Automatic Evaluation of Language Generation Technology Based on Structure Alignment

Katsuki Chousa and Tsutomu Hirao
NTT Communication Science Laboratories, NTT Corporation
{katsuki.chousa, tsutomu.hirao}@ntt.com

Abstract

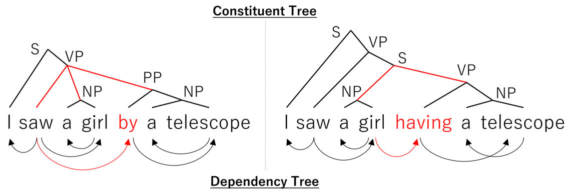
Language generation techniques require automatic evaluation to carry out efficient and reproducible experiments. While n-gram matching is standard, it fails to capture semantic equivalence with different wording. Recent methods have addressed this issue by using contextual embeddings from pre-trained language models to compute the similarity between reference and hypothesis. However, these methods frequently disregard the syntax of sentences, despite its crucial role in determining meaning, and thus assign unjustifiably high scores. This paper proposes an automatic evaluation metric that considers both the words in sentences and their syntactic structures. We integrate syntactic information into the recent embedding-based approach. Experimental results obtained from two NLP tasks show that our method is at least comparable to standard baselines.

1 Introduction

To promote the development of natural language generation (NLG) technologies, we need an automatic evaluation that enables efficient and reproducible experiments. Matching n-grams between reference and hypothesis is still the standard practice for automatic evaluation, e.g., BLEU (Papineni et al., 2002) for machine translation and ROUGE (Lin, 2004) for text summarization.

However, although recent NLG models can generate various sentences to convey a particular meaning, previous metrics seriously penalize a hypothesis that uses different words from those in the reference text. To address this issue, BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2022) use the contextual embeddings of words and sentences from pre-trained models, such as BERT (Devlin et al., 2019), to compute the similarity between reference and hypothesis.

Automatic evaluations based on similarities of embedding often ignore the syntax of sentences despite its importance in determining meaning. In Figure 1, we present an example of two sentences with nearly identical words but different meanings. The distinction between them lies in a single word. The sentence on the left-hand side can be derived from the one on the right-hand side by replacing *having* with *by*. However, when we focus on their structures, their different meanings become quite evident. This is because *by* is directly connected to *saw* in the left-hand sentence, while in the right-hand sentence, *having* (which occupies the same position as *by*) is linked to *girl*.



BLEU = 50.00, BERTScore = 98.01, Ours = 72.10

Figure 1: A sentence pair with nearly identical words but different meanings and syntax structures. Scores obtained by existing automatic evaluation methods are given at the bottom.

Unfortunately, both of the previous automatic evaluation metrics give unreasonably high scores because they focus on words rather than the syntactic structure. In fact, BLEU and BERTScore give high scores[1] to the two sentences in Figure 1. The syntactic structure can offer crucial evidence for disambiguation and capturing distinctions between sentences; however, its utility for embedding-based approaches remains not yet fully elucidated.

In this paper, we propose an automatic evaluation metric that not only focuses on words in sentences but also on their syntactic structure. After obtaining parse trees for the reference and hypothesis, we align the subtrees between reference and

[1] A BLEU score of 50 or higher is commonly interpreted as a high-quality, fluent translation.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7663–7670
January 19–24, 2025. ©2025 Association for Computational Linguistics

hypothesis and give a score based on the partial scores from the alignments. The experimental results based on two NLG tasks, data-to-text and machine translation, demonstrate that our method can achieve correlations that are comparable to or higher than those of previous methods.

2 Related Works

Despite the significant progress made in NLG, the development of automatic evaluation methods has not kept pace with the field's advances. Even today, the de facto standards for automatic evaluation are still BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which were proposed over 20 years ago. The main benefit of the n-gram matching-based approach is its simplicity, but it tends to underestimate paraphrases.

Several automatic evaluation methods using contextual word embeddings have been proposed to address the current limitations. BERTScore (Zhang et al., 2020) determines alignments between words in a reference and a hypothesis based on the cosine similarity of the embeddings. Subsequently, the scores between reference and hypothesis are computed based on the alignments.

COMET (Rei et al., 2020) computes scores using embeddings of source, reference, and hypothesis. Their model is made by adding a linear layer on top of a cross-lingual language model and then training on a regression task with manually annotated data.

