Studio sul sistema di generazione automatica di sottotitoli in italiano su Youtube

Silvia Di Giampasquale, Chiara Segala
silvi.digiampasquale@studio.unibo.it
chiara.segala@studio.unibo,it

Abstract

Questo progetto si propone di analizzare il sistema di generazione automatica dei sottotitoli su YouTube. Per raggiungere questo obiettivo, sono stati selezionati casualmente quindici video da una playlist di un canale in cui l'autore fornisce sottotitoli in italiano. L'analisi ha confrontato i sottotitoli forniti dal creatore con quelli generati automaticamente dalla piattaforma, al fine di identificare e classificare gli errori. Lo scopo è comprendere quali tipologie di errori siano più comuni nel processo di generazione automatica e valutare la qualità complessiva dei sottotitoli generati dal sistema.

1 Introduzione

L'idea di questo progetto nasce dal desiderio di valutare la qualità e l'accuratezza del sistema di generazione automatica dei sottotitoli di YouTube. La curiosità è stata stimolata dall'osservazione di errori evidenti riscontrati attivando i sottotitoli automatici in vari video; spesso, infatti, risultano presenti inesattezze che ne compromettono la comprensibilità e l'efficacia comunicativa. Partendo da queste osservazioni, l'obiettivo principale del progetto è quello di esplorare e quantificare la frequenza e il tipo di errori presenti nei sottotitoli generati automaticamente.

Questa analisi offre anche l'opportunità di applicare la programmazione in Python all'ambito del Natural Language Processing (NLP), settore dalle numerose applicazioni pratiche e con un potenziale di utilizzo significativo nel contesto professionale. Nello specifico, abbiamo indagato se gli errori nei sottotitoli automatici tendano a seguire schemi ricorrenti oppure se siano di natura più casuale.

Per eseguire un'analisi accurata, abbiamo selezionato un campione di quindici video da una playlist di un canale YouTube che fornisce sottotitoli accurati in italiano. I sottotitoli forniti dal creatore sono stati considerati come un gold standard, ovvero un modello di riferimento rispetto al quale valutare la precisione dei sottotitoli generati automaticamente. Gli errori nei sottotitoli automatici sono stati quindi confrontati con questo standard per identificare eventuali tendenze o tipologie di errori comuni, come problemi di trascrizione, errori grammaticali o mancanze semantiche.

Attraverso questo confronto, ci proponiamo di ottenere una panoramica più chiara delle capacità attuali del sistema di sottotitolazione automatica di YouTube, individuandone i limiti e i margini di miglioramento.

2 Prima fase di selezione dei video

Il primo passo del progetto è stato selezionare con attenzione i video più adatti agli obiettivi della ricerca. La prima operazione consisteva nel trovare un canale che fornisse sottotitoli per diversi video. Per questo scopo, abbiamo scelto il canale Progetto Happiness. Da questo creatore, abbiamo deciso di selezionare i video dalla playlist Latest Episodes.

Per la nostra analisi, sono stati scelti quindici video in modo casuale, seguendo i criteri di selezione riportati di seguito:

1. Durata: abbiamo limitato la selezione ai video di almeno quindici minuti, in modo da garantire una quantità sufficiente di dati per un confronto significativo.

2. Presenza di entrambe le tipologie di sottotitoli: per rispondere alle necessità

del progetto, era fondamentale che tutti i video selezionati avessero sia i sottotitoli forniti dal creatore sia quelli generati automaticamente. È emerso che, prima del 2023, il creatore non forniva i sottotitoli in italiano in alcuni video, mentre in altri mancavano i sottotitoli automatici.

3. Disponibilità: sono stati esclusi i video in première o programmati per una data futura, in quanto non erano ancora accessibili al momento della selezione.

Abbiamo quindi sviluppato un codice in grado di generare un elenco casuale di quindici video che soddisfacessero i tre criteri stabiliti. Al termine dell'esecuzione, il codice ha prodotto il seguente elenco:

