Machine Learning per la generazione automatica dei sottotitoli

Mauro **Rossini**[1], Carmen **Marino**[1], Mauro **Cettolo**[2], Leonardo **Badino**[2], Riccardo **Lucianer**[2]
[1] **Rai -** Centro Ricerche, Innovazione Tecnologica e Sperimentazione
[2] **PerVoice**

L' Intelligenza Artificiale e le tecnologie di Machine Learning per la Trascrizione Automatica permettono di ipotizzare nuovi scenari applicativi rivolti a servizi per le persone con disabilità cognitive e sensoriali. I contesti dove l'Intelligenza Artificiale incontra le necessità e aspirazioni di un settore di pubblico che desidera esplorare nuove soluzioni per migliorare la propria identità sociale diventano un terreno di sfida per le aziende che erogano servizi al cittadino. In questi scenari si incontrano le richieste per un incremento di contenuti accessibili e comprensibili per tutti a cui il servizio pubblico deve rispondere con innovazione tecnologica e investimenti sulla ricerca.

La Rai ha voluto accettare questa sfida tecnologica nel settore dei Servizi per le persone con disabilità sensoriali, rispondendo alla forte richiesta di incremento delle ore sottotitolate per le persone sorde, ipoacusiche e anziane.

Il Centro Ricerche Innovazione Tecnologica e Sperimentazione ha attivato una sperimentazione con PerVoice sulla Sottotitolazione Automatica dei Telegiornali Regionali, un progetto innovativo, sfidante e sicuramente non privo di complessità. L'analisi delle prestazioni del sistema di sottotitolazione automatica, seppur richiedendo un minimo intervento di correzione manuale prima della messa in onda, ha evidenziato elevati valori di accuratezza sulle parole trascritte. L'adozione di tale soluzione tecnologica, che prevede l'impiego nei processi di validazione/correzione di figure professionali senza specifiche competenze di stenotipia, potrebbe ridurre i costi del servizio di sottotitolazione e garantire la sostenibilità del progetto.

L'evoluzione dell'Intelligenza Artificiale nel campo dei Servizi Televisivi rivolti alle Persone con disabilità sensoriali e cognitive

L' evoluzione dell'*Intelligenza Artificiale* ha portato negli ultimi anni allo sviluppo e al miglioramento di nuovi servizi e tecnologie in vari contesti applicativi, tra questi il contesto televisivo con l'erogazione di servizi rivolti alle persone con disabilità cognitive e sensoriali.

Machine Learning e Reti Neurali

Il *machine learning (ML)*, o *apprendimento automatico*, è una branca dell'*Intelligenza Artificiale* che studia come le macchine possano apprendere dai dati.

Esistono differenti aree del *machine learning*, due delle quali, il *supervised learning* ed il *reinforcement learning*, nell'ultimo decennio hanno ottenuto risultati veramente sorprendenti, portando i calcolatori ad eguagliare, se non addirittura a superare, gli esseri umani in alcuni specifici compiti, quali il riconoscimento di oggetti, il riconoscimento del parlato o la capacità di giocare ad un gioco complesso come *GO*. Il presente articolo si sofferma sul supervised learning che è alla base dei sistemi allo stato dell'arte del riconoscimento del parlato e della traduzione automatica.

Nel supervised learning, al sistema di apprendimento vengono presentati *esempi* in cui ad un *segnale di input* (per esempio, un'immagine di un'automobile) è associato un *output desiderato* (ad esempio, la categoria *automobile*). Dati questi esempi, il sistema deve imparare la relazione tra input e output, per cui a ciascun input deve associare il corretto output. Da notare che questa non è una semplice operazione di memorizzazione in quanto il sistema deve essere in grado di *generalizzare*, ovvero di eseguire la corretta associazione dell'output anche ad un input che non ha mai visto nei dati di addestramento. Per esempio, deve essere in grado di classificare come *automobile* l'immagine di un'auto arancione anche se nei dati di addestramento non c'erano automobili arancioni.

Nella traduzione automatica, un esempio di addestramento è dato dalla frase nella lingua di origine (input) e la stessa frase tradotta nella lingua target (output). Nel riconoscimento automatico del parlato (o *speech-to-text*), l'input dell'esempio è il file audio, mentre l'output è la trascrizione testuale dell'audio.

L'enorme successo osservato a partire dagli anni '10 del nuovo millennio del *machine learning* si deve alla (re)introduzione di una particolare famiglia di tecniche di addestramento, le cosiddette *reti neurali*, in combinazione con la disponibilità di enormi dataset di esempi di addestramento e di più potenti unità di calcolo in parallelo (*Graphical Processing Units*, *GPUs*).

Oggi l'approccio basato su reti neurali è nettamente prevalente nel mondo scientifico del *machine learning* a giudicare dal numero di pubblicazioni che lo riguardano. Non è questa la sede per descrivere in maniera approfondita un argomento tanto complesso quanto quello dei modelli neurali. Nondimeno, nei paragrafi che seguono si cercherà di fornire un'idea, per quanto semplificata e pertanto qua e là necessariamente imprecisa, di come funziona una rete neurale.

Innanzitutto, vediamo come lavora un neurone, il componente base delle reti neurali. Un neurone (artificiale) consta (Fig. 1):

a) di un certo numero di *ingressi* che permettono al mondo esterno di fornire dei valori numerici (stimoli, X_i nella figura);
b) dei *pesi* (anch'essi numeri) associati a ciascun ingresso (W_i nella figura);
c) di un *aggregatore* che combina in un unico numero (risposta) i valori in ingresso pesati.

Una rete neurale è semplicemente un insieme di neuroni connessi tra di loro (Fig. 2). L'addestramento ha l'obiettivo di assegnare ai pesi dei valori tali che la rete si comporti come voluto su un insieme di coppie <*stimoli, risposta*>.

Tipicamente le reti neurali sono organizzate a strati: ogni strato contiene dei neuroni che ricevono gli stimoli dallo strato precedente e inviano le loro risposte allo strato successivo. Il primo strato (*di ingresso*) riceve gli stimoli dal mondo esterno, l'ultimo strato (*di uscita*) fornisce la risposta complessiva della rete al mondo esterno.

Fig. 1 – Schema funzionale di un neurone artificiale.

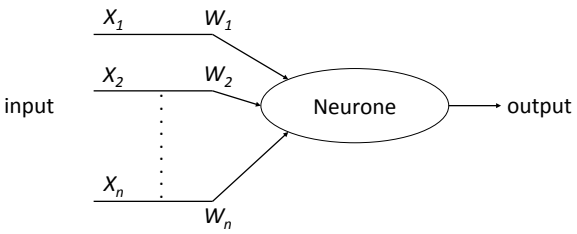
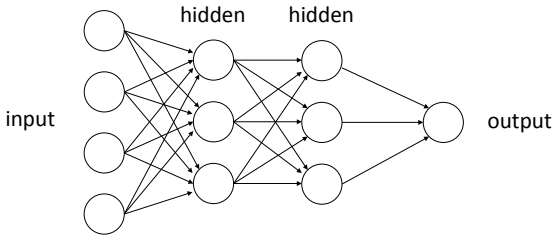


Fig. 2 – Rete neurale artificiale multistrato.



Una semplice rete costituita da un unico neurone potrebbe, ad esempio, decidere se l'atmosfera di un locale è adatta al mantenimento delle qualità organolettiche di un prodotto alimentare: gli ingressi potrebbero essere i valori di temperatura, umidità, quantità di ossigeno e di anidride carbonica presenti nell'aria, misurati da sensori appositi. L'uscita potrebbe essere il valore *1* se l'atmosfera è adatta alla conservazione dell'alimento, *0* se non lo è. Un set di addestramento dovrebbe contenere un certo insieme di stimoli a cui la rete deve rispondere col valore *1* e un altro insieme a cui la rete deve rispondere col valore *0*.

