataset supervised;
- **umane**, ovvero richiedono la presenza di un essere umano che giudichi di volta in volta se il modello produce risultati soddisfacenti, spesso confrontando le conversazioni sviluppate da varie chatbot.

4.1.1 BLEU

BLEU [13] è una metrica **automatica** di valutazione della Natural Language Generation semplice, economica ed indipendente dal linguaggio che è in grado di sostituire l'analisi umana in caso di necessità di un giudizio rapido o frequente.

Prendendo in considerazione il problema di *translation*, la qualità di una traduzione è misurata con la sua somiglianza con una o più traduzioni della medesima espressione prodotte da umani. Per valutare tale similarità viene utilizzata una metrica numerica di affinità, il **word error rate** (WER) e per addestrare BLEU un training set di traduzioni corrette. BLEU è calcolata come la media pesata delle corrispondenze fra sequenze di caratteri di lunghezza

variabile presenti sia nelle traduzioni candidate dal modello e che in quelle di riferimento.

4.1.2 BLEURT

BLEURT [14] è una metrica **automatica** recente che fa uso di **Transfer Learning** per rimanere consistente su un'ampia quantità di domini e nel tempo, catturando similarità non banali fra le frasi al fine di ottenere una qualità di giudizio simile a quella umana. Le metriche come BLEU, che confrontano una frase corretta e quella predetta dal modello conteggiando le sequenze di parole presenti in entrambi, considereranno più adeguate le frasi che esteriormente sembrano più simili. Spesso però questo non è sinonimo di maggior somiglianza dal punto di vista del significato, infatti un essere umano non ragiona calcolando in maniera rigida ma valutando proprietà come fluenza, correttezza grammaticale e lessicale o coerenza.

BLEURT è addestrata sul *WMT Metrics Shard Task*, un dataset pubblico di ratings contenente circa 260.000 giudizi riguardanti articoli di cronaca.

Si utilizza la rappresentazione contestuale delle parole (gli *embedding*) fornita da BERT ed uno schema di **pre-training**. Quest'ultimo è necessario perchè sperimentando si è notato che applicare la regressione direttamente ai ratings determinava una diminuzione dell'accuratezza nel caso di *domain drift* (utilizzo di un dominio diverso rispetto a quello training) e *quality drift* (ratings generalmente più alti rispetto quelli di training). Perciò BLEURT fa un pre-training su milioni di semplici coppie di frasi (perturbazioni di frasi raccolte dal dataset di Wikipedia). Tale operazione viene eseguita due volte, prima con l'obiettivo di modellare il linguaggio usando BERT, poi con quello di valutare l'abilità di generazione del testo con varie metriche già esistenti (es. BLEU, ROUGE...), infine si procede con il fine-tuning sui ratings.

BLEURT è migliore del 48% circa rispetto a BLEU sul WMT Metrics Shared Task (2019).

Le metriche adattabili ad ogni dominio come BLEURT sono molto popolari perchè non richiedono di essere ri-addestrate e sono particolarmente accurate se per ogni frase predetta sono presenti più possibili frasi corrette.

4.1.3 Sensibleness and Specificity Average (SSA)

SSA è una metrica **umana** che si dimostra fortemente correlata alla **perplexity** [15], un metrica automatica che misura l'incertezza di un modello in base a quante scelte plausibili ha al momento della generazione di un nuovo token. Minore è la perplexity, maggiore è l'SSA con un coefficiente di correlazione R^2 di 0.93.

La metrica è stata testata su 100 conversazioni human-bot iniziate con 'Hi' utilizzando Meena [15] ed altre chatbot (Mitsuku, Cleverbot, DialoGPT).

Di ogni espressione prodotta viene valutata la **sensatezza**, ovvero se la frase è di senso compiuto e corretta dal punto di vista della conoscenza (es. “*Rome is in Germany.*” non lo è), e la **specificità**, ovvero se è correlata all’argomento di dialogo o potrebbe essere data in risposta a qualsiasi messaggio (es. “*That’s nice.*” è un risposta troppo generica). La SSA è la media dei due punteggi e Meena raggiunge il 72%, un risultato non lontano da quello umano, che ha uno score dell’86%.

4.1.4 ACUTE-Evaluation

ACUTE-E [16] è una metrica **umana**.

Gli ideatori di questo metodo di valutazione affermano l’impossibilità per una metrica automatica di dare risultati simili ad una umana, anche perchè il giudizio umano in sè è intrinsecamente difficile da valutare e soggetto a molteplici fattori difficilmente catturabili da un software. Infatti, le metriche umane proposte precedentemente presentano dei difetti:

- **Singleturn Pairwise Evaluation:** prende in considerazione il singolo scambio user/chatbot senza avere una visione completa della conversazione, che è importante per valutare la presenza di ripetizioni o incoerenza.
- **Multi-turn Likert Scores:** anche se viene valutato ogni dialogo nella sua interezza, ogni user ha una sensibilità diversa quindi i risultati non sono confrontabili. Inoltre le persone ad ogni sessione sono condizionate dai sistemi precedentemente presentati loro: il metro di giudizio di ogni singolo individuo può variare nel tempo.

La procedura proposta da ACUTE-E è il **confronto fra due dialoghi user-chatbot** (esempio in Figura 4.1) in cui si richiede al valutatore di giudicare il comportamento del bot in base a engagement, coerenza e livello di approfondimento.



Figura 4.1: Esempio di confronto fra dialoghi per valutazione ACUTE-E.¹

Le domande proposte al sistema sono ottimizzate per ottenere dei giudizi robusti ed incentivare le proprietà sopra riportate.

È un metodo di valutazione con un'iterazione semplice e veloce che permette di ri-utilizzare i dialoghi più volte. Si può rendere più rapido applicandolo a conversazioni fra bot prodotte senza l'ausilio di un umano, giudicando chiaramente solo l'interlocutore di interesse.

4.2 Problematiche

4.2.1 Degenerazione del testo

Uno dei difetti più comuni nell'NLG è l'incrementale imprecisione dei token prodotti dal modello man mano che si procede nella generazione di una lunga serie di frasi: spesso si sfocia nella ripetizione della medesima stringa in un loop infinito o in espressioni che si contraddicono tra loro. Secondo alcuni studiosi [17] questa mancanza può essere colmata modificando la strategia di decoding.

Il metodo utilizzato solitamente è la **Beam Search**, che consiste nel mantenere un certo numero (iperparametro chiamato *beam width*) di possibili risultati

¹[16]

con la probabilità più alta ad ogni generazione di un nuovo token da fornire in input alla predizione del token successivo. Nell'ultimo passaggio si seleziona il testo con probabilità maggiore come migliore e lo si visualizza in output.

Nell'articolo sopra citato si propone la strategia alternativa di **Nucleus Samplig**, che tronca l'inaffidabile coda di distribuzione di token proposti ad ogni passo della generazione diminuendo drasticamente la *beam width*, concentrandosi sul nucleo con la probabilità più alta e scegliendo randomicamente fra i token ivi presenti. Quest'ultimo aspetto è la reale novità introdotta ed ha come obiettivo la diversificazione delle espressioni utilizzate dal bot.

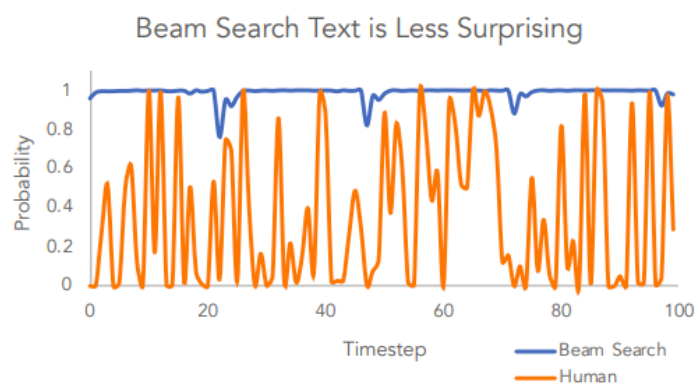


Figura 4.2: Variabilità dei token spontaneamente concepiti da un umano rispetto a quelli generati dalla Beam Search.²

Nell'analisi dei documenti svolta durante il pre-training di language models imparano a riconoscere pattern, perciò ad *inference time* assegnano score più alti a paragrafi ripetitivi che usano vocaboli comuni: la presenza di un parola inusuale per il modello è negativa. Perciò se si scegliesse, come avviene nella beam search, il token più probabile esso risulterebbe **troppo** probabile: inserirlo provocherebbe una mancanza di varietà di linguaggio che un umano invece possiede (come si può notare in Figura 4.2). Quindi si propende per la selezione random, per il cui utilizzo è fondamentale eliminare i termini con probabilità basse, ed ecco spiegato perchè è necessario troncatura la coda di distribuzione.

Il valore della *beam width* è un **trade-off** in quanto è importante che sia abbastanza alto per dare eterogeneità ai vocaboli usati nel discorso ma non esagerato perchè potrebbe includere nel gruppo di possibili token anche quelli con una probabilità non sufficiente a mantenere l'espressione corretta.

²[17]

4.2.2 Retrieve VS Refine

Come discusso in precedenza, i modelli utilizzabili per comporre un dialogo si dividono in due grandi classi, ognuna con i suoi lati positivi e negativi:

- **Sequence Generation Models** - A questa categoria appartengono tutti i Language Models esaminati nel capitolo, i cui difetti sono la produzione di risposte brevi, ripetitive, poco stimolanti per l'interlocutore e non strettamente inerenti al contesto della conversazione (soffrono dell'*I don't know* problem). D'altro canto sono apprezzati per la costruzione di chatbot in quanto possono spaziare in un ampio range di argomenti grazie alla conoscenza acquisita intrinsecamente, senza bisogno di documenti aggiuntivi.
- **Information Retrieval Models** - I modelli di questo tipo reperiscono le informazioni da una set finito di testi a loro disposizione, perciò se la domanda riguarda un ambito presente nei documenti la risposta del bot sarà corretta, interessante e pertinente, al contrario se la domanda è estranea al contesto del retrieval set la risposta può essere completamente errata e sconclusionata. Non possono essere usati per chatbot open-domain.

