DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence

Wei Zhao[12] **Michael Strube**[1] **Steffen Eger**[3]
[1]Heidelberg Institute for Theoretical Studies [2]Technische Universität Darmstadt
www.h-its.org/research/nlp/
[3]NLLG, Faculty of Technology, Bielefeld University
{wei.zhao, michael.strube}@h-its.org
steffen.eger@uni-bielefeld.de
nl2g.github.io

Abstract

Recently, there has been a growing interest in designing text generation systems from a discourse coherence perspective, e.g., modeling the interdependence between sentences. Still, recent BERT-based evaluation metrics are weak in recognizing coherence, and thus are not reliable in a way to spot the discourse-level improvements of those text generation systems. In this work, we introduce DiscoScore, a parametrized discourse metric, which uses BERT to model discourse coherence from different perspectives, driven by Centering theory. Our experiments encompass 16 non-discourse and discourse metrics, including DiscoScore and popular coherence models, evaluated on summarization and document-level machine translation (MT). We find that (i) the majority of BERT-based metrics correlate much worse with human rated coherence than early discourse metrics, invented a decade ago; (ii) the recent state-of-the-art BARTScore is weak when operated at system level—which is particularly problematic as systems are typically compared in this manner. DiscoScore, in contrast, achieves strong system-level correlation with human ratings, not only in coherence but also in factual consistency and other aspects, and surpasses BARTScore by over 10 correlation points on average. Further, aiming to understand DiscoScore, we provide justifications to the importance of discourse coherence for evaluation metrics, and explain the superiority of one variant over another. Our code is available at https://github.com/AIPHES/DiscoScore.

1 Introduction

In discourse, coherence refers to the continuity of semantics in text. Often, discourse relations and lexical cohesion devices, such as repetition and coreference, are employed to connect text spans, aiming to ensure text coherence. Popular theories in the linguistics community on discourse were provided by Grosz et al. (1995) and Mann and Thompson (1988). They formulate coherence through the lens of readers' focus of attention, and rhetorical discourse structures over sentences. Later on, coherence models as computational approaches of these theories emerged to judge text coherence in discourse tasks such as sentence ordering and essay scoring (Barzilay and Lapata, 2008; Lin et al., 2011; Guinaudeau and Strube, 2013).

While humans also often use text planning at discourse level prior to writing and speaking, up until recently, the majority of natural language generation (NLG) systems, be it text summarization or document-level MT, has performed sequential word prediction without considering text coherence. For instance, MT systems mostly do not model the interdependence between sentences and translate a document at sentence level, and thus produce many incoherent elements such as coreference mistakes in system outputs (Maruf et al., 2021). Only more recently has there been a surge of interest towards discourse based summarization and MT systems, aiming to model inter-sentence context, with a focus on pronominal anaphora (Voita et al., 2018; Liu et al., 2021) and discouse relations (Miculicich et al., 2018; Xu et al., 2020).

However, there appears a mismatch between discourse based NLG systems and non-discourse NLG evaluation metrics such as MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2020) which have recently become popular for MT and summarization evaluation. As these metrics base their judgment on semantic similarity (and lexical overlap (Kaster et al., 2021)) between hypotheses and references—which by design does not target text coherence—it is not surprising that they do not correlate well with human rated coherence (Fabbri et al., 2021; Yuan et al., 2021; Sai et al., 2021). Recently, BARTScore (Yuan et al., 2021) receives increasingly attention, which uses sequence-to-sequence language models to measure the likeli-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3865–3883
May 2-6, 2023

Hypothesis

Chelsea have made an offer for FC Tokyo forward Yoshinori Muto. The 22-year-old will join Chelsea 's Dutch partner club Vitesse Arnhem on loan next season if he completes a move to Stamford Bridge. Chelsea signed a £200million sponsorship deal with Japanese company Yokohama Rubber in February.