Automatic Speech-to-Text

L' *automatic speech-to-text* (*STT*, o anche detto *automatic speech recognition*, *ASR*) è la tecnologia che permette di trascrivere il parlato in testo. Un *sistema STT* è la componente principale in applicazioni quali dettatura automatica (ad es., dettatura di SMS o di referti medici), trascrizioni di meeting o sottotitolazione automatica. In applicazioni più complesse, quali call center automatico o assistente virtuale (ad esempio, *Amazon Alexa* o *Google Now*), il sistema STT viene combinato con un sistema di comprensione automatica del parlato (*spoken language understanding*, *SLU*) che permette di identificare nel testo l'esatta richiesta dell'utente.

La gran parte dei sistemi commerciali STT sono *speaker-independent*, ossia trascrivono il parlato di qualsiasi parlatore entro un certo range di accuratezza, al contrario dei sistemi *speaker-dependent* che sono in grado di riconoscere accuratamente solo il parlato di uno specifico parlatore.

Tipicamente l'STT opera in una di due modalità: *live* (detta anche *real time* o *online*) o *batch* (ossia *offline*). Nella modalità online il parlato viene trascritto in tempo reale, con una tolleranza di poche centinaia di millisecondi tra quando una parola viene detta e quando viene trascritta. Nella modalità offline non esiste tale vincolo temporale (o almeno è molto meno stringente) e il sistema STT può sfruttare il maggior tempo a disposizione per produrre una trascrizione più accurata che in modalità offline.

Un po' di storia

I primi sistemi STT furono sviluppati negli anni '50 ed erano in grado di riconoscere un numero molto limitato di parole, come il sistema sviluppato nei *Bell Labs* che era in grado di riconoscere le dieci cifre del sistema numerico decimale in contesti controllati (le cifre dovevano essere scandite introducendo un silenzio tra una cifra e la successiva).

Questi sistemi si basavano sul fatto che i suoni base hanno, in media, dei pattern tipicamente riconoscibili, specialmente in una rappresentazione *tempo-frequenza* (anche detta *spettro-temporale*). Per esempio, vocali diverse (ad es.. */a/* vs. */o/*) sono caratterizzate da un diverso pattern delle *formanti*, dove le formanti sono le regioni in tempo-frequenza dove il segnale ha piu' energia (Fig. 3).

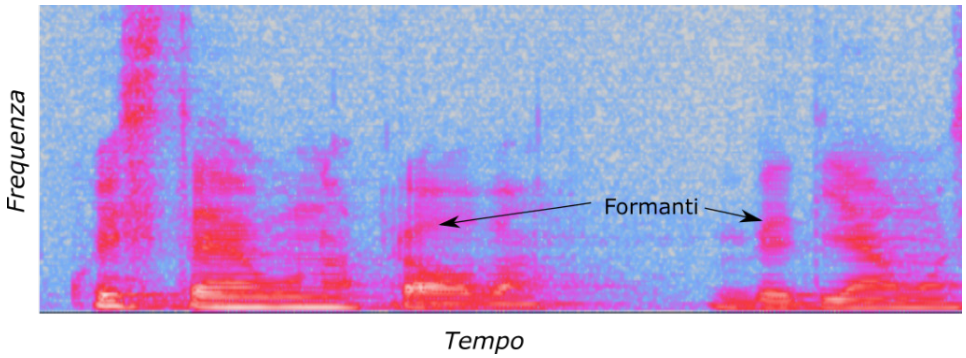


Fig. 3 – Esempio di spettrogramma di segnale vocale ed esempi di alcune formanti.

Per identificare le cifre numeriche, per esempio, gli algoritmi di riconoscimento cercavano di identificare le vocali delle cifre confrontando le formanti in input con dei *template*, ossia pattern predefiniti e rappresentativi di ciascuna vocale.

Questi approcci basati su *template matching*, oltre a essere limitati al riconoscimento di poche parole, presentavano diverse limitazioni, principalmente dovute al fatto che i template possono variare considerevolmente tra parlatori diversi e ancor di più a seconda del contesto fonetico in cui si trovano. Quest'ultimo aspetto, chiamato *co-articolazione*, fa sì che uno stesso suono, per esempio una vocale */a/* possa avere un pattern ben diverso a seconda che sia preceduta, ad esempio, dalla vocale */u/* (come nella parola *quando*) piuttosto che dalla consonante */n/* (come in *nave*). Questa differenza non viene *percepita* dal nostro cervello (che la annulla) quando ascolta le parole o le frasi, ma in realtà se si estraggono i segmenti audio di una stessa vocale da contesti diversi essi *suoneranno* molto diversi.

Un'ulteriore complicazione per quei sistemi di riconoscimento era la difficoltà a gestire la variabilità della durata di un suono: una stessa parola pronunciata da uno stesso parlatore in momenti diversi può avere durate molto diverse. Questa complicazione rendeva difficile il confronto con i pattern spettro-temporali di riferimento. Per ovviare a questo limite negli anni '60 venne introdotta la tecnica del *dynamic time warping* che rimase in auge sino agli anni '70.

Negli anni '70 i centri di ricerca degli *AT&T Bell Labs* e di *IBM* introdussero due elementi che sono alla base di gran parte dei sistemi di riconoscimento tutt'ora in uso. In entrambi i casi venne utilizzato ed avanzato un *approccio statistico* e *data-driven* al problema del riconoscimento del parlato.

IBM sviluppò il concetto di *n-grams* per i modelli del linguaggio. Un modello del linguaggio modella la struttura del linguaggio di una data lingua e stima la plausibiltà di una frase. Mentre in precedenti approcci la plausibilità di una frase poteva essere calcolata sulla base di regole sintattiche, IBM introdusse gli n-grams per stimare la plausibilità di una frase in forma di probabilità (la frase "il mio cane abbaia" è molto più probabile di "il mio cane cucina") calcolando da dataset testuali la probabilità di occorrenza di una parola ("abbaia") date le precedenti n-1 parole ("il mio cane").

Gli AT&T Bell Labs invece introdussero il framework degli *Hidden Markov Models* (*HMMs*) combinati con *modelli Gaussiani* (*HMM-GMMs*) per il *modellamento acustico*. I modelli acustici hanno la funzione di assegnare a pezzetti di segnale acustico il fonema corrispondente (ad es., */a/*) (e le tecniche di template matching descritte sopra sono da considerarsi un primo approccio al modellamento acustico) combinando l'evidenza acustica (ossia le features acustiche in input) con conoscenza a priori riguardo la durata dei fonemi (estratta dai dati di addestramento). L'approccio HMM-GMM consentiva, rispetto agli approcci precedenti, di modellare molto più accuratamente la grande variabilità del segnale del parlato dovuta a fattori citati sopra (come differenze tra parlatori, coarticolazione, accenti diversi, ecc.).

L'uso di n-grams ed HMMs si è evoluta negli anni '80 e '90 permettendo, ad esempio, di creare accuratissimi sistemi di dettatura speaker-dependent e raggiunse un plateau nel primo decennio del terzo millennio.

Gli anni '90 videro l'introduzione delle reti neurali per i modelli acustici e poco più tardi per i modelli del linguaggio. Tuttavia i computer utilizzati per l'addestramento dei modelli e la quantità di dati di addestramento non erano ancora sufficienti per sfruttare al meglio le enormi potenzialità delle reti neurali e le stesse reti neurali necessitavano di qualche perfezionamento. Ed è alla fine del primo decennio di questo millennio che le reti neurali (ri-)entrano in scena nel mondo delle tecnologie vocali e del linguaggio, portando in 4-5 anni a risultati che si pensava si sarebbero potuti raggiungere solo in 10-20 anni. Questi risultati hanno portato ad un forte consolidamento di applicazioni già esistenti ma non sempre affidabili ed al rapido sviluppo di assistenti virtuali.