As another approach, Prism (Thompson and Post, 2020) involves paraphrase generation using an encoder-decoder model. The evaluation score between reference and hypothesis is computed based on the scores of paraphrase generation from references to hypotheses or vice versa. They trained a multilingual translation model using a large-scale multilingual parallel corpus, and they applied this model as a zero-shot paraphraser.

On the other hand, some automatic evaluation methods explicitly leverage text structure. Liu and Gildea (2005) introduced metrics that rely on the agreement rates of subtrees of syntax trees between references and hypotheses. Anderson et al. (2016) proposed metrics using a graph-based semantic representation of references and hypotheses for image captioning tasks. They were able to obtain better correlations than conventional n-gram matching methods, but these methods are still limited in their ability to recognize paraphrasing. Their findings indicate that a text's structure can be advantageous in evaluation metrics; nevertheless, its effectiveness in the context of recent neural-based approaches requires further investigation.

3 Proposed Method

We extend the syntax-based approach by introducing contextual word embeddings. The overview of out method is shown in Figure 2. Initially, we decompose the syntax trees of the reference and hypothesis into subtrees and then align them by comparing their similarity using the embeddings. When calculating the similarity, we also consider the agreement of the syntactic features of subtrees. Then, we compute a score by normalizing the similarity to determine the final result.

Suppose we have a hypothesis $x = \langle x_1, \cdots, x_i \rangle$, reference $y = \langle y_1, \cdots, y_j \rangle$, and their contextual word embeddings, $\langle \boldsymbol{x}_1, \cdots, \boldsymbol{x}_i \rangle$ and $\langle \boldsymbol{y}_1, \cdots, \boldsymbol{y}_j \rangle$. First, we parse them into syntax trees, which are used to identify and obtain all of the subtrees to be aligned. We then obtain the vector representation of a subtree $s_m \in I_x$ from x as follows:

$$\boldsymbol{s}_m = \mathrm{max\_pooling}(\{\boldsymbol{x}_k \mid \boldsymbol{x}_k \in s_m\}), \quad (1)$$

where $\mathrm{max\_pooling}$ represents the element-wise max-pooling of the word embeddings in subtree s_m. We also conduct the same process to calculate the embedding for subtree $t_n \in I_y$ from y.

Next, while greedily determining the pair of subtrees with maximum similarity, we align subtrees from the reference and the hypothesis. We then calculate the score F based on the alignments between subtrees from the reference and the hypothesis: [2]

$$F = 2PR \,/\, (P + R), \quad (2)$$

$$P = \textstyle\sum_{s_m \in I_x} \max_{t_n \in I_y} \mathrm{sim}(s_m, t_n) \,/\, |I_x|, \quad (3)$$

$$R = \textstyle\sum_{t_n \in I_y} \max_{s_m \in I_x} \mathrm{sim}(s_m, t_n) \,/\, |I_y|. \quad (4)$$

To calculate similarity $\mathrm{sim}(s, t)$ between subtrees, we apply a filtering process based on the agreement of a syntactic feature of subtrees $\mathrm{feat}(s)$:

$$\mathrm{sim}(s,t) = \begin{cases} \frac{\boldsymbol{st}}{\|\boldsymbol{s}\|\|\boldsymbol{t}\|} & (\mathrm{feat}(s) = \mathrm{feat}(t)) \\ 0 & (\mathrm{otherwise}) \end{cases}. \quad (5)$$

This filtering uses syntactic features to calculate similarity and prevents low similarities from being

[2] The significance of spans would intuitively correlate with their depth in the syntactic trees. However, our experiments did not support this assumption. We present an ablation study on span weighting in Appendix B.

