TABLE I
STRUCTURE-AWARE DATASET SIZES. 1. FOR XSUM, THE FIRST COLUMN LISTS THE TRAINING DATASET SIZE WITHOUT CR-ANNOTATED DATA; THE SECOND COLUMN SHOWS THE TRAINING DATASET SIZE WITH CR-ANNOTATED DATA. VALIDATION AND TEST DATASETS REMAIN THE SAME RESPECTIVELY SINCE WE DO NOT USE STRUCTURE DATA FOR VALIDATION AND TEST EVALUATIONS.

Dataset	**Size**		
	CNNDM	**XSum**[1]	
Train	149634	186873	65698
Validation	7782	10391	10391
Test	11483	11328	11328

I. SUPPLEMENTARY MATERIAL

A. Data Preparation

Two benchmark datasets, CNN/Daily Mail (CNNDM) and XSum, are used.

a) Structure-Aware CNNDM Datasets: We download CNNDM datasets (train, validation, and test) as JSON files using HuggingFace's datasets package. We use Stanford CoreNLP (v4.4.0) parsing package[1] to acquire dependency parsing, constituency parsing, and coreference resolution (CR) parsing. We then adopt an implementation[2] to extract SD labels, dependency structures, and the corresponding words from the dependency parsing. We have also developed a tool to extract entity referents, endophoric references, and their attributes from the CR parsing. The attributes include animacy (e.g., ANIMATE and INANIMATE), gender (e.g., MALE and FEMALE), number (e.g., SINGULAR and PLURAL), and type (e.g., PROPER and PRONOMINAL). We further build the vocabularies of the SD labels and the CR attributes respectively.

The CoreNLP, by default, uses Penn Treebank to tokenize document word sequences for building structure parsings. We thus build our article and summary datasets (train, validation, and test) from these tokens. Note that we build validation and test sets from annotated tokens to keep consistent token distributions with the training dataset even though validation and test inferences do not apply structure data. We consider such tokens as words to distinguish them from the runtime token encodings by the BART model tokenizer[3].

To save the token encoding time during training runs, we preprocess the encodings[4] of the built documents using the model tokenizer. The number of encoded output tokens may exceed the length limit of the model. To apply the syntactic structures (e.g., SD and CR) properly, we exclude these samples in the training dataset instead of truncating them.

We also build the encoded token graph to facilitate the super token representation learning as described in Section I-B. Table I lists the preprocessed dataset sizes.

[1]https://stanfordnlp.github.io/CoreNLP.

[2]https://github.com/KaiQiangSong/struct_infused_summ.

[3]https://huggingface.co/transformers/v4.9.2/model_doc/bart.html#barttokenizer.

[4]Vocabulary is from https://huggingface.co/facebook/bart-base.

b) Structure-Aware XSum Datasets: We also use HuggingFace's datasets package to download XSum datasets. We build XSum datasets in the same way as we build the annotated CNNDM datasets. It yields a smaller XSum training dataset with CR-annotated data. The small dataset may not be adequate to train models with the SD structure-aware semantic similarity regression. Thus, we also build a larger training dataset without CR-annotated data. We first train our models without the CR-based margin ranking task. We then include the CR-annotated data to further fine-tune the trained models along with the CR-based regression task. The dataset sizes are listed in Table I.

B. Key Implementation

Our implementation is developed using Python and Pytorch. We extend HuggingFace's BART-base implementation for our experiments.

a) Super Token Representation Learning: The BART adopts byte-pair encoding method ([3]) to deal with oversized vocabulary issues. On the other hand, the CoreNLP produces syntactic structures on words. To apply the word-level structures to the segmented word tokens, we follow the aggregated token representation approach (e.g., [1]). To do so, we employ a graph neural network (GNN)-based super token representation learning by utilizing a generic GNN model ([6]). In detail, we first build a token graph for each word in which the subsequent tokens have edges to the leading token during data preprocessing. To deal with a single-token word case, we add a self-loop edge to the leading token of each word. Inspired by the positional embeddings ([5]), we further build the distance embeddings of vocabulary size equal to the max model-allowed length and use them as edge features. During training, we apply the GNN to aggregate the tokens of each word to a super token node at the leading token position of each word. The word-level structures are applied to the super tokens thereafter.