Sistemi Speech-to-Text e Reti Neurali

Un tipico sistema di riconoscimento vocale ha cinque moduli principali, rappresentati nella Fig. 4 (Sistema STT **PerVoice**).

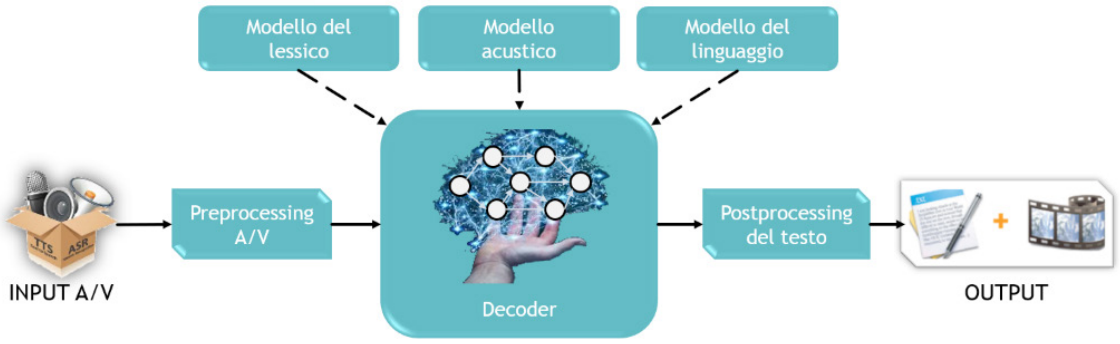
- *Modulo di pre-processamento del segnale acustico*. Questo modulo processa il segnale acustico in ingresso, separa segmenti di parlato da segmenti di silenzio o rumore ed estrae features acustiche spettrali dal segnale acustico;
- *Decoder*. Converte il segnale pre-processato in testo. Il decoder è basato su algoritmi di programmazione dinamica che attingono all'informazione combinata delle 3 componenti descritte di seguito: modello lessicale, acustico e del linguaggio;
- *Modello lessicale*. Si tratta di un dizionario di trascrizioni fonetiche. Tipicamente consiste di centinaia di migliaia di parole e alcune parole possono avere più di una trascrizione per gestire le varianti di accenti diversi;
- *Modello acustico*. Il modello acustico modella statisticamente il segnale spettro-temporale prodotto da una sequenza di fonemi e più in generale di parole. Ha la funzione di segmentare in pezzetti il segnale e di assegnare a ciascuno di essi il fonema corrispondente (anche se più che il fonema vengono considerate unità più piccole, parti di fonema dipendenti dal contesto dette *senoni*).
- *Modello del linguaggio*. Modella il contesto in cui ciascuna parola può tipicamente trovarsi in una determinata lingua. Stima la verosimiglianza di una sequenza di parole e quindi la probabilità di occorrenza di una parola date le precedenti.

In questa architettura le reti neurali (deep) svolgono un ruolo fondamentale in due moduli: il *modello acustico* e quello *del linguaggio*.

Per il modello acustico le reti neurali hanno soppiantato i modelli Gaussiani e superato qualsiasi altra tecnica di *machine learning* nel classificare fonemi o parti di fonemi dato il segnale acustico, anche in condizioni severe di rumorosità. Ma non è solo questo il segreto del loro successo. Come detto, le reti neurali non sono solo dei classificatori (o dei regressori) ma anche degli eccezionali estrattori di informazioni, dove queste informazioni vengono immagazzinate in *feature gerarchiche multilivello*. Queste feature possono essere manipolate in modo semplice ed efficace per adattarsi quasi istantaneamente alle caratteristiche timbriche di un parlatore o ad un rumore di sottofondo. Oppure possono essere riutilizzate per compiere task affini: una rete neurale per un modello acustico dell'italiano può essere utilizzata e riadattata per velocizzare la creazione di un modello acustico dello spagnolo.

Per i modelli del linguaggio le reti neurali, soprattutto quelle ricorrenti, si sono rivelate più efficaci

Fig. 4 – Componenti principali del sistema di riconoscimento del parlato di **PerVoice**. L'input al sistema può essere non solo audio ma anche video con audio.



delle precedenti tecniche per stimare gli n-grams e quindi per stimare la probabilità di una parola date le (n-1) precedenti. E questo, di nuovo, per la loro capacità di estrarre informazioni strutturate, che in questo caso si riflette nella loro capacità di estrarre il grado di similitudine (o diversità) tra parole. Dato un grande dataset testuale di addestramento, una rete neurale, addestrata per predire una parola date le precedenti, è in grado di riconoscere che le parole *mela* e *pera* sono molto più simili tra loro rispetto alla parola *macchina* e quindi hanno molti contesti condivisi ("*Oggi ho mangiato una mela/pera*"). Questa capacità di estrarre informazioni strutturate dai dati può essere visualizzata guardando al suo interno, ai suoi neuroni, con tecniche che consentono la visualizzazione 2D o 3D a partire da migliaia di neuroni.

In questi ultimi anni, nella comunità scientifica del riconoscimento vocale c'è una forte tendenza verso lo sviluppo di *sistemi end-to-end*, ossia sistemi in cui tutto il processo che va dal segnale acustico alla trascrizione testuale è gestito da un'unica rete neurale.

I sistemi di Traduzione Multilingua

Assumendo di sapere cosa significa *tradurre*, con *traduzione automatica* (*TA*) si intende la *traduzione eseguita dalle macchine*. Come il termine *traduzione* indica sia l'atto del tradurre sia il testo tradotto, anche la locuzione *traduzione automatica* possiede il doppio significato.

Un po' di storia

Le prime idee per meccanizzare la traduzione apparvero già nel XVII secolo, quando sia Leibniz sia Cartesio formularono al riguardo delle teorie basate su una lingua universale. Alla fine dell'800 furono definiti degli schemi di lingue internazionali, quali l'esperanto, i cui simboli logici furono utilizzati nella macchina traduttrice brevettata nel 1933 dal russo Pëtr Smirnov-Trojanskij; nello stesso anno, in maniera indipendente, anche il franco-armeno Georges Artsrouni depositò il brevetto di un traduttore meccanico.

Se proprio si vuol fissare una data precisa per la nascita della *TA* moderna, allora il 4 marzo 1947 è un buon candidato: quel giorno Warren Weaver, direttore della Divisione Scienze Naturali della Fondazione Rockefeller, scrisse al cibernetico Norbert Wiener:

> *"...mi chiedo se non sia possibile progettare un calcolatore che traduca [...] se il problema della traduzione non possa essere concettualmente trattato come un problema di crittografia. Quando osservo un articolo scritto in russo, dico 'È scritto in inglese, ma codificato in strani simboli'..."*

Nel 1954 venne inscenata una dimostrazione pubblica di un traduttore automatico, frutto del lavoro congiunto dell'*Università di Georgetown* e dell'*IBM*, che ricevette notevole eco sulla stampa (Fig. 5).

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

Fig. 5 – Esperimento Università di Georgetown-IBM del 1954: un'operatrice scrive le frasi in russo su schede perforate che poi vengono inserite nel traduttore automatico per ottenere la traduzione inglese.

Da allora, la ricerca sulla *TA* ha continuato a progredire senza soluzione di continuità, sebbene si siano alternati momenti di eccitazione derivati dall'introduzione di metodologie nuove e dirompenti a periodi più o meno lunghi di stasi. Negli anni '50 e '60 si indagarono principalmente due approcci, quello empirico basato su regole per la traduzione diretta tra due lingue e quello più teorico che si concentrò sui fondamenti linguistici della traduzione. Negli anni '70 comparvero i primi sistemi che fornivano un minimo di qualità in contesti specifici ed erano basati su regole linguistiche (*Systran*).

