UNIVERSITÀ DI PISA

DIPARTIMENTO DI
FILOLOGIA, LETTERATURA E LINGUISTICA

CORSO DI LAUREA IN INFORMATICA UMANISTICA

TESI DI LAUREA MAGISTRALE

Transformer and logical inference: The case of bridging anaphora

CANDIDATO
Francesco Gracci

RELATORE
Chiar.mo Prof. Alessandro Lenci

CONTRORELATORE
Chiar.mo Prof. Felice Dell'Orletta

ANNO ACCADEMICO 2020/2021

Abstract: L'obiettivo della presente tesi è quello di utilizzare un *challenge set* creato manualmente per stabilire se dei modelli neurali pre-addestrati su task di *Natural Language Inference* (NLI) siano in grado di riconoscere e rappresentare il fenomeno linguistico della *bridging anaphora*. Il progetto ha previsto la creazione di un dataset di 300 coppie di frasi nelle quali fosse riscontrabile il suddetto fenomeno. In seguito, il dataset così ottenuto è stato utilizzato come test set per valutare le prestazioni di tre modelli neurali (BART, RoBERTa e DistilBERT) e analizzare le loro performance in relazione al fenomeno della *bridging anaphora*.

Indice

1. Introduzione

La capacità dei modelli neurali nel sapere apprendere i meccanismi del linguaggio umano offre da anni risultati sempre più interessanti. L'ambito del *Natural Language Processing* (NLP), letteralmente l'Elaborazione del Linguaggio Naturale, ha infatti utilizzato sempre di più modelli basati su reti neurali, spesso con architettura complesse. Riuscire a comprendere quale conoscenza linguistica venga effettivamente appresa dai modelli è diventato un tema centrale in molti degli studi proposti negli ultimi anni, favoriti dall'adozione di metodologie di indagine come quelle dei *probing task* o dei *challenge set*. Approcci come questi permettono di analizzare le abilità inferenziali dei modelli neurali, che possono essere considerate in relazione a diversi fenomeni linguistici, come per esempio la *bridging anaphora* a cui sono dedicati gli studi proposti nel presente elaborato.

Il presente studio è quindi finalizzato ad analizzare le abilità inferenziali di alcuni modelli neurali con particolare riferimento al fenomeno della *bridging anaphora*. Per uno studio di questo tipo è stato costruito manualmente un dataset di 300 coppie di frasi, tutte contenenti casi di *bridging anaphora*, su cui sono poi stati eseguiti esperimenti di *Natural Language Inference* (NLI) tramite l'utilizzo di modelli neurali di tipo *transformer*.

Il primo capitolo di questa tesi è dedicato allo stato dell'arte di NLP e considera nello specifico il task di NLI (cfr. ¶ 2.2), illustrandone le caratteristiche e le problematiche, e proponendo le metodologia denominate rispettivamente *probing task* e *challenge set* (cfr. ¶ 2.4). Nel secondo capitolo si presenta il fenomeno della *bridging anaphora* sia per quanto riguarda gli studi relativi alla *bridging resolution* (cfr. ¶ 3.4), sia in relazione al task di riconoscimento del *Textual Entailment*. Il terzo capitolo è dedicato al dataset di *bridging anaphora* costruito appositamente per gli studi proposti nel presente elaborato, con particolare attenzione verso la

classificazione dei tipi di relazioni di *bridging* (cfr. ¶ 4.4). Infine, il quarto capitolo è interamente dedicato agli esperimenti condotti sul dataset e al confronto dei risultati ottenuti dai diversi modelli neurali (cfr. ¶ 5.6).

2. Stato dell'arte

In questo primo capitolo viene fornito un quadro sintetico dello stato dell'arte del Natural Language Processing (NLP), con particolare attenzione verso i *word embedding* e i *sentence embedding* (sezione 2.1). Si introduce il Natural Language Inference (NLI) e il task *Recognizing Textual Entailment* (RTE) (sezione 2.2). Infine si presenta la metodologia del probing task (sezione 2.3) e rispettivamente i Benchmark dataset e i Challenge set (sezione 2.4).

2.1. Natural Language Processing

Negli ultimi anni la Linguistica Computazionale e il campo del *Natural Language Processing* (NLP) sono stati protagonisti di progressi eccezionali, che hanno contribuito a renderli loro una componente essenziale nell'ambito dell'Intelligenza Artificiale (AI). Tra i fatti che hanno contribuito maggiormente a tale cambiamento vi sono da una parte la sempre maggiore disponibilità di grandi quantità di dati testuali, mentre dall'altra un approccio teorico con un occhio di riguardo verso le discipline statistiche e probabilistiche (Lenci, Montemagni, Pirelli; 2016). In particolare, quest'ultima novità è stata una conseguenza della rivoluzione statistica che si è venuta a creare negli ultimi decenni del Novecento. Fino ad allora, infatti, si parlava soprattuto di un approccio di tipo *grammar-based*, fondato quindi sull'utilizzo di grammatiche e di regole annotate manualmente, volte a sottolineare un approccio di tipo deterministico al linguaggio, che è stato quindi sostituto dal quello *model-based*, basato invece sull'utilizzo di modelli statistici e probabilistici.

Un'altra novità è stata l'avvento delle cosiddetta Ipotesi Distribuzionale (Harris; 1954), secondo cui per conoscere il significato di una parola è necessario conoscere i contesti in cui essa viene utilizzata. Se noi consideriamo quindi due espressioni linguistiche, l'Ipotesi Distribuzionale può essere così riassunta: *"The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear"* (Lenci; 2008). Se quindi due parole hanno significato simile, allora tenderanno ad avere una distribuzione statistica piuttosto simile, oltre che occorrere nei medesimi contesti linguistici. Ciò ha portato alla nascita dei Distributional Semantic Models (DSM), detti anche Word Space Models, basati sull'utilizzo di *word embedding*. Se infatti si considera l'Ipotesi Distribuzionale, due parole possono essere viste come due vettori le cui componenti codificano la rispettiva distribuzione delle due parole nei vari contesti in cui esse sono state osservate. Un DSM non fa altro che, sulla base di dati statistici elaborati dai corpora, proiettare in uno spazio vettoriale dove ciascuna

parola è un vettore, un *word embedding*. In tal caso, per confrontare due parole basterà misurare la distanza tra i loro rispettivi vettori distribuzionali.

2.1.1.Word Embedding

I *word embedding* sono quindi stati utilizzati in modo sempre più frequente nello svolgimento di task di NLP. In particolare, i suddetti vettori possono essere generati sia a partire da una matrice di co-occorrenza, oppure mediante l'utilizzo delle reti neurali. Quest'ultime possono presentare architetture diverse e costituiscono la base nell'ambito del *machine learning*. Le più utilizzate nei task di NLP sono le *deep neural network* ("reti neurali profonde") che presentano un'architettura fatta di molti *hidden layer* ("livelli nascosti") e sono quindi capaci di estrarre le *feature* dai corpora utilizzati per l'addestramento. In particolare tale estrazione avviene in maniere autonoma e non supervisionata, senza quindi un'intervento manuale che possa "aiutare" la rete durante l'apprendimento. Da tale processo si ottengono i *pre-trained word embedding* che possono essere utilizzati in altri task di NLP.

I primi embedding utilizzati erano di tipo statico e ogni parola veniva considerata indipendentemente dal contesto in cui essa ricorreva. L'estrazione avveniva utilizzando modelli basati su algoritmi non supervisionati come Word2Vec (Mikolov et al; 2013), GloVe (Global Vectors , Pennington et al.; 2014) e FastText (Bojanowski et al.; 2017).

In particolare Word2vec è costituito da due modelli differenti (vedi figura 1). Il primo, chiamato CBOW (*Continuous Bag of Words*), viene addestrato su un corpus linguistico così da predire una parola in un dato contesto; il secondo, *Skip-gram*, viene addestrato per l'operazione contraria, ovvero predire il contesto di un singolo *word embedding*. Entrambi gli algoritmi utilizzano una finestra di controllo che stabilisce quante parole considerare prima e dopo la parola target, e quindi la dimensione del contesto.

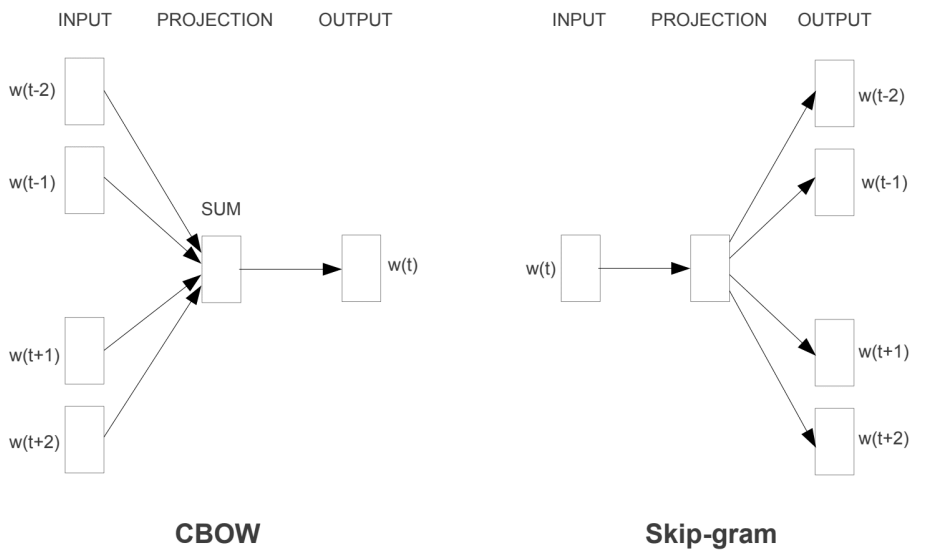

Figura 1: CBOW e Skip-Gram

In seguito si sono diffusi i *word embedding* contestualizzati, di modo che potessero tenere conto della possibilità che una stessa parola abbia più di un significato (polisemia) e del cosiddetto *meaning shift*, letteralmente "cambiamento di significato", causato dal contesto in cui la parola viene inserita. Tra gli esempi possiamo considerare il caso degli embedding generati da ELMo (Embedding from Language Models , Peters et al.; 2018) che tengono conto sia delle parole che seguono sia di quelle che precedono la parola da rappresentare, oppure il caso di BERT (Devlin et al.; 2018) che vedremo nel dettaglio in seguito. Inoltre, gli embedding di ELMo sono *character-based* e riescono a catturare informazioni di tipo morfologico, utili nei casi in cui si debbano rappresentare delle parole che non erano presenti nel vocabolario del corpus utilizzato per l'addestramento.

2.1.2. Sentence Embedding

La proiezione del significato di una parola su uno spazio vettoriale ha portato in seguito a considerare il significato di un'intera frase, generando i cosiddetti s*entence embedding*. Anche in questo caso ci sono più modi per ottenerli e, come per *word embedding*, anche qui dipende dal considerare o meno l'informazione

contestuale. Se si sceglie di ignorare quest'ultima si può utilizzare il metodo più semplice (Mitchell e Lapata; 2010), ovvero tramite una rappresentazione additiva di tipo *Bag-Of-Words* (BOW), dove le parole vengono considerate indipendentemente l'una dall'altra e per ottenere il *sentence embedding* di una frase non bisogna far altro che sommare i *word embedding* delle parole che costituiscono le frase. Una strada più complessa è quella di utilizzare modelli neurali non supervisionati, in particolare modelli di tipo *Encoder-Decoder* (ED) basati su reti neurali ricorrenti (RNN) e che, data una frase, tentano di ricostruire le relative frasi circostanti. Tra i suddetti modelli troviamo per esempio *Skip-Thought Vectors* (Kiros et al.; 2015). Se consideriamo invece degli algoritmi supervisionati non troviamo dei risultati degni di nota precedenti il lavoro di Conneau at al. (2017), in cui è stato implementato il metodo *InferSent*. Secondo quest'ultimo, la qualità delle rappresentazioni generate dipende fortemente dal dataset utilizzato per il training. Ciò è stato confermato dimostrando che i risultati ottenuti dai modelli addestrati su task di riconoscimento delle inferenze contenute nel corpus SNLI (Stanford Natural Language Inference corpus , Bowman et al.; 2015), con un BiLSTM e con la tecnica del Max pooling (vedi figura 2), fossero migliori rispetto a quelli basati sul metodo *Skip-Thoughts-Vectors*.

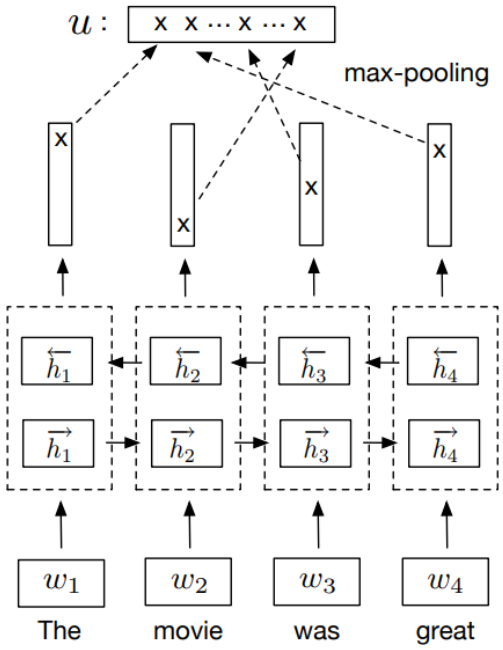

Figura 2: rete Bi-LSTM con max-pooling

La difficoltà di un approccio supervisionato risiede inoltre nel dover avere a disposizione un dataset annotato, che quindi richiede una dimensione e una composizione adeguate per ottenere un *pre-trained sentence embedding* di qualità. Nel 2018 Subramanian et al. hanno proposto il metodo *multi-task learning*, che unisce gli approcci *Skip-Thought* e *InferSent* e si basa su un addestramento dei modelli contemporaneamente su vari di tipi di task, cercando di sfruttare le similarità e le differenze di quest'ultimi. Tra gli esempi di questa applicazione troviamo il *Google Universal Sentence Encoder* (2018), basato su un'architettura neurale *sequence-to-sequence* di tipo *transformer* (Vaswani at al.; 2017).

La stato dell'arte del NLP è quindi dato dai modelli che utilizzando i *pre-trained word embedding* e i *pre-trained sentence embedding*, permettono di ottenere risultati notevoli nei diversi task computazionali. C'è però da dire che la loro complessità cresce in base ai risultati delle loro performance: definire infatti l'effettivo livello di conoscenza linguistica appreso dai vettori è sempre più difficile (Liu at al.; 2019). Anche le stesse reti, come abbiamo sottolineato nelle righe precedenti, hanno mano a mano sviluppato una struttura sempre più complessa, opaca (Belinkov, Glass; 2019), culminando nelle cosiddette *Deep Neural Network* (DNN), centrali in numerosi studi degli ultimi anni. Le DNN riescono ad estrarre regole e pattern semplicemente mediante l'addestramento su *high dimensional data* (dataset di grandi dimensioni). La loro efficienza dipende però da un numero considerevole di iperparametri, che implica un'impossibilità nel tener conto di tutte le possibili combinazioni di valori assunti da quest'ultimi. Riuscire a definire i meccanismi che influenzano il comportamento delle rete e stabilire il livello di conoscenza di quest'ultima, sono di conseguenza compiti molto complessi.

Il problema dell'opacità degli *embedding* e delle architetture neurali è al giorno d'oggi uno dei casi di studio principali del NLP (Ettinger et al.; 2018). Ciò ha contribuito al diffondersi di studi in cui si applicano metodologie dei cosiddetti *probing task* (*Conneau et al.; 2018*), che vedremo tra qualche paragrafo.

2.1.3.Transformer

Nel 2018 c'è stato un punto di svolta per lo sviluppo dei modelli neurali per NLP. In quel anno Google ha infatti rilasciato BERT (Devlin et al.; 2018), che è stato acclamato da tutta la comunità accademica. In particolare, BERT risultava essere un nuovo tipo di modello che utilizzava l'architettura *Transformer*.

I *transformer* sono stati proposti per la prima volta nel 2017 (Vaswani et al., 2017) e sono composti da un *encoder* e da un *decoder*. L'aspetto più interessante è però il meccanismo di *self-attention* che permette alla macchina di catturare meglio le informazioni riguardanti le parole, in particolare tenendo conto delle informazioni relative al contesto. Questo significa che dato l'encoding di una singola parola, il *transformer* utilizza anche le informazioni derivanti dalle parole che la precedono o che la seguono.

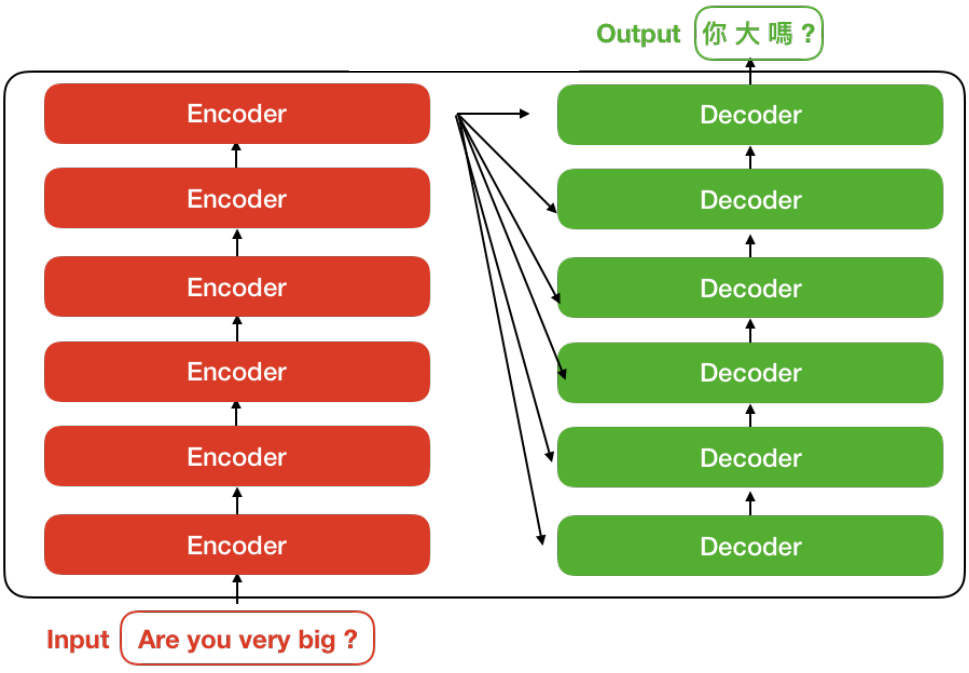

Figura 3: Architettura del transformer

In figura 3 possiamo osservare la composizione del *trasnformer*, che è costituito da due pile di encoder e di decoder. Gli encoder sono composti da un S*elf-Attention layer*, ovvero quello che prende un input, ne evidenzia le parti importanti e lo passa

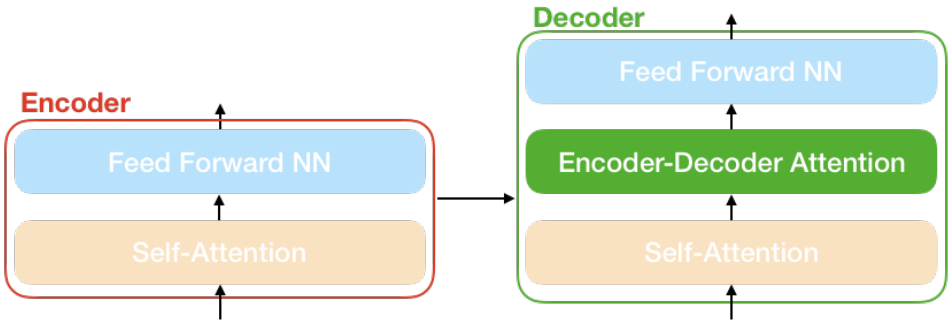

Figura 4: composizione dell'encoder e del decoder

ad una rete *feed-forward*, mentre il decoder ha una struttura *Encoder-Decoder* relativa all'*attention* (vedi figura 4).

Per meglio comprendere il funzionamento dell'*Attention layer* consideriamo la figura 5. Nello specifico, la parola "it" nell'immagine (evidenziata in grigio) è un embedding, dove ogni parola contenuta nelle frase assume un valore di rilevanza diverso. Se infatti consideriamo la coppia "the animal", le parole che la costituiscono saranno quelle che conterranno di più nella produzione dell'embedding, mentre una parola come "because" conterà molto meno in quanto con rilevanza minore in relazione alla menzione "it".

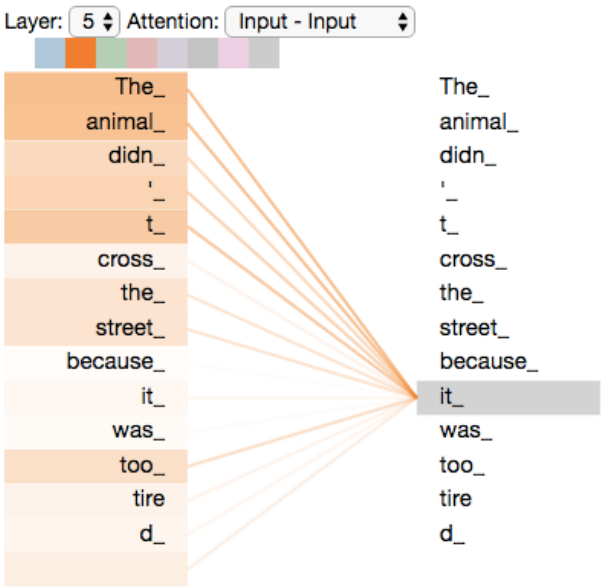

Figura 5: meccanismo di *attention*

Tutti i processi sopra descritti sono rappresentati nella figura 6, che descrive nel dettaglio il comportamento di un *transformer*.

Innanzitutto ogni parola viene convertita in un vettore, che viene poi passato al *Self-Attention layer* così che possa identificare le parole del contesto più rilevanti. Successivamente, un *Normalization layer* (Ba et al.; 2016) somma l'output del S*elf-Attention layer* con l'input originale così da unire le informazioni relative al contesto a quelle presenti nell'input di partenza. Il risultato di questo passaggio viene poi passato a un *feed forward networks* e l'output ottenuto da quest'ultimo ad un altro N*ormalization layer*. L'output così prodotto viene passato al successivo livello di encoding ripetendo gli stessi passaggi fino a che l'informazione così processata non arriva al decoder.

Il decoder lavora in modo molto simile all'encoder, se non per alcun differenze. Il S*elf-Attention layer* considera infatti solo le posizioni precedenti, mentre l'E*ncoder-Decoder Attention layer* lavora come un *multiheaded self-attention*.

L'output prodotto del decoder viene poi processato da un L*inear layer*, che è una rete neurale *fully connected* che produce il cosiddetto vettore *logits* contenente le probabilità di ciascuna parola presente nel dizionario. La parola più probabile viene infine selezionata dal S*oftmax layer*.

