Chair Data Science and Artificial Intelligence for Digitalized Industry and Services

Internship project

Subject

Summarization of legal text

Possibility to continue as a PhD candidate

YES (Funding to be confirmed)

About the chair

The Chair Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS), lead by Florence d'Alché-Buc, a Professor in the department Image, Data, Signal of Telecom Paris, unites five industrial partners: Airbus Defence & Space, Engie, Idemia, Safran et Valeo. It's general objective is to develop, in collaboration with the partners, teaching and research of the international level.

Its four principal research directions are:

1. Building predictive analytics on time series and data streams.
2. Exploiting large scale, heterogeneous, partially labeled data.
3. Machine Learning for trusted and robust decision.
4. Learning through interactions with environment.

Description of the internship

Supervision

Nils Holzenberger (https://perso.telecom-paristech.fr/holzenberger/)

Location and dates of the internship

Address : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date of the beginning of the internship : 2024

Team where the thesis will be written

Department INFRES, Team Data, Intelligence and Graphs (DIG)

Keywords

Summarization, Law, Natural Language Processing

Detailed subject

Legal cases generally deal with sets of documents, and decisions usually summarize vast amounts of information. In particular, judges often write summaries of the facts of a case when justifying a decision, and so do lawyers when defending a case in front of a jury. To some extent, a summary is also an interpretation of facts. The ability to summarize thus plays a crucial role in the legal domain. In the context of Natural Language Processing (NLP), summarization involves choosing what information to keep and what to discard.

Summarization is of interest from a research perspective — how does one produce summaries that

introduce a minimal amount of interpretation? — and from a practical perspective — can an NLP model reliably summarize large amounts of documents for legal purposes? A summary also makes it possible to efficiently compare cases, to find similarities and differences. A further refinement of summarization is contrastive summarization: given two (or more) cases, summarize them, while emphasizing what makes them different. This can help in retrieving relevant case law, and in finding inconsistencies between cases.

There are a couple of datasets ready to be used for legal summarization.

• **Civil Rights Litigation Clearinghouse** (CRLC). The CRLC curates summaries of civil rights cases, at multiple levels of granularities. See Shen et al. (2022) for initial attempts at summarization.

• **Legal tl;DR curated by lawyers** (https://www.tldrlegal.com/). This website contains summaries of software licenses.

The goal of this project is to determine how well current state-of-the-art NLP models can summarize legal documents, at multiple levels of granularity. This involves several questions, which this project will tackle in the following order:

1. Clearly define the summarization task
    - What are we summarizing? Cases, laws, documents used as proofs?
    - Determine the metrics that will serve to measure summarization performance
2. Analyze how interpretation is part of summarization
    - What kind of interpretation occurs in legal summaries?
    - Design a metric to measure the amount of interpretation
3. Experiment with generating summaries automatically
    - How well do state-of-the-art NLP models do?
    - Can we control the amount of interpretation that they introduce?

Candidate profile

- M2 student, interested in research
- Coursework in statistical machine learning, probabilities
- Good level of programming in Python
- Good command of English
- Optional: familiarity with deep learning and NLP libraries

Application

To send to nils.holzenberger@telecom-paris.fr:

- Curriculum Vitae
- Personalized motivation letter that explains interest of the candidate in the subject (can be directly in the body of the email)
- Grade reports for recent years
- Contact of a person willing to give recommendation

Incomplete applications will not be considered.

References

Shen, Z., K. Lo, L. Yu, N. Dahlberg, M. Schlanger, and D. Downey, 2022: Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. CoRR, abs/2206.10883.