The discussed GNN model is implemented by Pytorch Geometric package[5]. We adopt the model implementation from the package and configure it with our settings.

b) SD Structure-Aware Semantic Similarity Regression: We default the GNN model to a two-layer and undirected message-passing configuration after experimenting with several configurations (e.g., four-layer and directed).

Our SD graph-based token representation learning adopts the PNA implementation from the Pytorch Geometric package too. We extend the PNA to include our depth-based scaling function.

c) CR-Based Margin Ranking Regression: As the relations are encoded in the coreference representations, the scoring function of margin ranking in the main script is thus simplified as:

$$s_c = \mathrm{Re}(E_s \overline{E_o}^{\mathrm{T}}). \tag{1}$$

Referents and references may consist of multiple words. To simplify the learning task without loss of representational

[5]https://pytorch-geometric.readthedocs.io.

TABLE II
KEY TRAINING SETTINGS

Model latent dim	768
FFN dim	3072
Attention heads	12
Number of layers	6
AdamW optimizer	lr = $5e-5$ lr decay = linear weight decay = $1e^{-6}$
Max epoch	50
Early stop criterion	4

discriminative power, we adopt a super node representation learning similar to the super token representation learning described earlier. In short, we center on the keyword of an entity or endophoric reference and apply a GNN to aggregate its neighboring (up to) two words at each side of the keyword if applicable.

To choose a proper margin, we experiment with margins of 40, 50, and 60. The best performance is achieved at 50.

Our CR margin ranking regression refactors a number of functions from the knowledge graph learning package PyKeen[6]. The changes are made related to the complex vector-based model, interaction function, and margin ranking loss function.

d) Index Mapping from Word-Level Structure to Token: The model-encoded tokens and the word-level structures are not aligned by their sequential position indices. That is, their corresponding indices are not the same. To apply the structures to the super tokens, we have developed source codes for index mapping.

e) Weighting Learning Objectives: We set λ_{ssr} to 0.5 and default all other loss weighting factors to 1.0.

f) Training Setting Summary: We develop multi-GPU running procedure based on the reference runtime script[7], including AdamW optimizer settings except for the weight decay, which we adopt from [4]. Table II summarizes the key training settings.

We use an early-stop training approach up to the configured maximum epoch. The criterion is the ROUGE metric-based validation evaluation. The same ROUGE metrics for test time inference evaluation are used. The training stops when the validation ROUGE scores flatten consecutively for the number of configured times.

A model with the fully configured MLLTs has a size of 599.98MB. We train models using a dual-GPU setting with shared computational resources - two NVIDIA Quadro RTX 8000/48GB cards. A training session of the model with CNNDM datasets on the configured early stop setting takes about 60 hours. A training session with XSum datasets without the CR-based margin ranking regression takes roughly 95 hours on the same early stop setting. Further fine-tuning the model with the CR-based margin ranking regression takes about 6 hours on the same early stop setting.

g) ROUGE Metrics: The used ROUGE metrics output scores at low, medium, and high confidence intervals per metric. We report the scores on the high confidence interval.

[6]https://github.com/pykeen/pykeen/tree/master/src/pykeen.

[7]https://github.com/huggingface/transformers/blob/master/examples/pytorch/summarization/run_summarization_no_trainer.py.

TABLE III
ABLATION STUDY (ON CNNDM). 1. SUMMARY-LEVEL R-L SCORES (EQUIVALENT TO THE R-LSUM METRIC THIS PAPER USES). 2. SD LABEL CLASSIFICATION IS EXCLUDED. 3. SD STRUCTURE-AWARE SEMANTIC SIMILARITY IS EXCLUDED. 4. CR-BASED MARGIN RANKING IS EXCLUDED. 5. BOTH SD STRUCTURE-AWARE SEMANTIC SIMILARITY AND CR-BASED MARGIN RANKING ARE EXCLUDED. 6. THE MODEL TRAINED WITH FULLY CONFIGURED MLLTS FOR COMPARISON.

