NLP in legal domain

Context

- NLP in Legal domain
  - Legal Datasets
  - NLP to legal tasks
  - Legal Text Analysis
  - System Demos

- Experiments

Legal 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

- LEDGAR dataset
- DistilBERT on LEDGAR dataset
- Use legal-bert model as a baseline

- NEXT

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

- Format - {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"}

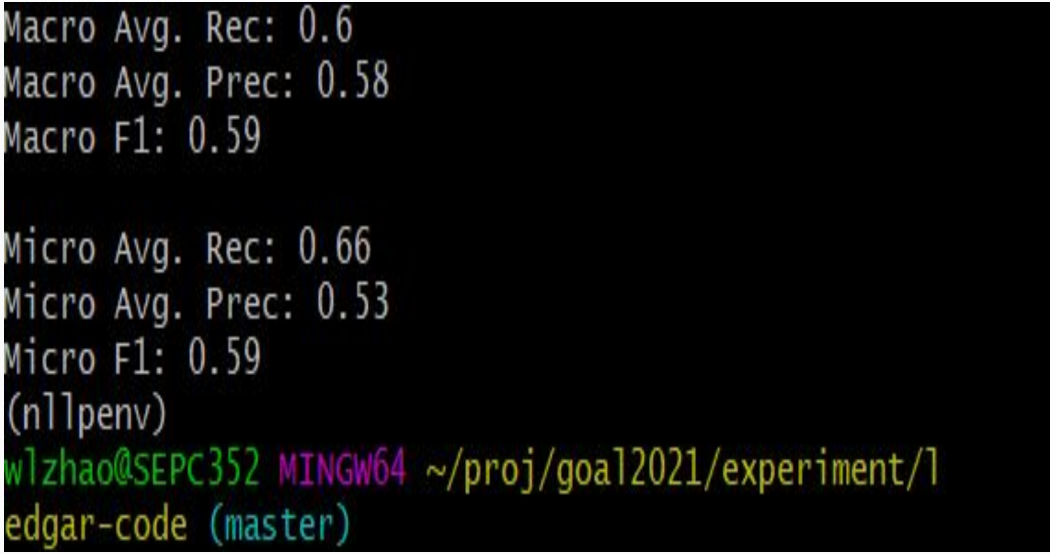
{"provision": "This Amendment may be executed in any number of counterparts (and by different parties hereto in separate counterparts), each of which when so executed and delivered shall constitute an original, but all such counterparts together shall constitute but one and the same instrument; signature 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



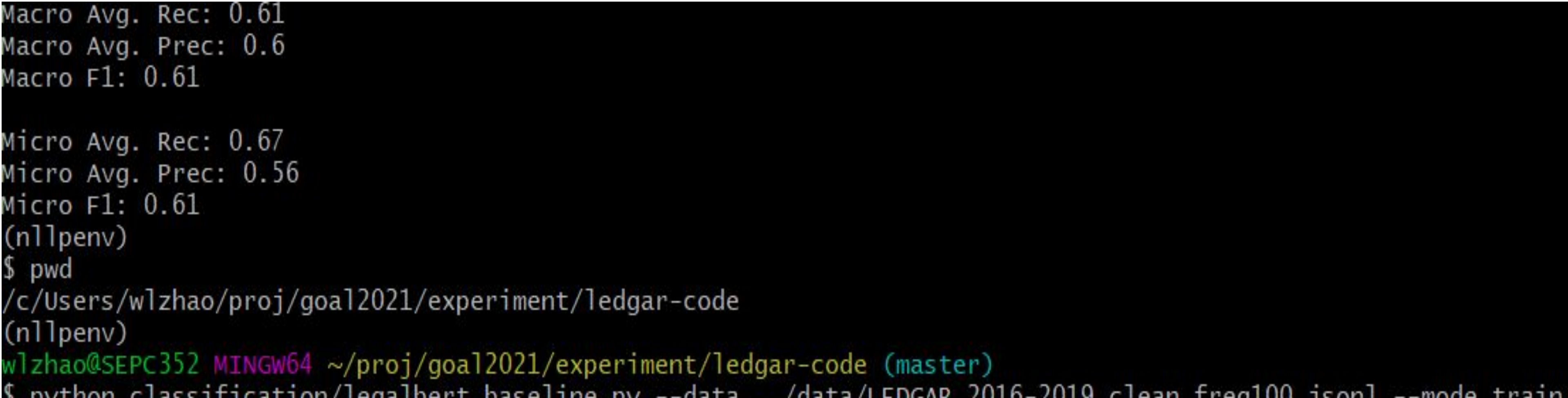
```
Macro Avg. Rec: 0.6
Macro Avg. Prec: 0.58
Macro F1: 0.59

Micro Avg. Rec: 0.66
Micro Avg. Prec: 0.53
Micro F1: 0.59
(nllpenv)
wlzhao@SEPC352 MINGW64 ~/proj/goal2021/experiment/l
edgar-code (master)
```

