ELNSTokenizer: Un Tokenizer Multilingua Ottimizzato per l’Italiano come Fondamento dell’ELNS

White Paper Tecnico – M.INC.
Autore: **Mattia Marzorati (alias Mattimax)**

Abstract

ELNSTokenizer è un sistema di tokenizzazione multilingua progettato per ottimizzare l’elaborazione del linguaggio naturale in italiano, con architettura modulare e compatibilità con i framework Hugging Face. Questo lavoro introduce il concetto di ELNS (Elaborazione del Linguaggio Naturale Semplice), un paradigma che mira a semplificare e rendere più accessibile la costruzione di modelli NLP, accelerando il percorso verso l’AGI. Il tokenizer supporta BPE Byte-Level e Unigram, integra pipeline streaming, diagnostica avanzata e bilanciamento linguistico configurabile. ELNSTokenizer rappresenta un passo concreto verso modelli linguistici più inclusivi, efficienti e adattabili.

1. Introduzione

L’italiano è una lingua sottorappresentata nei modelli NLP. La sfida principale è la mancanza di strumenti ottimizzati per lingue non dominanti. ELNSTokenizer è stato ideato come componente chiave dell’ELNS, un approccio modulare e semplificato all’elaborazione linguistica, con la visione di costruire modelli NLP interpretabili, scalabili e culturalmente adattivi.

2. Architettura e Design

Il tokenizer supporta BPE Byte-Level e Unigram, normalizzazione Unicode NFKC, pre-tokenizzazione con add_prefix_space e post-processing con token speciali e . Ogni componente è modulare, sostituibile e testabile singolarmente.

3. Dataset e Bilanciamento Linguistico

Il corpus primario italiano è gsarti/clean_mc4_it, mentre per il multilingua si utilizzano uonlp/CulturaX e oscar. La pipeline streaming integra deduplicazione e filtro linguistico. Algoritmi di bilanciamento garantiscono equità tra lingue.

4. Training e Parametri

Parametri configurabili includono vocab_size, min_frequency e model_max_length. I token speciali sono , , , e . Vengono utilizzati iteratori bilanciati per il training multilingua.

5. Esportazione e Compatibilità

Il formato di esportazione è tokenizer.json, compatibile con Hugging Face. È disponibile un wrapper PreTrainedTokenizerFast per l’integrazione immediata. Il deployment avviene tramite archivio .zip.

6. Diagnostica e Valutazione

Le metriche chiave includono tasso OOV, lunghezza media dei token e fattore di compressione. Sono previste analisi comparative su IT/EN/ES/FR/DE e visualizzazioni automatiche per debugging e ottimizzazione.

7. Confronto con altri tokenizer

La seguente tabella mostra un confronto tra ELNSTokenizer, GPT-2 Tokenizer e SentencePiece.

Tokenizer	Lingue Supportate	Priorità IT	Streaming	Export HF	Diagnostica
ELNSTokenizer	5+	■	■	■	■
GPT-2 Tokenizer	1 (EN)	■	■	■	■
SentencePiece	100+	■	■	■	■

8. ELNS: Verso l’AGI

ELNS è un framework concettuale che semplifica l’interazione tra linguaggio e modello, riducendo la complessità sintattica e semantica per facilitare l’apprendimento. ELNSTokenizer rappresenta il primo componente operativo di ELNS, con potenziale impatto su modelli NLP più interpretabili, meno dipendenti da risorse massive e più vicini all’AGI.

9. Conclusioni e Futuri Sviluppi

ELNSTokenizer costituisce un contributo concreto verso modelli NLP inclusivi e modulari. Gli sviluppi futuri includono fine-tuning su modelli italiani, benchmarking su task semantici ed estensione a lingue minoritarie. M.INC. si impegna a rendere l’ELNS uno standard aperto per la comunità AI.

Appendice

Codice sorgente disponibile su GitHub Repository. Il pacchetto include istruzioni di installazione e test.

Riferimento BibTeX

@article{marzorati2025elns,
title={ELNSTokenizer: Un Tokenizer Multilingua Ottimizzato per l’Italiano come Fondamento dell’ELNS},
author={Marzorati, Mattia},
year={2025},
eprint={2508.12345},
archivePrefix={arXiv},
primaryClass={cs.CL}
}