Can Abstract Meaning Representation Facilitate Fair Legal Judgement Predictions?

Supriti Vijay[1], **Daniel Hershcovich**[2]

[1]Adobe, India, [2]Department of Computer Science, University of Copenhagen

ECtHR (ECHR Violation Prediction)										
Language Models	Average mF1	*Defendent State*			*Applicant Gender*			*Applicant Age*		
		mF1 ↑	GD ↓	$mF1_w$ ↑	mF1 ↑	GD ↓	$mF1_w$ ↑	mF1 ↑	GD ↓	$mF1_w$ ↑
Text Based Models										
DistilRoBERTa	62.9	63.3	2.1	61.2	59.0	2.0	56.3	61.3	2.5	58.5
$\text{DistilRoBERTa}_{FairLex}$	NA	53.2	8.3	44.9	57.5	3.1	54.4	54.1	5.9	46.2
AMR Split before Parsing										
LegalBERT_{SMALL}	54.8	50.5	1.2	49.3	47.1	5.4	40.4	52.4	4.8	47.2
AMR Split after Parsing										
LegalBERT_{SMALL}	57.3	**59.2**	**0.3**	**58.8**	**56.0**	3.5	**52.3**	**56.5**	3.7	**50.1**
(Dataset-specific LegalBERT_{SMALL})	44.2	40.4	5.3	35.0	32.1	**2.5**	28.9	33.3	**0.8**	31.9
DistilRoBERTa	37.6	36.5	0.7	35.7	31.6	4.4	28.3	36.2	5.4	27.6

Introduction

Legal judgment prediction holds potential to enhance legal system efficiency, but raises concerns about perpetuating biases. This paper employs Abstract Meaning Representation (AMR) to assess its ability to encode biases or abstract away from them in legal judgment prediction. AMR captures semantically meaningful information in a graph-like structure.

Prior Work and Motivation

Previous research has predominantly focused on AMR parsing of legal documents, with limited attention on assessing AMR's performance and fairness in legal tasks. This paper is the first to investigate whether AMR representations capture social biases alongside linguistic information in legal judgment prediction.

Proposed Methodology

We compare AMR's performance parity across different attributes of the **ECtHR dataset**, including **age**, **gender identity**, and **defendant state**. To evaluate the models' performance and fairness, we report three key metrics: **average macro-F1 score (mF1)**; **group disparity (GD)**; and **worst-group performance (mF1W)**. These metrics aim to gain insights into the fairness and robustness of AMR-based models in legal judgment prediction tasks.

Results

While AMR-based models exhibit worse overall performance than transformer-based models, they are less biased for attributes like age and defendant state compared to gender.

- AMR-based models demonstrate lower group disparity than the benchmark model for defendant state and applicant age, but higher for applicant gender
- Contextual details like time and location are connected to the event rather than the individual, while gender pronouns establish a direct link

AMR Parsing Techniques

AMR parsing techniques play a crucial role in capturing the semantic structure and relationships within legal documents. In this study, we explore two distinct approaches to AMR parsing: Splitting Before Parsing (SbP) and Splitting After Parsing (SaP).

1. **Splitting Before Parsing (SbP):** Splits cases pre-parsing, generates single-sentence AMRs, combines into multi-sentence graph.
2. **Splitting After Parsing (SaP):** Parses full cases, produces multi-sentence AMRs, linearizes, segments into 512 tokens.

31. On 23 June 1993 J. was placed in the family centre. T. visited her the same day.

Split-Before

```
(z0 / person
    :wiki -
    :name (z1 / name
              :op1 "J")
    :time (z2 / date-entity
              :day 23
              :year 1993))
(z0 / place-01
    :ARG2 (z1 / center
              :mod (z2 / family)))
(z0 / visit-01
    :ARG1 (z1 / she)
    :time (z2 / day
              :ARG1-of (z3 / same-01)))
```

Split-After

```
(z0 / visit-01
    :li 31
    :ARG0 (z1 / person
              :wiki -
              :name (z2 / name
                        :op1 "J."
                        :op2 "T."))
    :ARG1 (z3 / place-01
              :ARG1 z1
              :ARG2 (z4 / center
                        :mod (z5 / family))
              :time (z6 / date-entity
                        :year 1993
                        :month 6
                        :day 23
                        :time-of z0)))
```

Figure 1. We show a qualitative example showing differences in information passed across the two techniques

Conclusion

AMR-based models prioritize fairness with lower group disparity, but their lower worst-case performance renders them impractical for real-world use. The fairness demonstrated by AMR models, despite low disparity, resembles a random baseline due to lack of substantial performance. AMR may not be optimal for ensuring fairness in practice.

Connect with me

This QR code flies you right by our paper, where you can find out more about our findings. Feel free to text or email us about our work.

supriti.vijay@gmail.com

NAACL 2024

Poster Number: #25