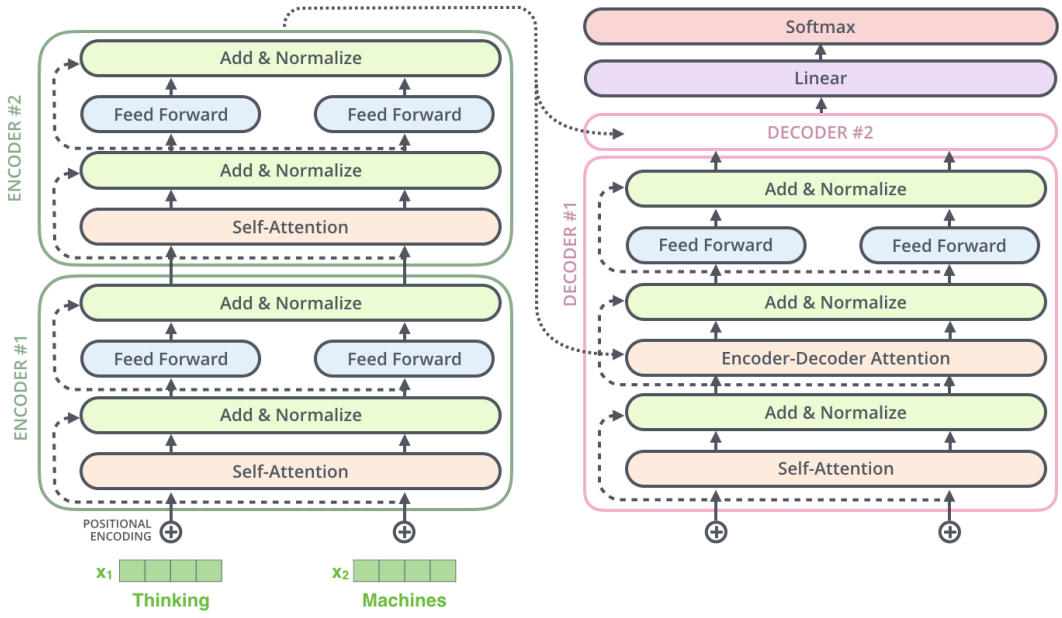

Figura 6: Esempio di funzionamento di un transformer

2.1.4.BERT

Dopo aver approfondito il funzionamento dei *Transformer*, in questo paragrafo vedremo come addestrare BERT per la creazione di un *Language Model* (LM). In particolare, l'aspetto più interessante di questo processo è la possibilità di poter creare un LM con un approccio semi-supervisionato.

BERT utilizza due diverse strategie di training. La prima (vedi figura 7) prende il nome di *Masked LM* e consiste nel mascherare il 15% delle parole presenti nel testo da passare alla rete, così che il modello sia addestrato per predire le suddette parole. In questo modo il modello riesce ad apprendere una buona rappresentazione del linguaggio pur non utilizzando dei dati etichettati.

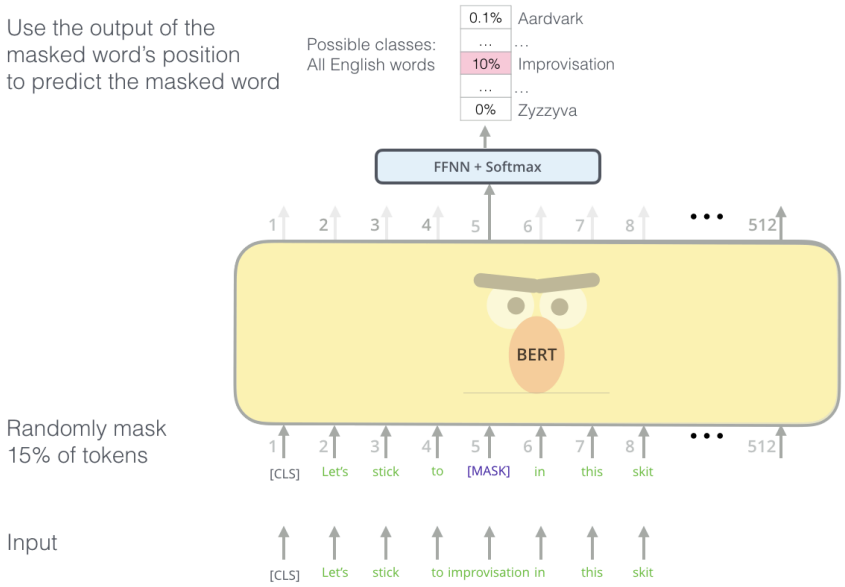

Figura 7: Masked LM in BERT

Il secondo approccio (vedi figura 8) è simile al primo, ma si concentra sulle frasi. Esso consiste infatti nel passare al modello una coppia di frasi così da ottenere una predizione sull'ordinamento delle stesse nel documento originale, ovvero se la seconda frase segua o meno la prima. In questo modo non risultano necessari dati strutturati, ma è sufficiente disporre di un numero adeguato di documenti.

Queste due procedure sono i training task, che permettono di generare un ampio *Language Model* che è in grado di adattarsi a diversi tipi di task. Questo fenomeno

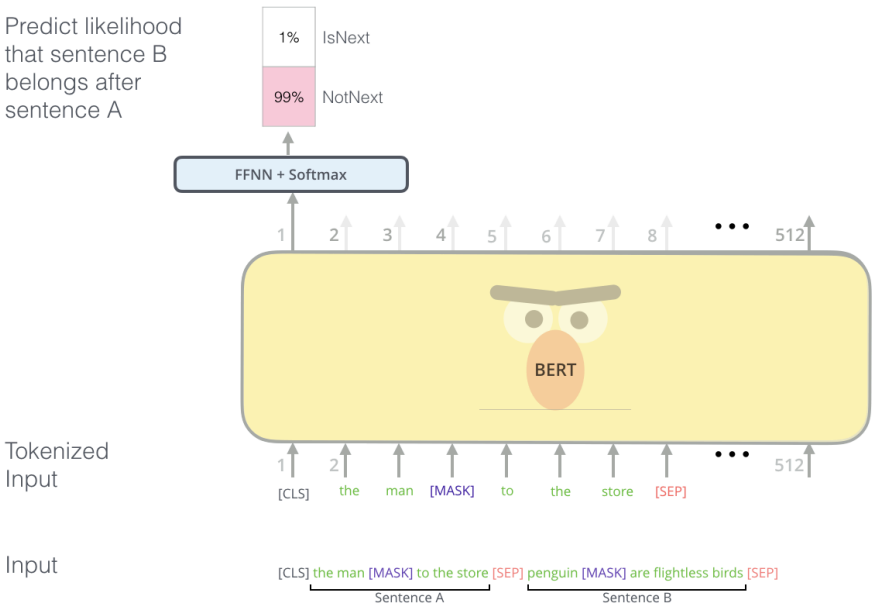

Figura 8: Next Sentence Prediction in BERT

prende il nome di *Transfer Learning* ed è uno dei più importanti traguardi raggiunti da NLP negli ultimi anni. Al giorno d'oggi non conviene infatti addestrare un modello *from scratch* (da zero), ma piuttosto utilizzare un modello pre-addestrato con i metodi descritti sopra e successivamente eseguire un *fine tuning* per insegnare alla macchina come elaborare diversi tipi di task (vedi figura 9). Per farlo, ci sono tre strade possibili:

- addestrare il modello nella sua interezza, aggiornando i parametri;
- *freeze* (congelare) alcuni pesi del modello originale e poi eseguire un nuovo training per il task differente;
- lasciare il modello così com'è, costruire una nuova rete e utilizzare l'output del modello originale come input per la rete incaricata di eseguire il task.

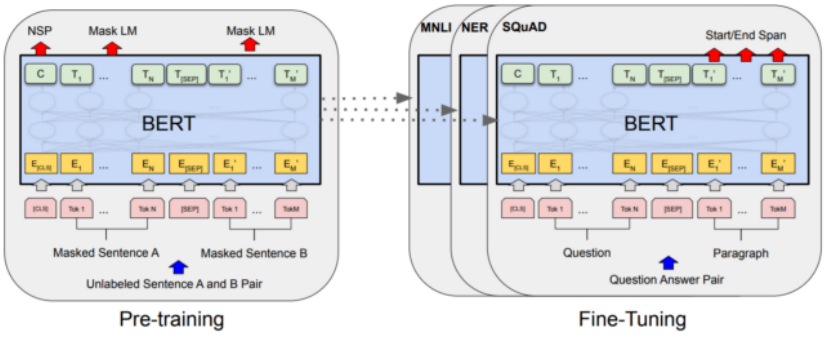

Figura 9: Procedure di pre-training e di fine-tuning per BERT

Su internet è possibile trovare un gran numero di modelli pre-addestrati per eseguire task differenti. Molti di questi sono raccolti sulla pagina web di *Hugging Face*[1]. Per gli scopi di questa tesi abbiamo utilizzato alcuni di questi modelli addestrati per eseguire task di *Natural Language Inference* (NLI), e li abbiamo testati per capire se fossero in grado di riconoscere i casi di *bridging anaphora*.

2.2. Natural Language Inference

Il Natural Language Inference (NLI) è uno dei principali task di NLP. L'idea è che, da una coppia di frasi, l'informazione espressa dalla seconda possa essere inferita a partire dall'informazione espressa dalla prima. Nello specifico la prima frase prende il nome di "premessa" mentre la seconda di "ipotesi". Questa tipologia di task è divenuta sempre più interessante negli ultimi anni tanto da essere oggetto di molti studi e lavori (Bowman et al.; 2015, Marelli et al.; 2014, Nangia et al.; 2017). Se una macchina è in grado di comprendere correttamente le implicazioni tra frasi del linguaggio naturale, allora si può affermare che essa abbia una buona conoscenza di come funziona la lingua (MacCartney, 2009).

Tra i task da segnalare, che vanno dal *question answering* fino alle valutazioni di *machine traslation system*, troviamo *Recognizing Textual Entailment* (RTE) ovvero il "riconoscimento della implicazione testuale", dove ogni coppia di frasi può essere

Relationship	Premise & Hypothesis	
Entailment	Premise:	This church choir sings to the masses as they sing joyous songs from the book at a church.
	Hypothesis:	The church is filled with song.
Neutral	Premise:	This church choir sings to the masses as they sing joyous songs from the book at a church.
	Hypothesis:	The church has cracks in the ceiling.
Contradict	Premise:	This church choir sings to the masses as they sing joyous songs from the book at a church.
	Hypothesis:	A choir singing at a baseball game.

Figura 10: Esempio di RTE

classificata come *entailment*, *contradiction* o *neutral*. Queste tre etichette indicano

[1] https://huggingface.co

quando l'ipotesi può essere inferita dalla premessa (*entailment*), quando contraddice l'informazione data dalla premessa (*contradiction*) o se l'ipotesi non è legata alla premessa e quindi nessun tipo di assunzione può essere fatta (*neutral*) (vedi figura 10).

Per sottolineare l'interesse generato da NLI, negli ultimi anni sono stati rilasciati due dataset interamente dedicati al suddetto task: SNLI (*Stanford Natural Language Inference*) (Bowman et al.; 2015), e MultiNLI (*Multi-genre Natural Language Inference*) (Williams et al.; 2018). Il primo è stato pubblicato dalla Stanford University nel 2015 ed è composto da 570k coppie di frasi in lingua inglese che sono state etichettate manualmente con i casi *entailment*, *contradiction* e *neutral*. L'interesse generato ha fatto sì che solo due anni dopo vedesse la luce il MultiNLI, Multi-genre Natural Language Inference corpus (Williams at al.; 2017), che è composto da 433k coppie di frasi, annotate con le stesse tre etichette utilizzate nell'SNLI. Anche il formato è il medesimo, ma il MultiNLI copre una maggior varietà di generi testuali, che vanno dall'inglese scritto a quello parlato.

I modelli neurali allo stato dell'arte basati sulla codifica dei *sentence embedding* sono stati testati su entrambi i corpora, ottenendo degli ottimi risultati in termini di accuratezza (Nie e Bansal; 2017, Conneau et al.; 2017). Ciò potrebbe far pensare che essi siano in grado di svolgere il task di NLI su tutti i generi testuali, se non fosse che i valori di accuratezza così ottenuti dipendono dai dati su cui i modelli sono stati addestrati. Si ottengono infatti delle performance peggiori quando i dati provengono da un'altra risorsa o sono stati costruiti diversamente, sottolineando i limiti della generalizzazione del modello (Naik et al.; 2018). Le reti neurali, infatti, apprendono le regolarità statistiche dei dati su cui vengono addestrati , ottenendo in seguito dei risultati ottimali quando ritrovano nel test dei pattern appresi che hanno la stessa distribuzione dei dati del training. Ciò comporta che la tendenza generale dei modelli sia limitarsi all'apprendimento della casistiche più semplici e più frequenti, tralasciando le proprietà linguistiche, ovvero ciò che dovrebbero catturare (McCoy at al.; 2019). È stato infatti dimostrato che i modelli tendano ad imparare le "*shallow*

heuristic" , andando a tener conto solo delle situazioni più comuni e frequenti all'interno del training set. Un esempio abbastanza comune è il caso della negazione, spesso indicata dal presenza di "not", e che, per le ragioni sopra descritte, tende ad essere classificata come *contradiction* dal modello. Un altro caso comune è quello in cui la premessa e l'ipotesi condividano gran parte delle parole. In questo caso il modello sarà portato ad individuare una relazione di *entailment*, senza alcun tipo di ragionamento in merito al significato delle due proposizioni.

2.3. Probing task

Quali sono le informazioni linguistiche che vengono realmente apprese e inserite nelle rappresentazioni vettoriali? Come abbiamo visto nel paragrafo precedente rispondere a questa domanda non è così semplice e, per questa ragione, sono nate le metodologie dei *probing task* (Conneau et al.; 2018) e in particolare, all'interno del *Natural Language Inference* (NLI), dei *challenge set* (Lehmann et al.; 1996).

"*A probing task is a classification problem that focuses on simple linguistic properties of sentences*" (Conneau et al.; 2018). In particolare l'approccio dei *probing task* consiste nell'utilizzare un modello già addestrato su un particolare task, congelare gli *embedding* con i pesi prodotti dal suddetto addestramento e utilizzarli come input su un nuovo task di classificazione. Se il nuovo classificatore ha successo significa che il modello usato alla base è in grado di catturare le informazioni richieste dal nuovo task e di memorizzarle nei suoi *embedding*. In altre parole, se il classificatore è in grado di comprendere il significato delle frase allora il modello su cui è stato addestrato ha memorizzata quel tipo di informazione. In caso contrario, il fenomeno linguistico richiesto non è stato catturato dagli *embedding* durante l'addestramento (Ettinger et al.; 2018).

Un lavoro che presenta l'applicazione deI *probing task* è quello pubblicato da Adi et al. (2017), che si è concentrato sulla capacità del modello nel saper codificare informazioni linguistiche relative alla struttura delle frasi catturate dai *sentence*

embedding, mediante l'ausilio di tre tipi di *probing task*: *length task*, predire la lunghezza della frase *s*; *word-content task*, determinare se una parola *w* è contenuta nella frase *s*; *word order task*, ricostruire l'ordinamento delle parole presenti in *s*. Questi elementi sono stati valutati sugli *embedding* ottenuti mediante CBOW e su quelli ottenuti mediante l'architettura *sequence-to-sequence* di un *Encoder-Decoder* (ED). CBOW si è dimostrato efficiente sui *length task* e *word order task*, ma non è riuscito ad ottenere performance sufficienti su *word-content task*. Il modello ED è invece efficace sui task *word-order* e *word content* già con vettori di piccole dimensioni, senza presentare grandi miglioramenti all'aumentare di quest'ultimi. Una simile osservazione va quindi a rafforzare l'idea secondo cui *"larger is not always better"*. Un simile approccio ha quindi permesso di comprendere fino a che punto i modelli "capiscono" e memorizzano la feature considerata.

Tra i lavori degli ultimi anni che vale la pena di segnalare ci sono poi quelli di Ettinger et al. (2018) in merito alla composizionalità delle frasi, quelli relativi alle abilità sintattiche di BERT (Goldberg (2019) o Jawahar et al. (2019)), quelli incentrati sulla rappresentazione degli alberi sintattici (Hewitt e Manning (2019)), e l'interessante proposta di Liu at al. (2019) che tramite l'utilizzo di 16 probing task indaga sulla *transferability dei contextualized word embedding* (CWRs). Troviamo inoltre il lavoro di Tenney et al. (2019), quello di Jiang e De Marneffe (2019) e quello di Richardson et al. (2019), oltre a tanti altri.

Quest'ultima rassegna di lavori sottolinea quanto siano importanti i *probing task*, considerando che permettono di esaminare quali caratteristiche linguistiche i modelli sono in grado di incorporare e in che misura. Tali metodologie presentano comunque un limite, legato all'approccio della classificazione. Non si ha infatti la certezza che la rete neurale abbia realmente usato l'informazione linguistica, catturata durante l'addestramento, per risolvere il task. Per esempio, il lavoro di Vanmassenhove et al. (2017) relativo alla *Neural Machine Translation* (NMT) mostra come il classificatore pre-addestrato con i *sentence embedding* abbia predetto correttamente il tempo verbale nel 90% dei casi, ma che lo abbia poi tradotto correttamente solo nel 79% dei

tempi verbali. Senza considerare che le reti neurali, seppur tendano a catturare un grande quantitativo di informazioni linguistiche, spesso sono solite apprendere solo i casi più frequenti. Tutto ciò implica che la situazione del *Natural Language Inference* (NLI) non sia poi tanto diversa, in particolare per quanto riguarda i training e test set utilizzati dai modelli neurali per il riconoscimento del *Textual entailment* (TE).

2.4. Benchmark dataset e challenge set

Per valutare e produrre modelli NLI si utilizzano i cosiddetti *Benchmark* dataset. Quest'ultimi sono degli insiemi di dataset relativi a vari tipi di task che vengono utilizzati per misurare le performance di un modello. Vengono ottenuti a partire da corpora testuali, cercando di rappresentare la distribuzione naturale del linguaggio. Il loro obiettivo è quello di fornire una valutazione complessiva dei modelli a cui sono sottoposti, senza quindi concentrarsi su un fenomeno linguistico in particolare. Per questa ragione risultano poco efficaci nella rappresentazione di informazioni linguistiche specifiche. I benchmark dataset più comuni sono per esempio quelli relativi al *POS-Tagging* e alla *Name-Entity recognition*.

I limiti sopra descritti portano alla nascita dei *challenge set*, ovvero dei dataset più specifici rispetto ai benchmark tradizionali (Belinkov, Glass; 2019). Consideriamo la seguente frase:

"The city councilmen refused the demonstrators a permit because they ***feared violence****"*.

In questo caso è impossibile stabilire a chi si riferisca "feared violence" (city councilmen o demostrators) basandoci esclusivamente sul contesto; è necessario essere in possesso di una conoscenza esterna per inferire che probabilmente è il consigliere a temere la violenza. I challenge set aiutano la rete neurale e cogliere la suddetta relazione, andando a rappresentare fenomeni linguistici meno frequenti, e quindi poco presenti nei dataset tradizionali, che cercano invece, come visto nel caso dei benchmark, di "*reflect naturally occurring data*"(Lehmann et al.; 1996).

Generalmente si tratta di dataset costruiti ed annotati a mano che, come i probing task, cercando di stimare la qualità degli embedding. Oltre all'NLI vengono utilizzati anche in ambiti specifici del NLP, come per esempio la *Machine Translation* (MT).

Alcuni benchmark tradizionali sono stati dotati di challenge set interamente dedicati al task di NLI. Il caso più noto è quello GLUE (Wang et al.; 2018), che sta per *General Language Understanding Evaluation*, ed è composto da nove differenti task per la comprensione delle frasi, di cui quattro sono task di inferenza basati su MNLI (Williams et al., 2018), QNLI (Rajpurkar et al. 2016), RTE e WNLI (Levesque; 2011). GLUE e il suo successore SuperGLUE (Wang at al.; 2019) sono dotati di un dataset "diagnostico" dedicato a 30 fenomeni riscontrabili negli ambiti di NLI, e suddivisi nella categorie *lexical semantics, predicate-argument structure , logic e knowledge*. Il suddetto dataset è costruito manualmente e cerca di coprire il più ampio spettro possibile di proprietà linguistiche. Costituito da 1000 coppie di frasi, i fenomeni che rappresenta vanno dall'anafora alla negazione, dell'iponimia all'iperorimia, passando per tanti altri casi specifici del linguaggio. Purtroppo la ridotta dimensione rischia di far sì che il modello incontri solo le forme più frequenti, impedendo quindi di valutarne le capacità, e quindi decidere quando esso sia destinato a fallire. Tutto questo tenendo conto di quanto i modelli neurali siano sensibili ai casi specifici e poco frequenti.

Un altro modo per testare le capacità del modello nel comprendere il linguaggio umano è attraverso il cosiddetto *Adversarial attack*, ovvero degli input annotati manualmente e creati per confondere la macchina. L'idea è quella di "attaccare" la buona previsione di un modello così da perturbarlo, e osservare quanto sia suscettibile a un cambiamento dei dati in input. Questo approccio può essere utilizzato anche per aumentare i dati di training al fine di ottenere prestazioni migliori. Da segnalare, a tal proposito, il lavoro TextAttack (Morris et al.; 2020).

L'approccio che abbiamo visto in merito ai *challenge set* è stato adottato per la costruzione del dataset (vedi cap. 3) utilizzato per gli esperimenti presentati nel presente elaborato (vedi cap. 4). Nello specifico, è stato considerato il fenomeno

linguistico della *bridging anaphora*, che verrà descritto nel dettaglio nel prossimo capitolo.

3. Bridging Anaphora

In questo capitolo verrà presentato il fenomeno della *bridging anaphora* con le analisi e gli studi compiuti in relazione ad esso. In particolare, dopo un'introduzione relativa al task di riconoscimento del *Textual Entailment* (TE) (sezione 3.1), definiremo la *bridging anaphora* offrendo una prospettiva storica sul fenomeno linguistico (sezioni 3.2 e 3.3). Andremo poi a presentare i corpora relativi allo studio del fenomeno (sezione 3.4) che saranno oggetto degli studi di *bridging resolution* che descriveremo in seguito (sezione 3.5).

3.1. Definizione di Bridging Anaphora

In linguistica esistono diverse funzioni di coesione linguistica dei testi, ovvero tutte quelle funzioni che possono essere utilizzate per collegare fra loro le componenti di un testo. In questo modo è possibile percepire il testo come un'unica entità, senza avvertire la distinzione tra enunciati diversi. Tra le suddette funzioni, una delle più importanti è senza dubbio l'anafora, che tramite il rapporto anaforico, è in grado di creare legami tra porzioni di testo più o meno vaste e più o meno distanti tra loro. La definizione enciclopedica della Treccani ci dice che "*l'anafora è il fenomeno per cui per interpretare alcuni sintagmi del testo occorre riferirsi a un altro costituente che compare nella parte precedente del testo stesso*". Consideriamo quindi il seguente esempio:

*Ho incontrato Giovanni in stazione e **gli** ho offerto un passaggio*

In questo caso il pronome "gli" rappresenta la ripresa anaforica dell'entità "Giovanni", che viene denominata "antecedente" (Lo Duca; 2008). In questo modo diventa possibile riferirsi a un costituente precedente senza dover ripetere nuovamente la stessa entità, magari come in questo caso rappresentata da un sintagma nominale. Può capitare che ad uno stesso antecedente possano corrispondere più riprese anaforiche, andando a creare la cosiddetta catena anaforica, dove ogni ripresa prende il nome di "anello" mentre l'elemento a cui si riferiscono viene denominato "capo-catena" (Angela Ferrari, *anafora*, Enciclopedia dell'Italiano, Treccani).