Negli anni '80 fece capolino la statistica, all'inizio solo timidamente in sistemi che recuperavano da basi di dati esempi di traduzioni precedenti. Nei primi anni '90, scienziati dei laboratori IBM pubblicarono degli articoli in cui definivano formalmente l'applicazione alla *TA* di modelli e metodi statistici fino ad allora in uso nel riconoscimento del parlato. Queste basi teoriche, insieme alla disponibilità per la prima volta nella storia di grandi quantità di dati in formato elettronico, quindi adatto al calcolo computerizzato delle statistiche, e di calcolatori sempre più potenti, permisero lo sviluppo di sistemi di traduzione la cui qualità cominciava finalmente ad avvicinarsi alle aspettative dei pionieri della *TA*. Nell'aprile del 2006 *Google* lanciò il suo famoso traduttore che in quella prima versione era basato proprio sui fondamenti teorizzati in quegli articoli una dozzina d'anni prima.

A metà della prima decade del nuovo millennio, grazie all'ulteriore progresso dell'hardware che ha permesso la realizzazione di GPU ad alto parallelismo e l'elaborazione general purpose su di esse, la *TA* è stata oggetto di una vera e propria rivoluzione data la possibilità di sfruttare il *deep learning* (apprendimento profondo), ovvero delle reti neurali artificiali. Il costituente di base delle reti neurali artificiali, il neurone, fu proposto nel 1943 da Walter Pitts, mentre negli anni '50 e '60 esse furono oggetto di approfondimenti teorici. Risale al 1974 la tesi di dottorato di Paul Werbos nella quale venne formulato l'algoritmo ancor oggi utilizzato per l'addestramento delle reti neurali multistrato (*profonde*).

È quindi curioso notare come le reti neurali artificiali, nonostante siano conosciute e utilizzate in ambito scientifico da oltre mezzo secolo, solo oggi siano entrate a far parte del vocabolario comune grazie ai notevoli risultati ottenuti in diversi settori applicativi (non solo l'elaborazione del linguaggio naturale ma anche il riconoscimento delle immagini, la bioinformatica, la diagnosi medica, la guida autonoma...) resi possibili dal progresso tecnologico. Comunque sia, esse hanno permesso alla *TA* di ricevere sempre maggior interesse da parte dei media, grazie agli innegabili miglioramenti dei sistemi disponibili online. Mentre resta ancora molto da fare, soprattutto per coppie di lingue o per domini in cui ci sono poche risorse utili per l'addestramento, la qualità nei casi più favorevoli, come la traduzione di notiziari dall'inglese al francese o dal tedesco all'inglese, ha raggiunto livelli senza precedenti, portando alcuni entusiasti a sostenere che essa ha ormai raggiunto il livello delle traduzioni professionali.

Reti Neurali Artificiali e Traduzione Automatica

Rispetto alle reti neurali, nel caso dell'elaborazione del linguaggio parlato, un primo ostacolo da superare è che le parole sono entità simboliche, non numeriche. Questo problema è facilmente superato con l'uso di indici numerici associati univocamente a ciascuna parola.

Nel caso l'elaborazione riguardi frasi, come quando si vuole tradurre, una seconda difficoltà nasce dal fatto che sia lo stimolo in ingresso sia la risposta del sistema sono delle sequenze di parole di lunghezza che non può essere predeterminata. Una soluzione a questo problema è fornita dal cosiddetto *modello sequence-to-sequence* in cui spiccano due componenti principali: il *codificatore* (*encoder*) e il *decodificatore* (*decoder*). L'encoder è una rete multistrato che si occupa di codificare la frase da tradurre in ingresso in un valore unico, indipendentemente dalla sua lunghezza, accumulando iterativamente le informazioni fornite dalle singole parole, dalla prima all'ultima (un simbolo fittizio che indica la fine della frase). Quando la codifica della frase in ingresso è

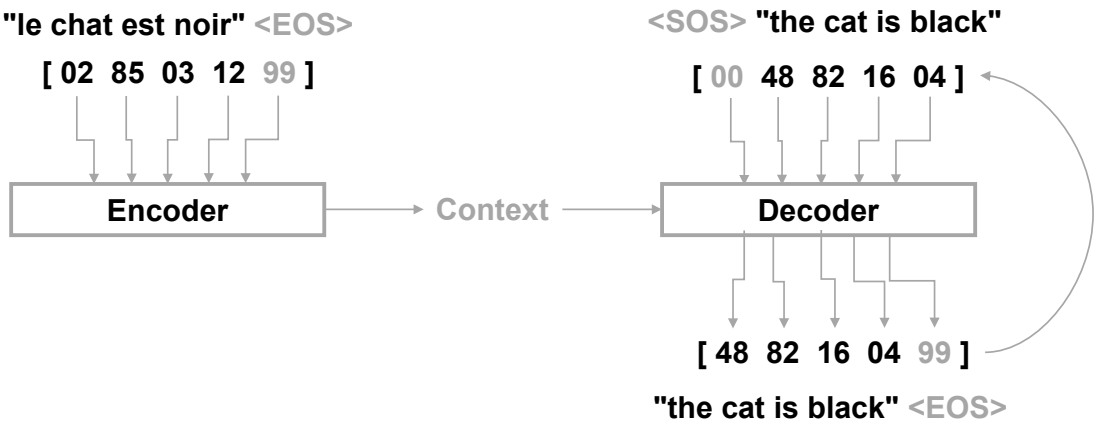
pronta, essa viene passata al decodificatore, un'altra rete multistrato, che inizia a generare una parola alla volta della traduzione, sulla base della codifica della frase da tradurre e di ciò che ha generato fin lì; la generazione termina nel momento in cui viene emesso il simbolo di fine frase (Fig. 6).

È quindi facilmente intuibile come questo meccanismo mal si adatti alle applicazioni di traduzione simultanea, ovvero quelle in cui le frasi da tradurre sono disponibili in maniera incrementale e si vorrebbe che analogamente il traduttore fornisse le traduzioni parziali, via via estese fino al loro completamento nel momento in cui anche l'ultima parola da tradurre è stata inserita. La difficoltà sta appunto nel fatto che il traduttore ha bisogno di codificare l'intera frase da tradurre prima di poter cominciare a generare la traduzione. Un'approssimazione usata oggigiorno in questi casi è quella di considerare le frasi parziali da tradurre come fossero frasi complete, di generare la loro traduzione automatica per poi sostituirla con le traduzioni successive delle frasi parziali a cui sono state aggiunte le nuove parole che nel frattempo si sono rese disponibili. Ma è appunto solo un modo piuttosto inefficiente di aggirare un limite intrinseco del modello sequence-to-sequence.

I sistemi di Auto Apprendimento

Abbiamo nei paragrafi precedenti già incontrato il concetto di apprendimento automatico e visto che esso si basa sulla capacità dei modelli di imparare dagli esempi. Tipicamente i modelli vengono addestrati sulla maggior quantità possibile di dati per aumentarne la capacità di generalizzare. È possibile però che in ambiti specifici questi modelli siano utilizzati su stimoli che hanno in comune certe caratteristiche: ad esempio, un solo parlatore con la sua specifica voce, un ambito linguistico delimitato e gergale (cronache sportive) o con vincoli esogeni (traduzione di manuali tecnici). In questi casi una successiva specializzazione dei modelli generali può consentire una loro maggior accuratezza nell'ambito specifico, al prezzo di una minor capacità di generalizzazione. Il processo di specializzazione, noto con il termine di *adattamento*, di un modello generale, addestrato su una grande quantità di dati, ad un ambito specifico per il quale è disponibile un set di dati di piccole dimensioni consiste sostanzialmente nel continuare l'addestramento sui dati specifici, prendendo precauzioni (ad esempio *metodi di regolarizzazione*) per evitare un eccesso di adattamento.

