Riassumere testi giuridici con GPT-3

Andrea Bolioli, Manola Cherubini[1], Francesco Romano[1], Nazareno De Francesco[2]

[1] *Istituto di Informatica Giuridica e Sistemi Giudiziari del CNR*
[2] *MAIZE srl*

Abstract

In questa presentazione descriviamo sinteticamente una sperimentazione di *summarization* di testi giuridici effettuata con GPT-3 (*Generative Pre-trained Transformer 3*). Gli autori (ricercatori di informatica giuridica e di Natural Language Processing) hanno realizzato l'esperimento su un set eterogeneo di atti amministrativi e di norme di vario livello. L'obiettivo della sperimentazione è stato verificare le performance di un servizio aperto al pubblico e basato su LLM proprietari nella creazione di riassunti di testi giuridici in italiano.

Keywords

Informatica giuridica, summarization, GPT-3, large language models

1. Introduzione e stato dell'arte

Le recenti tecnologie di intelligenza artificiale applicate al linguaggio naturale, come i modelli linguistici pre-addestrati, o *large language models* (LLM), stanno mostrando miglioramenti significativi in task classici di Natural Language Processing (NLP). Tra questi, la *summarization* è un task classico complesso che si sta rivelando "alla portata" degli LLM.

Una rassegna dei metodi di automatic text summarization (single-document o multi-document, extractive o abstractive) si trova in [1]. In [2], è stato utilizzato il modello BERT (Bidirectional Encoder Representations from Transformers) per l'extractive summarization del contenuto di lezioni, in inglese. In [3], BERT è stato utilizzato nella summarization extractive e abstractive di notizie in inglese.

Anche i documenti legali in inglese sono oggetto, da tempo, di sperimentazioni di summarization (principalmente estrattiva), e i risultati stanno migliorando negli ultimi tempi sotto vari aspetti. Oltre alla rassegna presentata in [4], segnaliamo tra i lavori più recenti l'analisi comparativa di sistemi di summarization su testi legali descritta in [5].

Le sperimentazioni di LLM sulla lingua italiana sono ovviamente meno numerose. In [6], gli autori presentano IT5, "the first family of encoder-decoder transformer models pretrained specifically on Italian", i test realizzati su vari task NLP tra i quali la summarization di articoli di Wikipedia e di notizie (in italiano), e i risultati ottenuti, migliori rispetto ai modelli T5 multilingua. In [7], gli autori presentano il modello BART-IT per la lingua italiana e i test di abstractive summarization effettuati.

Per quanto di nostra conoscenza, le uniche sperimentazioni di summarization su testi giuridici in italiano sono descritte in [8] dagli autori del presente contributo. In questa presentazione riassumiamo la sperimentazione effettuata con GPT-3 (Generative Pre-trained Transformer 3), l'analisi dei metodi e dei risultati, le relative conclusioni temporanee e le attività in corso.

Per una presentazione più generale di tecnologie di IA applicate al dominio legale e giuridico, rimandiamo a [9].

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29-31, 2023, Pisa, Italy
EMAIL: andrea.bolioli@gmail.com (A. 1)

2. Large Language Models e GPT-3

GPT-3 (Generative Pre-trained Transformer 3) è un modello linguistico autoregressivo con 175 miliardi di parametri, sviluppato dall'azienda OpenAI nel 2020 e descritto nell'articolo scientifico [10]. È il modello di terza generazione della serie GPT, che utilizza l'apprendimento profondo per produrre testo simile al linguaggio naturale umano.

L'architettura si basa su una rete di Transformer con dimensioni del contesto di 2048 token. Il task del metodo di pre-addestramento utilizzato è (solamente) prevedere qual è il prossimo token in una sequenza ("pre-allenamento generativo"). Per quanto riguarda il set di dati testuali su cui è stato addestrato, complessivamente è composto da circa 499 miliardi di token. Il set di training è parzialmente multilingua: circa il 93% delle parole del corpus sono in inglese, mentre il restante 7% sono in altre lingue, tra le quali il francese, il tedesco, lo spagnolo. La lingua italiana è rappresentata nello 0,6% del corpus, con più di 1 miliardo di parole.