Generalmente il rapporto anaforico coinvolge soprattutto sintagmi nominali e pronomi, ma si possono individuare anche casi in cui vengono utilizzati sinonimi o perfino ellissi del soggetto. La ripresa anaforica è infatti caratterizzata dal cosiddetto grado di trasparenza, che ci dice quanto essa sia più o meno esplicita (Lo Duca; 2008).

Di seguito un elenco delle diverse forme di ripresa anaforica ordinate in base al grado di trasparenza, dal maggiore al minore[2]:

- ripetizione dell'antecedente:

  *Stamani ho incontrato il falegname. Il **falegname** sta lavorando ad una porta.*

- sinonimo:

  *Hai provato dal dottor Bianchi? Il miglior **medico** da cui sia mai stata.*

- iperonimo:

  *Qui c'era un salice, ma la **pianta** purtroppo è morta.*

- nome generale:

  *Mi passeresti il martello? Ho bisogno di quello **strumento** per finire il lavoro.*

- perifrasi:

  *Sono stato a Roma lo scorso fine settimana. La **città eterna** è sempre bellissima.*

- sinonimo testuale:

  *Il capo ha assegnato un sacco di lavoro da fare a Paolo. Il **poveretto** non potrà godersi il fine settimana.*

- pronome tonico :

  *Il direttore arrivò in ritardo. **Egli** dichiarò di aver perso l'autobus.*

- pronome atono:

  *Ho comprato un libro qualche giorno fa e non riesco a fare a meno di legger**lo**.*

- ellissi del soggetto:

  Non riesco a trovare Tommaso. Dobbiamo muoverci prima che Ø si faccia del male.

- anafora zero:

  Giacomo ha preferito rinunciare al posto. Ø Dover rispettare quel orario sarebbe stato peggio.

In tutti i suddetti casi le riprese anaforiche si riferiscono sempre alla stessa entità, e rappresentano quindi casi di coreferenza. Ci sono però delle forme di anafora che

[2] Elenco ripreso da Lo Duca 2008, con però esempi diversi

non presentano la coreferenza, come per esempio i casi limite del cosiddetto "incapsulatore anaforico" e quello della "anafora associativa" o "*bridging anaphora*". Nel caso dell'incapsulatore anaforico vi è un'anafora in cui l'antecedente è richiamato da una "capsula", che può anche essere un nome astratto ed essere dotato di aggettivi valutativi (Lo Duca; 2008):

> *Un giovane di Torino ha cercato di togliersi la vita questa mattina. Si pensa che il motivo dell'**insano gesto** sia stata una delusione amorosa.*

L'anafora associativa, invece, consiste in una ripresa anaforica tramite l'uso di nuovi referenti, che risultano funzionali grazie alle circostanze e al contesto già evocati (Kleiber, 1990; 2001):

> *Ieri sera Giovanni e Paola sono andati a mangiare ad un ristorante. I **camerieri** erano molto cortesi e lo **chef** è venuto a salutarli a fine serata.*

Nell'esempio sopra riportato troviamo due frasi contenenti un collegamento anaforico. Nello specifico, i termini "camerieri" e "chef" si riferiscono entrambi alla menzione "ristorante", e assumono quindi un significato specifico in relazione al suddetto termine. La ripresa anaforica consiste infatti nell'indicare come i referenti introdotti nella seconda frase, in questo caso "camerieri" e "chef", siano legati all'antecedente menzionato nella prima frase, ovvero "ristorante". Così facendo il lettore (o l'ascoltatore) saprà per certo che i camerieri e lo chef presentati nel discorso sono quelli del ristorante in cui Giovanni e Paola sono andati a cena, poiché troveranno corrispondenza grazie all'utilizzo dell'anafora associativa.
L'anafora associativa o *bridging anaphora* richiede quindi che ci sia un collegamento nascosto, o un'ancora, che sia stato introdotto in precedenza. In questo modo, le espressioni sono interpretate come se il collegamento fosse esplicito e anaforico (Sebastian Löbner; 1998). Risulta quindi molto importante saper individuare il legame che esiste tra l'anafora e il suo antecedente, anche perché in caso contrario risulterebbe impossibile interpretare correttamente la frase o il testo considerati (Kobayashi at al.; 2020). Vediamo un esempio:

*Even if baseball triggers losses at CBS – and he doesn't think it will – "I'd rather see **the games** on our air than on NBC and ABC," he says.*

in questo caso c'è un collegamento anaforico tra "the games" e il suo antecedente "baseball", che permette di interpretare correttamente la frase, evidenziando come le partite ("games") che il soggetto parlante vedrà su NBC e ABC sono di baseball. Un'informazione di questo tipo diventa esplicita solo se la stessa ripresa anaforica risulta tale.

3.2. Prospettiva storica

La prima definizione di *associative/bridging anaphora* risale a Hawkins (1978) e indica i casi di "*definite descriptions whose referent is uniquely identifiable based on general knowledge about associations with entities evoked by antecedents*". Hawkins introduce anche i termini *trigger* e *associate* per indicare rispettivamente l'antecedente e la sua *associated definite description*. Per esempio, se considero la coppia di frasi:

*Bill found himself in the middle of a forest. The **trees** were tall and sturdy,*

la parola "forest" sarà il t*rigger*, mentre "trees" sarà l'*associate*. Per quanto riguarda il tipo di relazione, nel suddetto esempio la *bridging anaphora* copre un caso di meronimia (in un rapporto tra due parole una designa una parte e una il tutto; *trees* e *forest*), ma le relazioni coinvolte dal *bridging* sono molteplici, compresi i casi complessi come per esempio "Auschwitz" e "victims".

Nel corso degli anni la definizione di *bridging* si è a poco a poco evoluta, soprattuto in merito ai tipi di relazioni che il *bridging* dovrebbe coprire e ai tipi di espressioni linguistiche che possono essere utilizzate come *bridging anaphors* (Kobayashi at al.; 2020).

Per quanto riguarda "*che tipi di relazioni dovrebbero essere coperte dal bridging?"* i primi studi relativi al fenomeno sono stati fatti considerandolo da un punto di vista prettamente linguistico, in particolare con i lavori di Clark (1975), Prince (1981) e

Gundel (1993). Clark (1975) ha inaugurato questa area di ricerca e ha introdotto un ampio concetto di *bridging* caratterizzato dalla coreferenza (le riprese anaforiche si riferiscono sempre alla stessa entità). Una particolare attenzione verso la coreferenza si trova poi negli studi relativi ai casi in cui due coreferenti non condividono la stessa testa come *bridging* (Poesio e Vieira, 1998; Vieira e Poesio, 2000; Bunescu, 2003). La maggior parte degli studi più recenti si è invece concentrata su casi di *bridging non-identity*, molto più vicini alla definizione di anafora associativa di Hawkins (1978). In questo caso il *bridging* copre vari tipi di relazioni semantiche, distaccandosi ulteriormente da quei casi di studio che si limitano a considerare relazioni predefinite come quella di sottoinsieme, quella di appartenenza e quella di possessione (Poesio and Vieira, 1998; Poesio et al., 2004b). Gli studi più recenti tendono a sottolineare come il fenomeno del *bridging* non possa essere semplicemente descritto da un numero limitato di relazioni predefinite (Markert et al., 2012; R¨osiger, 2018a).

Per quanto riguarda "*quali tipi di espressioni linguistiche possono essere utilizzate come bridging anaphors*?" gli studi tradizionali (Hawkins, 1978; Poesio e Vieira, 1998; Lassalle e Denis, 2011; Rösiger, 2016) si limitano a considerare le espressioni definite (escludendo invece quelle indefinite) poiché esse permettono di introdurre nuove informazioni che possono essere tranquillamente interpretabili al di fuori del contesto generato dal discorso. Solo con il lavoro di Löbner del 1998 viene affermato che le *bridging anaphors* possono essere anche indefinite, poiché tali espressioni, all'interno di un discorso, possono avere delle relazioni semantiche con le espressioni precedenti (Poesio e Artstein, 2008; Markert et al., 2012; Rösiger, 2018a). Generalmente, i casi indefiniti di *bridging* sono relazione di tipo *part-of* o *part-of-event* (Rösiger, 2018a). Consideriamo il seguente esempio:

> *The Soviets announced that their last soldier would leave Afghanistan in February.* ***Millions of refugees*** *would have rushed home.*

In questo caso esiste una relazione semantica tra i termini “Afghanistan” e “Millions of refugees”, poiché i rifugiati di cui si parla nella seconda frase si trovano in Afghanistan, luogo geografico che è stato menzionato nella frase precedente.

3.3. Corpora

I corpora utilizzati per indagini relative al *bridging* sono diversi, seppur ne vengano utilizzati in particolare quattro, tutti in lingua inglese:

- ISNotes (composto da 50 articoli del Wall Street Journal in OntoNotes) (Markert et al.; 2012)
- BASHI (Bridging Anaphors Hand-annotated Inventory, composto da altri 50 articoli del Wall Street Journal in OntoNotes) (Rösiger, 2018a)
- ARRAU (composto da articoli appartenenti a 4 domini, RST, GNOME, PEAR e TRAINS) (Poesio e Artstein, 2008; Uryupina et al.; 2020)
- SciCorp (Scientific Corpus, composta da articoli scientifici di linguistica computazionale e genetica) (Rösiger, 2016).

Nella tabella 2 i suddetti corpora vengono comparati secondo cinque dimensioni, corrispondenti al tipo di dominio, alla dimensione (in termini di numero di documenti, token e menzioni), al numero di *bridging anaphors*, al tipo di *anaphors*, e al tipo di antecedenti.

Corpora	Domain Type	Size			Number of anaphors	Anaphor type	Antecedent type
		Docs	Tokens	Mentions			
ISNotes	WSJ news	50	40292	11272	663	All NPs	entity, event
BASHI	WSJ news	50	57709	18561	459	All NPs	entity, event
ARRAU RST	news	413	228901	72013	3777	All NPs	entity
ARRAU GNOME	medical, art history	5	21458	6562	692	All NPs	entity
ARRAU PEAR	spoken narratives	20	14059	4008	333	All NPs	entity

ARRAU TARINS	dialogues	114	83654	16999	710	All NPs	entity
SciCorp	scientific text	12	61045	9407	1366	Definite NPs	entity

Tabella 2: Confronto dei corpora in lingua inglese più utilizzati per studi di bridging.

Esistono comunque diversi *bridging corpora* non in lingua inglese. DIRNDL (Björkelund et al., 2014) e GRAIN (Schweitzer et al., 2018) in tedesco, DEDE (Gardent e Manuélian, 2005) e PAROLE in francese (Gardent et al., 2003), Caselli/ Prodanof (Caselli e Prodanof, 2006) e Italian Live Memories Corpus (Rodrìguez et al., 2010) in italiano, e molti altri che vanno a coprire lingue come lo spagnolo, il giapponese e il russo.

Ci sono inoltre i cosiddetti *parallel bridging corpora*, ovvero dei corpus che contengono testi in lingue diverse. Tra questi troviamo il Copenhagen Dependency Treebank (Korzen e Buch-kromann, 2011), un corpus parallelo che coinvolge testi in lingua tedesca, inglese e russa.

Infine, seppur non sia particolarmente utilizzato, segnaliamo il GUM corpus in lingua inglese annotato con *bridging links* dagli studenti della Georgetown University (Zeldes, 2017). Quest'ultimo, insieme al corpus BASHI, è stato utilizzato per costruire il dataset (vedi capitolo 3) descritto nella presente relazione.

3.4. Bridging Resolution

La *bridging resolution* è un'*anaphora resolution task* che consiste nell'identificare e risolvere *bridging/associative anaphors*, che sono riferimenti anaforici ad antecedenti non identici ad essi associati. Rispetto all'*entity co-reference resolution*, il task che consiste nel determinare quali entità menzionate in un testo si riferiscono alla stessa entità del mondo reale, la *bridging resolution* presenta delle difficoltà maggiori. Innanzitutto l'*entity co-reference resolution* è avvantaggiato dalla presenza di vincoli di natura grammaticale (genere e numero), sintattica (*bridging theory*), semantica (*semantic class agreement*), e legati al livello del discorso

(iniziale, centrale, finale). Tali possibilità non sono invece presenti nel *bridging resolution* e, seppur alcune volte l'antecedente possa essere identificato comparando la similarità lessicale con l'anafora, nella maggior parte dei casi non ci sono indizi sintattici o di superficie che permettano di identificare l'antecedente di un'anafora. Generalmente la risoluzione di task di questo tipo richiede sia l'utilizzo del contesto sia delle cosiddette *commonsense inferences*. Quest'ultime sono legate al concetto di *commonsense knowledge*, ovvero quell'insieme di conoscenze a cui gli umani possono accedere mentre processano un testo; questo aiuta loro a produrre delle inferenze (*commonsense inferences*) in merito a delle informazioni che non sono menzionate nel testo, ma che si presume siano note a tutti (Ostermann et al.; 2019). Per esempio, consideriamo la seguente conversazione:

Max: "It's 1 pm already, think we should get lunch."
Dustin:"Let me get my wallet."

Max non sarà sorpreso dal fatto che Dustin debba prendere il portafoglio per andare a pranzo, poiché sa bene che *paying* è parte di *get lunch*. Inoltre Max sa anche che il portafoglio è necessario per pagare, e quindi Dustin dovrà avere con sé il proprio portafoglio per andare a pranzo. Queste associazioni legate a delle conoscenze comuni dovrebbero essere note sia a Max sia Dustin, come alla maggior parte delle persone. Il discorso cambia se invece di considerare un soggetto umano si prende in considerazione una macchina, per cui, inferire dei fatti non menzionati come quello visto nell'esempio diventa una sfida tutt'altro che banale (Ostermann et al.; 2019).
Infine, mentre gli antecedenti nell'*entity co-reference* sono *noun phrases* (NPs), gli antecedenti nel *bridging* possono anche essere non-NPs come i *verb phrases* (VPs) o clausole, che incrementano in modo sensibile il numero di possibili candidati come antecedenti per ogni anafora.
Per risolvere un task di bridging resolution esistono diversi approcci che possono essere distinti in *rule-base approaches* e in *learning-based approaches*.

3.4.1.Rule-based Approaches

L'approccio a regole consiste nell'utilizzare un set di regole create e curate manualmente per risolvere il task, nel nostro caso la *bridging resolution*. I primi studi relativi sono stati compiuti da Vieira e Teufel (1997) che cercano di risolvere delle *bridging anaphors* utilizzando un'euristica basata su relazioni di sinonimia, di iponimia e meronimia ottenute da WordNet[3] 1.6. In seguito, Poesio at al. (1997) migliorano il sistema andando a limitare l'uso di alcune relazioni di WordNet e inserendo la strategia di ricerca dell'antecedente. Questo lavoro è stato ulteriormente migliorato (Poesio et al.; 2002) andando a completare la copertura di WordNet con un'altra risorsa lessicale relativa a relazioni di meronimia, acquisita interrogando il British National Corpus.

Dopo questa prima fase di *rule-based bridging system* è stata la volta dei sistemi più recenti, composti da regole che eseguono sia il task di riconoscimento sia quello di risoluzione allo stesso tempo. Per esempio, nel 2014 Hou et al. propongono un sistema di otto regole per ISNotes. Rösiger et al. (2018a) riesce ad applicare il sistema così prodotto al corpus BASHI aggiungendo solo una regola addizionale, scoprendo però che lo stesso sistema non può essere applicato ad ARRAU, costituito per lo più da *bridging* lessicali. Questi ultmi non sono presenti in ISNotes, corpus su cui sono state sviluppate le regole di Hou et al., che contiene invece solo *bridging* referenziali. Per questa ragione Rösiger et al. scelsero di mantenere solo tre regole di quelle sviluppate da Hou et al. (Relative alla cattura dei pattern comuni che apparivano sia in ISNotes sia in ARRAU) e di aggiungere otto regole create appositamente per ARRAU. Questa scelta sottolinea come uno degli svantaggi del *rule-based approach* sia che potrebbe essere necessario progettare nuove regole per ogni nuovo corpus annotato con uno schema diverso. Di conseguenza l'insieme di regole che vengono messe assieme in base al contenuto di un determinato corpus sarà

[3] WordNet è un database semantico-lessicale per la lingua inglese che si propone di organizzare, definire e descrivere i concetti espressi dai vocaboli.

valido soprattuto per quel corpus, mentre probabilmente non riuscirà a coprire tutti i casi di altri corpus selezionati, e quindi a generalizzare.

Le regole prodotte da Hou et al. (2014) e da Rösiger at al. (2018a) per la *bridging resolution* sul corpus ISNotes sono composte da due condizioni: una per l'anafora e una per l'antecedente. Se le due menzioni soddisfano queste condizioni, la regola stabilisce un *bridging link* tra le due. Nella tabella 3 riportiamo una regola per meglio comprendere il funzionamento di tale approccio.

Rule	**Condition on anaphor**	**Condition on antecedent**	**Motivation**	**Recognition**		**Resolution**	
				P(%)	**R(%)**	**P(%)**	**R(%)**
Set: Percentage	Percentage NPs in subject position	Closest NP modifying another percentage NP via "of"	Percentage expressions can indicate bridging	I: 100 B: 0.0 A: 100	I: 0.8 B: 0.0 A: 0.2	I: 100 B: 0.0 A: 100	I: 0.8 B: 0.0 A: 0.2

Tabella 3: regola per la percentuale di Hou et al. (2014) e di Rösiger et al. (2018a) per la bridging resolution.

Ogni regola è espressa in termini di nome, condizione per l'anafora, condizione per l'antecedente, motivazione dietro la sua implementazione, e i valori percentuali di *precision* (P(%)) e *recall* (R(%)) (vedi cap. 4) calcolati rispettivamente sui corpora ISNotes (I), BASHI(B) E ARRAU RST (A). Un esempio della applicazione della regola in tabella 3 potrebbe essere "22% of the firms" (22% delle imprese), che stabilisce un *bridging link* tra "22%", che rispetta la regola dell'anafora in quanto si trova nella posizione di soggetto, e "the firms", che è l'NP più vicino che modifica il *percentage* NP tramite la preposizione "of" e che quindi rispetta la condizione dell'antecedente.

3.4.2.Learning-based Approaches

I learning-based approaches possono essere suddivisi in tre categorie: *feature-based approaches*, *embedding approaches* e *neural models*.

Feature-based approaches. In questo caso si utilizza un *pairwise classifier* (classificatore a coppie), noto come *mention-pair model* nella *coreference resolution literature* (Soon at al., 2001; Ng and Cardie, 2002), che viene addestrato per stabilire quando due menzioni presentano una relazione di *bridging* tra loro. Per l'addestramento si utilizzano una serie di coppie costituite da due menzioni, una corrispondente alla *bridging anaphora* e l'altra al possibile antecedente. Ciascuna coppia viene classificata con l'etichetta "POSITIVE", nel caso in cui l'antecedente sia corretto, o in caso contrario con quella "NEGATIVE". Per stabilire l'etichetta da associare a ciascuna coppia si utilizzano le cosiddette feature, che sono portatrici di un'informazione lessicale, sintattica o semantica, oltre ai casi speciali di *salience features* e *sibling anaphor features*. Nella tabella 4 riportiamo alcune delle feature utilizzate per addestrare il *mention-pair model*.

Feature	Description	Paper
Lexical Features		
Head match	se m_i e m_j hanno la stessa testa	Hou et al. (2013b)
Compound premodification	se m_j è un nome composto la testa di m_i sta pre modificando m_j	Hou et al. (2013b)
Syntactic features		
Co-argument	se m_i e m_j sono rispettivamente soggetto e complemento oggetto dello stesso verbo	Hou et al. (2013b)
Parallel structure	se m_i has the lo stesso ruolo sintattico e è nella stessa frase (ma non nella stessa clausola) di m_j	Hou et al. (2013b)
Semantic features		
WordNet query	se m_i e m_j hanno una relazione "part-of" in WordNet	Hou et al. (2013b)
Google distance	numero di risultati della query "the X of the Y" restituita da Google, dove X è la testa di m_j e Y è la testa di m_i	Poesio et al. (2004a)
Salience features		
Utterance distance	la distanza tra l'enunciato contenente m_j e l'enunciato contenente m_i	Poesio et al. (2004a)

Sibling anaphor features		
Similar anaphors	se le due anafore sono collegate tramite congiunzione, hanno la stessa testa, hanno un alto punteggio di somiglianza o sono sintatticamente parallele.	Hou et al. (2013b)

Tabella 4: Features per la bridging resolution dove m_j è una bridging anaphora, e m_i è un possibile antecedente di m_j.

Embedding approaches. L'utilizzo degli embedding per la *bridging resolution* nasce da un'osservazione di Hou (2018b) secondo cui le più comunemente usate *word representations*, come per esempio GloVe (Pennington et al. 2014), catturano somiglianze e relazioni "genuine", ma nel caso del *bridging resolution* è richiesta la conoscenza dell'associazione lessicale piuttosto che informazioni di similarità semantica tra sinonimi o iperonimi. Questo ha fatto sì che venissero addestrati degli embedding specifici per il task di *bridging resolution*. Hou (2018b) osserva che la struttura preposizionale (es., *X di Y*) e quella possessiva (es., *Y's X*) di NPs codificano una varietà di relazioni *bridging* tra le anafore e gli antecedenti. Per esempio, consideriamo le seguenti frasi:

> *The window of the room* implica una relazione "*part-of*" tra *the window* e *the room*.
>
> *Japan's prime minister* presenta una relazione di *bridging* tra *Japan* e *prime minister.*

In seguito Hou estrae del *parsed* Gigaword corpus le coppie di nomi coinvolte nelle suddette relazioni sintattiche e le usa come segnali di supervisione per addestrare un *embedding model* chiamato *embeddings_PP*. Questo modello può essere utilizzato per selezionare un antecedente per una *bridging anaphora* calcolando la similarità tra i vettori rispettivamente della *head* della *anaphora* e della *head* del possibile antecedente. In seguito tale modello è stato combinato con il GloVe embeddings così da poter coprire molti più casi (Hou, 2018a).

Neural models. Il primo modello neurale completamente dedicato alla *bridging resolution* è stato proposto da Yu e Poesio (2020) sfruttando un modello neurale *span-based* originariamente sviluppato per l'*entity coreference resolution* da Kantor e Globerson (2019). Questo modello è un *mention-ranking model* (Denis e Baldrige, 2008) e cioè un modello addestrato per associare un rank ad ogni possibile antecedente per un'anafora, così che l'antecedente corretto abbia il rank più alto. Yu e Poesio forniscono innanzitutto al modello delle *gold mentions* (annotate a mano) come input, e secondariamente propongono di addestrare il modello per eseguire *coreference* e *bridging* in un *multi-task learning* (MTL) framework. In quest'ultimo, lo *span representation layer* è condiviso dai due task, così che l'informazione appressa da uno possa essere utilizzata durante l'apprendimento del secondo. Inoltre, questo modello neurale utilizza solo due feature (contrariamente a quanto accade nel *feature-based approach*) che sono la lunghezza di una menzione e la distanza tra gli elementi di una coppia di menzioni.

Recentemente, Hou (2020) ha proposto un approccio neurale al *bridging resolution* basato su un task di *question answering* (QA). Data un gold anafora:

1. Si crea una domanda dall'anafora nella forma "anaphor of what?";
2. a partire dagli antecedenti se ne sceglie uno come possibile risposta;
3. si usa BERT-based QA system pre-addestrato sul corpus SQuAD (Joshi at al., 2020) per scegliere la risposta più probabile (i.e., l'antecedente).