Fig. 6 – Modello sequence-to-sequence di una rete neurale artificiale.



L'adattamento può essere suddiviso in due categorie principali: *batch* e *live*. Queste due modalità si differenziano principalmente per la latenza con cui i modelli adattati sono disponibili. Nel caso batch, l'adattamento richiede più tempo, è più accurato, viene fatto su una quantità comunque significativa di dati e quindi il risultato è tipicamente ottimale. Nel caso live, al contrario, si cerca di adattare velocemente il modello durante il suo utilizzo in modo che la versione specializzata sia disponibile immediatamente.

Adattamento modelli SST

Nei sistemi del riconoscimento del parlato l'adattamento può interessare sia il modello acustico che quello del linguaggio.

Per l'adattamento batch del modello acustico possono essere sufficienti pochi minuti di parlato per ottenere un adattamento apprezzabile. Esistono molte tecniche di adattamento, da quelle che introducono nella rete neurale un *layer di adattamento* a quelle che operano un limitato riaddestramento sui dati di adattamento. Quest'ultimo approccio viene usato anche per l'adattamento di modelli del linguaggio, dove il numero di frasi testuali di adattamento può essere consistente (dell'ordine di decine di migliaia).

L'adattamento live (da non intendersi come real time) riguarda il modello acustico. Una tecnica efficace è quella in cui una *fingerprint acustica* del parlatore (nella forma di un vettore di lunghezza fissa di numeri reali) viene estratta da pochi secondi di parlato dello stesso (per esempio, per mezzo di un'altra rete neurale) e data come input aggiuntivo alla rete neurale del modello acustico. Ciò consente alla rete di adattarsi dopo pochi secondi alle caratteristiche timbriche del parlatore, con conseguente miglioramento dell'accuratezza del riconoscimento.

Adattamento modelli TA

Nel caso della *TA* basata su reti neurali, l'adattamento batch è fattibile nelle modalità ben note ed ereditate dal *machine learning*.

Prendendo i seguenti numeri come semplice ordine di grandezza, essendo numerose le variabili in gioco che possono determinare uno scostamento significativo in altri casi, possiamo dire che specializzare un modello di traduzione può richiedere qualche ora e va fatto su una quantità di dati che deve essere non inferiore a 100 volte quella usata per addestrare il modello generale di partenza.

L'adattamento live è invece intrinsecamente problematico proprio per la natura delle reti neurali: un modello di traduzione può contenere da decine fino a centinaia di milioni di parametri il cui valore viene assegnato durante la fase di apprendimento; è ingenuo pensare di poterli modificare velocemente e in maniera pertinente su pochi dati.

Recentemente, però, *ModernMT*, il cui pacchetto software per lo sviluppo di motori di traduzione è stato anche per questo adottato da **PerVoice**, ha sviluppato una tecnica per l'adattamento live. L'unicità e la novità di ModernMT consiste nella capacità di adattare la sua traduzione in tempo reale sul feedback umano e/o su nuovi dati, senza la necessità di riaddestrare il modello originale. Questo è possibile poiché ModernMT memorizza i dati di adattamento in *Memorie di Traduzione* che possono essere continuamente aggiornate con i nuovi contributi e che vengono utilizzate per *orientare* la traduzione generata dal modello originale.

L'Intelligenza Artificiale per la Sottotitolazione Automatica delle News

Le tecnologie di *Machine Learning* per la *Trascrizione Automatica* permettono di ipotizzare nuovi scenari applicativi rivolti essenzialmente a servizi per le persone con disabilità cognitiva e sensoriale. I contesti dove l'*Intelligenza Artificiale* incontra le necessità e le aspirazioni di un settore di pubblico che desidera esplorare nuove soluzioni per migliorare la propria identità sociale diventano un terreno di sfida per le aziende che erogano servizi al cittadino cercando di includere tutte le tipologie di pubblico a cui si rivolgono.

In questi scenari si incontrano le richieste per un incremento di contenuti accessibili e comprensibili per tutti a cui il servizio pubblico deve rispondere con innovazione tecnologica e investimenti sulla ricerca.

La **Rai** ha voluto accettare questa sfida tecnologica nel settore dei Servizi per le persone con disabilità sensoriali, rispondendo alla forte richiesta di incremento delle ore sottotitolate per le persone sorde, ipoacusiche e anziane.

La televisione è uno strumento sia di intrattenimento sia di approfondimento culturale e accompagna la persona nel suo percorso di socializzazione e inclusione nella vita quotidiana. La possibilità di incrementare le ore di trasmissioni televisive sottotitolate tramite l'adozione di soluzioni innovative che introducano tecnologie atte a produrre contenuti ad oggi non disponibili diventa un'opportunità di sperimentare sul campo le potenzialità di questi sistemi, seppur evidenziandone tutte le criticità ancora irrisolte.

Il mondo delle news è sicuramente il contesto più sfidante e riveste un ruolo di condivisione delle informazioni vitali in una società globalizzata. L'obbligo di diffondere un'informazione corretta diventa una sfida etica e contemporaneamente tecnologica, dove la *semantica*, ovvero la ricerca del significato del messaggio, riveste un ruolo fondamentale per la comprensione della notizia.

Per questo motivo, la trascrizione automatica e la successiva sottotitolazione applicata alle news non devono solamente focalizzarsi su riportare correttamente la parola ma devono polarizzarsi nel trasferire il corretto messaggio che si intende veicolare.

In questo ambito è stata attivata una sperimentazione **Rai** sulla *Sottotitolazione Automatica delle news regionali*, un progetto innovativo, sfidante e sicuramente non privo di complessità.

Il Progetto RAI

Il nuovo *Contratto di Servizio Rai* [Nota 1] prevede, tra le altre cose, l'estensione dell'offerta di contenuti sottotitolati e audio.

L'Articolo 25, che riguarda il tema dei Servizi rivolti alle Persone con disabilità, riporta:

> *"la Rai, ai fini dell'espletamento del servizio pubblico radiofonico, televisivo e multimediale, è tenuta a estendere progressivamente la sottotitolazione e le audiodescrizioni anche alla programmazione dei canali tematici, con particolare riguardo all'offerta specificamente rivolta ai minori ed estendere progressivamente la fruibilità dell'informazione regionale".*

La **Rai** pertanto ha attivato una politica che prevede l'introduzione progressiva della sottotitolazione dei *Telegiornali Regionali* per tutte le regioni. La sede di Bolzano dal mese di marzo 2017 provvede già alla sottotitolazione delle notizie dell'edizione delle ore 20 del *Tagesschau*, il telegiornale locale in lingua tedesca.

Nota 1 - Il *CONTRATTO NAZIONALE DI SERVIZIO TRA IL MINISTERO DELLO SVILUPPO ECONOMICO E LA RAI-RADIOTELEVISIONE ITALIANA S.P.A. 2018-2022* è disponibile all'indirizzo http://www.rai.it/dl/doc/1521036887269_Contratto%202018%20testo%20finale.pdf

Per rispondere a questa indicazione editoriale nasce il *Progetto Sperimentazione Sottotitolazione automatica TGR* che ha l'obiettivo di verificare, in termini tecnici, operativi, qualitativi ed economici, la possibilità di utilizzare una *piattaforma automatica di generazione sottotitoli* nell'ambito della produzione dei sottotitoli dei TG Regionali.

Il **Centro Ricerche Innovazione Tecnologica e Sperimentazione Rai** ha recentemente condotto un'analisi delle prestazioni di alcuni sistemi commerciali di *trascrizione automatica del parlato* in lingua italiana che, seppur richiedendo un processo di correzione manuale e inserimento della punteggiatura prima della messa in onda, hanno dimostrato elevati valori di accuratezza sulle parole trascritte e di *accuratezza semantica*, indice della correttezza del significato del messaggio trascritto rispetto all'originale. La tesi della sperimentazione è che l'adozione di una soluzione tecnologica che prevede l'impiego, nei processi di validazione/correzione, di figure professionali senza specifiche competenze di stenotipia, potrebbe potenzialmente ridurre i costi del servizio di sottotitolazione e garantire la sostenibilità del progetto.