Model	**CNNDM Test Set**		
	R-1	R-2	R-L[1]
BART-base/MLLTs $\nexists$(SD)[2]	42.85	19.59	39.80
BART-base/MLLTs $\nexists$(SD-SSR)[3]	42.79	19.50	39.73
BART-base/MLLTs $\nexists$(CR-MRR)[4]	42.80	19.48	39.76
BART-base/MLLTs $\nexists$(SD-SSR,CR-MRR)[5]	42.76	19.39	39.73
BART-base*/MLLTs[6]	**43.00**	**19.67**	**39.91**

C. Ablation Study

We have also run our ablation experiments on CNNDM[8] to assess the impact and effectiveness of our multi-level learning tasks. Table III shows the results. The ablation experiments indicate that excluding the SD structure-aware semantic similarity and the CR-based margin ranking tasks has the most negative impact on the backbone model's performance. Although omitting the SD label classification has the least impact, incorporating it can enhance the proposed regression tasks and yield even better model performance.

It is worth noting that the different learning tasks are configured with the different latent layers of the backbone model. Since the CR-annotated data is sparse and less contextual while the SD data annotates contextual sentence structures, we have taken into account previous studies (e.g., [7]; [2]) and formulated the CR-based margin ranking regression using the lower layer latent states of the encoder in our experiments. On the other hand, both the SD structure-aware semantic similarity regression and the SD label classification fit well with the outputs of the encoder and/or the decoder.

D. Human Evaluation Criteria and Definition

There are a range of human evaluation criteria seen in prior works. To relate to our work, we choose fluency and faithfulness. But fluency has a close link to coherence while faithfulness is inevitably associated with relevance. So, we provide annotators with the following guidance:

[8]The ablation experiments are only conducted on CNNDM due to computational resource constraints.

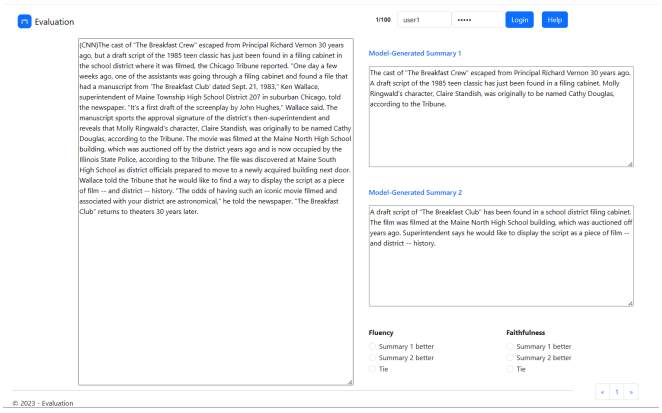

Fig. 1. An illustration of human evaluation user interface.

1) The grammatical or factual errors propagated from articles are discounted in deciding the respective choices.
2) If information in a summary is not evidenced in an article but can be reasonably deduced from the article, they are considered factual, not an error.
3) If both summaries of an article have no apparent factual errors, key information summarized from articles may be considered in deciding faithfulness.
4) If both summaries of an article contain factual errors, errors in key information have more significance in deciding faithfulness.
5) Missing punctuation marks may be less concern if they do not obscure comprehension or alter the facts from articles.
6) If unsure of which summary is better, 'tie' is the option to choose.
7) An annotator (evaluator) has the freedom to decide what are the key information that an article conveys.

E. Human Evaluation User Tool and Procedure

Fig. 1 illustrates the main user interface of the human evaluation tool. The user interface presents each article and its model-generated summaries side-by-side on a page. The interface asks an annotator to select the better summary for each of the two criteria. An annotator can also rank the summaries as 'tie' otherwise. The tool draws 50 random samples from CNNDM and XSum test sets, respectively. A total of 100 samples are shown to an annotator page-by-page. To make the evaluation as fair as possible, the tool presents summaries in a randomly shuffled order. So, the numbers "1" and "2" shown in "Summary 1" and "Summary 2" have no fixed corresponding to models throughout the evaluation. The user interface also does not indicate which dataset a sample is drawn from. The tool records each annotator's choices in a back-end database from which the queried analysis is conducted thereafter.

