**UNIVERSITÀ POLITECNICA DELLE MARCHE**
**FACOLTÀ DI INGEGNERIA**
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica e dell'Automazione


**TESI DI LAUREA**

# Esperienze con applicazioni di Intelligenza Artificiale a supporto del Cognitive Computing: traduttori, trascrittori e chatbot

# Experiences with Artificial Intelligence applications supporting Cognitive Computing: translators, transcribers and chatbots

Relatore

Prof. Domenico Ursino

Candidato

Giansimone Coccia

**ANNO ACCADEMICO 2022-2023**

*Non hai bisogno di vedere l'intera scalinata.*
*Inizia semplicemente a salire il primo gradino*

Martin Luther King

**Sommario**

Negli ultimi anni l'Intelligenza Artificiale, in particolare il Cognitive Computing, è diventata sempre più centrale, fino a diventare parte integrante della nostra vita. Questo grazie a dispositivi che ci semplificano la vita quotidiana; basti pensare ai trascrittori presenti nel web e sui social, ai traduttori, oppure ai chatbot con i quali dobbiamo comunicare, ad esempio, per servizi di assistenza al cliente. Questi strumenti, sono pensati per funzionare con l'input dell'uomo; ecco, dunque, lo scopo del Cognitive Computing, ovvero cercare di ottenere soluzioni informatiche che consentano di interagire con i computer in modo molto più immediato e naturale di oggi, praticamente "parlando" con le macchine, e sfruttando la loro capacità di imparare dall'esperienza. In questa tesi vengono analizzati tre servizi, legati ai sistemi cognitivi, come quelli di traduzione, di trascrizione e chatbot, messi a disposizione dai principali provider Cloud, come Google, Microsoft e Amazon. Questi sono presentati in linea teorica, congiuntamente ad alcuni esempi pratici, così da effettuare un confronto e mettere in risalto caratteristiche e peculiarità di ciascuno di essi.

**Keyword**: Cognitive Computing, Intelligenza Artificiale, Traduttori, Trascrittori, Chatbot, Cloud Computing, Reti Neurali, Natural Language Processing (NLP), Big Data, Question Answering, Machine Learning

# Indice

# Elenco delle figure

# Elenco delle tabelle

# Introduzione

L'universo dell'*Intelligenza Artificiale (IA)*, anche detta *Artificial Intelligence (AI)*, ha catturato la nostra curiosità fin da quando ha iniziato a diffondersi a gran voce negli ultimi anni. Molte aziende, infatti, si sono cimentate nello sviluppo di software e applicazioni che hanno alla base questa speciale tecnologia, come chatbot, trascrittori, traduttori, prodotti utilizzati in ambito della medicina o nel customer service.

Siamo stati affascinati dalle promesse e dalle potenzialità di questa rivoluzionaria tecnologia, soprattutto quando questa, unita alle capacità dell'essere umano, è in grado di fornire soluzioni straordinarie e sorprendenti in ambito scientifico. Il panorama, sul quale attualmente si sta spingendo molto, è quello del *Cognitive Computing*, che concentra la sua forza sull'unione tra Intelligenza Artificiale e capacità intellettuali dell'uomo.

Abbiamo così deciso di intraprendere un progetto che permettesse di esplorare a fondo i servizi maggiormente legati all'interazione uomo-macchina, come quelli di traduzione, trascrizione e chatbot, offerti da alcune delle principali piattaforme di Intelligenza Artificiale: Microsoft Azure, Google Cloud e Amazon AWS.

La possibilità di creare macchine capaci di apprendere e adattarsi autonomamente, migliorando le prestazioni nel tempo, ci ha spinto a voler comprendere come queste soluzioni di Intelligenza Artificiale potessero apportare maggiore valore in settori importanti come nel campo della traduzione multilingue, nella trascrizione di contenuti audio e nella creazione di chatbot avanzati in grado di rispondere a determinate nostre esigenze.

Abbiamo così deciso di dedicare parte del nostro tempo a studiare i servizi di Intelligenza Artificiale offerti da Microsoft Azure, Google Cloud e Amazon AWS, non solo per soddisfare la nostra curiosità personale, ma anche per comprendere appieno il potenziale di queste piattaforme e come possano essere sfruttate, in modo innovativo, nei vari contesti aziendali e sociali.

Questo studio ci ha fornito una visione approfondita delle caratteristiche distintive di ciascun servizio, e ci ha permesso di valutarne le prestazioni in situazioni reali, siano esse positive o negative. Inoltre, ci ha offerto preziose indicazioni su come poter approfondire il percorso accademico, e su quali studi magistrali intraprendere, visti anche i numerosi ambiti toccati dall'informatica: Intelligenza Artificiale, Data Scienze, Cybersecurity e così via.

Nel corso di questo studio, abbiamo esaminato attentamente le soluzioni proposte da Microsoft Azure, Google Cloud e Amazon AWS eseguendo una serie di test ed esempi pratici per valutare le prestazioni di ciascun servizio. Durante questo processo, abbiamo identificato i punti di forza distintivi di ogni piattaforma e gli aspetti che potrebbero essere migliorati per garantire risultati ottimali, così da soddisfare le esigenze specifiche di ciascun progetto.

La parte riguardante ciacun servizio inizia con una breve presentazione su di esso, spiegando quali siano le sue caratteristiche principali e fornendo una descrizione di massima dello stesso. Per ciascun servizio, poi, abbiamo eseguito vari esempi basati sullo stesso campione di test, in modo da poter rilevare le principali differenze e le peculiarità che lo contraddistinguono.

Al termine dell'esame dei tre servizi, congiuntamente al capitolo sui chatbot, verrà presentato, brevemente, uno dei Customer Relationship Management (CRM) più in voga al momento, ovvero *Salesforce*.

Successivamente, prima di ricavare le conclusioni, abbiamo deciso di fornire qualche dettaglio in più sulla differenza tra *Cloud* e *on-premise*, visto che l'utilizzo di questi servizi è avvenuto prevalentemente attraverso le piattaforme Cloud messe a disposizione dai singoli provider.

Infine, abbiamo tratto le nostre conclusioni, spiegando perché abbiamo preferito determinate caratteristiche, e delineando alcuni aspetti futuri per noi interessanti, che potrebbero tornare utili.

La presente tesi è composta da sei capitoli strutturati come di seguito specificato:

- Nel Capitolo 1 saranno introdotti i concetti di Cognitive Computing, aspetto centrale del nostro lavoro, Intelligenza Artificiale, con alcuni cenni storici a cui fare riferimento, Machine Learning e Deep Learning; verranno, altresì, presentate le tecniche di Intelligenza Artificiale adoperate.

- Nel Capitolo 2 verrà approfondito maggiormente il discorso sul Cognitive Computing, in particolare riportando quali sono i concetti base e le sue caratteristiche. Verranno, poi, introdotti i principali servizi di Cognitive Computing, dedicando particolare attenzione a Watson di IBM e ai suoi servizi, per poi accennare ai provider da noi utilizzati, ovvero Google, Microsoft ed Amazon.

- Nel Capitolo 3 verrà introdotto il primo servizio scelto, ovvero quello della traduzione, messo a disposizione dalle varie piattaforme. In particolare, vengono presentate caratteristiche generali sul servizio, nonché esempi esplicativi utili nel confronto tra i vari provider.

- Nel Capitolo 4 verrà analizzato il secondo dei tre servizi, ovvero la trascrizione. Come nel caso precedente, vi sarà una presentazione generale del servizio di trascrizione, nonché alcuni esempi acquisiti effettuando delle prove.

- Nel Capitolo 5 verrà esaminato l'ultimo dei servizi, ovvero quello riguardante i chatbot, diversi in base ai vari provider utilizzati. Anche quì sono stati inseriti esempi adottati in fase di sperimentazione, così da evidenziare i tratti più importanti di ciascuno di essi. Infine, verrà proposto un breve accenno ad uno dei maggiori fornitori di soluzioni CRM del momento, ovvero Salesforce.

- Nel Capitolo 6 verranno tratte le conclusioni, considerando anche la differenza tra i due pilastri dell'attuale trasformazione digitale, ossia il Cloud Computing e le soluzioni on-premise. Per concludere, verranno delineati alcuni possibili scenari futuri.

CAPITOLO 1

# Introduzione al cognitive computing

*In questo primo capitolo verrà introdotto il concetto di Cognitive Computing. Dunque si parlerà in primis dell'Intelligenza Artificiale in senso generale, riportando qualche estratto di storia, per poi passare alle sotto-categorie di questa. In particolare verranno presentati i concetti di Machine Learning e Deep Learning, con le relative tecniche e metodologie che maggiormente vengono applicate in questi casi. Infine sono stati proposti alcuni prototipi interessanti che negli ulimi anni sono stati sviluppati. Superata questa fase di infarinatura generale, si passerà al concetto cardine, cioè il Cognitive Computing. È prevista una nota introduttiva che spiega l'argomento prima di entrare nel dettaglio. Dopodiché, la parte di storia per comprendere l'evoluzione dei sistemi che hanno portato a coniare questo termine. Successivamente verranno analizzate le peculiarità di quest'ultimo, con le sue tre caratteristiche principali e le tre capacità cognitive, per poi terminare con un accenno sulle possibili implementazioni future ed i limiti.*

## 1.1 L'Intelligenza Artificiale

L'Intelligenza Artificiale, spesso abbreviata con l'acronimo AI (Artificial Intelligence), è la disciplina che studia e tenta di riprodurre le facoltà *cognitive* umane mediante sistemi informatici e con l'ausilio di particolari tecnologie. Il termine fu usato per la prima volta nel 1956 dal matematico americano John McCarthy, durante un seminario svoltosi nel College di Dartmouth, nel New Hampshire (Figra 1.1).

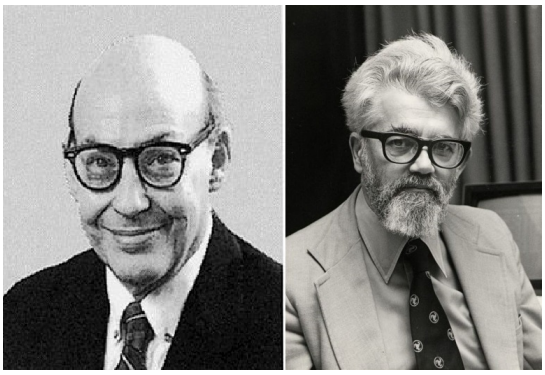

**Figura 1.1:** Marvin Lee Minsky a sinistra, John McCarthy a destra

L'obiettivo primario dell'AI è quello di creare macchine (hardware/software) in grado di pensare e agire come gli esseri umani attraverso l'individuazione di modelli, ovvero descrizioni dei problemi da risolvere, e di algoritmi, ovvero procedure effettive per risolvere i

modelli. L'Intelligenza Artificiale si focalizza su tre abilità cognitive tipiche dell'essere umano, ovvero:

- apprendimento;
- ragionamento;
- auto-correzione.

Quindi l'AI tenta di capire sempre meglio l'intelligenza umana così da risolvere i problemi in modo analogo a come verrebbero risolti dall'uomo.

> *"La domanda non è se le macchine intelligenti possano avere emozioni, bensì se le macchine possano essere intelligenti senza avere emozioni", The Society of Mind, Marvin Minsky, 1927-2016.*

Partendo da questa citazione è possibile pensare ad una possibile suddivisione dell'Intelligenza Artificiale, in tre principali categorie:

- *Artificial Narrow Intelligence (AI Debole).* Sono sistemi progettati ed addestrati per portare a termine un task specifico limitandosi a simulare il comportamento umano in base ai dati ricevuti in input.
- *Artificial General Intelligence (AI Forte).* Sono sistemi in grado di comportarsi alla pari di un altro essere umano comprendendone anche il tono e le emozioni.
- *Artificial Super Intelligence (Super AI).* Quando la macchina diventa consapevole di se stessa superando le capacità dell'essere umano.

### 1.1.1 Metodiche e tecniche utilizzate dall'AI

Dopo aver introdotto il concetto generale di "Intelligenza Artificiale", siamo ora in grado di analizzare i metodi e le tecniche che sono alla base di questo principio.

Molti degli algoritmi di Intelligenza Artificiale sono basati sull' *Analisi delle serie temporali* e sulle *Reti neurali ricorrenti (RNN, Recurrent neural networks)*.

L'analisi delle serie temporali consiste nello studiare i dati temporali in possesso così da prevedere possibili scenari futuri o anomalie che potrebbero scaturire. Gli algoritmi basati sulle serie temporali acquisiscono un insieme di dati, la cui natura non deve essere strettamente legata al tempo, così da apprendere sequenze temporali.

L'obiettivo principale dell'analisi delle serie temporali è quello di sviluppare modelli matematici che forniscano descrizioni plausibili da dati campione.

Le reti neurali ricorrenti sono state ideate per l'analisi delle sequenze. Sono infatti in grado di sviluppare una sorta di memoria legata agli input passati che andranno poi ad influenzare quelli futuri.

È una classe di reti neurali artificiali che include neuroni collegati tra loro in un ciclo.

Tipicamente i valori di uscita di uno strato di un livello superiore sono utilizzati in ingresso di uno strato di livello inferiore. Quest'interconnessione tra strati permette l'utilizzo di uno di essi come memoria di stato, e consente, fornendo in ingresso una sequenza temporale di valori, di modellarne un comportamento dinamico temporale dipendente dalle informazioni ricevute agli istanti di tempo precedenti.

Un'altra tecnica particolare utilizzata negli algoritmi di Intelligenza Artificiale è quella del *Natural Language Processing (NLP)*, che si occupa di determinare il significato delle frasi espresse in linguaggio naturale.

Il dialogo tra uomo e macchina coinvolge diversi aspetti, quali fonetica, fonologia, morfologia, sintassi, semantica, pragmatica e il discorso nel suo complesso. Di conseguenza, sono numerosi i task di NLP che automatizzano queste aree. Per svolgere questi ultimi le aziende adottano molti task di Natural Language Processing, quali ad esempio:

- *Text Analysis:* analisi di un testo e, laddove richiesto, individuazione di elementi chiave.
- *Text Classification:* interpretazione di un testo per classificarlo in una categoria predefinita.
- *Sentiment Analysis:* rilevamento dell'umore all'interno di un testo.
- *Language Translation:* traduzione di testi scegliendo, volta per volta, il significato migliore a seconda del contesto.

Per far ciò viene eseguita sulla frase prima un'analisi sintattica e subito dopo un'analisi semantica per l'identificazione dei concetti espressi. Il Natural Language Processing costituisce il principio cardine di molti chatbot intelligenti o assistenti virtuali come Alexa, Siri o Google Assistant.

### 1.1.2 Machine Learning e Deep Learning

Oltre a saper interpretare i significati delle frasi espresse in linguaggio naturale, gli algoritmi di AI riescono anche a comprendere ciò che un'immagine rappresenta. Questo è il settore della *Computer Vision* che si concentra sull'acqusizione, elaborazione ed analisi di immagini e video.

Gli algoritmi di Computer Vision immagazzinano grandi quantità di immagini ad alta risoluzione e sono poi in grado di restituire delle rappresentazioni simboliche dei contenuti rappresentati al loro interno. Il procedimento completo consiste nello scomporre prima l'immagine in piccoli patch; ciascuno di questi viene, poi, analizzato così da riuscire a cogliere delle caratteristiche più semplici usando le quali si verranno a creare delle feature più complesse.

Gli algoritmi di Computer Vision possono effettuare indagini più o meno approfondite su un'immagine, a seconda delle tecniche utilizzate, della tipologia di immagine e del tipo di task effettuato. Tra i possibili task si individuano i seguenti:

- *Image Classification:* analisi del contenuto dell'immagine e attribuzione di un'etichetta.
- *Object Detection:* identificazione di una o più entità all'interno di un'immagine.
- *Face Recognition:* riconoscimento di volti di persone.
- *Emotion Recognition:* rilevamento del sentiment di un'immagine.

Per quanto riguarda la Computer Vision, ci sono alcune difficoltà da affrontare, come, ad esempio quella di creare un dataset sufficientemente ampio che consenta un addestramento ottimale dell'algoritmo, oppure il dover insegnare all'algoritmo a riconoscere anche immagini in presenza di trasformazioni.

Quello spiegato finora non è altro che uno dei primi algoritmi di *Machine Learning*, cioè un sottoinsieme dell'Intelligenza Artificiale che fornisce metodi ed algoritmi per permettere alle macchine di imparare autonomamente dalle loro esperienze e dai dati passati.

Le tecniche di Machine Learning utilizzano approcci come gli alberi di decisione, cioè un grafo di decisioni e delle loro possibili conseguenze, oppure il *reinforcement learning*, cioè un metodo in cui i programmatori dividono i comportamenti positivi da quelli negativi.

Il metodo consiste nell'assegnare valori positivi alle azioni desiderate e valori negativi ai comportamenti indesiderati.

Gli algoritmi di Machine Learning, di cui sopra ne sono riportati alcuni, lavorano bene quando i dati sono buoni, puliti e coerenti. Nel momento in cui le dimensioni dei dati aumentano, entra in gioco il *Deep Learning*, che trova il suo punto di forza proprio nella sua capacità di gestire questi dati "profondi".

Il Deep Learning è visto come un'evoluzione del Machine Learning che utilizza dati meno strutturati, più grandi e di diversa natura. Esso si basa sulle *Reti neurali artificiali*.

La prima rete neurale artificiale risale al 1949, in particolare al "Percettrone" di Rosenblatt. Nella Figura 1.2 viene riportato un esempio di funzionamento di una rete neurale.

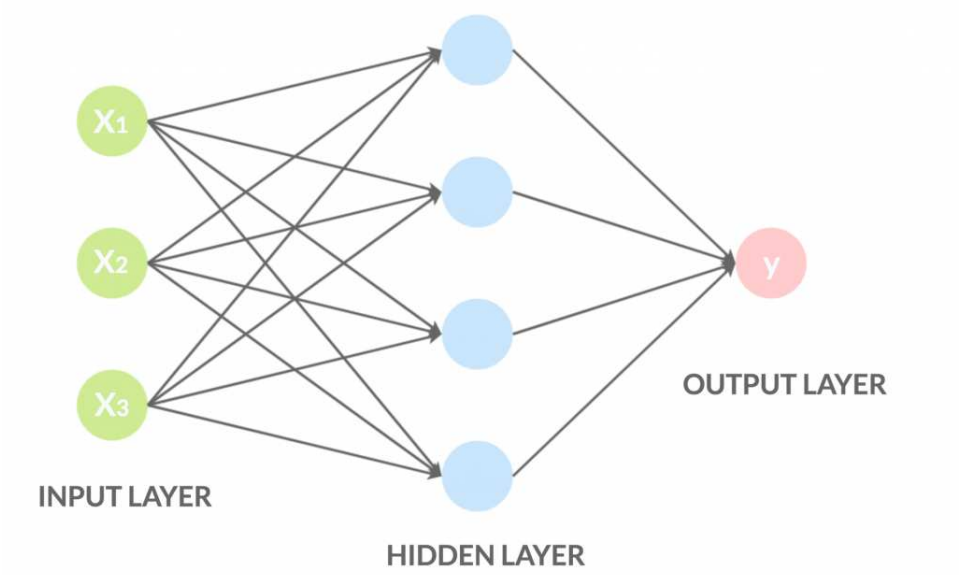


**Figura 1.2:** Schema di funzionamento di una rete neurale

In pratica la rete neurale è una funzione definita come composizione di altre funzioni, che, a loro volta, possono essere definite come composizione di ulteriori funzioni.

Il modello più semplice a cui fare riferimento è costituito da tre livelli di elaborazione, ovvero il primo in cui vengono forniti i dati (input layer), il secondo, in cui vengono elaborati i dati (hidden layer), ed il terzo, in cui vengono restituiti i risultati ottenuti (output layer). Aumentando il numero di strati intermedi, ovvero gli "hidden layers", si avranno una complessità maggiore ed anche uno sforzo computazionale maggiore.

### 1.1.3 Invenzioni e risultati ottenuti

Mostriamo ora una serie di risultati ottenuti durante gli ultimi anni utilizzando le tecniche di Machine Learning e Deep Learning.

- *1997, DeepBlue sconfigge il campione mondiale di scacchi Kasparov*. DeepBlue è un computer IBM Risk 2000 in grado di valutare duecento milioni di mosse al secondo e conosce seicento mila aperture di partita differenti (Figura 1.3).
- *2011, Watson vince a Jeopardy!*. Watson è un sistema di Intelligenza Artificiale, in grado di rispondere a domande espresse in un linguaggio naturale, sviluppato all'interno del progetto DeepQA di IBM. Alla base della vittoria di Watson vi sono il Cognitive Computing ed il Natural Language Processing (Figura 1.4).


**Figura 1.3:** DeepBlue vs Kasparov


**Figura 1.4:** Watson vince a Jeopardy!

- *2017, AlphaGo batte il numero 1 al mondo nel gioco del Go*. Go è un gioco che richiede particolare creatività ed intuito. AlphaGo è basato sulla tecnica del reinforcement learning citata precedentemente (Figura 1.5).

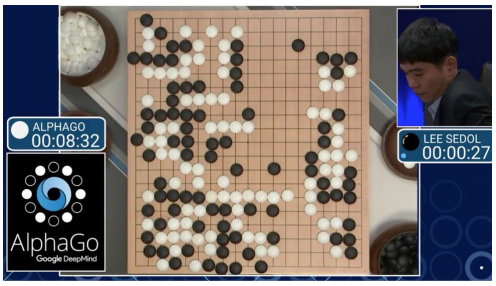


**Figura 1.5:** AlphaGo vince nel gioco del Go

- *2017, Libratus vince un milione e mezzo di dollari giocando a poker*. Il poker è ritenuto complesso in quanto l'informazione è in parte nascosta. Libratus è stato addestrato con la tecnica del reinforcement learning ed ha iniziato a giocare numerose partite contro sé stesso (Figura 1.6).

## 1.2 Cognitive Computing

Il *Cognitive Computing* è un approccio tecnologico che consente agli esseri umani di collaborare con le macchine, arrivando ad ottenere un processo di pensiero umano computerizzato.

Il Cognitive Computing rappresenta sistemi di autoapprendimento automatico per imitare il funzionamento del cervello umano, quest'ultimo rappresentato da una rete neurale che ricalca le funzionalità del sistema nervoso centrale degli esseri umani.

Si tratta di macchine che operano ad un livello diverso dai normali sistemi informatici in quanto necessitano di analizzare grandi quantità di dati, strutturati o meno, cercando di apprendere da questi.

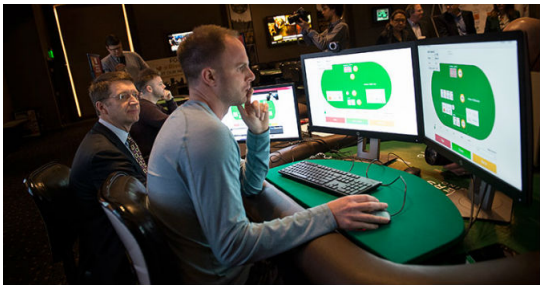

**Figura 1.6:** Libratus

Un sistema cognitivo ha tre principi fondamentali come descritti di seguito:

1. *Learn*. Un sistema cognitivo impara, cioè sfrutta i dati per trarre conclusioni su un dominio, un argomento o un problema basati su informazioni e osservazioni.

2. *Model*. Per imparare, il sistema necessita di costruire un modello o una rappresentazione del dominio e presupposti per determinare quali algoritmi di apprendimento utilizzare.

3. *Generate hypothesis*. Un sistema cognitivo ipotizza che non esistano risposte corrette. La risposta migliore è basata sul dato in sé, per questo un sistema cognitivo è probabilistico. Il sistema sfrutta i dati per addestrare, testare o assegnare un punteggio ad una ipotesi.

La "cognizione" è il processo che l'uomo utilizza per apprendere attraverso pensieri, esperienze e dati sensoriali. Attualmente l'Intelligenza Artificiale in uso può essere considerata debole, poiché focalizzata sull'esecuzione di un solo task per volta.

Molti ricercatori sono al lavoro per portare l'AI al livello superiore facendola diventare "forte", in modo che apprenda come un essere umano e possa risolvere problemi tipici dell'umano.

L'uomo è capace di applicare il buon senso, la morale e la ragione per risolvere i problemi che si trova davanti. Intuisce le migliori idee che lo portano a superare le difficoltà, ma è limitato dalla quantità di tempo che deve impiegare per imparare.

Il Cognitive Computing è una delle sottodiscipline dell'Intelligenza Artificiale che unisce i punti di forza dell'essere umano con quelli della macchina, quindi tenta di fondere capacità umane come emozioni e immaginazione, con capacità informatiche, come la capacità di calcolo.

Il Cognitive Computing utilizza i punti di forza dell'essere umano per simulare i processi del pensiero umano in un modello computerizzato.

Per fare ciò esso sfrutta tecniche come il Machine Learning, il Natural Language Processing, il Data Mining ed il Pattern Matching.

Stiamo entrando in una nuova era del computing che trasformerà il modo in cui gli umani collaborano con le macchine. È chiaro che le innovazioni tecnologiche hanno trasformato industrie ed anche il modo di vivere dei singoli individui.

Nei prossimi anni i sistemi cognitivi verranno utilizzati in molti settori, come ad esempio la sicurezza, oppure per risolvere problemi legati all'ottimizzazione industriale, come il determinare la via migliore per anticipare le esigenze dei clienti nella vendita al dettaglio.

Anche nei settori legati alla medicina si stanno facendo grandi passi in avanti introducendo sistemi che riconoscono e sono in grado di diagnosticare una malattia.

In molti casi l'informatica cognitiva si è già fatta largo; basti pensare ai numerosi ChatBot ideati dai grandi player dell'informatica e che troviamo quando tentiamo di metterci in contatto con un call center.

### 1.2.1 Storia del Cognitive Computing

Per inquadrare al meglio il concetto del Cognitive Computing, è necessario riportare alcuni cenni storici.

Nonostante il concetto di macchine intelligenti esisteva da molto tempo, nel 1950 fu Alan Turing il primo ad affrontare questo problema proponendo il "test di Turing", riportato nell'articolo *"Computing machinery and intelligence"*.

Il test di Turing è un criterio per stabilire se una macchina sia intelligente. Esso fa riferimento al gioco dell'imitazione in cui si hanno tre partecipanti, un uomo A, una donna B ed un terzo soggetto C. Quest'ultimo è tenuto separato dagli altri due e tramite una serie di domande deve stabilire qual è l'uomo e quale la donna. Dal canto loro anche A e B hanno dei compiti: A deve ingannare C e portarlo a fare un'identificazione errata, mentre B deve aiutarlo.

Turing propose di seguire questo procedimento prima normalmente e poi sostituendo A con una macchina. Se la percentuale di volte in cui C indovina chi sia l'uomo e chi la donna è simile prima e dopo la sostituzione di A con la macchina, allora quest'ultima dovrebbe essere considerata intelligente, dal momento che è "indistinguibile" dall'essere umano.

Per quanto riguarda le date precise, possiamo distinguere tre diverse ere dell'informatica, corrispondenti ad invenzioni e periodi storici ben precisi. La Figura 1.7 mostra i tre range in cui è diviso l'arco temporale.

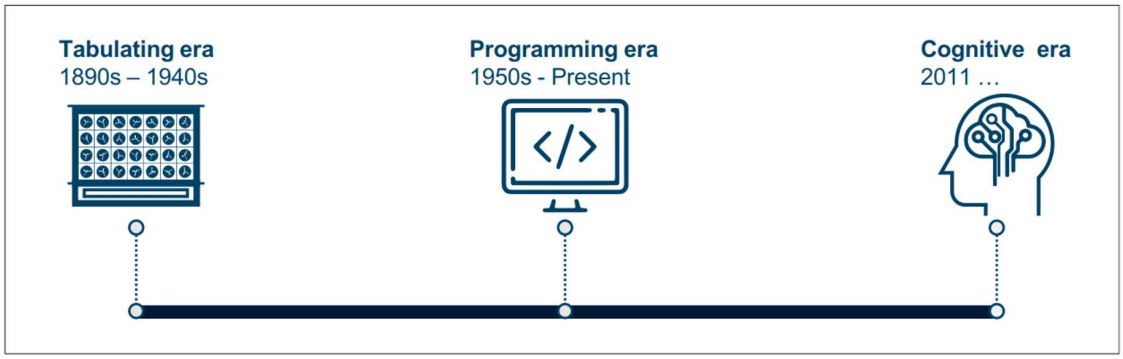


**Figura 1.7:** Ere dell'Informatica sulla linea del tempo

Come è possibile vedere nella figura, ciascuna era fa riferimento ad un periodo temporale ben preciso; analizziamole più nel dettaglio.

*L'era della tabulazione* deve il suo nome al funzionamento dei primi sistemi elettromeccanici. Questi venivano utilizzati per memorizzare dati attraverso l'utilizzo di schede perforate che, opportunamente lette dalla macchina, identificavano il compito da svolgere. Sorprendentemente, l'uso delle schede perforate risale a prima dei computer, basti pensare che già nel 1725 venivano utilizzate da Basile Bouchon per regolare l'andamento dei telai sulla stoffa.

Questo venne a sua volta migliorato nel 1726 dal suo collega Jean-Baptiste Falcon utilizzando sequenze di schede. Nel 1837 Charles Babbage, artefice dell'idea di una macchina calcolatrice programmabile, adottò il sistema a schede perforate di Jacquard (1801, a sua volta basatosi sulle due precedenti) per il controllo della sequenza di calcoli nel progetto della sua macchina analitica.

Il problema di queste macchine era il fatto di poter svolgere un solo compito per volta. Le macchine di tabulazione si sono evolute nelle "unit record equipment" (apparecchiature periferiche collegate al computer che leggono o scrivono record di unità), e hanno dato luogo al settore industriale dell'elaborazione dei dati.

*L'era della programmazione* segna il passaggio dai tabulatori meccanici ai sistemi elettronici, sviluppati durante la Seconda Guerra Mondiale per scopi bellici.

A partire dall'immediato dopoguerra, questi "computer" digitali si sono evoluti molto velocemente arrivando ad interferire con organizzazioni importanti di imprese e governi.

Questo diede il via all'era dell'informatica programmabile, principalmente dovuta alla possibilità di riprogrammare i sistemi informatici per adattarli a problemi diversi dai precedenti.

Come appare evidente, la programmazione richiede la capacità dell'essere umano. Un ulteriore passo in avanti si è fatto in questi utlimi anni con l'introduzione di sistemi intelligenti che hanno dato l'input per *l'era cognitiva*.

Nonostante molti esperti credano che l'era della programmazione continuerà ad esistere indipendentemente, è naturale ammettere che l'essere umano, che si contraddistingue per le sue qualità profonde legate al pensiero, alla capacità di analizzare i dati e nella risoluzione di problemi particolarmente complessi, abbia qualche lacuna nell'elaborare enormi moli di dati.

Dunque l'informatica cognitiva punta ad estendere il cervello umano grazie all'intervento di sistemi intelligenti che sono in grado di eseguire milioni di calcoli al secondo.

John E. Kelly III, nel suo articolo *"Computing, cognition and the future of knowing: How humans and machines are forging a new age of understanding"* definisce così l'informatica cognitiva:

> *"L'informatica cognitiva si riferisce a sistemi che imparano su scala, ragionano con un obiettivo e interagiscono con gli esseri umani in modo naturale."*
>
> *"Piuttosto che essere esplicitamente programmati, imparano e ragionano dalle loro interazioni con noi e dalle loro esperienze con il loro ambiente."*

### 1.2.2 Caratteristiche generali del cognitive computing

L'obiettivo del Cognitive Computing è quello di realizzare strutture informatiche in grado di risolvere problemi senza l'intervento da parte dell'uomo.

A tal proposito il Cognitive Computing Consortium ha raccomandato che i sistemi cognitivi abbiano le seguenti caratteristiche:

- *Adattività*. Primo passo da realizzare per un sistema basato sul Machine Learning le cui soluzioni dovrebbero imitare le capacità del cervello umano di imparare dall'ambiente circostante e di adattarsi all'ambiente circostante. Questo perché i sistemi non devono essere applicabili ad un singolo task ma debbono essere dinamici nel raccogliere i dati, analizzarli e fornire soluzioni.

- *Interattività*. Equivale a dire che un sistema di Cognitive Computing deve sfruttare ed interagire con tutte le potenzialità messe a sua disposizione, i dispositivi, i servizi cloud, gli utenti etc. I sistemi cognitivi dovrebbero essere interattivi fra loro, cogliere l'input umano e fornire risultati attendibili attraverso tecniche di Deep Learning e Natural Language Processing.

- *Iteratività e gestione degli strati*. Il sistema dovrebbe ricordare gli input precedenti e restituire risultati affini in quel momento. Dovrebbe essere in grado di formulare il problema ponendo domande o trovando una fonte alternativa. Questo diventa plausibile applicando particolari metodologie e validazione dei dati così da assicurarsi che il sistema abbia sempre una risorsa a cui attingere e soprattutto, che le fonti forniscano input affidabili ed aggiornati.

- *Contestualizzazione*. I sistemi di Cognitive Computing devono capire, identificare ed estrarre elementi contestuali come il tempo, il luogo, il dominio, il significato o la sintassi. Possono attingere a fonti di informazione, incluse informazioni digitali strutturate e non, così come gli input sensoriali.

L'introduzione della tecnologia cognitiva ha consentito la risoluzione di problemi che non si erano risolti precedentemente, quali il riconoscimento di singoli oggetti nelle immagini o la comprensione del linguaggio naturale.

Secondo uno studio condotto dall'IBM Institute for Business Value - *"Your cognitive future. How next-gen computing changes the way we live and work"*, esistono tre ampie aree di capacità per i sistemi cognitivi.

Aprendo nuovi orizzonti per l'innovazione, queste tre aree sono strettamente correlate al modo di pensare e lavorare delle persone. È importante notare che queste capacità non sono esclusive. La Figura 1.8 fornisce una rappresentazione di tali capacità.

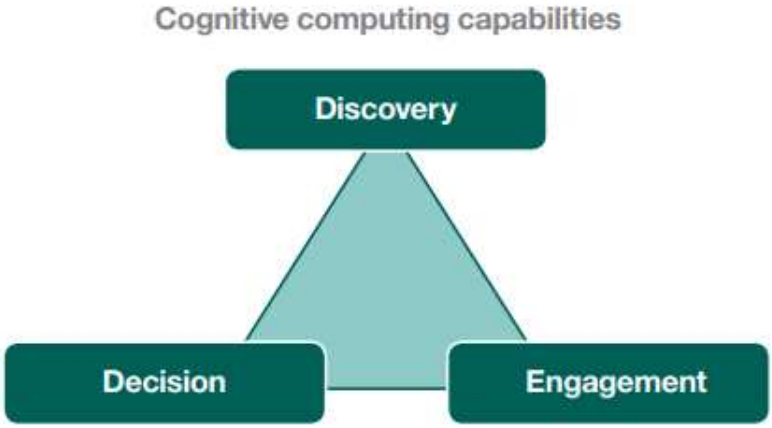


**Figura 1.8:** Le tre "capability areas" per un sistema cognitivo

In particolare notiamo come queste siano: *Ingaggio (Engagement), Decisione (Decision) e Scoperta (Discovery)*. Analizziamole più nel dettaglio:

- *Ingaggio*. Questi sistemi cambiano letteralmente il modo in cui essi interagiscono con gli umani, ed estendono le capacità dell'umano sfruttando le sue abilità per fornire assistenza e comprensione. Possono conciliare dati ambigui e, persino, autocontraddittori. Sono sistemi in grado di impegnarsi in un dialogo con l'uomo; si pensi, ad esempio, ai ChatBot, molti dei quali sono modelli pre-addestrati con la conoscenza del dominio. Quí i sistemi cognitivi assumono un ruolo di assistenti. In questa partnership, la coppia uomo-macchina è più efficace rispetto al singolo uomo o alla singola macchina. Un esempio importante è costituito da quanto messo in atto negli USA, dove è stato proposto un sistema cognitivo che assiste i membri militari durante la transizione nella loro vita civile. In pratica, USAA, una compagnia finanziaria, fornisce servizi bancari e di assistenza a 10,4 milioni di membri della "U.S Armed Forces" ed ai loro nuclei familiari, inclusi veterani, che spesso risentono della difficoltà nel passaggio dalla vita militare a quella civile. Per meglio supportare i propri assistiti, USAA ha implementato un innovativo sistema cognitivo che sfrutta IBM Watson. Il sistema permette ai membri di visitare una pagina web in cui è possibile chiedere domande su argomenti ben specifici del settore.