La Squadra di Progetto

Un progetto sfidante, basato su metodologie di trascrizione e sottotitolazione completamente automatiche, che possa essere efficace ed efficiente, prevede una fase di studio e analisi approfondita di tutte le componenti della pipeline di sviluppo, in modo da realizzare una proposta progettuale chiara e perfettamente calata sulle necessità. Per soddisfare questo requisito principe la collaborazione delle specifiche competenze aziendali risulta la vera scelta vincente per ottimizzare i processi e ottenere il risultato desiderato: competenze verticali e trasversali hanno contribuito a identificare un percorso di sviluppo e le specifiche caratteristiche dei moduli da adottare nella soluzione finale. In questi termini, le **Direzioni Rai** sia editoriali sia tecnologiche, affiancate alla professionalità di **PerVoice**, hanno contribuito fattivamente alla completa modellazione del disegno di progetto.

La Sfida

La sfida di utilizzare sistemi di trascrizione e di generazione automatica dei sottotitoli per il campo delle news è avvalorata dalla continua evoluzione in termini di precisione dei sistemi di *machine learning*, con la consapevolezza, però, della criticità del campo specifico di applicazione.

Nel configurare uno scenario applicativo così particolare occorre tener presente alcune considerazioni nodali:

- nel *contesto news* non solo è importante l'*accuratezza sulle parole* (*Word Accuracy-WA*) che il sistema di trascrizione automatica permette di raggiungere, ovvero la percentuale di parole trascritte correttamente, ma riveste un ruolo fondamentale anche l'*accuratezza semantica* (*Semantic Accuracy-SA*) che dà indicazione di quanto il significato del messaggio trascritto sia affine a quello originale. In un servizio giornalistico, anche se si registrasse un valore molto alto di *WA*, un solo errore di trascrizione di una singola parola potrebbe risultare critico e compromettere il significato dell'intero messaggio, portando la Semantic Accuracy a un valore bassissimo;
- la specializzazione del lessico e dei dizionari diventa in questo contesto una attività imprescindibile per poter disambiguare velocemente le parole e utilizzare il lessico specifico per il dominio della notizia.

Lo scenario è molto complesso, dovendo anche prevedere l'integrazione della soluzione applicativa in un contesto consolidato di sorgenti e flussi editoriali che rappresentano la struttura portante dell'edizione di un telegiornale.

Tutti questi elementi devono poter essere orchestrati in maniera scrupolosa per poter immaginare una soluzione completa e automatica, orientata alla generazione di un flusso continuo di sottotitoli finalizzato alla corretta comunicazione del messaggio verso l'utente televisivo finale.

Piattaforma di sottotitolazione automatica: Analisi e Progettazione

Il cuore della *Piattaforma di sottotitolazione automatica per le news*, rappresentata in Fig. 7, è il *modulo Speech to Text* che opera la trascrizione dell'audio in testo, affiancato da un modulo che formatta opportunamente il trascritto generando i sottotitoli, corredati di tutte le informazioni di sincronizzazione con il segnale audio/video di riferimento.

Le fasi di analisi e modellazione di una soluzione per il sistema di sottotitolazione automatica, che fosse completamente integrata con l'infrastruttura e con le differenti sorgenti necessarie per la messa in onda delle news, hanno richiesto un approccio di tipo bottom-up nell'affrontare le attività di analisi e progettazione.

La metodologia applicata ha consentito di partire da una soluzione che si potesse inserire in un flusso di produzione già attivo, passando poi a soluzioni sempre più complesse per integrare le differenti componenti dell'infrastruttura esistente, incrementandone e modellandone le complessità.

L'attività di analisi e progettazione, pertanto, si è articolata in tre diverse fasi:

- ***Fase 1***: approccio *FULL LIVE*
- ***Fase 2***: approccio *BATCH*
- ***Fase 3***: approccio integrato *TEXT*

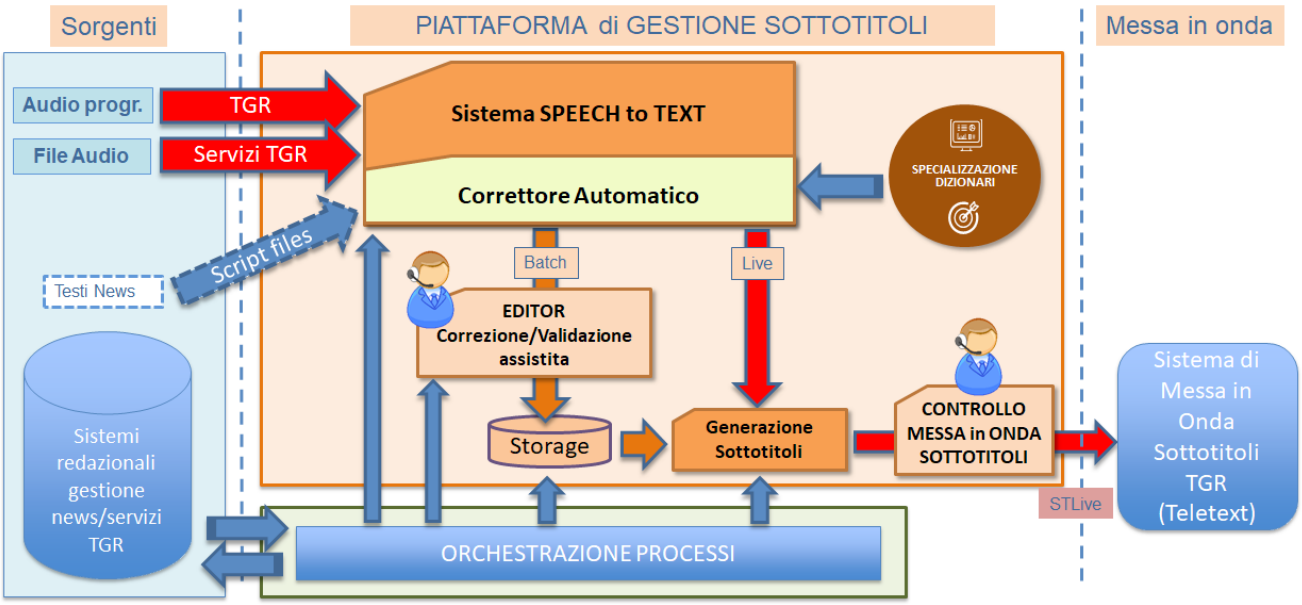
Queste tre fasi individuano tre diverse modalità di funzionamento ed erogazione dei sottotitoli della piattaforma e implementano logiche differenti per interfacciarsi e gestire correttamente i sistemi sorgente e i flussi di produzione e di pubblicazione delle news presenti in **Rai**.

Fase 1: approccio FULL LIVE

La *Piattaforma di sottotitolazione automatica per le news* nella modalità detta *FULL LIVE* fornisce in real time il flusso di sottotitoli in uscita.

Questa modalità permette di introdurre la sottotitolazione indipendentemente dall'infrastruttura di gestione dei flussi delle informazioni esistente, dalle sorgenti, con le loro tipicità, e dalle procedure editoriali che governano la messa in onda di un Telegiornale Regionale.

Fig. 7 – Piattaforma di sottotitolazione automatica per le news.



In questa specifica modalità il parlato viene trascritto in tempo reale, con una tolleranza di poche centinaia di millisecondi tra il momento in cui una parola viene pronunciata e il momento in cui è disponibile la sua trascrizione. La sequenza di più parole così trascritte forma il *Sottotitolo* che, formattato secondo specifiche regole che ne garantiscono la leggibilità (numero di parole per riga, tempo di visualizzazione, dimensione e posizione carattere), è inviato per la visualizzazione tramite il *Servizio Televideo* alla pagina 777.

