Assessing the Difficulty of Spans Within Nested Text Markup Data Models

Tabachenkov Andrei M.

417 group

Scientific supervisor: Maysuradze Archil I.

October 31, 2025

Language model training and validation

Table: Validation of language model

Encoder and method	Accuracy	Precision-Macro	Recall-Macro	F1-micro
BERT + mean	0.887	0.761	0.699	0.720
BERT + attention	0.900	0.652	0.648	0.650
BERT + max pooling	0.914	0.717	0.647	0.667
BERT + endpoint	0.915	0.750	0.663	0.689
BERT + diff-sum	0.914	0.761	0.701	0.724
BERT + coherent	0.914	0.788	0.702	0.732
DeBERTa + mean	0.899	0.705	0.640	0.657
DeBERTa + attention	0.898	0.704	0.614	0.624
DeBERTa + max pooling	0.915	0.716	0.646	0.666
DeBERTa + endpoint	0.917	0.810	0.704	0.739
DeBERTa + diff-sum	**0.919**	**0.812**	**0.705**	**0.740**
DeBERTa + coherent	0.916	0.701	0.656	0.671
SpanBERT + endpoint	0.911	0.714	0.618	0.631

Validation of methods assessing difficulty

Table: ROC-AUC for difficulty assessing methods applied to different difficult datasets

Encoder + method	**Noisy dataset**				**Drug-dosage**				**Drug-effect**			
	VoG	$-M(x)$	$-\lvert M(x)\rvert$	Gaus.	VoG	$-M(x)$	$-\lvert M(x)\rvert$	Gaus.	VoG	$-M(x)$	$-\lvert M(x)\rvert$	Gaus.
BERT + mean	0.563	0.584	0.590	0.999	0.737	0.766	0.802	0.948	0.694	0.698	0.732	0.933
BERT + attention	0.542	0.571	0.573	0.999	0.796	0.830	**0.846**	0.953	0.758	0.783	0.801	0.936
BERT + max pooling	0.560	0.594	0.593	0.456	0.855	0.826	0.820	0.958	0.863	0.841	**0.828**	0.950
BERT + endpoint	0.581	0.610	0.610	**1.000**	0.826	0.830	0.801	0.960	0.842	0.788	0.764	0.957
BERT + diff-sum	0.571	0.611	0.610	**1.000**	**0.869**	**0.865**	0.805	**0.967**	0.856	**0.863**	0.808	**0.965**
BERT + coherent	0.546	0.605	0.603	0.472	0.848	0.845	0.806	0.935	**0.867**	0.852	0.810	0.925
DeBERTa + mean	0.578	0.598	0.599	0.999	0.762	0.742	0.779	0.935	0.767	0.683	0.716	0.925
DeBERTa + attention	**0.617**	0.603	0.606	0.999	0.572	0.818	0.842	0.945	0.539	0.759	0.769	0.921
DeBERTa + max pooling	0.572	0.613	0.613	0.424	0.839	0.807	0.796	0.910	0.846	0.780	0.751	0.891
DeBERTa + endpoint	0.553	0.616	0.616	**1.000**	0.800	0.831	0.767	0.941	0.827	0.837	0.752	0.937
DeBERTa + diff-sum	0.580	**0.616**	**0.616**	**1.000**	0.845	0.809	0.781	0.913	0.835	0.831	0.789	0.920
DeBERTa + coherent	0.571	0.609	0.608	0.448	0.805	0.830	0.712	0.900	0.801	0.815	0.685	0.873
SpanBERT + endpoint	0.560	0.612	0.611	**1.000**	0.841	0.848	0.725	0.955	0.864	0.846	0.695	0.946

Correlations between difficulty assessing methods

Table: Correlations between different difficulty assessing methods for different models

Encoder + method	**VoG**			$-M(x)$			$-\lvert M(x)\rvert$			**Gaussian**		
	$-M(x)$	$-\lvert M(x)\rvert$	Gaus.	VoG	$-\lvert M(x)\rvert$	Gaus.	VoG	$-M(x)$	Gaus.	VoG	$-M(x)$	$-\lvert M(x)\rvert$
BERT + mean	**0.886**	**0.835**	**0.621**	**0.886**	0.936	0.560	**0.835**	0.936	0.521	**0.621**	0.560	0.521
BERT + attention	0.829	0.789	0.607	0.829	0.944	0.602	0.789	0.944	0.592	0.607	0.602	0.592
BERT + max pooling	0.725	0.689	0.526	0.725	0.958	0.398	0.689	0.958	0.391	0.526	0.398	0.391
BERT + endpoint	0.739	0.701	0.476	0.739	0.963	0.314	0.701	0.963	0.302	0.476	0.314	0.302
BERT + diff-sum	0.718	0.679	0.538	0.718	0.963	0.289	0.679	0.963	0.278	0.538	0.289	0.278
BERT + coherent	0.702	0.670	0.600	0.702	0.964	0.428	0.670	0.964	0.422	0.600	0.428	0.422
DeBERTa + mean	0.806	0.762	0.226	0.806	0.940	0.238	0.762	0.940	0.216	0.226	0.238	0.216
DeBERTa + attention	0.575	0.490	0.142	0.575	0.936	**0.604**	0.490	0.936	**0.602**	0.142	**0.604**	**0.602**
DeBERTa + max pooling	0.662	0.633	0.276	0.662	0.962	0.062	0.633	0.962	0.054	0.276	0.062	0.054
DeBERTa + endpoint	0.589	0.557	0.474	0.589	0.968	0.186	0.557	0.968	0.180	0.474	0.186	0.180
DeBERTa + diff-sum	0.727	0.697	0.409	0.727	0.968	0.081	0.697	0.968	0.071	0.409	0.081	0.071
DeBERTa + coherent	0.685	0.665	0.386	0.685	**0.970**	0.211	0.665	**0.970**	0.212	0.386	0.211	0.212
SpanBERT + endpoint	0.676	0.644	0.398	0.676	0.965	0.203	0.644	0.965	0.196	0.398	0.203	0.196