- *Decisione*. Questi sistemi hanno capacità decisionali. Le decisioni prese da sistemi cognitivi sono basate su nuove informazioni, risultati ed azioni. Esse sono, inoltre, prive

di pregiudizi. Questi sistemi sono modellati utilizzando la tecnica del reinforcement learning. Le decisioni prese si evolvono continuamente grazie all'acquisizione di nuove informazioni. La confidenza nell'abilità di scelta di un sistema cognitivo dipende anche dall'attitudine di domandare ed avere tracciabilità sul perché una particolare decisione è stata presa. Il valore di confidenza è l'output prodotto da un sistema cognitivo in merito ad una decisione presa dopo aver valutato molteplici opzioni. Un esempio è quello in cui vengono applicate soluzioni di sistemi cognitivi per supportare decisioni volte a migliorare la "patient care". WellPoint, una delle più vaste compagnie di salute e benessere negli Stati Uniti, inoltra un grande numero di soluzioni legate al benessere della salute attraverso la sua rete nazionale. Il management degli infermieri spende in media il 40/60 percento del proprio tempo accumulando informazioni ricevute, per decidere se le richieste di procedure debbano essere approvate o rifiutate in base alla propria politica. Per decisioni più complesse i pazienti devono spesso attendere settimane e, questo dispendio di tempo può apportare ritardi nei trattamenti o addirittura errori. Per rimediare a ciò, WeelPoint ha implementato un sistema cognitivo introdotto da IBM Watson per fornire supporto nelle decisioni. Il sistema si basa sulla sua abilità di interpretare significati ed analizzare domande in ambito medico e con l'interazione umana.

- *Scoperta*. Questi sistemi possono scoprire intuizioni che farebbero fatica ad essere scoperte anche dalle migliori menti umane. La "scoperta" include la ricerca di approfondimenti e collegamenti, nonché la coprensione della grande quantità di informazione disponibile. Questi modelli sono costruiti sul deep learning e sull'unsupervised machine learning. Con l'aumento della mole di dati, è chiaro che si ha bisogno di sistemi che ci aiutino nel ricercare l'informazione necessaria molto più facilmente di come sapremmo fare da soli. Passi in avanti in questo ambito sono stati fatti nella ricerca medica. Baylor College of Medicine, leader delle università scientifiche, è alla costante ricerca di approcci innovativi per migliorare ed accelerare le ricerche legate alla medicina. Il tempo richiesto ai ricercatori per testare una ipotesi e formulare una conclusione, in genere, può essere di qualche giorno così come di qualche anno. I biologi e i data scientist dell'univeristà di Baylor hanno sfruttato un sistema cognitivo basato su IBM Watson per il loro *"Baylor Knowledge Integration Toolkit (KnIT)"*, così da accelerare le ricerche. Il sistema è allenato per pensare come un ricercatore umano arrivando ad approfondimenti, visualizzando possibilità e validando teorie a velocità maggiori. Un ulteriore esempio è quello relativo alla *"Shell Cognitive Information Management (CIM)"*, della Louisina State University. Quì gli agenti intelligenti presenti nel modello raccolgono dati in streaming, come testo e video, e li utilizzano per creare un sistema interattivo di rilevamento, ispezione e visualizzazione per il monitoraggio e l'analisi in tempo reale.

Come si evolveranno queste tre capacità cognitive dipenderà principalmente da cinque fattori riportati nella Figura 1.9:

Esse sono:

- *Personalized interaction (Interazione personalizzata)*. Gli attuali sistemi cognitivi sono prevalentemente passivi, ossia necessitano di un primo input da parte dell'uomo per generare una risposta. Questa interazione è basata su scambi di messaggi, ma nulla esclude il fatto che in futuro essa possa diventare più interattiva e, quindi, sfruttare direttamente la voce.

- *Learning (Apprendimento)*. Attualmente i sistemi cognitivi sono pre-addestrati (supervised learning). Essi fanno affidamento sull'uomo e su esperti del settore richiesto per

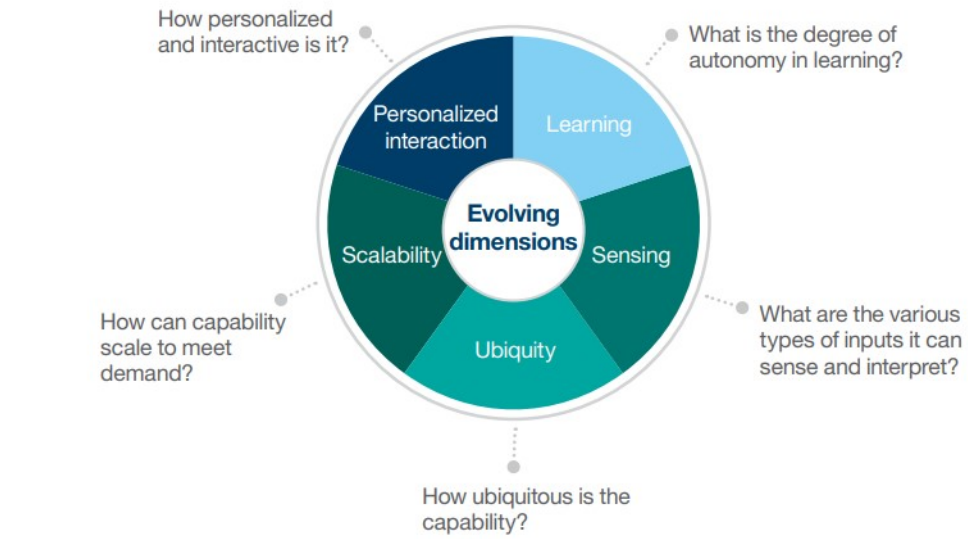


**Figura 1.9:** Le cinque chiavi legate all'evoluzione delle capacità cognitive

addestrarli. In futuro i sistemi adotteranno l'unsupervised learning che richiederà meno coinvolgimento da parte dell'uomo durante il processo di addestramento.

- *Sensing (Rilevamento)*. I sistemi odierni lavorano principalmente con il Natural Language Processing e richiedono abilità legate a quest'ultimo per un determinato linguaggio. I futuri sistemi saranno in grado di gestire una varietà di media ben oltre il testo (audio, immagini, video, etc).

- *Ubiquity (Onnipresenza)*. I moderni sistemi cognitivi vengono sviluppati per garantire una fruizione attraverso portali web, app mobile e servizi cloud. In futuro, questi sistemi saranno, per così dire, "onnipresenti". Potrebbero esserci marketplace con milioni di agenti di sistemi cognitivi o avatar.

- *Scalability (Scalabilità)*. I sistemi cognitivi continuano ad aumentare in scalabilità per supportare vasti campi di applicazione. Basti pensare che nell'ultima versione, IBM Watson è 24 volte più veloce, ha avuto un incremento del 2400 percento in performance ed è diventato più piccolo del 90 percento rispetto alla prima versione che, nel 2011, si laureò campione del game show "Jeopardy!".

### 1.2.3 Uno sguardo al futuro

Il futuro del Cognitive Computing è influenzato da fattori esterni, come l'evoluzione tecnologica e dei suoi trend.

Sei forze saranno quelle che influenzeranno maggiormente il futuro del Cognitive Computing; esse sono mostrate nella Figura 1.10:

Tali forze sono:

- *Society*. A livello sociale vi saranno due fronti opposti.
  - Il primo in favore della tecnologia, che è propenso all'adozione di macchine intelligenti con la possibilità di interagire con esse anche attraverso l'utilizzo di dispositivi mobili.

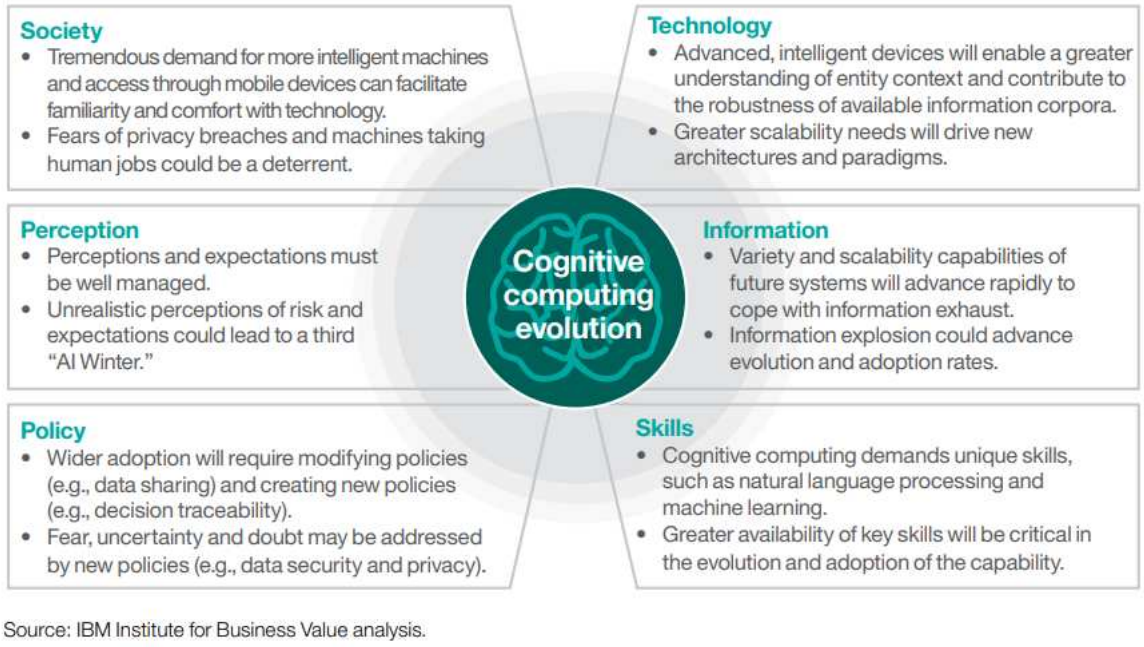


**Figura 1.10:** Le sei forze che influenzeranno il futuro del cognitive computing

  - Il secondo, più moderato, che tenta di rallentare lo sviluppo di questi sistemi cognitivi così da permetterne una più ampia comprensione.

- *Technology*. C'è una forte credenza, da parte degli esperti, secondo i quali le attuali architetture ed i paradigmi di programmazione dovrebbero progredire così da portare il Cognitive Computing ad un livello superiore. Queste tecnologie a cui si fa riferimento, che includono il Natural Language Processing, algoritmi di unsupervised Machine Learning e dispositivi di realtà virtuale potrebbero aiutare in questa evoluzione.

- *Perception*. La proposta del Cognitive Computing è avvincente, e molte organizzazioni stanno già realizzando un valore economico. Tuttavia, bisogna essere realistici, la presenza di due opinioni contrastanti e la disinformazione potrebbero condurre ad un altro "AI Winter", stereotipo per indicare un periodo segnato dalla riduzione di interesse in ambito di ricerca legata all'intelligenza artificiale.

- *Information*. Fu previsto che l'universo digitale avrebbe raggiunto i 40 zettabyte (ZB, settima potenza di mille) entro il 2020, pari a 57 volte la quantità di granelli di sabbia di tutte le spiagge presenti sulla terra. Questa esplosione di informazione ha accelerato la crescita del calcolo cognitivo. Per la gestione di questa informazione, il calcolo cognitivo sarà, quindi, forzato ad evolvere più velocemente.

- *Policy*. Un più largo utilizzo del Cognitive Computing attraverso i domini richiederà l'avanzamento delle politiche (condivisione dei dati, sicurezza dei dati e privacy).

- *Skills*. Un fattore chiave per il progresso del Cognitive Computing sarà la disponibilità di personale specializzato. I sistemi cognitivi avanzati e l'implementazione di questi richiedono conoscenze uniche nel settore, come, ad esempio, quelle legate al Machine Learning ed al Natural Language Processing.

### 1.2.4 Limiti del Cognitive Computing

Una volta analizzati gli aspetti positivi del Cognitive Computing, vediamo, ora, quali sono i suoi limiti:

- *Analisi dei rishi limitata*. I sistemi cognitivi non sono in grado di analizzare particolari rischi se questi non sono esplicitamente espressi nei dati analizzati. A tal fine, l'intervento umano per un'analisi completa del rischio è sempre richiesta.

- *Processo di training meticoloso*. Affinché questi sistemi arrivino a sviluppare comportamenti a noi utili, è necessario che abbiano alle proprie spalle un bagaglio di conoscenza per comprendere completamente il processo e migliorare costantemente. Ciò richiede un processo di addestramento piuttosto laborioso e, probabilmente, è questo uno dei motivi per i quali essi non sono ancora pienamente utilizzati. A conferma di ciò basta ripensare al caso della WellPoint (Sezione 1.2.2), in cui il personale infermieristico continua ad alimentare i casi fino a quando il sistema capisce completamente una determinata condizione medica. Questa nota negativa è incrementata dall'elevato costo del processo di utilizzo di questi sistemi.

- *Più aumento dell'intelligenza che un'intelligenza artificiale*. La situazione attuale dei sistemi cognitivi è limitata dall'Engagement (Ingaggio) e dalla Decision (Decisione). I sistemi di questo tipo sono molto più efficaci come assistenti, quindi è richiesto più un incremento di intelligenza che una intelligenza artificiale. In più sono sempre dipendenti dagli esseri umani, infatti sono basati su un'interazione uomo-macchina che integra il pensiero e l'analisi umana.

Il prossimo passo è sicuramente rivolto verso il cognitive computing, ma l'impossibilità di applicare l'Intelligenza Artificiale in situazioni di incertezza o rapidi cambiamenti non consente una rapida diffusione di essa.

La complessità è direttamente proporzionale alla quantità di dati a disposizione; infatti è un procedimento oneroso aggregare ed analizzare molti dati non strutturati.

La gestione di dati dinamici richiede una soluzione cognitiva complessa che dovrebbe avere molte tecnologie a disposizione in grado di fornire intuizioni profonde del dominio.

CAPITOLO 2

# Le principali funzionalità del Cognitive Computing

*Nel secondo capitolo è ripreso il tema del Cognitive Computing già accennato nel precedente, nel tentativo di esplicitare maggiormente questo argomento ed eliminare ulteriori incertezze. In particolare, verranno espressi i concetti base del Cognitve Computing e le sue caratteristiche, concludendo con la descrizione dei principali sistemi di Cognitive Computing già presenti sul mercato. Per quest'utlima parte, verranno analizzati i servizi più importanti di cui i sistemi citati sono provvisti, con vista particolare a quelli forniti da Watson di IBM. In chiusura, è prevista la descrizione di ciascuno di questi servizi, con piccoli accenni e riferimenti ad altri sistemi, messi a punto da altre aziende come Google, Microsoft ed Amazon, che verranno utilizzati e discussi anche nei capitoli successivi.*

## 2.1 Concetti base del Cognitive Computing

In un modello di Cognitive Computing sono presenti le seguenti tecnologie:

- *Big Data:* raccolta di dati informatici così estesa in termini di volume, velocità e varietà, da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore o conoscenza.
- *Quetion Answering:* QA, consiste nel rispondere automaticamente ad una domanda espressa in un linguaggio naturale.
- *Machine Learning*.
- *Natural Language Processing (NLP)*.
- *Cloud Computing:* indica un'erogazione di servizi offerti su richiesta da un fornitore ad un utente finale attraverso la rete Internet (come l'archiviazione, l'elaborazione o la trasmissione dati), a partire da un insieme di risorse preesistenti, configurabili e disponibili in remoto, sotto forma di architettura distribuita.
- *Application Programming Interface (API):* si indica un insieme di procedure atte a risolvere uno specifico problema di comunicazione tra diversi computer, o tra diversi software, o tra diversi componenti di software; spesso tale termine designa le librerie software di un linguaggio di programmazione, sebbene, più propriamente, le API sono il metodo con cui le librerie vengono usate per sopperire ad uno specifico problema di scambio di informazioni.

Esse sono combinate, togliendo la necessità per gli utenti di essere esperti in discipline differenti e consentendo loro di focalizzarsi nella ricerca di soluzioni migliori.

Il modello di Cognitive Computing mira ad avere capacità importanti in vari settori, cosicché gli utenti non debbano concentrarsi ad apprendere i dettagli relativi agli strumenti richiesti, ma possano, invece, spendere il loro tempo per individuare soluzioni alternative, prendere decisioni ed intraprendere azioni volte a migliorare i processi aziendali ed operativi.

Come descritto inizialmente, il Cognitive Computing imita il pensiero umano, dunque la qualità dell'output è buona solo nel caso in cui gli algoritmi vengano utilizzati fin dall'inizio; questi modelli sono poi migliorati grazie al Machine Learning.

In ambito di ricerca e di studio, un esperto umano impiegherebbe settimane per analizzare enormi volumi di dati, viceversa il modello informatico potrebbe farlo in pochi secondi. Per ottenere una risposta più soddisfacente è spesso richiesto fornire ulteriori fonti di dati che, a loro volta, richiederebbero maggiori quantità di tempo per essere studiati e compresi.

Tralasciando per un attimo il problema legato alla quantità dei dati ed al tempo impiegato per analizzarli, l'utilizzo di strumenti analitici, richiesti per mettere in atto questi procedimenti, richiede che esperti del settore diventino esperti di computer.

In merito a quest'ultima osservazione, il Cognitive Computing ha, tra gli obiettivi primari futuri, quello di richiedere solo capacità di conversazione da parte dell'esperto del settore, così da permettergli di trarre preziose intuizioni.

Con il trascorrere del tempo, tecnologie come il Natural Language Processing ed il Question Answering hanno fatto grandi passi in avanti nell'identificazione dei modelli vocali e nella comprensione di ciò che l'utente esprime.

La gestione di queste grandi quantità di dati è semplificata dal Cloud Computing. Diversi fornitori, infatti, hanno creato ambienti di Cloud Computing in cui gli utenti possono richiedere i servizi di cui hanno bisogno, e tali piattaforme cloud forniscono agli utenti l'accesso ai dati in loro possesso.

Svariate sono le API che consentono l'accesso ai servizi, permettendo così, un utilizzo rapido, facile ed intuitivo. La maggior parte di queste API sono indipendenti dal linguaggio di programmazione, concedendo ai programmatori di lavorare in qualsiasi linguaggio di programmazione desiderino.

L'utilizzo delle API per la condivisione dei dati, dei servizi e delle funzioni aziendali tra endpoint (ad esempio applicazioni, dispositivi o siti web) permette la riduzione dei costi e dei tempi di integrazione.

### 2.1.1 Sistemi di Cognitive Computing

L'attuale panorama del Cognitive Computing è dominato da grandi player come IBM, Microsoft e Google. IBM ha investito ben 26 miliardi di dollari in *"Big data e Analytics"*, spendendo, attualmente, oltre un terzo del suo budget in ricerca e sviluppo.

IBM e Google hanno acquisito alcuni dei loro rivali, portando il mercato verso un consolidamento. Elenchiamo, di seguito, alcuni principali prodotti in questo mercato:

- *IBM Watson*. Originariamente Watson era un supercomputer IBM basato sull'Intelligenza Artificiale ed un software analitico per la vittoria del quiz televisivo Jeopardy!. Attualmente, Watson sfrutta un insieme di tecnologie come l'elaborazione del linguaggio naturale, il riconoscimento delle immagini, l'analisi del testo e gli agenti virtuali. Esso si basa su una profonda analisi del contenuto e sul ragionamento basato sulle prove che, combinato con tecniche di elaborazione probabilistica, può consentire ad esso di migliorare il processo decisionale, ridurre i costi ed ottimizzare i risultati.

- *Microsoft Cognitive Services*. Precedentemente noti come "Project Oxford", sono una raccolta di API, SDK (Software Development Kit) e servizi cognitivi che sono messi a

disposizione degli sviluppatori per rendere le loro applicazioni più intelligenti. Con i Cognitive Services, gli sviluppatori possono aggiungere alle applicazioni servizi intelligenti, come rilevamento di emozioni e del sentimento, la visione ed il riconoscimento vocale, la conoscenza, la ricerca e la comprensione del linguaggio. Inoltre Microsoft, come IBM, fornisce anche tool innovativi per la realizzazione di chatbot.

- *Google DeepMind*. DeepMind, leader nel settore dell'AI, grazie anche ad i suoi esperti molto rinomati nel campo delle deep neural network, del reinforcement learning e dei modelli ispirati alle neuroscienze, fu acquisita da Google nel 2014. DeepMind diventò popolare grazie ad AlphaGo, primo programma di AI a battere un giocatore umano nel gioco del Go nell'ottobre del 2015.
- *CognitiveScale*. Fu fondata da ex membri del team IBM Watson. Essa fornisce un software cognitivo cloud per le imprese, supportandole nella gestione dei dati disordinati, disparati, di prime e terze parti. Questa piattaforma ricava intuizioni utilizzabili da tali dati e li utilizza in un processo di apprendimento continuo. In particolare, la piattaforma di intelligenza aumentata di CognitiveScale, offre insights-as-a-service e accelera la creazione di applicazioni cognitive in ambiti come la sanità, il retail, i viaggi e i servizi finanziari.
- *SparkCognition*. È una startup fondata ad Austin nel 2014. Essa sviluppa software di cyber-sicurezza potenziato dall'Intelligenza Artificiale per la safety, la security e l'affidabilità di IoT (Internet of Things). La sua tecnologia è ben vista soprattutto in ambito manifatturiero, in quanto è in grado di sfruttare i dati restituiti dai sensori in tempo reale, e di imparare da essi continuamente, consentendo una più accurata mitigazione del rischio e politiche di prevenzione per prevenire disastri.

Il successo ottenuto da Watson e DeepMind ha ispirato altre aziende a sviluppare piattaforme cognitive, tra queste citiamo *Qualcomm* e *Intel*, che stanno muovendo i primi passi per cercare di includere soluzioni cognitive in settori industriali specializzati.

*Uber*, che ha istituito un settore di ricerca dedicato all'AI ed al Machine Learning, acquisendo *Geometric Intelligence* (basata sulla generazione di risposte utilizzando dati minori in algoritmi di Machine Learning) e *Otto* (Startup per il trasporto e camion autonomi).

Oppure *Gamalon*, che è riuscita a sviluppare una tecnica di AI denominata *"Bayesian Program Synthesis"*, che permette di addestrare il sistema avendo a disposizione solo pochi dati, e garantendo gli stessi livelli di precisione delle reti neurali.

Anche la sanità prospera di soluzioni cognitive. Basti pensare a startup come *Lumiata* e *Enlitic*, che sono riuscite a sviluppare piccole e potenti soluzioni analitiche in grado di assistere gli operatori sanitari nella diagnosi e nella previsione delle condizioni di una malattia.

Molto recente è l'emissione, da parte di OpenAI, di *ChatGPT*. È un software progettato per simulare una conversazione con un essere umano per la generazione di testo.

Il suo funzionamento si basa su GPT-3 (Generative Pre-trained Transformer 3), un modello di elaborazione del linguaggio naturale (o NLP) sviluppato dalla stessa OpenAI. Attualmente, il software presenta già notevoli evoluzioni, dopo l'introduzione del livello successivo basato su GPT-4.

## 2.2 Principali servizi offerti dai sistemi di Cognitive Computing

Per spiegare le caratteristiche ed i principali servizi offerti dai sistemi di Cognitive Computing, prendiamo in considerazione uno dei primi sul mercato, vale a dire *Watson di IBM*.

Watson combina cinque funzionalità principali, ovvero:

1. Interagisce con le persone in modo più naturale, tenendo in considerazione le loro preferenze.
2. Grazie alla collaborazione con gli esperti, acquisisce velocemente le skill di ciascun settore industriale così da elevare le proprie competenze.
3. Consente ai nuovi prodotti e servizi di percepire, ragionare e conoscere i propri utenti ed il mondo che li circonda.
4. Utilizza i dati per migliorare i processi e le previsioni aziendali, favorendo un aumento dell'efficacia operativa.
5. Potenzia l'esplorazione e la scoperta individuando pattern, opportunità, ed ipotesi fattibili.

Dalle precedenti caratteristiche appare chiaro come Watson sia in grado di comprendere tutte le forme di dati, di interagire in modo naturale con le persone e di imparare e ragionare in scala.

Dati, informazione ed esperienza creano le basi per lavorare con Watson. In Figura 2.1, sono riportate tutte le tipologie di dati che Watson è in grado di analizzare.

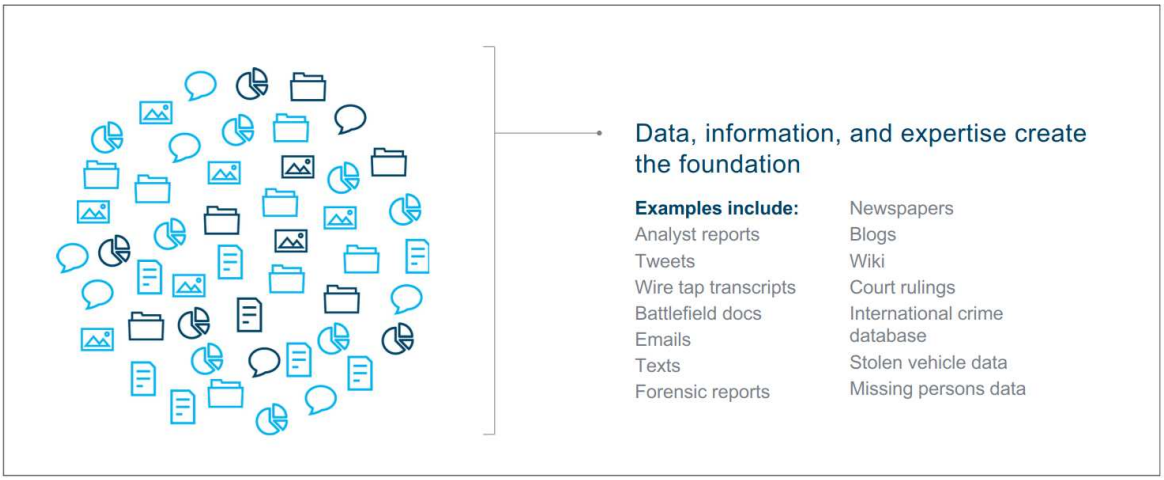


**Figura 2.1:** Esempi di dati ed informazioni che Watson è in grado di analizzare

Watson è fruibile come un insieme di soluzioni *"Software as a Service (SaaS)"* e come un insieme di API, quete ultime fornite tramite *IBM Bluemix*, la *Platform as a Service (SaaS)* in cloud sviluppata da IBM.

Attualmente, le API di Watson disponibili sono le seguenti (riassunte nella Figura 2.2):

- *Language:*
  - *Conversation;*
  - *Document Conversion;*
  - *Language Translator;*
  - *Natural Language Classifier;*
  - *Natural Language Understanding;*
  - *Personality Insights;*
  - *Retrieve and Rank;*

  - *Tone Analyzer;*

- *Speech:*
  - *Speech to Text;*
  - *Text to Speech;*
- *Vision:*
  - *Visual Recognition;*
- *Discovery:*
  - *Discovery;*
  - *Discovery News.*

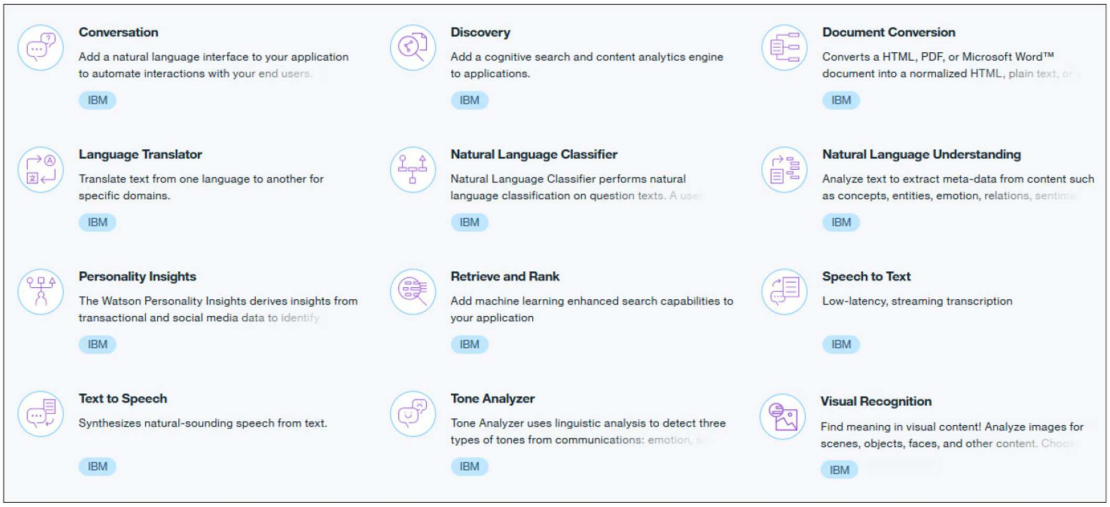


**Figura 2.2:** Servizi offerti da Watson

Prima di analizzare ciascuno di questi servizi, è bene puntualizzare che gli stessi sono ugualmente offerti da altre piattaforme cloud basate su sistemi cognitivi, come *Google Cloud*, *Amazon Web Services (AWS)* e *Microsoft Azure*, che tratteremo in seguito.

È ormai scontato il fatto che molte industrie gestiscano le loro attività utilizzando dati; proprio per questo uno dei principali investimenti di IBM Watson è proprio nel settore industriale, seguito da quello finanziario e da quello scientifico.

Il Cognitive Computing viene applicato in svariati settori; vediamo, di seguito, le aree in cui la tecnologia Watson è maggiormente richiesta:

- *IBM Watson Commerce*. Esso combina l'esperienza aziendale con soluzioni leader del settore, integrate con capacità cognitive. Watson Commerce capisce, ragiona e impara dalla conoscenza collettiva dell'organizzazione e dai trend del settore. Le aziende sono, quindi, in grado di comprendere il comportamento dei clienti e del business, consentendo loro di prendere decisioni tempestive e poco rischiose in ambito di mercato.
- *IBM Watson Education*. Esso comprende soluzioni ed app che contengono al loro interno capacità di analisi e capacità cognitive, che possono aiutare gli insegnanti nel conoscere i loro studenti in maniera globale. In questo modo, educatori e studenti possono dar luogo ad un'esperienza di apprendimento individualizzata, costruendo relazioni e momenti di collaborazione, scambiandosi e accrescendo le competenze nell'ambito dell'istruzione.

- *IBM Watson Financial Services*. È possibile sfruttare la capacità cognitiva di Watson per migliorare la gestione delle conformità alle normative. Questa sua capacità ci consente di oltrepassare i tradizionali criteri basati su regole ed aspetti demografici per la comprensione della redditività, delle preferenze e del ciclo di vita dei clienti, così da proporre offerte ed esperienze innovative e personalizzate.
- *IBM Watson Health*. Esso consente di capire, ragionare ed imparare, aiutando a tradurre informazioni in conoscenza. Watson è in grado di generare nuove intuizioni che accelerano il processo di diagnosi per un paziente. Watson Health fornisce, a sua volta, delle soluzioni più specifiche e dettagliate, quali, ad esempio:
  - *IBM Watson for Genomics*. Watson for Genomics consente ai laboratori di patologia molecolare di scalare i propri programmi oncologici di precisione per soddisfare le esigenze presenti e sempre crescenti di cure oncologiche personalizzate. Watson for Genomics ottiene questo risultato sfruttando l'Intelligenza Artificiale per estrarre dati non strutturati dalla letteratura già recensita, così da aumentare la propria conoscenza.
  - *IBM Watson for Drug Discovery*. Watson for Drug Discovery è una soluzione cognitiva basata sul cloud, che fornisce visualizzazioni dinamiche e previsioni classificate, supportate da prove inerenti ad un largo set di contenuti privati ed eterogenei, come articoli di riviste mediche, libri di testo e brevetti.
  - *IBM Watson Health Patient Engagement*. Le soluzioni automatizzate relative a questo servizio estendono le risorse a disposizione per migliorare gli esiti dei pazienti. Watson identifica le lacune nell'assistenza, coinvolge i pazienti prioritari e misura i progressi.
  - *IBM Watson for Oncology*. IBM Watson for Oncology è un Software as a Service (SaaS) che offre una capacità avanzata per analizzare il significato ed il contesto di dati strutturati e non, in note e report clinici, semplicemente assimilando le parole chiave dei pazienti espresse in un inglese semplice.
  - *IBM Watson Care Manager*. IBM Watson Care Manager è una soluzione di gestione dell'assistenza basata sul cloud, per aiutare le organizzazioni a focalizzarsi sull'assistenza individuale. I team di assistenza possono accedere a dati strutturati e non, selezionare programmi mirati, e creare piani di assistenza personalizzati. Questi ultimi possono essere adattati per affrontare le mutevoli esigenze biologiche, psicologiche e sociali.
- *IBM Watson nel settore assicurativo*. La gestione dei reclami riguardanti le norme e le polizze assicurative richiede la manodopera di valutatori altamente qualificati, che devono rivedere testi, appunti, blog ed altre fonti. Per evitare questa perdita di tempo, le compagnie di assicurazione stanno istruendo Watson nel comprendere le interazioni, le regole e la logica dietro le polizze, così che esso possa analizzare da solo dati strutturati e non strutturati, per formulare raccomandazioni approfondite che possano aiutare i dipendenti nel determinare quali reclami sono ammissibili e quale percentuale deve essere pagata.
- *IBM Watson Internet of Things (IoT)*. Un IoT cognitivo può scegliere le proprie fonti di dati e decidere a quali pattern e relazioni prestare attenzione. Esso utilizza il Machine Learning e l'elaborazione avanzata per organizzare i dati e generare intuizioni.
- *IBM Watson Cognitive Video*. Watson produce video grazie alle sue funzionalità di riconoscimento visivo e di analisi del testo. Esso analizza e impara da altri video per

crearne di nuovi in base alle esigenze dell'utente. Il riconoscimento visivo consente di capire meglio il contenuto delle immagini, consentendone l'individuazione dei volti umani, la determinazione dell'età approssimata e del sesso. Utilizzando alcune API di Watson e alcune tecniche di Machine Learning, gli scienziati della IBM Research, in collaborazione con la 20th Century Fox, hanno creato il primo trailer cognitivo per il film "Morgan". Questo, solo dopo che il sistema ha analizzato centinaia di film horror e thriller, comprendendo cosa affascinava gli spettatori, e riportando le sue intuizioni in questo trailer.

- *IBM Watson for Cyber Security*. I sistemi cognitivi possono facilitare il compito dell'analista fornendo informazioni chiave come visualizzazioni avanzate e analisi interattive delle vulnerabilità. Essi sono in grado di individuare le anomalie ed i difetti logici. I sistemi cognitivi interpretano i dati, aggiungono alla loro base di conoscenze praticamente ogni interazione, soppesano le probabilità sulla base di una profonda conoscenza, e considerano le variabili rilevanti per aiutare l'esperto ad agire, permettendo così una riduzione dei costi e della complessità alla lotta contro la criminalità informatica.

Nell 2007 IBM accettò la sfida di costruire un sistema informatico in grado di competere con i campioni del quiz televisivo *Jeopardy! (Figura 1.4)*. Nel 2011, Watson, sistema di domande e risposte a dominio aperto, batté i due giocatori meglio classificati.

Jeopardy! andò in onda per la prima volta negli Stati Uniti nel 1964; è un quiz in cui tre concorrenti tentano di rispondere a domande di cultura generale.

Ovviamente, l'obiettivo del progetto Watson non era quello di vincere il famoso quiz televisivo, ma quello di fare ricerche nel campo della comprensione del linguaggio naturale e del Machine Learning.

I progressi portati avanti nella tecnologia QA (Question Answering) possono aiutare gli esperti nel prendere decisioni tempestive e corrette.

Naturalmente, per ottenere questi risultati, è necessario addestrare precedentemente questi sistemi cognitivi alla comprensione di nuovi domini. Tale addestramento è impartito da esperti del settore che forniscono una supervisione umana e basi di conoscenze specifiche del dominio.