Questo particolare approccio risulta snello nell'implementazione e di veloce applicabilità in architetture complesse ed eterogenee. Il motore di trascrizione e sottotitolazione non prevede di modificare i flussi produttivi audio e video, ma si può affiancare ad architetture esistenti collocandosi parallelamente alle pipeline di messa in onda del segnale televisivo.

Al sistema, come segnale d'ingresso, è sufficiente fornire un segnale audio normalizzato, tipicamente il segnale audio proveniente dallo studio del TGR.

Il segnale di uscita del sistema di trascrizione e sottotitolazione è rappresentato da un flusso di dati conforme al protocollo *STLive Formatted Input Protocol*, definito da **Rai** per l'erogazione dei servizi di sottotitolazione. STLive è il protocollo con il quale un *Client STLive* invia i sottotitoli, già formattati secondo lo standard teletext, al *Server STLive* che opera come *gestore dei servizi di sottotitolazione*.

Come descritto, l'approccio *FULL LIVE* ha la sua più importante potenzialità nel poter essere inserito in un flusso editoriale di gestione delle news senza introdurre specifiche variazioni e adattamenti. In questo caso l'analisi delle sorgenti e del workflow di lavoro redazionale non impatta sul risultato finale della sottotitolazione e pertanto non è possibile migliorare le prestazioni del sistema se non migliorando la specializzazione del modulo di trascrizione, inserendo nel dizionario ulteriori informazioni legate, ad esempio, alla geografia dei luoghi, alla toponomastica e alla odonomastica.

Nell'ottica di erogare un servizio di sottotitolazione automatica, è contemplata la possibilità da parte di un operatore di correggere il testo durante la fase di trascrizione e in corrispondenza della messa in onda.

Fase 2: approccio BATCH

La *Piattaforma di sottotitolazione automatica per le news* gestisce una modalità di lavoro complementare, denominata *BATCH*, che prevede il processamento dei contenuti audio relativi ai *servizi chiusi* della redazione TGR, ovvero tutti i servizi già montati che hanno ottenuto la validazione editoriale e che vengono resi disponibili al sistema prima della messa in onda del TGR. Il sistema in tale modalità esegue *offline* la trascrizione dal parlato in testo e genera il flusso dei sottotitoli che sarà opportunamente richiamato e visualizzato tramite il *Servizio Televideo* alla pagina 777, in corrispondenza della messa in onda dello specifico contenuto.

La gestione della modalità *BATCH*, che si alterna con la modalità *LIVE*, consente di incrementare la qualità del servizio migliorando l'accuratezza della trascrizione del prodotto finale in termini di *word accuracy*, misurata come percentuale di parole trascritte correttamente rispetto al numero totale di parole trascritte dal sistema automatico. In particolare, la trascrizione in modalità *BATCH*, operando su un file audio interamente disponibile al sistema, consente di fornire su un servizio TGR un'accuratezza mediamente superiore del 2% rispetto a quella generata in modalità *LIVE*. Inoltre il sistema è in grado di gestire la punteggiatura fornendo così dei sottotitoli più leggibili e quindi di più facile comprensione.

In aggiunta, il fatto di trascrivere e sottotitolare servizi chiusi prima della messa in onda consente, anche, di prevedere operazioni di modifica, correzione e validazione dei sottotitoli generati automaticamente da parte di un operatore, fornendo così una sottotitolazione con un valore di accuratezza pari al 100%.

Nell'ottica di erogare un servizio di sottotitolazione automatica affidabile, è quindi previsto l'impiego di

un'applicazione web per effettuare offline questo tipo di operazioni di modifica, correzione e validazione. Queste funzionalità sono fornite dal modulo *EDITOR*, collocato valle del sistema *Speech to Text*.

Il modulo *EDITOR* è completamente integrato all'interno della piattaforma di generazione automatica dei sottotitoli ed è gestito dal modulo *Orchestrazione Processi* che, durante la messa in onda dell'edizione TGR, accede alle trascrizioni corrette e validate fornendo, così, un servizio di sottotitolazione con un elevatissimo grado di qualità.

Le funzionalità fornite dal modulo *EDITOR* permettono ad un operatore di:

- ascoltare l'audio del servizio giornalistico ed operare con semplicità funzioni di start-stop-repeat (es: tramite pedaliera);
- effettuare delle correzioni sul testo generato dal modulo *Speech To Text* e inserire la punteggiatura. L'interfaccia grafica consente una correzione agevole delle parole errate e fornisce dei suggerimenti per una scelta rapida di quelle che, con maggiore probabilità, potrebbero essere corrette;
- salvare i termini maggiormente ricorrenti o specifici di un particolare contesto in un'area *suggerimenti* (es: nomi di personaggi o località)

Fase 3: approccio integrato TEXT

La *Piattaforma di sottotitolazione automatica per le news* contempla una terza modalità operativa, denominata *TEXT*, che non prevede l'attivazione del sistema di trascrizione ma la gestione dei contenuti testuali attinti dai sistemi redazionali TGR, i quali vengono formattati opportunamente e trasformati in sottotitoli.

Tale modalità, che viene attivata tipicamente in corrispondenza dei titoli e dei lanci dei servizi, consente di garantire un'accuratezza della sottotitolazione pari al 100%, di valorizzare fonti dati già disponibili e validate e di mantenere l'integrità del messaggio che il giornalista, nella scrittura del testo, intende comunicare.

Piattaforma di sottotitolazione automatica: funzionalità

Le principali funzionalità fornite dalla *Piattaforma di sottotitolazione automatica per le news*, rappresentata nella già citata Fig. 7, possono essere sintetizzate in:

- *Acquisizione*. La piattaforma acquisisce i contenuti testo/audio dalle diverse tipologie di sorgenti tramite specifiche interfacce opportunamente definite;
- *Trascrizione*. Il *sistema Speech to Text* è il modulo della piattaforma responsabile della trascrizione automatica del segnale audio. Esso trasforma il parlato in un testo scritto corredato di tutte le informazioni di sincronizzazione con l'audio originale;
- *Correzione Assistita flusso Live*. Per la modalità *Live* è previsto un tool grafico usato da un operatore per correggere eventuali errori di trascrizione e inserire la punteggiatura prima della messa in onda;
- *Correzione mediante editor dei testi*. Per la modalità *Batch* è previsto un editor per correggere e validare il testo generato dalla trascrizione automatica, garantendo così una percentuale di accuratezza pari al 100%;
- *Generazione Sottotitoli*. Formattazione dei sottotitoli, sincronizzazione e interazione con i sistemi di messa in onda.

Le ulteriori funzionalità di *Specializzazione del modello del linguaggio* e di *Arricchimento dei Dizionari* contribuiscono, nel tempo, a migliorare le prestazioni del sistema in termini di accuratezza delle trascrizioni.

Il sistema di trascrizione è dotato a priori di un modello del linguaggio *generalista*, ovvero non specifico di alcun ambito.

Per il progetto *Sottotitolazione Automatica TGR* il modello del linguaggio viene costantemente arricchito, rispetto a specifici argomenti del contesto desiderato (es. politica, sport, meteo, traffico, ecc.) e dello specifico lessico regionale.

Specializzazione del modello del linguaggio

Nel corso della sperimentazione, al fine di migliorare il livello di accuratezza nel riconoscimento del parlato, è prevista la realizzazione di specifici *modelli di linguaggio*, che vengono arricchiti e declinati per, ad esempio, ciascuna regione e/o rispetto a specifici argomenti trattati nel contesto desiderato (es. politica, sport, meteo, traffico, ecc.). È possibile, pertanto, selezionare il modello del linguaggio specializzato in funzione del dominio da trascrivere.