Interpretability of methods assessing difficulty

VoG In these examples, there are 3 typical situations when span objects has high VoG score. In the first case such span is poly-semantic ("Open") and we observe poly-semantic difficulty of object. In the second case object's difficulty is caused by over-fitting of the model: for example, model thinks, that "II" is a part of person's name instead of event. Third case is when labelling of span is controversial: "Lanka" may be both geo-political and geographical object. Thus, VoG scores are interpretable but many high scores are simply connected with model over-fitting.

Margin If we use $-M(x)$ as difficulty score then in most cases high scores are because model's over-fitting ("Greek", "Time", "Liechtenstein"), so margins are less interpretable. By the way, this method is still able to detect difficult spans: "People", "vote".

Absolute value of margin In this method we observe that objects with lowest absolute values of margins are quite often correctly classified by the model but this method helps us detect named entities with difficult words and structure: "High Commissioner for Human Rights", "EU Foreign Policy", "Srebrenica" and others. Thus absolute values of margins are interpretable in their own way.

Gaussian method This method has connection with span embeddings anomalies. So there are two different cases: model's over-fitting ("Wimbledon", "Canal") and difficulty of named entities' by themselves ("Please").

Connection between difficulty and consistency

Table: Correlation between inconsistency scores and difficulty scores

Encoder and method	VoG	$-M(x)$	$-\lvert M(x)\rvert$	Gaussian
BERT + mean	0.000	0.013	0.009	0.014
BERT + attention	0.004	0.010	0.004	**0.015**
BERT + max pooling	0.003	**0.028**	**0.027**	0.008
BERT + endpoint	0.001	0.028	0.026	0.002
BERT + diff-sum	0.002	0.028	0.025	0.002
BERT + coherent	0.000	0.028	0.026	0.014
DeBERTa + mean	0.004	0.010	0.003	0.015
DeBERTa + attention	**0.016**	0.008	0.004	0.007
DeBERTa + max pooling	-0.003	0.012	0.009	0.007
DeBERTa + endpoint	-0.003	0.020	0.016	0.001
DeBERTa + diff-sum	0.004	0.021	0.017	0.003
DeBERTa + coherent	-0.001	0.020	0.016	0.000
SpanBERT + endpoint	0.012	0.016	0.015	0.006

Connection between difficulty and consistency

Table: ROC-AUC applied to difficulty scores of consistent objects (as negative ones) and inconsistent objects (as positive ones)

Encoder and method	VoG	$-M(x)$	$-\lvert M(x)\rvert$	Gaussian
BERT + mean	0.503	0.628	0.581	0.634
BERT + attention	0.458	0.599	0.541	0.641
BERT + max pooling	0.530	**0.772**	**0.755**	0.573
BERT + endpoint	0.508	0.765	0.746	0.479
BERT + diff-sum	0.484	0.766	0.743	0.478
BERT + coherent	0.499	0.765	0.746	0.630
DeBERTa + mean	0.535	0.596	0.525	**0.647**
DeBERTa + attention	0.351	0.578	0.538	0.435
DeBERTa + max pooling	0.472	0.613	0.589	0.439
DeBERTa + endpoint	0.473	0.688	0.649	0.507
DeBERTa + diff-sum	0.542	0.706	0.660	0.475
DeBERTa + coherent	0.492	0.693	0.655	0.504
SpanBERT + endpoint	**0.613**	0.650	0.642	0.562

Connection between difficulty and consistency

Table: The p-values from the Kolmogorov-Smirnov test for checking the hypothesis that difficulty scores of consistent and inconsistent objects are from the same distribution

Encoder and method	VoG	$-M(x)$	$-\lvert M(x)\rvert$	Gaussian
BERT + mean	0.940	0.399	0.355	0.386
BERT + attention	0.960	0.307	0.264	0.224
BERT + max pooling	0.838	**0.012**	**0.009**	0.392
BERT + endpoint	0.837	0.028	0.024	0.677
BERT + diff-sum	0.865	0.049	0.044	0.835
BERT + coherent	0.939	0.025	0.021	**0.128**
DeBERTa + mean	0.900	0.822	0.783	0.389
DeBERTa + attention	0.120	0.606	0.506	0.456
DeBERTa + max pooling	0.702	0.736	0.703	0.869
DeBERTa + endpoint	0.751	0.272	0.447	0.861
DeBERTa + diff-sum	0.718	0.132	0.243	0.783
DeBERTa + coherent	0.781	0.415	0.404	0.966
SpanBERT + endpoint	**0.113**	0.125	0.089	0.576