Experiments using LegalBert

- Paper: 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
    - Source code
    - Result from ser



```
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)
$ python classification/legalbert_baseline.py --data ../data/LEDGAR_2016-2019_clean_freq100.jsonl --mode train
```

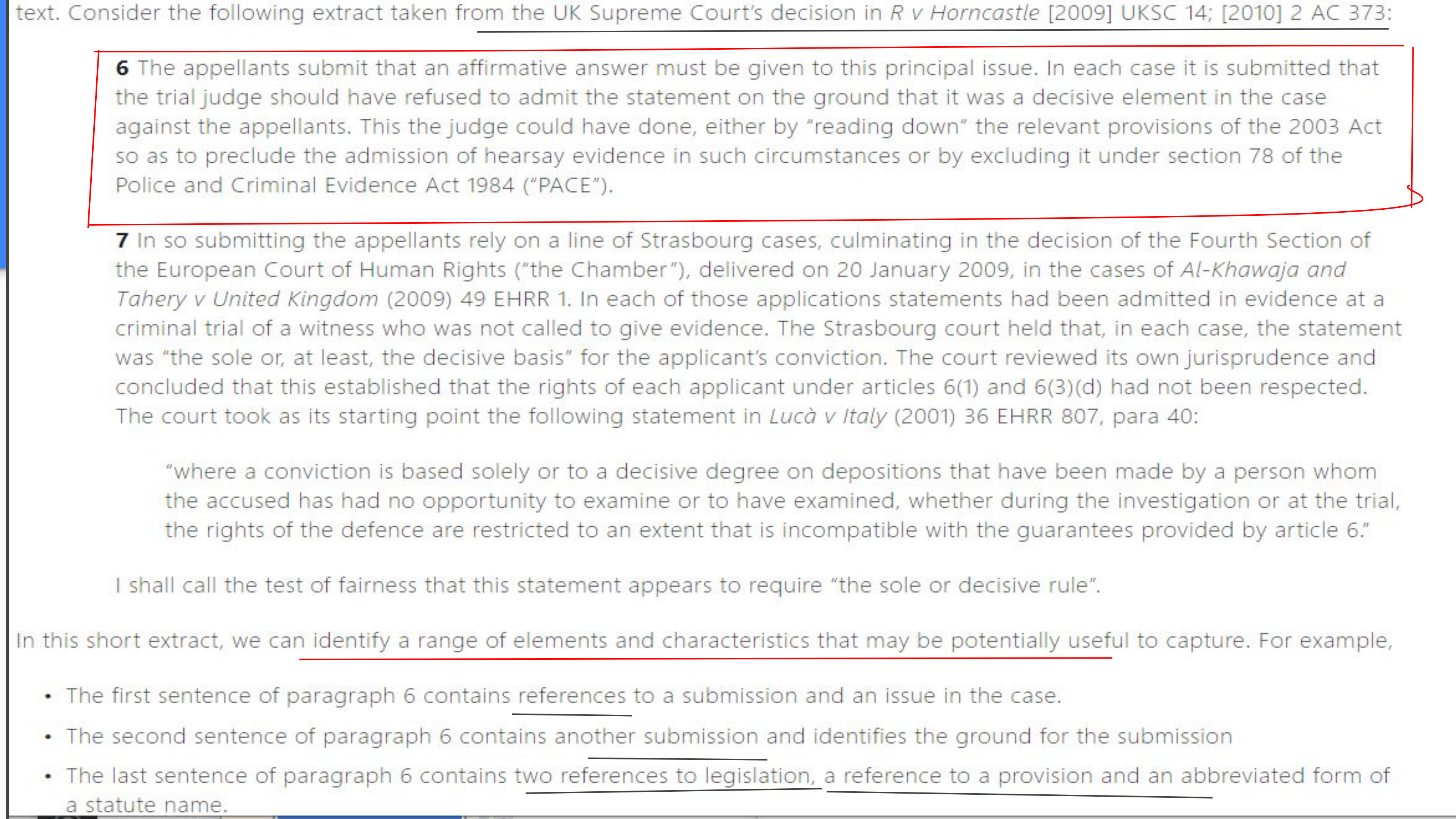
NEXT: Legal text such judgments?

