September 2019

Identification of Rhetorical Roles of Sentences in Indian Legal Judgments

Supplementary Information

1. Sentence level agreement scores between the annotators

As described in the paper, we have three human annotators A_1,A_2,A_3. To measure the sentence-level agreement of their annotations, we compute the sentence-level agreement matrix between each pair of annotators, as described in the paper. The sentence level agreement matrix for the annotators A_2 and A_3 has been given in the paper. Here we shown the corresponding matrices for the other two annotator-pairs.

The sentence level agreement matrix (computed as described in the paper) for annotators A_1 and A_3 is presented in Table 1. This annotator pair has an overall agreement of 0.874 (average F-score) as measured by the GATE tool. This pair has the highest agreement among all the 3 pairs.

Similarly, the sentence level agreement matrix for annotators A_1 and A_2 in shown in Table 2. They have an overall agreement of 0.81 (average F-score) as measured by the GATE tool.

Table 1. Sentence level agreement between annotators A1 and A3

A1 ↓ A3 →	**FAC**	**ARG**	**PRE**	**STA**	**Ratio**	**RLC**	**RPC**
FAC	**2154**	8	0	3	34	11	0
ARG	29	**827**	0	0	0	0	0
PRE	0	0	**1464**	0	19	0	0
STA	0	0	0	**639**	10	0	0
Ratio	6	0	4	4	**3511**	1	0
RLC	36	1	0	0	25	**305**	0
RPC	6	0	0	0	21	0	**262**

Table 2. Sentence level agreement between annotators A1 and A2

A1 ↓ A2 →	FAC	ARG	PRE	STA	Ratio	RLC	RPC
FAC	**2207**	12	0	0	10	2	0
ARG	3	**816**	16	1	0	0	0
PRE	0	11	**1429**	0	28	0	0
STA	0	0	0	**642**	2	2	0
Ratio	2	13	0	1	**3604**	0	0
RLC	16	0	0	0	0	**301**	0
RPC	0	0	0	0	0	0	**262**

2. Average Inter-Annotator Agreement (IAA) across domains

As stated in the paper, we have documents from five domains of Law. We report in Table 3 the average IAA F-score for the labels across each domain. We can observe that inter-annotator agreement is uniform across different domains (as mentioned in the paper).

Table 3. Average IAA across different domains, and across different labels, in terms of F-score as measured by GATE

Domain ↓ Labels→	ARG	FAC	PRE	Ratio	RLC	RPC	STA	*Macro Average (across domains)*
Land & Property	0.883	0.808	0.823	0.79	0.841	0.789	0.888	0.831
Constitutional	0.851	0.865	0.837	0.8125	0.945	0.807	0.926	0.863
Criminal	0.784	0.8265	0.801	0.7925	0.777	0.833	0.931	0.821
Intellectual Property	0.786	0.802	0.913	0.764	0.840	0.742	0.944	0.827
Labour & Industrial	0.825	0.829	0.758	0.7995	0.800	0.858	0.858	0.818
Macro Average (across labels)	0.826	0.826	0.826	0.792	0.841	0.806	0.909	–

3. Dataset Statistics

The final curated gold standard dataset used in the experiments contained 9,308 sentences in total. The gold standard label for each sentence was decided based on the majority agreement among the three annotators, as stated in the paper. Some statistics on the dataset are presented in Table 4.

Table 4. Statistics of the gold standard corpus

Labels	Ratio	FAC	PRE	ARG	STA	RLC	RPC
% of total sentences	38.63%	23.13%	15.65%	9.00%	6.88%	3.36%	2.79%
Avg. length of sentences (#words)	26.28	22.29	25.04	29.00	32.13	28.32	16.61