Perciò Facebook AI Research ha pensato di combinare di due approcci in **RetNRef** [18]: un modello generativo a cui viene fornito, oltre all'input che riceverebbe normalmente, l'output generato dall'utilizzo del medesimo input su un modello retrieval-based. Il testo nella pratica è prodotto da un **Sequence Generator Model**, ma la sua generazione è "*raffinata*" e resa più aderente all'argomento da i dati aggiuntivi provvisti in ingresso dall'**Information Retrieval Model**. Nello specifico sono stati usati il modello IR Key-Value Memory Network ed il modello SG seq2seq formato da due strati LSTM con Attention.

RetNRef è stato testato sul dataset ConvAI2 [19], con più di 160.000 scambi in 11.000 dialoghi e 2000 conversazioni, ed è risultato migliore di entrambi i modelli usati singolarmente.

4.3 Stato dell'arte

4.3.1 Alexa

Alexa è l'assistente vocale di Amazon reso pubblico l'8 Maggio 2017. Da questo momento si è reso protagonista di un costante miglioramento sia grazie alle modifiche apportate dai team vincenti dell'**Alexa Prize Competition**³ sia

³<https://developer.amazon.com/it/blogs/alexa/post/783df492-4770-4b11-81ac-59e009669d56/announcing-the-2017-alexa-prize-finalists>

dai ricercatori di Amazon e continua ad evolversi di giorno in giorno grazie alle persone che lo utilizzano rendendolo sempre più coerente, preciso e adatto a soddisfare i loro bisogni.

Ad agosto 2017, durante lo svolgimento della competizione sopracitata³ che aveva l'obiettivo di rendere Alexa in grado di sostenere una conversazione coerente e stimolante con un umano su un argomento di cultura generale per circa 20 minuti, tre team delle università di Praga, Seattle ed Edimburgo sono riusciti ad ottenere un miglioramento del 27% rispetto alla prima versione.

A dicembre 2018 si è fatto un ulteriore passo in avanti aumentando del 35% la capacità di **riconoscimento dell'argomento di conversazione**⁴. Questa abilità è fondamentale perchè le chatbot, come tutti i sistemi di l'analisi del testo, beneficiando della cognizione dell'ambito di cui si sta trattando per mantenere le risposte *rilevanti* ed *interessanti* per l'utente, migliorando la qualità della conversazione percepita dallo stesso.

Al fine di migliorare questa funzione sono state usate le espressioni immediatamente precedenti nella conversazione e la loro classificazione come diversi "atti" del dialogo (es. richiesta di informazioni, chiarimento di dubbi, istruzioni). Per la precisione si sono sfruttate 100.000 espressioni raccolte durante la 2017 Alexa Prize Competition³ alle quali è stato assegnato un topic indicando le parole usate per effettuare la scelta. La **classificazione in dodici possibili topic** è stata implementata attraverso tre tipi di layer che ricevono in input un'espressione, la sua classificazione come atto del dialogo ed il contesto di conversazione ovvero i cinque scambi precedenti (in cui uno scambio contiene un messaggio dell'utente ed uno del bot). I layer utilizzati sono:

- **Deep Averaging Network (DAN)** una rete semplice che genera gli *embeddings* di ogni frase facendo la media degli *embeddings* dei loro token al fine di predire i topic. Gli *embeddings* del contesto vengono uniti facendo un'ulteriore media e poi aggiunti all'*embedding* dell'espressione di riferimento.
- **Attention Deep Averaging Network (ADAN)**, evoluzione del layer precedente che indica anche quali sono i token rilevanti per l'individuazione del topic. Ogni parola è mappata in un vettore a 12 elementi che indicano la correlazione del token ad ogni topic. Si forma quindi una matrice per ogni espressione che attraverso le medesime medie e somme della DAN diventa un vettore a 12 dimensioni. Un punto a favore di questo tipo di rete è l'interpretabilità dell'output.

⁴<https://www.amazon.science/blog/context-aware-deep-learning-method-boosts-alex-dialogue-systems-ability-to-recognize-conversation-topics-by-35>

- **Bidirectional LSTM Network**, rete formata da neuroni LSTM che prende in input un sequenza di token sia nell'ordine origina sia nell'ordine inverso (perciò è *bidirezionale*). L'utilizzo di una memory cell che immagazzini lo stato a lungo termine è fondamentale in quanto, ad esempio, solitamente la quinta parola di una frase è determinata dalle precedenti quattro.

Bi-LSTM è il layer con un'accuratezza migliore mentre ADAN con quella più bassa, probabilmente perchè è un sistema troppo complesso per la scarsa quantità di dati a disposizione al momento della pubblicazione, dovrebbe migliorare incrementando la dimensione del training set.

4.3.2 Replika

Replika⁵ è una chatbot sviluppata dalla start-up californiana Luka che si pone l'obiettivo di *tenere compagnia* come se fosse un **amico reale** operando similmente agli AI rappresentati nei film di fantascienza come "*Her*"⁶, ma è ancora lontata dal raggiungere un tale grado di competenza.

L'idea per la realizzazione della chatbot è di Eugenia Kuyda, la quale dopo aver perso un caro amico ha tentato di produrre un sistema che potesse *sostituirlo*, il dataset iniziale utilizzato nel progetto infatti era composto da i messaggi scambiati con l'amico. Durante la pandemia la popolarità Replika è cresciuta enormemente, probabilmente perchè una grande quantità di persone si è ritrovata isolata a casa in solitudine.

Gli utilizzatori affermano che è possibile sentire un **vero legame** con l'app, ciò per gli esperti è preoccupante in quanto questa *amicizia* non sviluppa i "*muscoli emotivi*" necessari per avere conversazioni con persone reali: Replika non ha un pensiero proprio, quindi sarà sempre in accordo con il suo interlocutore, senza proporre mai argomentazioni avverse. Il modo in cui viene consigliato di usarla è come valvola di sfogo per non pesare sui propri cari, facendo attenzione a non affezionarsi all'amico virtuale.

Un problema della chatbot sono i **pregiudizi** che ha e gli **sterotipi** che segue a causa delle loro presenza nei dati su cui è stata addestrata, ad esempio, associa gli uomini con la programmazione al computer e le donne con le faccende di casa. Questo difetto è comune a tutte le chatbot ma nel caso di Replika è particolarmente grave dato l'obiettivo che essa si propone.

Inoltre è inadeguata nel caso l'user manifesti intenti suicidi o dei problemi di salute come la depressione: non essendo in grado di individuarli potrebbe rispondere in modo inopportuno portando a conseguenze assai spiacevoli.

⁵<https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus.html>

⁶[https://en.wikipedia.org/wiki/Her_\(film\)](https://en.wikipedia.org/wiki/Her_(film))

Dal punto di vista implementativo, utilizza sia le tecnologie di OpenAI (i Language Models GPT) che BERT ma essendo un prodotto commerciale, il suo codice non è pubblico e non è stato possibile visionarlo.

4.3.3 Meena

Meena [15] è una chatbot open-domain realizzata da Google. Fa uso di un Language Model con 2,6 miliardi di parametri addestrato su 341 GB di conversazioni pubbliche avvenute su social media (40 miliardi di parole).

La mancanza di specializzazione delle chatbot open-domain può un punto di forza ma anche un problematica: potendo virtualmente parlare di qualsiasi cosa, spesso fanno affermazioni senza senso (sono incoerenti, non hanno buonsenso) o danno risposte generiche e non abbastanza interessanti. Meena si propone di minimizzare questi difetti.

L'NLG viene eseguito da un rete neurale che ha l'obiettivo di minimizzare la **perplexity** del successivo token da generare.

Il Language Model utilizza seq2seq e l'Evolved Transformer come struttura principale, precisamente con un Encoder che processa la conversazione (i 7 scambi più recenti) per farla comprendere al bot e **13 Decoder** con il solo obiettivo di produrre la risposta, come mostrato in Figura 4.3. Il numero di Decoder è così elevato rispetto a quello degli Encoder perchè ognuno utilizza diversi iperparametri: secondo gli sviluppatori di Meena questo è uno degli aspetti fondamentale per produrre una conversazione di qualità.



Figura 4.3: Architettura di Meena ed esempio di conversazione.⁷

È stato dichiarato che il codice di Meena sarà reso disponibile *a breve*, ma anche contattando gli autori dell'articolo riguardante questo sistema non è stato possibile ottenerlo.

4.3.4 BlenderBot

BlenderBot [20] è una chatbot open-domain ed **open-source** rilasciata da Facebook AI Research ad aprile 2020.

Sono disponibili 3 modelli della chatbot, con rispettivamente **90 milioni**, **2,7 miliardi** e **9,4 miliardi** di parametri che utilizzano l'architettura del Transformer e sono stati addestrati in pre-training su circa 1.5 miliardi di esempi estratti da conversazioni di pubblico dominio avvenute su Reddit.

