Introduzione

Il progetto di NER tagging multilingua rappresenta un'importante sfida nell'ambito del Natural Language Processing (NLP), in quanto richiede lo sviluppo di un sistema in grado di identificare e classificare entità di interesse in testi scritti in lingue diverse. La NER (Named Entity Recognition) è un'area cruciale del NLP che si occupa dell'individuazione e dell'etichettatura di entità come nomi di persone, luoghi, organizzazioni, date e quantità all'interno di un testo.

L'obiettivo di questo progetto è stato quello di sviluppare un sistema di NER tagging multilingua utilizzando un approccio basato sull'algoritmo di Viterbi. Il sistema è stato addestrato e valutato su tre diverse lingue: inglese, italiana e spagnola. A tal fine, sono stati utilizzati tre dataset distinti, ognuno contenente un insieme di testi annotati con le entità di interesse, suddivisi in train, validation e test set.

Contesto teorico

Prima di entrare nei dettagli del progetto, è importante comprendere il contesto teorico della NER e dell'algoritmo di Viterbi nel contesto del NLP.

La NER è una tecnica fondamentale nel NLP, utilizzata in una vasta gamma di applicazioni, tra cui l'estrazione di informazioni da testi, l'analisi del sentiment, l'elaborazione del linguaggio naturale assistita da computer e molto altro ancora. La sua importanza deriva dalla capacità di identificare e classificare entità di interesse all'interno di un testo, fornendo così informazioni utili per una serie di compiti di analisi del testo.

L'algoritmo di Viterbi è una tecnica di programmazione dinamica utilizzata per trovare la sequenza più probabile di stati nascosti in un modello Markov nascosto. Nel contesto del NER tagging, l'algoritmo di Viterbi è utilizzato per determinare la sequenza ottimale di etichette di entità per una determinata sequenza di parole, tenendo conto delle probabilità di transizione tra le etichette e delle probabilità di emissione delle parole.

Valutazione dei risultati

Nel corso del progetto, la performance del sistema di NER tagging è stata valutata considerando diverse metriche, tra cui l'accuratezza complessiva del tagging e la precisione e il recall sulle entità identificate. Queste metriche sono state confrontate con quelle ottenute da altri approcci di tagging, come un tagger di tipo Naive Bayes e un tagger di tipo MEMM (Maximum Entropy Markov Model), al fine di valutare l'efficacia del nostro sistema rispetto a metodi di riferimento.

Nelle prossime sezioni, approfondiremo la filiera del lavoro coinvolta nello sviluppo del sistema di NER tagging multilingua e l'implementazione dell'algoritmo di Viterbi per affrontare questa sfida complessa.