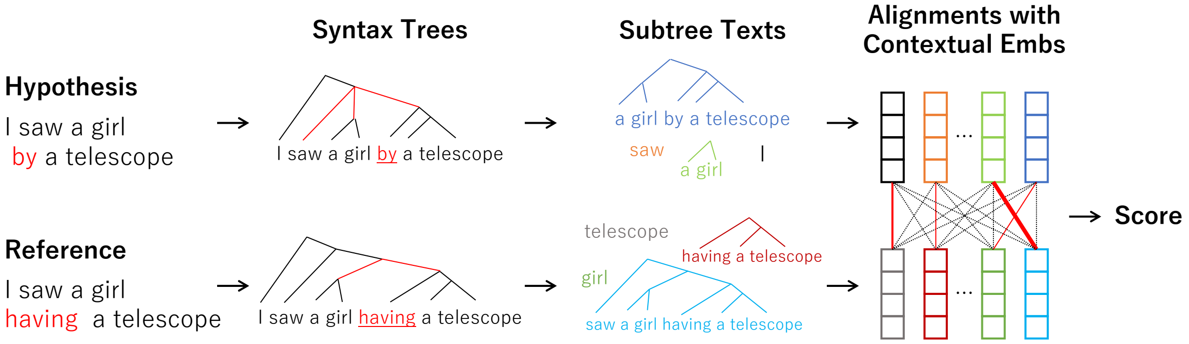

Syntax Trees
Subtree Texts
Alignments with Contextual Embs
Hypothesis
I saw a girl by a telescope
I saw a girl by a telescope
a girl by a telescope
saw
a girl
I
Reference
I saw a girl having a telescope
I saw a girl having a telescope
telescope
having a telescope
girl
saw a girl having a telescope
...
...
Score

Figure 2: Overview of the score's calculation process. Given the reference and hypothesis, we parse them into syntax trees and align subtrees based on the contextual embeddings.

assigned to non-corresponding alignment candidates, thus reducing noisy alignments.

4 Experiments

We conducted a meta-evaluation of the datasets used in two NLG tasks: the WMT Metric Shared Task and the WebNLG Challenge.

4.1 Tasks and Datasets

WMT Metrics Shared Task We employed the WMT20 Metric Shared Task (Mathur et al., 2020), which contains the outputs of the participating systems in the WMT News Translation Task as well as human ratings for each output ranging from 0 to 100 in terms of adequacy. Since our method needs a syntactic parser, we selected source and target language pairs whose target side is English.

WebNLG Challenge We used the WebNLG Challenge 2020 (Castro Ferreira et al., 2020), along with human ratings, for RDF-to-text tasks. It consists of 178 instances, including generated sentences from 15 systems. Ratings range from 0 to 100 based on these criteria: Data Coverage (Cov), Relevance (Rel), Correctness (Cor), Text Structure (Str), and Fluency (Flu). The first three criteria evaluate how well the outputs reflect the relationship between subject, predicate, and object in the RDF, while the remaining two criteria assess the outputs for being grammatical, well-structured, and logically sound.

4.2 Settings

We used spaCy[3] with RoBERTa to perform dependency parsing and Berkeley Neural Parser (Kitaev and Klein, 2018) for constituency parsing. To obtain the vector representation of subtrees, we used DeBERTa-v2-xxlarge (He et al., 2021) for the constituent tree and SpanBERT (Joshi et al., 2020) for the dependency tree.[4] As the subtree for the dependency tree, we specified (1) all single nodes, (2) paths from each node to its leaves, and (3) sets of each node and its descendants.[5] For the syntactic feature in Eq.(5), we used the head word and phrase label of the subtree, since these features play the central role of a subtree in terms of syntax. The head of the constituent tree was not directly modeled; therefore, we applied the head rules of Collins (2003) to determine the head.

As baseline methods, we employed BLUE (Papineni et al., 2002), Prism (Thompson and Post, 2020), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020), all of which are frequently used for evaluation of text generation tasks. We evaluated automatic evaluation methods based on system-level correlations with Pearson's correlation coefficient for human ratings by following the WMT's official settings. To compute the correlation, we standardized the ratings of the human evaluation as reference scores by following the WMT.[6]

4.3 Results

The results obtained from the WMT (Table 1) demonstrate that our methods achieve at least comparable performance to baseline methods on all language pairs. Ours (dep) + head filtering achieved the best correlation on cs-en and ru-en, and Ours (con) + head filtering obtained the best on ja-en and km-en. Furthermore, Ours (con) + head fil-

[3] `https://spacy.io/models/en#en_core_web_trf`

[4] In preliminary experiments, we found that this setting worked best; however, the performance differences between these models are slight.

[5] Refer to Appendix C for details.

[6] For the WMT, we used **mt-metrics-eval** to obtain the dataset and evaluate the systems. `https://github.com/google-research/mt-metrics-eval`

Language Pair (# of systems)	cs-en (10)	de-en (9)	iu-en (9)	ja-en (7)	km-en (7)	pl-en (13)	ps-en (6)	ru-en (10)	zh-en (13)
sentBLEU	.800	.786	.469	.851	.969	.284	.888	.833	**.950**
COMET	.694	.773	.605	.828	.971	.345	.941	.836	.931
Prism	.720	.775	.616	.869	.950	.269	**.966**	.839	.945
BERTScore	.743	**.794**	**.643**	.856	.948	.293	.938	.838	.942
Ours (con)	.638	.634	.468	.674	.959	.307	.928	.829	.942
+ label filtering	.693	.653	.545	.712	.946	.266	.914	.846	.945
+ head filtering	.783	.741	.486	**.886**	**.983**	.193	.917	.882	.924
Ours (dep)	.776	.778	.383	.833	.978	**.488**	.882	.860	.939
+ head filtering	**.824**	.754	.477	.862	.977	.269	.889	**.893**	.904