Un processo analogo deve essere eseguito per Watson. L'adattamento ad un dominio consiste nelle attività necessarie per adattare un sistema aperto ad un dominio specifico. Il meccanismo fondamentale alla base dell'adattamento legato a Watson, è il *supervised learning* (sottocategoria del Machine Learning e dell'Intelligenza Artificiale). Quest'ultimo è definito dall'uso di un set di dati etichettati, per addestrare algoritmi che classificano i dati o prevedono i risultati in modo accurato; esso è alla base della classificazione nel Data Mining.

Un modo alternativo per l'adattamento di Watson a nuovi domini è quello di assorbire le basi di conoscenza, ovvero strutture dati che rappresentano informazioni strutturate utilizzabili da un sistema informatico per effettuare inferenze. Questi due metodi, le basi di conoscenza ed il supervised learning, si complementano a vicenda. In generale, maggiore è la conoscenza fornita e minore sarà l'addestramento richiesto dal sistema per raggiungere buone performance.

In Figura 2.3 sono riportati i servizi di Watson che non possono essere addestrati dall'utente, e per i quali IBM si è assunta la responsabilità di addestramento.

Poiché gran parte dei servizi di Watson si basa su un approccio di supervised learning, per addestrarli potrebbe essere richiesta la possibilità di fornire dati etichettati maunualmente.

In Figura 2.4 invece, sono riportati i servizi Watson che possono essere addestrati per adattarli ad un dominio chiuso.

Nelle prossime sottosezioni esamineremo alcuni di questi servizi.

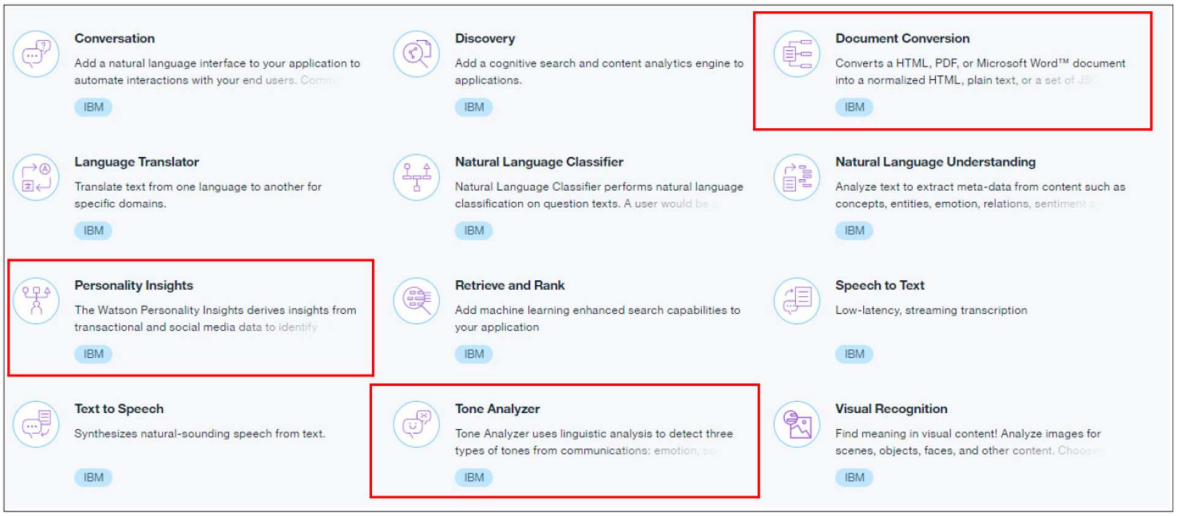


**Figura 2.3:** Servizi di Watson non addestrabili dall'utente

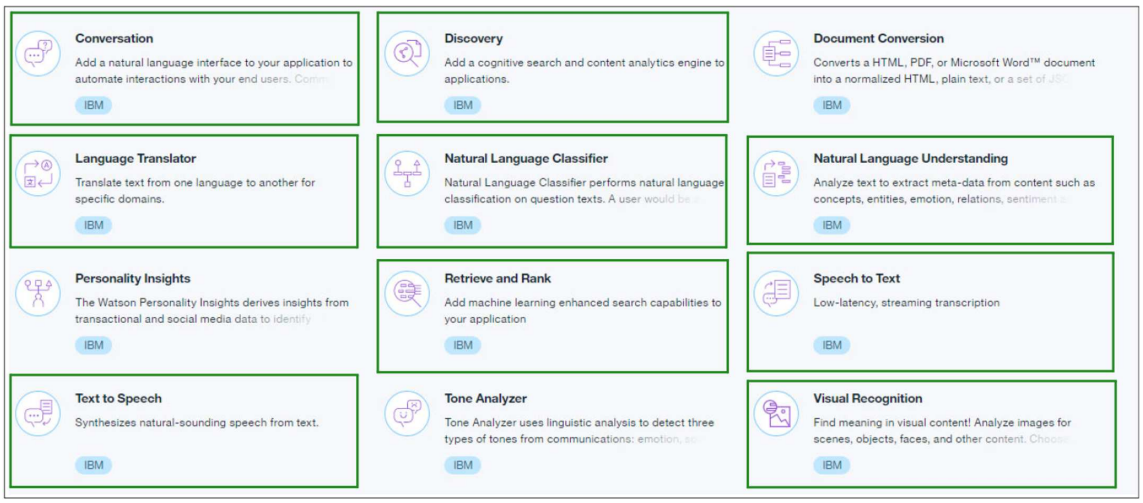


**Figura 2.4:** Servizi di Watson addestrabili dall'utente

### 2.2.1 Watson Conversation

È l'equivalente dei seguenti altri servizi:

- *QnA Maker di Microsoft Azure;*
- *Amazon Lex di Amazon Web Services;*
- *DialogFlow di Google Cloud.*

Con *Watson Conversation* è possibile creare un'applicazione, ed un agente, che comprendono l'input del linguaggio naturale e comunicano con gli utenti simulando una reale conversazione umana. Ciò è reso possibile grazie all'utilizzo di tecniche di Deep Learning.

Il servizio mette a disposizione dell'utente un servizio web che consente di gestire il proprio chatbot attraverso l'utilizzo di tre concetti chiave, ovvero:

1. *#intents*: rappresentano l'obiettivo dell'input dell'utente, ossia ciò che quest'utlimo vuole ottenere.

2. *@entity*: rappresenta un termine o un oggetto rilevante per l'*intent*; esso fornisce il contesto.

3. *Dialog*: consente al servizio di rispondere agli utenti in base ai due elementi precedenti.

L'obiettivo primario per coloro che si cimentano nello sviluppo di un chatbot è quello di anticipare ogni possibile modo in cui gli utenti cercheranno di comunicare con esso.

A sostegno di ciò, vi è una componente, *Improve*, la quale mette a disposizione la cronologia della conversazione con gli utenti; quest'utlima può essere sfruttata per migliorare la comprensione dell'input dell'utente da parte del chatbot.

In Figura 2.5 sono riportati i passi principali per adattare il Conversation service.

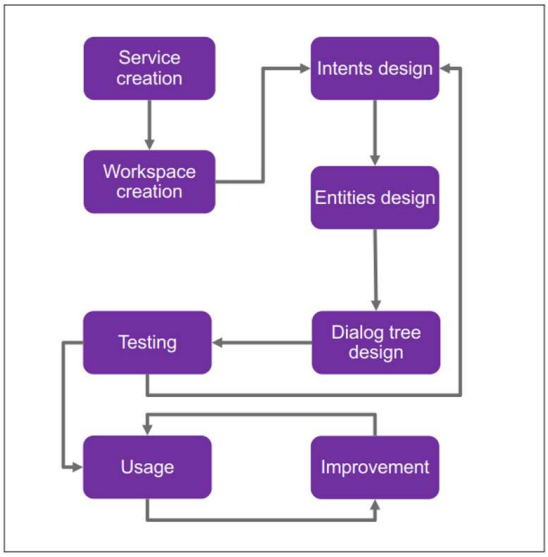


**Figura 2.5:** Passi da seguire per impostare il Conversation service

### QnA Maker di Microsoft Azure

Il QnA Maker è un servizio cloud basato sul Natural Language Processing (NLP), il quale consente di creare conversazioni naturali sui dati. Questo strumento viene utilizzato per trovare la risposta più appropriata per qualsiasi tipo di input attraverso una knowledge base personalizzata, costituita da una serie di domande e risposte.

Solitamente, il QnA Maker viene utilizzato per lo sviluppo di applicazioni client conversazionali, tra cui social media, chatbot e applicazioni desktop con riconoscimento vocale.

Tuttavia, va notato che tale strumento non archivia i dati dei clienti. Tutte le informazioni relative alle risposte inerenti alle domande ed ai log delle conversazioni sono archiviate nell'area di distribuzione del servizio, dipendente dal cliente.

### Amazon Lex di Amazon Web Services

Amazon Lex è un servizio di Intelligenza Artificiale (AI) completamente gestito con modelli avanzati di linguaggio naturale, per progettare, costruire, testare e distribuire interfacce di comunicazione nelle applicazioni.

In Figura 2.6 è riportato un esempio di funzionamento generale che include le funzionalità e le risorse di Amazon Lex.

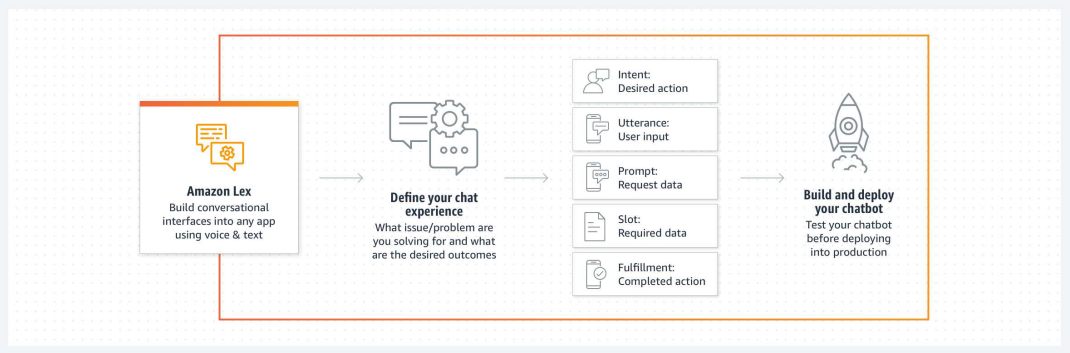


**Figura 2.6:** Funzionamento generale di Amazon Lex

Esso si basa sulle stesse tecnologie impiegate per Alexa, e offre gli strumenti adatti per affrontare i problemi più comuni del Deep Learning, ad esempio il riconoscimento vocale e la comprensione del linguaggio, con un servizio intuitivo completamente gestito.

Amazon Lex offre una soluzione end-to-end intuitiva, sicura e scalabile per la creazione, la pubblicazione e il monitoraggio dei bot. In altre parole, Amazon Lex permette agli sviluppatori di creare facilmente interfacce di comunicazione vocali o testuali, capaci di comprendere le intenzioni dell'utente e di fornire risposte adeguate. Tale servizio è ideale per lo sviluppo di chatbot, assistenti virtuali e altri tipi di applicazioni che richiedono l'interazione con l'utente in modo naturale e intuitivo.

#### DialogFlow di Google Cloud

DialogFlow è un'Intelligenza Artificiale di conversazione naturale che utilizza agenti virtuali avanzati in grado di supportare conversazioni avanzate con i clienti.

Questa piattaforma si basa sulla comprensione del linguaggio naturale, semplificando la progettazione e l'integrazione di un'interfaccia utente conversazionale per app, applicazioni web, dispositivi, bot e sistemi di risposta interattivi vocali.

DialogFlow è in grado di analizzare diversi tipi di input, tra cui input di testo o audio, provenienti da un telefono o da una registrazione vocale.

Grazie alla sua capacità di comprensione del linguaggio naturale, la piattaforma è in grado di comprendere le intenzioni degli utenti e di fornire loro risposte adeguate in modo coerente e preciso.

In questo modo, DialogFlow può aiutare le aziende a migliorare l'esperienza degli utenti, fornendo un'interazione naturale e intuitiva con i propri prodotti e servizi.

DialogFlow può analizzare più tipi di input, inclusi input di testo o audio (come da un telefono o da una registrazione vocale).

### 2.2.2 Watson Language Translator

È l'equivalente dei seguenti altri servizi:

- *Translator di Microsoft Azure;*
- *Amazon Translate di Amazon Web Services;*
- *Translation AI di Google Cloud;*
- *DeepL.*

Watson Language Translator traduce il testo espresso in una determinata lingua, in un'altra. Di seguito una vista dei tre modelli linguistici che sono forniti dal servizio di Watson Language Translator:

1. *Notizie*. Destinato ad articoli e trascrizioni di notizie, è in grado di tradurre l'inglese in varie lingue, tra cui l'arabo, il francese, il tedesco etc. , nonché di tradurre lo spagnolo da e verso il francese.

2. *Conversazioni*. Destinato alle conversazioni ed ai colloqui. Anche quì, come nel caso precedente, è in grado di tradurre l'inglese in molteplici lingue.

3. *Brevetti*. Indirizzato alla terminologia tecnica e giuridica dei brevetti. Traduce lingue come il portoghese brasiliano, cinese e lo spagnolo, verso l'inglese.

Questo servizio è destinato a migliorare nel tempo attraverso tecniche di addestramento in cui il Language Translator impara dalle traduzioni precedenti.

Watson Language Translator è uno tra i migliori translate, vista anche la sua capacità di comprendere termini e frasi specifici.

Inoltre, esso fornisce vari strumenti e meccanismi utili, quali, ad esempio, un meccanismo per addestrare il servizio a migliorare la traduzione in un determinato dominio, oppure la funzionalità per aggiornare modelli preesistenti e migliorare la loro qualità.

I modelli di traduzione forniti per l'utilizzo di questo servizio vengono aggiornati con l'aggiunta di un file sorgente di input, e la personalizzazione dipende dal tipo e dal contenuto di quest'ultimo.

#### Translator di Microsoft Azure

Translator è un servizio di traduzione automatica neurale basato sul cloud, che fa parte della famiglia delle API REST di Servizi cognitivi di Azure, e può essere usato con qualsiasi sistema operativo.

Integra facilmente le funzionalità di traduzione di testi in tempo reale nei siti Web, negli strumenti o in qualsiasi soluzione dell'applicazione che richiede un supporto multilingue, ad esempio localizzazione di siti Web, e-commerce, supporto tecnico, applicazioni di messaggistica, comunicazione interna ed altro ancora.

Di seguito sono riportate le funzionalità supportate dal servizio Translator:

- *Traduzione di testo*. Eseguire la traduzione del testo tra le lingue di origine e di destinazione supportate, in tempo reale.

- *Traduzione di documenti*. Tradurre file batch (è un file di testo che contiene una sequenza di comandi per l'interprete dei comandi del sistema) e complessi, mantenendo la struttura e il formato dei documenti originali.

- *Custom Translator*. Creare modelli personalizzati per tradurre il linguaggio, la terminologia e lo stile specifici del settore.

#### Amazon Translate di Amazon Web Services

Amazon Translate è un servizio di traduzione automatica neurale completamente gestito, che utilizza modelli di Deep Learning per produrre traduzioni di alta qualità e a costi contenuti. Questo servizio supporta la traduzione di testo in tempo reale tra diverse lingue, consentendo a utenti e aziende di raggiungere un pubblico globale in modo efficiente e preciso.

I modelli di Deep Learning utilizzati da Amazon Translate sono in grado di analizzare il contesto della frase, e di tradurre le parole e le espressioni in modo più preciso e naturale rispetto ai tradizionali algoritmi di traduzione basati su regole. Inoltre, Amazon Translate offre una vasta gamma di funzionalità personalizzabili tra cui la possibilità di creare vocabolari personalizzati, di tradurre documenti interi e di integrare il servizio con altre applicazioni e piattaforme.

Grazie ad Amazon Translate, le aziende possono raggiungere un pubblico globale in modo efficiente e preciso, offrendo prodotti e servizi in diverse lingue, senza dover affrontare le spese e la complessità associate alla traduzione manuale.

Ecco alcuni dei possibili casi d'uso:

- *Localizzazione linguistica*. Per i team di traduzione umana è difficile rimanere al passo dei contenuti generati dagli utenti. Grazie ad Amazon Translate è possibile tradurre enormi quantità di dati in tempo reale.
- *Analisi del testo*. Il servizio, infatti, consente di apprendere le tendenze relative ad un marchio, un prodotto o un servizio sui social, monitorando messaggi in diverse lingue. È sufficiente tradurre il testo in inglese ed utilizzare un'applicazione di elaborazione del linguaggio naturale, ad esempio Amazon Comprehend, per analizzare testi in una serie di lingue.
- *Comunicazione*. La traduzione automatica fornita da Amazon Translate consente agli utenti che parlano lingue diverse di comunicare tra loro. Aggiungendo la traduzione in tempo reale a chat, e-mail, helpdesk e sistemi di ticketing, un agente o un dipendente che parla inglese può comunicare con i clienti in più lingue.

**Translation AI di Google Cloud**

Il sistema *Translation AI* di Google è disponibile principalmente in quattro varianti:

1. *Translation Hub*. Traduce facilmente i contenuti in 135 lingue, con un'interfaccia utente intuitiva e facile da usare, e integra il feedback umano quando necessario.
2. *AutoML Translation*. Sviluppatori, traduttori ed esperti di localizzazione con conoscenza del dominio, possono utilizzare la tecnologia AutoML per creare modelli di traduzione personalizzati senza scrivere una sola riga di codice.
3. *API Translation*. L'API Translation Basic utilizza la tecnologia di traduzione automatica neurale di Google per tradurre istantaneamente i testi in più di 100 lingue. La versione Advanced dell'API Translation offre gli stessi risultati rapidi e dinamici proposti dalla versione Basic, oltre a funzionalità aggiuntive per la personalizzazione.
4. *API Media Translation*. L'API Media Translation offre la traduzione dell'audio in tempo reale direttamente nei contenuti e nelle applicazioni, con maggiore precisione e integrazione semplificata.

**DeepL**

DeepL è considerato uno dei migliori traduttori in circolazione, grazie anche alle sue svariate funzionalità innovative che introduce.

Infatti ne sono un esempio le traduzioni alternative, che permettono di risparmiare il tempo di cercare parole o frasi diverse per la traduzione. Basta cliccare su una parola, scegliere tra le opzioni alternative, e il testo circostante sarà adattato automaticamente.

Un'altra funzionalità importante, è quella del poter tradurre interi file di vari formati, la cui dimensione però, non superi i 20MB.

Altra comodità importante, ad esempio, è quella di poter specificare come debbano essere tradotti alcuni termini, ovvero, risparmiare tempo nella traduzione impostando delle regole per tradurre dei termini specifici sempre allo stesso modo.

### 2.2.3 Watson Natural Language Classifier

Il servizio Natural Language Classifier di IBM Watson utilizza tecniche di elaborazione del linguaggio naturale, di Cogitive Computing e di apprendimento automatico per classificare brevi input di testo in classi predefinite.

Ad esempio, un utente può inviare una domanda, ed il servizio classificherà la domanda in una classe predefinita che meglio risponde a quest'ultima, oppure fornirà un'azione appropriata per l'applicazione.

Questo servizio può essere utilizzato in molte applicazioni, come chatbot, assistenti virtuali e automazione del supporto clienti, per migliorare l'esperienza dell'utente e aumentare l'efficienza operativa.

Affinché possa essere utilizzato questo servizio all'interno di un'applicazione, è necessario che questo venga prima addestrato seguendo quattro passi riportati in Figura 2.7.

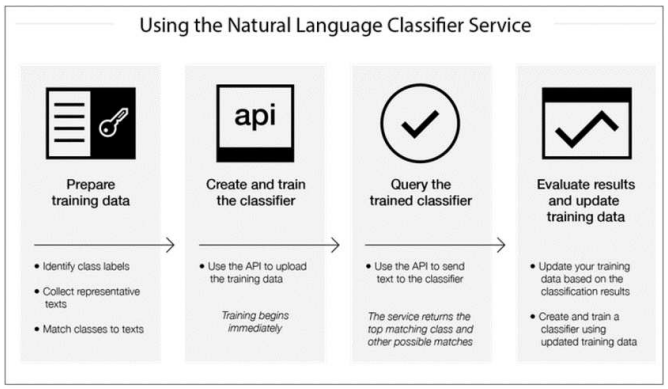


**Figura 2.7:** Quattro fasi del Natural Language Classifier

Essi sono:

1. preprarare i dati di training;
2. creare ed addestrare il classificatore;
3. interrogare il classificatore addestrato;
4. valutare i risultati e aggiornare i dati.

Soffermandoci sulla preparazione dei dati di training, per portare completamente a termine questo step, è necessario:

- *Identificare le etichette di classe*. Queste ultime rappresentano, in pratica, l'output di un classificatore addestrato. Esse descrivono l'intento del testo inserito.
- *Raccogliere testo rappresentativo*. Bisogna individuare testi rappresentativi per ciascuna etichetta di classe per effettuare il training.
- *Effettuare il matching delle classi con il testo*. Bisogna creare i dati di training facendo corrispondere il testo con le rispettive classi. La tecnica migliore è quella di creare un file CSV di training che viene usato quando si crea il classificatore.

### 2.2.4 Watson Retrieve and Rank

Il servizio Retrieve and Rank di Watson utilizza Apache Solr [1] e tecniche di Machine Learning per fornire risultati di ricerca più rilevanti e personalizzati.

In particolare, il servizio permette di creare e addestrare un modello di Machine Learning personalizzato per il proprio dominio di interesse, che può essere utilizzato per classificare e ordinare i risultati di ricerca in base alla rilevanza rispetto alla query dell'utente. In questo modo, si ottengono risultati di ricerca più precisi e utili per l'utente.

Per l'utilizzo di questo servizio non è necessario apportare alcuna modifica; esso è funzionale già nello stesso modo in cui viene fornito da IBM tuttavia, se necessario, è sempre possibile apportare modifiche legate alla personalizzazione del servizio per ottenere risultati migliori.

### 2.2.5 Watson Visual Recognition

Watson Visual Recognition utilizza algoritmi di Deep Learning per l'analisi di immagini per scene, oggetti, volti ed altri contenuti.

L'output include, inoltre, parole chiave che forniscono informazioni sul contenuto.

Nonostante vi sia la possibilità di addestrare e creare un classificatore personalizzato, l'insieme di classi predefinite fornisce, comunque, risultati accurati senza che vi sia bisogno di ulteriore addestramento.

Se necessario, tuttavia, l'addestramento si basa sull'analisi ed il riconoscimento di vecchie immagini di esempio, a partire dalle quali è possibile definire opportune classi che possono essere utilizzate sia per la classificazione sia per la definizione di opportuni score.

Fornito un insieme di immagini all'interno di un file, di queste ultime, quelle contenenti esempi positivi vengono sfruttate per la creazione di classi che definiscono ciò che il classificatore rappresenta.

Successivamente, il prefisso specificato per ogni parametro di esempio positivo, viene utilizzato come nome della classe, mentre è obbligatoria l'aggiunta del suffisso *_positive_examples*.

Non vi è alcun limite al numero di file di esempio positivi che possono essere caricati.

Viceversa, i file di esempio negativi non vengono utilizzati per creare una classe ma definiscono semplicemente ciò che non è il classificatore.

### 2.2.6 Watson Speech to Text

È l'equivalente dei seguenti altri servizi:

- *Servizio Voce di Microsoft Azure;*
- *Amazon Transcribe di Amazon Web Services;*
- *Speech-to-Text di Google Cloud.*

Watson Speech to Text è in grado di convertire il parlato in testo, in base alla lingua specificata dall'utente. Inoltre, è in grado di trascrivere il parlato in tempo reale da varie lingue e formati audio.

Il servizio utilizza funzionalità di riconoscimento vocale e ha un vocabolario di base che contiene molte parole utilizzate nella conversazione quotidiana.

[1] Server di ricerca basato su Apache Lucene, una libreria di recupero informazioni open source basata su Java. È progettato per guidare potenti applicazioni di information retrival

Tuttavia, per migliorare l'accuratezza del riconoscimento del parlato in settori specifici, come la medicina o il diritto, è necessario creare un nuovo modello linguistico personalizzato.

L'interfaccia di personalizzazione del modello linguistico consente agli utenti di adattare il servizio alle loro specifiche esigenze e di migliorare la precisione del riconoscimento del parlato in settori specifici.

**Servizio Voce di Microsoft Azure**

Il servizio Voce di Microsoft Azure è un servizio cloud che offre diverse funzionalità per la trasformazione e l'elaborazione del suono e del linguaggio.

In particolare, il servizio è in grado di convertire il parlato in testo leggibile, di aggiungere traduzioni vocali in tempo reale alle app ed ai servizi e di convertire il testo in audio quasi in tempo reale.

Inoltre, esso consente di creare rapidamente app e attività abilitate alla sintesi vocale utilizzando i linguaggi di programmazione già in uso.

Infine, è possibile personalizzare i sistemi vocali per ottimizzare la qualità in scenari specifici, grazie alle funzionalità di personalizzazione messe a disposizione dal servizio.

**Amazon Transcribe di Amazon Web Services**

Amazon Transcribe è un servizio che offre il riconoscimento vocale automatico, fornendo funzionalità avanzate di sintesi vocale per qualsiasi applicazione.

Tale strumento è stato sviluppato per semplificare l'aggiunta di queste funzionalità alle applicazioni e fornisce una varietà di opzioni, come l'importazione di input audio, la produzione di trascrizioni facili da leggere e rivedere e l'opzione di personalizzare il servizio per migliorare la precisione della trascrizione.

Inoltre, Transcribe garantisce la privacy dei clienti filtrando i contenuti.

Amazon Transcribe è in grado di elaborare sia file audio che video, dal vivo e registrati, offrendo trascrizioni di alta qualità per l'analisi e la ricerca.

Questo servizio è in grado di identificare automaticamente la lingua dominante in un file audio e produrre trascrizioni, il che può essere utile se si dispone di una libreria multimediale contenente file audio in diverse lingue.

È, inoltre, possibile utilizzare questa funzionalità per la classificazione dei contenuti multimediali, in modo da garantire che la lingua principale parlata nei video o podcast sia etichettata correttamente.

**Speech-to-Text di Google Cloud**

Speech-to-Text è un servizio di riconoscimento vocale e trascrizione del parlato in 125 lingue. Per il suo funzionamento sono stati utilizzati avanzati algoritmi di Reti Neurali e Deep Learning messi a punto da Google stessa.

È possibile personalizzare il proprio servizio Speech-to-Text grazie alla comoda interfaccia utente messa a disposizione sulla piattaforma.

Esso usa le classi per convertire automaticamente i numeri vocali in indirizzi, anni, valute e molto altro ancora.

Vi è la possibilità di scegliere da una selezione di modelli addestrati per il controllo vocale le chiamate telefoniche e la trascrizione di video, ottimizzati per i requisiti di qualità specifici del dominio.

### 2.2.7 Watson Text to Speech

Watson Text to Speech è un'API che permette di sintetizzare il testo scritto in parlato udibile, in diverse lingue e dialetti. L'input può essere testo semplice o scritto nel linguaggio SSML (Speech Synthesis Markup Language); il servizio gestisce stili di conversazione, pronuncia, tono e velocità di conversazione.

Le funzioni Voices offrono diverse voci maschili e femminili per ogni lingua e dialetto, e l'audio prodotto utilizza una cadenza e un'intonazione appropriate. Durante la sintesi, il servizio applica regole di pronuncia dipendenti dalla lingua per convertire lo spelling ordinario in uno spelling fonetico.

### 2.2.8 Watson Natural Language Understanding

Watson Natural Language Understanding è un servizio di analisi del linguaggio naturale fornito da IBM. Esso consente di analizzare il testo di input e di estrarre informazioni utili come categorie, concetti, emozioni, entità, parole chiave, relazioni, ruoli semantici e sentimenti.

Il servizio può essere personalizzato utilizzando modelli di annotazione sviluppati tramite IBM Watson Knowledge Studio, consentendo di identificare entità e relazioni specifiche per un dominio di interesse.

Attualmente, Watson Natural Language Understanding supporta una vasta gamma di lingue, tra cui l'arabo, il cinese, l'olandese, l'inglese, il francese, il tedesco, l'italiano, il giapponese, il coreano, il portoghese, il russo, lo spagnolo e lo svedese.

Il servizio è stato introdotto per sostituire l'API AlchemyLanguage, che non è più supportata. Watson Natural Language Understanding è stato progettato per fornire un'analisi del linguaggio naturale avanzata e personalizzabile per applicazioni aziendali.

### 2.2.9 Watson Discovery

Watson Discovery permette agli sviluppatori di aggiungere alle applicazioni un motore di ricerca e analisi dei contenuti di tipo cognitivo in grado di identificare qualsiasi intuizione possa portare ad un migliore processo decisionale.

Esso fornisce delle linee guida per l'acquisizione, l'arricchimento e la memorizzazione di grandi quantità di dati non strutturati.

Il servizio è in continuo miglioramento grazie alla possibilità di training su documenti precedenti tratti da interazioni con clienti, log di chat e risposte nei forum. Tutte queste funzionalità sono accessibili attraverso la sua API specifica.

Come nei servizi precedenti, anche Discovery consente la personalizzazione per il proprio dominio, attraverso l'addestramento di un modello personalizzato sfruttando Watson Knowledge Studio, che permette l'insegnamento da parte di esperti, attraverso l'utilizzo di esempi ed eliminando la necessità di codice, e collegando il modello con Discovery.

### 2.2.10 Watson Document Conversion

Il servizio Document Conversation di Watson fu ritirato nell'ottobre 2017. Nonostante ciò, le abilità di conversione furono portate avanti e sviluppate dalla IBM attraverso Watson Discovery.

### 2.2.11 Watson Personality Insights

Il servizio Personality Insights di Watson consente di desumere informazioni dai Social Media per identificare le caratteristiche psicologiche di un individuo.

Gli obiettivi a cui si vorrebbe arrivare sono i seguenti:

- *Ottenere ritratti dettagliati della personalità:* ciò avviene mediante l'utilizzo di analisi linguistiche, così da ottenere le caratteristiche della personalità degli individui a partire dalle comunicazioni digitali (e-mail, blog, tweet, etc.).
- *Capire le preferenze di consumo:* osservare l'inclinazione degli utenti nella ricerca di diversi prodotti, servizi e attività.
- *Ritagliare un'esperienza mirata per il cliente:* comprendere i singoli clienti per la realizzazione di prodotti personalizzati e consigli mirati.

### 2.2.12 Watson Tone Analyzer

Tone Analyzer utilizza l'analisi linguistica cognitiva per individuare una varietà di toni a livello sia di "frase" che di "documento".

In Figura 2.8 viene riportato il procedimento di analisi svolto dal servizio.

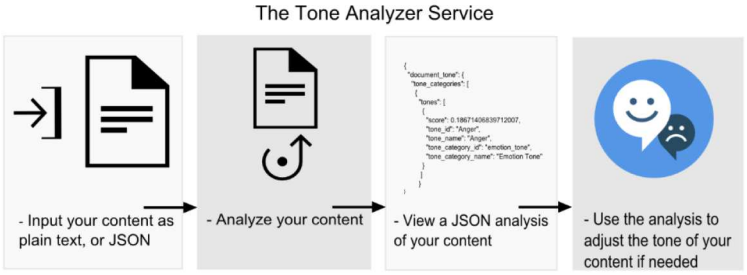


**Figura 2.8:** Funzionamento generale di Watson Tone Analyzer

Analizziamo le funzionalità consentite da Tone Analyzer:

- *Condurre ascolto sociale:* analizzare le emozioni ed i toni associati a ciò che le persone esprimono online, come tweet e recensioni, e da questi analizzare lo stato d'animo degli utenti.
- *Migliorare il servizio clienti:* monitorare il servizio clienti e le conversazioni di supporto così da riuscire a fornire ai clienti risposte appropriate e su una vasta gamma di argomenti, analizzare opinioni dei clienti e gli atteggiamenti degli agenti.
- *Integrarsi con chatbot:* permettere ad un chatbot di rilevare i toni dei clienti così da poter costruire strategie di dialogo per controllare la conversazione di conseguenza.

# CAPITOLO 3

## Traduzione: implementazione in AWS, Google, Azure e DeepL

*Nel capitolo corrente si analizzerà il servizio di traduzione messo a disposizione dalle varie piattaforme Cloud utilizzate. L'idea è stata quella di fornire lo stesso input per i quattro servizi scelti e valutare ciò che si è ottenuto in output. In particolare, si vedrà come, per ciascuno di questi, gli output potranno essere simili o meno in base anche alle metodologie adoperate dai loro produttori. Per ciascuno sarà spiegato il modello di funzionamento, verranno presentati alcuni esempi prodotti e si elencheranno i principali vantaggi o svantaggi.*

## 3.1 Implementazione in AWS

Amazon Translate è un servizio di traduzione di testi che utilizza tecnologie avanzate di Machine Learning per fornire traduzioni di alta qualità su richiesta. È possibile utilizzare Amazon Translate per tradurre documenti di testo non strutturati o per creare applicazioni che funzionano in più lingue.

Il servizio Amazon Translate si basa su reti neurali addestrate per la traduzione linguistica. Consente di tradurre tra una lingua di partenza (la lingua originale del testo da tradurre) e una lingua di arrivo (la lingua in cui il testo viene tradotto).

Quando si lavora con il servizio di Amazon Translate bisogna fornire una risorsa di testo per ottenere in output un'ulteriore risorsa di testo. Più specificatamente, introduciamo le seguenti definizioni:

- *Risorsa di testo:* il testo che si desidera tradurre. Bisogna fornire la risorsa nel formato UTF-8 [1].
- *Testo in uscita:* il testo che Amazon Translate ha tradotto nella lingua di destinazione. Anche il testo di output è in formato UTF-8. A seconda delle lingue di origine e di destinazione, il testo di output potrebbe contenere più caratteri rispetto al testo di input.

Il modello di traduzione ha due componenti, ovvero l'encoder e il decoder. L'encoder legge una frase di partenza una parola alla volta e costruisce una rappresentazione semantica che ne cattura il significato. Il decoder (decodificatore) utilizza la rappresentazione semantica per generare una traduzione una parola alla volta nella lingua di destinazione.

[1] UTF-8 (Unicode Transformation Format, 8 bit) è una codifica di caratteri Unicode in sequenze di lunghezza variabile di byte, creata da Rob Pike e Ken Thompson. UTF-8 usa gruppi di byte per rappresentare i caratteri Unicode, ed è particolarmente utile per il trasferimento tramite sistemi di posta elettronica a 8-bit.

Amazon Translate utilizza meccanismi per comprendere il contesto. Questo lo aiuta a decidere quali parole del testo di partenza sono più rilevanti per generare la parola di destinazione successiva. I meccanismi di attenzione consentono al decodificatore di concentrarsi sulle parti più rilevanti di una frase di partenza. Ciò garantisce che quest'ultimo traduca correttamente parole o frasi ambigue.

La parola target generata dal modello diventa l'input del decodificatore. La rete continua a generare parole finché non raggiunge la fine della frase.

Amazon Translate può rilevare automaticamente la lingua utilizzata nel testo di origine. Per utilizzare il rilevamento automatico della lingua, è necessario specificare *auto* come lingua di origine. Amazon Translate chiama Amazon Comprehend per conto dell'utente per determinare la lingua utilizzata nel testo di origine.

Se si specifica una lingua di origine o di destinazione non supportata, Amazon Translate restituisce le seguenti eccezioni:

- *UnsupportedLanguagePairException:* Amazon Translate garantisce la traduzione tra tutte le lingue supportate. Questa eccezione viene restituita se la lingua di origine o di destinazione non è supportata.
- *DetectedLanguageLowConfidenceException:* se si utilizza il rilevamento automatico della lingua e Amazon Translate ha una bassa certezza di aver rilevato la lingua di origine corretta, restituisce questa eccezione. Se un livello di confidenza basso è accettabile, è possibile utilizzare la lingua di origine restituita nell'eccezione.

### 3.1.1 Spiegazione del funzionamento

Per selezionare il servizio di traduzione di AWS, ovvero *Amazon Translate*, è sufficiente ricercare quest'ultimo nella barra di ricerca in alto a sinistra, oppure selezionarlo direttamente dalla schermata dei servizi visualizzati di recente, come mostrato nella Figura 3.1.

