NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

DiP Benchmark

A unified platform for discourse evaluation of MT models.

Student: Shen Youlin **Supervisor:** Prof Joty Shafiq Rayhan

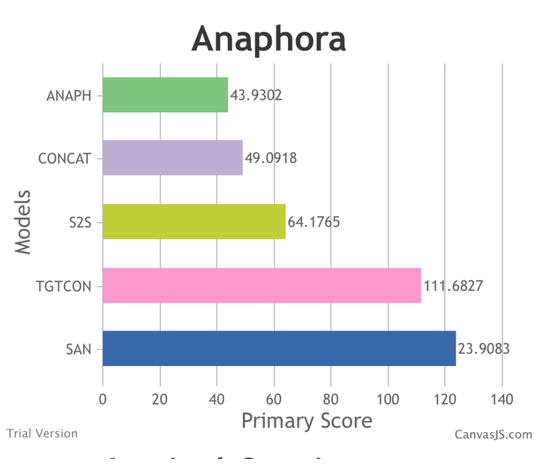
Testsets

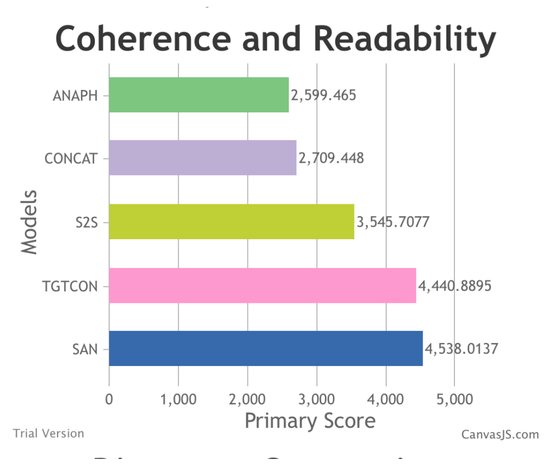
Testset_name	Phenomena	Test Size	Source Lang	Target Lang	Details	Evaluation
chinese_anaphora	Anaphora	1,540	Chinese	English	Details	Upload
german_anaphora	Anaphora	2,564	German	English	Details	Upload
russian_anaphora	Anaphora	2,368	Russian	English	Details	Upload
german_lexcon	Lexical Consistency	619	German	English	Details	Upload
russian_lexcon	Lexical Consistency	733	Russian	English	Details	Upload
russian_coherence	Coherence and Readability	331	Russian	English	Details	Upload

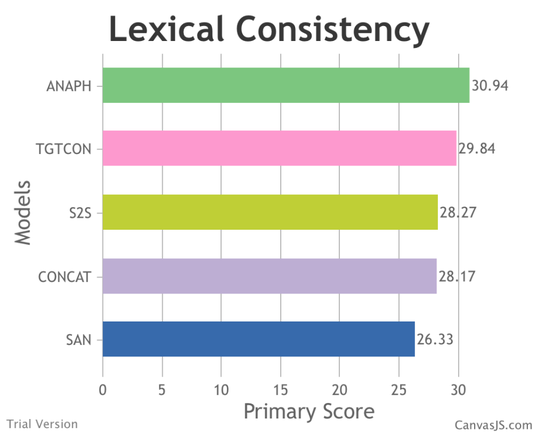
Testsets

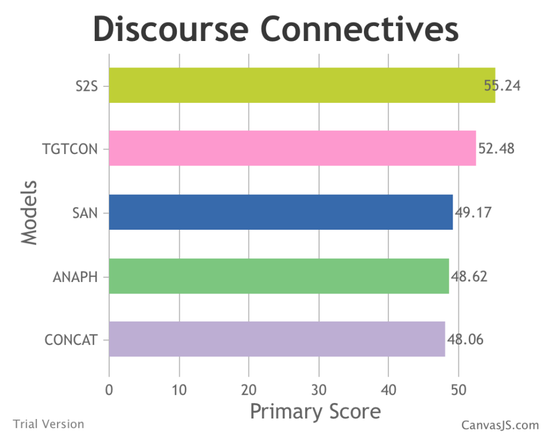
Project objectives:

The DiP Benchmark application aims to give NLP researchers more evidence on the quality of the MT system output by the comprehensive benchmark framework proposed by Jwala et. al. 2019. The framework checks 4 discourse phenomena, namely Anaphora, Lexical Consistency, Coherence, and Readability. The application features testset downloads (shown above), automatic model output evaluation, enhanced visualizations (bottom left figure), a leaderboard for researchers to compare their model performances (bottom right figure).









Leaderboard

Source Language: German

Rank	Model Name	Source Language	Target Language	BLEU	Anaphora
1	ANAPH	German	English	29.94	129.6621
2	CONCAT	German	English	31.96	112.5835
3	HAN	German	English	29.69	118.0674
4	S2S	German	English	31.65	113.7828
5	SAN	German	English	29.32	117.8379
6	TGTCON	German	English	29.94	131.6987