L'aspetto più interessante di questo approccio è che esso non richiede alcun tipo di *gold mention* o di *system mention* come invece accade nel caso di Yu e Poesio. Hou successivamente ha ipotizzato che i risultati potrebbero migliorare se il modello fosse pre-addestrato su un *bridging* corpus piuttosto che su un QA corpus. Come già detto in precedenza tutti i *bridging corpora* esistenti sono però troppo piccoli per addestrare un modello neurale. Per superare questo limite, Hou ha impiegato un *distant supervision method* per generare un *automatically labelled bridging corpus*, e dimostrare che un modello pre-addestrato su quest'ultimo offra performance migliori di quello pre-addestrato su SQuAD.

Nella tabelle sottostante riportiamo i migliori risultati raggiunti nella *bridging resolution* e nella f*ull bridging resolution* con diversi approcci e utilizzando i tre dataset più comuni (ISNotes, BASHI, e ARRAU RST). La differenza tra i due tipi di task sta nel fatto che, nella *full bridging resolution* a un sistema non venga dato in input solo il documento ma anche le *gold mentions* nel documento. L'obiettivo è quindi quello di identificare il sottogruppo di *gold mentions* che sono *bridging anaphhors* e che risolvono nei loro antecedenti, anche questi scelti tra le *gold mentions.*

System	**Approach**	**Gold coref?**	**Dataset**		
			ISNotes	**BASHI**	**ARRAU**
Hou et al. (2018)	Feature based	Feature extraction	50.7	-	-
Hou (2018a)	Embedding	No	39.5	27.4	32.4
Yu and Poesio (2020)	MTL	No	40.7	34	49.3
Hou (2020)	QA	No	50.1	38.7	-

Tabella 5: Accuratezza dei bridging resolvers.

System	**Approach**	**Gold coref?**	**Dataset**					
			ISNotes		**BASHI**		**ARRAU**	
			Rec	**Res**	**Rec**	**Res**	**Rec**	**Res**
Rösiger et al. (2018b)	Rule based	Anaphor filtering	29.3	20.4	28.7	14.1	30.8	19.5
Hou et al. (2018)	Feature based	Feature extraction	46.1	21.6	-	-	-	-
Yu and Poesio (2020)	MTL	Anaphor filtering	43.6	23.2	27.2	14.4	36.7	24.0

Tabella 6: Recognition (Rec) e resolution (Res) F-scores dei full bridging resolvers.

Seppur, come dimostra la tabella 5, ci siano dei buoni progressi per quanto riguarda la *bridging resolution*, considerando che il miglior valore di Accuracy si aggira intorno al 50%, è anche vero che entrambi i casi di studio siano lontani dall'essere risolti.

Nelle presente relazione abbiamo voluto considerare il fenomeno della *bridging anaphora* da un altro punto di vista. Come abbiamo visto il caso di studio più diffuso in relazione al fenomeno è senza dubbio la *bridging resolution*, e per questa ragione abbiamo scelto di presentarlo in modo dettagliato, anche per far comprendere al meglio la natura del fenomeno. Ciò che però abbiamo cercato di fare è stato inserire la *bridging anaphora* nel contesto del *Natural Language Inference* (NLI) e in particolare nella task di riconoscimento del *textual entailment* (TE). Per farlo abbiamo costruito un dataset che sarà presentato nel prossimo capitolo, per poi arrivare alla fase sperimentale tramite l'utilizzo di modelli neurali (vedi cap. 4).

4. Il Dataset

In questo capitolo andremo ad illustrare il dataset utilizzato per gli studi sperimentali contenuti nella presente tesi (sezione 4.1). In particolare, il dataset è composto da circa trecento coppie di frasi, ciascuna etichettata rispettivamente come "Premise" o come "Hypothesis" (sezioni 4.2 e 4.3), così da poter essere utilizzata in un task di *Recognizing Textual Entailment* (RTE). Per ogni coppia troviamo poi una *label* che stabilisce la presenza o meno di inferenza logica, e che quindi assumerà un valore fra "entailment", "contradiction" e "neutral". Infine, il campo "type" definisce la relazione tra le due menzioni legate dal *bridging link* (sezione 4.4). La costruzione del dataset è profondamente legata ai task di NLI ed esso è stato costruito per verificare l'apprendimento del fenomeno linguistico della *bridging anaphora* da parte degli attuali modelli neurali.

Forniremo poi una serie di informazioni di natura statistica legate alla distribuzione dei dati (sezione 4.5).

4.1. Composizione e struttura

Per costruire il dataset abbiamo preso come modello la struttura del dataset di diagnostica del *benchmark SuperGlue*[4].

Tutti i dati sono stati inseriti in un framework tabellare con le seguenti colonne:

Premise: premessa per il task di *recognizing textual entailment* (TE), in questo campo si trova una frase (o una coppia di frasi) corrispondente a un caso di *bridging anaphora*, ovvero presenta due menzioni che sono legate da un *bridging link*;

Hypothesis: ipotesi per il task di RTE;

Label: contiene il tag che definisce la relazione tra l'ipotesi e la premessa (*entailment*, *contradiction* o *neutral*)

Type: questa colonna stabilisce il tipo di *bridging link* tra le due menzioni presenti nella premessa. Per esempio, se consideriamo la frase "Nonna ha fatto una torta, aveva delle mele da utilizzare" il *bridging link* sarà tra "torta" e "mele" e corrisponderà ad una relazione di possesso, che sarà quindi indicata con il tag "Have".

In tabella 7 è possibile osservare un paio di casi estratti dal dataset. In particolare, se consideriamo il primo caso, la premessa presenta una relazione di possesso tra le parole "ringtone" e "cell phone", mentre l'ipotesi va a contraddire la premessa, generando quindi un caso di *contradiction*.

Premise	Hypotheis	Label	Type
Anna heard the ringtone and she took the cell phone from her bag.	Anna had left the phone at home	Contradiction	Have
Mom digs out the inhaler and Cara takes a hit.	Cara needs to take a hit at the inhaler.	Entailment	Use

Tabella 7: Esempi dal dataset.

[4] https://super.gluebenchmark.com/diagnostics

Nel secondo caso troviamo invece una relazione d'uso tra le parole "inhaler" e "hit" per quanto riguarda la premessa, e un'implicazione logica (*entailment*) dell'ipotesi.

4.2. Premise e Bridging Anaphora

Come abbiamo visto nel capitolo precedente, per il presente elaborato abbiamo scelto di considerare il fenomeno della *bridging anaphora* all'interno della premessa (colonna "premise"). In particolare, abbiamo preferito considerare una frase o al più una coppia di frasi contenenti una relazione di *bridging* per poi andare a valutare la possibile implicazione logica con una nuova frase. In questo modo si è cercato di valutare se risultasse possibile per un modello neurale riconoscere il *bridging link* nella premessa così da poter etichettare correttamente la relazione creatasi con l'ipotesi.

Per quanto riguarda le frasi da inserire nel campo "premise" abbiamo scelto di utilizzare sia dei casi di *bridging* di tipo gold sia degli esempi di *bridging anaphora* nuovi e creati manualmente per l'indagine descritta nella presente relazione. In particolare, per i *gold bridging link* abbiamo utilizzato il corpus BASHI (Rösiger, 2018) e il GUM corpus (Zeldes, 2017), già trattati nel capitolo 2 (vedi par. 2.4). Il BASHI corpus è un corpus in lingua inglese contenente 50 articoli del Wall Street Journal, per un totale di 57,709 tokens, annotati con 459 *bridging links*. Quest'ultimi si dividono in 3 categorie: *definite bridging links*, ovvero casi di relazione ben definita tra un'anafora e il suo antecedente, *indefinite bridging links*, ovvero caratterizzati da un'espressione indefinita che introduce relazioni di *part-whole* o di *part-of-event*, e *comparative bridging links*, che indicano una relazione alla pari tra l'antecedente e l'anafora, priva quindi di una gerarchia (Rösiger, 2018).

I. ***Definite****:*

   I met a man yesterday. ***The bastard*** *stole all my money.*

II. ***Indefinite****: I bought a bicycle.*

   A tire *was already flat.*

III. ***Comparative:***

> *About 200,000 East Germans marched in Leipzig and thousands more staged protests in* ***three other cities****.*

Il GUM corpus (Zeldes, 2017) è un corpus di 22,656 tokens annotato con oltre 180,00 annotazioni di diversa natura. È suddiviso in 4 sottogruppi in base alla tipologia dei testi contenuti, che derivano a loro volta da diverse fonti (vedi tabella 8). Abbiamo scelto di utilizzare principalmente i casi di *bridging anaphora* appartenenti al campo "narrative", poiché caratterizzati da *bridging links* meno "ambigui" rispetto agli altri casi, primo fra tutti quello di tipo Interview (conversational).

Text type	Source	Documents	Tokens
News (narrative)	Wikinews	6	5051
Interview (conversational)	Wikinews	7	6535
Hot-wo (instructional)	wikiHow	7	6701
Travel guide (informative)	Wikivoyage	5	4369
Total		25	22656

Tabella 8: documenti nel GUM corpus

Oltre ai *gold bridging links* estratti dai corpus sopra descritti, abbiamo costruito delle premesse in modo manuale, cercando di utilizzare delle frasi che fossero il più semplici possibile. Si è cercato quindi di fornire al modello un contesto nitido e privo di ambiguità, che sapesse bilanciare sia i testi del Wall Street Journal, che sono stati comunque sottoposti ad una revisione manuale al fine di renderli più accessibili, sia quelli del GUM corpus che, vista la loro natura principalmente narrativa, hanno richiesto anch'essi un *restyling*. Per meglio comprendere le operazioni di revisione che sono state eseguite, riportiamo di seguito un esempio di un caso di *bridging* estratto dal BASHI corpus (I) e uno estratto dal GUM corpus (II):

I. *Concerns about the pace of the Vienna talks also are being registered at the Pentagon. Mr Bush has called for* ***an agreement by next September at the latest****.*

   è diventata la seguente premessa:

   *The Vienna talks are underway. Bush has called for a****n agreement by next September****.*

II. *The inside of the church smelled like damp wood and furniture polish, not alive at all. My father took off his object coat and draped it over the edge of* ***the pew*** *and when I came back from communion I stole his glove.*

   è diventata:

   The church smelled like damp wood. Phil's father put his coat on the edge of ***the pew****.*

Oltre a delle semplificazioni a livello sintattico e lessicale, abbiamo scelto (come è possibile osservare nel caso (II) sopra descritto) di evitare i casi di pronomi personali per indicare dei soggetti generici, preferendo invece inserire dei nomi propri così da migliorare ulteriormente la comprensione delle frasi da parte dei modelli.
Le applicazioni sopra descritte ci hanno portato ad ottenere circa un centinaio di premesse, da cui poi sono state costruite più ipotesi per ciascuna di esse come andremo a descrivere nel prossimo paragrafo.

4.3. Hypothesis

Il campo "Hypothesis" contiene le frasi che possono rappresentare casi di implicazione logica rispetto al contenuto del campo "Premises", e rappresenta il fulcro del task di riconoscimento di *Textual Entailment*.
Nell'indagine proposta nel presente elaborato abbiamo cercato di realizzare per ogni premessa almeno un'ipotesi che creasse una relazione di *entailment*, una che creasse una *contradiction* e una che generasse una caso di neutralità (etichettato come

"neutral"). Tutte le ipotesi presenti nel dataset sono state quindi costruite in modo manuale, tenendo conto del tipo di relazione definita dalle tre label caratteristiche di TE e cercando di produrre dei casi attinenti. Nello specifico, per ciascun tipo di relazione, abbiamo evitato di utilizzare delle scelte sintattiche, lessicali e semantiche che impedissero un riconoscimento del fenomeno. Per esempio, nei casi di contraddizione ("contradiction") abbiamo preferito evitare la negazione, utilizzando invece una sostituzione lessicale in aggiunta alla conservazione della struttura affermativa (I). Come abbiamo visto nell'Introduzione, la presenza della negazione in numerosi casi etichettati come "contradiction" porterebbe il modello ad associare tale relazione solo ai casi che presentino la struttura "VB + not", perdendosi invece tutti quei casi in cui la contraddizione si crea comunque tramite un'affermazione. Inoltre, per i casi di *entailment* (II) abbiamo cercato di mantenere le stesse componenti lessicali presenti nella premessa, evitando l'uso di sinonimi o di forme dal significato simile, preferendo che il modello fosse in grado di riconoscere le stesse menzioni già individuate precedentemente. Infine, per i casi di neutralità ("*neutral*") abbiamo mantenuto una parte delle forme lessicali presenti nella premessa, così da creare una relazione, ma ne abbiamo inserite di nuove, generando quindi un maggior livello di sorpresa e creando quindi una frase che non andasse né a confermare né a negare la premessa (III).

Premise: *This is a picture of the kitchen in the house Mary used to live in. The coffe table was in* ***the living room***.

I. **Hypothesis - contradiction:**

In Mary's house there was a coffee table in the kitchen.

La menzione "the living room" viene sostituita con "the kitchen", generando così una contraddizione rispetto a quanto affermato nella premessa.

II. **Hypothesis - entailment**:

In Mary's house there was a coffee table in the living room.

Vengono utilizzate le menzioni "Mary's House" e "the living room", entrambe presenti nella premessa, che permettono di inferire che la *living room* di cui si parla sia quella della casa di Mary.

III. **Hypothesis - neutral:**

In Mary's house there were two bathrooms.

Affermare che nella casa di Mary ci siano due bagni non crea alcun tipo di conflitto con la premessa, ma non permette nemmeno di confermarla.

Come nel caso delle premesse, abbiamo preferito sostituire i pronomi personali con nomi propri, evitando i soggetti "generici" e dotando ciascuna persona presente nella frase di un identificativo non ambiguo, ovvero un nome proprio. Tale scelta nasce naturalmente dall'idea di voler conservare le menzioni presenti nella premessa anche nell'ipotesi.

4.4. Type

La colonna "Type" contiene i valori corrispondenti al tipo di *bridging link* presente tra l'anafora e l'antecedente contenuti nella premessa. In particolare, definisce il tipo di relazione semantica tra le due menzioni. Ci sono due tipi di relazioni semantiche, quelle persistenti e quelle contestuali. Le prime sono quelle che fanno parte della conoscenza lessicale e che sono valide nel contesto di un particolare discorso così come in tutti gli altri discorsi. Le relazioni contestuali sono invece transitorie, e possono essere valide solo nell'ambito di un singolo discorso, come per esempio "a cup on a table" o "John has a knife" (Nand e Yeap, 2013). Per risolvere una *bridging anaphora* è necessario identificare sia le relazioni persistenti sia quelle contestuali.

Abbiamo scelto di utilizzare gli undici tipi di relazione proposti da Nand e Yeap (2013), che sono gli stessi utilizzati per descrivere le relazioni tra un modificatore e il

nome *head* di un nome composto. Nand e Yeap (2013) sono quindi partiti dall'intuizione secondo cui l'uso della *bridging anaphora* rappresenti una sorta di scorciatoia molto simile a quella relativa all'uso dei nomi composti, osservando che la maggioranza (circa il 98%) delle relazioni di *bridging* possa essere descritta dai suddetti tipi di relazioni. Abbiamo quindi scelto di considerare anche noi tali tipologie proprio per avere una copertura pressoché totale di tutti i casi considerati. Di seguito descriviamo nel dettaglio ciascuno degli undici casi proposti.

CAUSE: le due menzioni si trovano in una relazione di causalità l'una rispetto all'altra (terremoto/tsnunami).

The battle lasted three days. The fatigue was unbearable.

HAVE: include le nozioni di possesso, che indica generalmente casi in cui un'entità è parte di un'altra, come serpente/veleno oppure torta/mele.

Jack recently bought a house in the mountains. In the living room there is a large fireplace.

MAKE: un'entità è necessaria per la realizzazione fisica dell'altra, come per esempio l'asfalto per fare una strada.

Simon really likes the basil. Simon wants to order some seeds.

USE: una delle due entità viene utilizzata per far funzionare la seconda (nave/ vapore)

The washing machine does not work. The power must have gone out.

IN: questa relazione cattura gruppi di entità che condividono proprietà fisiche e temporali, e possono essere sia oggetti come il caso di lampada/tavolo sia luoghi geografici come Toscana/Italia

Carroll met Rachel in the College library while she sat at the checkout desk.

FOR: un'entità è lo scopo di un'altra, ovvero la prima viene utilizzata per eseguire la seconda. Questa relazione lega quindi un nome ad un verbo come i casi di penna/scrivere e di palla/giocare

A TV network is airing a celebrity interview. There are a lot of people behind the cameras.

FROM: in questo caso un'entità viene derivata dall'altra (grano/farina)

The olive harvest was abundant. This year there will be a lot of oil.

ABOUT: descrive il caso in cui un'entità è un argomento dell'altra, come per esempio viaggio/storia

Physics was Tom's favourite subject. Tom liked relativity very much.

Le restanti tre relazioni richiedono di essere trattate in modo più approfondito per essere comprese al meglio. In particolare le prime due derivano da una relazione già presente in Levi (1978), mentre la terza è stata introdotta nel lavoro di Nand e Yeap (2013), che ha comunque ripreso da Levi le otto relazioni sopra descritte.
Per indicare l'esistenza di plurali in forma diversa Levi utilizza la relazione BE. Nand e Yeap (2013) scelgono di dividere tale relazione in BE-OCCR e BE-INST per distinguere i casi di *direct co-reference* o di *identity relation* dai casi di *instance relation*. Nella prima relazione (BE-OCCR) un NP forma una coreferenza uno ad uno con un altro NP, come per esempio *John*/*he* e *John*/*the driver*. La relazione BE-INST rappresenta invece i casi dove un'anafora si riferisce ad un antecedente plurale, come per esempio il caso di *both trucks*/*northbound truck*. In questo caso *northbound truck* è un'istanza di *both trucks*, che è diverso rispetto a una relazione di coreferenza (Nand e Yeap, 2013).

BE-OCCR:

CBS hires featured actors. A documentary will be made with a cast of thousands.

BE-INST:

Tommy gets along well with many professors. John is the best.

Quando due o più entità di un discorso partecipano allo stesso evento o a un evento simile, possono essere indicate come un'unità da un NP collettivo nel contesto del discorso. Tali entità possono essere espresse tramite la congiunzione "and" (I) o descritte da due diverse clausole (II).

I. The coastguard and Lion Foundation Rescue helicopter were called out.

II. the truck rolled down the hill
the ball rolled down the hill

In particolare, il caso II presenta una relazione di tipo contestuale tra *truck* e *ball*, valida quindi solo nel contesto del discorso. Questa relazione viene chiamata ACTION e lega delle entità che partecipano allo stesso evento o a eventi simili (Nand e Yeap, 2013). Una possibile applicazione in un caso di *bridging anaphora* potrebbe essere:

The fox and Paul were running. The runners were thirsty.

la relazione ACTION viene utilizzata per descrivere il fatto che l'NP *runners* dell'ipotesi si riferisca sia a *fox* sia a *Paul* presenti nella premessa.

4.5. Distribuzione dei dati

Durante la fase di costruzione del dataset abbiamo cercato di mantenerlo il più possibile bilanciato per quanto riguarda la distribuzione delle label (*entailment*, *contradiction* e *neutral*). Volevamo evitare di ottenere una rappresentazione del fenomeno priva di equilibrio, che avrebbe portato anche a dei risultati peggiori e poco interessanti nella fase di sperimentazione. Per questa ragione, come abbiamo

già visto nei paragrafi precedenti, abbiamo scelto di costruire per ogni premessa almeno tre ipotesi, corrispondenti alle label caratteristiche del task di riconoscimento del *textual entailment* (TE). Per un totale di 96 premesse abbiamo costruito manualmente 300 ipotesi, che sono state poi etichettate secondo i tipi di relazioni descritti nel paragrafo precedente.

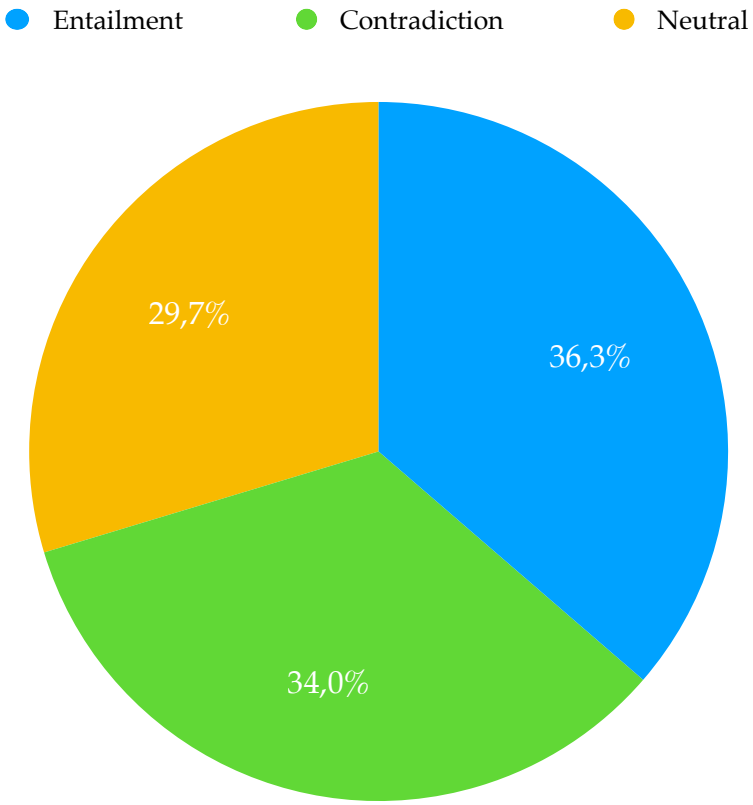

Figura 11: Textual Entailments Labels Distribution

Come possiamo osservare in figura 11 le tre classi sono ben bilanciate. Abbiamo il caso *entailment* che copre circa il 36% dei valori totali, un numero di contradiction pari al 34% e infine il restante 29,7 % è rappresentato dalla neutralità.

Non possiamo però dire la stessa cosa per quanto riguarda la colonna “Type”. Come abbiamo visto in precedenza, le *gold bridging anaphora* sono state estratte a partire dal Corpus BASHI (Rösiger, 2018) e dal GUM corpus (Zeldes, 2017). Nel primo caso parliamo di un raccolta di articoli del Wall Street Journal, mentre nel secondo abbiamo considerato principalmente i casi narrativi. Non sorprende quindi che la maggior parte delle relazioni dei *bridging link* così estratti fossero per lo più di

natura causale (CAUSE) e soprattuto possessiva (HAVE). In particolare, su 299 casi 51 sono etichettate come “CAUSE”, mentre 149 come “HAVE”, per un totale di 200 casi, e quindi più di 2/3 del dataset totale. Per questa ragione abbiamo cercato di presentare gli altri tipi di relazioni nei casi creati in modo manuale, cecando in questo modo di coprire tutto il repertorio di relazioni possibili (vedi figura 12).

