Sottotitolazione e doppiaggio intelligente tramite tecniche di intelligenza artificiale

Marco Turchi[*], **Matteo Negri**[*], **Daniele Falavigna**[*], **Sébastien Bratières**[+]
Fondazione Bruno Kessler ([*]), Translated.com ([+])
{turchi, negri, falavigna}@fbk.eu, sebastien@translated.com

Abstract

Il traffico quotidianamente generato dalla visione di contenuti audio/video (A/V) detiene ormai stabilmente la quota più alta nel consumo di banda tra le attività svolte tramite Internet e, secondo le previsioni, questa quota aumenterà nei prossimi anni raggiungendo entro il 2021 l'82% di tutto il traffico online. All'aumento della produzione di contenuti A/V corrisponde inevitabilmente un aumento della richiesta di servizi di sottotitolazione e doppiaggio in altre lingue per rendere i contenuti originali fruibili a target linguistici diversi da quello sorgente. L'ingente quantitativo di dati da processare richiede l'utilizzo di tecniche automatiche basate su intelligenza artificiale (e.g. modelli sequence-to-sequence) capaci di sottotitolare e doppiare in breve tempo film o video generati dagli utenti. Lo sviluppo di tali sistemi automatici presenta nuove sfide scientifiche che richiedono, tra le altre, la corretta conversione di un segnale audio in una lingua direttamente in un testo o un segnale audio in un'altra lingua, l'adattamento dinamico allo stile dei parlanti e la sincronizzazione con il video. Questo articolo presenta le sfide proposte all'interno del progetto Smart Subtitling and Dubbing System per la realizzazione di un sistema per la generazione automatica di sottotitoli e doppiaggio.

1 Introduzione

L'aumento della banda di connessione, il perfezionamento di tecnologie esistenti, l'ingresso di nuove tecnologie e prodotti dedicati alla registrazione e alla condivisione di contenuti multimediali, sono solo alcuni dei fattori che stanno contribuendo alla rapida crescita della produzione di contenuti audiovisivi. Questa crescita è naturalmente sostenuta da una domanda in continuo aumento. Analisi[1] effettuate sulla distribuzione del tempo medio per giorno speso su diverse attività digitali evidenziano che, dal 2011 al 2015, il tempo impiegato nella visualizzazione di video digitali è passato da 39 minuti a 115 minuti (+195%), facendo registrare non solo la crescita più veloce, ma anche il valore più alto, se confrontato ad altri servizi dello stesso settore (e.g. il tempo passato su Facebook). Un riscontro evidente della crescita della domanda e dell'offerta di contenuti audiovisivi lo si può trovare nella crescita del numero di aziende che ne veicolano il consumo su internet (e.g. YouTube o Netflix). Questo incremento sia della domanda che dei contenuti deve pero' fronteggiare la capacita' di scalare il contenuto ad un pubblico internazionale. Le difficoltà che le aziende distributrici di questi servizi stanno affrontando per localizzare i propri contenuti sono perlopiù derivanti dalla scarsa capacità di assorbimento della domanda da parte delle aziende di localizzazione stesse, che si avvalgono di tecnologie e processi tradizionali per rispondere a necessità sempre più complesse da gestire. I requisiti necessari per poter sostenere questa crescita sono sempre più stringenti: il mercato, per questo genere di contenuti, richiede infatti localizzazione di elevata qualità e tempistiche ristrette.

Il progetto Smart Subtitling and Dubbing System (SSDS) finanziato all'interno del bando Creatività2020 da LazioInnova si occupa di sviluppare un prototipo per la generazione di sottotitoli e doppiaggio automatico basato su intelligenza artificiale, con lo scopo di abbattere i tempi necessari per la localizzazione e di offrire, ad un costo competitivo, un'esperienza migliore al cliente rispetto alle tecnologie tradizionali. Questo è ottenuto attraverso la realizzazione di un sistema automatico capace di prendere in ingresso un segmento audio e di generare automaticamente o la traduzione testuale (sottotitolo) o un segmento audio contenente la vocalizzazione della traduzione automatica (doppiaggio).

2 Stato dell'arte prima del progetto

La creazione di un sistema capace di tradurre un segnale audio in testo o in un altro segnale audio è un problema ben noto in letteratura [Cettolo *et al.*, 2014]. I principali approcci si sono basati sulla concatenazione dei singoli componenti [Zhang e et al., 2004], ovvero un riconoscitore del parlato, un traduttore automatico e, se necessario, un sintetizzatore vocale. Il segmento audio viene passato al riconoscitore che produce un testo nella stessa lingua del parlato in ingresso. Tale testo è poi tradotto nella lingua di arrivo e il risultato passato ad un sintetizzatore. Questo approccio chiamato "a cascata" è tuttora uno dei più usati, in quanto permette di sfruttare al meglio

[1] https://contently.com/strategist/2015/07/06/the-explosive-growth-of-online-video-in-5-charts/

i) la tecniche più avanzate di ogni singolo componente e ii) le grosse quantità di dati disponibili per ogni settore.