F. Qualitative Assessment

This section provides qualitative assessment samples in Table VI and Table VII for CNNDM and XSum, respectively.

TABLE IV
KEY PACKAGE VERSIONS

Package	Version
Python	3.8.10
Pytorch	1.11.0+cu113
Hugging Face (Transformers)	4.9.2
Pytorch Geometric (PyG)	2.0.4
PyKeen	1.9.0
Rouge Metrics	0.0.4
Stanford CoreNLP	4.4.0

TABLE V
DATASET VERSION AND LICENSE

Dataset	Version	License
CNNDM	3.0.0	Apache-2.0
XSum	1.0.0	N/A

G. Key Packages

1) Versions: Table IV summarizes the versions of key third-party packages used in this paper.

2) Licenses: The licenses related to these package versions are described as follows.

a) Python: Python license is accessible here[9].

b) Pytorch: Pytorch uses a collective license[10].

c) Hugging Face: Hugging Face[11] uses Apache-2.0 License and covers its packages, including Transformer-based models (e.g., BART), pre-trained BART-base tokenizer, datasets package (incl. CNNDM and XSum), and Accelerate package.

d) PyG: Pytorch Geometric[12] is a graph neural network modeling package and uses MIT License.

e) PyKeen: PyKeen[13] is a knowledge graph learning package and uses MIT License.

f) Rouge Metrics: Rouge Metrics uses Apache-2.0 License.

g) Stanford CoreNLP: CoreNLP[14] uses GNU General Public License v3.0.

h) SummaC: SummaC[15] uses Apache-2.0 License.

H. Datasets

Table V lists the version information of both CNNDM and XSum datasets.

Limitations

Although the chosen datasets in this paper represent typical characteristics of ATS, both abstractive/extrinsic and extractive/intrinsic, our investigation is confined to English news

[9]https://docs.python.org/3.8/license.html.
[10]https://github.com/pytorch/pytorch/blob/master/LICENSE.
[11]https://github.com/huggingface.
[12]https://github.com/pyg-team/pytorch_geometric.
[13]https://github.com/pykeen/pykeen.
[14]https://github.com/stanfordnlp/CoreNLP/tree/v4.4.0.
[15]https://github.com/tingofurro/summac.

articles. Our future work may apply our learning tasks to other document types.

Ethics Statement

To the best of our knowledge, we have attributed our work to prior works and implementations that this paper adopts either in the main script or in the supplementary material. Also, two authors participate in human evaluation.

REFERENCES

[1] Adam Ek and Jean-Philippe Bernardy. 2020. Composing Byte-Pair Encodings for Morphological Sequence Classification. In Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 76–86. Association for Computational Linguistics.

[2] Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 55–65. Association for Computational Linguistics.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725. Association for Computational Linguistics.

[4] Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-Infused Copy Mechanisms for Abstractive Summarization. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1717–1729. Association for Computational Linguistics.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

[6] Jiaxuan You, Zhitao Ying, and Jure Leskovec. 2020. Design Space for Graph Neural Networks. In Advances in Neural Information Processing Systems, volume 33, pages 17009–17021. Curran Associates, Inc.

[7] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 563–578. Association for Computational Linguistics.

TABLE VI

GENERATIVE SUMMARY ASSESSMENT ON CNNDM. WE USE ELLIPSES TO OMIT THE LENGTHY CONTENT NOT RELEVANT OR IMPORTANT TO OUR ILLUSTRATION. WE UNDERLINE THE RELATED FACTS IN BOTH SOURCE DOCUMENTS AND THE GENERATED SUMMARIES WITH STRAIGHT BLUE LINES. WE ALSO HIGHLIGHT THE FACTUAL ISSUES IN THE GENERATED SUMMARIES WITH RED WAVE SYMBOLS.