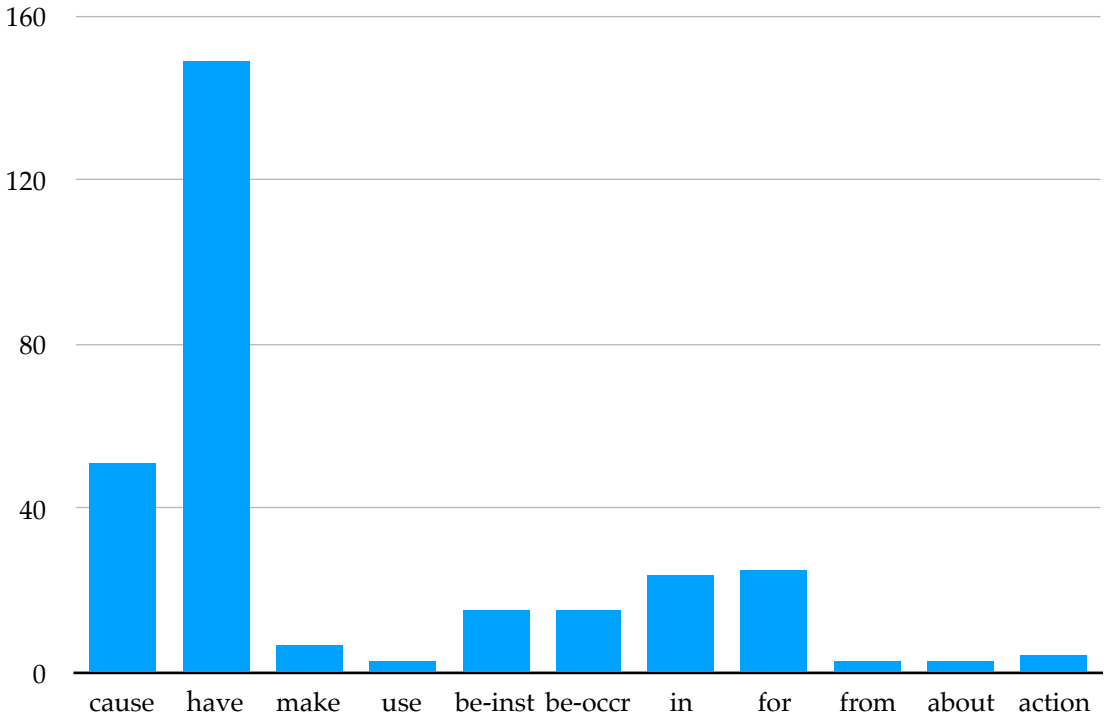

Figura 12: Relation Types Distribution

Anche in questo dataset certi casi presentano una frequenza maggiore degli altri, ma questo dipende anche dal fatto che alcuni risultino davvero poco comuni rispetto a casi di *bridging anaphora* che possano trovare corrispondenza nello scritto e nel parlato. In particolare i casi “make”, “use”, “from” e “about” sono piuttosto rari per quanto riguarda la creazione di *bridging link*. Se pensiamo infatti ad un oggetto e ciò di cui esso è realizzato, oppure a ciò che gli permette di funzionare o ancora alla sua origine a livello materiale, risulta difficile pensare che nel parlato come nello scritto si realizzino *bridging link* tra entità legate da simili relazioni. Per esempio, sarà molto comune inserire il nome “farina” in un discorso senza che si sia parlato di grano, oppure parlare di un apparecchio elettronico senza sottolineare che necessiti di

elettricità per funzionare, o per finire presentare un'entità come una penna o una scrivania senza sottolineare la presenza di inchiostro nella prima o di legno nella seconda. In merito all'etichetta "ABOUT" vale un discorso simile considerando che i casi in cui una menzione risulta essere il topic di un'altra sono poco frequenti. Tutto ciò dipende probabilmente dal fatto che tali etichette relative al tipo di relazione siano le stesse utilizzate da Levi in merito ai nomi composti nel 1978. Nel corso degli anni il linguaggio è infatti cambiato drasticamente e l'uso di certe relazioni è andato pian a piano a diminuire, facendo sì che circa l'80% dei bridging links gold estratti da corpus esistenti copra un paio delle categorie utilizzate. Ciò dimostra chiaramente quanto certe meccaniche linguistiche siano in disuso, ma allo stesso tempo, al fine di un'analisi quanto più ampia, abbiamo comunque scelto di considerare tutti e gli 11 casi presentati nei paragrafi precedenti e di fornire per ciascuno di essi quanto meno un'applicazione. Nelle tabella seguente riportiamo il numero di occorrenze per ciascuno tipo di relazione presente nel dataset.

Type	**Numero di occorrenze**
cause	51
have	149
make	7
use	3
be-inst	19
be-occr	12
in	24
for	25
from	3
about	3
action	4
Totale	300

Tabella 9: Types Distribution

5. Esperimenti

In questo capitolo vedremo i risultati degli esperimenti realizzati a partire dal dataset descritto nel capitolo precedente. Presenteremo i modelli neurali utilizzati (sezione 5.1) fornendo le nozioni relative alle metriche di accuratezza (sezione 5.2). In seguito vedremo i risultati dei singoli modelli, analizzandoli singolarmente nel dettaglio (sezioni 5.3, 5.4 e 5.5). Offriremo poi un confronto tra i modelli (sezione 5.6) per poi analizzare gli errori di quello con i risultati più soddisfacenti (sezione 5.7).

5.1. Modelli neurali

Nelle prima fase del lavoro sperimentale abbiamo testato le performance dei seguenti modelli, tutti pre-addestrati per task di NLI con il dataset MNLI:

- *facebook/bart-large-mnli*
- *roberta-large-mnli*
- *huggingface/distilbert-base-uncased-finetuned-mnli*

Per ogni modello abbiamo eseguito una *forward phase* utilizzando i casi presenti nel dataset creato. Abbiamo misurato i risultati nei termini di *Accuracy* e di *Score*:

	MACRO-PRECISION	**MACRO-RECALL**	**MACRO-F1-SCORE**	**ACCURACY**	**MNLI ACCURACY**
bart-large-mnli	0.87	0.87	0.87	0.87	0.90
roberta-large-mnli	0.84	0.84	0.84	0.84	0.90
distilbert-base-uncased.finetuned-mnli	0.77	0.77	0.77	0.77	0.82

Tabella 10: report delle performance dei modelli sull'intero dataset e su MNLI

Come possiamo osservare ci sono delle performance molto buone, come per esempio i casi di *roberta-large-mnli* e di *bert-large-mnli*. Il miglior modello risulta essere *bart-large-mnli* con valori di accuracy e di score pari a **0.87**, seguito da *roberta-large-mnli* con dei valori leggermente più bassi (**0.84** per lo score e **0.84** per l'accuracy), e infine il caso di *distilbert-base-uncased.finetuned-mnli* che offre comunque dei risultati interessanti in relazione al valore di *accuracy* misurato su MNLI (**0.77** contro **0.82**).

Nel grafico in figura 13 vengono mostrati i valori di *accuracy* per ciascun modello in forma percentuale, tenendo conto dei casi correttamente riconosciuti rispetto a quelli classificati erroneamente.

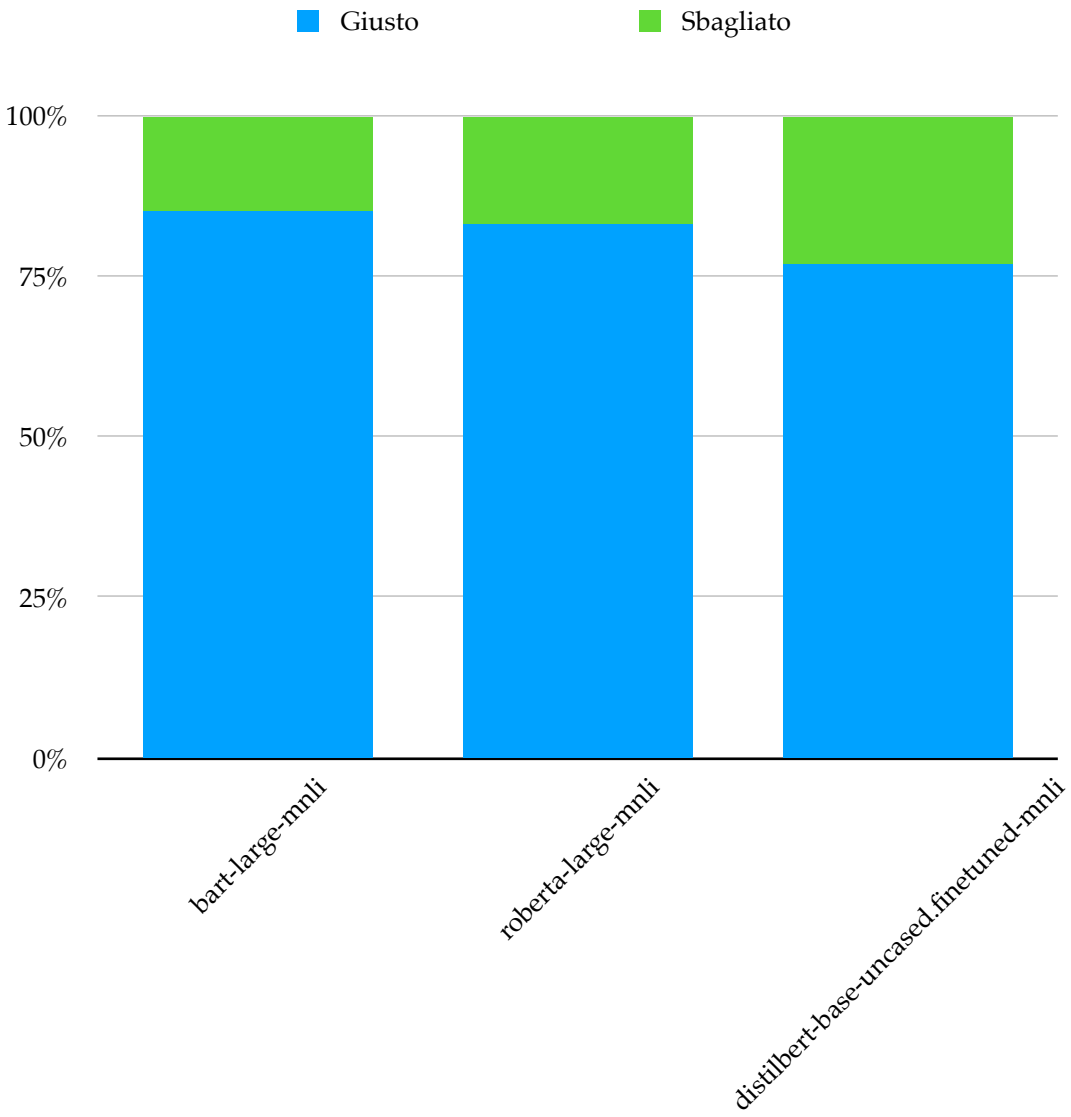

Figura 13: accuratezza dei modelli sull'intero dataset

In seguito andremo a considerare le performance di ciascuno dei modelli sopra riportati, così da valutare gli errori e soprattuto offrire un'analisi degli stessi, che consenta in seguito un confronto più preciso tra i vari approcci sperimentali.

5.2. Misure di accuratezza

Le performance dei modelli sono state misurate in termini di *precision*, *recall*, *accuracy* e *f1-score*. Per descrivere le suddette misure è necessario introdurre il

concetto di Matrice di Confusione. La matrice di confusione è una tabella utilizzata per descrivere le performance di un classificatore su un set di test data di cui sono conosciute le classi corrette, corrispondenti al cosiddetto gold standard (Jurafsky; 2018). In essa troviamo 4 parametri che ci permettono di derivare rispettivamente i valori di *precision*, *recall*, *accuracy* e *f1-score*.

	Classi Predette		
Classi Attuali		Classe = sì	Classe = no
	Classe = sì	True Positive	False Negative
	Classe = no	False Positive	True Negative

Tabella 11: matrice di confusione

True positive e *true negative* sono le osservazioni che sono state predette correttamente e sono mostrate in verde nella tabella 11. L'obiettivo per un buon classificatore è quello di andare a minimizzare i *false positive* e i *false negative*, ovvero i valori evidenziati in rosso nella tabella 11. Vediamo ciascun caso nel dettaglio:

- **True Positives (TP)**: questi sono tutti i valori positivi predetti correttamente, ovvero i casi in cui il valore della classe attuale e di quella predetta è in entrambi i casi uguale a "sì".
- **True Negatives (TN)**: questi sono i valori negativi predetti correttamente, ovvero i casi in cui il valore della classe attuale e di quella predetta è in entrambi i casi pari a "no".
- **False Positives (FP)**: in questo caso la classe attuale è uguale a "no" mentre la classe predetta è uguale "sì". Ovvero il modello classifica come "sì" un caso che invece dovrebbe essere classificato come "no"
- **False Negatives (FN)**: quando la classe attuale è uguale a "sì" mentre quella predetta è pari a "no". In questo caso il modello classifica come "sì" un caso che, per essere classificato correttamente, dovrebbe avere valore pari a "no".

Tenendo conto dei suddetti parametri è quindi possibile calcolare le misure di nostro interesse:

- **Accuracy**: la misura più intuitiva fra tutte, si limita a calcolare il rapporto tra le osservazioni predette correttamente e il totale delle osservazioni. Per questa ragione, tale misura risulta realmente affidabile solo nei casi in cui si abbia un dataset simmetrico, ovvero con un numero di *false positive* e di *false negative* piuttosto simile. In caso contrario è necessario ricorrere alle altre misurazioni.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision**: il rapporto tra i valori positivi predetti correttamente e il totale delle osservazioni positive.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**: il rapporto tra le osservazioni positive predette correttamente sul totale delle osservazioni nella classe attuale "sì".

$$Recall = \frac{TP}{TP + FN}$$

- **F1 score**: la media pesata tra *Precision* e *Recall*. Questa misura prende in esame sia i *false positive* sia i *false negative*, divenendo molto più utile rispetto alla semplice *Accuracy*, specialmente quando si ha una distribuzione di classe irregolare.

$$F1Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

5.3. Bart

BART è un modello pre-addestrato sulla lingua inglese ed è stato introdotto da Lewis et al. nel 2019. BART è un *transformer encoder-encoder* (*seq2seq*) con un encoder bidirezionale (simile al caso di BERT) e un decoder autoregressivo (simile a GPT (*Generative Pre-trained Transformer*)). Questo modello viene pre-addestrato andando a modificare il testo con una funzione di rumore arbitraria, che permette al modello di imparare a ricostruire il testo originale. BART è particolarmente efficiente nel *fine-tuned* per la generazione di testi (come per esempio i casi di traduzione), ma anche per studi relativi al *comprehension task* (*text classification*, *question and answering*) (Lewis et al.; 2019). In particolare a noi interessa proprio quest'ultimo caso, ovvero quello relativo al *text classification*. Per questo ragione la versione utilizzata nelle analisi sperimentali proposte nella presente relazione è quella denominata *bart-large-mnli*, ovvero il modello ottenuto dopo il *finetuning* sul MultiNLI (MNLI) dataset. Quest'ultimo è un corpus di 433k coppie di frasi annotate con informazioni relative al *textual entailment* (TE). Rispetto al corpus SNLI, che è stato utilizzato come modello per la creazione, MNLI copre una maggior gamma di generi testuali sia per quanto riguarda il parlato che per lo scritto, e supporta inoltre una valutazione distintiva della generalizzazione tra i generi (Williams at al.; 2017).

Abbiamo sottoposto a *bart-large-mnli* il dataset costruito ad hoc e descritto nel capitolo precedente. Nello specifico, abbiamo passato al modello la coppia premessa-ipotesi così che potesse assegnare un valore percentuale a ciascuna delle tre label caratteristiche del *textual entailment* (TE). Al valore maggiore è associata naturalmente la classe predetta dal modello, che deve poi essere confrontata con quella attuale così da valutare la performance del modello. Così facendo otterremo i valori corrispondenti alla matrice di confusione (vedi figura 14), che permetteranno a loro volta di misurare l'*accuracy*, la *precision*, la *recall* e l'*f1 score*.

Tutti i modelli presentati nella prima parte del seguente capitolo sono stati utilizzati allo stesso modo di quanto descritto in merito a *bart-large-mnli*.

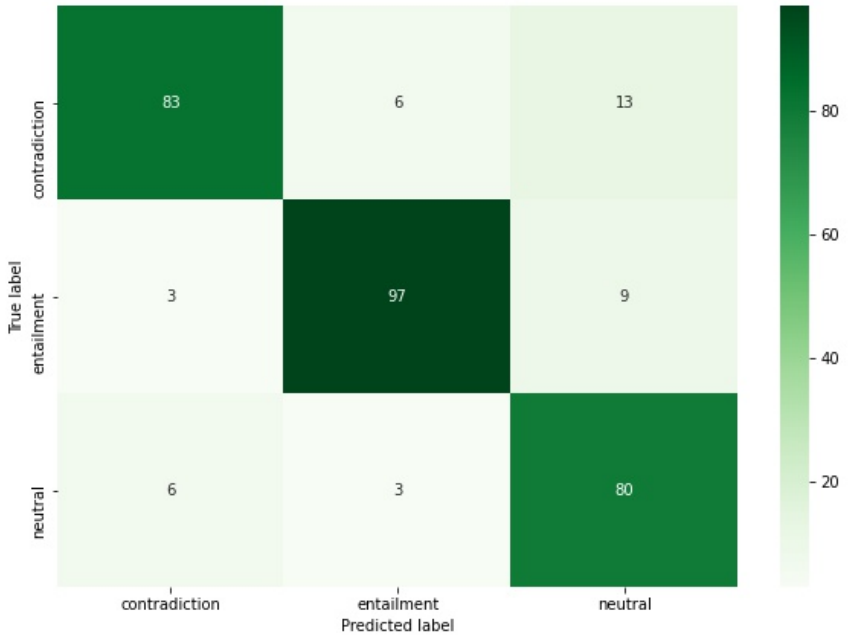

Figura 14: matrice di confusione sul dataset di bridging anaphora, testato su bart-large-mnli

Osservando la matrice di confusione in figura 14 è possibile riscontrare che *bart-large-mnli* abbia ottenuto dei risultati molto accurati. Per quanto riguarda il caso *entailment*, il modello ha classificato correttamente **97** casi su **109**, con 3 casi che sono stati interpretati come *contradiction* e i restanti 9 come *neutral*. Anche i casi rappresentanti una contraddizione sono stati riconosciuti in larga parte (**83** casi su **102**), se non fosse per 6 casi a cui è stato assegnato il valore di *entailment* e 13 casi corrispondenti alla label *neutral*. È interessante notare come la maggior parte degli errori, relativi ai casi di inferenza e di contraddizione, corrisponda alla scelta errata del caso *neutral*. Il modello sembra quindi distinguere chiaramente i casi di contraddizione da quelli di *entailment*. Questo aspetto sottolinea la bontà del modello e risulta inoltre una nota di merito per il dataset e la sua costruzione. Infine, per quanto riguarda il campo *neutral*, qua troviamo il numero minore di errori, con solo 9 casi interpretati in modo errato (6 *contradiction* e 3 *entailment*), seppur si parli comunque della label con la distribuzione minore fra tutti (**88** casi nel dataset totale, di cui **79** classificati correttamente da BART). Nella tabella 12 è possibile osservare i

valori corrispondenti alle metriche di valutazione relative alla performance di *bart-large-mnli*.

Bart-large-mnli	**Precision**	**Recall**	**F1-score**	**Accuracy**
Contradiction	0.90	0.81	0.88	
Entailment	**0.92**	0.89	**0.90**	
Neutral	0.78	**0.90**	0.84	
macro-average	0.87	0.87	0.87	
weighted-average	0.87	0.87	0.87	
total				**0.87**

Tabella 12: metriche di valutazione relative alla performance di bart-large-mnli

Come possiamo osservare le inferenze (*entailment*) presentano il valore di *precision* più alto, molto simile a quello delle *contradiction*, sottolineando come il numero di casi classificati come *entailment* siano per il 90% realmente *entailment*. Il valore di *recall* più alto spetta invece al caso *neutral* (0.90) dimostrando che la rete è riuscita riconoscere questi casi piuttosto che le inferenze o le contraddizioni. Se consideriamo infine la media armonica tra queste due metriche, ovvero l'*f1-score*, essa privilegia la performance eseguita sulle inferenze (0.90), dimostrando che il modello ha predetto in larga parte delle *entailment* reali riconoscendone inoltre la maggior parte.

Per quanto riguarda le medie, riportiamo la *macro-average*, che dà lo stesso peso a tutte le classi, e le *weighted-average*, che tiene conto del numero di elementi appartenenti a ciascuna classe così da restituire la media pesata.

Considerando che il modello bart-large-mnli nel dataset MNLI presenta un'accuratezza pari al 90%, il valore complessivo di **0.87** ci permette di affermare che il modello sia in grado di riconoscere il fenomeno della *bridging anaphora*.

5.4. RoBERTa

RoBERTa è un modello pre-addestrato in lingua inglese utilizzando un *masked laguage modelling* (MLM). Gli autori Liu et al. (2019) lo hanno creato come evoluzione del *transformer* BERT, sostenendo che quest'ultimo risultasse essere u*nder trained* ("poco addestrato") e che riuscisse comunque a superare con grandi distacchi i modelli pre-esistenti. RoBERTa è un modello *transformer* pre-addestrato su un ampio corpus in lingua inglese in modo non supervisionato. Questo significa che l'apprendimento avviene solo tramite i *raw texts* ("testi grezzi"), senza alcun tipo di etichetta umana e con un processo automatico per generare input ed etichette dai suddetti testi. Più precisamente, RoBERTa è stato pre-addestrato con il *Masked Language Modelling* (MLM), che consiste nel far sì che il modello "mascheri" casualmente il 15% delle parole in input, ottenendo un frase priva di alcune parole che devono essere predette dal modello stesso. In questo modo il modello è in grado di apprendere una rappresentazione bidirezionale della frase, discostandosi dalle tradizionali rete neurali ricorrenti (RNNs) o dai modelli auto-regressivi come GPT. In questo modo il modello è in grado di apprendere una rappresentazione interna della lingua inglese che può essere utilizzata per estrare feature relative a *downstream task*. La differenza maggiore rispetto a BERT risiede nel *dynamic masking* ("masking dinamico"), per cui i token mascherati non rimangono gli stessi per tutto l'addestramento, ma cambiano ad ogni sequenza, facendo sì che il modello non veda troppe volte lo stesso *masking pattern*.

RoBERTa è stato *finetuned* sul corpus MNLI così da poter essere usato in *text classification task* relativi al riconoscimento del *textual entailment* (TE). Per questa ragione si parla di *roberta-large-mnli*, ovvero un modello che date due frasi (premessa e ipotesi) restituisce l'etichetta di TE ad essa associata (*entailment*, *contradiction*, *neutral*). I risultati ottenuti durante l'addestramento su MNLI mostrano un'*accuracy* pari al 90%.

Abbiamo sottoposto a RoBERTa il dataset di *bridging anaphora*, andando a misurare le performance ottenute in merito al task di *recognizing textual entailment* (RTE).