A differenza dei sistemi precedenti, GPT-3 mostra risultati sorprendenti nei casi di "few-shot", "one-shot" e "zero-shot learning".

GPT-3 è utilizzabile come servizio a pagamento tramite un'interfaccia web o tramite API, per gli utenti registrati al servizio. Nel sito web di OpenAI si trova la documentazione che spiega come utilizzare il servizio e alcuni esempi di task specifici (rispondere a domande, riassumere testi, scrivere recensioni di ristoranti, generare codice software, ecc). A differenza di altri modelli linguistici pre-addestrati, non è possibile accedere al codice sorgente e ai modelli linguistici sottostanti.

3. Summarization di testi giuridici

Sfruttando la capacità di Zero-shot Learning di GPT-3, sono stati sottoposti al modello 20 porzioni di testi normativi e amministrativi, con il task specifico di generare dei riassunti. ll campione di testi è stato selezionato in modo da permettere di testare lo strumento con atti con caratteristiche anche molto diverse tra loro. Il dataset è composto, infatti, da cinque leggi statali, due decreti legislativi, un decreto-legge, un decreto del Presidente del Consiglio dei Ministri, una Circolare del Ministero dell'Interno, una circolare del Ministero della Salute, una direttiva del Ministero del Lavoro, una direttiva della Presidenza del Consiglio dei Ministri, una direttiva del Ministero dell'Interno, una Ordinanza del Ministro della Sanità, due leggi regionali, una circolare della Regione Toscana e due ordinanze comunali. Circa le caratteristiche formali di tali risorse bisogna segnalare che, a parte quattro atti che sono stati emanati tra il 1983 e il 1998, il campione restante riguarda tutti atti successivi al 2000. Per una descrizione dettagliata del campione di testi, si rimanda a [8].

Il compito richiesto a GPT-3 è stato la creazione di un riassunto con un massimo di 500 caratteri per ognuno dei 20 documenti selezionati. In termini NLP, questo compito viene definito Abstractive text summarization (riassunto di tipo astrattivo). Per ogni riassunto è stato richiesto un numero massimo di 500 caratteri, parametro indicativo ma non vincolante per GPT-3. Non è stato usato nessun esempio di riassunto a supporto, quindi GPT-3 ha funzionato in modalità Zero-shot. I riassunti sono stati generati utilizzando i seguenti parametri di generazione nel Playground di GPT-3 (https://beta.openai.com/playground): modello: text-davinci-002; temperature: 0.7; frequence penalty: 0.0; presence penalty: 0.0; best_of: 3.

Le metriche di valutazione di task di generazione di testo spesso non sono esaustive e in alcuni casi manifestano scarsa correlazione con la valutazione umana. Nel caso della summarization di testi legali in italiano non era disponibile un gold standard di riferimento, per cui è stata prima effettuata una validazione manuale dei riassunti generati da GPT-3, e successivamente i due autori con competenze giuridiche hanno scritto i 20 riassunti, a loro giudizio "corretti", per poter effettuare anche una valutazione quantitativa (confrontando il riassunto automatico con il riassunto manuale).

La validazione manuale del riassunto, secondo la letteratura scientifica in questo ambito, tiene conto soprattutto dei seguenti aspetti:

- contenuto del riassunto (*Informativeness & Conciseness*): valutare se i sommari generati sono in grado di restituire le informazioni più rilevanti e se presentano informazioni ridondanti;
- forma del riassunto (*Coherence*): verificare che sia preservata la leggibilità dell'output generato, quindi che non vi siano errori grammaticali o informazioni in contrasto tra di loro.

I parametri, dunque, in base ai quali sono stati valutati i venti testi generati in output da GPT-3, sono stati i seguenti:

1. errori di scrittura;
2. errori di interpretazione del testo;
3. informazioni mancanti;
4. informazioni ridondanti.

