Titolo progetto borsa di ricerca

Studio e sperimentazione di metodologie innovative di intelligenza artificiale per l'analisi e la classificazione dei testi anche su base semantica.

Oggetto dell'attività della borsa

Per esigenze tecniche, legate anche all'utilizzo di attrezzature specifiche in dotazione al dipartimento, il progetto di ricerca richiede necessariamente che l'attività sia svolta in presenza nella sede indicata nel bando, con possibilità di incontri in presenza anche presso l'azienda committente su indicazioni fornite dal supervisore della borsa di ricerca.

Il progetto si propone di individuare e valutare un modello di rappresentazione semantica del linguaggio capace di supportare una serie di task, in prima battuta il supporto alla ricerca semantica, la classificazione di testi, l'estrazione di informazioni. Le fonti documentali alle quali attingere si caratterizzano per la molteplicità di lingua, ad ora comprendente Francese, Inglese, Spagnolo, e Tedesco. L'ambito di provenienza del dato presenta un ulteriore punto di attenzione: in particolare si valuterà la capacità dei modelli di adattarsi ad ambiti semantici specifici, quali manuali tecnici o procedure, fortemente distanti tra loro perché relativi a verticali produttivi distinti e separati. L'articolo "Attention Is All You Need" [Vaswani A. et al. 2017] decreta l'introduzione dell'architettura di Transformer nel panorama del processamento del linguaggio naturale (NLP), e successivamente in altri ambiti, per esempio il trattamento di immagini. Questi approcci, caratterizzati da una forte richiesta di risorse di calcolo per il training iniziale, possono essere adattati a un ampio numero di task attraverso un approccio detto di fine-tuning. Data l'alta capacità che li caratterizza, questi modelli necessitano di una mole elevata di dati in input, tali per cui il task iniziale è scelto in ottica non supervisionata. Alcuni approcci, prevedono l'utilizzo di corpora multi-lingua [Devlin J. et al. 2019, Alexis C. et al. 2019]. Ancora da valutare è sia l'effettiva capacità dei modelli sia di gestire tali lingue con uguale efficacia [Pires T. et al. 2019], che la vera e propria costruzione di uno spazio semantico multi-lingua [Karthikeyan K. et al. 2020], obiettivo invece perseguito da modelli agnostici alla lingua di ingresso [Artetxe M. et al. 2018]. Se le tempistiche di progetto e le risorse computazionali disponibili lo permetteranno, sarà preso in considerazione anche il modello generativo T5 pre-addestrato anche sulla lingua italiana [Raffel C. et al.] e possibilmente anche la recentissima versione per documenti lunghi [Guo G. et al.]. La valutazione in oggetto parte dal presupposto di poter utilizzare un modello di embedding allenato con tecniche di self-supervision per adattarlo a task di interesse applicativo. In particolare, l'approccio verrà considerato all'interno di un framework di gestione documentale per corredare l'applicativo principale di funzionalità di classificazione e scoring.

L'obiettivo di questo progetto concerne la valutazione di un modello di rappresentazione del linguaggio secondo precise direttive:

- il grado di riutilizzo su un parco limitato di lingue (Francese, Inglese, Spagnolo, e Tedesco),
- la capacità di generare una rappresentazione multilingua capace di sostenere una metrica di similarità semantica su frasi della stessa lingua,
- la valutazione di tale rappresentazione semantica in un'ottica più estesa di similarità multi-lingua,

- l'applicabilità della soluzione su un numero di task "downstream" quali classificazione, Entity Recognition attraverso un processo di fine-tuning.

Bibliografia non esaustiva

Vaswani A. et al. (2017), *Attention is All you Need*, NIPS 2017

Devlin J. et al. (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, NAACL-HLT 2019

Alexis C. et al. (2019), *Unsupervised Cross-lingual Representation Learning at Scale*, ACL 2019

Pires T. et al. (2019), *How multilingual is Multilingual BERT ?*, ACL 2019

Karthikeyan K. et al. (2020), *Cross-lingual ability of Multilingual BERT: an Empirical Study*, ICLR 2020

Artetxe M. et al. (2018), *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond*, TACL 2019

Raffel C. et al. (2020), Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. volume 21, pp. 140:1--140:67, http://jmlr.org/papers/v21/20-074.html

Guo G. et al. (2021), LongT5: Efficient Text-To-Text Transformer for Long Sequences, CoRR volume abs/2112.07916, https://arxiv.org/abs/2112.07916