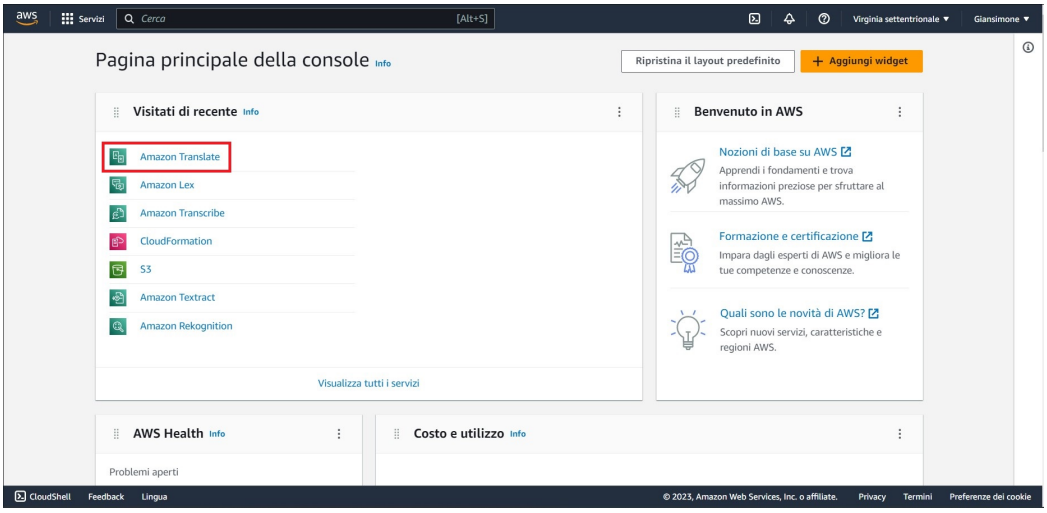


**Figura 3.1:** Selezione del servizio Amazon Translate

Fatto ciò si aprirà una nuova finestra di traduzione in tempo reale che permetterà l'inserimento di un testo da tradurre.

Il funzionamento è molto semplice in confronto agli altri servizi utilizzati. Basta scrivere il testo che si vuole tradurre nell'opportuna area di testo per vedersi comparire di fianco il corrispondente testo tradotto.

È sufficiente selezionare la sola lingua di destinazione poiché la lingua di origine viene rilevata automaticamente nel momento di inserimento del testo.

### 3.1.2 Esempi svolti

Nelle Figure 3.2 e 3.3 sono riportati due esempi di traduzione. Sono state scelte delle frasi di alcuni dei più celebri personaggi che verranno utilizzate per il confronto tra i servizi di traduzione di AWS, Google Cloud, Microsoft Azure e DeepL.

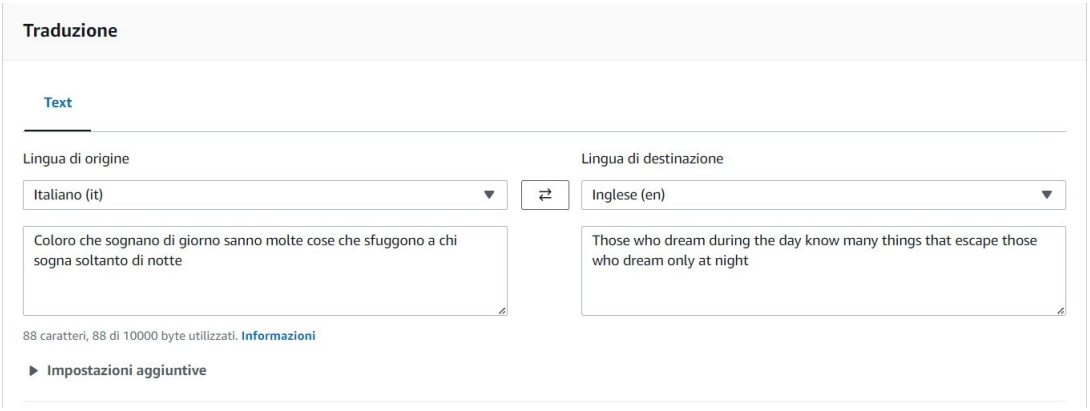


**Figura 3.2:** Traduzione della frase di Edgar Allan Poe

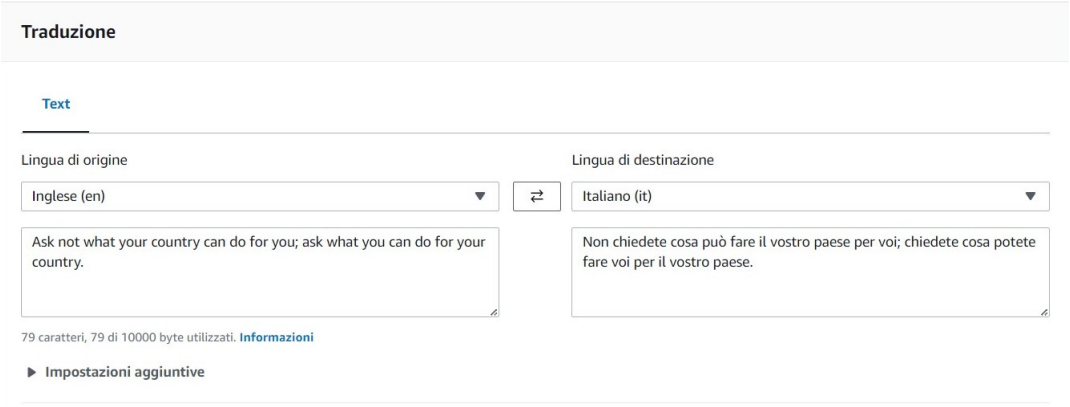


**Figura 3.3:** Traduzione della frase di John Fitzgerald Kennedy

Come è possibile notare, entrambe le frasi risultano tradotte in maniera corretta e prive di errori. Un ulteriore dettaglio da segnalare è quello di una traduzione molto rigida e scandita parola per parola.

### 3.1.3 Vantaggi e svantaggi del servizio

Passiamo, dunque, ad analizzare quali possono essere i principali vantaggi o svantaggi che il servizio Amazon Translate riporta. Tra i vantaggi possiamo sicuramente annoverare:

- *Scalabilità:* sia che si tratti di poche parole o di grandi quantità di documenti, Amazon Translate si adatta facilmente alle esigenze di traduzione. Qualsiasi contenuto elaborato da Amazon Translate è crittografato in transito e a regime.
- *Facilità d'uso:* questa è una caratteristica che la maggior parte degli utenti richiedono, ossia la possibilità di interagire con il servizio in maniera veloce ed efficace. Personalmente abbiamo trovato Amazon Translate molto efficace sotto questo punto di vista e ciò ha permesso di sfruttare a pieno le sue potenzialità.

- *Batch asincrono e in tempo reale:* Amazon Translate è ideale per eseguire traduzioni in batch quando si dispone di grandi quantità di testo preesistenti da tradurre, e traduzioni in tempo reale quando si desidera offrire traduzioni on demand di contenuti.
- *Sicurezza:* qualsiasi contenuto elaborato da Amazon Translate è crittografato in transito e a riposo.
- *Qualità elevata:* Amazon Translate fornisce traduzioni di alta qualità che soddisfano le esigenze di un'ampia gamma di settori.

Viceversa gli svantaggi non sono rilevanti; tuttavia la traduzione strettamente legata alle singole parole può indurre il processo in errore (si veda, a tal proposito, la Sezione 3.5).

## 3.2 Implementazione in Google Cloud

*Translation AI* consente ai siti web ed alle applicazioni di tradurre in modo dinamico il testo da programma, tramite un'API. Per la traduzione del testo, il servizio di Google utilizza un modello di Machine Learning preaddestrato o personalizzato. Per impostazione predefinita, Translation AI utilizza un modello NMT (Nural Machine Translation) preaddestrato di Google, che viene aggiornato con cadenza semiregolare quando vengono resi disponibili dati di addestramento o tecniche migliori.

Ci sono varie funzionalità possibili per il servizio translate di Google; in particolare evidenziamo le seguenti tre:

1. API Translation Basic;
2. API Translation Advanced;
3. AutoML Translation.

Ciascuno di questi presenta prezzi differenti in base alle funzionalità messe a disposizione, che possono variare, ad esempio, sulla base della tipologia di input che sono in grado di acquisire per tradurre, sulla base della possibilità di addestramento personalizzato o meno, sulla base della possibilità di selezionare modelli di traduzione personalizzati, e così via.

### 3.2.1 Spiegazione del funzionamento

Per quanto riguarda il funzionamento del servizio, Google mette a disposizione un'API alla quale è possibile accedere attraverso una chiave fornita da Google stessa in fase di attivazione del servizio.

L'API è disponibile per i seguenti linguaggi: *Node.js, Java, Go, Python*. In particolare, per il nostro lavoro di tesi, abbiamo deciso di adottare il linguaggio Python.

È, poi, possibile utilizzare questa API, completa di chiave per l'accesso, per la traduzione di qualsiasi testo. In Figura 3.4 è possibile osservare il codice Python utilizzato per il funzionamento del servizio attraverso API Google.

Notiamo come il tutto si riconduca ad un'unica funzione `translate_text` la quale riceve in input i seguenti due parametri:

- `text`: il testo da tradurre e da passare tra apici in formato stringa.
- `project_id`: l'identificativo del progetto sul quale si sta lavorando, passato tra doppi apici in formato stringa.

main.py

```python
#TranslationAI di google cloud
from google.cloud import translate

def translate_text(text="Hello, Mr Davies", project_id="my-project-translation-382319"):

    client = translate.TranslationServiceClient()
    location = "global"
    parent = f"projects/{project_id}/locations/{location}"

    response = client.translate_text(
        request={
            "parent": parent,
            "contents": [text],
            "mime_type": "text/plain",
            "source_language_code": "en-US", #en-US
            "target_language_code": "it", #it
        }
    )

    for translation in response.translations:
        print("Translated text: {}".format(translation.translated_text))


translate_text()
```

**Figura 3.4:** Codice Python per il funzionamento dell'API

Come prima istruzione viene utilizzata la libreria `translate` importata precedentemente, per la creazione di un oggetto `client` di tipo `TranslationServiceClient`.

È bene notare come i parametri passati alla funzione iniziale, ovvero `text` e `project_id` vengano ora utilizzati per il completamento dell'URI all'interno della variabile `parent`.

A seguire ci serviremo dell'oggetto `client` creato all'inizio della funzione in questione, per la generazione di una richiesta nella quale è possibile specificare sia la lingua di origine, nel nostro caso l'inglese (en-US), che la lingua di destinazione nella quale vogliamo venga tradotto il nostro testo, in questo caso l'italiano (it).

Infine la parte conclusiva della funzione si avvale di un blocco `for` per la generazione in output del messaggio tradotto.

La Figura 3.5 riporta l'immagine della console sulla quale viene visualizzato il testo di esempio tradotto nella lingua desiderata.

```
Translated text: Salve, signor Davies

Process finished with exit code 0
```

**Figura 3.5:** Testo tradotto dal servizio

### 3.2.2 Esempi svolti

Anche per il servizio di Google sono state svolte varie prove d'esempio. In particolare nelle Figure 3.6 e 3.7 sono riportati i due esempi basati sulle celebri fasi di Edgar Allan Poe e di John Fitzgerald Kennedy.

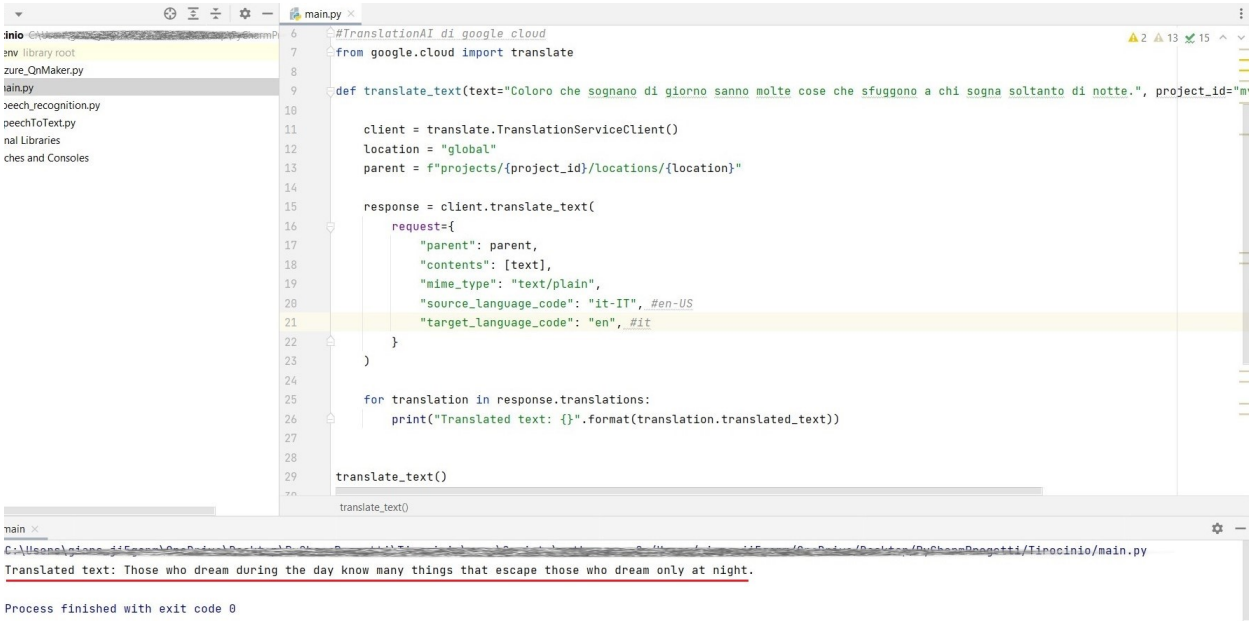


**Figura 3.6:** Traduzione della frase di Edgar Allan Poe

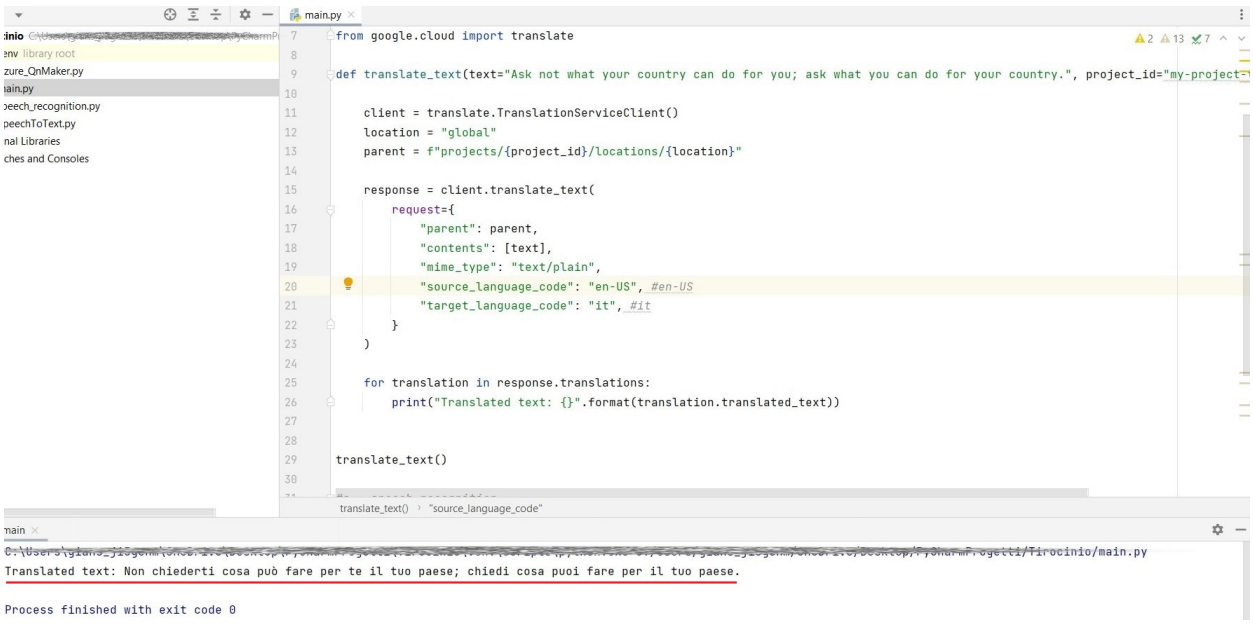


**Figura 3.7:** Traduzione della frase di John Fitzgerald Kennedy

Come possiamo notare, il risultato delle traduzione è identico a quello restituito dal servizio di traduzione di AWS. Come nel caso precedente, infatti, si procede con una traduzione testuale molto puntuale e basata sul significato delle singole parole. La differenza sostanziale tra questi primi due servizi consiste, dopotutto, soltanto nella gestione del servizio in sé, che richiede la scrittura di un piccolo pezzo di programma con cui richiamare l'API, nel caso di

Google, e nell'immediato utilizzo sulla stessa piattaforma, per quanto riguarda il servizio proposto da AWS.

### 3.2.3 Vantaggi e svantaggi del servizio

Passiamo, ora, all'analisi dei principali vantaggi e svantaggi che caratterizzano il servizio *Translation AI* di Google.

Tra le note positive di quest'ultimo notiamo:

- *Supporto linguistico senza precedenti:* esso usa la traduzione automatica per rilevare più di 100 lingue. Crea modelli personalizzati in più di 50 combinazioni linguistiche utilizzando la tecnologia AutoML.
- *Qualità:* Google vanta una vasta esperienza nella fornitura di servizi di traduzione rivolti a consumatori e organizzazioni. I corrispettivi modelli e strumenti collaudati offrono l'esperienza di Google nella traduzione con un'accuratezza garantita da un leader del settore.
- *Specificità del dominio:* Translation AI consente una personalizzazione dei servizi di traduzione per comprendere il gergo del settore o i termini specifici del dominio. Esso mantiene il contesto e il significato nelle traduzioni di documenti tecnici, descrizioni di prodotti e contenuti social.

Quest'utlima caratteristica è una proprietà che abbiamo apprezzato particolarmente poiché, anche per coloro che sono alle prime armi in qualsiasi ambito, sia esso informatico, elettronico o altro, la possibilità di tradurre documentazioni e testi con termini specifici aiuta notevolmente l'utente nell'apprendimento e nella comprensione della materia in esame.

Per quanto riguarda gli svantaggi, non vi sono caratteristiche negative che hanno causato problemi o richiesto grande manodopera in fase di attivazione. Nonostante ciò, il fatto di dover utilizzare l'API tramite un particolare script scritto in un determinato linguaggio di programmazione può comportare problemi per coloro che non sono ancora entrati completamente nella logica di funzionamento di questi sistemi; infatti è richiesta una conoscenza base del linguaggio nel quale si è deciso di implementare il codice di funzionamento per l'utilizzo del servizio.

## 3.3 Implementazione in Microsoft Azure

Il servizio di traduzione di Microsof Azure, *Translator*, traduce il testo immediatamente o in batch in oltre 100 lingue. Esso supporta una vasta gamma di casi d'uso, ad esempio la traduzione per call center, agenti di conversazione in più lingue o le comunicazioni in-app.

### 3.3.1 Spiegazione del funzionamento

Per accedere al servizio di traduzione di Microsoft Azure, dopo aver effettuato l'accesso, è possibile selezionare il servizio tra quelli proposti o, in caso contrario, ricercarlo direttamente dalla barra di ricerca, come riportato in Figura 3.8.

Una volta selezionato il servizio, si arriva ad una schermata in cui è possibile direttamente usufruire del servizio con le stesse modalità proposte da AWS. Accedendo al servizio osserviamo la presenza di due aree di testo; nella prima è possibile inserire il testo da tradurre (la lingua di origine viene rilevata automaticamente); nella seconda verrà visualizzato il testo tradotto nel linguaggio desiderato (Figura 3.9).

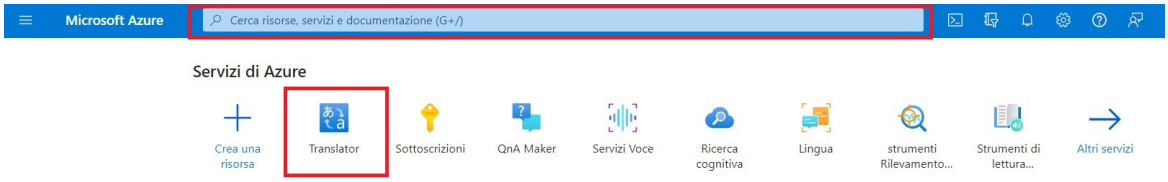


**Figura 3.8:** Selezione del servizio Translator

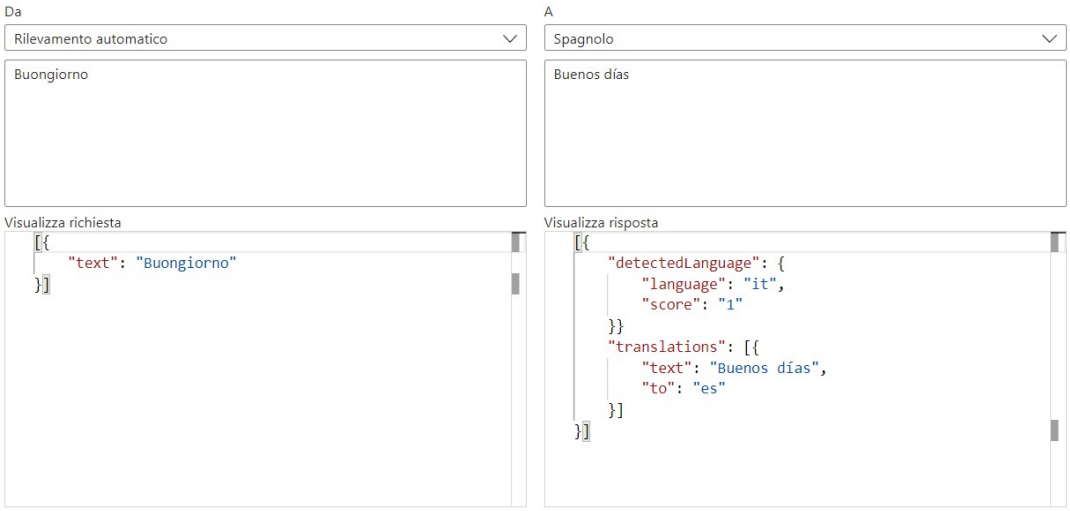


**Figura 3.9:** Come si presenta il servizio Translator di Microsof Azure

### 3.3.2 Esempi svolti

Tra gli esempi proposti ne analizziamo due, in particolare quelli già visti per i servizi proposti da Google ed AWS, così da evidenziare meglio le differenze o similitudini presenti. Le Figure 3.10 e 3.11 riportano gli esempi citati.

Dai dati riportati emerge come la prima frase sia rimasta invariata e tradotta ugualmente da tutti e tre i traduttori; per la seconda invece, il traduttore di Microsoft Azure propone una traduzione differente.

In dettaglio, nella prima parte della frase in Figura 3.11 viene cambiato il soggetto, che, però, rimane invariato come nei due servizi precedenti per la seconda parte della frase. È chiaro, dunque, che si tratti di un errore di comprensione che, sebbene sia irrilevante nel nostro caso, può comportare problemi per altri utilizzi in ambiti differenti.

### 3.3.3 Vantaggi e svantaggi del servizio

Concentriamoci, ora, sulle possibili caratteristiche positive o negative del servizio appena analizzato. Tra gli aspetti vantaggiosi troviamo:

- *Supporto per più lingue:* Translator traduce accuratamente il testo in oltre 100 lingue.
- *Traduzioni personalizzabili:* Esso crea modelli personalizzati per gestire la terminologia specifica del dominio.
- *Sicurezza incorporata:* l'input di testo non viene registrato durante la traduzione.


**Figura 3.10:** Traduzione della frase di Edgar Allan Poe

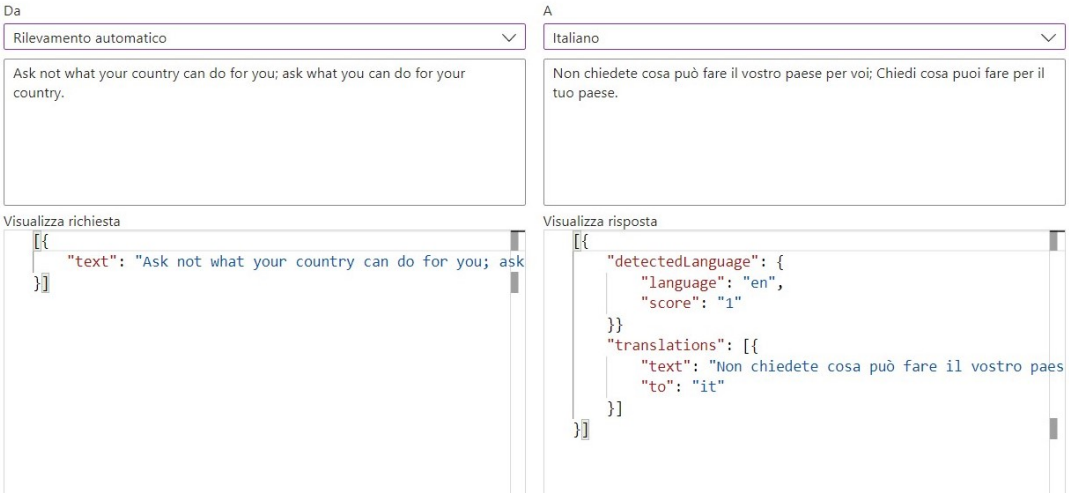


**Figura 3.11:** Traduzione della frase di John Fitzgerald Kennedy

- *Facilità di utilizzo:* così come il servizio di AWS, l'interfaccia amichevole consente una fruizione del servizio più rapida e veloce, a differenza di quello che accade nel caso di Translation AI di Google Cloud.

Tra gli aspetti sfavorevoli non ci sono da annoverare molti dettagli, se non il fatto di dover sopportare una fase di attivazione del servizio piuttosto rigida e severa da parte di Microsoft Azure, cosa che non è accaduta nei casi precedenti per le piattaforme già analizzate.

## 3.4 Implementazione in DeepL

*DeepL Translator* è un servizio di traduzione gratuito multilingue, alimentato dalla base di conoscenza di Linguee, servizio creato dalla stessa azienda, DeepL GmbH (in precedenza chiamata Linguee). Attualmente, il portale supporta 24 lingue. Lanciato ad agosto 2017 dalla società tedesca DeepL GmbH, esso offre una soluzione base che non richiede iscrizione, ed una premium, con un'API che consente l'integrazione negli applicativi desktop e aziendali.

L'algoritmo e il codice sorgente sono proprietari. La tecnologia di DeepL si basa su reti neurali convoluzionali [2] e su una serie di superserver [3].

Nel 2018 sono anche arrivate le estensioni gratuite per Google Chrome e Firefox, mentre nel 2019 DeepL ha rilasciato il software per Windows e macOS.

### 3.4.1 Spiegazione del funzionamento

Il funzionamento di DeepL è relativamente semplice e questo, unito alla grande capacità di traduzione, lo porta ad essere uno dei migliori traduttori al mondo. La sua capacità di discostarsi dalla singola parola e di valutare il contesto generale lo rendono capace di fornire risultati adeguati e ben tradotti, merito anche della differente tecnologia alla base del servizio, ovvero la rete neurale convoluzionale, che presenta delle peculiarità che la contraddistinguono dalle tecnologie precedenti.

In Figura 3.12 è riportata la schermata di presentazione del servizio DeepL, molto facile ed intuitiva. Vi sono due aree di testo nelle quali, a sinistra, è possibile inserire il testo da tradurre (la cui lingua verrà automaticamente rilevata); a destra verrà fornito in output il risultato tradotto.

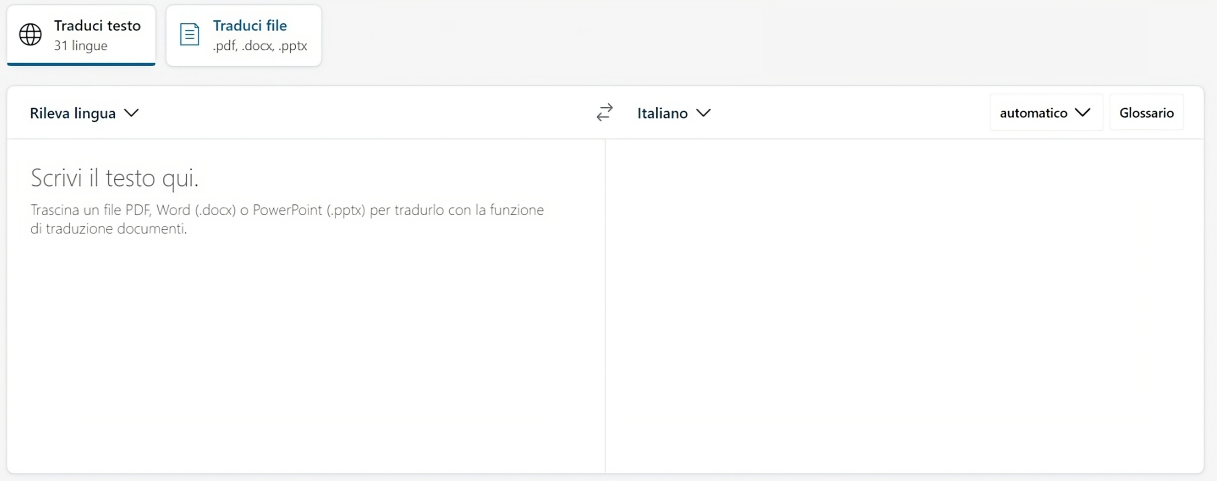


**Figura 3.12:** Come si presenta DeepL

### 3.4.2 Esempi svolti

Passiamo, ora, all'analisi delle due solite frasi, per evidenziare alcuni aspetti importanti del servizio riportate in Figura 3.13 e 3.14.

Dalla Figura 3.13 scaturisce un dettaglio rilevante. DeepL utilizza l'avverbio di tempo "during" mentre alcuni dei servizi precedenti, come ad esempio Microsoft Azure, utilizzavano "by". Questo dipende dal fatto spiegato precedentemente, cioè dal fatto che DeepL tiene conto della frase in generale e del suo contesto e non effettua una traduzione parola per parola.

[2] Una rete neurale convoluzionale (CNN o ConvNet, dall'inglese Convolutional Neural Network) è un tipo di rete neurale artificiale feed-forward (ovvero, una rete neurale artificiale dove le connessioni tra i nodi non formano cicli, differenziandosi dalle reti neurali ricorrenti) in cui il pattern di connettività tra i neuroni è ispirato dall'organizzazione della corteccia visiva animale, i cui singoli neuroni sono disposti in maniera tale da rispondere alle regioni di sovrapposizione che tassellano il campo visivo

[3] Il supercomputer è un tipo di sistema di elaborazione progettato per ottenere capacità di calcolo con un numero di cifre, dette bit, estremamente elevate rispetto a quelle dei calcolatori base, presenti prima della realizzione (assemblaggio) della macchina superiore che si sta realizzando.

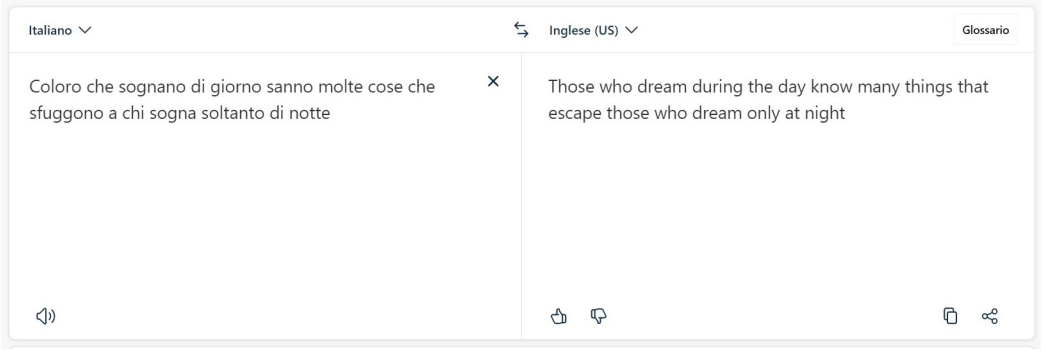


**Figura 3.13:** Traduzione della frase di Edgar Allan Poe

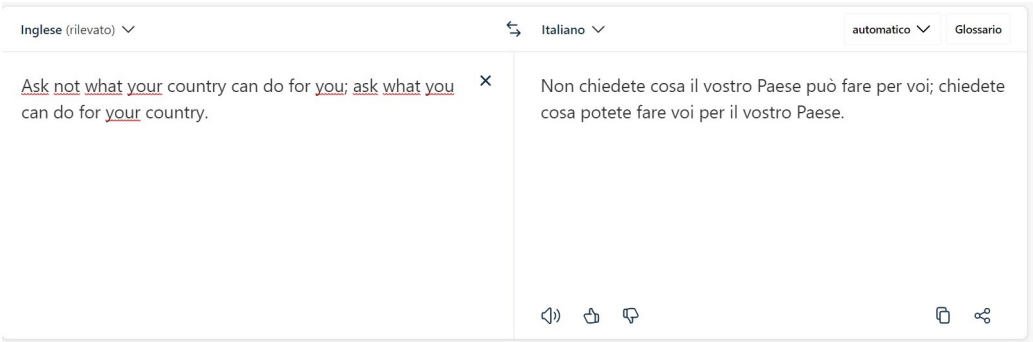


**Figura 3.14:** Traduzione della frase di John Fitzgerald Kennedy

Nella seconda Figura 3.14, invece, possiamo notare come, in questo caso, il soggetto è variato rispetto ai risultati forniti dai servizi precedenti, ma rimane costante al suo interno e non viene alterato, come nel caso di Microsoft Azure, rispecchiando, ancora una volta, il concetto espresso poco fa sulla traduzione generale.

### 3.4.3 Vantaggi e svantaggi del servizio

Il servizio DeepL ha numerose capacità aggiuntive che lo portano ad essere considerato uno dei migliori al mondo; tra queste citiamo:

- *Facilità d'uso:* l'interfaccia amichevole proposta aiuta decisamente nell'utilizzo del servizio.
- *Traduzione più precisa:* i dettagli finora elencati, come la capacità di valutare il contesto generale della frase e tradurre le parole in base a quest'ultimo, comportano una traduzione molto efficace e precisa.
- *Possibilità di caricare documenti:* la possibilità di caricare documenti in vari formati è una nota addizionale da non sottovalutare, nonstante ci sia la medesima possibilità anche in alcuni dei servizi già discussi precedentemente, ma con la richiesta di un corrispettivo in denaro.

Non abbiamo evidenziato particolari note negative durante l'utilizzo del sistema di traduzione di DeepL.

## 3.5 Confronto critico tra i quattro sistemi di traduzione

Dalle sezioni precedenti sono affiorate alcune differenze in termini di risultati tradotti; ciò è dovuto a vari fattori come, ad esempio, il metodo con cui sono stati addestrati i vari servizi, o anche il modello di tecnologia utilizzata.

Di seguito abbiamo deciso di riportare alcuni casi eclatanti riguardanti i quattro servizi.

Il primo riguarda la traduzione della seguente frase: *1 - Bill is a very educated boy*, la cui corretta traduzione sarebbe la seguente: *1 - Bill è un ragazzo ben istruito*.

Possiamo riscontrare, nelle Figure 3.15, 3.16, 3.17 e 3.18, come sia stata effettivamente tradotta questa frase.

Principalmente solo il servizio Translation AI di Google Cloud non è riuscito a cogliere il reale significato della parola "educated", in inglese anche detta *"false friend"*, poiché ricorda l'aggettivo "educato", ma in realtà il suo reale significato rappresenta il livello di istruzione di una persona.

