NLP in Legal domain

Context

- Introduction
  - Datasets
  - NLP to legal tasks
  - Legal Text Analysis
  - System Demos

- Experiments

Datasets

- Legal Datasets

- Or New annotated dataset

NLP Applications to legal tasks

- Legal Judgment prediction
- Similar Case Matching (in civil law)
- Legal Answering questions
- Bias and Privacy
- … ...
- **Challenge**
  - New Neural Models (Embeddings / Pretrained models) has a better performance (F1)
  - Incorporating legal knowledge into neural models for the performance and **interpretability**

Legal Text Analysis

- Text Classification
  - Toward Domain-Guided Controllable Summarization of Privacy Policies
- Text Summarization and Generation, or style transfer
  - Plain English Summarization of Contracts
- Sentiment analysis
  - Explaining potentially unfair clauses to the consumer with the CLAUDETTE tool
- Topic models
- Relation and Event Extraction
- Parsing
- Document relevance scoring

System Demos

- Demo to Use NLP for legal text
  - CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service => demo
  - Quick Check: A Legal Research Recommendation System

Methods

- Embedding Methods
  - Embedding space or legal knowledge graph
  - Pretrained models (BERT) + legal knowledge

- Symbol Methods
  - Legal element extraction

Experiments on Legal Text Classification

- Experiments on **Legal Text Classification**
  - **LEDGAR dataset** => **Done**
  - **DistilBERT on LEDGAR dataset** => **Done**
  - **Use legal-bert model as a baseline** => **Done**, need to check more

LEDGAR: A large-Scale Multilabel legal corpus

- LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts (2020)

- Data Source

LEDGAR: A large-Scale Multilabel legal corpus

- clean data (text, label, source)

{"provision": "Section and Subsection headings in this Amendment are included herein for convenience of reference only and shall not constitute a part of this Amendment for any other purpose or be given any substantive effect.", "label": ["headings"], "source": "2019/QTR1/000119312519044328/d691151dex101.htm"}

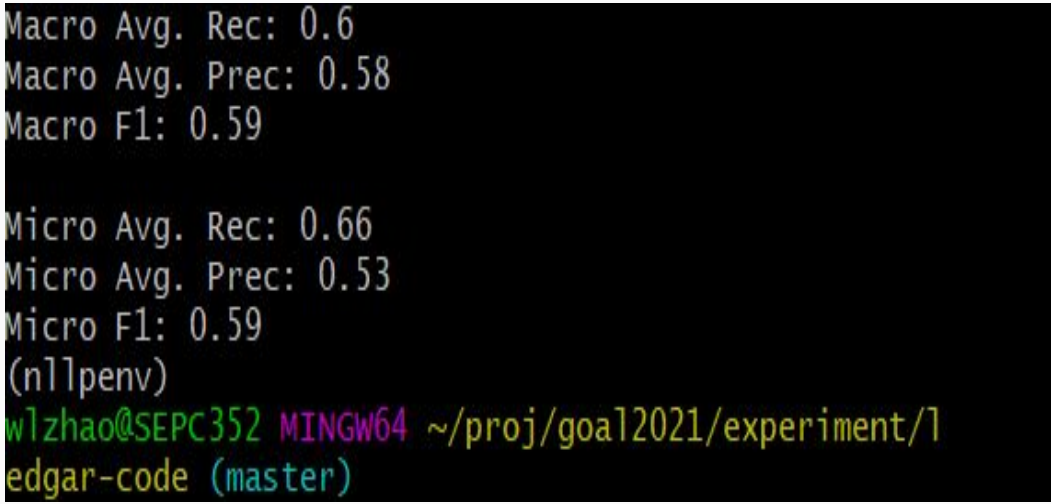
{"provision": "THIS AMENDMENT AND THE RIGHTS AND OBLIGATIONS OF THE PARTIES HEREUNDER SHALL BE GOVERNED BY, AND SHALL BE CONSTRUED AND ENFORCED IN ACCORDANCE WITH, THE LAWS OF THE STATE OF NEW YORK.", "label": ["applicable laws"], "source": "2019/QTR1/000119312519044328/d691151dex101.htm"}

{"provision": "This Amendment may be executed in any number of counterparts (and by different parties hereto in separate counterparts), each of which when so executed and delivered shall constitute an original, but all such counterparts together shall constitute but one and the same instrument; signature pages may be detached from multiple separate counterparts and attached to a single counterpart so that all signature pages are physically attached to the same document. Delivery of an executed counterpart by facsimile or other electronic transmission (e.g., “pdf” or “tif”) shall be effective as delivery of a manually executed counterpart of this Amendment.", "label": ["counterparts"], "source": "2019/QTR1/000119312519044328/d691151dex101.htm"}

- Original files from SEC

Experiments using DistilBert

- DistilBERT (a distilled version of BERT) Model on LEDGAR_2016-2019_clean_freq100.jsonl
- Source code



Experiments using LegalBert

- LEGAL-BERT: The Muppets straight out of Law School (2020)
- Model: nlpaueb/legal-bert-small-uncased

Update the DistilBERT source code for legal-bert-small-uncased model

- Legalbert_baseline.py
- NEED CHECK more details like the parameters

```
Macro Avg. Rec: 0.61
Macro Avg. Prec: 0.6
Macro F1: 0.61

Micro Avg. Rec: 0.67
Micro Avg. Prec: 0.56
Micro F1: 0.61
(nllpenv)
$ pwd
/c/Users/wlzhao/proj/goal2021/experiment/ledgar-code
(nllpenv)
wlzhao@SEPC352 MINGW64 ~/proj/goal2021/experiment/ledgar-code (master)
```

Next (on 2020-06-04)

TODO

- Large-Scale corpus
- LexNLP toolkit
- How to use GPU

Reference

- “How does NLP Benefit Legal System” by Haoxi ZHONG et al.
- NLLP-Natural Legal Language Processing
- LEGAL-BERT: The Muppets straight out of Law School (2020) by I Chalkidis et al.
- LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts by Don Tuggener et al.