L'analisi qualitativa degli errori e la valutazione quantitativa (ROUGE-n e BERT score) si trovano in [8].

4. Conclusioni e attività in corso

In questa presentazione abbiamo descritto brevemente una sperimentazione di abstractive summarization di testi giuridici in italiano con GPT-3. Sul campione di testi utilizzato, gli output del sistema sono stati in generale coerenti e pertinenti. Gli errori riscontrati sono stati analizzati dagli esperti giuristi. Il primo esperimento, le analisi e i risultati sono descritti in modo dettagliato in [8].

Il limite principale di questa sperimentazione è ovviamente la non riproducibilità dell'esperimento. L'azienda OpenAI infatti non ha rilasciato finora i suoi LLM in open source ma mette solo a disposizione degli utenti un servizio web le cui prestazioni stanno cambiando nel corso dei mesi. Inoltre, quando abbiamo effettuato più volte la richiesta di summarization di un singolo documento, il sistema non ha restituito output identici anche nei casi in cui il prompt (la richiesta) e i parametri erano invariati.

Ci è sembrato comunque interessante verificare le performance di GPT in un task di summarization di testi giuridici, per poterlo confrontare con altri sistemi. Il modello linguistico di GPT-3 non è stato addestrato esplicitamente su testi del dominio giuridico e ha funzionato in modalità Zero-shot, quindi senza esempi di riassunti a supporto. Inoltre, il modello GPT-3 utilizzato era stato pre-addestrato principalmente sulla lingua inglese e solo parzialmente sull'italiano.

Visti i primi risultati positivi, gli autori stanno proseguendo nella sperimentazione di summarization di testi giuridici in italiano nelle direzioni seguenti:

- espansione del dataset di testi giuridici su cui effettuare la summarization;
- espansione del golden standard di riassunti scritti dai giuristi;
- sperimentazione del task di semplificazione dei testi giuridici unito alla summarization, ovvero generazione di un riassunto che utilizzi solo frasi brevi e parole semplici;
- sperimentazione della modalità Few-shot learning in GPT-3.5;
- confronto con LLM open source;
- analisi del tema dell'explainability dei modelli;

5. References

[1] El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. "Automatic text summarization: A comprehensive survey." Expert systems with applications 165 (2021): 113679.

[2] Miller, Derek. "Leveraging BERT for extractive text summarization on lectures." arXiv preprint arXiv:1906.04165 (2019).

[3] Liu, Yang, and Mirella Lapata. "Text summarization with pretrained encoders." arXiv preprint arXiv:1908.08345 (2019).

[4] Kanapala, Ambedkar, Sukomal Pal, and Rajendra Pamula. "Text summarization from legal documents: a survey." Artificial Intelligence Review 51 (2019): 371-402.

[5] Núñez-Robinson, Daniel, Jose Talavera-Montalto, and Willy Ugarte. "A Comparative Analysis on the Summarization of Legal Texts Using Transformer Models." In Advanced Research in Technologies, Information, Innovation and Sustainability: Second International Conference, ARTIIS 2022, Santiago de Compostela, Spain, September 12–15, 2022, Revised Selected Papers, Part I, pp. 372-386. Cham: Springer Nature Switzerland, 2022.

[6] Sarti, Gabriele, and Malvina Nissim. "It5: Large-scale text-to-text pretraining for italian language understanding and generation." arXiv preprint arXiv:2203.03759 (2022).

[7] La Quatra, Moreno, and Luca Cagliero. "BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization." Future Internet 15, no. 1 (2022): 15.

[8] Cherubini, Manola, Francesco Romano, Andrea Bolioli, Nazareno De Francesco, Irene Benedetto. "La summarization di testi giuridici: una sperimentazione con GPT-3." Rivista italiana di informatica e diritto. 5, 1

(2023)
DOI:https://doi.org/10.32091/RIID0103.
[9] Cherubini, Manola, and Francesco Romano. "Legiferare con l'Intelligenza Artificiale." Journal of Ethics and Legal Technologies 4, no. 1 (2022).
[10] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.