L'avvento delle tecniche basate su intelligenze artificiale ha fondamentalmente rimpiazzato i vecchi modelli statistici con approcci neuronali, capaci di generalizzare meglio e produrre risultati di più alta qualità. Tali modelli sono noti come modelli sequence-to-sequence basati su una architettura encoder-decoder [Bahdanau *et al.*, 2014], i quali processano il segnale in ingresso trasformandolo in un vettore numerico che viene usato dal decoder per produrre l'output richiesto. Generalmente vengono identificati con il termine "end-to-end", in quanto un solo componente è addestrato per risolvere uno specifico task.

Per quanto riguarda i sistemi automatici di riconoscimento vocale, l'attuale stato dell'arte [Saon e et al., 2017] fa uso sia di modelli ibridi, attraverso i quali modello del linguaggio e modello acustico vengono stimati indipendentemente utilizzando insiemi di dati disgiunti, che di modelli "end-to-end" con cui si può stimare una distribuzione di probabilità congiunta (audio e corrispondente testo). La tendenza della ricerca attuale è quella di investigare maggiormente l'uso di sistemi "end-to-end". Per quanto riguarda i sistemi di traduzione automatica, i modelli neurali "end-to-end" [Cho *et al.*, 2014] hanno repentinamente rimpiazzato i vecchi modelli a frasi. Recentemente, una nuova architettura neurale chiamata Trasformer [Vaswani e et al., 2017], basata sul concetto del self-attention, ha ottenuto i migliori risultate in molte campagne di valutazione mostrando notevoli miglioramenti sia dal punto di vista della qualità che dell'efficienza. Per quanto riguarda la sintesi vocale, i principali sistemi allo stato dell'arte usano approcci neurali [Wang e et al., 2018]. Tali approcci richiedono molte ore di dati di addestramento, dove l'audio è registrato da attori in studi di registrazione. Questi sistemi, seppur usati, tendono a produrre un audio con bassa espressività, limitata prosodia e stile della voce.

3 Problemi aperti

I modelli "a cascata", seppur molto usati, sono affetti da due principali problematiche. La prima è dovuta alla propagazione dell'errore da un componente all'altro che può creare un effetto a catena producendo una traduzione completamente diversa in significato dall'audio in ingresso. Il secondo è legato al fatto che i sistemi di traduzione automatica e di sintesi della voce prendo in ingresso del testo che è stato privato di informazioni utili come la prosodia o aspetti caratteristici del parlatore, che sono state perse durante la procedura di riconoscimento del parlato.

Per sopperire a queste problematiche e produrre dei sottotitoli e doppiaggi realmente utili all'industria, il progetto SSDS si occupa di sviluppare nuove tecnologie basata sull'intelligenza artificiale fatte su misure per questo mercato. Le principali sfide del progetto sono:

Architetture "end-to-end": la propagazione dell'errore tra i vari componenti è la principale fonte di errore. Per eliminare questo problema, prendendo vantaggio dalla adattabilità dei sistemi sequence-to-sequence, il progetto svilupperà dei modelli che partendo dall'audio in ingresso, direttamente produrranno del testo o dell'audio nella lingua di arrivo. Questo richiede sia il cambiamento delle strategie degli attuali modelli sequence-to-sequence che la necessita' di sopperire alla carenza di dati per l'addestramento.

Adattamento: nel settore del doppiaggio, un ruolo di cruciale importanza è quello del direttore del doppiaggio, il quale ha il compito di gestire l'adattamento della traduzione (audio e testuale) alla lunghezza della porzione del video dove un certo personaggio appare, al labiale e alla voce del parlatore e ad aspetti socio-culturali di una particolare nazione dove un film o un video viene ambientato. I sistemi automatici generano audio e testo che sono fedeli al segnale di ingresso, ma mancano completamente di questi aspetti di adattamento. Il progetto si focalizza sullo sviluppo di tecniche che permettono ai sistemi "end-to-end" di adattarsi velocemente con limitate quantità di dati, cercando di imitare il più possibile le principali caratteristiche del parlatore e del contesto originale.

Alla fine del progetto, i componenti sviluppati saranno testati grazie al supporto della Società Edizioni Italiane Film, la quale valuterà i contributi del progetto all'interno della sua catena di produzione.

4 Conclusioni

Questo breve articolo mostra le principali sfide che si presentano nell'applicare tecniche di intelligenza artificiale al mondo del sottotitolaggio e del doppiaggio. Il progetto ambisce ad automatizzare il lavoro dei traduttori e doppiatori professionisti, al fine di ridurre i tempi e sopperire alla sempre più pressante richiesta del mercato dei contenuti audiovisivi.

Riferimenti bibliografici

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, e Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Cettolo *et al.*, 2014] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, e Marcello Federico. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of IWSLT*, page 57, 2014.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, e Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[Saon e et al., 2017] George Saon e et al. English conversational telephone speech recognition by humans and machines. In *Proceedings of Interspeech*, pages 132–136, 2017.

[Vaswani e et al., 2017] Ashish Vaswani e et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 2017.

[Wang e et al., 2018] Yuxuan Wang e et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *CoRR*, abs/1803.09017, 2018.

[Zhang e et al., 2004] Ruiqiang Zhang e et al. A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation. In *Proceedings of COLING*, page 1168, 2004.