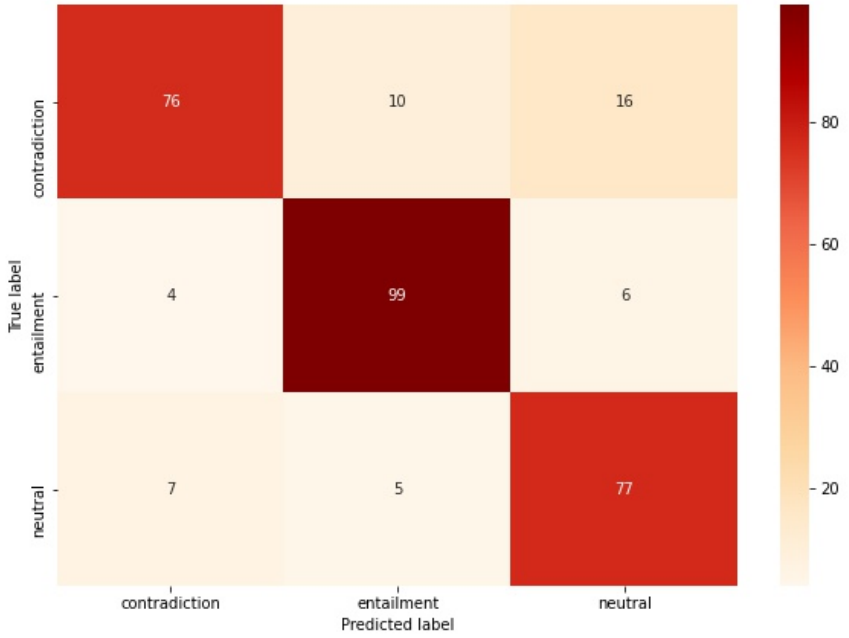

Figura 15: matrice di confusione sul dataset di bridging anaphora, testato su roberta-large-mnli

La matrice di confusione in figura 15 mostra dei risultati meno accurati rispetto a quanto ottenuto per il caso di *bart-large-mnli*. Per quanto riguarda le contraddizioni, il *transformer* ne ha classificate correttamente **99** su **109**, ottenendo quindi una performance anche più soddisfacente rispetto a BART. Al contrario, il caso neutrale ha visto riconosciuti **77** casi su **88**, mentre le contraddizioni sono state il caso più insoddisfacente fra tutti, con **76** casi classificati correttamente su **102**, ovvero 7 casi in meno rispetto a quanto fatto da BART.

Per quanto riguarda l'andamento del modello, la tabella 13 mostra le metriche di valutazione relative alla presente performance. Si può notare come il valore più alto di precisione sia in corrispondenza delle contraddizioni, che abbiamo visto però essere il caso con il maggior numero di errori di classificazione, aspetto che viene confermato dal valore di *recall*. Proprio in merito a quest'ultima, il risultato più

soddisfatene è quello relativo agli *entailment*, che sottolinea come il modello sia stato in grado di riconoscere le inferenza reali. Infine, l'f1-score è anche esso superiore per il caso delle inferenze.

Roberta-large-mnli	**Precision**	**Recall**	**F1-score**	**Accuracy**
Contradiction	**0.87**	0.75	0.80	
Entailment	0.87	**0.91**	**0.89**	
Neutral	0.78	0.87	0.82	
macro-average	0.84	0.84	0.84	
weighted-average	0.84	0.84	0.84	
total				**0.84**

Tabella 13: metriche di valutazione relative alla performance di roberta-large-mnli

Complessivamente RoBERTa (nel dataset MNLI presenta un'accuratezza pari al 90%) ha ottenuto un *accuracy* pari all'**84%**, leggermente inferiore rispetto a quanto visto nel caso di BART.

5.5. DistilBERT

DistilBERT è una versione distillata di BERT realizzata da Sanh et al. (2019). Parliamo quindi di un *transformer*, più piccolo e più veloce di BERT, che è stato pre-addestrato sullo stesso corpus in modo non supervisionato. Ciò significa il modello è stato addestrato solo con i testi grezzi, senza etichettature umane, con un processo automatico per generare input ed etichette da quei testi utilizzando il modello di base BERT. In particolare, tale modello è stato pre-addestrato con tre obiettivi:

- *Distillation loss*: il modello è stato addestrato per restituire le stesse probabilità del modello BERT.
- *Masked language modeling* (MLM): fa parte del *training loss* originale del modello BERT di base (vedi paragrafo relativo a RoBERTa).
- *Cosine embedding loss*: il modello è addestrato per generare *hidden states* il più vicino possibile al modello BERT di base.

DistilBERT apprende quindi le stesse rappresentazioni della lingua inglese apprese a loro volta da BERT, ma risultando più veloce per l'inferenza e per *downstream task*.
La versione di distilBERT utilizzata negli esperimenti contenuti nel presente elaborato è quella sottoposta al processo di *finetuning* sul MultiNLI dataset, e prende il nome di *distilbert-base-uncased-finetuned-mnli*. Anche il suddetto modello è stato utilizzato in modo analogo ai due precedentemente descritti, ovvero utilizzando il dataset di *bridging anaphora* per predire la classe di appartenenza di ciascuna coppia di frasi, facendo poi il confronto con il gold standard e mostrando i risultati nella matrice di confusione (vedi figura 16), misurando quindi le metriche di valutazione relative alla performance (vedi tabella 14).
DistilBERT risulta bilanciato fin dalle prime osservazioni essere il modello con la performance meno ottimale rispetto ai modelli visti finora. Ciò che però si nota maggiormente, dettaglio che sarà poi oggetto di analisi nel paragrafo successivo, è che le differenze maggiori riguardano sempre i casi di contraddizione e di neutralità, mentre gli *entailment* vengono predetti in modo molto simile dai vari modelli

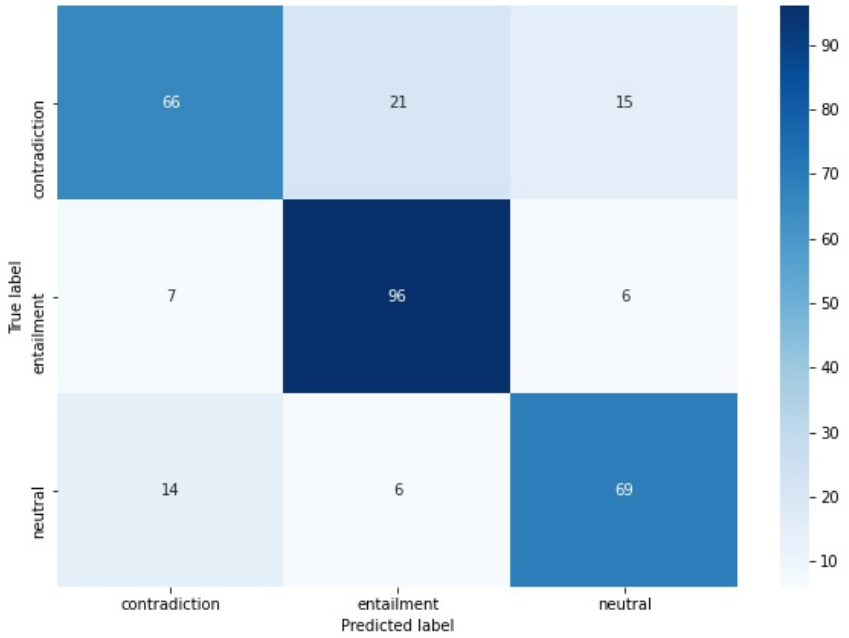

Figura 16: matrice di confusione sul dataset di bridging anaphora, testato su distilbert-base-uncased-finetuned-mnli

considerati. Anche in questo caso troviamo infatti **96** casi predetti correttamente rispetto ai **109** totali.

La situazione cambia in modo evidente se si vanno invece a considerare gli altri due casi. In particolare, la classe *neutral* presenta una classificazione corretta di **69** casi su **88**, mentre le contraddizioni sono sicuramente il risultato più deludente, con **66** casi predetti correttamente rispetto ai **102** totali. Parliamo di ben 36 casi non riconosciuti come *contradiction*, corrispondenti a circa il 37 % delle contraddizioni presenti nel dataset.

L'andamento del modello è invece valutabile a partire dalla tabella 14, che mostra le metriche di valutazione ottenute in merito alla performance di *distilbert-base-uncased-finetuned-mnli* sul dataset di *bridging anaphora*. Contrariamente ai casi precedenti (BART e RoBERTa), qua troviamo per ognuna delle misure considerate (*precision*, *recall*, *f1-score*) il valore massimo in corrispondenza dei casi di *entailment*. Ciò che quindi era stato osservato già a partire dalla matrice di

confusione viene ulteriormente confermato dalle metriche di valutazione. Nello specifico conviene soffermarsi sul valore di recall relativo alle contraddizioni, ovvero 0.65, che è il valore più basso riscontrato sino ad ora. Quest'ultimo ci dice che il modello è stato in grado di riconoscere solo il 65% delle contraddizioni reali, mostrando quindi una maggiore difficoltà rispetto ai modelli precedenti in relazione al caso *contradiction*.

Distilbert-base-uncased-finetuned-mnli	**Precision**	**Recall**	**F1-score**	**Accuracy**
Contradiction	0.76	0.65	0.70	
Entailment	**0.78**	**0.88**	**0.83**	
Neutral	0.77	0.78	0.77	
macro-average	0.77	0.77	0.77	
weighted-average	0.77	0.77	0.77	
total				**0.77**

Tabella 14: metriche di valutazione relative alla performance di roberta-large-mnli

Il valore di *accuracy* relativo a DistilBERT è pari al 77%, dimostrandosi un modello non completamente in grado di riconoscere il fenomeno della *bridging anaphora*. É necessario però sottolineare come tale modello abbia un valore di accuratezza più basso anche su MNLI, rispetto ai casi precedenti, e corrispondente all'82%.

5.6. Confronto tra i modelli

Nei paragrafi precedenti abbiamo visto e commentato le performance dei vari modelli utilizzati nella fase sperimentale del presente elaborato. In particolare, abbiamo visto come il *transformer* BART sia il modello che offre la prestazione migliore in relazione al dataset di *bridging anaphora*. Questa osservazione è riscontrabile anche nella tabella 15, che mostra il risultato delle metriche adottate in questa indagine per ciascuno dei modelli utilizzati.

Modello	m-Precision	w-Precision	m-Recall	w-Recall	m-F1	w-F1
BART	**0.87**	**0.87**	**0.87**	**0.87**	**0.87**	**0.87**
RoBERTa	0.84	0.84	0.84	0.84	0.84	0.84
DistilBERT	0.77	0.77	0.77	0.77	0.77	0.77

Tabella 15: confronto tra le metriche dei risultati ottenuti dai modelli sul dataset di bridging anaphora. “m” sta per per macro-average e “w” sta per weighted average.

La prestazione di BART è sicuramente la migliore fra tutte, ed anche quella con il valore di accuratezza più vicino a quanto raggiunto dal modello sul corpus MNLI (vedi tabella 15). Se infatti *bart-large-mnli* raggiunge sul corpus relativo un valore di *accuracy* poco al di sopra del 90%, è anche vero che la performance offerta sul dataset di *bridging anaphora* presenta un’accuratezza pari all’87 %, solo il 3% in meno rispetto a quella registrata su MNLI.

Come già detto nelle righe precedenti, il modello DistilBERT offre comunque una buona performance, seppur i valori mostrati nella tabella 15 siano più deludenti rispetti agli altri due modelli. Bisogna infatti considerare che il modello presenta un’*accuracy* pari all’82% sul corpus MNLI, aspetto che rende il modello meno preciso rispetto agli altri due casi, caratterizzati invece da valori di *accuracy* pari al 90% (vedi tabella 16). Tenendo conto di questo aspetto, anche nel caso della versione distillata di BERT troviamo una lacuna del 5% rispetto all’accuratezza misurata utilizzando MNLI. Questo dimostra che seppur il modello non sia tra i più indicati per i task di riconoscimento di TE, la sua prestazione sul dataset di *bridging* risulti comunque buona, in particolare per quanto riguarda il riconoscimento delle

inferenze, che risulta estremamente simile a quanto fatto da BART e da RoBERTa. Questi ultimi offrono dei risultati migliori in merito alle contraddizioni e al caso neutrale, dando prova della differenza nei valori di Accuracy tra i tre modelli sul corpus MNLI (vedi tabella 16).

Model	Dataset Brdging Accuracy	MNLI Accuracy
BART	**87**%	**90**%
RoBERTa	84%	90%
DistilBERT	77%	82%

Tabella 16: confronto tra l'accuracy raggiunta dai modelli sul dataset di bridging e su MNLI.

In sintesi, possiamo affermare che le performance dei tre modelli considerati in fase sperimentale siano piuttosto interessanti, dimostrando come il fenomeno linguistico della *bridging anaphora* possa essere compreso e conseguentemente riconosciuto e interpretato correttamente. Naturalmente, il decremento riscontrabile nelle performance dei vari modelli rispetto a MNLI è indice della complessità del fenomeno del *bridging*, ed è quindi perfettamente comprensibile che la misura di accuratezza ottenuta sul dataset di *bridging anaphora* sia leggermente inferiore a quanto ottenuto su MNLI.

5.7. BART: analisi degli errori

Abbiamo visto ed analizzato le performance dei tre modelli presentati nei paragrafi precedenti, confrontando le metriche di misura e constatando come BART sia il modello con la performance migliore, con un valore di accuracy pari all'87%. Ciò che andremo a fare in questo paragrafo sarà analizzare gli errori commessi dal suddetto modello, così da determinare le ragioni per cui la classificazione non è riuscita, individuando, magari, dei casi specifici che permettano in seguito una generalizzazione.

Come abbiamo visto nel paragrafo relativo, BART ha classificato correttamente **255** casi a fronte dei **300** contenuti nel dataset. Nello specifico, il modello ha prodotto **97** casi di *entailment* su **109**, **83** contraddizioni (*contradiction*) su **102**, e **79** casi neutrali (*neutral*) su **88**. Il grafico in figura 17 ci da un'idea più precisa dei risultati appena descritti. Come possiamo notare, la categoria che è stata riconosciuta di meno è quella delle contraddizioni, che in 19 casi sono state etichettate in modo errato, come già visto nel paragrafo relativo ai risultati ottenuti da BART.

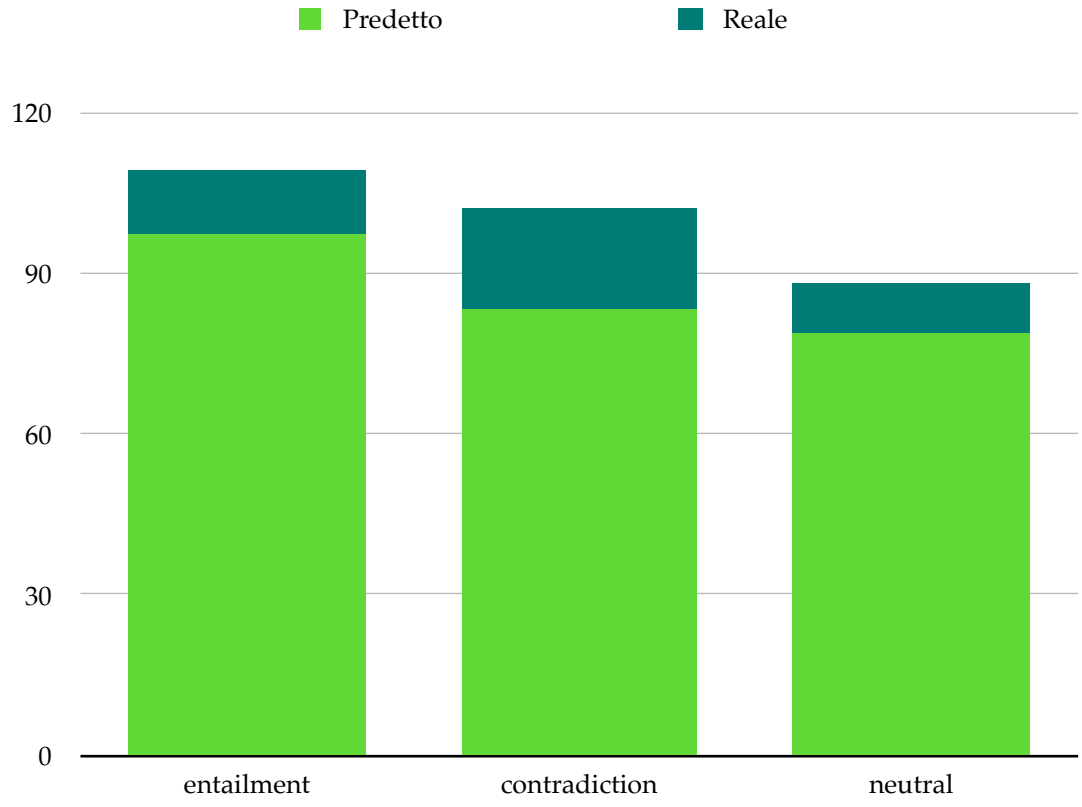

Figura 17: grafico relativo ai risultati della classificazione compita da bart-large-mnli

5.7.1.Contradiction

Andando ulteriormente nel dettaglio possiamo osservare che tra i casi etichettati erroneamente 13 sono stati riconosciuti dal modello come casi neutrali, mentre 6 come inferenze logiche. Questo primo dato permette di fare una prima stima degli errori compiuti, dimostrando indicativamente che il 68% dei casi errati corrisponde al caso *contradiction* ⇒ *neutral*. Possiamo quindi affermare che il modello tenderà a interpretare più facilmente una contraddizione come un caso di neutralità, piuttosto che come un caso di *entailment* (vedi figura 18).

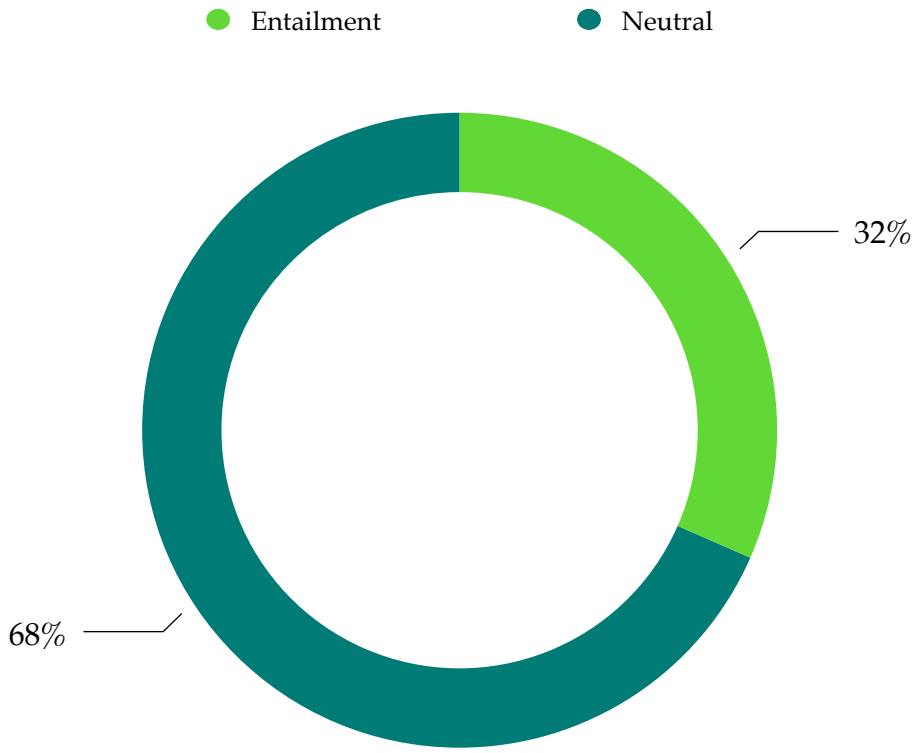

Figura 18: grafico che illustra la percentuale di casi di contraddiction classificate erroneamente.

Contradiction ⇒ neutral

Consideriamo quindi un caso di classificazione errata compiuto da BART sul dataset di *bridging anaphora*. In particolare cominciamo con un caso in cui una

contraddizione viene interpretata dal modello come un caso neutrale, così da interpretare le ragioni dietro l'errata classificazione (vedi tabella 17).

Premessa	Ipotesi	Reale	Predetta
An ABC's television interview was interrupted. The network will have to repay the money to the advertisers.	Advertisers were happy for the interruption on the ABC.	Contradiction	Neutral

Tabella 17: caso di classificazione errata (contradiction ⇒ neutral) da parte di BART sul dataset di bridging anaphora.

In questo caso parliamo di una contraddizione abbastanza sottile, che senz'altro ha messo a dura prova il modello. In particolare, parliamo di uno stato d'animo che non viene riconosciuto dal modello come contrario a quanto affermato nell'ipotesi. Gli inserzionisti ("advertisers") infatti, non possono essere felici dell'interruzione avvenuta sull'ABC, perché se comunque ci sarà un rimborso, l'acquisto dello spazio televisivo è avvenuto e l'intervista non è stata mostrata come da programma. È quindi lecito aspettarsi che gli inserzionisti non siano felici di quanto accaduto. Il modello non riesce a cogliere questa sfumatura e preferisce puntare sul caso *neutral*, interpretando lo stato d'animo degli inserzionisti come un informazione aggiuntiva che non conferma né smentisce la premessa. Vediamo ora i valori di score assegnati dal modello ad ognuna delle tre label, in relazione al caso nella tabella 18:

	Contradiction	Neutral	Entailment
Score	0.357106	**0.642567**	0.000325

Tabella 18: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per la frase in tabella 17

Come possiamo notare, la scelta del caso neutrale si fonda su un valore di score di circa il 64%, dimostrando che quindi la suddetta scelta non sia compiuta con una sicurezza particolarmente elevata. Risulta comunque chiaro che la preferenza del modello sia l'etichetta *neutral* e che non si parli di un caso in cui ci siano dei valori percentuali piuttosto simili in due o più casi.

Un altro dato interessante che possiamo controllare è il tipo di relazione di *bridging* presente nella premessa del caso analizzato. In particolare, è ragionevole pensare che certe relazioni siano riconosciute con più difficolta rispetto ad altre, aspetto che potrebbe essere determinato anche dal fatto che nel dataset di *bridging anaphora* non ci sia un bilanciamento in merito al campo "type" (come descritto nel capitolo 3). Il caso preso in esame ha in corrispondenza di "type" il valore "have", quindi il più comune tra tutti i tipi considerati. Questo potrebbe spiegare i valori di score visti nella tabella 18, implicando che l'errore del modello dipenda probabilmente da quanto detto in relazione alla presenza di uno concetto complesso come quello di uno stato d'animo, piuttosto che da un legame di *bridging* meno comune o comunque presente in misura minore nel dataset.

C*ontradiction* ⇒ *entailment*

Vediamo adesso un caso di classificazione errata in cui una contraddizione viene interpretata come un'inferenza:

Premessa	Ipotesi	Reale	Predetta
Anna had a very high fever last night. She needs to take an antibiotic.	Anna needs to take a pill because she had a very high fever last night	Contradiction	Entailment

Tabella 19: caso di classificazione errata (contradiction ⇒ entailment) da parte di BART sul dataset di bridging anaphora.

In questo caso è probabile che il modello non faccia distinzione tra le menzioni "pill" e "antibiotic", ma tenda a generalizzare il tutto sono il concetto di "medicinale". L'analisi semantica della premesse e delle ipotesi ci dice in entrambi i casi che Anna ha bisogno di cure perché ha avuto la febbre alta, ma la natura di queste cure non sembra diventare oggetto di dubbio per il modello, almeno così a prima vista.