No.	Source	Text
1	Article	Matthew Hall, ... in Manchester's fashionable Northern Quarter district ... after scaling the walls of trendy apartment blocks where ...
	BART-base	Matthew Hall, ... scaled the walls of Manchester's fashionable Northern Quarter district. ...
	BART-base*/MLLTs	Matthew Hall, ... after scaling walls of trendy apartment blocks. ...
2	Article	... Tim Sherwood and Chris Ramsey know each other inside out. ... The pair worked together at Spurs with Les Ferdinand (second right), now QPR's director of football. ...
	BART-base	... Pair worked together at Spurs, now QPR's director of football.
	BART-base*/MLLTs	Tim Sherwood and Chris Ramsey know each other inside out. ... The pair worked together at Spurs with Les Ferdinand, now QPR's director of football. ...
3	Article	... Bikram Choudhury built an empire. ... after the guru was accused of rape or sexual assault by six of his former students. ... Sarah Baughn, a former student who now accuses Choudhury of sexual assault. ... He said he's guided by a deep calling to help others. ...
	BART-base	Bikram Choudhury says he wants to set the record straight. ... Sarah Baughn says he's guided by a deep calling to help others.
	BART-base*/MLLTs	Bikram Choudhury says he never sexually assaulted anyone. ... He says he feels sorry for his accusers, claiming they've been manipulated to lie.
4	Article	... In 2011, al Qaeda took Warren Weinstein hostage. ... his family paid money to his captors, ... the captors ... began demanding prisoners be released in exchange for Weinstein, ...
	BART-base	Al Qaeda took Warren Weinstein hostage in 2011, then paid money to his captors. ...
	BART-base*/MLLTs	After al Qaeda took Warren Weinstein hostage in 2011, his captors began demanding prisoners be released in exchange for Weinstein. ...

TABLE VII
GENERATIVE SUMMARY ASSESSMENT ON XSUM. WE USE THE SAME ILLUSTRATION APPROACH AS WITH TABLE VI.

No.	Source	Text
1	Article	... But Mr Farage, ... "We must be completely mad, as a country, to be giving people from Eastern Europe in-work benefits," he told BBC News. ...
	BART-base	UKIP Leader Nigel Farage has said the government should be "completely mad" to cut immigration from Eastern Europe by claiming in-work benefits.
	BART-base*/MLLTs	The UK must be "completely mad" to be giving migrants from Eastern Europe in-work benefits, former UKIP leader Nigel Farage has said.
2	Article	The test investigates whether people can detect if they are talking to machines or humans. ... The 65-year-old Turing Test is successfully passed if a computer is mistaken for a human more than 30% of the time during a series of five-minute keyboard conversations. On 7 June Eugene convinced 33% of the judges at the Royal Society in London that it was human. ... The event was organised by Reading University's School of Systems Engineering in partnership with RoboLaw, ...
	BART-base	An artificial intelligence system won a competition to see if it was man or woman during a series of five-minute keyboard conversations.
	BART-base*/MLLTs	An artificial intelligence system called Eugene has passed a Turing test, which was conducted at the University of Reading.
3	Article	In an interview marking five years as first minister, ... "If we see the Tory plans, ... " I do not know what the NHS in Wales would look like by 2020, or the education services, if we see those levels of cuts. ...
	BART-base	The first minister has said there will be no further cuts to the NHS until 2020 if the Conservatives win the general election.
	BART-base*/MLLTs	Wales could have no NHS by 2020 if Conservative plans for further cuts go ahead, the first minister has said.
4	Article	The closure of the tunnel near Linlithgow in West Lothian will mean rail services between Edinburgh and Glasgow will be severely hit. ... Engineers said the tunnel closure was necessary ... Train operator ScotRail has a dedicated website to advise on the disruption, which will mainly hit travel between Glasgow - or Stirling/Dunblane - and Edinburgh ... programme director of EGIP for Network Rail, said the tunnel work over the next six weeks ... He said: "We can't avoid the work in Winchburgh tunnel. ... Work on the ... M74 Motorway Improvements Project in Lanarkshire is already leading to disruption as traffic restrictions are in force. ...
	BART-base	Work on the main Edinburgh to Glasgow railway line is due to begin next month with the closure of the M74 motorway tunnel.
	BART-base*/MLLTs	Work on the Winchburgh railway tunnel is expected to cause major disruption to Glasgow-Edinburgh rail services over the next six weeks.