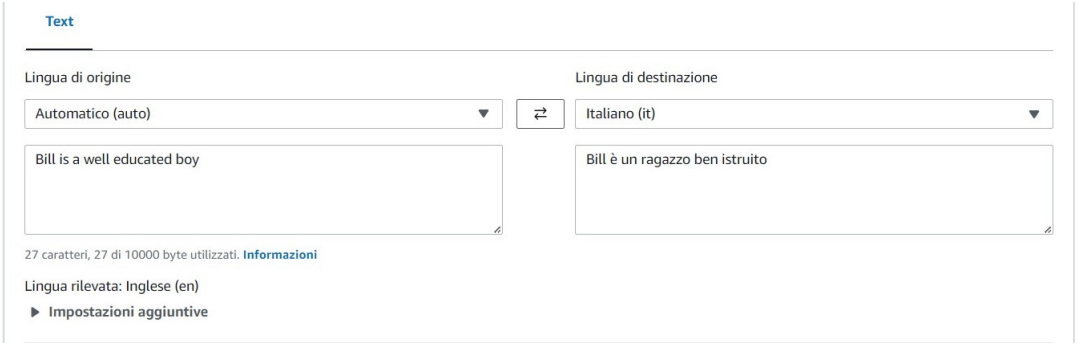


**Figura 3.15:** Traduzione della frase n.1 con il servizio AWS

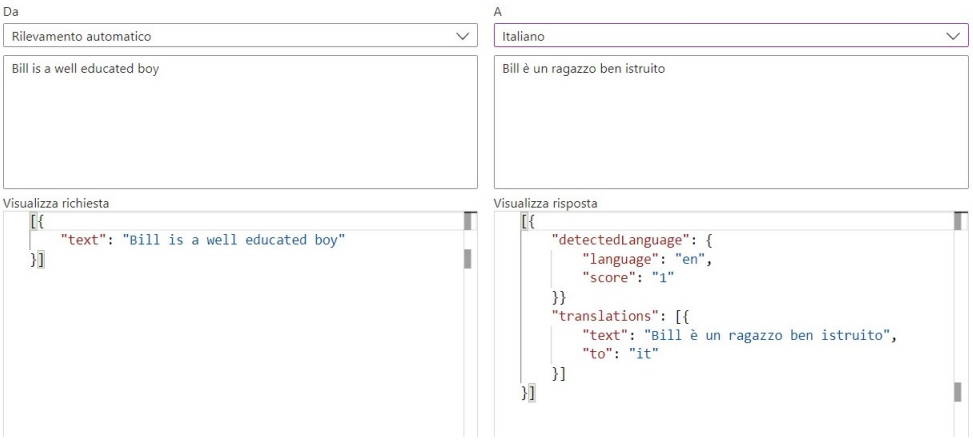


**Figura 3.16:** Traduzione della frase n.1 con il servizio Microsoft Azure

Verosimilmente, esempi del genere su parole che possono trarre in inganno il servizio sono numerosi. Un ulteriore caso che consideriamo rilevante è quello riguardante la traduzione della seguente frase: *2 - Poor Harry, he was a very polite boy*, corrispondente alla frase tradotta in italiano: *2 - Povero Harry, era un ragazzo molto educato*.

L'aggettivo "polite", infatti, identifica una persona educata. Questo non è stato recepito, però, dal servizio di traduzione di AWS, il quale ha frainteso il reale significato con un ulteriore aggettivo, gentile (in inglese "kind"). L'esempio in questione è riportato nella Figura 3.19.

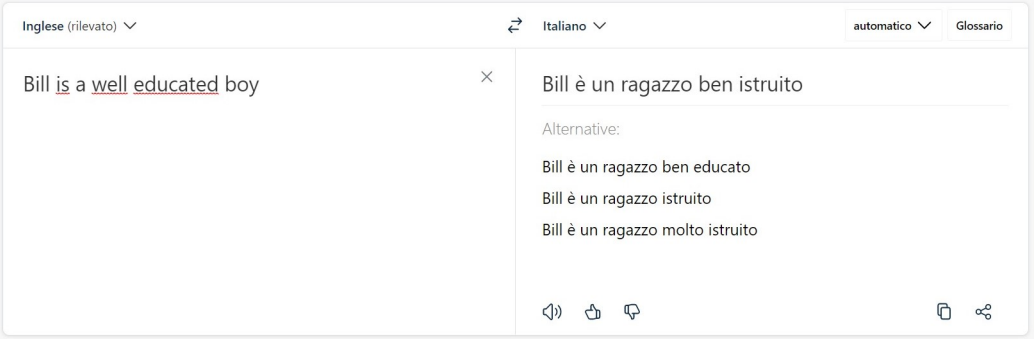


**Figura 3.17:** Traduzione della frase n.1 con il servizio DeepL

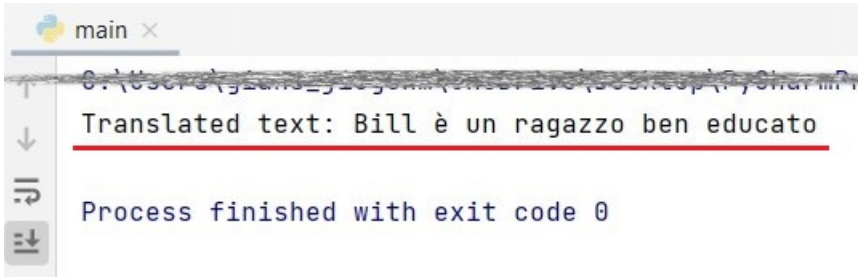


**Figura 3.18:** Traduzione della frase n.1 con il servizio di Google Cloud

Possiamo, dunque, ancora una volta, sottolineare le varie caratteristiche differenti che contraddistinguono i vari servizi, che hanno, poi, portato ad avere un risultato in alcuni casi differente da quello atteso.

Personalmente abbiamo trovato i servizi di traduzione molto interessanti. L'aver studiato ed analizzato come questi sono impostati e come questi gestiscono il processo di traduzione ci ha permesso di comprendere l'importanza di questi ultimi per una possibile implementazione all'interno dei sistemi di Cognitive Computing.

Tali sistemi, infatti, che si basano sulla comunicazione e scambio di messaggi con l'utente, richiedono lo sviluppo di questi modelli di traduzione così da consentire loro l'interfacciamento con un'ampia scala di utenti, dislocati in aree diverse del mondo.

È bene, dunque, continuare a lavorare nello sviluppo e sul miglioramento di questi servizi nella speranza di contribuire, anche, ad un sostanziale miglioramento dei sistemi di Cognitive Computing.

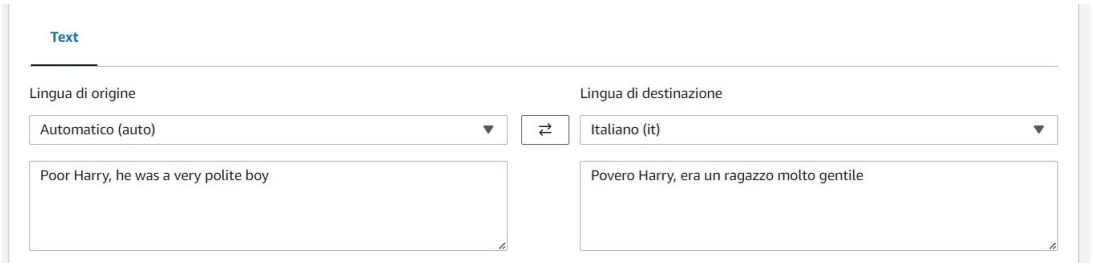


**Figura 3.19:** Traduzione della frase n.2 con il servizio di AWS

CAPITOLO 4

# Trascrizione: implementazione in AWS, Google ed Azure

*Nel capitolo in questione verrà esaminato il servizio di trascrizione offerto dalle diverse piattaforme Cloud utilizzate. L'obiettivo è stato, così come per il servizio precedente, quello di confrontare i risultati ottenuti utilizzando lo stesso input per i tre servizi selezionati. Verrà fornita una spiegazione del modello di funzionamento di ciascun servizio, seguita dalla presentazione di alcuni esempi prodotti, e l'indicazione dei principali vantaggi e svantaggi di ognuno. Infine saranno messi in evidenza i tratti caratteristici di ogni singolo servizio e verranno confrontati per ottenere una panoramica generale.*

## 4.1 Implementazione in AWS

*Amazon Transcribe* è un servizio avanzato di riconoscimento vocale automatico che semplifica l'integrazione delle funzionalità di conversione da parlato a testo in qualsiasi applicazione. Gli strumenti di Transcribe consentono di elaborare input audio con facilità, generare trascrizioni chiare e leggibili, migliorare l'accuratezza attraverso la personalizzazione e garantire la privacy dei clienti filtrando i contenuti sensibili. Con Amazon Transcribe, puoi arricchire le tue applicazioni con la potenza del riconoscimento vocale automatico, offrendo un'esperienza di conversione da parlato a testo affidabile e di alta qualità.

Transcribe è stato appositamente sviluppato per gestire sia input audio che video, in tempo reale o registrati, al fine di fornire trascrizioni di elevata qualità per scopi di ricerca e analisi. Inoltre, AWS offre API specifiche per comprendere in modo accurato le chiamate dei clienti, grazie ad Amazon Transcribe Call Analytics, e le conversazioni mediche con Amazon Transcribe Medical. Queste soluzioni separate consentono una comprensione univoca e mirata di contesti specifici, offrendo un supporto ancora più personalizzato per le esigenze individuali dei clienti.

È possibile utilizzare Transcribe per elaborare registrazioni audio preesistenti o per effettuare la trascrizione in tempo reale di flussi audio in diretta. Grazie ad una connessione sicura è possibile inviare l'audio al servizio e ricevere in risposta un flusso di testo corrispondente. Questo processo consente di convertire rapidamente l'audio in testo, offrendo una soluzione efficiente e accurata per le esigenze di trascrizione.

Grazie ad Amazon Transcribe è possibile riconoscere automaticamente la lingua predominante in un file audio e generare trascrizioni corrispondenti. Questa funzionalità risulta particolarmente utile quando si lavora con una libreria multimediale che contiene file audio in diverse lingue. È possibile sfruttare questa funzione anche per la classificazione dei contenuti multimediali, assicurando che la lingua principale parlata nei video e nei podcast sia

correttamente identificata ed etichettata. In questo modo si otterranno informazioni precise sulla lingua utilizzata nei contenuti multimediali, e sarà possibile gestirli in modo efficiente.

Amazon Transcribe è in grado di arricchire automaticamente le trascrizioni con la punteggiatura e la formattazione corretta dei numeri. Questo significa che l'output finale sarà pari alla qualità di una trascrizione effettuata manualmente, consentendo di risparmiare tempo e costi. I numeri vengono trascritti utilizzando le cifre numeriche o la loro "forma normale" anziché essere rappresentati in forma verbale. Questo miglioramento nella precisione e nella formattazione dei numeri permette di ottenere trascrizioni più accurate e pronte all'uso, senza dover dedicare risorse aggiuntive per la revisione o la correzione.

Grazie ad Amazon Transcribe ogni parola nella trascrizione corrisponde ad un timestamp, che indica l'istante temporale in cui la parola è pronunciata nella registrazione originale. Questo rende semplice individuare una parola o una frase specifica nel file audio di origine. Si possono utilizzare questi timestamp per navigare rapidamente attraverso le registrazioni, per trovare le sezioni desiderate, o per aggiungere sottotitoli accurati ai propri video. La presenza dei timestamp consente una migliore gestione e organizzazione dei contenuti audio, semplificando il processo di ricerca, revisione e accessibilità per i file multimediali.

Come già accennato precedentemente, Amazon Transcribe presenta implementazioni apposite per comprendere chiamate dei clienti o conversazioni mediche. Questi due sottoservizi specializzati sono:

1. *Amazon Transcribe Call Analytics:* con Amazon Transcribe Call Analytics, si può ottenere molto più di una semplice trascrizione delle conversazioni. Il servizio consente di estrarre informazioni dettagliate sulle conversazioni, come il sentiment delle chiamate e l'intensità del parlato. Questi dati possono tornare utili per migliorare la produttività degli agenti e l'esperienza complessiva con i clienti.

   Inoltre, Amazon Transcribe Call Analytics genera riepiloghi delle chiamate, che aiutano gli agenti a concentrarsi sugli aspetti più importanti durante l'interazione con i clienti. Questi riepiloghi catturano automaticamente le parti chiave della conversazione, consentendo agli agenti di identificare rapidamente i punti salienti e fornire un'esperienza eccellente. I manager possono accedere a questi riepiloghi per comprendere il contesto di un'interazione senza dover revisionare l'intera trascrizione. Ciò consente loro di indagare su eventuali problemi da parte dei clienti e prendere misure correttive in modo tempestivo.

   In definitiva, Amazon Transcribe Call Analytics offre un'ampia gamma di strumenti per analizzare e trarre valore dalle conversazioni, migliorando l'efficienza delle operazioni e garantendo un'esperienza di alta qualità per i clienti.

2. *Amazon Transcribe Medical:* è possibile semplificare la trascrizione delle conversazioni mediche con Amazon Transcribe Medical, un servizio di riconoscimento vocale automatico (ASR) che rispetta gli standard di conformità HIPAA [1].

   Amazon Transcribe Medical offre un modo semplice e sicuro per convertire le conversazioni mediche registrate in testo scritto. Grazie alla conformità HIPAA, i dati sensibili dei pazienti vengono trattati nel rispetto delle normative sulla privacy e della sicurezza.

   Utilizzando Amazon Transcribe Medical si possono ottenere trascrizioni accurate e dettagliate delle conversazioni mediche, semplificando la gestione delle informazioni e consentendo una documentazione completa e precisa.

[1] L'Health Insurance Portability and Accountability Act (HIPAA) è una legge federale degli Stati Uniti che definisce i requisiti per il trattamento dei dati sanitari protetti dei privati.

Questo servizio rappresenta un importante aiuto per gli operatori sanitari, consentendo loro di dedicare più tempo all'assistenza ai pazienti anziché alla trascrizione manuale delle informazioni.

### 4.1.1 Spiegazione del funzionamento

Amazon Transcribe è relativamente semplice da utilizzare, merito anche della sua interfaccia amichevole con la quale bisogna interagire.


**Figura 4.1:** Home del servizio Amazon Transcribe di AWS

Come è possibile osservare in Figura 4.1, è possibile scegliere tra tre possibili alternative, due delle quali sono già state spiegate precedentemente. Queste sono:

1. *Crea una trascrizione.*
2. *Crea un processo di analisi delle chiamate:* anche conosciuto come Amazon Transcribe Call Analytics.
3. *Crea una trascrizione medica:* anche conosciuto come Amazon Transcribe Medical.

Poiché non avevamo bisogno di scendere nel dettaglio in merito alle opzioni due e tre, abbiamo optato per una normale trascrizione in tempo reale.

La pagina alla quale si viene rimandati è riassunta in Figura 4.2 ed in Figura 4.3.

In particolare, nella prima figura notiamo come sia possibile avviare una trascrizione in streaming cliccando sull'apposito pulsante, e fin da subito il servizio inizierà la trascrizione di ciò che viene recepito. Nonostante sia possibile lasciare al servizio stesso il compito di rilevare automaticamente la lingua, noi abbiamo deciso di preimpostarla così da evitare eventuali errori di trascrizione. Questo può essere fatto semplicemente selezionando la lingua desiderata nel menù a tendina così come mostrato in Figura 4.3.

### 4.1.2 Esempi svolti

Così come per il precedente servizio, anche per la trascrizione ci siamo avvalsi di alcuni esempi che abbiamo sottoposto al servizio di trascrizione di ciascuna piattaforma Cloud, così da evidenziare aspetti caratteristici e differenze per ciascuna di loro.

Nello specifico, ci siamo muniti di due diverse registrazioni da sottoporre ai vari trascrittori, una in lingua inglese ed una in lingua spagnola. La registrazione in lingua inglese recita

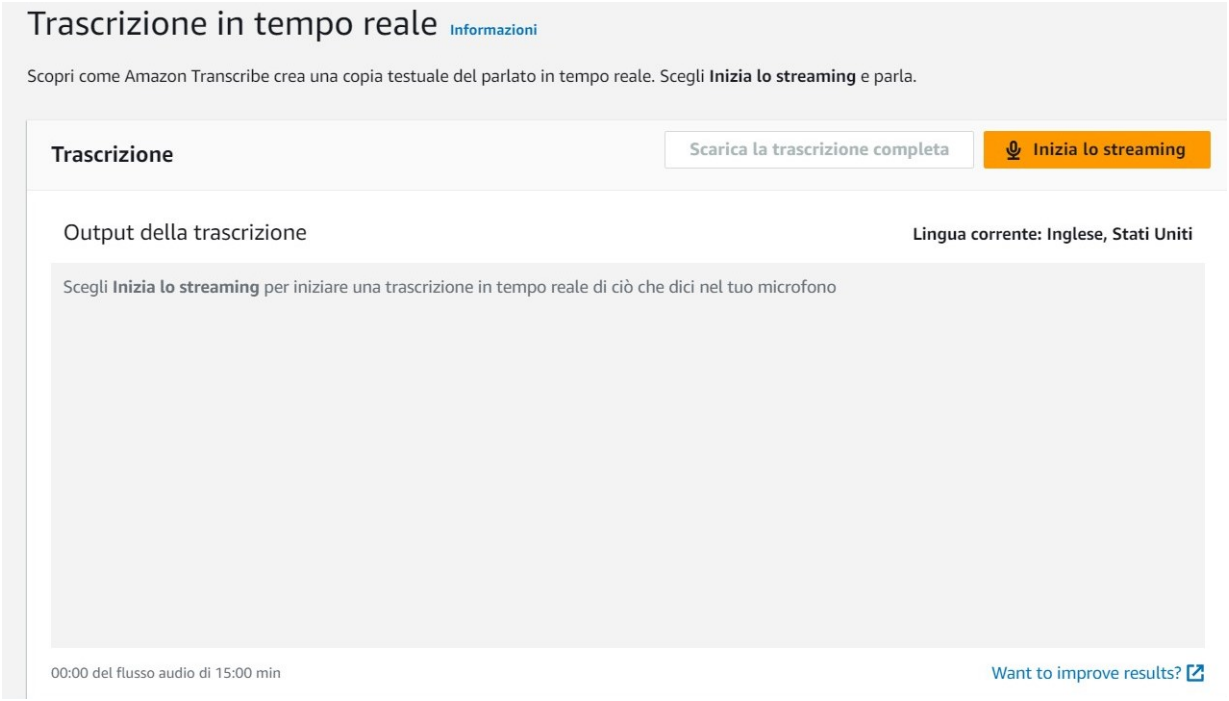


**Figura 4.2:** Home del servizio Amazon Transcribe di AWS per trascrizione in tempo reale

il seguente testo: *Talking in the library is not allowed*, la cui traduzione in lingua italiana corrisponde a: *Non è consentito parlare in biblioteca.* Viceversa, la registrazione in lingua spagnola da noi adoperata esprime quanto segue: *Hola mamá, hoy no vuelvo a casa par almorzar, estoy en el parque con mis amigos, nos vemos esta tarde al cine*, la cui traduzione in lingua italiana è la seguente: *Ciao mamma, oggi non torno a casa per pranzo, sono al parco con i miei amici, ci vediamo al cinema oggi pomeriggio.*

Come possiamo notare dalle immagini riportate sui due esempi svolti, in particolare la Figura 4.4 e la Figura 4.5, solo la prima di queste due frasi risulta correttamente trascritta ed è priva di errori sintattici. Per quanto riguarda la seconda frase, essa risulta essere mancante di alcune parti del discorso, quindi il servizio non ha colto al meglio la registrazione proposta, comportando a sua volta una serie di errori sintattici e grammaticali.

### 4.1.3 Vantaggi e svantaggi del servizio

Dopo aver analizzato il funzionamento del servizio Amazon Transcribe ed alcuni esempi presi in considerazione durante lo studio di quest'ultimo, riportiamo alcune caratteristiche peculiari di questo strumento;

- *Possibilità di scegliere il tipo di trascrizione:* consideriamo questo punto fondamentale, soprattutto dopo aver introdotto i due sottoservizi maggiormente specializzati come Amazon Transcribe Call Analytics ed Amazon transcribe Medical. Questo perché l'avere a disposizione uno strumento strettamente legato al campo nel quale verrà adoperato, significa avere a disposizione uno strumento addestrato e preparato solo ed esclusivamente per quell'intento, il che lo porta ad essere notevolmente preciso ed accurato.
- *Modelli di riconoscimento vocale all'avanguardia, completamente gestiti e costantemente addestrati:* consideriamo anche questa cosa come caratteristica positiva, in quanto l'aggior-

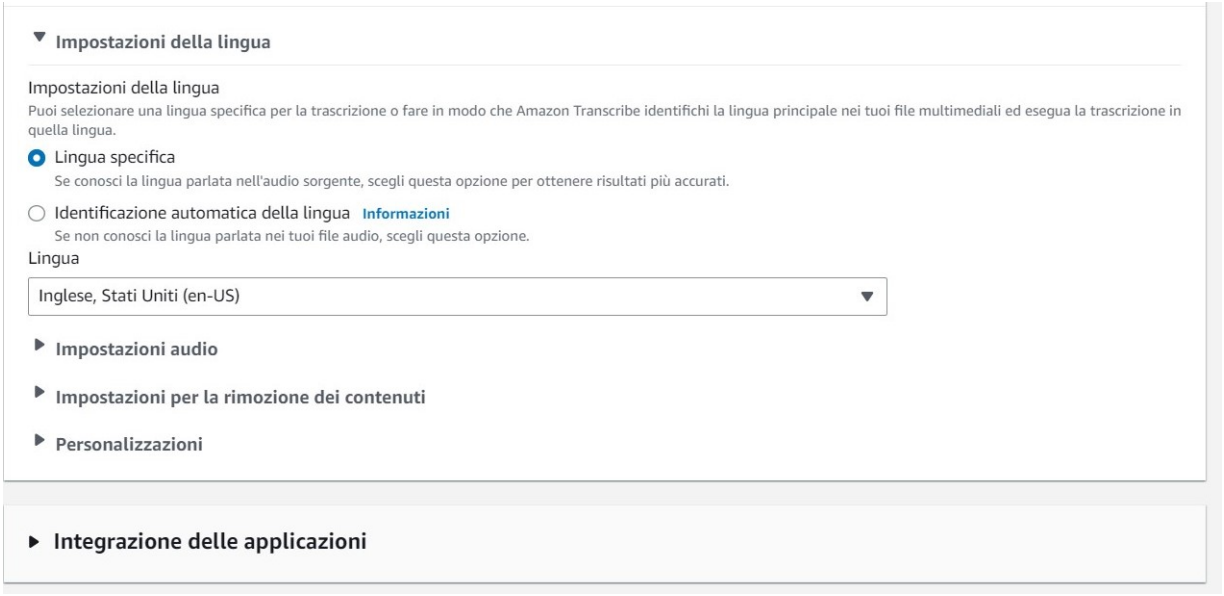


**Figura 4.3:** Home del servizio Amazon Transcribe di AWS per trascrizione in tempo reale


**Figura 4.4:** Trascrizione della prima frase in lingua inglese

namento deve essere nota integrante di qualsiasi sistema o servizio, così da consentire agli utilizzatori di essere sempre all'avanguardia in ciò che fanno.

- *Migliora la precisione grazie a modelli personalizzati che comprendono il vocabolario specifico del dominio:* questo punto riprende in parte il discorso precedentemente affrontato sulla possibilità di avere a disposizione degli utenti sistemi in grado di esprimere al meglio i risultati per quello specifico settore.
- *Assicura la privacy e la sicurezza dei clienti proteggendo le informazioni sensibili:* quest'ultimo punto è ciò che la maggior parte delle aziende richiede, soprattutto in ambiti di ricerca e sviluppo, dove la conoscenza e la capacità sono la marcia in più per arrivare primi negli obiettivi preposti.

In merito ad eventuali svantaggi nell'utilizzo di questo servizio, non abbiamo notato carenze che debbano essere messe in risalto.

## 4.2 Implementazione in Google Cloud

*Speech-to-Text* è il servizio di trascrizione messo a disposizione da Google sulla sua piattaforma Cloud. Speech-to-Text consente l'integrazione facile delle tecnologie di riconoscimento vocale di Google nelle applicazioni per sviluppatori. È possibile inviare i dati audio all'API Speech-to-Text, che restituisce una trascrizione del testo del file audio.

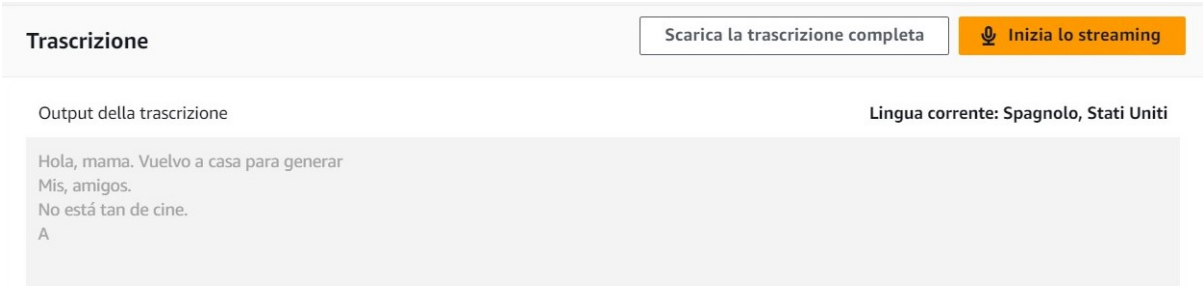


**Figura 4.5:** Trascrizione della seconda frase in lingua spagnola

### 4.2.1 Spiegazione del funzionamento

Prima di poter inviare una richiesta all'API Speech-to-Text, è necessario aver completato le seguenti azioni.

- *Abilitare Speech-to-Text su un progetto Google Cloud.*
- *Impostare la variabile di ambiente di autenticazione.*
- *(Facoltativo) Creare un nuovo bucket Google Cloud Storage per archiviare i dati audio.*

Dopo aver svolto questi primi passi, è sufficiente implementare poche righe di codice per far funzionare l'API e caricare il file audio desiderato. Noi abbiamo deciso di implementare il tutto utilizzando come linguaggio di programmazione Python; di seguito le immagini e la spiegazione del codice riportato.

In Figura 4.6 possiamo notare la definizione del metodo `transcribe_file`, il quale prende come parametro uno `speech_file` equivalente al percorso, in formato stringa, del file audio da caricare. Successivamente viene importata la libreria `speech` che ci permette di sfruttare il servizio di trascrizione, e viene creato l'oggetto `client`.

Fatto ciò, attraverso il comando `open` apriamo e leggiamo il file ricevuto come parametro. Il passo successivo consiste nell'utilizzare dei metodi preimpostati della libreria importata precedentemente, per configurare alcuni parametri di nostro interesse come frequenza, numero canali e la lingua utilizzata.

In Figura 4.7 è riportata la seconda parte di codice. Nello specifico, continuiamo ad utilizzare la solita libreria importata per avviare il processo di trascrizione richiamando sull'oggetto `client` creato precedentemente, il metodo `long_running_recognize`, il quale prende come parametri l'oggetto `config` nel quale abbiamo impostato prima i parametri di nostro interesse, e l'oggetto `audio`.

Il tutto si conclude con una stampa a schermo di una stringa per l'attesa del caricamento del servizio, ed un ciclo `for` per formattare l'output trascritto.

È bene precisare che il formato audio supportato è di tipo .FLAC e non .mp3, nonostante siano presenti dei parametri all'interno dell'oggetto `config` che possano essere utilizzati anche per la conversione con formati .mp3.

Infine, affinché risulti possibile richiamare il metodo appena descritto dal file `main.py`, è necessario importare la libreria così come riportato in Figura 4.8, e creare un oggetto di tipo `SpeechToText` sul quale richiamare il metodo finora descritto, al quale passeremo il percorso del file audio.

### 4.2.2 Esempi svolti

Per il confronto accurato con gli altri servizi messi a disposizione dalle altre piattaforme Cloud, abbiamo utilizzato i soliti due esempi adoperati anche per il precedente servizio messo

SpeechToText.py main.py

```python
# Imports the Google Cloud client library
#Documentazione: https://cloud.google.com/speech-to-text/docs/reference/rest/v1/RecognitionConfig

def transcribe_file(speech_file):
    """Transcribe the given audio file asynchronously."""
    from google.cloud import speech

    client = speech.SpeechClient()

    speech_file = ""+speech_file+""
    with open(speech_file, "rb") as audio_file:
        content = audio_file.read()

    """
     Note that transcription is limited to a 60 seconds audio file.
     Use a GCS file for audio longer than 1 minute.
    """
    audio = speech.RecognitionAudio(content=content)

    config = speech.RecognitionConfig(
        #encoding=speech.RecognitionConfig.AudioEncoding.LINEAR16, #per MP3
        encoding=speech.RecognitionConfig.AudioEncoding.FLAC, #per FLAC
        sample_rate_hertz=48000, #per FLAC
        audio_channel_count=2,
        #sample_rate_hertz=16000, #per MP3
        language_code="es-es",
    )
```

**Figura 4.6:** Implementazione funzione `trascribe_file` parte 1

a disposizione da AWS. In partciolare, la registrazione in lingua inglese recita il seguente testo: *Talking in the library is not allowed*, mentre la registrazione in lingua spagnola da noi adoperata esprime quanto segue: *Hola mamá, hoy no vuelvo a casa par almorzar, estoy en el parque con mis amigos, nos vemos esta tarde al cine* (per le corrispondenti traduzioni in lingua italiana fare riferimento alla Sezione 4.1.2).

Analizzando attentamente la Figura 4.10 e la Figura 4.11, possiamo notare alcune discrepanze tra le due trascrizioni. Nello specifico, la trascrizione dell'audio in lingua inglese non è stata compresa correttamente dal servizio, come appare evidente dalla stringa monotona restituita.

Viceversa la registrazione fornita al servizio in lingua spagnola è stata compresa sicuramente meglio rispetto a quella fornita precedentemente nell'altra lingua, ma si sono comunque verificati errori nella trascrizione, quali ad esempio la mancanza di alcune parti del discorso, o la presenza di alcuni errori grammaticali e sintattici.

### 4.2.3 Vantaggi e svantaggi del servizio

Avendo analizzato in completezza il servizio ed avendo fornito degli esempi pratici, siamo ora in grado di elencare alcuni elementi che mettono in risalto il servizio Speech-to-Text di Google Cloud, o altri che possono essere, viceversa, migliorati ulteriormente per garantire maggiore professionalità.

Tra gli aspetti positivi del servizio abbiamo riscontrato i seguenti:

- *Adattamento vocale:* offre suggerimenti per ottimizzare la precisione nella trascrizione di parole e frasi appartenenti a settori specifici o di utilizzo meno comune. È possibile

```
31
32      operation = client.long_running_recognize(config=config, audio=audio)
33
34      print("Waiting for operation to complete...")
35      response = operation.result(timeout=90)
36
37      # Each result is for a consecutive portion of the audio. Iterate through
38      # them to get the transcripts for the entire audio file.
39      for result in response.results:
40          # The first alternative is the most likely one for this portion.
41          print("Transcript: {}".format(result.alternatives[0].transcript))
42          print("Confidence: {}".format(result.alternatives[0].confidence))
43
```

**Figura 4.7:** Implementazione funzione `trascribe_file` parte 2

SpeechToText.py × main.py ×

```
1   #TranslationAI di google cloud
2   from google.cloud import translate
3   import SpeechToText
4   #import speech_recognition
5
```

**Figura 4.8:** Implementazione `main` parte 1

utilizzare le classi per convertire automaticamente i numeri in forma verbale in indirizzi, anni, valute e altre informazioni simili, al fine di migliorare l'accuratezza complessiva della trascrizione.

- *Modelli specifici del dominio:* consente di scegliere da una selezione di modelli addestrati per il controllo vocale, le chiamate telefoniche e la trascrizione di video, ottimizzati per i requisiti di qualità specifici del dominio.
- *Speech on-device:* permette di eseguire gli algoritmi di riconoscimento e sintesi vocale di Google Cloud localmente su qualsiasi dispositivo, indipendentemente dalla connessione a internet. Caratteristica molto importante al giorno d'oggi, dove, con la potenza degli attuali dispositivi e della rete 5G, è possibile avere sempre a disposizione questo servizio.
- *Modello di base per Speech-to-Text:* si possono creare applicazioni vocali per un pubblico globale con modelli vocali basati su Chirp, il modello di base di Google Cloud per la sintesi vocale, addestrato con milioni di ore di dati audio e miliardi di frasi di testo.

Viceversa, gli aspetti negativi non sono numerosi, ma la necessità di dover scrivere un piccolo programma in Python per l'esecuzione dell'API potrebbe scoraggiare coloro che non sono del settore in merito all'utilizzo del servizio proposto da Google Cloud.

```
33
34  sp = SpeechToText
35  sp.transcribe_file(r"C:\Users\...\Transcribe\Audio\Library-EN.flac")
36
```

**Figura 4.9:** Implementazione `main` parte 2

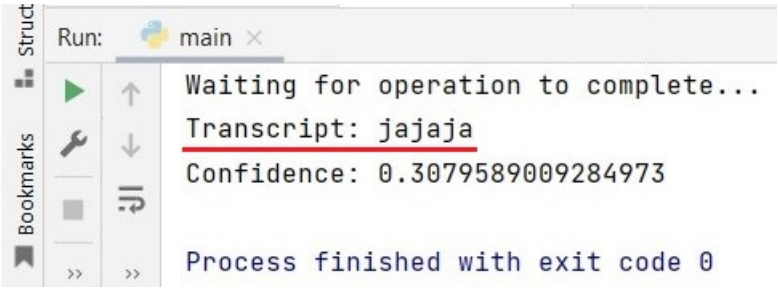


**Figura 4.10:** Trascrizione della frase in lingua inglese proposta dal servizio Speech-to-Text di Google Cloud

## 4.3 Implementazione in Microsoft Azure

Il prodotto *Servizio Voce* di Mircosoft Azure offre un servizio di trascrizione rapido ed accurato dell'audio in testo, supportando oltre 100 lingue e varianti. Inoltre, è possibile personalizzare i modelli per migliorare la precisione della terminologia specifica del dominio di interesse. Questo permette di ottenere un valore aggiunto dall'audio parlato, consentendo la ricerca o l'analisi del testo trascritto, nonché di facilitare l'automazione delle azioni desiderate, tutto ciò utilizzando uno dei linguaggi di programmazione scelto dall'utente.

### 4.3.1 Spiegazione del funzionamento

Anche in questo caso è stato necessario ricorrere alla stesura di un piccolo programma in Python per l'utilizzo dell'API messa a disposizone da Microsoft Azure. In Figura 4.12 possiamo osservare il codice scritto per il funzionamento del servizio, che si riduce ad una semplice funzione `recognize_from_microphone()`.

Le prime due righe sono utilizzate per configurare alcuni parametri fondamentali per il servizio, come ad esempio la `SPEECH_KEY` e la `SPEECH_REGION`, cioè la chiave di accesso al servizio e la regione impostata in fase di attivazione, ma anche la lingua scelta per una determinata trascrizione, in questo caso la lingua inglese, `en-US`.

Fatto ciò, si utilizzano i parametri appena impostati, per la creazione dell'oggetto `speech_recognizer`, che verrà impiegato per richiamare alcuni metodi predefiniti ed impostare lo `speech_recognition_result`.

Successivamente, attraverso una serie di condizioni `if` in cascata, è possibile verificare quale delle condizioni si è verificata, come ad esempio l'andata a buon fine, la mancata comprensione ("matchig)", oppure se sono avvenuti alcuni errori.

Per testare il servizio e la chiamata all'API tramite questa funzione, basta semplicemente richiamarla scrivendo `recognize_from_microphone()` e mandare in input al microfono del proprio dispositivo un segnale audio nella lingua specificata.