Vediamo se i valori di score ci dicono qualcosa in più:

	Contradiction	Neutral	Entailment
Score	0.000154	0.003134	**0.996710**

Tabella 20: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per la frase in tabella 19

Ci troviamo chiaramente di fronte ad un caso in cui il modello non ha praticamente nessun dubbio su come classificare il rapporto tra premessa e ipotesi, con un score quasi del 100% per l'etichetta *entailment*. Questo non fa che rendere ancora più probabile l'ipotesi prodotta dall'analisi linguistica delle frasi, evidenziando come il modello non sembri individuare alcuna distinzione tra le parole "pill" e "antibiotic".
Per quanto riguarda invece il campo "type", il caso preso in esame presenta una relazione di tipo "cause", anch'essa tra le più comuni e con un percentuale di ricorrenza abbastanza elevata all'interno del dataset, proprio come il precedente caso di "have".

Valori di score bilanciati

Vediamo infine un ultimo caso di *contradiction* classificato in modo errato in cui i valori di score trovano un maggior bilanciamento rispetto ai casi visti finora, o comunque rispetto alla maggior parte dei casi in cui il modello ha fallito. Nella tabella 21 si trovano rispettivamente la premessa e l'ipotesi, tra cui si crea una contraddizione poiché la necessità di Sara di dover andare dal parrucchiere va in conflitto con il fatto che i capelli di Sara sembrino appena usciti dal parrucchiere.

Premessa	Ipotesi	Reale	Predetta
Sara needs to get her hair done. Sara has the gray roots.	Sara's hair looks like it just came out of the hairdresser.	Contradiction	Entailment

Tabella 21: caso di classificazione errata (contradiction ⇒ entailment) da parte di BART sul dataset di bridging anaphora.

In questo caso non è così semplice riuscire ad interpretare il comportamento del modello, anche perché i valori relativi allo scorre, come detto precedentemente, sono molto più bilanciati rispetto a quanto visto finora (vedi tabella 22).

	Contradiction	**Neutral**	**Entailment**
Score	0.175000	0.365204	**0.459794**

Tabella 22: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per la frase in tabella 21

La scelta dell'inferenza come classe più plausibile è quindi preferita rispetto alle altre due, seppur come valore non sia poi troppo distante rispetto alla neutralità. In questo caso, la difficoltà maggiore da parte del modello sta probabilmente nel non riuscire a collocare temporalmente i due eventi, avendo quindi difficoltà a individuare la contraddizione che si crea nel sostenere allo stesso tempo di aver sia bisogno di andare dal parrucchiere sia di esserci appena stati. Per questa ragione è probabile che il modello veda un'inferenza tra la premessa e l'ipotesi, e che quindi legga l'"essere andato al parrucchiere" come un qualcosa implicato dalla necessità di andarci. In ogni caso, tralasciando le suddette considerazioni, possiamo senza dubbio affermare che i valori di score facciano pensare ad uno dei casi di *contradiction* più difficili da classificare dell'intero dataset di *bridging anaphora*, o comunque uno di quelli in cui il modello ha mostrato una maggiore indecisione.

In merito al campo *type* non ci sono sorprese nemmeno in questo caso e la relazione di *bridging* presente nella premessa è una relazione di possessione ("have").

Il campo type nei casi errati

Come abbiamo visto, i casi considerati nelle righe precedenti non offrivano spunti particolarmente interessanti per quanto riguarda il tipo di relazione di *bridging*, che è risultata essere poco incisiva per quanto riguarda lo studio degli errori. Osservare infatti che alcuni casi classificati erroneamente corrispondano ad un tipo per cui la stragrande maggioranza dei casi è stata classificata in modo corretto, fa supporre che le difficoltà del modello non dipendano del campo "type".

Vediamo però nel dettaglio i tipi di relazioni di *bridging* caratterizzanti i casi di *contradiction* classificati erroneamente.

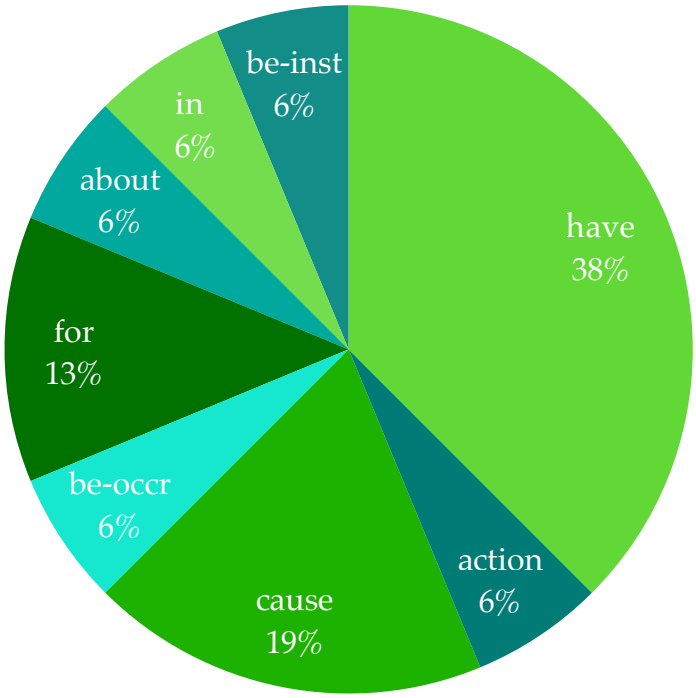

Figura 19: distribuzione dei valori del campo "type" nei casi classificati in maniera errata

Come possiamo notare nel grafico in figura 19, quasi il 40% dei casi classificati erroneamente è caratterizzato da una relazione di *bridging* di tipo *have* (possessione). Questo dato non è sorprendente, ma anzi estremamente ragionevole da un punto di vista puramente statistico, considerando che che i casi di tipo *have* presenti nel dataset sono **149**, ovvero circa il 50% sul totale. Stesso discorso per quanto riguarda il tipo *cause*, che nei casi errati trova **3** corrispondenze a fronte comunque delle 51 presenti nel dataset. Ciò su cui vogliamo invece soffermarci sono i casi più particolari, ovvero quelli etichettati rispettivamente con i tipi *action*, *be-occr* e *be-inst*. Come visto nel capitolo 3, queste tipologie di relazioni di *bridging* hanno trovato poche corrispondenze all'interno dei corpus utilizzati per la costruzione del dataset (BASHI e GUM) e sono quindi state create mentalmente delle premesse che contenessero loro. Il grafico in figura 19 ci dice che ciascuno di questi casi trova una

distribuzione nell'errore pari al 6%, che corrisponde ad **1** solo caso errato per ciascuno tipo, considerando che i casi di *contradiction* classificati erroneamente sono **19**. Questo dato dimostra che comunque, seppur si parli di legami meno comuni e quindi più complessi, il modello sia comunque in grado di riconoscerli e di classificarli in modo corretto. Purtroppo il giudizio sulla label *action* richiederebbe un maggior numero di casi per farsi un'idea più precisa in merito al riconoscimento del fenomeno, ma i due casi relativi alla tipologia *be* presentano solo 1 caso su 15 classificato erroneamente, valore piuttosto incoraggiante se proporzionato alle distribuzioni dei tipi *cause* e *have*. Possiamo quindi considerare l'idea che le difficoltà del modello nel classificare alcuni dei casi presenti nel dataset di *bridging* non dipendano, almeno per quanto riguarda la *contradiction*, dal tipo di relazione di *bridging* presente nella premessa, ma piuttosto da fattori legati, come abbiamo già visto nel paragrafo precedente, al contenuto semantico dell'ipotesi o a fattori di natura lessicale.

5.7.2.Entailment

Nel caso delle inferenze abbiamo visto come BART abbia classificato correttamente **97** casi a fronte dei **109** presenti nel dataset. Anche in questo caso il numero di inferenze classificate come *neutral* è maggiore di quello di inferenze classificate come *contradiction*, rispettivamente pari a 9 e a 3 casi (vedi figura 20). Risulta quindi chiaro come gli errori di riconoscimento di un caso di *entailment* da parte del modello nascano per lo più dalla lettura di quest'ultimi come casi di neutralità.

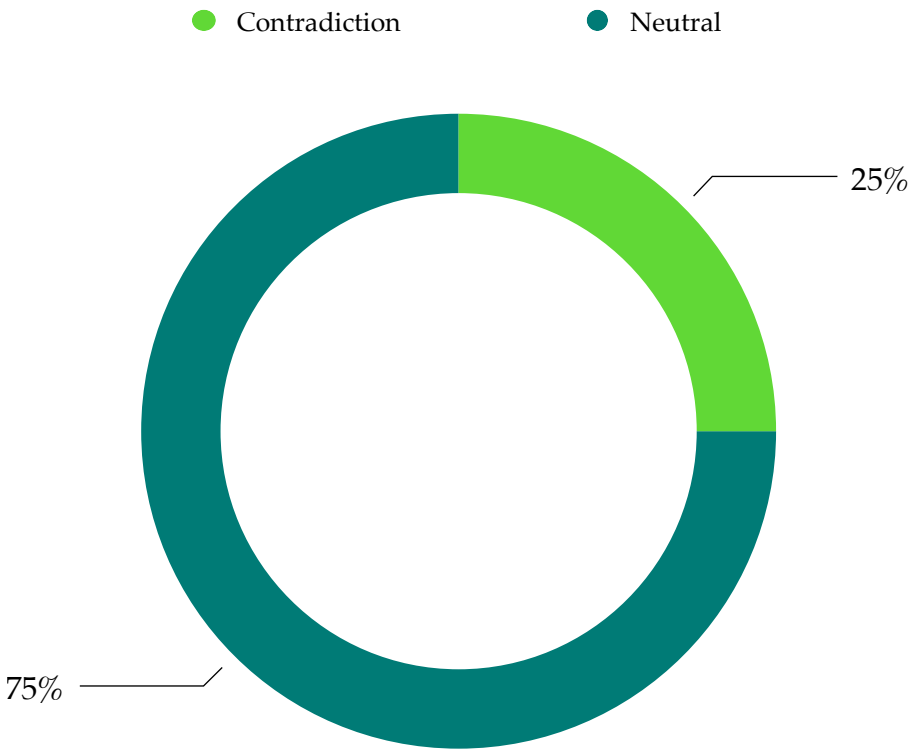

Figura 20: grafico che illustra la percentuale di casi di entailment classificate erroneamente.

Entailment ⇒ Neutral

Il caso in tabella 23 consiste in una classificazione errata da parte del modello *bart-large-mnli* di un'inferenza, in particolare quest'ultima viene interpretata come un caso di neutralità. La premessa ci dice che la mamma di Cara ha tirato fuori l'inalatore così che la figlia possa utilizzarlo. L'ipotesi è che quindi Cara abbia bisogno di utilizzare l'inalatore. Il modello legge però questa seconda frase come un

caso di neutralità, probabilmente separando a livello temporale la premessa l'ipotesi. In questo modo il fatto che Cara abbia bisogno di utilizzare l'inalatore non nasce come implicazione del fatto che la mamma glielo faccia utilizzare in quel dato istante, ma questi ultimi diventano due eventi separati che generano quindi un caso di neutralità. In sostanza il modello non vede nell'ipotesi una conferma al fatto che Cara utilizzi l'inalatore, potrebbe quindi utilizzarlo anche nel caso in cui non ne avesse bisogno. Come possiamo notare parliamo di un caso con tante sfumature, che risulta essere piuttosto difficile per un modello neurale, in particolare per quella che è la concezione di "necessità", ben diversa da quella di "scelta".

Premessa	Ipotesi	Reale	Predetta
Mom digs out the inhaler and Cara takes a hit.	Cara needs to take a hit at the inhaler	Entailment	Neutral

Tabella 23: caso di classificazione errata (entailment ⇒ neutral) da parte di BART sul dataset di bridging anaphora.

Le considerazioni presentate sopra trovano corrispondenza anche in merito ai valori di score che il modello produce per ciascuna classe di TE. Se osserviamo infatti la tabella 24, possiamo notare come valori di score relativi alla label *neutral* e a quella *entailment* siano rispettivamente pari al 64% e al 36% circa. Non c'è quindi una classe con una sicurezza pari al 90% o sopra, che non fa che sottolineare l'indecisione del modello nella classificazione del caso preso in esame.

	Contradiction	Neutral	Entailment
Score	0.003551	**0.638146**	0.358302

Tabella 24: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per la frase in tabella 23

In merito al campo "type" parliamo di un caso definito dal valore "for", che risulta essere meno comune rispetto ai più utilizzati *have* e *cause*. Ciò che però evidenzia come probabilmente la difficoltà del modello nel classificare l'esempio in tabella 23 non sia riconducibile al "type", è il fatto che gli altri casi che condividono la stessa premessa (e quindi lo stesso tipo) siano stati classificati correttamente dal modello. A

tal proposito, di seguito riportiamo un caso caratteristico in merito a quanto detto nelle ultime righe; un caso in cui la stessa premessa con due ipotesi differenti, che generano due casi di inferenza, sia classificata erroneamente in entrambi i casi (vedi tabella 25).

Premessa	Ipotesi	Reale	Predetta
A small bureaucrat died suddenly. A son sacrificed his career so that his brother could be successful.	One of the two sons of the bureaucrat sacrificed his career so that his brother would be successful.	Entailment	Neutral
A small bureaucrat died suddenly. A son sacrificed his career so that his brother could be successful.	The deceased bureaucrat had two children.	Entailment	Neutral

Tabella 25: due casi di classificazione errata (entailment ⇒ neutral) da parte di BART sul dataset di bridging anaphora che condividono la stessa premessa

Questo indica che probabilmente il modello non sia riuscito a riconoscere la *bridging anaphora* presente nella premessa, e quindi, indipendentemente dalla ipotesi, abbia fallito nel classificare i due casi correttamente. Si potrebbe pensare che il modello non trovi un legame tra "bureaucrat" e "son", e che quindi entrambe le informazioni espresse dalle ipotesi generino casi di neutralità. Il campo "type" è infatti occupato dal valore "have", che sottolinea il legame tra padre e figlio, ma che non sembra essere riconosciuto da BART. Nel primo caso, infatti, il modello sembra non associare la menzione "bureacraut" a quanto detto nella prima parte delle premessa, andando forse a pensare che i bureaucraut siano due, e non la stessa entità. Questo aspetto lo si ritrova anche nei valori di score, dove la label *neutral* ottiene un valore pari al 94% circa, dimostrando quindi poca indecisione da parte del modello nel compito di classificazione. Per quanto riguarda il secondo caso, si assiste probabilmente a un a sorta di ribaltamento del primo, poichè il modello sembra non riuscire in questo caso a far coincidere le menzioni "son" e "children", che diventano così i due figli del burocrate, ma distinti rispetto a quei fratelli menzionati nella seconda parte della premessa.

	Contradiction	Neutral	Entailment
Score primo caso	0.000401	**0.938168**	0.061429
Score secondo caso	0.019194	**0.656639**	0.324165

Tabella 26: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per i casi in tabella 25

La score relativo al secondo caso indica poi una maggiore indecisione, dovuta probabilmente alla capacità del modello nel sapere far coincidere le due menzioni di buraucraut.

Entailment ⇒ Contradiction

Come abbiamo visto nelle righe precedenti, le inferenze classificate erroneamente appartengono per lo più alla classe *neutral*. Ci sono però alcuni casi che vengono considerati delle vere e proprie contraddizioni, alcuni senza il minimo dubbio. Nella tabella 27 consideriamo proprio uno di questi.

Premessa	Ipotesi	Reale	Predetta
Carl got out of the car. Cara was watching him through the watery glass.	Cara stayed in the car after Carl went out.	Entailment	Contradiction

Tabella 27: caso di classificazione errata (entailment ⇒ neutral) da parte di BART sul dataset di bridging anaphora

Nella tabella 28, invece, mostriamo i valori di score per il caso che andremo a discutere a breve.

	Contradiction	Neutral	Entailment
Score	**0.998532**	0.001312	0.000155

Tabella 28: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per la frase in tabella 27

Cominciamo dalla premessa. In questo caso troviamo una scelta lessicale da non sottovalutare, ovvero l'utilizzo del termine "watery glass" per indicare il finestrino bagnato della macchina. Questa associazione è possibile solo tenendo conto che il

bridging link presente nella premessa sia di tipo contestuale, e che possa quindi essere interpretato correttamente solo all'interno della premessa stessa. In caso contrario, "watery glass" diventerà chiaramente "bicchiere d'acqua", perdendo il suo valore semantico, necessario per stabilire se ci sia un'implicazione tra premessa e ipotesi. La sicurezza con cui il modello sceglie la classe *contradiction* (vedi il valore di score nella tabella 28) fa pensare che il modello non riesca a dare il giusto valore alla menzione "watery glass" individuando quindi un contraddizione con l'ipotesi successiva. È infatti irragionevole pensare che Cara possa guardare Carl attraverso un bicchier d'acqua trovandosi dentro la macchina, mentre Carl sia invece uscito da quest'ultima. La plausibile errate lettura da parte del modello della menzione "watery glass" fa sì che il *brdiging link* di possessione ("have") non venga riconosciuto dal modello, e che quindi anche la successiva inferenza venga letta come una contraddizione.

Il campo type nei casi errati

Per quanto riguarda il campo "type" c'è molto meno da dire rispetto a quanto fatto nel paragrafo relativo alla contraddizioni classificate erroneamente. Il grafico in figura 21 mostra come la maggior parte dei casi di *entailment* non riconosciuti come

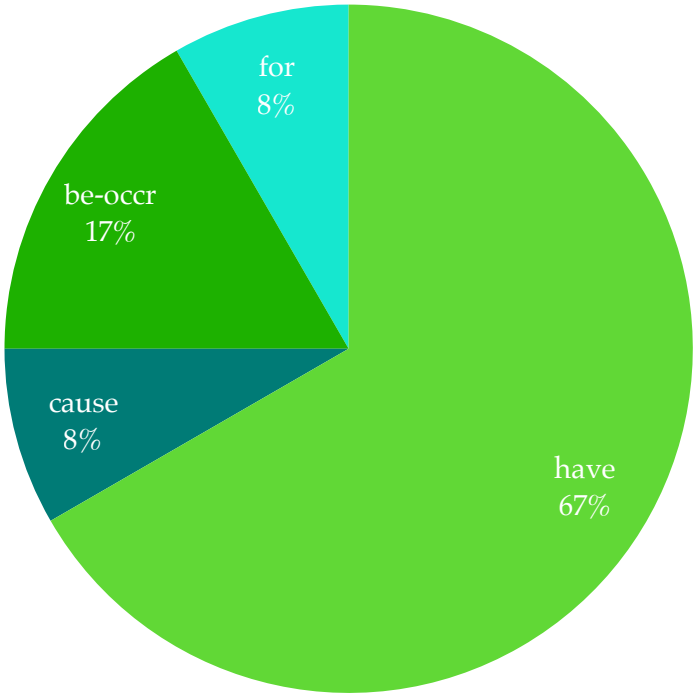

Figura 21: distribuzione dei valori del campo "type" nei casi classificati in maniera errata

tali sia caratterizzata da relazioni di possessione (“have”) e da qualche altro caso che non consente però di individuare delle relazioni che risultino più difficili da classificare come inferenze rispetto alle altre.

5.7.3. Neutral

Nei paragrafi precedenti abbiamo visto alcuni dei casi più interessanti relativi alla classificazione errata da parte di BART di inferenze e contraddizioni. In quest'ultimo paragrafo andremo a considerare il caso *neutral* che, come detto nelle considerazioni generali, risulta essere quello con il numero minore di casi errati con **9** errori su **88** rapporti di neutralità totali. Il grafico in figura 22 mostra la distribuzione dei casi *neutral* classificati in maniera errati rispetto alle label *entailment* e *contradiction*.

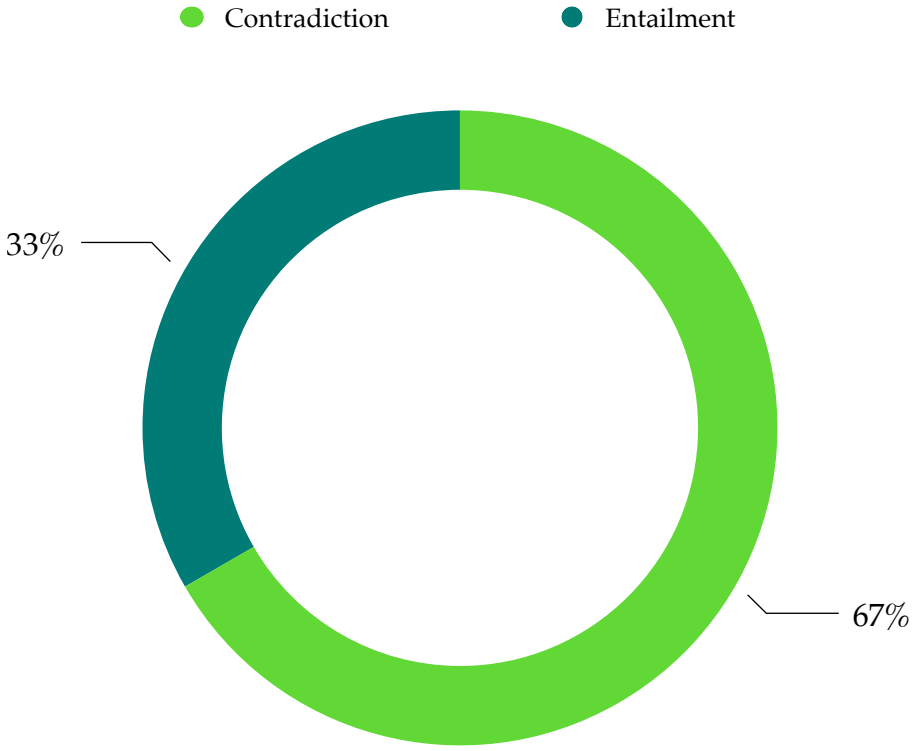

Figura 22: grafico che illustra la percentuale di casi di entailment classificate erroneamente.

Possiamo osservare come il modello tenda maggiormente a considerare i casi di neutralità che non riconosce come contraddizioni, piuttosto che come inferenze. Come sappiamo, una classificazione neutrale di una coppia premessa/ipotesi è data da un caso in cui l'ipotesi non offra né una conferma né una smentita della premessa. Il modello BART sembra preferire la seconda opzione, con il 67% dei casi di

neutralità erroneamente classificati che vengono interpretati come contraddizioni. Cominciamo quindi da questa tipologia nella nostra analisi degli errori.

Neutral ⇒ Contradiction

Consideriamo il caso proposto nella tabella 29. La premessa ci dice che l'università del Minnesota ha fatto dei test su 75 persone con livelli alti di colesterolo, mentre l'ipotesi ci dice che il gruppo che ha preso la medicina sta migliorando. La *bridging anaphora* è quindi data dall'interpretazione della menzione "the group" in relazione alle "75 people" che hanno partecipato allo studio. In particolare, l'informazione che ci viene fornita dalla frase, e che quindi possiamo inferire, è che i 75 soggetti siano stati divisi in due gruppi, di cui solo uno sia stato poi sottoposto al medicinale in esame ottenendo degli effetti benefici.

Premessa	Ipotesi	Reale	Predetta
The University of Minnesota tested 75 people with raised cholesterol levels. The group that took the medicine are getting better.	The medicine test groups consisted of 35 and 40 subjects respectively	Neutral	Contradiction

Tabella 29: caso di classificazione errata (neutral ⇒ contradiction) da parte di BART sul dataset di bridging anaphora

Si potrebbe presupporre, osservando in aggiunta i valori di score presenti nella tabella 30, che il modello non riesca a fare la considerazione per cui solo una parte dei soggetti scelti per il test abbia assunto il medicinali.