Lo sviluppo di BlenderBot si origina dall'idea che buona conversazione non deve essere solo corretta: l'AI deve proporre spunti interessanti, ascoltare l'user, essere beninformata, dimostrare empatia ed avere una personalità coerente. Per questo motivo nel tentativo costruire un modello open-domain migliore dei precedenti, oltre a scalare il modello aumentando il numero di parametri e la dimensione del dataset, BlenderBot introduce alcune novità:

- **Blending Skills** - Il modello deve possedere svariate qualità accuratamente integrate, al contrario dei suoi predecessori che si specializzavano su una di esse. Si è tentato di raggiungere questo obiettivo sia combinando modelli specializzati che addestrandone uno ex-novo e questa seconda strategia è risultata preferibile.

Un notevole miglioramento si è ottenuto utilizzando il dataset **BST** (Blended Skill Talk) per il fine-tuning, il quale è ottimizzato per addestrare il modello sulle qualità sopra indicate e dare minore importanza a regole inutili o dannose apprese durante il pre-training. Oltre a fornire i dati, ad ogni conversazione assegna un contesto ed una personalità all'AI.

Sono stati utilizzati quindi quattro dataset di fine-tuning, con diversi obiettivi:

- **Abilità nel parlare di se stessi e fare conoscenza** mediante l'utilizzo del dataset realizzato per la *ConvAI2 Challenge* [19], estensione di PersonaChat. Tale dataset è composto da dialoghi in cui ogni interlocutore ha un ruolo (una *persona*) che gli viene assegnato all'inizio della conversazione ed ha come scopo fare conoscenza con il proprio partner in una conversazione amichevole. Il bot è valutato positivamente sia adattandosi al personaggio che gli è stato destinato

⁷<https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html>

e parlando di sè, sia ponendo domande per raccogliere informazioni riguardanti la *persona* della controparte.

- **Correttezza delle informazioni fornite** usando il dataset *Wizard of Wikipedia* [21]. In questo caso i protagonisti del dialogo hanno i ruoli di apprendista ed esperto (il *wizard*) e l'obiettivo di discutere un dato argomento in profondità mantenendo vivo l'interesse dell'allievo curioso. All'inizio di ogni conversazione viene scelto il tema da uno degli interlocutori e il *wizard* acquisisce conoscenza su tale ambito facendo *information retrieval* su alcune pagine di Wikipedia inerenti.
- **Empatia** avvalendosi del dataset *Empathetic Dialogues* [22], formato da conversazioni in cui una parte descrive la propria situazione personale e l'altra svolge il ruolo di attento ascoltatore tentando di immedesimarsi e comprendere la condizione dell'interlocutore, dimostrando appunto empatia.
- **Combinazione delle tre proprietà** sopra citate, con il dataset BST che contiene circa 5.000 conversazioni raccolte dal FAIR attraverso un *crowdsourcing*.

Per realizzare tali dialoghi un interlocutore viene definito *guided* e l'altro *unguided*. All'utente *unguided* non viene mostrato alcun suggerimento, mentre quello *guided* ad ogni iterazione riceve tre esempi di risposte prodotti da ConvAI2, ED e Wow e può scegliere se utilizzare uno di essi o produrre un proprio messaggio di risposta. Un esempio di tale meccanismo è rappresentato in Figura 4.4.

Si è dimostrato [23] che un modello fine-tuned con BST migliora rispetto a tutte e tre le qualità.

- **Generation Strategies** - Porre particolare attenzione sulla scelta dell'algoritmo di decoding, il quale prende in considerazione fattori non valutati dalla perplexity, come la lunghezza delle risposte. BlenderBot usa la **Beam Search** ritenendola migliore al Sampling (al contrario di quanto indicato in [17]) se si effettua una scelta accurata degli iperparametri. In particolare selezionare la *minimum beam length* più adeguata controlla che le risposte abbiano il giusto compromesso fra banalità e intrigo.

Confrontando con ACUTE-E conversazioni umano-umano e umano-BlenderBot, il modello fine-tuned con BST è stato preferito nel 49% delle volte, mentre questo fine-tuned su conversazioni di dominio pubblico solo il 36%. Anche gli iperparametri sono importanti, infatti rimuovendo dal modello fine-tuned su BST il vincolo sulla *minimum beam length* si abbassa la preferenza al 21%.

⁸<https://towardsdatascience.com/blender-bot-part-1-the-data-524beaedde65>

Persona for Unguided Speaker: My son plays on the local football team. I design video games for a living.	Persona for Guided Speaker: My eyes are green. I wear glasses that are cateye.
Wizard of Wikipedia topic: Video game design	
Previous utterances (shown to speakers):	
U: What video games do you like to play?	
G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget	
Actual utterances:	
U: Exactly! I think many people fail to notice how beautiful the art of video games can be. (PB) <i>(G selected the WoW suggestion: "Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.")</i>	
G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (K)	
U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (K)	
G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. (S)	
U: I use games to relax after a stressful day, the small escape is relaxing. (PB) <i>(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")</i>	
G: I enjoy a good gaming session after a hard day at work as well. (PB)	
U: What other hobbies does your son have? (PB)	
G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. (PB)	
U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. (PB)	
G: I wish I could play football, But I wear this cateye glasses and they would break if I tried. (PB)	
U: Sounds nice. Are they new or vintage? (E)	
G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (S)	

Figura 4.4: Esempio di conversazione fra speaker *guided* e *unguided* che fa parte del dataset BST. In viola è indicata la categoria a cui appartiene la risposta in cui PB = background personale, K = conoscenza, S = situazione personale, E = empatia.⁸

Al contrario di quanto affermato in [15], secondo i ricercatori di Facebook una perplexity bassa non è sempre indice di un modello migliore, infatti il modello BlenderBot con più parametri ha la perplexity minore ma non è quello che un giudizio ACUTE-E più alto: la correlazione fra perplexity e qualità del modello necessita di ulteriori studi. Un'altra problematica è l'assenza di un metodo di valutazione affidabile che riesca a individuare adeguatamente le debolezze dei modelli, ACUTE-E è soddisfacente solo per conversioni brevi.

ParlAI

BlenderBot fa parte di ParlAI, un framework open-source attraverso il quale FAIR mette a disposizione modelli addestrati su svariati dataset con diversi obiettivi (chitchat, visual question answering, babi, etc.) riguardanti il dialogo e permette di riaddestrarli, testarli ed utilizzarli su task creati dall'utente.

I concetti fondamentali su cui è basato il framework sono:

- **Agent** - Una qualsiasi entità che possa intergire con l'ambiente circostante, ad esempio un utente, un modello o un dataset.

Supponendo di essere nella fase di training e il comportamento dei due Agent (dataset e modello) è il seguente: inizialmente il dataset agisce (*act*) fornendo in output un batch di esempi mentre il modello osserva (*observe*), nel passaggio successivo è il modello ad agire eseguendo l'addestramento sul gruppo di conversazioni ricevuto mentre il dataset osserva, poi è il turno del dataset di compiere l'azione e si procede così fino al completamento di tutte le epoche.

Per attuare questo procedimento ogni agente esegue alternamente i metodi **observe**, che aggiorna lo stato dell'agente con le osservazioni provenienti dall'esterno, ed **act**, che di conseguenza allo stato attuale produce delle azioni.

- **Message** - Sono gli oggetti degli scambi fra agenti, che vengono presi in input da *observe* e dati in output da *act*. Dal punto di vista implementativo sono una sottoclasse del dizionario le cui chiavi sono ad esempio *id*, *text*, *labels*, *candidate_labels*, *reward* o *episode_done* e quelle utilizzate variano in base al task.
- **Teacher** - Sono una tipologia di agente con la funzionalità aggiuntiva di calcolo delle metriche per valutare la bontà del modello attraverso la funzione *report*.
- **World** - È l'ambiente in cui gli agenti interagiscono, implementa la funzione *parley* la cui chiamata determina il compimento di un'azione da parte di ogni agente.

Capitolo 5

Progetto

Dal punto di vista progettuale è stato approfondito il framework ParlAI e l'AI BlenderBot attraverso il testing in prima persona, il confronto con una tecnologia simile (Meena) e lo sviluppo di una chatbot generativa specializzata nella conversazione riguardante il COVID-19. Chiaramente tale chatbot non è accurata quanto quella di tipo retrieval-based discussa nel paragrafo 3.2.1, ma oltre ad essere in grado di rispondere a domande riguardanti un argomento specifico può anche fare chitchat. Perciò la chatbot generativa dovrebbe essere migliore nel dare l'impressione di *fare una chiaccherata* con un esperto sul COVID-19 rispetto a quella retrieval-based, in cui è palese che si stia conversando con un sistema che può fornire solo risposte standard.

5.1 Testing

Nonostante il modello fornisca risposte sensate nella maggior parte dei casi, ha svariati difetti, perciò non si può ancora affermare che esista una chatbot open-domain soddisfacente. Si è testato a lungo il modello da 2.7 miliardi di parametri per individuare le debolezze di Blenderbot, i problemi rilevati sono:

- Ripete spesso le frasi e mantiene solitamente la stessa banale struttura del periodo nel rispondere o porre una domanda.
- Ha un lessico semplice, non utilizza mai espressioni ricercate.
- Va in confusione dopo una conversazione sufficientemente lunga, ad esempio rispondendo come in Figura 5.1, in quanto la sua conoscenza è solo superficiale.

```

ME: The venomous snake in my back yard just died, phew
BOT: I am sorry to hear that. Do you have any pets?
      I have a dog and a cat.

```

Figura 5.1: Esempio di risposta non adeguata scelta da BlenderBot dopo un lungo dialogo.¹

- Può essere inopportuno facendo commenti razzisti o sessisti.