1. DENTRO IL TEMPIO DEI MONACI SHAOLIN
2. GLI UOMINI PESCE: la Tribù dei Badjao mutati geneticamente
3. GLI UOMINI RENNA - La tribù che sopravvive a -50° tra orsi e lupi
4. SULLE STRADE PIÙ PERICOLOSE DEL MONDO - I camionisti immortali del Pakistan
5. IL POPOLO DELLE MONTAGNE MAROCCHINE
6. IL REGNO DEI NANI: Dentro il parco giochi più controverso del mondo
7. KALASH: l'ultimo matriarcato che combatte i Talebani
8. L'ULTIMO BEDUINO del deserto del Wadi Rum
9. L'ULTIMO CACCIATORE DI BALENE delle Azzorre
10. LA CITTÀ SEGRETA DENTRO IL CIMITERO DI MANILA
11. LA VITA DI UNA DEA VIVENTE: bimba vergine non può più toccare terra con i piedi
12. MADE IN BANGLADESH - la storia dei bambini operai nel Fast Fashion
13. PAG PAG - Cucinare la spazzatura per sopravvivere a Manila
14. RAZZISMO AMBIENTALE NEL CUORE DEL BRASILE: una realtà sconosciuta
15. SETTA SEGRETA GIAPPONESE NASCOSTA TRA LE MONTAGNE

3 Download e pre-processing dei sottotitoli

Abbiamo sviluppato due codici distinti per scaricare i sottotitoli dei video: uno per i sottotitoli forniti dall'autore e l'altro per quelli generati automaticamente. I file sono stati salvati nei formati .vtt e .srt e successivamente convertiti in formato .txt. Successivamente, abbiamo effettuato una pulizia dei dati, rimuovendo il minutaggio e le annotazioni relative a elementi ambientali (come [musica] e [applauso]), poiché non rilevanti per il nostro confronto.

In seguito, abbiamo applicato le tecniche standard di elaborazione del linguaggio naturale, eseguendo la tokenizzazione, la lemmatizzazione e il Part-of-Speech tagging su tutti i testi, al fine di prepararli adeguatamente per il confronto nelle diverse versioni. Per eseguire queste operazioni, sono state utilizzate due librerie principali: NLTK (Natural Language Toolkit), una delle librerie più popolari per il processamento del linguaggio naturale, per la tokenizzazione e Stanford Stanza, una suite di strumenti di NLP sviluppata dalla Stanford NLP Group, che offre funzionalità avanzate utilizzando modelli di deep learning pre-addestrati per linguaggi diversi, per la lemmatizzazione e il Part-of-Speech tagging.

4 Confronto

La terza e ultima fase del progetto è stata dedicata al confronto tra le due tipologie di testi ottenuti. Questo confronto è stato applicato ai testi tokenizzati, lemmatizzati e annotati con il Part-of-Speech tagging.

Abbiamo utilizzato Pandas, creando dataframe che sintetizzano i dati per ogni video. Ogni riga rappresenta un video, mentre le colonne includono: il numero di elementi unici nei sottotitoli dell'autore, il numero di

elementi unici nei sottotitoli automatici, e due ulteriori colonne contenenti gli elenchi di questi elementi distinti per ciascuna tipologia di sottotitolo.

A completamento dell'analisi, abbiamo aggiunto due tabelle riassuntive. La prima mostra, per ciascun video, le percentuali di occorrenza di ogni categoria grammaticale (PoS). La seconda fornisce una visione complessiva, elencando ogni categoria grammaticale con la relativa percentuale di presenza nei sottotitoli dell'autore e in quelli generati automaticamente, calcolata su tutti i quindici video analizzati.

5 Risultati

Un aspetto evidente fin dal momento del download delle due tipologie di testi, ancor prima di effettuare qualsiasi confronto, è stata l'assenza totale di punteggiatura nei sottotitoli automatici, a differenza di quelli forniti dall'autore, in cui era correttamente inclusa.

L'analisi dei confronti ha ulteriormente confermato questa osservazione iniziale. Come si può osservare dalle tabelle, infatti, i token relativi alla punteggiatura sono presenti in tutte le versioni di testo analizzate (tokenizzate, lemmatizzate e annotate con PoS tagging) esclusivamente nei sottotitoli d'autore, mentre risultano completamente assenti in quelli generati automaticamente.

Questa discrepanza emerge chiaramente anche dalla tabella finale che riassume le percentuali delle categorie grammaticali (PoS): nei sottotitoli forniti dal creator, la punteggiatura rappresenta il 12,44% del totale, mentre nei sottotitoli automatici la sua presenza scende a un trascurabile 0,01%.

Un'altra significativa differenza emersa nelle percentuali riguarda il tag X, che nel tagset delle Universal Dependencies viene attribuito ai token che non possono essere classificati in nessun'altra categoria grammaticale. Dall'analisi dell'ultima tabella, si nota che la percentuale di questo tag nei sottotitoli d'autore è pari allo 0,12%, mentre nei sottotitoli generati automaticamente sale al 6,54%.