	Contradiction	Neutral	Entailment
Score	**0.996775**	0.003075	0.000149

Tabella 30: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per la frase in tabella 29

Probabilmente, infatti, BART associa alla menzione "group" l'intero blocco di 75 elementi, non facendo quindi la divisione che si è portati a realizzare avendo una conoscenza più approfondita di come funzionino i test di nuovi medicinali. Per questa ragione, il modello classifica con indecisione pressoché nulla (score = 0.99) il presente esempio come caso di contraddizione. Questo accade perché nella premessa

si parla di un gruppo di 75 soggetti (lettura errata) mentre nell'ipotesi si parla di due gruppi rispettivamente di 35 e 40 soggetti, ovvero i due gruppi in cui l'insieme totale è stato diviso per poter effettuare i test medici. Per quanto riguarda il campo "type", in questo caso troviamo un valore pari a "be-inst" che, come abbiamo visto nelle righe precedenti, non riesce ad essere interpretato correttamente dal modello, senza considerare che anche il caso di contradction che condivide la stessa premessa viene classificato erroneamente. L'unico caso riconosciuto correttamente è l'inferenza generata con la stessa premesse e contenuta nel dataset di *bridging anaphora*, che ha come ipotesi la seguente:

People with high cholesterol levels who took the medicine are getting better.

In questo caso si parla in modo generico di "people" senza far riferimento ai gruppi, generando quindi un'informazione derivata per inferenza a partire della premessa, ma allo stesso tempo un qualcosa che possa essere facilmente riconosciuto da un modello neurale.

Un caso come quello sopracitato richiede quindi di saper estrarre un'informazione contestuale e di essere in possesso di uno conoscenza del mondo risultando quindi un caso di classificazione abbastanza complicato per un modello come BART.

Neutral ⇒ entailment

Vediamo adesso un caso di *neutral* classificato erroneamente come *entailment* (vedi tabella 31). Nello specifico, parliamo di un caso di *bridging anaphora* di tipo "have" in cui la menzione "classes" viene interpretata correttamente in relazione alla menzione "school" precedente nella premessa.

Premessa	Ipotesi	Reale	Predetta
John went back inside the school because it was raining. There were two classes of each grade.	It was raining inside one of the school classrooms.	Neutral	Entailment

Tabella 31: caso di classificazione errata (neutral ⇒ entailment) da parte di BART sul dataset di bridging anaphora

BART interpreta il fatto che stia piovendo dentro uno delle classi della scuola, probabilmente per una finestra rimasta aperta, come un'inferenza ottenuta a partire dall'informazione contenuta nella premessa. Probabilmente l'errore da parte del modello in questo caso consiste nell'assenza di una maggiore conoscenza del mondo. BART potrebbe infatti pensare che il fenomeno atmosferico della pioggia stia avvenendo all'interno della scuola stessa. In tal caso si spiegherebbe la scelta della label "entailment" e del valore di score associato a quest'ultima (vedi tabella 32).

	Contradiction	**Neutral**	**Entailment**
Score	**0.996775**	0.003075	0.000149

Tabella 31: score assegnati dal modello bart-large-mnli a ciascuna delle tre label per la frase in tabella 32

Vediamo cosa succede se modifichiamo la premessa del caso in tabella 28 come segue:

> *John went back inside the school because **outside** it was raining. There were two classes of each grade.*

Abbiamo aggiunto il termine "outside" così da distinguere l'interno della scuola rispetto all'esterno. In questo caso il modello restituisce il valore di *contradiction* con uno score pari a 0.924, contrariamente a quanto accadeva nella frase originale (vedi tabella 29). BART legge quindi il fatto che stia piovendo dentro una classe come una contraddizione delle premessa, in cui viene affermato che invece stia piovendo all'esterno della scuola. Seppur il modello non riesca a cogliere il caso di neutralità (il fatto che piova dentro una classe non è un'implicazione della premessa, ma nemmeno contraddice quest'ultima) la classificazione prodotta è comunque più ragionevole rispetto a quanto accade nella frase in tabella 28. L'apparente incapacità del modello nel riuscire a comprendere che affermare che piova si riferisca all'esterno viene compensata dall'utilizzo della parola "outside", portando ad un'analisi inferenziale più razionale, seppur al tempo stesso sempre errata.

Questo caso dimostra quanto le modifiche lessicali, sintattiche o semantiche, anche le meno invasive, possano creare dei risultati diametralmente opposti in alcuni casi. Uno studio come quello presentato in questi ultimi paragrafi consente di approfondire numerosi aspetti del t*extual entailment* (TE) a partire proprio dall'analisi degli errori compiuti dai modelli.

Il campo type nei casi errati

Per concludere quest'ultimo paragrafo presentiamo infine la distribuzione dei valori del campo "type" nei casi di neutralità classificati erroneamente (vedi figura 23).

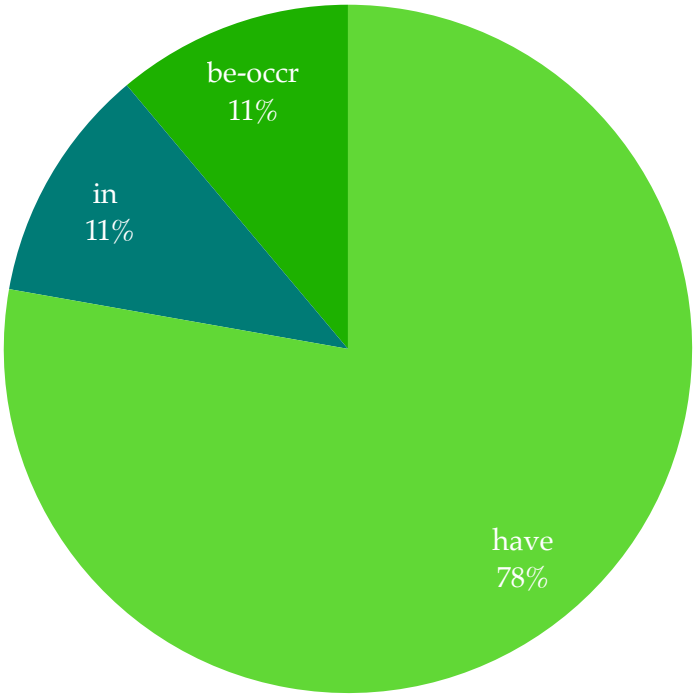

Figura 23: distribuzione dei valori del campo "type" nei casi classificati in maniera errata

Come abbiamo visto nel caso descritto nel paragrafo precedente (e in altri prima di esso), se il legame descritto dal campo "type" non viene riconosciuto dal modello è piuttosto probabile che il modello tenderà a classificare erroneamente tutti quei casi che nell'ipotesi presentano un'informazione derivabile solo e soltanto dalla comprensione della *bridging anaphora* contenuta nella premessa. Questo fatto sottolinea quanto il riconoscimento e la comprensione del *bridging link* da parte del

modello sia uno degli aspetti fondamentali per la riuscita della task centrale degli esperimenti descritti nella presente relazione.

In generale, dopo aver affrontato nel dettaglio i tre casi di errori possibili nel processo di classificazione, possiamo affermare che la maggior parte dei tipi di *bridging link* venga riconosciuta correttamente dal modello, e che inoltre tale riconoscimento sia necessario per la scelta della label associati ad ogni caso.

6. Conclusioni

Il presente elaborato ha considerato il fenomeno linguistico della *bridging anaphora* inserito nell'ambito del *Natural Language Inference* (NLI) e in particolare nel task di riconoscimento del *Textual Entailment* (TE). Nello specifico, l'idea di base da cui si è partiti è stata quella di considerare quali informazioni linguistiche vengano catturate dalle rappresentazioni distribuzionali, un argomento particolarmente dibattuto nelle ricerche di NLI e, in generale, di *Natural Language Processing* (NLP). Per eseguire correttamente un task di riconoscimento delle inferenze è necessario avere a disposizione degli embedding di qualità, così che i modelli possano raggiungere risultati ottimali. A tale scopo è quindi fondamentale avere a disposizione delle rappresentazioni distribuzionali che contengano le informazioni necessarie per il conseguimento dei task.

L'obiettivo di questa tesi è stato valutare se alcuni dei modelli neurali attualmente più utilizzati in studi di NLI fossero in grado di registrare e rappresentare il fenomeno della *bridging anaphora*. In particolare, all'interno di un task di riconoscimento di TE sono state valutate le rappresentazioni distribuzionali prodotte da dei modelli *transformer*, al fine di stabilire se avessero catturato l'informazione relativa al *bridging*. Quest'ultima consiste in un collegamento di tipo anaforico (all'interno di una stessa frase o tra una coppia di frasi) tra due menzioni, che possono a loro volta essere una singola parola o un'intera porzione di frase. Riconoscere il fenomeno linguistico così proposto permette di comprendere correttamente la\e frase\i, risalendo a un'informazione di tipo contestuale o relativa ad una conoscenza del mondo.

Per questa indagine è stato realizzato un dataset di 300 coppie di frasi contenenti molteplici casi di *bridging anaphora*, che sono stati a loro volta distinti in base al tipo di relazione. Il *challenge set* così ottenuto è stato sottoposto alla classificazione di tre modelli (BART, RoBERTa e DistilBERT) pre-addestrati sul corpus MNLI (Williams et al.; 2017). In seguito sono stati analizzati i risultati ottenuti facendo un confronto tra i modelli e cercando di individuare le analogie e le differenze tra le varie performance. Inoltre, è stato tenuto conto dei risultati degli stessi modelli sul corpus di addestramento, così da poter eseguire un confronto con i risultati ottenuti sul dataste di *bridging anaphora*. Gli esiti degli esperimenti hanno dimostrato che i modelli sono in grado di riconoscere il fenomeno del *bridging*, con alcuni risultati che mostrano un valore di accuratezza poco lontano da quanto raggiunto nei test set di MNLI, come per esempio il caso di BART (87% di *accuracy* contro il 90% misurato su MNLI).

Per quanto riguarda l'analisi degli errori, abbiamo scelto di considerare nel dettaglio i risultati ottenuti sul migliore dei tre modelli, ovvero BART, che ha messo in luce come esso sia naturalmente vulnerabile di fronte a casistiche più rare. La maggior parte degli errori è infatti da ricondurre a casi di *bridging* particolarmente complessi, in cui l'informazione concettuale che ci aspetteremmo fosse derivata dal modello non venga in realtà identificata, segnale che sottolinea inoltre i tipi di relazione maggiormente complessi, e per questo meno diffusi. Gli esperimenti infatti tendono a dimostrare come relazioni di bridging poco presenti sia nel parlato sia nello scritto, risultino presenti nei casi per cui il modello produce il maggior numero di errori di classificazione, al di là naturalmente di una serie di costruzioni linguistiche che, come già detto, non vengono riconosciute in quanto non corrispondenti a euristiche superficiali e/o frequenti.

Per quanto concerne gli sviluppi futuri legati al dataset descritto nel presente elaborato, sarebbe interessante arricchirlo, andando da una parte ad aumentare il numero di occorrenze, mentre dall'altra cercando di inserire un numero maggiore di casi rari e poco comuni, così da poter avere, in merito ad essi, dei risultati più esaurienti. Inoltre, visti gli ottimi risultati ottenuti dai modelli, si potrebbe perfino aumentare la complessità dei casi, così da produrre un dataset privo di semplificazioni o *bias*. Infine, si potrebbe inserire all'interno del training utilizzato per l'addestramento dei modelli, una porzione di dati relativi al fenomeno della *bridging anaphora*, così da valutare ulteriori possibili miglioramenti nella fase sperimentale.

7. Bibliografia

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., Goldberg, Y., 2017. *Fine-grained analysis of sentence embeddings using auxiliary prediction tasks.* ICLR 2017.

Amir Zeldes. 2017. *The GUM corpus: Creating multilayer resources in the classroom.* Language Resources and Evaluation, 51(3): 581–612.

Anders Bj¨orkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. *The extended dirndl corpus as a resource for automatic coreference and bridging resolution.* Proceedings of the 9th International Conference on Language Resources and Evaluation, pages 3222–3228.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton (2016). *Layer Normalization*. en. In: CoRR. arXiv: 1607.06450.

Belinkov, Y., Glass, J. R., 2019. *Analysis Methods in neural language processing: a survey.* Transactions of the Association for Computational Linguistics.

Ben Kantor and Amir Globerson. 2019. *Coreference resolution with entity equalization.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 673–677.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. *Enriching word vectors with subword information.* Transactions of the Association for Computational Linguistics, 5: 135-146.

Bowman, S. R., Angeli, G., Potts, C., Manning, C. D., 2015. *A large annotated corpus for learning natural language inference.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Claire Gardent and Hélène Manuélian. 2005. *Création d'un corpus annoté pour le traitement des descriptions définies*. Traitement Automatique des Langues, 46(1):115–140.

Claire Gardent, H´el`ene Manu´elian, and Eric Kow. 2003. *Which bridges for bridging definite descriptions?* Proceedings of 4th International Workshop on Linguistically Interpreted Corpora.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A., 2017. *Supervised Learning Representations from Natural Language Inference Data.* Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M., 2018. *What you can cram into a single vector: Probing sentence embeddings for linguistic properties.* Proceedings of Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805.

Ellen Prince. 1981. *Toward a taxonomy of given-new information. In P. Cole, editor, Radical Pragmatics*, page 223–255. Academic Press, New York, N.Y.

Emmanuel Lassalle and Pascal Denis. 2011. *Leveraging different meronym discovery methods for bridging resolution in French.* The 8th Discourse Anaphora and Anaphor Resolution Colloquium, pages 35–46.

Ettinger, A., Elgohary, A., Phillips, C., Resnik, P., 2018. *Assessing Composition in Sentence Vectors Representations.* Proceedings of the 27th International Conference on Computational Linguistics.

Ettinger, A., Elgohary, A., Phillips, C., Resnik, P., 2018. *Assessing Composition in Sentence Vectors Representations.* Proceedings of the 27th International Conference on Computational Linguistics.

Goldberg, Y., 2019. *Assessing BERT's Syntactic Abilities* . arXiv: 1901.05287v1.

Harris, Z., 1954. *Distributional structure.Word* , 10(23): 146-162.

Herbert H. Clark. 1975. *Bridging. In Theoretical Issues in Natural Language Processing*.

Hewitt, J., Manning, C. D., 2019. *A Structural Probe for Finding Syntax in Word Representations*. Proceedings of NAACL-HLT 2019, Association for Computational Linguistics.

Hideo Kobayashi and Vincent Ng. 2020. *Bridging Resolution: A Survey of the State of the Art*. Proceedings of the 28th International Conference on Computational Linguistics, pages 3708-3721.

Ina R¨osiger. 2016. *SciCorp: A corpus of English scientific articles annotated for information status analysis.* Proceedings of the Tenth International Conference on Language Resources and Evaluation, pages 1743–1749.

Ina Rösiger. 2018a. *BASHI: A corpus of wall street journal articles annotated with bridging links.* Proceedings of the Eleventh International Conference on Language Resources and Evaluation.

Iørn Korzen and Matthias Buch-kromann. 2011. *Anaphoric relations in the Copenhagen Dependency Treebanks.* Proceedings of Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena, pages 83–98.

Jawahar, G., Sagot, B., Seddah, D., 2019. *What does BERT learn about the structure of language?* Proceedings of the 57th Annual Meeting for Computational Linguistics.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. *Cognitive status and the form of referring expressions in discourse*. Language, pages 274–307.

Jiang, Nanjiang and Marie-Catherine de Marneffe (2019). *Evaluating BERT for Natural Language Inference: A Case Study on the CommitmentBank*. en. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 6085–6090. doi: 10.18653/v1/d19-1630.

John A Hawkins. 1978. *Definiteness and indefiniteness: A study in reference and grammaticality prediction.* Journal of Linguistics, 27:405–442.

Juntao Yu and Massimo Poesio. 2020. *Multi-task learning based neural bridging reference resolution*. arXiv:2003.03666 [cs.CL].

Jurafsky, D., Martin, J. H., 2018. *Speech and Language Processing. Prentice-Hall.*

Katja Markert, Yufang Hou, and Michael Strube. 2012. *Collective classification for fine-grained information status.* Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 795–804.

Katrin Schweitzer, Kerstin Eckart, Markus G¨artner, Agnieszka Falenska, Arndt Riester, Ina R¨osiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn. 2018. *German radio interviews: The grain release of the sfb732 silver standard collection*. Proceedings of the Eleventh International Conference on Language Resources and Evaluation.

Kepa Joseba Rodrìguez, Francesca Delogu, Yannick Versley, Egon W. Stemle, and Massimo Poesio. 2010. *Anaphoric annotation of Wikipedia and blogs in the live memories corpus*. Proceedings of the Seventh International Conference on Language Resources and Evaluation, pages 157–163.

Kiros, R., Zhu, Y., Salakhutdinov, R. R, Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. *Skip-Thought Vectors.* Advances in neural Information Processing Systems 28.

Lehmann, S., Oepen, S., RegnierProst, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., et al., 2019. *TSNLP: Test Suites for Natural Language*. Proceedings of Association for Computational Linguistics.

Lenci, A., 2008. *Distributional semantics in linguistic and cognitive research*. Italian Journal of Linguistics, gennaio 2008.

Lenci, A., Montemagni, S., Pirrelli, V., 2016 . *Testo e Computer. Elementi di Linguistica Computazionale*. Carocci editore, Roma.

Levesque, Hector J., Ernest Davis, and Leora Morgenstern (2012). *The Winograd Schema Challenge*. Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. KR'12. Rome, Italy: AAAI Press, pp. 552–561. isbn: 978- 1-57735-560-1.

Levi, J.N.: *The syntax and semantics of complex nominals*. Academic Press, New York : (1978)

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2019. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation and Comprehension*. arXiv: 1910.13461

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., Smith, N. A, 2019. *Linguistic Knowledge and Transferability of Contextual Representations*. Proceedings of NAACL-HLT 2019, Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach* . arXiv: 1907.11692.

Maria Lo Duca, *Lingua italiana ed educazione linguistica*, ed. Carocci, 2008 (8a ristampa della 1ª ed.), ISBN 978-88-430-2646-3

MacCartney, Bill (2009). *Natural Language Inference*. en. Ph.D. Dissertation. Stanford University.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. *SpanBERT: Improving pre-training by representing and predicting spans*. Transactions of the Association for Computational Linguistics, 8: 64–77.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., 2014. *A SICK cure for the evaluation of compositional semantic models*. Language Resources and Evaluation Conference.

Massimo Poesio and Renata Vieira. 1998. *A corpus-based investigation of definite description use*. Computational Linguistics, 24(2): 183–216.

Massimo Poesio and Ron Artstein. 2008. *Anaphoric annotation in the ARRAU corpus*. Proceedings of the Sixth International Conference on Language Resources and Evaluation, pages 1170–1174.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. *Resolving bridging references in unrestricted text*. Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004b. *Centering: A parametric theory and its instantiations*. Computational linguistics, 30(3):309–363.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. 2002. *Acquiring lexical knowledge for anaphora resolution*. Proceedings of the Third International Conference on Language Resources and Evaluation.

McCoy, R. T., Pavlick, E., Linzen, T., 2019. *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. *Distributed representations of words and phrases and their compositionality*. Advances in Neural Information Processing Systems.

Mitchell, J., Lapata, M., 2010. *Composition in Distributional Models of Semantics* . Cognitive Science, vol. 34.

Morris, John et al. (2020). *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*. en. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics, pp. 119–126. doi: 10/gjpbxq.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., Neubig, G., 2018. *Stress Test Evaluation for Natural Language Inference*. Proceedings of the 27th International Conference on Computational Linguistics.

Nangia, N., Williams, A., Lazaridou, A., Bowman, S. R., 2017. *The repeval 2017 shared task: Multi-genre natural language inference with sentence representations* . Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, pp. 1-10.

Nie, Y., Bansal, M., 2017. *Shortcut-stacked sentence encoders for multi-domain inference*. Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, pp. 41-45.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, and Massimo Poesio. 2020. *Annotating a broad range of anaphoric phenomena, in a variety of genres: The ARRAU corpus*. Natural Language Engineering, 26(1): 95–128.

Parma Nand and Wai Yeap. 2013. *A Framework for Interpreting Bridging Anaphora.* Communications in Computer and Information Science 358: 131-144. Proceedings of the 2013 International Conference on Agents and Artificial Intelligence.

Pascal Denis and Jason Baldridge. 2008. *Specialized models and ranking for coreference resolution*. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 660–669.

Pennington, J., Socher, R., Manning, C., 2014. *GloVe: Global Vectors for Word Representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. *Deep Contextualized Word Representations*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1.

Razvan Bunescu. 2003. *Associative anaphora resolution: A web-based approach*. Proceedings of the EACL Workshop on The Computational Treatment of Anaphora, pages 47–52.

Renata Vieira and Massimo Poesio. 2000. *An empirically based system for processing definite descriptions*. Computational Linguistics, 26(4): 539–593.

Renata Vieira and Simone Teufel. 1997. *Towards resolution of bridging descriptions*. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics, pages 522–524.

Richardson, Kyle et al. (2019). *Probing Natural Language Inference Models through Semantic Fragments*. en. In: arXiv:1909.07521 [cs]. arXiv: 1909.07521 [cs].

Sebastian Löbner. 1998. *Definite associative anaphora*. http://user.phil-fak.uniduesseldorf. de/˜loebner/publ/DAA-03.pdf.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. *MCScript2.0: A Machine Comprehension Corpus Focused on Script Events and Participants*. Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019), pages 103–117.

Subramanian, S., Trischler, A., Bengio, Y., Pal, C. J., 2018. *Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning*. ICLR 2018. arXiv:1804.00079.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., Pavlick, E., 2019. *What do you learn from context? Probing for sentence structure in contextualized word epresentations* . ICLR 2019.

Tommaso Caselli and Irina Prodanof. 2006. *Annotating bridging anaphors in Italian: in search of reliability*. Proceedings of the Fifth International Conference on Language Resources and Evaluation.

Vanmassenhove, E., Shterionov, D., Way, A., 2017. *Lost in Translation: Loss and Decay of Linguistics Richness in Machine Translation*. Proceedings of Machine Translation Summit XVII 1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I., 2017. *Attention is all you need*. arXiv:1706.03762

Vincent Ng and Claire Cardie. 2002. I*mproving machine learning approaches to coreference resolution*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 104–111.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. R., 2019. SuperGLUE: *A Stickier Benchmark for General-Purpose Language Understanding Systems*. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. R., 2018. GLUE: A *Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353-355.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. *A machine learning approach to coreference esolution of noun phrases*. Computational linguistics, 27(4):521–544.

Williams, Adina, Nikita Nangia, and Samuel Bowman (2018). *A Broad- Coverage Challenge Corpus for Sentence Understanding through Inference*. en. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. doi: 10.18653/v1/n18-1101.

Yufang Hou, Katja Markert, and Michael Strube. 2014. *A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pages 2082–2093.

Yufang Hou. 2018a. *A deterministic algorithm for bridging anaphora resolution*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1938–1948.

Yufang Hou. 2018b. *Enhanced word representations for bridging anaphora resolution*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 1–7.

Yufang Hou. 2020. *Bridging anaphora resolution as question answering*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1428–1438.