- LexNLP: Natural language processing and information extraction for legal and regulatory texts
- Blackstone in Hong Kong
  - extract information from unstructured legal texts such as **judgments, scholarly articles, arguments and pleadings**, not only commercial applications

    => like reference, quote, submission
  - DATA: HKLII Databases

Example

- Assumption in common law
  - “the long and unstructured legal texts contain elements and characteristics that can be harnessed in a systematic way to improve our understanding of the meaning of the text and how it might fit into a larger corpus of text” (from ICLR)



text. Consider the following extract taken from the UK Supreme Court's decision in *R v Horncastle* [2009] UKSC 14; [2010] 2 AC 373:

> **6** The appellants submit that an affirmative answer must be given to this principal issue. In each case it is submitted that the trial judge should have refused to admit the statement on the ground that it was a decisive element in the case against the appellants. This the judge could have done, either by "reading down" the relevant provisions of the 2003 Act so as to preclude the admission of hearsay evidence in such circumstances or by excluding it under section 78 of the Police and Criminal Evidence Act 1984 ("PACE").
>
> **7** In so submitting the appellants rely on a line of Strasbourg cases, culminating in the decision of the Fourth Section of the European Court of Human Rights ("the Chamber"), delivered on 20 January 2009, in the cases of *Al-Khawaja and Tahery v United Kingdom* (2009) 49 EHRR 1. In each of those applications statements had been admitted in evidence at a criminal trial of a witness who was not called to give evidence. The Strasbourg court held that, in each case, the statement was "the sole or, at least, the decisive basis" for the applicant's conviction. The court reviewed its own jurisprudence and concluded that this established that the rights of each applicant under articles 6(1) and 6(3)(d) had not been respected. The court took as its starting point the following statement in *Lucà v Italy* (2001) 36 EHRR 807, para 40:
>
>> "where a conviction is based solely or to a decisive degree on depositions that have been made by a person whom the accused has had no opportunity to examine or to have examined, whether during the investigation or at the trial, the rights of the defence are restricted to an extent that is incompatible with the guarantees provided by article 6."
>
> I shall call the test of fairness that this statement appears to require "the sole or decisive rule".

In this short extract, we can identify a range of elements and characteristics that may be potentially useful to capture. For example,

- The first sentence of paragraph 6 contains references to a submission and an issue in the case.
- The second sentence of paragraph 6 contains another submission and identifies the ground for the submission
- The last sentence of paragraph 6 contains two references to legislation, a reference to a provision and an abbreviated form of a statute name.

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.
- Large-Scale Multi-Label Text Classification on EU Legislation by Ilias Chalkidis et al.
- LexNLP: Natural language processing and information extraction for legal and regulatory texts by Michael J Bommarito II et al.