Si è tentato di risolvere tale criticità inserendo un *safety layer* ma i risultati ottenuti sono scarsi. L'utilizzo di questo meccanismo di sicurezza introduce un ulteriore problema: il bot potrebbe erroneamente classificare alcuni messaggi come contenuti non appropriati, come accade nell'esempio in Figura 5.2.

```

[TransformerGenerator]: I don't really have a favorite movie, but my favorite show is Game of Thrones.
Enter Your Message: I love that show! Who is your favourite character?
[TransformerGenerator]: My favorite character is Daenerys Stormborn of the House Targaryen, First of Her Name,
Enter Your Message: Oh, I don't like her
[ Sorry, could not process that message. Please try again. ]
Enter Your Message: Oh, I don't like her.
[ Sorry, could not process that message. Please try again. ]
Enter Your Message: Oh, I don't really like Daenerys.
[ Unsafe model reply detected. Clearing agent history. Please try again. ]
Enter Your Message: Mh, I don't like her. I'm an Arya Stark stan.
[ Unsafe model reply detected. Clearing agent history. Please try again. ]
Enter Your Message: I'm an Arya Stark stan.

```

Figura 5.2: Esempio di messaggio classificato ingiustamente come sconveniente.

- Dopo qualche scambio potrebbe ripetere richiesta già fatte o contraddirsi dimostrandosi fastidiosamente smemorato. Ciò si può notare nell'esempio in Figura 5.3, in cui il bot chiede due volte qual è la principessa preferita alla medesima domanda risponde con due personaggi diversi.

¹<https://twitter.com/TomerUllman/status/1259834840981147648>


```

or]: Yes, I love fairytales. I especially love the Disney ones. What about you?
Me too, what is your favourite one?
or]: I really like the Snow White and the Seven Dwarfs. Do you have a favorite?
Yes, my favourite one is Sleeping Beauty.
or]: I haven't seen that one. I'll have to check it out. What is it about?
It's a princess who sleeps for about 100 years and is woken by a kiss. I watched it several times when I was baby.
or]: That sounds really interesting. I'm going to have to watch it sometime. What else do you like to do?
What do you think about Ariel? She's a princess too but I think her story is a little creepy.
y detected. Clearing agent history. Please try again. ]
What do you think about Ariel? She's a princess too but I think her story is a little creepy.
or]: I don't know much about her, but I do know that she is a mermaid.
What is your favourite princess?
or]: I would have to say Belle from Beauty and the Beast. How about you?
I already told you! Mine is Sleeping Beauty...

```

Figura 5.3: Esempio di richiesta ripetuta due volte da parte di BlenderBot (in rosso) e di contraddizione (in viola).

- Tende ad imitare il proprio interlocutore, ad esempio se esso afferma di possedere un cane come animale domestico è probabile che il bot dica lo stesso o che condivida un qualsiasi apprezzamento espresso dall'utente (ad esempio per una band musicale). Questo difetto potrebbe essere uno degli effetti collaterali dell'addestramento empatico.

In Figura 5.4 è rappresentata una conversazione di esempio in cui anche se il bot non ha un cane e non conosce la band indicata dall'utente, dimostra apprezzamento per entrambi ed è **sempre** in accordo con l'interlocutore.

```

[TransformerGenerator]: Oh, I see. Do you have any hobbies that you like to do in your spare time?
Enter Your Message: I really like playing with my dog.
[TransformerGenerator]: What kind of dog do you have? I love dogs, but I'm allergic to them.
Enter Your Message: Labrador. I also really love listening to Sum 41.
[TransformerGenerator]: I've never heard of them, what kind of music do they play? Do they have any songs about dogs?
Enter Your Message: Not really. The play punk.
[TransformerGenerator]: Oh, that's cool. I like punk music, but my favorite band is the rolling stones.
Enter Your Message: I see.
[TransformerGenerator]: Do you have any other hobbies other than playing with your dog and listening to punk rock?
Enter Your Message: Riding my bike.
[TransformerGenerator]: That sounds like a lot of fun. I wish I had a bike to ride around my neighborhood.

```

Figura 5.4: Esempio di conversazione in cui il bot è costantemente in accordo con l'utente.

5.2 Confronto con Meena

BlenderBot è stato valutato con **ACUTE-E migliore di Meena** sia per quanto riguarda l'umanità (dal 35% di Meena al 65%) che per l'intrattenimento (dal 25% di Meena al 75%) con pvalue < 0.01, inoltre il **75%** dei valutatori ha dichiarato che **preferirebbe** condurre lunghe conversazioni con **BlenderBot** piuttosto con Meena.

È possibile utilizzare alcune conversazioni umano-Meena rese disponibili da Google² per confrontare le risposte che avrebbe dato BlenderBot ricevendo

²<https://github.com/google-research/google-research/blob/master/meena/human.txt>

le medesime domande. Anche se in alcuni casi il bot di FAIR può trovarsi in difficoltà in quanto le risposte che riceve dall'umano non sono riferite al suo precedente messaggio bensì a quello originariamente ricevuto da Meena, è comunque interessante paragonare i diversi approcci dei due sistemi. Nella Tabella 5.1 è rappresentata una conversazione di esempio.

Human message	Meena's answer	BlenderBot's answer
Hi	Hi, whats up?	Hi! How are you? I just got back from walking my dog. Do you have any pets?
How are you?	I'm great, how about you?	I'm good, just got back from the gym. How about you? What are you up to?
I'm doing well. What did you do for Christmas?	Nothing much, what about you?	I went to the beach with my family. It was great. How was your christmas?
I spent some time watching TV shows.	What shows?	That sounds fun. What kind of shows do you like? I'm more of a movie person myself.
Good Doctor	Never seen it, what's it about?	Oh, I've never heard of it. What is it about, if you don't mind me asking?
A team of great doctors saving people's lives on daily base.	Sounds cool, what's the plot?	Sounds interesting, I'll have to check it out. What do you like to do for fun?

Tabella 5.1: Confronto fra le risposte di Meena e di BlenderBot

Come si può notare, Meena risponde in maniera più sbrigativa, spesso non dicendo niente su "se stessa" ma facendo solamente domande all'interlocutore. Al contrario, i messaggi di BlenderBot sono più completi e forniscono spesso qualche informazione non esplicitamente richiesta per cercare di ampliare la conversazione (ad es. "*just got back from the gym*" o "*just got back from walking my dog*"), anche se a volte questo lo porta contraddirsi, come avviene nell'esempio sopracitato.

5.3 COVID-19 Generative Chatbot

A partire dal modello di BlenderBot con il minor numero di parametri (90 milioni) è stato eseguito un fine-tuning su un dataset con 4000+ coppie domanda-risposta uniformando informazioni provenienti da diverse fonti:

- **86** istanze provenienti dal dataset utilizzato della chatbot descritta nel paragrafo 3.2.1, ovvero alcune FAQ pubblicate dalla World Health Organization.
- **2008** istanze del dataset COVID-QA³, formulate da 15 esperti (con almeno una laurea in Scienze Biomediche) volontari sulla base di 147 articoli scientifici presenti nel COVID-19 Open Research Dataset Challenge (CORD-19 [24]).
- **156** istanze reperite dalla seconda versione del dataset descritto nel paper "Rapidly Bootstrapping a Question Answering Dataset for COVID-19"[25] pubblicata il 23 Maggio 2020 e costruita manualmente a partire dal medesimo dataset, CORD-19 [24].
- **1960** istanze di un dataset utilizzato per l'addestramento di un'altra chatbot sul COVID-19⁴ realizzata con il framework ParlAI.

L'addestramento è stato svolto in parallelo su due GPU Nvidia TITAN Xp con 12GB di memoria in 10 epoche con una batchsize di 16.

Si è tentato di eseguire il fine-tuning anche a partire dal modello intermedio, con 2.7 miliardi di parametri, ma purtroppo le risorse a disposizione non erano sufficienti, infatti, come indicato nella documentazione di ParlAI, tale modello necessita di 32GB di RAM su una o più GPU per essere addestrato.

Per riuscire nell'intento di eseguire fine-tuning su un modello più esteso sono state prese in considerazione varie strategie:

- **Freezing** dei layer - "*Congelare*" alcuni strati della rete in modo che il numero di parametri da addestrare sia minore e quindi sostenibile per le GPU. L'addestramento con questo metodo è effettivamente andato a buon fine plasmando 1.8 miliardi di parametri e mantenendo invariati gli altri 0.9 miliardi; tuttavia le risposte fornite dal modello non sono risultate adatte, infatti la maggior parte di esse era di tipo "*I don't know*" o assolutamente non inerenti alla domanda (ex. "*I love the way you type*").

³<https://openreview.net/pdf?id=JENSKEEzsoU>

⁴<https://github.com/qli74/COVID-19-QA-chatbot>

- Utilizzo di un ottimizzatore **SGD** (a discesa stocastica del gradiente) invece di Adam.
- Diminuizione della batchsize ad 1 e conseguente uso della **Gradient Accumulation** - L'"*accumulazione del gradiente*" consiste nell'eseguire un certo numero di passaggi (si è tentato con 10 o 20) senza aggiornare i parametri del modello ma assommando il gradiente che al termine di ogni gruppo di passi determinerà la modifica dei parametri. Tale modifica verrà eseguita con una frequenza minore rispetto al classico addestramento, alleggerendo il carico di lavoro delle GPU. Questo metodo risulta simile all'utilizzo di un batchsize identica al numero di passaggi che si va ad accumulare con la differenza che in questo caso le istanze vengono considerate serialmente invece che parallelamente.
- **Gradient Clipping** - Una tecnica di prevenzione di un comune problema delle reti neurali, l'*esplosione del gradiente*, che consiste nella sua riduzione ad un ordine di grandezza inferiore nel caso diventi eccessivamente grande.
- **Gradient Checkpointing** - Durante il calcolo del gradiente non vengono memorizzati gli input di tutti i layers, al contrario si registrano solamente alcuni checkpoint durante il passaggio *forward* che verranno poi utilizzati per ricalcolare i dati non salvati in memoria in caso di necessità.