### 4.3.2 Esempi svolti

Come è possibile constatare dalla Figura 4.13 e dalla Figura 4.14 che riportano il risultato della trascrizione corrispondenti alle frasi in inglese ed in spagnolo, questa volta il servizio non è riuscito a comprendere l'audio trasmesso. Infatti il messaggio restituito afferma quanto

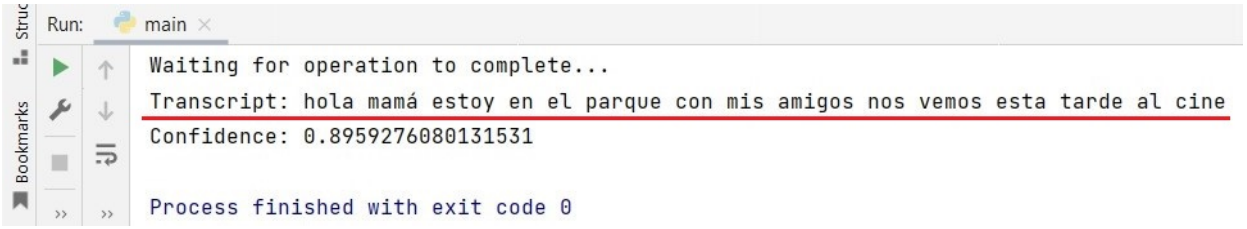


**Figura 4.11:** Trascrizione della frase in lingua spagnola proposta dal servizio Speech-to-Text di Google Cloud

```
def recognize_from_microphone():
    # This example requires environment variables named "SPEECH_KEY" and "SPEECH_REGION"
    speech_config = speechsdk.SpeechConfig(subscription=os.environ.get('SPEECH_KEY'), region=os.environ.get('SPEECH_REGION'))

    speech_config.speech_recognition_language="en-US"

    audio_config = speechsdk.audio.AudioConfig(use_default_microphone=True)
    #audio_config = speechsdk.audio.AudioConfig(filename="YourAudioFile.wav")
    speech_recognizer = speechsdk.SpeechRecognizer(speech_config=speech_config, audio_config=audio_config)

    print("Speak into your microphone.")
    speech_recognition_result = speech_recognizer.recognize_once_async().get()

    if speech_recognition_result.reason == speechsdk.ResultReason.RecognizedSpeech:
        print("Recognized: {}".format(speech_recognition_result.text))
    elif speech_recognition_result.reason == speechsdk.ResultReason.NoMatch:
        print("No speech could be recognized: {}".format(speech_recognition_result.no_match_details))
    elif speech_recognition_result.reason == speechsdk.ResultReason.Canceled:
        cancellation_details = speech_recognition_result.cancellation_details
        print("Speech Recognition canceled: {}".format(cancellation_details.reason))
        if cancellation_details.reason == speechsdk.CancellationReason.Error:
            print("Error details: {}".format(cancellation_details.error_details))
            print("Did you set the speech resource key and region values?")

recognize_from_microphone()
```

**Figura 4.12:** Codice di funzionamento del servizio Servizio Voce di Microsoft Azure

segue: *No speech could be recognized: NoMatchDetails(reason=NoMatchReason.InitialSilenceTimeout)*, ovvero: *Non è stato possibile riconoscere alcun discorso* e, tra parentesi, la possibile causa di ciò, ossia una mancata corrispondenza dovuta ad un iniziale intervallo di tempo privo di audio.

Dopo ciò, possiamo affermare che, a differenza dei servizi di trascrizione precedenti, quello proposto da Microsoft Azure porta con sé un difetto legato all'acquisizione del file audio da trascrivere.

### 4.3.3 Vantaggi e svantaggi del servizio

Di seguito riportiamo alcune note positive del Servizio Voce di Microsoft Azure:

- *Qualità leader di settore:* consente di ottenere riconoscimento vocale all'avanguardia, sintesi vocale realistica e riconoscimento del parlante ottimizzato.
- *Conformità e sicurezza:* i dati rimangono dell'utente. L'input vocale non viene registrato durante l'elaborazione. Questo caratteristica è importante soprattutto nell'ultimo periodo, in cui i dati e la questione privacy sono diventati fondamentali, e le leggi applicate a questo settore sono sempre più rigide.
- *Voci e modelli personalizzabili:* è possibile creare voci personalizzate, aggiungere parole specifiche al vocabolario di base o creare modelli personalizzati.

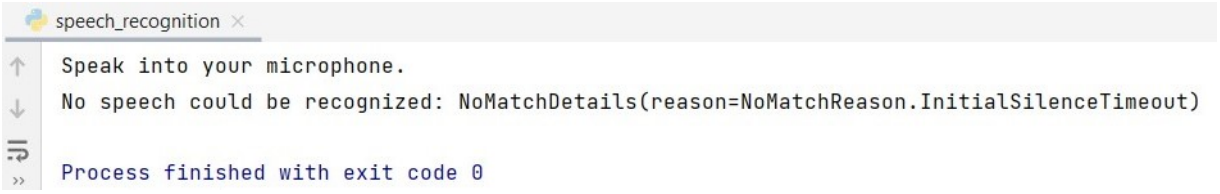


**Figura 4.13:** Trascrizione della frase in lingua inglese del servizio Servizio Voce di Microsoft Azure

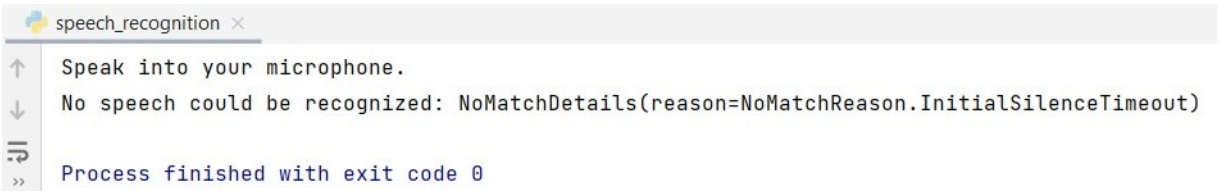


**Figura 4.14:** Trascrizione della frase in lingua spagnola del servizio Servizio Voce di Microsoft Azure

- *Distribuzione flessibile:* altra proprietà da sottolineare, ossia quella di eseguire il Servizio Voce ovunque, sul cloud o nella propria rete.

Viceversa, tra le caratteristiche negative che il servizio presenta, abbiamo deciso di inserire le seguenti:

- *Mancato apprendimento del file audio fornito in input:* il fatto che il servizio proposto da Microsoft Azure non sia riuscito a comprendere l'audio in ingresso lascia perplessi. Nonostante ciò, il servizio rimane comunque un'alternativa valida per coloro che necessitano di strumenti del genere.
- *Necessità di un programma per la chiamata dell'API:* questo punto ritrae in parte lo stesso problema sottolineato durante l'analisi del servizio Speech-to-Text di Google cloud. La necessità di scrivere un programma in uno dei linguaggi di programmazione attuali per l'utilizzo corretto dell'API può diventare fattore fondamentale nel caso di indecisioni sulla scelta dei servizi da utilizzare. Infatti, sotto questo punto di vista, il servizio Amazon Transcribe di AWS risulta il migliore.

## 4.4 Confronto critico tra i tre sistemi di trascrizione

Dopo aver esaminato i singoli servizi di trascrizione proposti dalle varie piattaforme Cloud, possiamo renderci conto, visti anche i numerosi esempi, delle differenze e/o similitudini presenti tra questi. L'aver ottenuto in output risultati distinti ci rimanda alla prima considerazione sulla tipologia/metodologia alla base di questi servizi e che li differenziano dagli altri. Infatti, il servizio Amazon trasncribe di AWS utilizza modelli di apprendimento automatico avanzati per ottenere risultati di trascrizione altamente precisi (come risulta dagli esempi riportati), ed è in grado di supportare diverse lingue e può riconoscere i parlanti.

Microsoft Azure, similmente, adotta la stessa tecnologia. Esso si affida a modelli di linguaggio avanzati e all'apprendimento automatico. Nonostante ci siano le stesse tecniche come fondamento per questi servizi, gli esempi hanno dimostrato che il servizio di AWS è di gran lunga sviluppato meglio rispetto a quello proposto da Microsoft Azure, che non è riuscito a comprendere l'audio.

In merito alle lingue supportate, invece, sia il Servizio Voce che Amazon Transcribe supportano bene o male le stesse lingue. Speech-to-Text di Google Cloud diversamente, è in grado di comprendere un maggior numero di lingue, inclusi i dialetti regionali e le varianti

linguistiche. Google ha fatto investimenti significativi nello sviluppo di tecnologie di riconoscimento vocale multilingue, e ha messo a disposizione un ampio supporto per lingue diverse in molti dei suoi prodotti e servizi. La loro infrastruttura di apprendimento automatico e di Intelligenza Artificiale è stata addestrata su un vasto corpus di dati multilingue, consentendo loro di estendere la copertura linguistica.

D'altra parte, AWS Amazon Transcribe e Microsoft Azure Servizio Voce potrebbero aver deciso di concentrarsi su un set di lingue iniziale più ristretto, magari per ottimizzare le risorse e fornire un servizio di alta qualità in quelle specifiche lingue. Potrebbe essere una strategia per concentrarsi sulla precisione e l'affidabilità della trascrizione in un numero più limitato di lingue, prima di espandersi a livello globale.

È importante notare che la copertura linguistica è un aspetto che può evolvere nel tempo. I fornitori di servizi cloud continuano a migliorare i loro modelli e ad aggiungere nuove lingue in base alla domanda e alle esigenze degli utenti. Quindi, è possibile che nel tempo la copertura linguistica di ciascun servizio possa cambiare e allinearsi maggiormente tra di loro.

Tutti e tre i servizi sono sviluppati al meglio per quanto riguarda scalabilità e prestazioni, infatti Amazon Transcribe di AWS è altamente scalabile, consentendo di gestire facilmente carichi di lavoro di grandi dimensioni. Offre anche la possibilità di elaborare più trascrizioni contemporaneamente. Google Cloud Speech-to-Text è altamente scalabile anch'esso, e può elaborare trascrizioni in tempo reale con latenza ridotta. È in grado di gestire volumi di lavoro elevati. Infine, il Servizio Voce di Microsoft Azure, offre scalabilità per supportare carichi di lavoro di grandi dimensioni, garantisce prestazioni affidabili anche in caso di picchi di traffico.

Tornando al tema principale, ossia quello del Cognitive Computing, con questi servizio a disposizione dell'uomo, è possibile migliorare ulteriormente i confini dell'Intelligenza Artificiale. Infatti, è possibile notare come Amazon Transcribe si integri facilmente con altri servizi AWS, consentendo di utilizzare le trascrizioni come input per ulteriori analisi e sviluppo di applicazioni basate sull'Intelligenza Artificiale. Google Speech-to-Text può essere integrato con altre API di Google Cloud, consentendo l'elaborazione di testi trascritti per scopi di Intelligenza Artificiale, ed il Servizio Voce di Microsoft Azure è parte di un'ampia suite di servizi di Intelligenza Artificiale offerti da Microsoft, consentendo l'integrazione con altri strumenti e tecnologie AI.

In conclusione, sia Amazon Transcribe, Google Speech-to-Text che Microsoft Servizio Voce offrono servizi di trascrizione altamente precisi e con funzionalità avanzate. La scelta dipenderà dalle specifiche esigenze del progetto e dalla preferenza per determinate funzionalità o integrazioni con altri servizi di Intelligenza Artificiale.

CAPITOLO 5

# Chatbot: implementazione in AWS, Google, Azure e Salesforce

*In un'era sempre più orientata alla tecnologia, i chatbot si sono affermati come una soluzione innovativa per migliorare l'interazione tra le aziende e i loro clienti. In questo capitolo, esploreremo i chatbot avanzati offerti dai principali fornitori di servizi cloud, tra cui AWS, Google Cloud, Azure e Salesforce, focalizzandoci sulle loro caratteristiche, funzionalità e vantaggi distintivi. Inizieremo esaminando le soluzioni di chatbot fornite da AWS, un leader nel settore dei servizi cloud. Esploreremo come Amazon Lex sfrutti le potenti funzionalità di Machine Learning per offrire esperienze conversazionali fluide e intuitive. Successivamente passeremo al servizio Dialogflow di Google Cloud, QnA Maker di Microsoft Azure, osservando come esso semplifichi la creazione e la gestione di chatbot intelligenti; infine, ci concentreremo sul chatbot di Salesforce, uno dei principali fornitori di soluzioni CRM.*

## 5.1 Introduzione ai chatbot

Un chatbot è un programma informatico che consente alle persone di interagire con la tecnologia attraverso vari metodi di input, come voce, testo, gesti e tocchi, disponibile 24 ore al giorno, 7 giorni su 7 e 365 giorni all'anno.

Per molti anni, i chatbot sono stati principalmente utilizzati nel servizio clienti, ma ora trovano impiego in diverse altre funzioni aziendali per migliorare l'esperienza del cliente e l'efficienza operativa.

Conosciuti con diversi nomi, come AI bot conversazionali, assistenti AI, assistenti virtuali intelligenti, assistenti clienti virtuali, assistenti digitali, agenti conversazionali, agenti virtuali, interfacce conversazionali, e altro ancora, i chatbot stanno diventando sempre più popolari. Tuttavia, così come i chatbot hanno diversi nomi, essi presentano anche diversi livelli di intelligenza. Un chatbot di base potrebbe essere semplicemente una soluzione per rispondere alle domande frequenti standard.

I chatbot creati utilizzando alcuni dei framework di chatbot attualmente disponibili possono offrire funzionalità leggermente più avanzate, come il completamento dei campi o altre semplici capacità transazionali. Ma sono solo i chatbot avanzati di Intelligenza Artificiale conversazionale che hanno l'intelligenza e la capacità di offrire l'esperienza sofisticata che la maggior parte delle imprese sta cercando di implementare.

Negli ultimi anni, il panorama tecnologico è stato trasformato dagli smartphone, dai dispositivi elettronici e dall'Internet of Things (IoT). Nonostante le dimensioni di questi dispositivi siano sempre più ridotte, la potenza di calcolo al loro interno è aumentata significativamente.

Tuttavia, le app mobili e le attività che richiedono grandi quantità di dati spesso non si integrano bene. Inoltre, i consumatori non sono più disposti ad essere limitati ai metodi

di comunicazione scelti da un'organizzazione ma, desiderano interagire con la tecnologia attraverso una vasta gamma di canali.

I chatbot offrono una soluzione a questi problemi, consentendo ai clienti di ottenere ciò di cui hanno bisogno in modo semplice, su diversi canali, ovunque si trovino, a qualsiasi ora desiderino.

Diamo uno sguardo al funzionamento dei chatbot.

Ad un livello basilare, c'è un'interazione tra un essere umano ed un chatbot. Se viene utilizzata la voce, il chatbot converte i dati vocali percepiti in testo, utilizzando la tecnologia di riconoscimento automatico del parlato, *Automatic Speech Recognition (ASR)*. Viceversa, i chatbot basati esclusivamente su testo, come i servizi di messaggistica, saltano questa prima fase.

Successivamente, il chatbot analizza il testo ottenuto in input, valuta la migliore risposta e la restituisce all'utente. L'output della risposta del chatbot può essere consegnato in diversi modi, come testo scritto, voce, attraverso strumenti di *Text to Speech (TTS)*, o completando un'azione specifica.

È importante notare che la comprensione umana non è un compito facile per una macchina. La modalità sottile e sfumata con cui gli esseri umani comunicano rappresenta una sfida complessa da ricreare artificialmente, ed è per questo motivo che i chatbot utilizzano diversi principi del linguaggio naturale. Tra questi principi evidenziamo:

- *Il Natural Language Processing (NLP):* l'elaborazione del linguaggio naturale viene utilizzata per suddividere l'input dell'utente in frasi e parole. Inoltre, essa standardizza il testo attraverso una serie di tecniche, come la conversione in minuscolo o la correzione degli errori di ortografia, prima di determinare se una parola è un aggettivo o un verbo. Durante questa fase, vengono anche considerati fattori come il sentiment.

- *Il Natural Language Understanding (NLU):* esso aiuta il chatbot a comprendere ciò che l'utente ha detto, utilizzando sia elementi linguistici generali che specifici del dominio, come lessico, sinonimi e temi. Questi elementi vengono, quindi, combinati con algoritmi o regole per creare flussi di dialogo che indicano al chatbot come rispondere.

- *Il Natural Language Generation (NLG):* affinché risulti possibile offrire un'esperienza significativa e personalizzata al di là delle risposte predefinite, è necessaria la generazione di linguaggio naturale. Ciò consente al chatbot di interrogare archivi di dati, compresi sistemi integrati di backend e database di terze parti, e utilizzare tali informazioni per creare una risposta.

- *La Conversational AI Technology:* esso migliora il Natural Language Processing e il Natural Language Understanding portandoli al livello successivo. La Conversational AI Technology consente alle aziende di creare sistemi di dialogo avanzati basati sull'utilizzo della memoria, delle preferenze personali e della comprensione del contesto, per fornire un'interfaccia di linguaggio naturale realistica e coinvolgente.

## 5.2 Implementazione in AWS

Amazon Lex è un servizio di Intelligenza Artificiale (IA) completamente gestito con modelli avanzati di linguaggio naturale per progettare, costruire, testare e distribuire interfacce di comunicazione nelle applicazioni. È il servizio di IA alla base di Amazon Alexa, il famoso assistente virtuale di Amazon. Di seguito alcune caratteristiche di Amazon Lex:

- *Elaborazione del linguaggio naturale (NLP):* Amazon Lex utilizza avanzate tecniche di elaborazione del linguaggio naturale per interpretare e comprendere l'input umano. Può riconoscere frasi e domande complesse e generare risposte appropriate.
- *Creazione di chatbot:* Amazon Lex semplifica la creazione di chatbot personalizzati senza la necessità di sviluppare codice da zero. Esso fornisce un'interfaccia di sviluppo visuale che consente di definire l'architettura delle conversazioni, creare gli intenti (le azioni che il chatbot può eseguire) e gestire i modelli di linguaggio.
- *Integrazione con altri servizi AWS:* Amazon Lex può essere facilmente integrato con altri servizi di AWS.
- *Scalabilità e disponibilità.*
- *Analisi delle conversazioni:* Amazon Lex offre funzionalità di analisi delle conversazioni che consentono di estrarre informazioni utili dai dialoghi utente-chatbot. Ciò può aiutare a identificare le esigenze degli utenti, migliorare le risposte del chatbot e ottenere insight sulle interazioni degli utenti.

### 5.2.1 Spiegazione del funzionamento

In breve il ciclo di funzionamento di Amazon Lex è mostrato nella Figura 5.1.

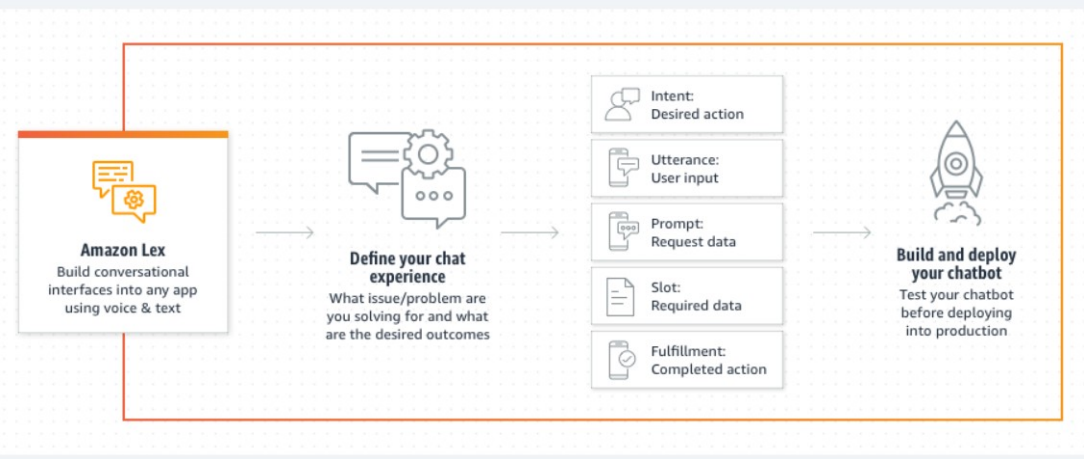


**Figura 5.1:** Schema di funzionamento di Amazon Lex

Tale ciclo consiste nei seguenti passi:

- *Progettazione del chatbot:* il primo passo da fare è quello di comprendere quali scopi o problemi il nostro Chatbot dovrà risolvere. Fatto ciò saremo poi in grado di continuare con i passi successivi, come, ad esempio, la creazione degli *intenti*, che rappresentano le azioni che il chatbot può eseguire, e la specifica delle frasi di esempio che gli utenti possono utilizzare per attivare ciascuno di essi.
- *Creazione dei modelli di linguaggio:* è opportuno insegnare al chatbot come comprendere l'input dell'utente attraverso la costruzione di modelli di linguaggio. Questi modelli definiscono le parole chiave, le frasi di esempio e i pattern di linguaggio che indicano l'intento dell'utente. A tal proposito, Amazon Lex utilizza tecniche di Machine Learning per addestrare i modelli di linguaggio in base ai dati forniti.

- *Configurazione delle risposte del chatbot:* verranno definite le risposte che il chatbot dovrà fornire agli utenti. Queste possono essere semplici stringhe di caratteri oppure possono contenere prompt per la raccolta di ulteriori informazioni da parte degli utenti.
- *Integrazione con altri servizi:* se necessario, è possibile integrare Amazon Lex con altri servizi AWS così da personalizzare il proprio codice affinché rispetti gli obiettivi prestabiliti.
- *Monitoraggio e analisi:* Amazon Lex fornisce strumenti di monitoraggio e analisi per valutare le prestazioni del chatbot. È possibile ottenere dati sulle conversazioni degli utenti, identificare eventuali problemi o aree di miglioramento e apportare modifiche al chatbot per ottimizzarne le prestazioni.

### 5.2.2 Esempi svolti

Vediamo ora un esempio svolto per la crezione di un intento riguardante una prenotazione. In Figura 5.2 è riportato il flusso generale dell'intento creato.

Il procedimento di messa in atto dell'intento inizia con un'affermazione che l'utente inserisce e che viene riconosciuta dal chatbot. Questo passo è evidenziato in Figura 5.3, nella quale sono state inserite quattro frasi di possibile apertura del discorso da parte dell'utente e, alle quali, il chatbot saprà rispondere correttamente.

Successivamente, una volta appreso l'avvio dell'intento da parte del chatbot, questo deve essere in grado di rispondere correttamente. Ciò è stato fatto impostando delle possibili risposte alle affermazioni viste precedentemente e che riportiamo in Figura 5.4.

Dopo aver stemperato l'approccio iniziale, il chatbot è stato progettato per l'emissione di un prompt con lo scopo di acquisire maggiori informazioni da parte dell'utente. In particolare, nella Figura 5.5, è possibile notare come ciò accada richiedendo il numero di telefono di quest'ultimo.

A questa azione seguono immediatamente alcune condizioni; se l'utente fornisce il proprio numero di telefono si procederà in una direzione, altrimenti, nel caso in cui l'utente decidesse di non fornirlo, si procederebbe in maniera differente. Questa duplice opzione vale sia nel momento di conferma del numero inserito da parte dell'utente, sia per la conferma dell'adempimento ad una delle due opzioni che potrebbe verificarsi (Figura 5.6).

Il discorso viene concluso con una risposta di chiusura che verrà proposta quando l'utente sarà stato soddisfatto del servizio ricevuto. Questa fase è stata riportata nella Figura 5.7.

Il processo generale di creazione e gestione di un intento è così concluso. È importante, però, notare che il discorso può essere maggiormente arricchito con molteplici opzioni ed interventi, quali, ad esempio, il caso in cui il chatbot non comprenda l'input inserito dall'utente (Figura 5.8), oppure situazioni in cui l'utente impiega molto tempo nel fornire una risposta. In questo caso è possibile settare determinate tempistiche, scadute le quali spetterà poi al chatbot intervenire e riproporre all'utente l'opzione corrente.

### 5.2.3 Vantaggi e svantaggi del servizio

L'utilizzo di Amazon Lex presenta diversi vantaggi e svantaggi. Ecco una panoramica dei principali vantaggi riscontrati:

- *Semplicità di sviluppo:* Amazon Lex fornisce un'interfaccia visuale intuitiva che semplifica la creazione e la gestione dei chatbot basati sul linguaggio naturale. Non è necessario avere una conoscenza approfondita di programmazione per iniziare ad utilizzare il servizio.

- *Integrazione con altri servizi AWS:* è possibile integrare Amazon Lex con altri servizi di AWS per arricchire e personalizzare la propria applicazione.
- *Supporto multi-canale:* Amazon Lex supporta l'integrazione con diverse piattaforme di messaggistica e canali di comunicazione, come applicazioni web o applicazioni mobile. Ciò consente di raggiungere gli utenti su diverse piattaforme con lo stesso chatbot.
- *Elaborazione del linguaggio naturale avanzata:* Amazon Lex utilizza tecniche di elaborazione del linguaggio naturale avanzate per comprendere e interpretare l'input utente in modo accurato. Esso può gestire frasi complesse, riconoscere sinonimi e rispondere in modo coerente.
- *Scalabilità:* Amazon Lex è altamente scalabile e può gestire carichi di lavoro di grandi dimensioni.

Viceversa, tra gli svantaggi possiamo notare i seguenti:

- *Limitazioni delle funzionalità predefinite:* le funzionalità predefinite di Amazon Lex potrebbero non essere sufficienti per alcuni casi d'uso avanzati. Potrebbe essere necessario estendere il chatbot utilizzando codice personalizzato o integrando altri servizi AWS.
- *Personalizzazione complessa:* per implementare ulteriormente il proprio chatbot oltre le funzionalità predefinite potrebbe essere necessario l'utilizzo di altre funzionalità, come AWS Lambda o altri servizi inerenti.
- *Dipendenza dalla piattaforma AWS:* l'utilizzo di Amazon Lex è prevalentemente consentito sulla piattaforma Cloud di AWS. Se si desidera spostare il chatbot su un'altra piattaforma o servizio, potrebbe essere necessario affrontare problemi di migrazione.
- *Costi:* così come gli altri servizi analizzati finora, l'utilizzo di Amazon Lex comporta costi in base all'utilizzo, come il numero di richieste e la quantità di dati elaborati. Per le applicazioni che richiedono un alto numero di richieste i costi possono aumentare.
- *Riconoscimento limitato delle intenzioni:* nonostante l'elaborazione del linguaggio naturale avanzata, il riconoscimento delle intenzioni potrebbe non essere sempre accurato. È possibile che il chatbot possa fraintendere o non comprendere correttamente l'input utente, richiedendo iterazioni aggiuntive come quelle di cui si è discusso alla fine della sezione precedente.

## 5.3 Implementazione in Google Cloud

Dialogflow è una piattaforma di sviluppo di chatbot e assistenti virtuali basata sull'elaborazione del linguaggio naturale (NLP) offerta da Google Cloud. Essa fornisce strumenti avanzati per la creazione di conversazioni interattive basate sul linguaggio naturale. Infatti Dialogflow utilizza sofisticate tecniche di elaborazione del linguaggio naturale per interpretare e comprendere l'input umano. Il tutto usufruibile sulla piattaforma di Google Cloud.

### 5.3.1 Spiegazione del funzionamento

Illustriamo, di seguito, il funzionamento del servizio dando alcune definizioni importanti.

Un *agente* Dialogflow è un agente virtuale in grado di gestire le interazioni con i suoi utenti finali. È un modulo basato sulla comprensione del linguaggio naturale, testo o audio, per poi essere tradotto in dati strutturati comprensibili per app e servizi.

Un *intent* è un'etichetta che identifica l'intenzione dell'utente finale in una singola conversazione. Per ogni agente, è necessario definire numerosi intent, in modo che, combinati insieme, possano gestire una conversazione completa. Quando un utente finale scrive o descrive qualcosa, detto anche "espressione utente finale", Dialogflow associa quello che scrive l'utente all'intent più appropriato dell'agente. Tale associazione è anche nota come *classificazione degli intent*. Ad esempio, volendo creare un agente meteo è logico pensare alla definizione di un intento legato a domande sulle previsioni meteo, come riportato nella Figura 5.9.

In generale, un intent di base contiene i seguenti quattro punti:

1. *Frasi di addestramento:* cioè frasi di esempio relative a ciò che gli utenti finali potrebbero dire. Quando un'espressione dell'utente finale è simile ad una di queste frasi, Dialogflow restituisce l'intent. Tuttavia, non è necessario definire tutti i possibili esempi, perché il Machine Learning integrato in Dialogflow si espande nell'elenco con altre frasi simili.

2. *Azione:* è possibile definire un'azione per ciascun intent. Quando viene trovata una corrispondenza con uno degli intent definiti, Dialogflow lancia l'azione prestabilita per quello specifico intent.

3. *Parametri:* quando un intent viene abbinato in fase di runtime, Dialogflow fornisce i valori estratti dall'espressione dell'utente finale come parametri. Ciascuno di questi parametri ha un tipo, detto tipo di entità, che esplicita esattamente come vengono estratti i dati.

4. *Risposte:* ovvero le risposte testuali, visive o vocali, da tornare all'utente finale.

Abbiamo poco fa accennato al tipo di entità riferito ai parametri degli intent. A tal proposito Dialogflow fornisce una serie di entità di sistema predefinite in grado di soddisfare gran parte dei dati comuni. Ci sono, infatti, corrispondenze per date, orari, colori, indirizzi email e così via. È anche possibile definire entità personalizzate legate ai corrispettivi dati personalizzati.

I *contesti* di Dialogflow sono simili al contesto del linguaggio naturale; essi necessitano di ulteriori spiegazioni per comprendere ciò a cui una persona fa riferimento nello specifico. Analogamente, per far sì che Dialogflow gestisca un'espressione dell'utente finale, è necessario fornirne il contesto per abbinare correttamente un intent.

È possibile configurare i contesti per un intent impostando i *contesti di input e di output*, che sono identificati da nomi di stringa. Quando un intent trova una corrispondenza, tutti i contesti di output configurati per quell'intent diventano attivi. Quando i contesti sono attivi, Dialogflow ha maggiori probabilità di corrispondere a intent configurati con contesti di input che corrispondono ai contesti attualmente attivi. In Figura 5.10 è riportato un esempio di utilizzo di un contesto, il cui funzionamento è il seguente:

- L'utente finale chiede informazioni sul proprio account di controllo.

- Dialogflow associa questa richiesta dell'utente finale all'intent "CheckingInfo". Questo intent ha un contesto di output chiamato "checking" che diventa attivo.

- L'agente chiede all'utente finale il tipo di informazioni specifiche relative al proprio account di controllo che desidera ottenere.

- L'utente finale risponde con "il mio saldo".

- Dialogflow abbina questa espressione dell'utente finale all'intent "CheckingBalance". Questo intent ha un contesto di input checking, che deve essere attivo per farlo. Potrebbe anche esistere un intent "SavingsBalance" simile per stabilire la corrispondenza con la stessa espressione dell'utente finale quando è attivo un contesto savings.
- Dopo aver effettuato le opportune query al database, l'agente fornisce la risposta con il saldo dell'account di controllo.

Risulta possibile utilizzare particolari tipi di intent, detti di *follow-up*, per "settare" automaticamente i contesti per coppie di intent. Un intent di follow-up è un intent aggiuntivo collegato all'intent principale. Quando si crea un intent di follow-up, viene automaticamente creato un contesto di output con lo stesso nome dell'intent principale. L'associazione di un intent di follow-up avviene solo quando l'intent principale è stato rilevato nella conversazione precedente. È anche possibile creare più livelli di intenti di follow-up nidificati.

Possiamo anche decidere di utilizzare risposte dinamiche agli intent attraverso i cosiddetti *fulfillment*. Quando si abilita il fulfillment per un intent, Dialogflow risponde a tale servizio chiamando un servizio da noi definito.

Per ogni intent esiste una configurazione per abilitare il fulfillment. Se un intent richiede un'azione da parte del sistema o una risposta dinamica, è necessario attivare il fulfillment per quell'intent. Se viene rilevata una corrispondenza con un intent che non ha il fulfillment abilitato, Dialogflow utilizzerà la risposta statica definita per esso. Quando viene trovata una corrispondenza con un intent che ha il fulfillment abilitato, Dialogflow invierà una richiesta al servizio webhook con le informazioni sull'intent corrispondente. Il sistema potrà poi eseguire le azioni richieste e rispondere a Dialogflow fornendo indicazioni su come procedere. Quando il fulfillment è abilitato, la risposta statica definita per quell'intent viene utilizzata solo se il servizio webhook non è in grado di fornire una risposta.

In Figura 5.11 è mostrato il flusso di elaborazione per l'evasione degli ordini. Questo è quanto accade:

- L'utente finale digita o pronuncia un'espressione.
- Dialogflow abbina l'espressione dell'utente finale a un intent e ne estrae i parametri.
- Dialogflow invia un messaggio richiesta webhook al servizio webhook. Questo messaggio contiene informazioni sull'intent corrispondente, sull'azione, sui parametri e sulla risposta definita per l'intent.
- Il servizio esegue le azioni necessarie.
- Il servizio invia un messaggio risposta webhook a Dialogflow. Questo messaggio contiene la risposta che deve essere inviata all'utente finale.
- Dialogflow invia la risposta all'utente finale.
- L'utente finale vede o sente la risposta.

### 5.3.2 Esempi svolti

Riportiamo, di seguito, alcuni esempi svolti in merito al servizio Dialogflow. In Figura 5.12 è riportato il workflow generale dell'esempio, mentre nella Figura 5.13 possiamo notare i dettagli della finestra "start".

Tra le cose rilevanti della finestra "start" notiamo la rotta di benvenuto, alcune rotte predefinite per i saluti e degli handler per la gestione degli errori, quali, ad esempio, la mancata corrispondenza oppure il mancato input da parte dell'utente.

Vale la pena soffermarsi anche sulla Figura 5.14, nella quale possiamo osservare, oltre alle solite rotte di benvenuto, anche altri gruppi di rotte. Ciò risulta più evidente e meglio espresso nella Figura 5.15, nella quale viene esteso il campo "Confirmation" che ci permette di cogliere alcuni degli intent definiti alla destra della finestra.

Infine, nella Figura 5.16, è stato riportato un breve scambio di messaggi avvenuto tra noi ed il chatbot.

### 5.3.3 Vantaggi e svantaggi del servizio

Dopo aver analizzato attentamente il funzionamento del servizio Dialogflow di Google, ed aver riportato alcuni esempi svolti, siamo ora in grado di elencare quali possono essere i principali vantaggi o svantaggi di questo servizio.

Tra i vantaggi notiamo i seguenti:

- *Elaborazione del linguaggio naturale avanzata:* Dialogflow utilizza algoritmi di elaborazione del linguaggio naturale avanzati di Google, consentendo una migliore comprensione dell'input utente e una maggiore accuratezza nella rilevazione delle intenzioni.
- *Facilità di utilizzo:* Dialogflow fornisce un'interfaccia utente intuitiva e user-friendly che semplifica la creazione e la gestione dei chatbot, anche per gli sviluppatori meno esperti.
- *Supporto per le conversazioni contestuali:* Dialogflow permette la gestione delle conversazioni contestuali, mantenendo lo stato delle interazioni precedenti e fornendo risposte coerenti con il contesto. Ciò consente conversazioni più naturali e dinamiche.
- *Ampia gamma di integrazioni:* Dialogflow offre una vasta gamma di integrazioni con piattaforme di messaggistica e canali di comunicazione, incluso Google Assistant.
- *Integrazione con servizi Google Cloud:* Dialogflow si integra nativamente con altri servizi di Google Cloud, come Google Cloud Functions e Google Cloud Storage, che offrono ulteriori funzionalità e opzioni di personalizzazione.

Viceversa, tra gli svantaggi possiamo annotare i seguenti:

- *Limitazioni delle funzionalità gratuite:* l'opzione gratuita di Dialogflow ha alcune limitazioni, come il numero di richieste al giorno e l'accesso a funzionalità avanzate, quali l'apprendimento automatico.
- *Dipendenza da Google Cloud:* l'utilizzo di Dialogflow implica una dipendenza dal cloud di Google. Questo potrebbe essere un fattore da considerare se si desidera evitare la dipendenza da un'unica piattaforma di servizi cloud.

## 5.4 Implementazione in Microsoft Azure

Il servizio di Microsoft Azure che consente di creare chatbot è QnA Maker. Esso rappresenta un servizio basato su cloud per l'elaborazione del linguaggio naturale (NLP, Natural Language Processing) che permette di creare un'interazione conversazionale naturale con i dati. È utilizzato per trovare la risposta più adatta a qualsiasi input in base alle informazioni

personalizzate contenute nella Knowledge Base (KB). Esso viene generalmente utilizzato per lo sviluppo di applicazioni come social media, chatbot o applicazioni desktop per la trascrizione del linguaggio.