Table 1: System-level Pearson Correlation on the WMT20: The languages are abbreviated into ISO 639-1 codes. Bold values indicate the best correlations across language pairs. Ours (dep) and Ours (con) denote our methods using dependency and constituent trees.

System	Cor	Cov	Flu	Rel	Str
sentBLEU	.650	.534	.907	.609	.912
Prism	**.913**	**.829**	.897	**.896**	.893
BERTScore	.857	.769	.926	.836	.912
Ours (con, head)	.768	.679	**.935**	.744	**.922**
Ours (dep, head)	.660	.535	.897	.658	.893

Table 2: System-level Pearson Correlations on WebNLG. Bold values indicate the best correlations across criteria.

tering outperformed the recent neural-based baselines, COMET, Prism, and BERTScore, in four of nine language pairs: cs-en, ja-en, km-en, and ru-en. These results imply that the subtree alignments using word embeddings are promising.

The results also suggest the effectiveness of filtering with syntactic features of the subtree. In most cases, our methods with such filtering obtained better scores than those without it. In particular, when employing constituent trees, we found that using the head as the filtering criterion outperformed using the label on six language pairs. The results indicate that filtering with the syntactic features, especially the head, is advantageous. Although there were no significant differences between constituency and dependency trees, the former outperformed the latter on five language pairs. We also show the representative examples on the WMT20 Ja-En in Appendix A.

Table 2 shows the results obtained from the WebNLG. We experimented with our method using head filtering since this achieved the best performance on the WMT. Again, our methods performed comparably to the baselines. In particular, Ours (con, head) obtained the best correlations in Flu and Str, which are related to the structures of sentences. However, our methods were outperformed by the embedding-based baseline, BERTScore, in Cor, Cov, and Rel, which involve the relationships in the RDF. The results suggest the effectiveness of syntactic structure when calculating sentence similarities. When comparing Ours (con) with Ours (dep), we found that the former obtained better correlations than the latter.

These results from the two datasets demonstrate that our methods' performances are at least comparable to the baselines. Unlike Prism and COMET, which require large multilingual parallel corpora or human-annotated evaluation data, our methods require only monolingual pre-trained language models, which are readily available, to obtain vector representation of words. This is a significant advantage of our methods over Prism and COMET.

On the other hand, BERTScore sometimes obtains better correlations against ours. The major difference between BERTScore and our methods lies in the lexical unit used for alignment: the former uses tokens, while the latter employs subtrees. The performance degradation may reflect the fact that alignment sometimes fails to work effectively when targeting subtrees larger than the token units. However, unlike previous methods, our approach can emphasize differences in the syntactic structure between references and hypotheses. We believe the ability to achieve comparable performance to that of a previous approach with different methodologies holds significant value in automatic evaluation.

4.4 Partial Correlations Between Metrics

To better understand the characteristics of our proposed method, we calculated partial correlations between different automatic evaluation metrics while controlling for human scores. Figure 3 presents the results from the WMT20 Ja-En dataset. We observed notably high partial correlations, around 0.8,

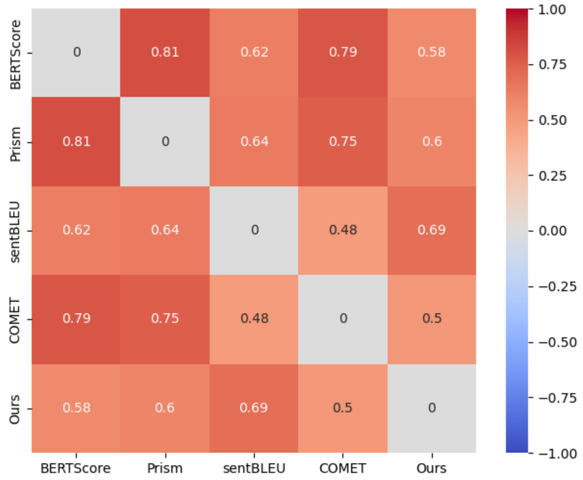


Figure 3: Partial Correlations Between Metrics on the WMT20 Ja-En. Ours denotes our method using constituent trees with head filtering.

among the embedding-based methods (BERTScore, COMET, and Prism). This suggests that these metrics share similar evaluation characteristics due to their dependence on word embeddings.