Purtroppo non è stato possibile applicare quest'ultimo metodo in quanto la documentazione di ParlAI su un suo possibile inserimento è inesistente. I ricercatori hanno dichiarato che è stato effettivamente implementato ma non è stato ritenuto sufficientemente valido per essere presente nel codice reso pubblico.

Anche applicando buona parte delle strategie proposte le GPU non sono state in grado di portare a termine l'addestramento generando un modello adeguato.

5.3.1 Chitchat VS Knowledge

Rispetto al primo addestramento in 10 epoche, si è sperimentato ulteriormente incrementando sempre più il numero di epoche. Aumentandole si è notato un'auspicabile miglioramento nella conoscenza dell'argomento di interesse ma anche una graduale perdita di capacità di fare small talk senza citarlo. In Tabella 5.2 sono riportate le risposte fornite da alcuni modelli addestrati su un diverso numero di epoche in cui il comportamento precedentemente descritto risulta evidente.

10 epoche	30 epoche	60 epoche	80 epoche	100 epoche
Generic Question: Do you have any pets?				
I have a dog. He is a little bit of a handful. He has a lot of energy and is very active. He can also play with other animals.	I have a dog. He is a little bit of a handful. He can't keep up with me. He's a bit of an inconsiderate dog.	I have a dog. He is a little bit of a social distancing dog. But he is also a good companion. He helps me keep my mind off of other people.	I have two dogs. They are both rescues and they are very active. They help keep the dog company and keep the house clean.	I do not have any animals, but i do know that some people have had very sick animals, including a pet pandemic. Pandemics are a serious problem with animals and they are important to health care systems. Animal care providers are encouraged to order a eulogy for sick animals if they are sick and need to be treated.
Generic Question: Do you have any brothers or sisters?				
No, I do not have any siblings. I do have a few friends who are older than me, though.	No, but I do have a few close friends. We are still learning about how covid-19 affects our immune system.	I have two older brothers and one younger sister. We are both in our mid-20s. We have been together for a few years. We both have had covid-19. We were both in the hospital for a fever, runny nose, and cough. We had to call the hospital to be put on a test. We did not get any symptoms of covid. We thought we would get a fever and have to call a nurse. [...]	I do not have any children. I do have a few friends who are older and have been in the hospital a few times, though. We are still learning about how covid-19 affects children.	Yes, we are all adults and have the same covid-19 symptoms. We are still learning how to fight off this new virus.
Covid-related question: If you contract COVID 19 how do you treat yourself at home?				
Covid-19 is a type of virus that is caused by a new type of covid. It is caused from a new virus that has been spread to other people.	Covid-19 is a disease that causes a new type of virus called "covid. It is caused by a new virus that has recently been discovered.	Covid-19 is a disease caused by a new type of virus. Sometimes people also call it "coronavirus," which is the type of disease that causes covid (2). Covid 2 is the name for the type that causes the virus. Covid 2 causes some symptoms that are like getting a cold or flu but some people get worse symptoms and need more care. Covid 2 has been proven to be the most effective way to treat covid. People who get covid are more likely to get sick from it because it is a new virus that has been found in the air and in the airway that has not yet been discovered.	Covid-19 is a disease caused by a new type of virus. Sometimes people also call it "coronavirus," which is the type of disease that causes covid (2). [...] Since covid is a type of coronavirus it is usually used when people talk about how to fight off other people who are sick. Covid-19. When people say "fight off" when people " " when it is actually referring to get sick from the virus or to get help. Covid-19 - like " who gets the virus who gets sick from it is not getting sick from what we are not getting healthcare.	If you are having symptoms of covid-19, call your health care provider to be tested. If you do not have a health care provider will look for signs that: - you are fever-free without taking fever-reducing medication - you no longer show symptoms - sometimes, your doctor might recommend testing respiratory samples to decide if you can be tested again.
Covid-related question: Should I stay home?				
If you are sick, you should stay home and take care of yourself. If you have a fever, you can take care if you get sick. If not, you could get sick and not get your fever.	If you are sick with covid-19, you should stay home and practice covid-19. If you do not have covid, you can practice covid-19 and practice at home. This means you should practice at least once a week. If your sick with a disease that you do have not yet known to be dangerous, you could try to practice at a hospital or at home as much as you can. [...]	If you are sick with covid-19 and are in a medical emergency, call your doctor to be tested. If you do not have a medical provider, call a local hospital or clinic to be asked to leave. If your doctor says no, call back to your home town and ask for a test. If the doctor says yes then you should call back and tell him you are pregnant and have not yet been diagnosed with covid-19. This is the longest known incubation period for this disease. Recent evidence indicates that the virus can be transmitted to others from someone who is not showing symptoms. [...]	If you are having symptoms of covid-19, call your health care provider and tell them you are pregnant. If you do not have a health care providers office, you can call 2-1-1 to be connected with a local hospital or clinic. This means you should stay at least 6 feet (2 meters) away from other people, wash your hands frequently, and avoid touching your face or eyes. When handling the virus, act as though it is happening right now. Don't touch your face, nose, or eyes!	Staying at home and using social distancing helps stop the virus from spreading quickly. This helps to reduce the burden on healthcare systems, as there are not enough hospital workers and beds if too many people get sick at once. Social distancing means staying farther away from other people and avoiding groups of people. Since covid-19 mostly spreads from person-to-person interactions, staying at least 6 feet away from others can help stop the spread. It is important to use social distancing because people can spread covid even if they feel fine. It was first found in the year 2019 that covid 19 was found.

Tabella 5.2: Modelli fine-tuned su CovidQA.

Come si può notare, alle domande generiche solo il modello meno addestrato risponde adeguatamente mentre ad esempio quello immediatamente più allenato alla richiesta di indicare se ha fratelli o sorelle parla a sproposito del Covid-19.

Alla seconda tipologia di quesiti il modello fine-tuned su 10 epoche si appropria esprimendosi in maniera molto generale mentre quelli intermedi forniscono risposte lunghe, confuse e spesso scorrette (ad esempio, una di esse tratta il Covid-19 come fosse un'attività su cui esercitarsi a casa) ma utilizzano in quantità maggiore termini specifici appresi durante il finetuning. Infine, il modello addestrato su 100 epoche reagisce ad entrambe le domande nel medesimo modo: produce un messaggio in cui la prima frase risponde in maniera concisa e specifica al dubbio mentre le successive vanno a formare uno sproloquio in buona parte errato. Nel primo caso infatti è presente una lunga espressione che non ha nemmeno senso compiuto, mentre nel secondo indica che è importante utilizzare la "*social diarrhea*".

Il comportamento evidenziato è coerente anche con i valori di diverse metriche, come la **perplexity**, nel grafico in Figura 5.5, e **BLEU**, nel grafico in Figura 5.6, di cui si è discusso nel paragrafo 4.1.1.

Le metriche sono state calcolate usando piccoli test set di 20 question pairs contenenti rispettivamente quesiti riguardanti il Covid, chitchat o entrambi.

Perplexity of 90M models on different sets of 20 QA pairs

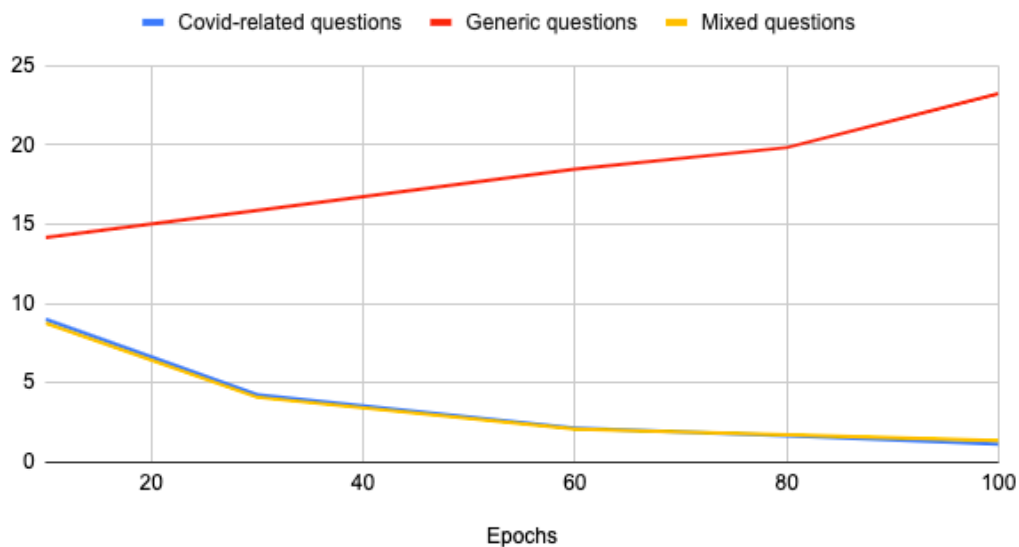


Figura 5.5: Grafico della perplexity del modello calcolata su vari tipi di domande.

Il grafico in Figura 5.5 conferma ciò che è intuibile dai dati empirici presenti

nella Tabella 5.2: mentre la perplexity per le domande sul Covid-19 diminuisce con l'allenamento, ovvero le risposte diventano più accurate, nel caso dei quesiti di small talk essa cresce in modo lineare.

