Chair Data Science and Artificial Intelligence for Digitalized Industry and Services

Internship project

Subject

Statutory reasoning with constrained language

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

Natural Language Processing, Large Language Models, Code generation

Detailed subject

Legal professionals routinely need to determine which laws apply to a specific legal case. *Statutory reasoning* is the task of determining whether a given legal rule applies to a case, both being expressed in natural language. Statutory reasoning is a basic skill for lawyers, and computational statutory reasoning is a fundamental task for legal AI. The core challenge is developing models with the ability to utilize prescriptive rules stated in natural language, and able to generalize to new rules. The SARA dataset is a benchmark dataset for statutory reasoning (Holzenberger et al., 2020).

Statutory reasoning can be solved using formal reasoners, for example by translating laws and cases

into Prolog. This translation has been done manually for the SARA dataset. One way to perform this translation automatically is to use semantic parsing (https://en.wikipedia.org/wiki/Semantic_parsing). Semantic parsing turns natural language into a logical form. The logical form can then be interpreted by a computer to perform e.g. reasoning.

Shin et al. (2021) paraphrase natural language utterances into a controlled language, using a Large Language Model (LLM). The controlled language is described by a strict grammar, and sounds like a subset of natural language. The paraphrases in controlled language can then be deterministically mapped to the semantic parse, because the controlled language is quite simple. The goal of this project is to leverage this rewriting into a constrained language to parse legal language into a logical form, for statutory reasoning. The plan is to go through the following steps:

1. Determine what the controlled language should be
    - Prolog?
    - Logical English (Kowalski, 2020) or its legal version (Kowalski and Datoo, 2022)?
    - Attempto Controlled English? (http://attempto.ifi.uzh.ch/site/index.html)
2. Parse SARA statutes into the controlled language, following Shin et al. (2021)
3. Do the same for SARA cases
4. Determine whether this process can be used to simplify and explain statutes

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

Holzenberger, N., A. Blair-Stanek, and B. Van Durme, 2020: A dataset for statutory reasoning in tax law entailment and question answering. Proceedings of the Natural Legal Language Processing Workshop 2020.

Kowalski, R., 2020: Logical english. Proceedings of Logic and Practice of Programming (LPOP).

Kowalski, R. and A. Datoo, 2022: Logical english meets legal english for swaps and derivatives. Artif. Intell. Law, 30(2), 163–197.

Shin, R., C. H. Lin, S. Thomson, C. Chen, S. Roy, E. A. Platanios, A. Pauls, D. Klein, J. Eisner, and B. Van Durme, 2021: Constrained language models yield few-shot semantic parsers. EMNLP 2021.