QnA Maker è consigliato:

- *Quando si hanno informazioni statiche:* QnA Maker viene impiegato quando è necessario utilizzare una knowledge base di risposte contenenti informazioni statiche. Tale knowledge base può essere personalizzata in base alle specifiche esigenze e può essere creata utilizzando documenti come PDF e URL.
- *Quando si vuole fornire la stessa risposta a una richiesta, a una domanda o a un comando:* se diversi utenti inviano la stessa domanda, viene restituita la stessa risposta.
- *Quando si vogliono filtrare informazioni statiche in base a meta informazioni:* l'aggiunta di tag di metadati consente di fornire ulteriori opzioni di filtro rilevanti per gli utenti dell'applicazione client e per le informazioni stesse. I metadati comuni includono informazioni quali chiacchiere, tipo o formato, scopo e aggiornamenti del contenuto.
- *Quando si vuole gestire una conversazione con bot che include informazioni statiche:* una knowledge base fornisce una risposta al comando o al testo inserito dall'utente durante la conversazione. Se la risposta fa parte di un flusso di conversazione predefinito, rappresentato nella knowledge base con un contesto a più turni, il bot può agevolmente seguire tale flusso.

### 5.4.1 Spiegazione del funzionamento

Dopo aver presentato una panoramica generale sul servizio proposto da Microsoft Azure, analizziamo più nel dettaglio il funzionamento di questo.

Per fare ciò è necessario fornire la definizione di *"knowledge base"*, una raccolta che QnA Maker importa e che risulta costituita da coppie domande-risposte. Il processo di importazione estrae informazioni sulla relazione tra le parti del contenuto strutturato e semistrutturato in modo da implicare le relazioni tra le coppie di domande e risposte. Queste coppie possono essere, comunque, modificate; è anche possibile aggiungere direttamente delle coppie nuove. Il contenuto di queste coppie contiene generalmente tutte le possibili forme alternative delle domande, i tag dei metadati usati per filtrare le opzioni di risposta ed una richiesta di completamento, per continuare l'affinamento della ricerca. In Figura 5.17 è riportato un esempio di knowledge base con tutte le caratteristiche appena descritte.

Dopo aver pubblicato la knowledge base, un'applicazione client invia la domanda di un utente all'endpoint. Il servizio QnA Maker elabora la domanda e restituisce la risposta migliore (generalmente fornita in formato JSON).

Per quanto riguarda la creazione di codice per un chatbot, le fasi che avvengono in ambito di comunicazione tra i servizi in gioco sono in genere le seguenti tre:

1. L'applicazione client invia la domanda dell'utente (testo in parole proprie) all'endpoint della knowledge base.
2. QnA Maker usa la knowledge base sottoposta a training per fornire la risposta corretta e le eventuali richieste di completamento che possono servire per affinare la ricerca della risposta migliore. QnA Maker restituisce una risposta in formato JSON.
3. L'applicazione client usa la risposta JSON per prendere decisioni su come continuare la conversazione. Queste decisioni possono includere la visualizzazione della risposta principale e la presentazione di più scelte per affinare la ricerca della risposta migliore.

Il portale QnA Maker offre un'esperienza completa per la creazione di knowledge base. È possibile importare i documenti nel loro formato corrente nella knowledge base. Questi documenti, come FAQ, manuali dei prodotti, fogli di calcolo o pagine Web, vengono convertiti in coppie domanda-risposta. È possibile eseguire la scansione di ogni coppia per suggerimenti sul completamento e collegamenti ad altre coppie. Il formato "Markdown" finale supporta presentazioni avanzate con immagini e collegamenti.

QnA Maker ha, tra le sue numerosi funzioni, quella di offrire richieste a più turni permettendo, così, un miglioramento delle coppie domande-risposte. Le richieste a più turni offrono la possibilità di collegare coppie di domande e risposte. Questa sorta di collegamento permette di definire una risposta principale ed utilizzare le restanti domande per raffinare il risultato finale.

Inoltre, QnA Maker offre anche molteplici suggerimenti in fase di utilizzo, come suggerimenti su quali modifiche apportare alla knowledge base per migliorare la qualità. Tale comportamento viene anche detto *apprendimento attivo*.

### 5.4.2 Esempi svolti

Di seguito riportiamo gli esempi svolti durante l'analisi del servizio. In Figura 5.18 notiamo l'impostazione delle domande e delle risposte per la fase di avvio del chatbot. In particolare, esso è stato programmato per rispondere a determinati input, quali ad esempio "Goodmorning", "Hi" oppure "Hey". Il chatbot produrrà, di conseguenza, una riposta di cortesia chiedendo come potrebbe tornare utile la sua interazione.

In aggiunta a ciò sono state previste anche delle opzioni selezionabili dall'utente (Figura 5.19), come, ad esempio, la possibilità di conoscere più approfonditamente il chatbot (scegliendo l'opzione numero uno, *About me*), oppure variando completamente, e quindi ottenere un racconto casuale da quest'utlimo.

Il chatbot può essere programmato secondo il proprio volere, prevedendo numerevoli opzioni e con un alto numero di strade percorribili. Nel nostro caso abbiamo aggiunto un numero di funzionalità in più limitate, come risulta evidente nella Figura 5.20.

Per accertarci della corretta implementazione, la piattaforma cloud di Microsof Azure mette a disposizione un'area nella quale è possibile testare il proprio chatbot con le impostazioni appena inserite (Figure 5.21 e 5.22).

### 5.4.3 Vantaggi e svantaggi del servizio

Descriviamo quelli che, secondo noi, possono rappresentare i vantaggi o gli svantaggi dell'utilizzo di un servizio come QnA Maker di Microsoft Azure. Tra i vantaggi inseriamo i seguenti:

- *Facilità di creazione di una base di conoscenza:* QnA Maker semplifica il processo di creazione di una base di conoscenza interattiva. Risulta, infatti, possibile importare facilmente le domande e le risposte da file o database esistenti, oppure utilizzare direttamente l'interfaccia utente per inserirle manualmente.

- *Personalizzazione delle risposte:* è possibile personalizzare le risposte generate da QnA Maker aggiungendo parole chiave, contesto o formattazione specifica. Ciò consente al servizio una migliore comprensione, e quindi la possibilità di fornire risposte più pertinenti e di alta qualità agli utenti.

- *Monitoraggio e ottimizzazione delle prestazioni:* QnA Maker offre strumenti per il monitoraggio delle metriche di utilizzo e la raccolta di feedback degli utenti. Grazie a tali

strumenti è possibile ottimizzare la base di conoscenza e migliorare le prestazioni del sistema di domande e risposte nel corso del tempo.

- *Integrazione con altri servizi di Azure:* QnA Maker si integra senza problemi con altri servizi di Intelligenza Artificiale di Azure, come Azure Cognitive Services e Bot Framework. Si possono sfruttare funzionalità avanzate, come il riconoscimento di immagini o la traduzione automatica, per arricchire le risposte.
- *Elaborazione avanzata del linguaggio naturale:* QnA Maker utilizza l'elaborazione del linguaggio naturale (NLP) compresa in Azure per comprendere e interpretare le domande degli utenti in modo accurato. Esso riconosce variazioni di parole, sinonimi e struttura delle frasi per fornire risposte pertinenti.

Al contrario, tra gli svantaggi incontrati, secondo noi è bene inserire i seguenti:

- *Dipendenza dall'ecosistema di Azure:* l'utilizzo di QnA Maker implica una dipendenza dall'ecosistema di servizi Azure di Microsoft. Sebbene ciò possa offrire vantaggi come l'integrazione con altri servizi, allo stesso tempo potrebbe rappresentare una limitazione se si preferisce utilizzare altre piattaforme o provider cloud per lo sviluppo della propria applicazione.
- *Richiede un'accurata creazione della base di conoscenza:* affinché risulti possibile ottenere risultati ottimali, è importante investire tempo e sforzo nella creazione e nell'aggiornamento accurato della base di conoscenza (knowledge base). Ciò può richiedere una fase di inizializzazione più lunga e un costante monitoraggio e aggiornamento nel tempo, così da garantire la massima puntualità ed efficienza.
- *Limitazioni delle funzionalità gratuite:* l'opzione gratuita di QnA Maker ha alcune limitazioni, come il numero di transazioni e le dimensioni della base di conoscenza consentite. Per utilizzi più intensivi, avanzati e professionali potrebbe essere necessario passare ad un piano a pagamento.
- *Dipendenza dalla qualità delle domande degli utenti:* l'efficacia di QnA Maker dipende anche dalla qualità delle domande poste dagli utenti. Se le domande sono vaghe o poco chiare, potrebbe risultare difficile fornire risposte accurate e pertinenti. Di questo abbiamo trattato poco fa; infatti vi è una funzionalità incorporata in QnA Maker che consente ad esso di fornire suggerimenti sulla corretta impostazione della knowledge base o sugli input forniti.

In definitiva, QnA Maker di Microsoft Azure offre un insieme di funzionalità avanzate per la creazione di sistemi di domande e risposte intelligenti. Tuttavia, come con qualsiasi servizio, ci sono considerazioni su cui riflettere, come la dipendenza dall'ecosistema di Azure e la necessità di un'adeguata creazione della base di conoscenza.

## 5.5 Confronto critico tra i tre sistemi di chatbot

Quando si tratta di creare chatbot e sistemi di domande e risposte intelligenti, tre dei principali servizi disponibili sono QnA Maker, Amazon Lex e Dialogflow. Questi servizi offrono funzionalità avanzate per l'elaborazione del linguaggio naturale e la creazione di interazioni conversazionali.

QnA Maker di Microsoft Azure è un servizio che semplifica la creazione di una base di conoscenza interattiva basata su domande e risposte. Utilizza l'elaborazione del linguaggio

naturale di Azure per comprendere le domande degli utenti e fornire risposte adeguate. È possibile importare i dati da file o database esistenti o inserirli manualmente. QnA Maker offre, anche, la possibilità di personalizzare le risposte aggiungendo parole chiave, contesto o formattazione specifica. È possibile integrare le conoscenze create con QnA Maker in una varietà di canali di comunicazione.

Amazon Lex di AWS è un servizio che consente di creare modelli di interazione conversazionale personalizzati. Utilizza l'elaborazione del linguaggio naturale di AWS per comprendere e interpretare le richieste degli utenti. Amazon Lex offre la possibilità di personalizzare le risposte attraverso la logica di conversazione personalizzata e le regole definite nell'applicazione. Esso supporta anche l'integrazione con una vasta gamma di servizi AWS e canali di comunicazione popolari. Amazon Lex è flessibile e consente di creare chatbot con funzionalità avanzate in base alle specifiche esigenze del progetto.

Dialogflow di Google Cloud è un servizio che offre un'elaborazione del linguaggio naturale avanzata per la creazione di agenti conversazionali. Utilizza l'elaborazione del linguaggio naturale di Google per comprendere e riconoscere le intenzioni degli utenti. Con Dialogflow è possibile creare agenti conversazionali con una vasta gamma di intenti e addestrarli con esempi di dialogo. Esso offre anche un'ampia personalizzazione delle risposte tramite la definizione di intenti, azioni e risposte specifiche. Dialogflow supporta numerose integrazioni di canali di comunicazione e piattaforme di messaggistica, inclusi Google Assistant, Facebook Messenger, Slack e altri.

Amazon Lex è molto flessibile, dunque adeguato per scopi che richiedono una certa duttilità nell'ambito di utilizzo del chatbot. Viceversa QnA Maker e Dialogflow hanno caratteristiche differenti; infatti Dialogflow gode di una certa professionalità, consentendo l'inserimento di numerosi intenti e la loro organizzazione in maniera gerarchica. Da non sottovalutare QnA Maker che, grazie alla sua semplice interfaccia, è in grado di fornire soluzioni ottimali per qualsiasi tipologia di utenti, dai meno esperti ai più ferrati in ambito informatico.

In definitiva, sia QnA Maker, sia Amazon Lex che Dialogflow offrono funzionalità avanzate per la creazione di chatbot e sistemi di domande e risposte intelligenti. La scelta tra i tre dipende dalle specifiche esigenze del progetto, dalle preferenze personali e dall'integrazione desiderata con l'ecosistema di servizi cloud di Azure, AWS o Google Cloud. Tutti e tre i servizi offrono funzionalità di elaborazione del linguaggio naturale e personalizzazione delle risposte, ma possono differire leggermente in termini di caratteristiche specifiche e integrazioni di canali.

## 5.6 Salesforce

Salesforce è una delle principali piattaforme cloud per la gestione delle relazioni con i clienti *(CRM, Customer Relationship Management)* disponibile sul mercato. È stato fondato nel 1999 e offre una vasta gamma di soluzioni per aiutare le aziende a gestire le loro interazioni con i clienti, le vendite, il marketing e il servizio clienti.

Salesforce fornisce un sistema CRM completo che consente alle aziende di gestire le relazioni con i clienti in tutti i punti di contatto. La piattaforma offre funzionalità per la gestione delle vendite, il marketing, il servizio clienti e la gestione delle operazioni. Inoltre, esso offre un'ampia personalizzazione per adattarsi alle esigenze specifiche di un'azienda. Le aziende possono configurare e personalizzare i moduli, i campi, i flussi di lavoro e le regole di business per adattare la piattaforma alle loro operazioni e di processi aziendali unici.

Salesforce fornisce strumenti di automazione delle vendite e del marketing che consentono alle aziende di gestire le opportunità di vendita, automatizzare le campagne di marketing, creare campagne e analizzare i risultati. Salesforce fornisce funzionalità di gestione del

servizio clienti, tra cui la gestione dei casi, la creazione di ticket di supporto, la gestione delle richieste dei clienti e la creazione di una base di conoscenza a risposta rapida.

Salesforce dispone di un ampio ecosistema di app di terze parti chiamato AppExchange. Le aziende possono facilmente integrare altre applicazioni e servizi nella piattaforma Salesforce per arricchire le funzionalità esistenti e soddisfare le esigenze specifiche del settore o dell'azienda. È una piattaforma basata su cloud, il che significa che le aziende possono accedere alle informazioni e lavorare ovunque ci sia una connessione Internet.

Salesforce prende molto seriamente la sicurezza dei dati. La piattaforma offre misure di sicurezza avanzate, come crittografia dei dati, autenticazione a due fattori e controlli di accesso per proteggere le informazioni dei clienti ed è ampiamente utilizzata da aziende di tutte le dimensioni e in tutti i settori per migliorare l'efficienza operativa, semplificare le vendite e offrire una migliore esperienza di servizio al cliente. La piattaforma è costantemente aggiornata e sviluppata per fornire nuove funzionalità e soluzioni per soddisfare le mutevoli esigenze aziendali.

### 5.6.1 Il chatbot di Salesforce

Salesforce fornisce un servizio di chatbot chiamato *"Einstein Bot"*. Esso rappresenta una soluzione di Intelligenza Artificiale basata su Salesforce che consente alle aziende di creare e implementare chatbot personalizzati per le loro esigenze di assistenza clienti. Questi chatbot possono essere integrati nelle piattaforme Salesforce, come Service Cloud e Sales Cloud, consentendo alle aziende di automatizzare le interazioni con i clienti e fornire supporto 24/7.

Gli Einstein Bot utilizzano l'Intelligenza Artificiale per comprendere e rispondere alle domande dei clienti in modo autonomo. Possono essere addestrati per gestire una vasta gamma di scenari e possono essere personalizzati per adattarsi alle esigenze specifiche dell'azienda. I bot possono interagire con i clienti attraverso canali come chat online, messaggi di testo e social media.

Utilizzando Einstein Bot, le aziende possono migliorare l'efficienza del servizio clienti, ridurre i tempi di risposta e offrire ai clienti un'esperienza self-service. I dati raccolti durante le interazioni con i chatbot possono anche essere utilizzati per analisi e miglioramenti futuri.

Gli Einstein Bot utilizzano algoritmi di Intelligenza Artificiale per comprendere e interpretare il linguaggio naturale dei clienti. Possono rispondere alle domande dei clienti, fornire informazioni, risolvere problemi comuni e avviare processi automatizzati all'interno del sistema Salesforce.

Questi bot possono essere integrati con varie soluzioni Salesforce, come Service Cloud e Sales Cloud, consentendo alle aziende di automatizzare le interazioni con i clienti attraverso vari punti di contatto, come siti Web, app mobili, messaggi di testo e canali di social media.

Essi possono essere addestrati utilizzando modelli di apprendimento automatico per migliorare la loro reattività nel tempo. Inoltre, si possono addestrare su set di dati specifici per fornire risposte personalizzate in base alle esigenze aziendali.

Gli Einstein Bot offrono, anche, un routing intelligente, nel senso che riconoscono le richieste dei clienti e le instradano al dipartimento o all'agente appropriato all'interno dell'azienda.

Si possono raccogliere i dati generati dalle interazioni con questi bot e utilizzarli per un'analisi approfondita dell'esperienza del cliente e per il miglioramento continuo delle prestazioni dei bot.

In sintesi, gli Einstein Bot sono chatbot di Intelligenza Artificiale di Salesforce progettati per migliorare l'efficienza del servizio clienti, fornire supporto 24 ore su 24, 7 giorni su 7, ridurre i tempi di risposta e offrire ai clienti un'esperienza self-service.