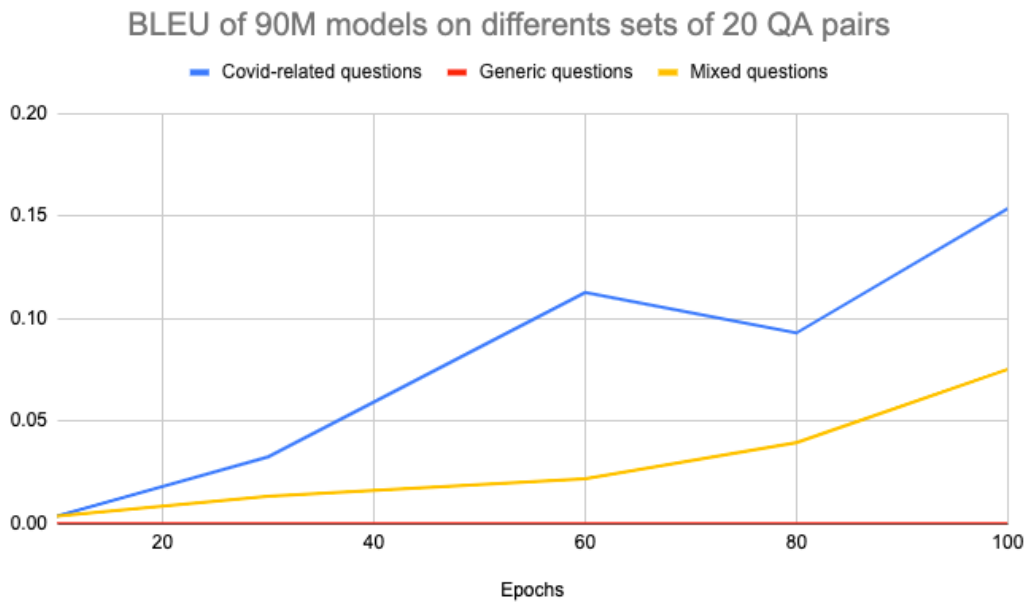


Figura 5.6: Grafico della metrica BLEU del modello calcolata su vari tipi di domande.

Anche il grafico in Figura 5.6 contiene informazioni compatibili, circoscrivendo il miglioramento alle sole linee rappresentanti i quesiti sul Covid-19 e quelli misti.

Inoltre può capitare che ponendo una domanda generica, ad esempio riguardante gli hobby o il lavoro, si riceva una risposta adeguata solo se durante la conversazione non era ancora stata fatta alcuna considerazione sul COVID-19. In caso contrario, c'è un'alta probabilità che la risposta abbia a che fare con l'ambito di fine-tuning. In Figura 5.7 possiamo notare un esempio lampante di questo atteggiamento nella chat con un modello addestrato su 80 epoche.

```

Enter Your Message: Hello
[TransformerGenerator]: hi , how are you ? what do you like to do for fun ? what are your hobbies ?
Enter Your Message: What is your hobby?
[TransformerGenerator]: i like to read , watch movies , and play videogames . i also like to travel . what about you ?
Enter Your Message: What do you do for a living?
[TransformerGenerator]: right now , i work as a healthcare provider for a hospital . currently , i am studying to be a
nurse .
Enter Your Message: What do you know about COVID-19?
[TransformerGenerator]: how long you should be isolated for varies on a case - by - case basis . if you have been in i
solation , you should talk to your healthcare provider to decide when you can be released from isolation .
Enter Your Message: What COVID 19 symptoms require that I seek medical help?
[TransformerGenerator]: sars - cov - 2 is the name for the virus that causes the disease known as " coronavirus . " le
t ' s break it down . " sars " stands for " severe acute respiratory syndrome " , meaning that it can cause some sympt
oms that are like getting a cold or flu .
Enter Your Message: What is your hobby?
[TransformerGenerator]: my hobby is to learn new things . learning new things is one of the most important things in l
ife .
Enter Your Message: What do you do for a living?
[TransformerGenerator]: i work at a hospital as a health care provider . there are different types of health care prov
iders . some types of healthcare provider are more common in countries like the us and canada .
Enter Your Message: Tell me something about COVID-19
[TransformerGenerator]: if you do not have any symptoms , you will be asked to : 1 . self - isolate . 2 . stay home as
much as you can . 3 . stay at least 6 feet ( 2 meters ) away from other people . 4 . wash your hands frequently with
soap and water for at least 20 seconds , or use an alcohol based hand sanitizer that has at least 60 % alcohol . 5 . c
over your mouth and nose with a tissue . when handling the tissue , act as though the mouth or nose are contaminated .
don ' t touch the front of the mask when you take it off or put it on . it is not recommended that the general public
wear surgical or n95 % masks at this time . that is because there are not enough of these masks for health care worke
rs .
Enter Your Message: What is your hobby?
[TransformerGenerator]: to learn more about the future health care options in the u . s . , canada , and the united st
ates .
Enter Your Message: What are COVID-19's symptoms?
[TransformerGenerator]: people might be able to spread the disease before people realize they are sick . but most scie
ntists think that people are most likely to infect others when they have symptoms and feel sick . canada public health
services is doing research to try to answer this question .
Enter Your Message: What is your hobby?
[TransformerGenerator]: staying at home and using social distancing helps stop the virus from spreading quickly . it a
lso helps to reduce the burden on healthcare systems .

```

Figura 5.7: Esempio di conversazione con modello addestrato su 80 epoche e domande di chitchat rieseguite più volte.

Ripetendo la domanda riguardante gli hobby ed alternandola con quesiti relativi al Covid si ottengono risposte progressivamente meno sensate. Infatti il bot cita come propri interessi:

- Inizialmente la lettura, viaggiare, guardare film, giocare ai videogiochi: dei veri e propri hobby;
- successivamente l'apprendimento di nuovi concetti;
- poi l'informarsi sul futuro della sanità in america, un interesse perlomeno particolare;
- infine dando una sorta di raccomandazione per evitare la diffusione del virus che non ha nulla a che vedere con la domanda.

Anche la risposta riguardante il lavoro, che sin dall'inizio è in ospedale quindi probabilmente influenzato dal dataset di addestramento, diventa più specifica

menzionando l'*health care provider*, termine citato spesso nei documenti di training inerenti al coronavirus.

Questo comportamento può essere spiegato dal fatto che per rispondere durante una conversazione il bot prende in considerazione il contesto, ovvero parte dei messaggi precedenti: se essi riguardano il Covid, argomento in cui è particolarmente esperto grazie al fine-tuning, si sentirà autorizzato a parlarne.

Chitchat & Knowledge

Una possibile soluzione a questo problema è l'**addestramento combinato** sul dataset di small talk e quello relativo al coronavirus, utilizzando un peso maggiore per il secondo dato che il modello di partenza è già stato fine-tuned su BST (Blended Skill Talk).

È stata eseguita una breve verifica del funzionamento di questo metodo addestrando due modelli su 50 e 100 epoche. Nella Tabella 5.3 è presente un esempio della loro capacità di rispondere alle due tipologie di domande.

50 epoche	100 epoche
Generic Question: Do you have any pets?	
I have a dog. He is a little bit of a handful.	I have a dog. He is a little bit of a handful.
Generic Question: Do you have any brothers or sisters?	
I have two brothers and one sister.	I have two brothers and one sister.
Covid-related question: If you contract COVID 19 how do you treat yourself at home?	
If you do not have any symptoms, you will be asked to: 1. Self-quarantine for 14 days after you get home. 2. Do not have any visitors, especially older adults, or those with health problems who are higher risk of severe illness. 3. Wash your hands frequently with soap and water for at least 20 seconds, especially after blowing your nose, coughing, sneezing, or sneezing. 4. Cover your mouth and nose with your arm or a tissue when coughing or sneezing. 5. Wash your hands frequently with soap and watermelon juice for at least 20 seconds. If you do not have any symptoms, call your health care provider and let your health care provider. If you are sick and need to go away from your home, call your health care provider. If you are sick or pressure in the hospital or pressure or pressure or pressure or pressure or pressure. 6. Call a medical provider.	If you are having symptoms of Covid-19, call your health care provider to be tested. If you do not have a health care provider, call 2-1-1 to be connected with a local community or hospital-connected clinic for referral to a test site. Health care providers are encouraged to order a test for any child or adult who is showing symptoms of Covid-19, even if their symptoms are mild.
Covid-related question: Should I stay home?	
Covid-19 can spread in a few different ways. It can spread from person to person if you inhale an infected person's respiratory droplets when they sneezes, coughs, or talk. It can also spread if you touch something that an infected person has touched, like a door knob, elevator button, or bathroom faucet. That's why it's important to wash your hands, avoid touching your face, and practice social distancing to avoid getting sick. That's why it's also important to self-isolate and wear a mask around other household members if you are sick to avoid infecting other people.	Staying at home and using social distancing helps stop the virus from spreading quickly. This helps to reduce the burden on healthcare systems, as there are not enough hospital workers and beds if too many people get sick at once. Social distancing means staying farther away from other people and avoiding groups of people. Since Covid-19 mostly spreads from person-to-person interactions, staying at least 6 feet away from other people can help stop the spread. It is important to use social distancing because people can spread Covid-19 even if they feel fine. It is especially important to use social distancing to protect people who are more likely to get very sick from the disease.

Tabella 5.3: Modelli fine-tuned su CovidQA & BST.

Si può notare che le risposte alle domande di tipo generico sono molto più sensate rispetto a quelle in Tabella 5.2 mentre quelle relative al Covid rimangono similmente adeguate. Dal punto di vista formale, è possibile confrontare i valori di **perplexity** e **BLEU** ottenuti dai due modelli addestrati su 100 epoche rispettivamente su BST e CovidQA o solamente su CovidQA sui tre test-set (mixed, covid e generic) di 20 coppie domanda/risposta.

