Chair Data Science and Artificial Intelligence for Digitalized Industry and Services

Internship project

Subject

Semantic parsing for statutory reasoning

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, Semantic Parsing, Law, 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.

There has been work on mapping cases to their Prolog representation, with significant success. There has not been any work on mapping statutes to their Prolog representation. Codex (Chen et al., 2021) is a Large Language Model (LLM) trained on code, which has the ability to map natural language instructions to computer code. This is a form of semantic parsing. The goal of this project is to determine how well LLMs like Codex can map statutory language and case descriptions to Prolog code. The project's milestones are as follows:

1. Use Codex to map:
    - SARA statutes to Prolog
    - SARA cases to Prolog
2. Experiment with:
    - mapping a whole section of the statutes to its Prolog counterpart
    - mapping a single paragraph to its Prolog counterpart
    - few-shot and zero-shot prediction
3. Analyze where and why Codex fails to perform the mapping (Pertierra et al., 2017)
4. Experiment with modifying the Prolog code to be easier to map for Codex
5. Try the generated Prolog code for statutory reasoning

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

Chen, M., J. Tworek, H. Jun, *et al*, 2021: Evaluating large language models trained on code. CoRR, abs/2107.03374.

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

Pertierra, M. A., S. Lawsky, E. Hemberg, and U. O'Reilly, 2017: Towards formalizing statute law as

default logic through automatic semantic parsing. Second Workshop on Automated Semantic Analysis of Information in Legal Texts @ ICAIL 2017.