In contrast, the partial correlations between our method and each embedding-based method are relatively lower, ranging from 0.5 to 0.6. This suggests that our approach, which leverages syntactic span information through tree structures, captures different aspects of translation quality compared to purely embedding-based methods, while still maintaining comparable performance levels in terms of correlation with human judgments. Interestingly, our method's partial correlation with sentBLEU is 0.69, which is relatively higher than its correlations with embedding-based methods. This higher correlation likely stems from certain similarities between n-grams and syntactic spans in capturing local structural information.

Recently, embedding-based methods have been regarded as definitive metrics for automatically evaluating NLG technologies. However, they exhibited low correlations with human judgments when applied to new domain data on which their methods were not trained (Kocmi et al., 2024). As demonstrated by the partial correlations, our method exhibits distinct characteristics from conventional embedding-based approaches while achieving comparable correlations with human evaluation. This finding suggests that integrating our method with existing approaches could lead to a more comprehensive and robust framework for automatic evaluation. The complementary nature of our syntax-based approach alongside embedding-based methods holds promise for providing a more reliable assessment of the quality of the generated text.

5 Conclusions

In this paper, we proposed an automatic evaluation metric for NLG based on the alignment between subtrees of sentences. We conducted meta-evaluations on two datasets: the WMT Metrics Shared Task and the WebNLG Challenge. The results demonstrate that although our method does not require training data as supervision, it is comparable to the baseline methods; of sentBLEU, Prism, COMET, and BERTScore. This finding suggests that incorporating syntactic information into the recent embedding-based approach is beneficial for automatic evaluation.

Limitations

While our method does not require large multilingual parallel corpora or human-annotated evaluation data, it requires syntactic parsers and pretrained language models for the reference-side language. A notable limitation of our method is its reliance on the availability of these resources. Unfortunately, such resources are not universally available across all languages, and thus our method's applicability remains limited to those languages possessing the above resources.

Additionally, the performance of the parsers will inevitably vary depending on the language. In cases where the parser's accuracy is low, the performance of our method may also be degraded.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, the Netherlands.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, Michigan. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202, Berlin, Germany. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Examples

We show pairs of references and hypotheses in Table 3. The first half of Table 3 shows examples where the human and COMET ranks are disagreed. We observed that our method can correctly penalize sentence pairs with incorrect modification relationships (as seen in No. 2 and No. 3). Additionally, our method demonstrates robustness in evaluating sentence pairs where the tokenization of proper nouns differs between the reference and candidate (shown in No. 1). These observations imply that the subtree alignments based on the syntax features are beneficial for automatic evaluation.

The second half of Table 3 shows examples where the human and our method ranks are disagreed. In these cases, our method assigned incorrect ratings due to the structure differences between references and candidates. However, these results indicate that our method can properly capture structural differences.

B Ablation Study of Weighting

We initially hypothesized that the significance of a subtree would correlate with its depth in the syntactic trees because subtrees closer to the root would contain key information about the sentence. If this assumption were correct, it would be rational to assign higher scores to the upper subtrees, and our method could carry out this approach by weighting subtrees. We experimented with the weight for a subtree s as:

$$w(s) = \log\left(\mathrm{len}(s)\right), \tag{6}$$

where $\mathrm{len}(s)$ denotes the token length of the subtree s. To calculate the weighted scores, we modified P (Eq.(3)) and R (Eq.(4)) as follows:

$$P = \frac{\sum_{s_m \in I_x} w(s_m) \cdot \max_{t_n \in I_y} \mathrm{sim}(s_m, t_n)}{\sum_{s_m \in I_x} w(s_m)}, \tag{7}$$

$$R = \frac{\sum_{t_n \in I_y} w(s_m) \cdot \max_{s_m \in I_x} \mathrm{sim}(s_m, t_n)}{\sum_{t_n \in I_y} w(s_m)}. \tag{8}$$

To validate this hypothesis, we compared our method with and without weighting using constituent trees. Table 4 shows the correlation coefficients derived from the WMT20. These results illustrate that the correlation coefficients of the "weighting" approach against evaluations without weighting remain almost the same or even degrade. Consequently, we decided not to employ such weighting in this paper.