Esaminando la tabella precedente, che riporta le percentuali delle categorie grammaticali (PoS) per ciascun video, emerge che le maggiori discrepanze per il tag X si concentrano su quattro video specifici: L'ULTIMO BEDUINO del deserto del Wadi Rum, L'ULTIMO CACCIATORE DI BALENE delle Azzorre, LA VITA DI UNA DEA VIVENTE: bimba vergine non può più toccare terra con i piedi, RAZZISMO AMBIENTALE NEL CUORE DEL BRASILE: una realtà sconosciuta.

Un'analisi più approfondita dei confronti effettuati per questi video evidenzia che la discrepanza è legata a token rilevati dal sistema nella loro lingua originale, che però non compaiono nei sottotitoli d'autore poiché tradotti e integrati direttamente come caption nel video. Di conseguenza, nei sottotitoli generati automaticamente, questi token non vengono assegnati a nessun'altra categoria grammaticale.

Un'analisi più qualitativa dei confronti ha evidenziato due ulteriori discrepanze. La prima riguarda gli elementi numerali: nei sottotitoli generati automaticamente, i numeri sono generalmente rappresentati in cifre, mentre nei sottotitoli forniti dall'autore sono espressi in parole.

La seconda anomalia consiste nella tendenza dei sottotitoli automatici a troncare le parole, generando termini incompleti o, in alcuni casi, persino singole lettere isolate.

6 Conclusioni

L'obiettivo di questo progetto era analizzare il sistema di generazione automatica di sottotitoli in italiano su YouTube. Questa analisi è stata condotta confrontando i testi generati automaticamente con quelli scaricati come sottotitoli forniti dall'autore, considerati il gold standard. Le principali differenze emerse riguardano: la completa assenza di punteggiatura nei sottotitoli automatici; la gestione delle parole in lingua straniera, rilevate dal sistema automatico ma assenti nei sottotitoli tradotti in italiano; i numerali, rappresentati in cifre nei sottotitoli automatici e in parole in quelli forniti dall'autore; e infine, l'incapacità del sistema automatico di riconoscere correttamente alcune parole, spesso troncate o ridotte a singole lettere.

I codici sviluppati per la prima e l'ultima

fase del progetto sono il risultato di una collaborazione costante tra le due partecipanti. Successivamente, l'elaborazione dei sottotitoli è stata suddivisa in base alla tipologia: Silvia si è occupata di quelli forniti dall'autore, mentre Chiara di quelli generati automaticamente. Nonostante la divisione del lavoro, il confronto tra le due studentesse è stato continuo e produttivo. Per maggiore praticità, tutti gli elementi di codice sono stati infine riuniti e combinati in un unico notebook.

Questo progetto ha permesso di esplorare le capacità e i limiti del sistema automatico di sottotitolazione di YouTube, mettendo in luce aspetti che potrebbero essere migliorati per raggiungere una maggiore accuratezza e affidabilità. Il lavoro svolto rappresenta un passo verso una riflessione più ampia sull'applicazione del Natural Language Processing in contesti reali e concreti.

7 Bibliografia

1. yt-dlp: https://github.com/yt-dlp/yt-dlp
2. ffmpeg: https://ffmpeg.org/documentation.html
3. os: https://docs.python.org/3/library/os.html
4. datetime: https://docs.python.org/3/library/datetime.html
5. random: https://docs.python.org/3.11/library/random.html
6. youtube-transcript-api: https://pypi.org/project/youtube-transcript-api/
7. subprocess: https://docs.python.org/3.11/library/subprocess.html
8. re: https://docs.python.org/3.11/library/re.html
9. counter: https://docs.python.org/3.11/library/collections.html#collections.Counter
10. string: https://docs.python.org/3.11/library/string.html
11. nltk: https://github.com/nltk/nltk
12. unicodedata: https://docs.python.org/3.11/library/unicodedata.html
13. stanza: https://stanfordnlp.github.io/stanza/neural_pipeline.html
14. numpy: https://numpy.org/doc/
15. sys: https://docs.python.org/3.11/library/sys.html
16. scipy: https://docs.scipy.org/doc/scipy-1.10.1/
17. gensim: https://pypi.org/project/gensim/
18. scikit-learn: https://scikit-learn.org/1.2/
19. statsmodels: https://www.statsmodels.org/v0.13.5/
20. numba: https://numba.readthedocs.io/en/stable/release-notes.html
21. scikit-image: https://scikit-image.org/docs/0.20.x/
22. plotly: https://plotly.com/python/
23. pandas: https://pandas.pydata.org/pandas-docs/version/1.5.3/