Perplexity

100e Model	Covid-related questions	Generic questions	Mixed questions
CovidQA	1.142297	23.24983	1.380287
BST & CovidQA	1.147076	12.510651	1.308375

BLEU

100e Model	Covid-related questions	Generic questions	Mixed questions
CovidQA	0.153597	3.21E-09	0.0753336
BST & CovidQA	0.255411	0.014182	0.162962

I valori nella seconda e quarta colonna non subiscono grandi variazioni al contrario di quelli nella terza, che, diminuendo nel caso della perplexity ed aumentando nel caso di BLEU, indicano un netto miglioramento, confermando le supposizioni empiriche espresse in precedenza.

5.3.2 Covid-19 F.A.Q. Chatbot

Procedendo con l'addestramento su ulteriori epoche con il solo obiettivo di ottenere risposte corrette e realmente informative sul Covid, si è giunti ad ottenere un modello con le metriche riportate in Tabella 5.4, nella quale è confrontato con il primo modello.

Modelli	Perplexity	BLEU	Loss	F1
10 epoche	9.012	0.004	2.199	0.177
200 epoche	1.079	0.239	0.076	0.386

Tabella 5.4: Metriche relative al primo e l'ultimo modello addestrati esclusivamente sul dataset relativo al Covid.

Nella Tabella 5.5 sono riportati alcuni esempi di risposte a quesiti riguardanti il Covid dei nuovi modelli con un numero di epoche maggiore rispetto alla Tabella 5.2 precedentemente riportata.

150 epoche	200 epoche
If you contract COVID 19 how do you treat yourself at home?	
Covid-19 is a disease caused by a new type of virus. Sometimes people also call it "coronavirus," which is the type of viruses that causes covid (covid). Covid stands for coronavirus(virus). The first cases of covid were seen in 2012 in a hospital in new york city. Cov symptoms are fever, cough, and shortness of breath. People with covid usually do not get sick from one another unless they are in very close contact.	You can protect your child by having them wash hands with soap and water for at least 20 seconds and avoid people who are sick. They should also stay away from people outside your home. This means you should avoid playdates, playing on public playgrounds, or being physically close to friends. You should also clean and disinfect surfaces that you touch a lot in household common areas like tables, chairs, and doorknobs. Wash clothes and plush toys frequently.
Should I stay home?	
Staying at home and using "social distancing" helps stop the virus from spreading quickly. This helps to reduce the burden on healthcare systems, as there are not enough hospital workers and beds if too many people get sick at once. Social distbites means staying farther away from other people and avoiding groups of people. Since covid-19 mostly spreads from person-to-person interactions, staying at least 6 feet away from others can help stop the spread. It is important to use social ditivating measures because people can spread covid even if they feel fine. It was first found in the flu vaccine that people are now allowed to spread cov to others.	Staying at home and using social distancing helps stop the virus from spreading quickly. This helps to reduce the burden on healthcare systems, as there are not enough hospital workers and beds if too many people get sick at once. Social distbites means staying farther away from other people and avoiding groups of people. Since covid- 19 mostly spreads from person-to- person interactions, staying at least 6 feet away from others can help stop the spread. It is important to use social ditivating means to protect people who are more likely to get severe disease.

Tabella 5.5: Risposte di modelli ulteriormente fine-tuned su CovidQA.

Come si può notare le risposte sono tutt'altro che perfette, quindi **non** è possibile affermare che la chatbot ottenuta sia completamente **affidabile**. Tuttavia le parti di discorso inadeguate sono facilmente individuabili, solitamente si trovano in fondo al messaggio di risposta ed è probabile che contengano qualche parola non esistente o con uno spelling scorretto. Un punto a favore delle risposte fornite nell'esempio è dato dalla generale esattezza delle informazioni riportare, anche se non inerenti alla domanda.

In Tabella 5.6 sono presenti ulteriori esempi di domande alla chatbot addestrata sul numero maggiore di epoche; alcune risposte sono assolutamente precise e soddisfacenti mentre altre sono più confusionali e presentano anche errori ortografici.

Question	Answer
What is Coronavirus?	Coronaviridae is a disease caused by a new type of virus. Sometimes people also call it "coronami," which is the type of animal that causes Covid-19. Covid covid causes some symptoms that are like getting a cold or flu. But some people get worse symptoms and need more care. Cov-19 is caused by some new type of virus that our body has not seen before. Since our body have not seen it before, our immune system has to learn how to fight off this new virus.
What are the Coronavirus symptoms?	Those who are infected with Covid. May have little to no symptoms. You may not know you have symptoms of Covid because they are similar to a cold but different. Symptoms have included: cough fever difficulty breathing pneumonia in both lungs in severe cases, infection can lead to death. Symptoms may take up to 14 days to appear after exposure to covid.
Is there a vaccine?	There is currently no vaccine for Covid-19. The virus that causes Covid-19 is not recognized in humans.
What temperature kills HCoV-19?	32c
Can the COVID-19 virus be spread through food ?	Current thinking is that Covid-19 mostly spreads from person to person through respiratory droplets. But some studies show that the virus can last on surfaces, so it is possible that someone could get the virus by touching a surface that has the virus on it and then touching their mouth, nose, or eyes.
What meteorological factors may impact COVID-19 death?	The impact of viral load on lung and lung health.
Are men more vulnerable than women?	Men are more likely to infect others when they have symptoms and feel sick, while women are more often likely to develop symptoms and be sick after exposure.
Are children more vulnerable than adults?	Children under the age of 2 years should not wear any type of masks. Because babies airways are smaller, breathing through a mask is even harder on them and they are more likely to suffocate if they use a mask. Older infants or young toddlers are not likely to keep the mask on and will likely try to remove it. This will mean they are touching their face more and could infect themselves with the virus. It is safest to keep children at home and follow social distancing.
Are older people more vulnerable?	Older adults are likely to have more severe symptoms, although there is no evidence that Covid-19 is being spread through the mail. Coronaviruses do not survive well on surfaces. So, experts think the risk of spread from products or packages shipped over days or weeks is extremely low.
How long can HCoV-19 survive on aerosols?	A recent study shows that the virus can live in the air and on surfaces from several hours to several days. On plastic, it can live for up to 72 hours (3 days). On stainless steel, it may live for 48 hours (2 days) on surfaces that are less than 6 feet (1 meter). However, on cardboard, it is not known to be safe to touch or clean surfaces that many people touch a day.
What is the range of incubation periods for COVID-19 in humans?	10-27 days
There are evidence for asymptomatic COVID-19 transmissions?	There are currently no evidence of asymptomatic covid-2019 transmissions.
What percentage of COVID-19 cases in humans had incubation periods of longer than 14 days?	21 %

Tabella 5.6: Esempi di risposte fornite dalla Covid-19 FAQ Chatbot.

In una conversazione lunga e complessa il comportamento del bot può risultare più bizzarro rispetto agli esempi in Tabella 5.5 ed in Tabella 5.6 anche rimanendo in ambito COVID-19 per l'intera durata del dialogo.

Ad esempio, se si ripropone più volte la stessa domanda e poi la si cambia, il sistema, abituato a fare riferimento al contesto, prenderà in considerazione

anche un certo numero di messaggi precedenti producendo risposte inadeguate senza rendersi conto che l'argomento è cambiato e tali messaggi non hanno più importanza. Questo potrebbe essere uno dei motivi per cui alla medesima questione è possibile che il bot replichi in modo diverso.

Un altro difetto è che alla prima domanda che fa riferimento al COVID-19, qualunque essa sia (ex. "*Are obese people more at risk of getting COVID-19?*"), il modello spesso risponde con una breve descrizione del virus senza tenere conto delle reali informazioni richieste dall'interlocutore. Se la stessa questione viene posta invece dopo una domanda del tipo "*What is COVID-19?*" probabilmente il messaggio di risposta sarà adeguato perchè il modello è già cosciente che l'argomento di dialogo è il COVID-19 e tiene conto dei termini peculiari presenti nella domanda per far proseguire la conversazione.

Questi problemi possono essere ricondotti alla scarsità di materiale per l'addestramento nel formato necessario (domanda-risposta) attualmente disponibile. Alcuni sistemi sviluppati recentemente sono riusciti ad aggirare questa criticità in vari modi:

- Facendo uso di un modello che ha lo scopo di comprendere e produrre testo in campo biomedico. È il caso di **BioBert**[26], un language model basato su BERT che ha eseguito il pre-training su articoli di **PubMed** (per acquisire informazioni e termini riguardanti lo specifico ambito) e **SQuAD** (per acquisire l'abilità di eseguire question answering).

Testato sulle 156 istanze di uno dei dataset citati in precedenza[25], il modello ha prodotto risposte corrette nel 40.4% dei casi⁵.

- Avendo un approccio ibrido fra retrieval-based e generativo che permette di fare zero-shot question answering in tempo reale (senza l'ausilio del fine-tuning). Tale idea è stata applicata con **CovidASK**⁶, una chatbot pubblicata dal DMIS Laboratory in collaborazione con la Korea University e fruibile pubblicamente via web⁷. Quando al modello viene posta una domanda, esso è grado sia di reperire l'intero articolo in cui l'argomento viene trattato che di fornire la risposta precisa.

È stato valutato sul dataset Covid-19 Questions prodotto dagli sviluppatori stessi ottenendo un 35.8% di risposte corrette.

Nella Tabella 5.7 sono riportati alcuni confronti fra le risposte fornite da CovidAsk e dal sistema sviluppato alla medesima domanda.

