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/DailyMail (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 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.

Note that the CoreNLP by default uses Penn Treebank to tokenize document word sequence for building structure parsings. We build our article and summary datasets from these tokens as a result. 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 as described in Section I-B to facilitate the super token representation learning. Table I lists the preprocessed dataset sizes.

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 on 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 the byte-pair encoding method ([2]) 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 ([4]). 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 ([3]), 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 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:

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

Given the complex number elements $e_i^s \in E_s$ and $e_i^o \in E_o$, the element form of the function is then reduced as:

$$\begin{aligned} e_i^s \overline{e_i^o} &= (a_i^s + jb_i^s)(a_i^o - jb_i^o) \\ &= (a_i^s a_i^o + b_i^s b_i^o) + j(b_i^s a_i^o - a_i^s b_i^o). \end{aligned} \tag{2}$$

[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]https://huggingface.co/facebook/bart-base.

[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

As seen, the imaginary part $(b_i^s a_i^o - a_i^s b_i^o)$ possesses the antisymmetric property.

Referents and references may consist of multiple words. To simplify the learning task without loss of representational discriminative power, we adopt a super node representation learning similar to the super token representation learning described in Section I-B. In short, we center on the keyword of an entity or endophoric reference and apply a GNN to aggregate its neighboring (up to) n-gram 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 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: Table II summarizes the key training settings. We develop multi-GPU running procedure based on the reference runtime script[7]. We also adopt the training optimizer setting from the script. We initialize the backbone BART-base using pre-trained weights[8].

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.

C. 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:

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) In case of 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.

D. Human Evaluation User Interface

Fig. 1 illustrates the main user interface of the human evaluation tool. The model-generated summaries shown in "Summary 1" and "Summary 2" are randomly shuffled such that the numbers "1" and "2" in the captions have no fixed corresponding to models throughout the evaluation.

E. Qualitative Assessment

We further assess generated summaries to gain a better understanding of the plausible reasons behind the ROUGE, factuality evaluation scores and human evaluation results. Table V and Table VI compare summary samples generated from CNNDM and XSum respectively. The generated summaries demonstrate that the MLLTs-trained models improve the factual consistency of sub-phrases. For CNNDM, the MLLTs-trained model improves the sub-phrasal factual consistency which tends to be localized and extractive. But, the summaries for XSum exhibit different characteristics as the models often reorganize sparse concepts document-wise into new sub-phrases. Therefore, these summaries are more abstractive

[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.

[8]https://huggingface.co/facebook/bart-base.

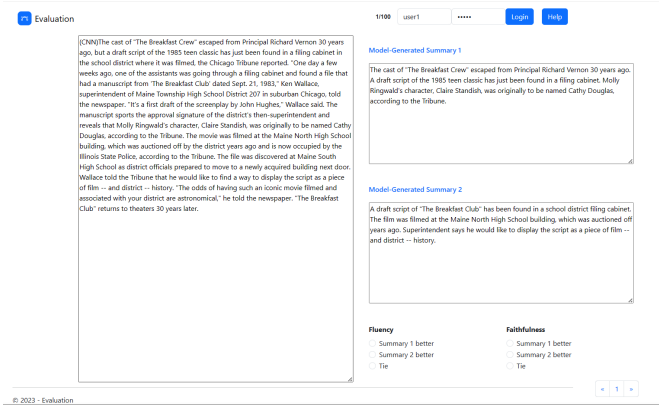

Fig. 1. An illustration of human evaluation user interface.

TABLE III
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 IV
DATASET VERSION AND LICENSE

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

and difficult to summarize, yet the MLLTs-trained model performs better than the baseline in extracting and reorganizing globally relevant concepts into concise and factually consistent summaries that display sound syntax and coherent semantics. Our observations and rationale might explain why the MLLTs-trained models score higher on three ROUGE metrics with CNNDM but only higher on R-L with XSum compared to the baseline. Meanwhile, endophoric reference problems are much less frequent, but some generated summaries may show evidence that our approach could help mitigate endophoric reference errors.

F. Key Packages

1) Versions: Table III 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 derived 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.

G. Datasets

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

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, Barcelona, Spain (Online). Association for Computational Linguistics.

[2] 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, Berlin, Germany. Association for Computational Linguistics.

[3] 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.

[4] 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.

[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.

TABLE V

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.

Source	**Text**
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. ...
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. ...
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.
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 VI
GENERATIVE SUMMARY ASSESSMENT ON XSUM. WE USE THE SAME ILLUSTRATION APPROACH AS WITH TABLE V.

Source	Text
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.
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.
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.
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.