Arricchimento dei Dizionari

Il sistema di trascrizione automatica è dotato di moduli specifici che consentono di aggiornare, migliorare ed ottimizzare il riconoscimento automatico di terminologie speciali ed attualizzate relative ad un ambito specifico:

- *Modulo attualizzazione automatico*: attualizza ed arricchisce automaticamente il dizionario del modello raccogliendo in autonomia nuovi termini da fonti aperte (es. feeds rss) permettendo di riconoscere sempre i nuovi termini;
- *Modulo arricchimento assistito*: permette, attraverso un'avanzata interfaccia utente web-based, l'inserimento e la modifica dei termini, quali, a titolo di esempio:
  - *lista nomi propri*: i nomi e cognomi di persona o di località specifici appartenenti ai contesti locali.
  - *lista parole*: tutte le parole peculiari di uno specifico contesto.

L'Orchestratore Processi

I vantaggi derivanti dalla gestione delle trascrizioni operate nelle modalità *BATCH* e *TEXT* hanno indotto la decisione di sviluppare e introdurre per il progetto *Sottotitolazione Automatica TGR* un orchestratore che, durante la messa in onda dell'edizione TGR, opera una commutazione del sistema di *Trascrizione e Sottotitolazione Automatica* tra le modalità di trascrizione e generazione di sottotitoli *Live*, *Batch* e *Text*, in base all'identificazione della tipologia di contenuto in onda.

L'*Orchestratore Processi* è il modulo software responsabile della gestione dei vari moduli interni alla piattaforma di generazione automatica dei sottotitoli, di tutti i semilavorati generati dai singoli moduli e garantisce il corretto interfacciamento della piattaforma con le differenti tipologie di sistemi sorgente e di sistemi di messa in onda.

Piattaforma di sottotitolazione automatica: sperimentazione e valutazione del servizio

La sperimentazione prevede un'ultima fase di valutazione della piattaforma per l'erogazione del servizio e del prodotto/servizio finale.

A corredo dell'analisi tecnica delle prestazioni dei singoli moduli della piattaforma di generazione dei sottotitoli, è importante valutare la piattaforma stessa nella sua interezza in termini di usabilità, velocità, impiego di risorse richieste e affidabilità. Tale valutazione sarà effettuata dagli operatori coinvolti nei processi di revisione delle trascrizioni *batch* e di presidio/intervento *live* durante la messa in onda delle diverse edizioni dei TGR.

La valutazione del prodotto finale si articola, invece, in due modalità di analisi:

- una *valutazione tecnico-oggettiva* delle prestazioni del sistema di trascrizione e generazione dei sottotitoli automatico sulla base di analisi sia di tipo *word accuracy*,ovvero basate sulla percentuale di parole trascritte correttamente sul totale delle parole trascritte, sia di tipo *semantic accuracy*, ovvero basate sulla percentuale di concetti comprensibili e corretti sul totale del numero di concetti espressi. E' doveroso precisare che, a differenza della word accuracy, la semantic accuracy, basandosi sull'interpretazione semantica dei concetti, intesi come sequenze di parole che veicolano un messaggio di senso compiuto, risente di una minima componente di interpretazione soggettiva;
- una *valutazione soggettiva* che prevede l'attivazione di focus group di potenziali utilizzatori del servizio a cui verranno presentate alcune edizioni dei TGR sottotitolate mediante l'uso

del sistema di sottotitolazione automatico e che forniranno un riscontro importante sull'indice di comprensibilità, gradevolezza, leggibilità dei relativi sottotitoli. Una parte consistente degli utenti selezionati saranno le persone alle quali è fortemente indirizzato il servizio di sottotitolazione, in particolare persone con disabilità uditive. È fondamentale pertanto il coinvolgimento delle relative Associazioni e una comunicazione corretta al fine di condividere ambizioni e finalità del progetto, mettendo anche in luce, con la massima trasparenza, i limiti e gli eventuali errori sui sottotitoli che un sistema automatico, anche se in minima misura, potrebbe commettere.

A titolo di esempio, in Fig. 8 è riportato un estratto dell'analisi di una tipica edizione di un TGR. Per ogni tipologia di contenuto viene messa in evidenza la modalità di funzionamento della piattaforma e la relativa *word accuracy (WA)*. Un'accuratezza del 100% si registra in corrispondenza dei sottotitoli generati in modalità *TEXT*, in cui non interviene il sistema di trascrizione. Nelle modalità *LIVE* e *BATCH* le percentuali di accuratezza sono comunque piuttosto elevate, mediamente intorno al 90-95%. Introducendo, per la correzione dei servizi TGR chiusi trascritti in modalità *BATCH*, l'operazione di revisione e correzione del trascritto si raggiunge, come indicato nella quarta colonna, un'accuratezza del 100%.

La valutazione della *Piattaforma di sottotitolazione automatica per le news* fornirà gli elementi per stimare costi e benefici di una soluzione che, grazie all'introduzione di algoritmi di intelligenza artificiale, può innovare e ottimizzare gli attuali processi di produzione dei sottotitoli.

Fig. 8 – Analisi di un'edizione del TGR (WA = Word Accuracy).

Scaletta TGR	Trascrizione	Word Accuracy	WA con BATCH Corretti
1 – Titoli	**TEXT**	100 %	100 %
2 – Giornalista	**TEXT**	100 %	100 %
3 – Servizio	**BATCH**	**94,34 %**	100 %
4 – Giornalista	**TEXT**	100 %	100 %
5 – Collegamento	**LIVE**	**94,34 %**	**94,34 %**
6 – Servizio	**BATCH**	**89,80 %**	100 %
7 – Giornalista	**TEXT**	100 %	100 %
8 – Servizio	**BATCH**	**95,19 %**	100 %
9 – Servizio	**BATCH**	**79,66 %**	100 %
10 – Giornalista	**TEXT**	100 %	100 %
11 – Servizio	**BATCH**	96.97 %	100 %
12 – Giornalista	**TEXT**	100 %	100 %
13 – Servizio	**BATCH**	**95,36 %**	100 %
14 – Giornalista	**TEXT**	100 %	100 %
15 – Collegamento	**LIVE**	**85,63 %**	**85,63 %**
16 - Saluti	**LIVE**	**99,10 %**	**99,10 %**
		95,65 %	**98,69 %**

Conclusioni e future evoluzioni

Alla luce di quanto fin qui esposto è lecito pensare che i sistemi di *Machine Learning* per la T*rascrizione Automatica* siano solo all'albore del loro sviluppo, essendo essi una realtà estremamente complessa e multiforme in cui coesistono aspetti contradittori, ma anche sviluppi interessanti per innovativi scenari di applicazione.

L'utilizzo sempre più spinto di queste tecnologie indica che tali sistemi, nonostante possibili errori e imprecisioni, sono in grado di fornire una trascrizione di elevata qualità, tale da poter ipotizzare di incentrare sempre più servizi per il cittadino su tali soluzioni.

Nel contesto giornalistico la specializzazione dei modelli del linguaggio riveste un importante tassello per il miglioramento della trascrizione del parlato ed è innegabile che, data l'importanza delle tematiche affrontate e della diffusione del messaggio veicolato, diventano fondamentali anche le attività di post-editing sul testo già trascritto per raggiungere percentuali elevatissime di accuratezza.

Le logiche di gestione del workflow, in relazione alle tipicità delle sorgenti coinvolte nel processo di creazione di una edizione di news, incrementano ulteriormente la qualità dei risultati prodotti dai sistemi automatici in termine di efficacia e di affidabilità del sistema.

Nei prossimi anni i sistemi basati su *Intelligenza Artificiale* saranno un elemento nodale dei processi produttivi coinvolti nella gestione dell'informazione e nell'ottimizzazione delle risorse aziendali.