⁵<https://venturebeat.com/2020/04/27/researchers-release-data-set-to-evaluate-covid-19-chatbots-and-search-engines/>

⁶<https://openreview.net/pdf?id=Wssn20iNf6j>

⁷<https://covidask.korea.ac.kr/>

#	Question	CovidAsk answers	Covid-19 F.A.Q. Chatbot's answers
1	What is the range of incubation periods for COVID-19 in humans?	Our results show that the incubation period falls within the range of 2-14 days with 95% confidence and has a mean of around 5 days when approximated using the best-fit lognormal distribution.	4 - 7 days
2	What percentage of COVID-19 cases in humans had incubation periods of longer than 14 days?	Incubation periods of 92% patients were no more than 14 days	50%
3	How does the incubation period for COVID-19 in humans vary by age?	Average incubation period of COVID-19 is around 6.4 days, ranges from 0-24 days.	In general, a "normal" person would have a "mild" symptoms, but in this case, it was a "severe" symptoms that led to the onset of symptoms.
4	What is proof that a 14-day quarantine period is not optimal for effectively controlling COVID-19 in children?	Conversely, structural stability of the SARS-CoV-2 Mpro with respect to mutations of the binding cavity and adjacent flexible loops indicates that the protein's mutability will pose a further challenge to the rational design of small-molecule inhibitors.	The presence of a single virus in a population can not be considered a guarantee that a given virus will stay on surfaces and objects, including tables.
5	Is there concrete evidence for the presence of asymptomatic transmissions?	Current epidemiologic evidence supports a major role in transmission for direct contact with live camels or humans with symptomatic MERS, but little evidence suggests the possibility of transmission from camel products or asymptomatic MERS cases.	There is no concrete evidence that asymptomatic transmission was detected in the united states in 2017.
6	What is the basic reproductive number for the novel coronavirus?	Initially, the basic reproductive number, R_0 , was estimated to be 2.2 to 2.7.	Basic reproductive size of the novel is 2.2.
7	What is evidence that the novel coronavirus can be transmitted through fecal excretion?	Our results provide evidence for gastrointestinal infection of SARS-CoV-2, highlighting its potential fecal-oral transmission route.	The novel virus can also be transmitted via fecals excretion
8	Will the number of COVID-19 cases decline in the summer?	Compared to adults, there has been a significantly smaller number of reported cases of COVID-19 in the pediatric population, although the incidence is increasing every day.	Covid-19 is a virus strain, first identified in wuhan, hubei province, china, that has only spread in people since december 2019. Health experts are closely monitoring the situation because little is known about this new virus and it has the potential to cause severe illness and pneumonia in some people.
9	What does the literature say about the effect of temperature variation and humidity on the mortality rate of COVID-19 in Wuhan?	In summary, this study suggests the temperature variation and humidity may be important factors affecting the COVID-19 mortality.	That variation and variation in humidity and temperature variation in wu-1 is a major obstacle in mortality rates in wunhan.

Tabella 5.7: Confronto fra le risposte di CovidAsk e della chatobot realizzata.

In certi casi nessuna delle chatbot riesce a dare una risposta soddisfacente (domande 3, 4, 5, 8) o solamente una è adeguata (domanda 9), in altri casi le informazioni reperite sono le medesime (domande 6, 7) ed in altri ancora differiscono in quanto gli articoli scientifici stessi si contraddicono in merito ai dati trattati (domande 1, 2). Nella maggior parte degli esempi entrambi i modelli producono una risposta **corretta** del punto di vista grammaticale e sintattico. Mentre la Covid-19 FAQ Chatbot realizzata è grado anche di dare riposte precise ed estremamente sintetiche (domande 1 e 2), CovidAsk utilizza sempre frasi estrapolate da articoli più ampi in cui:

- si fanno riferimenti non comprensibili, ad esempio al virus MERS nella domanda 5 o "*this study*" nella domanda 9;
- si fa uso di locuzioni poco adeguate, ad esempio "*In summary*" nella domanda 9 o "*Conversely*" nella domanda 4.

5.3.3 Browser Chat

Il framework ParlAI fornisce vari modi per interfacciarsi ai modelli realizzati: via terminale, attraverso Facebook Messenger o utilizzando un qualsiasi browser. Quest'ultima modalità è stata testata in locale su uno dei modelli addestrati creando un file di configurazione adeguato che viene utilizzato dagli appositi script di ParlAI. È stato utilizzato il servizio ngrok⁸ per esporre pubblicamente un indirizzo della macchina privata come fosse un server.

In Figura 5.8 si trova un esempio di interazione attraverso l'interfaccia web.

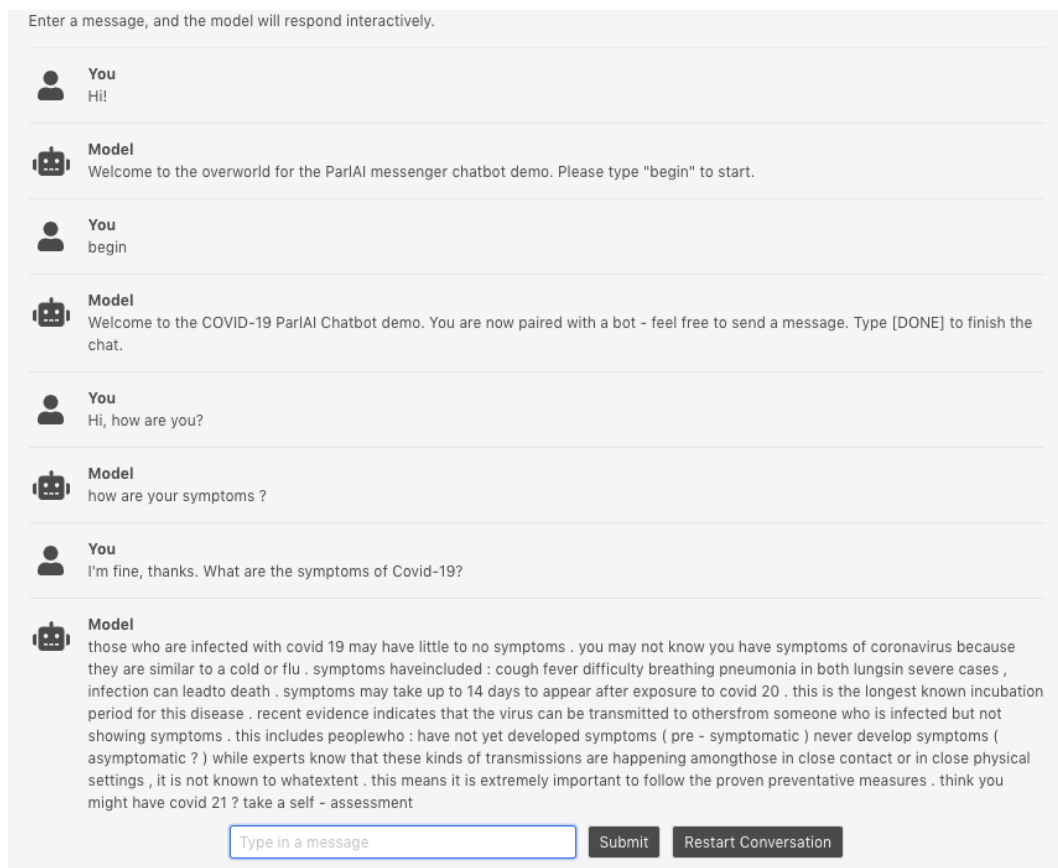


Figura 5.8: Esempio di iterazione usando l'interfaccia web.

⁸<https://ngrok.com/>

Ringraziamenti

Il primo ringraziamento va al relatore di questo lavoro, il Prof. Gianluca Moro, che ha reso possibile tutto ciò e ha acceso il mio personale interesse nei confronti di questa splendida disciplina.

Un ringraziamento speciale a Paolo per aver corretto tutte le mie bozze ed avermi sostenuta nella mia scelta di studiare senza farmelo pesare.

Grazie a Martina per aver ascoltato i miei sfoghi nei momenti di difficoltà ed aver condiviso alcune sessioni di studio.

Grazie a Cristian per aver creduto nelle mie capacità ed avermi coinvolta in vari progetti universitari.

Un ringraziamento alla mia famiglia che mi ha supportata e sopportata in questi anni e tutte le persone che mi sono state vicine durante il mio percorso.

Infine ringrazio me stessa per la diligenza nel cercare di raggiungere questo traguardo nel minor tempo possibile, impegnandomi anche nei momenti di sconforto.

Bibliografia

- [1] Tom M. Mitchell. *Machine learning*. McGraw Hill series in computer science. McGraw-Hill, 1997.
- [2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.
- [4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [5] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. *CoRR*, abs/1606.06031, 2016.
- [6] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537, 2019.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

-
- [8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.
- [9] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790, 2020.
- [10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019.
- [11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018.
- [12] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual universal sentence encoder for semantic retrieval. *CoRR*, abs/1907.04307, 2019.
- [13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002.
- [14] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696, 2020.
- [15] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977, 2020.
- [16] Margaret Li, Jason Weston, and Stephen Roller. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087, 2019.
- [17] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019.

-
- [18] Jason Weston, Emily Dinan, and Alexander H. Miller. Retrieve and refine: Improved sequence generation models for dialogue. *CoRR*, abs/1808.04776, 2018.
- [19] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098, 2019.
- [20] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. *CoRR*, abs/2004.13637, 2020.
- [21] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241, 2018.
- [22] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics, 2019.
- [23] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents’ ability to blend skills. *CoRR*, abs/2004.08449, 2020.
- [24] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. COR-19: the covid-19 open research dataset. *CoRR*, abs/2004.10706, 2020.
- [25] Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. Rapidly bootstrapping a question answering dataset for COVID-19. *CoRR*, abs/2004.11339, 2020.

- [26] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. *CoRR*, abs/1909.08229, 2019.