C Subtrees for Dependency Trees

In this paper, we used three types of subtrees of the dependency trees, as shown in Figure 4: (1) We first used all single nodes as subtrees, which are essentially word lists because each node in a dependency tree corresponds to a word in a sentence. We used this subtree to consider word-level alignment, such as BERTScore (Zhang et al., 2020). (2) Next, we identified paths from each node to its leaves as subtrees. For example, the word *girl* has two leaves, and there are two paths from *girl*: *girl→a* and *girl→having→telescope→a*. Such a path is widely used as a feature of a dependency tree (Liu and Gildea, 2005; Roth and Lapata, 2016; Bunescu and Mooney, 2005). (3) Finally, we listed parts of the whole tree that include a node and all of its descendants as subtrees.

No.		Reference and Hypothesis	Human	Ours	COMET
1.	x: $\hat{x}$:	Tsurugakehi defeats Hokuriku for Hokushinetsu ticket Tsuruga kehi defeats Hokuriku for Hokushinetsu ticket	218	214	1748
2.	x: $\hat{x}$:	However "government-prompted price cuts" will distort the market and it is difficult to think that it will promote healthy competition. However, I do not think that"official price cuts"will distort the market and encourage healthy competition.	4116	4126	2138
3.	x: $\hat{x}$:	The roughly 400 participants included residents and employees of 20 organizations including the prefecture and the three towns. Approximately 400 people, including staff and residents of 20 organizations such as roads and 3 towns, participated.	4416	4409	2561
4.	x: $\hat{x}$:	Conversely, since he was able to gain high points in the games of the first half his schedule likely becomes more flexible. Rather, the high points gained during the first half should have made it easier for them to adjust their schedule for the upcoming tournament.	4089	7412	4112
5.	x: $\hat{x}$:	Meanwhile, Usui attracted attention with their defeat of the powerful Fukui Industrial. Meanwhile, Hanesui defeated the powerful Fukui Institute of Technology and received attention from this tournament.	4163	8301	4145
6.	x: $\hat{x}$:	Targeted hospitals will be asked to consider actions such as abolishing or moving parts of some departments to other hospitals. The target hospitals will be asked to consider abolishing them or moving some departments to other hospitals.	771	2189	781

Table 3: Examples of Reference, Candidate, and Rankings on the WMT Ja-En: x and $\hat{x}$ denote gold reference and hypothesis of MT systems. The rankings are assigned based on the scores of Human, Our Method, and COMET.

Method	cs-en	de-en	iu-en	ja-en	km-en	pl-en	ps-en	ru-en
Ours (con)	.763	.733	.488	.889	.983	.190	.918	.884
+ weighting	.785	.739	.442	.863	.980	.169	.900	.872

Table 4: Ablation study of weighting using system-level Pearson Correlation on the WMT20 Metrics Shared Task.

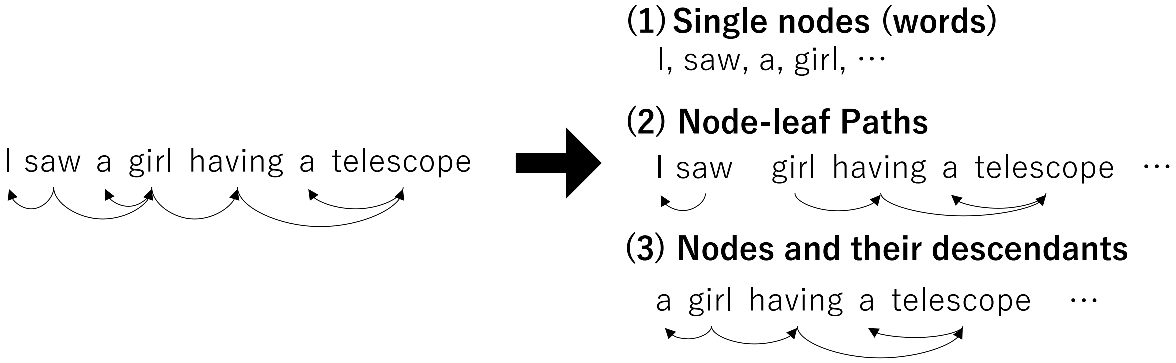

I saw a girl having a telescope
(1) Single nodes (words)
I, saw, a, girl, ⋯
(2) Node-leaf Paths
I saw girl having a telescope ⋯
(3) Nodes and their descendants
a girl having a telescope ⋯

Figure 4: Dependency Subtree