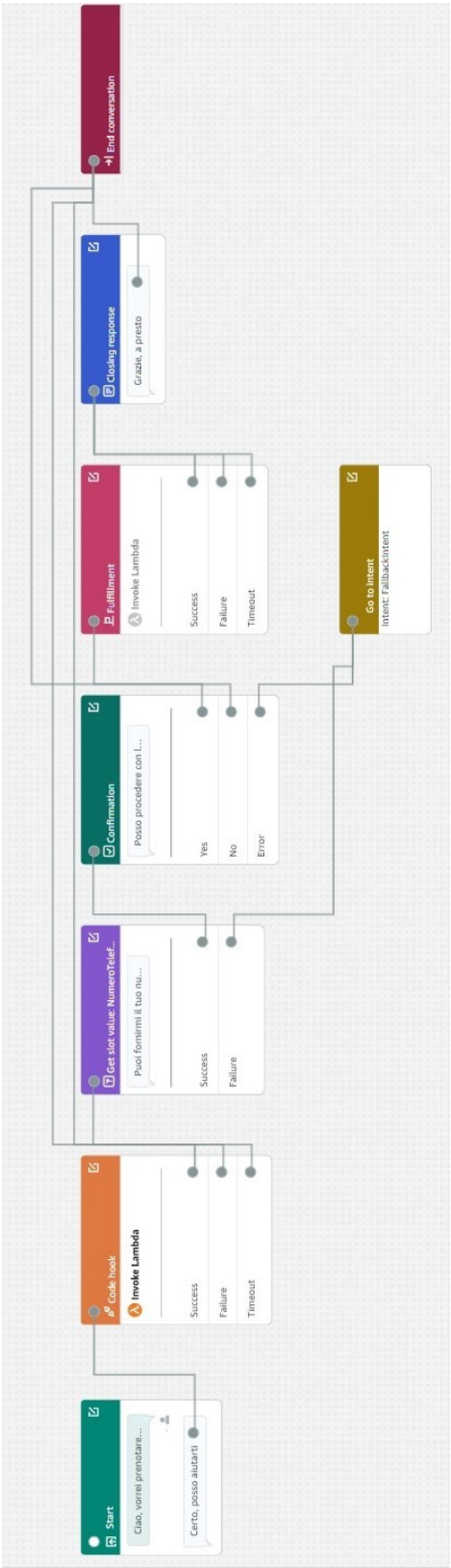


**Figura 5.2:** Workflow generale dell'intento creato

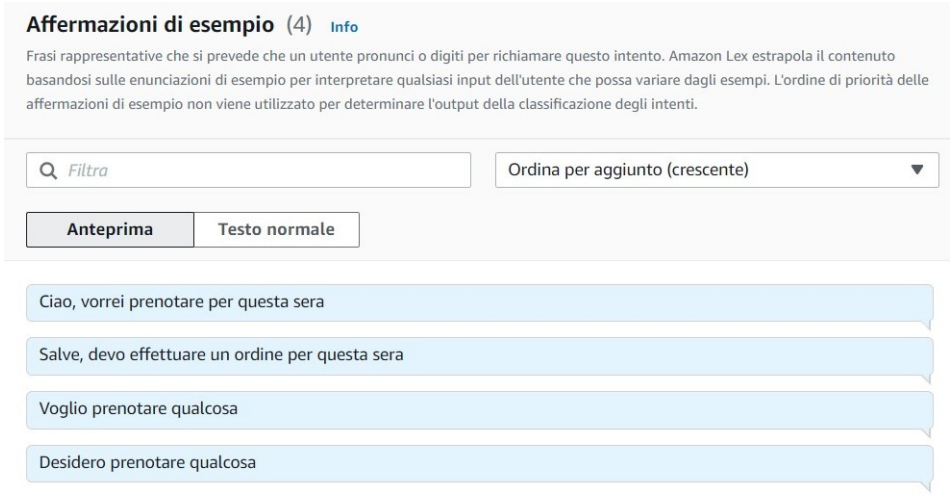


**Figura 5.3:** Affermazione di esempio

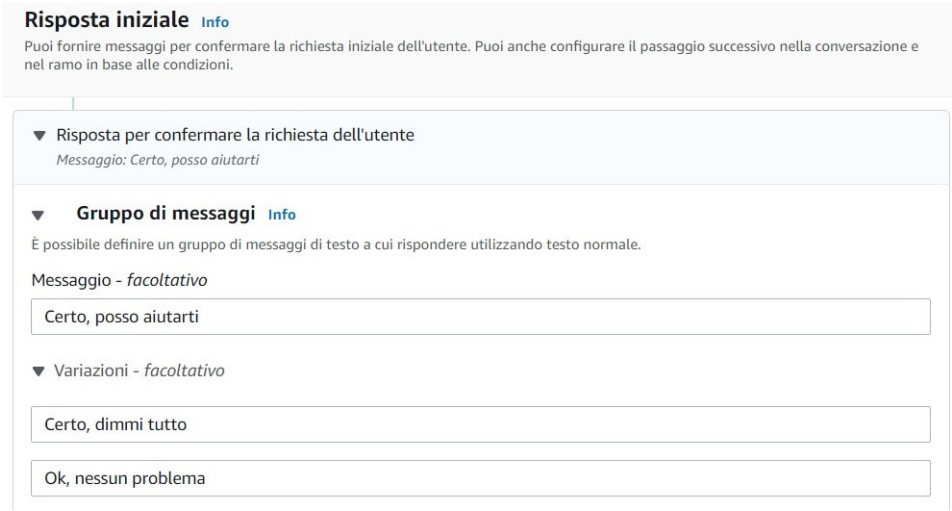


**Figura 5.4:** Risposta iniziale

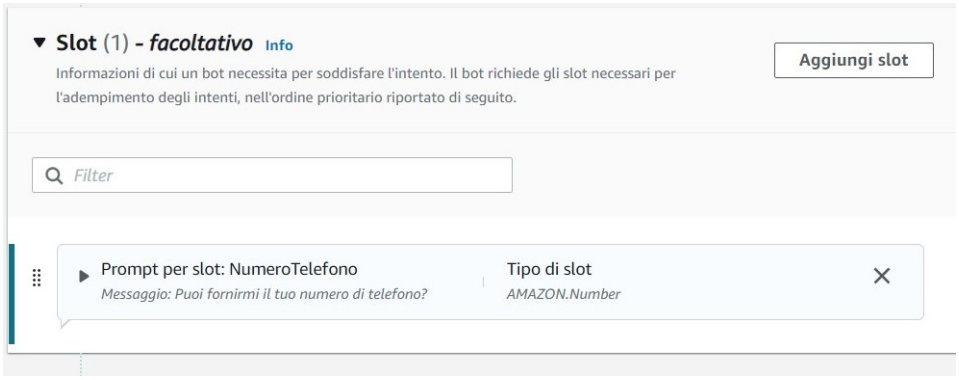


**Figura 5.5:** Prompt per l'inserimento del numero di telefono


**Figura 5.6:** Conferma del prompt sul numero di telefono

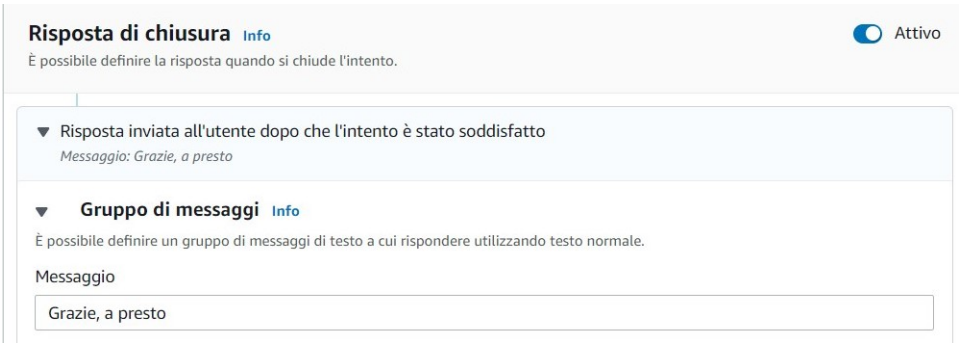


**Figura 5.7:** Risposta di chiusura


**Figura 5.8:** Risposta per errori o fallimenti


**Figura 5.9:** Esempio di intento sulle previsioni meteo

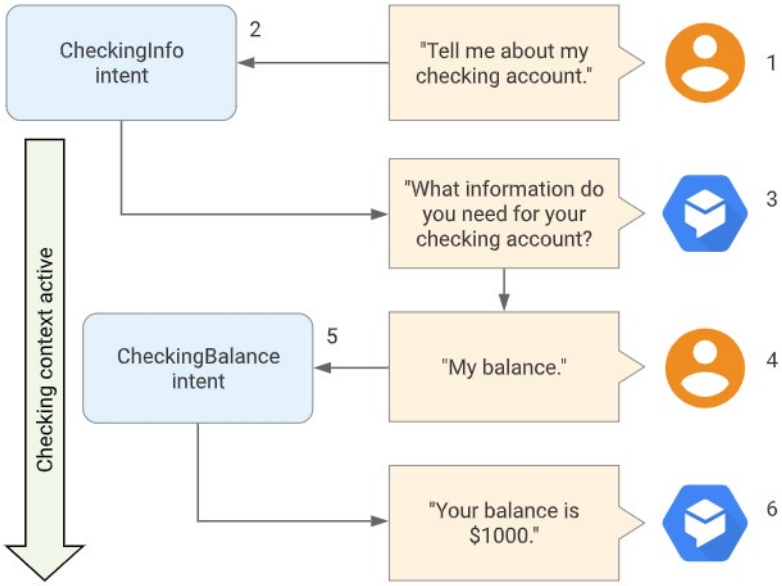


**Figura 5.10:** Esempio sull'utilizzo di un contesto per un agente bancario

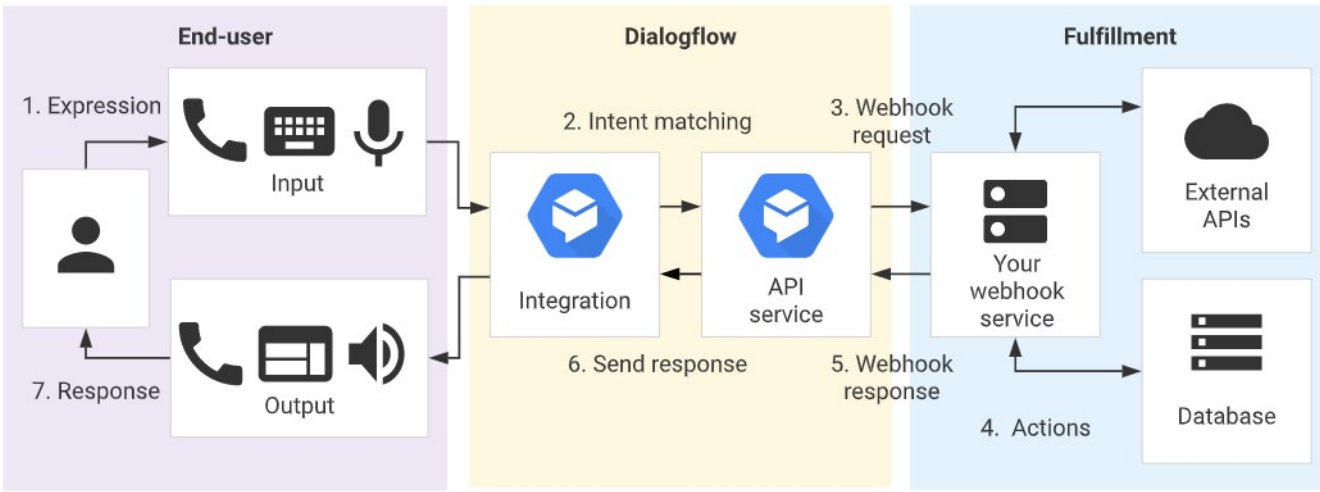


**Figura 5.11:** Esempio di utilizzo per l'evasione di ordini

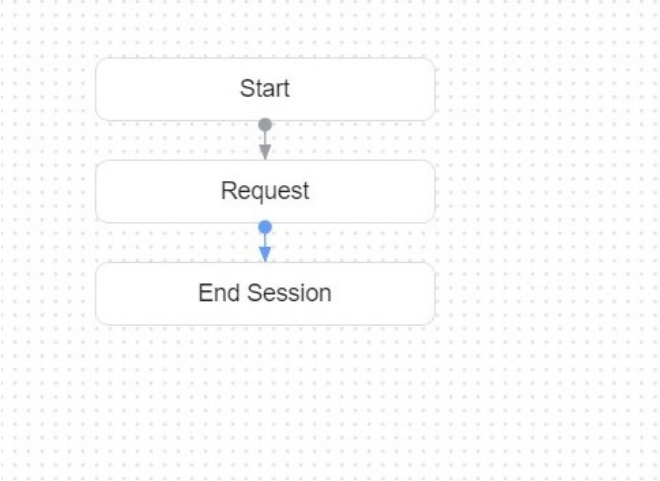


**Figura 5.12:** Workflow generale

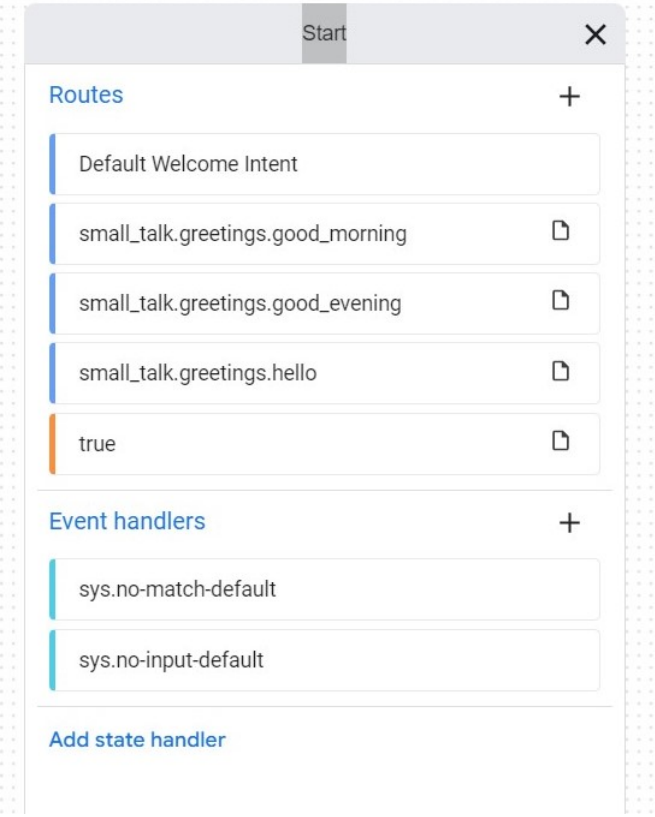


**Figura 5.13:** Visione completa "Start"

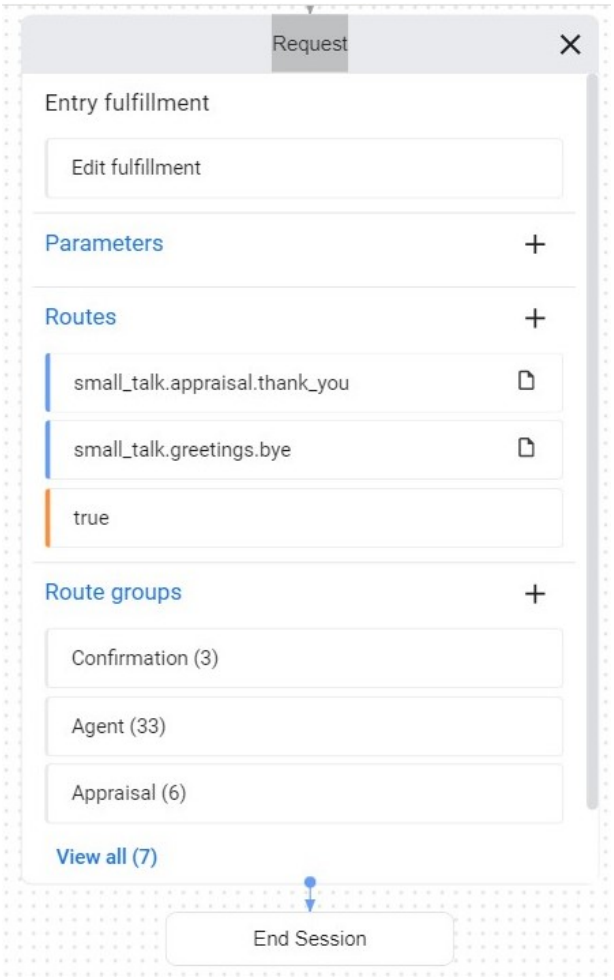


**Figura 5.14:** Visione completa "Request"

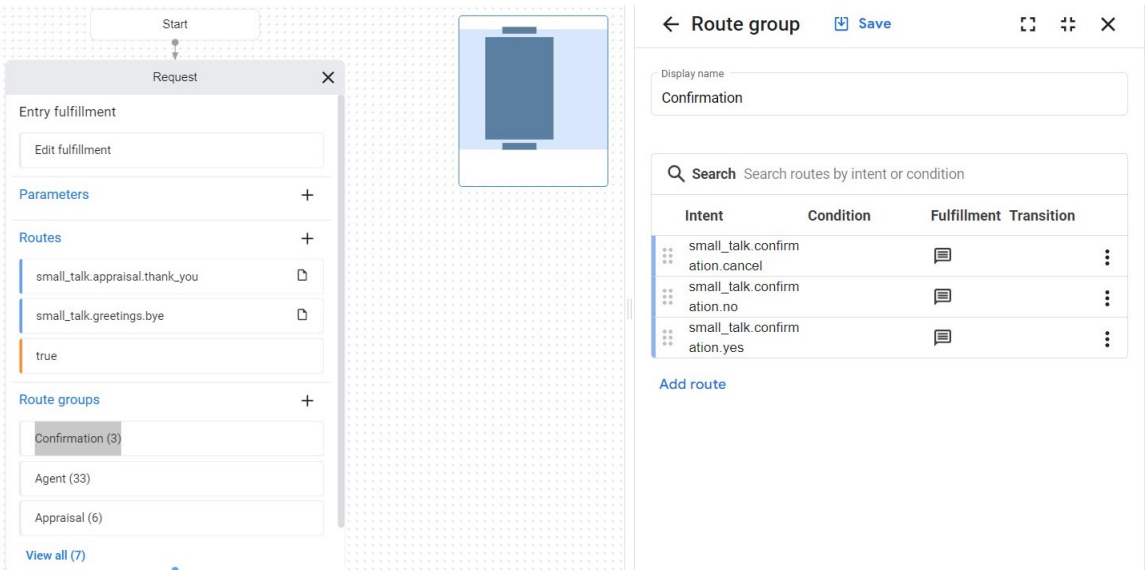


**Figura 5.15:** Visualizzazione degli intent definiti per "request"

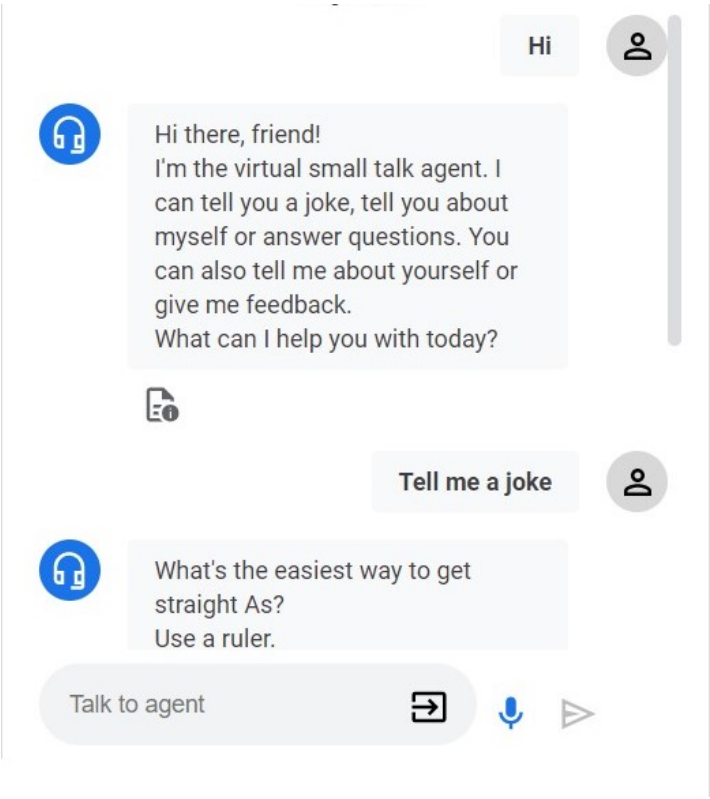


**Figura 5.16:** Esempio di utilizzo del chatbot

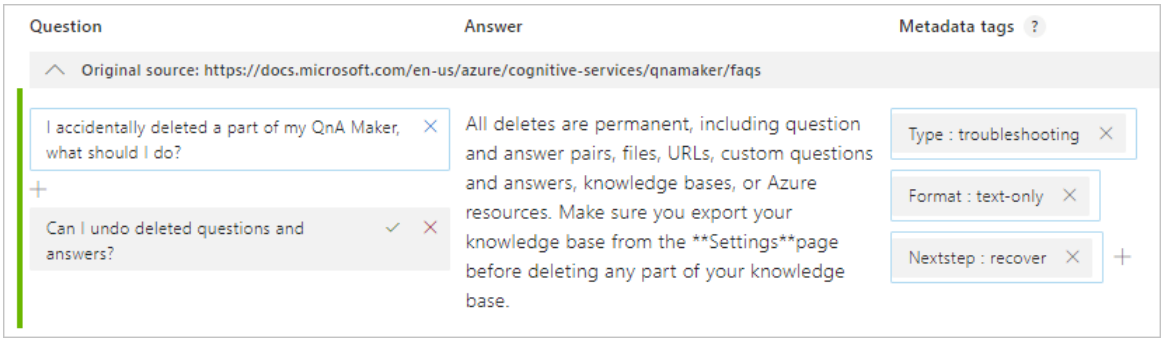


**Figura 5.17:** Esempio di knowledge base

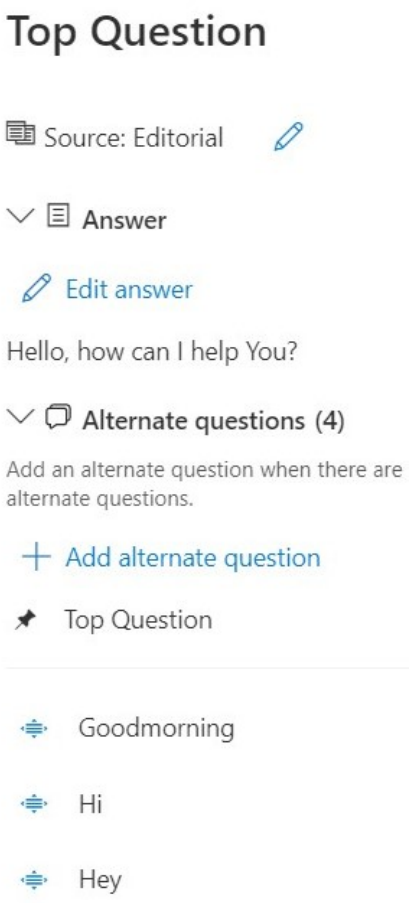


**Figura 5.18:** Esempio di avvio di QnA Maker

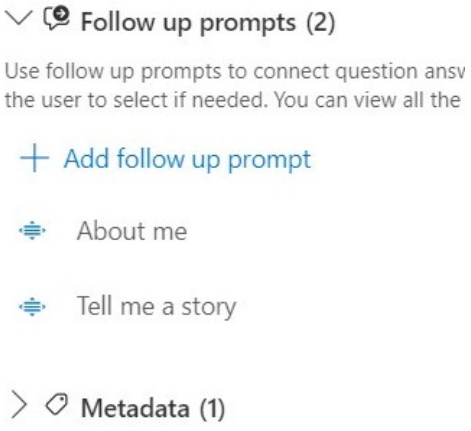


**Figura 5.19:** Esempio di risposte fornite QnA Maker

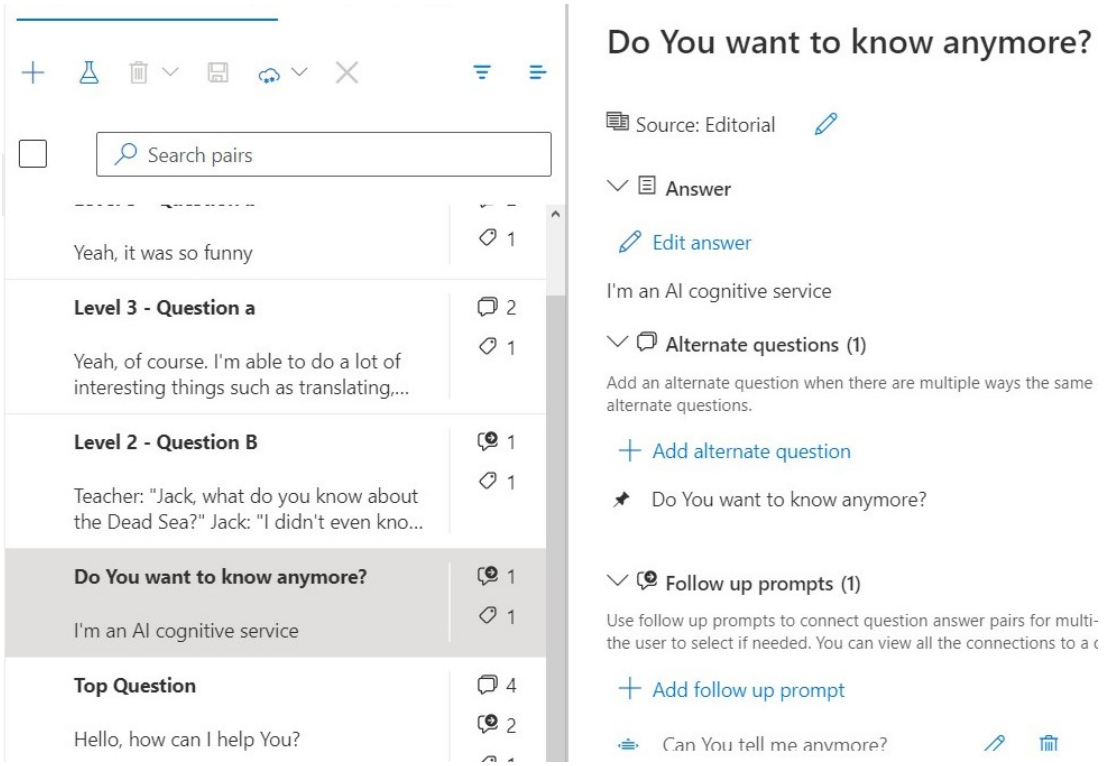


**Figura 5.20:** Esempio di risposte aggiuntive QnA Maker

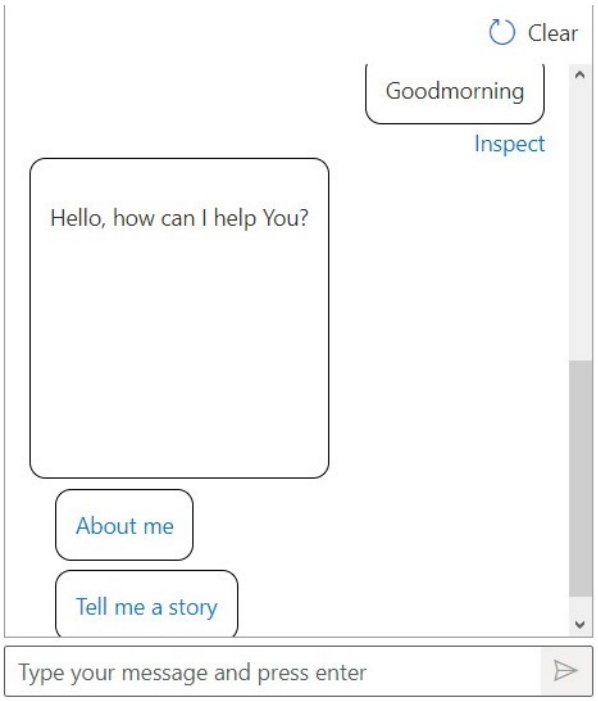


**Figura 5.21:** Messa in atto del servizio con le domande-risposte fornite

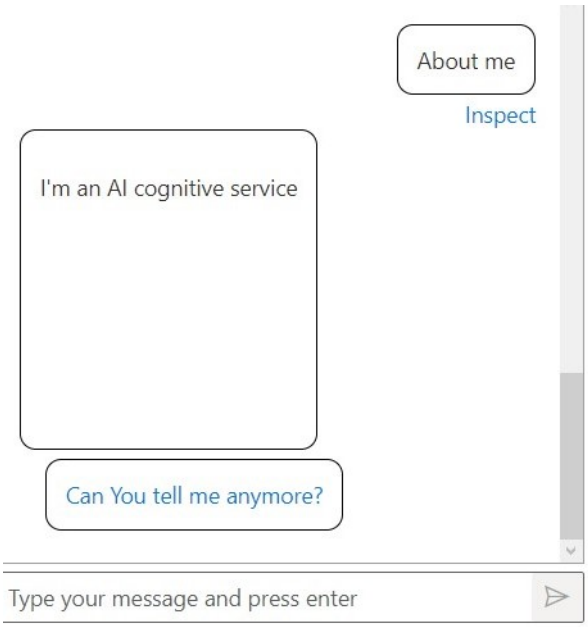


**Figura 5.22:** Messa in atto del servizio con le domande-risposte fornite

CAPITOLO 6

# Conclusioni

*L'epoca dell'informazione digitale in cui viviamo è in continua evoluzione; le organizzazioni stanno costantemente cercando nuovi modi per migliorare l'efficienza operativa, ridurre i costi e offrire un'esperienza utente sempre più avanzata. Due concetti che si sono affermati come pilastri fondamentali di questa trasformazione digitale sono il Cloud Computing e le soluzioni on-premise. In questo capitolo esploreremo le differenze tra queste due modalità di distribuzione dei servizi informatici, mettendo in luce le loro caratteristiche distintive e i vantaggi che offrono. Inoltre, daremo uno sguardo al futuro dei chatbot, una tecnologia emergente che sta rivoluzionando l'interazione tra le persone e le macchine.*

## 6.1 Discussione

Prima di esprimere opinioni sul lavoro svolto e sui vari servizi analizzati, poniamo l'attenzione sulla tecnologia alla base di questi prodotti, ovvero il Cloud Computing e la sua controparte, il modello *on-premise*.

Un data center on-premise è un gruppo di server detenuti e controllati privatamente. Il Cloud Computing tradizionale, invece, prevede il noleggio di risorse di data center da un service provider esterno.

Il concetto di "on-premise" fa riferimento ai data center privati che le aziende posseggono e gestiscono internamente presso la propria sede. Tali infrastrutture on-premise possono essere impiegate per creare Cloud privati, in cui le risorse di calcolo sono virtualizzate in modo simile a quanto avviene nei Cloud pubblici.

Quando si tratta di Cloud Computing, di solito si fa riferimento ai cloud "pubblici", o "tradizionali", un modello in cui un fornitore di servizi esterno mette a disposizione le risorse di elaborazione in base alle esigenze dell'acquirente. Il Cloud pubblico è un ambiente condiviso da più utenti o aziende, in cui le risorse di elaborazione vengono condivise e tutti i dati sono protetti utilizzando crittografia avanzata. In Figura 6.1 sono riportati gli elementi indispensabili per organizzare un sistema basato sul Cloud, come, ad esempio, connessione di rete, persistenza dei dati, crittografia dei dati, condivisione con altri utenti, etc.

Sia i data center on-premise che il Cloud Computing offrono all'azienda l'infrastruttura IT necessaria. La scelta del modello dipende principalmente dal livello di sicurezza richiesto per adempiere agli standard di conformità e dalle preferenze riguardo alla struttura dei costi.

Come accennato in precedenza, spostare la propria infrastruttura IT nel Cloud significa "affittare" l'infrastruttura di un provider esterno. Questo offre due importanti vantaggi. In primo luogo, ci consente di pagare in base alle risorse effettivamente utilizzate. In secondo


**Figura 6.1:** Principali elementi di un sistema basato su Cloud

luogo, ci consente di aumentare (o ridurre) tali risorse in modo efficiente, in tempo reale, man mano che le esigenze degli utenti e il business crescono.

I server Cloud utilizzano la tecnologia di virtualizzazione per ospitare le applicazioni di un'azienda; quindi, è ovvio che in merito ai servizi IT, gestiti come quelli sul Cloud, il fornitore di servizi si fa carico di problemi e costi che mettono a dura prova il reparto IT, come aggiornamenti software, patch di sicurezza e manutenzione del sistema.

Un discorso simile vale per la sicurezza; chi fornisce il servizio deve garantire la sicurezza, eliminando, così, la necessità di occuparsene da parte dell'azienda o del cliente. Il Cloud elimina anche i costi associati al consumo energetico e consente di risparmiare lo spazio necessario per installare l'infrastruttura IT e tutti i servizi successivi che riguardano il benessere dei dipendenti.

Infine, non bisogna dimenticare che il Cloud fornisce un provisioning quasi istantaneo perché tutto è già configurato. Così, ad esempio, qualsiasi nuovo software integrato nell'ambiente è pronto all'uso e immediatamente disponibile, una volta attivato l'abbonamento. Ciò elimina il tempo necessario per l'installazione e la configurazione.

Che un'azienda metta le sue applicazioni nel Cloud o decida di mantenerle on-premise, in entrambi i casi la sicurezza dei dati è sempre fondamentale. Per le aziende che operano in settori altamente regolamentati (come banche, finanza o assicurazioni), o che devono mantenere importanti segreti industriali (come industrie farmaceutiche, automobilistiche o manifatturiere), la scelta di ospitare internamente la propria infrastruttura IT può essere una strategia vincente per superare la concorrenza. Questo perché sapere che i dati risiedono sui propri server può garantire maggior tranquillità e, allo stesso tempo, una migliore conformità alle normative.

Lo svantaggio di un ambiente locale è che i costi associati alla gestione e alla manutenzione dell'intera soluzione possono essere molto più elevati rispetto alle soluzioni Cloud. Infatti, le configurazioni on-premise richiedono acquisti di server, computer, apparecchiature di rete e licenze software, ma non solo. Sono, inoltre, necessarie capacità di integrazione e personale IT qualificato per affrontare i problemi che potrebbero sorgere. Per non parlare del costo della manutenzione e degli arresti che devono essere implementati quando qualcosa si rompe o non funziona correttamente.

Inoltre, se l'azienda ha bisogno di espandere l'infrastruttura IT, le nuove macchine devono essere acquistate, ordinate, consegnate, configurate e messe in funzione prima che possano essere utilizzate. In questo caso, il Cloud è decisamente superiore all'on-premise.

Il dibattito sui pro e contro dell'on-premise rispetto al Cloud computing è intenso e coinvolge diverse aziende. Tuttavia, invece di scegliere un'unica soluzione, c'è un'altra opzione, ovvero la possibilità di avere il meglio di entrambi: si tratta della infrastruttura ibrida, che è un ponte tra on-premise e Cloud che ci permette di avere parti "statiche" della infrastruttura IT on-premise, e parti "dinamiche" sul Cloud. Ciò consente di avere sempre la giusta infrastruttura IT per le esigenze più diverse, ad esempio poter utilizzare applicazioni legacy strategiche on-premise non disponibili nel Cloud, o scalare in tempo reale quando è richiesta maggiore potenza di calcolo. Nel caso di un'infrastruttura mista, la connessione WAN [1] può fare la differenza, essa, inoltre, deve essere sufficiente a supportare grandi quantità di traffico dati.

Come specificato precedentemente, non crediamo che tra Cloud e on-premise vi sia una migliore dell'altra ma, semplicemente, la modalità di utilizzo dipende dalle esigenze per le quali si è voluto attivare un sistema del genere. Appare logico che, per grandi aziende, soprattutto software house, la possibilità di avere un sistema consistente e compatto, totalmente disponibile in rete come il Cloud, permette un controllo ed una gestione migliori. Questo è stato il caso delle piattaforme Cloud che mettevano a disposizione i servizi da noi utilizzati.

Viceversa, per piccole aziende, dove magari la qualità dei prodotti sul mercato diventa il punto di forza, è naturale pensare ad un sistema basato sulla tecnologia on-premise, legato, anche, al fatto di non voler espandere i propri dati e le proprie strategie di mercato.

Infine, per quanto riguarda i servizi di AI utilizzati, possiamo assolutamente considerarli come il presente ed il futuro dell'informatica. Essi, infatti, rappresentano la base per costruire qualcosa di più grande. Ne sono un esempio i chatbot che tuttora utilizziamo. Essi, per funzionare correttamente, hanno alla loro base molti sistemi di trascrizione e traduzione (che sono quelli che abbiamo analizzato precedentemente) che li aiutano nella trascrizione di messaggi vocali e nella traduzione del parlato o di semplici messaggi. Oltre questi ci sono numerosi altri servizi, come il riconoscimento delle immagini (image recognition), che meritano di essere studiati approfonditamente e compresi al meglio, anche e soprattuto per l'utilizzo nell'ambito della medicina e di tutti quei campi legati alla vita ed alla salute dell'uomo.

## 6.2 Uno sguardo al futuro

Da un recente studio commissionato a *451 Research* da *Oracle Cloud Infrastructure*, e che ha coinvolto più di duemila aziende, risulta che la stragrande maggioranza di queste sta adottando, o si adopererà a breve per farlo, il multicloud [2].

Negli ultimi anni, il cloud computing è diventato un pilastro fondamentale del settore IT, grazie alla crescente richiesta, da parte delle aziende, di maggiore flessibilità ed efficienza operativa.

Nonostante l'interesse verso le risorse Cloud esista già da tempo, più del 90% dei partecipanti all'indagine concordano sul fatto che la pandemia abbia ulteriormente stimolato l'interesse e gli investimenti per questa tecnologia.

[1] Una rete geografica, o Wide Area Network, è una rete di telecomunicazioni che si estende su una grande distanza geografica per lo scopo principale della rete di computer. Per definizione, la WAN è una rete che attraversa regioni, paesi, o addirittura il mondo.

[2] Il multi-cloud è un modello di cloud computing in cui un'organizzazione utilizza una combinazione di cloud, che possono essere due o più public cloud, due o più private cloud o una combinazione di public cloud, private cloud ed edge cloud, per distribuire applicazioni e servizi.

Con la crescita del lavoro da remoto e la necessità di collaborare con nuovi partner e fornitori, molte organizzazioni hanno optato per la strategia multicloud per ottenere la flessibilità e la scalabilità di cui avevano bisogno in questo nuovo contesto.

D'altronde, come viene evidenziato nella ricerca portata a termine da questa compagnia, il 98% delle aziende intervistate utilizza, o prevede di utilizzare, almeno due fornitori di infrastrutture Cloud e il 31% ne utilizza quattro o più.

La scelta nell'adoperare questo tipo di tecnologia multicloud, che è considerata il futuro della sua tecnologia madre, il Cloud, è spinta principalmente da considerazioni che riguardano la sovranità dei dati e l'ottimizzazione dei costi. Altri fattori importanti sono l'agilità e l'innovazione aziendale, l'accesso ai servizi e applicazioni di qualità più elevata e preoccupazioni per il rischio di lock-in con specifici fornitori.

Le strategie multicloud offrono alle aziende maggiore controllo e flessibilità su come e dove i dati vengono gestiti e utilizzati. In futuro, i dipartimenti IT prevedono di utilizzare cloud multipli per ridondanza dei dati (54%), mobilità dei dati (49%), ottimizzazione dei costi tra cloud pubblici (42%), mitigazione dei rischi per l'ambiente IT (40%) ed espansione geografica (38%).

Per completare la loro conversione nell'era digitale, le aziende si avvalgono di molteplici piattaforme cloud principalmente per i seguenti motivi:

- *Accelerare la trasformazione delle app e la distribuzione di nuove app.*
- *Evitare la dipendenza da un singolo vendor e garantire la sovranità dell'azienda:* i costi richiesti per la gestione di questi servizi Cloud, la sovranità dei dati e la dipendenza da un singolo provider incutono sempre maggior preoccupazione. Questo è uno dei principali motivi per cui si preferisce distribuire le risorse su più ambienti.
- *Distribuire applicazioni e servizi sull'edge:* questo prevalentemente nei settori come logistica e vendita al dettaglio, nei quali l'esperienza del cliente richiede che le applicazioni siano distribuite sull'edge, più vicine ai dispositivi e agli utenti fisici.
- *Supportare l'aumento della forza lavoro distribuita:* questa è la nuova sfida per le aziende. Garantire ai dipendenti una postazione stabile per lavorare, ovunque essi si trovino, è il punto di forza delle nuove aziende tecnologiche.

Ovviamente questi Cloud sono solo la base per l'utilizzo e la fruizione di ulteriori servizi come i chatbot. Vediamo dunque qual è e quale sarà l'evoluzione di questi chatbot.

Come risulta da un recente articolo di *indigo.ai* (startup del Gruppo Vedrai dal 2022, la cui missione è aiutare le aziende più innovative a evolvere la propria customer experience grazie all'Intelligenza Artificiale conversazionale), l'86% degli italiani continua a preferire l'interazione con un operatore umano a quella condotta tramite chatbot. Ma le tecnologie di nuova generazione, basate sull'Intelligenza Artificiale generativa, potrebbero ribaltare la situazione.

I chatbot in Italia non sono ancora apprezzatissimi, ed i consumatori preferiscono parlare ancora con un interlocutore umano; ciò è dovuto alla scarsa fiducia che essi ripongono nei chatbot. Questa sfiducia, aumenta nel caso di vendita di beni o servizi, nella quale l'utente potrebbe aver bisogno di assistenza specifica, e quindi di rivolgere al customer care dell'azienda domande complesse.

Il parere degli utenti, però, potrebbe cambiare con l'introduzione di chatbot intelligenti che utilizzano l'IA generativa [3].

[3] L'IA generativa è un algoritmo di IA che genera nuovi output in base ai dati su cui è stato addestrato. Se, infatti, un sistema di Intelligenza Artificiale è progettato per riconoscere modelli e fare previsioni, l'IA generativa crea nuovi contenuti sotto forma di immagini, testo, audio e altro ancora.

La differenza dalla normale IA è che l'IA generativa utilizza molta più potenza di elaborazione, ed è per questo che è incredibilmente più costosa (ciò limita anche il numero di soggetti che tentano di svilupparla, numero sempre più ristretto attorno ad attori miliardari del mercato come la nota OpenAI).

Altra caratteristica che la contraddistingue dalla normale IA è il tipo di output che è in grado di fornire come risultato: non solo testi o messaggi, ma anche immagini, audio e video (tra i più noti in questo settore vi è *Midjourney*). Ovviamenti, i modelli di Intelligenza Artificiale generativa sono diversi. In Figura 6.2 è riportato un esempio di cosa è in grado di creare l'IA generativa semplicemente chiedendoglielo a parole.

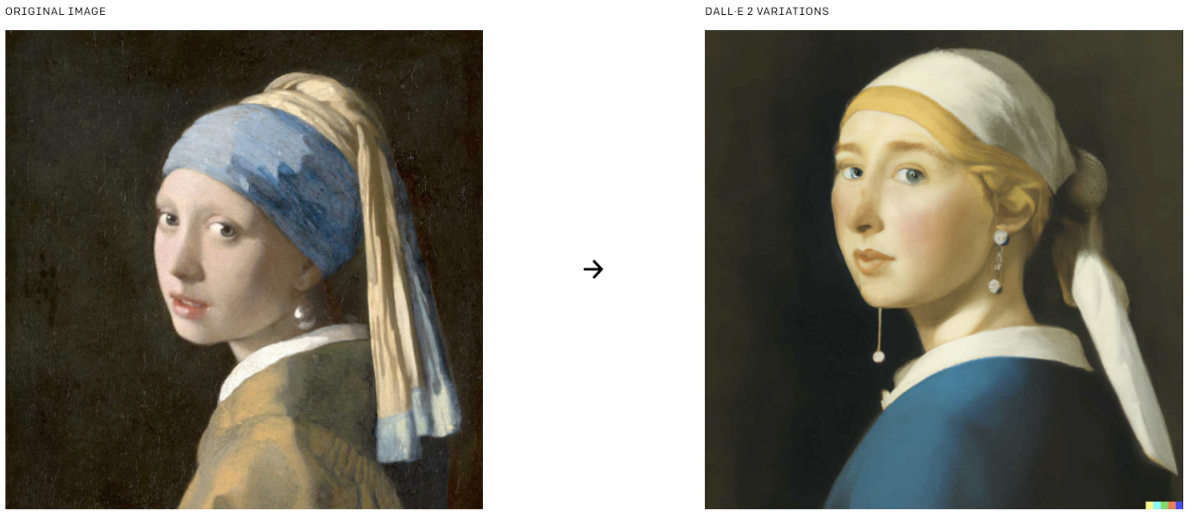


**Figura 6.2:** Esempio di cosa è in grado di creare l'IA generativa

Le *GAN (Generative Adversarial Networks)* sono composte da due modelli di Machine Learning che vengono addestrati contemporaneamente. Uno è chiamato generatore e l'altro è chiamato discriminatore. Il compito del generatore è creare nuovi output che assomigliano ai dati di addestramento, mentre il discriminatore deve valutare i dati generati e fornire un feedback al generatore per migliorarne l'output.

I Transformer, come ChatGPT e GPT-3.5 di OpenAI, sono reti neurali progettate per l'elaborazione del linguaggio naturale. Sono addestrati su grandi quantità di dati per apprendere le relazioni tra dati sequenziali, come parole e frasi, al fine di utilizzare per le attività di generazione del testo.

Ecco, dunque, come cambierà l'orizzonte dell'Intelligenza Artificiale, con questi sistemi che diventeranno sempre più intelligenti. Nonostante ciò, una delle maggiori cause che porta gli utenti a non fidarsi, è il timore di non essere capiti o di ricevere risposte inaccurate.

Non c'è da stupirsi che, secondo l'analisi di indigo.ai, la fiducia e il gradimento nei confronti dei chatbot diminuiscono con l'avanzare dell'età degli intervistati. Ne sono certi soprattuto i giovani, particolarmente fiduciosi nei confronti delle potenzialità messe a disposizioni da strumenti come ChatGPT, sempre più parte del quotidiano. Infatti, dallo stesso articolo, emerge che nel 55% dei giovani italiani, l'arrivo di ChatGPT ha migliorato la loro percezione nei confronti dei chatbot.

Dobbiamo essere consapevoli, dunque, che queste tipologie di servizi faranno sempre più parte della nostra vita in futuro, ed è per questo che crediamo sia indispensabile fornire documentazioni e conoscenze appropriate in modo tale che tutti possano usufruirne al meglio.

# Bibliografia


AGRAWAL, A. (2023), *No-Code Artificial Intelligence: The new way to build AI powered applications (English Edition)*, bpb.

AJAY AGRAWAL, A. G., JOSHUA GANS (2022), *Power and Prediction: The Disruptive Economics of Artificial Intelligence*, Harvard Business School Pr.

ALESSANDRO LONGO, G. S. (2020), *Intelligenza artificiale. L'impatto sulle nostre vite, diritti e libertà*, Mondadori Università.

ALTOBELLO, G. (2020), «Cos'è Bert, l'algoritmo che cambia il mondo del Natural Language Processing», *ai4business.it – Cos'è Bert, l'algoritmo che cambia il mondo del Natural Language Processing*.

BERNSTEIN, P. (2022), *Machine Learning: Architecture in the age of Artificial Intelligence*, RIBA Publishing.

CAFFARATTI, R. (2023), «Chat GPT: cos'è e come funziona la nuova frontiera dell'IA», *onlinesim.it*.

CAREW, J. M. (2019), «Reinforcement Learning», *techtarget.com*.

CHAILLOU, S. (2022), *Artificial Intelligence and Architecture: From Research to Practice*, Birkhauser Architecture.

COURSERA (2023), «Deep Learning vs. Machine Learning: Beginner's Guide», *coursera.org*.

DAVID L. POOLE, A. K. M. (2017), *Artificial Intelligence: Foundations of Computational Agents*, Cambridge University Press.

DR. JAGREET KAUR, N. S. G. (2019), *Artificial Intelligence and Deep Learning for Decision Makers: A Growth Hacker's Guide to Cutting Edge Technologies (English Edition)*, bpb.

FOCUS (2014), «Innovazione Che cos'è il cognitive computing?», *focus.it*.

INNOVATION, R. O. D. (2023a), «Computer Vision: definizione, funzionamento e applicazioni», *osservatori.net*.

INNOVATION, R. O. D. (2023b), «Natural Language Processing (NLP): come funziona l'elaborazione del linguaggio naturale», *osservatori.net*.

LUCA MASSARON (AUTORE), A. V. T., JOHN PAUL MUELLER (AUTORE) (2020), *Intelligenza artificiale for dummies*, Hoepli.

MARGARET BODEN (AUTORE), F. C. T., DIEGO MARCONI (A CURA DI) (2019), *L'intelligenza artificiale*, Il Mulino.

MARUZZELLA, G. (2023), «Chatbot, addio vecchia generazione, la strada per il futuro è l'AI generativa», *indigo.ai – Chatbot, addio vecchia generazione, la strada per il futuro è l'AI generativa*.

PETER NORVIG, S. R. (2021), *Artificial Intelligence: A Modern Approach, Global Edition*, Pearson.

STEFANO QUINTARELLI, P. A. (2020), *Intelligenza artificiale. Cos'è davvero, come funziona, che effetti avrà*, Bollati Boringhieri.

STUART J. RUSSELL, P. N. (2021), *Intelligenza artificiale. Un approccio moderno. Ediz. MyLab (Vol. 1)*, Pearson.

STUART J. RUSSELL, P. N. (2022), *Intelligenza artificiale. Un approccio moderno. Ediz. MyLab (Vol. 2)*, Pearson.

TAULLI, T. (2019), *Artificial Intelligence Basics: A Non-Technical Introduction*, Apress.

WEBER, H. (2019), *Artificial Intelligence and Life: A Complete Guide to the Basic Concepts in AI, Neural Networks, Machine Learning and Data Science*.

## Siti Web consultati

- DEFENSIS – `https://www.defensis.it/servizi/intelligenza_artificiale__introduzione.htm`
- Wikipedia – `https://it.wikipedia.org/wiki/BERT`
- Wikipedia – `https://it.wikipedia.org/wiki/Scheda_perforata`
- Wikipedia – `https://it.wikipedia.org/wiki/Test_di_Turing`
- IGA, International Geothermal Association – `www.geothermal-energy.org`
- Intelligenza artificiale italia – `https://www.intelligenzaartificialeitalia.net/post/cos-ÃÍ-la-previsione-delle-serie-temporali-o-time-series-forecasting`
- Wikipedia – `https://it.wikipedia.org/wiki/Rete_neurale_ricorrente`
- Wikipedia – `https://it.wikipedia.org/wiki/Question_answering`
- Wikipedia – `https://it.wikipedia.org/wiki/Cloud_computing`
- Wikipedia – `https://it.wikipedia.org/wiki/Application_programming_interface`

- IBM – https://www.ibm.com/common/ssi/ShowDoc.wss?docURL=/common/ssi/rep_sm/9/897/ENUS5737-B19/index.html
- IBM – https://researcher.watson.ibm.com/researcher/view_group.php?id=9809
- IBM – https://www.ibm.com/watson-health/about/phytel
- IBM – https://www.ibm.com/common/ssi/cgi-bin/ssialias?appname=skmwww&htmlfid=897%2FENUS5725-W51&infotype=DD&subtype=SM&mhsrc=ibmsearch_a&mhq=IBM%20WATSON%20ONcology
- IBM – https://www.ibm.com/docs/en/watson-care-manager?topic=SSRMV7/com.ibm.iwcm.doc/product_overview/c_cp_product_overview.html
- IBM – https://www.ibm.com/topics/supervised-learning#:textasciitilde:text=the%20next%20step-,What%20is%20supervised%20learning%3F,data%20\or%20predict%20outcomes%20accurately.
- Wikipedia – https://en.wikipedia.org/wiki/IBM_Watson#:textasciitilde:text=IBM%20Watson%20is%20a%20question,first%20CEO%2C%20industrialist%20Thomas%20J.
- Microsoft – https://learn.microsoft.com/it-it/azure/cognitive-services/qnamaker/overview/overview
- Google – https://cloud.google.com/dialogflow?hl=it
- Google – https://cloud.google.com/dialogflow/docs
- Microsoft – https://learn.microsoft.com/it-it/azure/cognitive-services/translator/translator-overview
- Amazon – https://aws.amazon.com/it/translate/
- Google – https://cloud.google.com/translate?hl=it#section-1
- Microsoft – https://azure.microsoft.com/it-it/products/cognitive-services/speech-services#layout-container-uid189e
- Amazon – https://aws.amazon.com/it/transcribe/features/?nc1=h_ls
- Google – https://cloud.google.com/speech-to-text?hl=it#section-3
- IBM – https://www.ibm.com/watson/services/document-conversion/
- Amazon – https://docs.aws.amazon.com/translate/latest/dg/what-is.html
- Amazon – https://docs.aws.amazon.com/translate/latest/dg/how-it-works.html
- Wikipedia – https://it.wikipedia.org/wiki/UTF-8
- Google – https://cloud.google.com/translate/docs/overview?hl=it

- Microsoft – https://azure.microsoft.com/it-it/products/cognitive-services/translator#content-card-list-oc803c
- Wikipedia – https://it.wikipedia.org/wiki/DeepL_Translator
- Wikipedia – https://it.wikipedia.org/wiki/Rete_neurale_convoluzionale
- Amazon – https://aws.amazon.com/it/transcribe/features/
- Amazon – https://aws.amazon.com/it/transcribe/?nc2=h_ql_prod_ml_ts
- Google – https://cloud.google.com/speech-to-text/docs/transcribe-client-libraries?hl=it
- Microsoft – https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/
- Google – https://cloud.google.com/dialogflow/es/docs/basics?hl=it
- Amazon – https://aws.amazon.com/it/lex/
- Microsoft – https://learn.microsoft.com/it-it/azure/cognitive-services/qnamaker/overview/overview
- Salesforce – https://www.salesforce.com/it/blog/2019/10/guida-uso-chatbot.html
- HewlettPackard Enterprise – https://www.hpe.com/it/it/what-is/on-premises-vs-cloud.html
- Quanture – https://blog.quanture.com/on-cloud-vs-on-premise-i-pro-e-i-contro-quando-si-parla-di-infrastruttura-it
- Wikipedia – https://it.wikipedia.org/wiki/Wide_Area_Network
- Tomshw – https://www.tomshw.it/business/mondo-lavoro-trend-2023/
- VMWare – https://www.vmware.com/it/topics/glossary/content/multi-cloud.html
- Smartworld – https://www.smartworld.it/guide/ia-generativa.html#generativa

# Ringraziamenti

Ringrazio il Prof. Ursino per l'opportunità datami, e per avermi seguito durante tutto il mio percorso di tirocinio e tesi.

Un ringraziamento speciale è rivolto alla mia famiglia, mamma, papà e mia sorella, che mi hanno costantemente sostenuto in questi tre anni, aiutandomi nei momenti più difficili, e guidandomi grazie ai loro fondamentali consigli.

Ringrazio e saluto con cuore l'altra mia grande famiglia, composta da tutti i miei nonni, gli zii, ed i cugini.

Infine, vorrei ringraziare i miei compagni di corso, Alessio, Alessandra, Edoardo, Laura, Luca, Sara, Valeria e Walter, che hanno reso questa avventura più divertente e meno stressante. A loro invio un grande in bocca al lupo per il futuro.