Reference

Naoki Ogane says that Chelsea have made an offer for Yoshinori Muto. The 22-year-old forward has one goal in 11 games for Japan. Muto admits that it is an 'honour' to receive an offer from the Blues. Chelsea have signed a £200m sponsorship deal with Yokohama Rubber. Muto graduated from university with an economics degree two weeks ago. He would become the first Japanese player to sign for Chelsea.

	t_1	t_2	t_3	t_4	t_5	...
Chelsea	1	0	0	0	0	1
offer	0	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) FocusDiff

	s_1	s_2	s_3
s_1	0	1	0.5
s_2	0	0	1
s_3	0	0	0

(b) SentGraph

Figure 1: Sample hypothesis and reference from SUMEval. Each focus[1] is marked in a different color, corresponding to multiple tokens as instances of a focus. Foci shared in Hypothesis and Reference are marked in the same color. (a)+(b) are adjacency matrices used to model focus-based coherence for Hypothesis; for simplicity, adjacency matrices for Reference are omitted. FocusDiff and SentGraph are the variants of DiscoScore. For FocusDiff, we use (a) to depict the relations between foci and tokens, reflecting focus frequency. For SentGraph, we use (b) to depict the interdependence between sentences according to the number of foci shared between sentences and the distance between sentences.

hood that hypothesis and reference are paraphrases, and that cannot contrast text pairs at discourse level.

In this work, we fill the gap of missing discourse metrics in MT and summarization evaluation, particularly in reference-based evaluation scenarios. We introduce DiscoScore, a parametrized discourse metric, which uses BERT to model discourse coherence through the lens of readers' focus, driven by Centering theory (Grosz et al., 1995). The DiscoScore variants can be distinguished in how we use *focus*—see Figure 1: (i) we model focus frequency and semantics, and compare their difference between hypothesis and reference and (ii) we use focus transitions to model the interdependence between sentences. Building upon this, we present a simple graph-based approach to compare hypothesis with reference.

We compare DiscoScore with a range of baselines, including discourse and non-discourse metrics, and coherence models on summarization and document-level MT datasets. Our contributions and findings are summarized as follows:

- Recent BERT-based metrics and the state-of-the-art BARTScore (Yuan et al., 2021) are all weak in system-level correlation with human ratings, not only in coherence but also in other aspects such as factual consistency. Most of them are even worse than very early discourse metrics, RC and LC (Wong and Kit, 2012)—which require neither source texts nor references and use discourse features to predict hypothesis coherence.
- DiscoScore strongly correlates with human rated coherence and many other aspects, over 10 points (on average across aspects) better than BARTScore and two strong baselines RC and LC in the single and multi-references settings. This indicates that either leveraging contextualized encoders or finding discourse features is not sufficient, suggesting to combine both as DiscoScore does.
- We demonstrate the importance of including discourse signals in the assessment of system outputs, as the discourse features derived from DiscoScore can strongly separate hypothesis from reference. Further, we show that the more discriminative these features are, the better the metrics perform, which allows for interpreting the performance gaps between the variants of DisoScore.
- We investigate two focus choices popular in the discourse community, i.e., noun (Elsner and Charniak, 2011) and semantic entity (Mesgar and Strube, 2016). Our results show that entity as focus is not always helpful, but when it helps, the gain is big.

2 Related work

Evaluation Metrics. Traditional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) measure lexical n-gram overlap between a hypothesis and a human reference. As they fail to measure semantic similarity in the absence of lexical overlap, several metrics have been proposed to overcome this issue, which carry out soft lexical matching with static word embeddings (Ng and Abrecht, 2015) and synonym matching (Lavie and Agarwal, 2007). However, none of those metrics

[1]The formal definition of focusing in discourse is given on two levels (Grosz et al., 1977): (i) readers are said to be *globally* focusing on a set of entities relevant to the overall discourse, and (ii) readers focus on a particular entity that an utterance *locally* concerns most. Section 3 elaborates on focus as a key ingredient of DiscoScore.

can properly judge text coherence (Kryscinski et al., 2019; Zhu and Bhat, 2020).

Recently, a class of novel metrics based on BERT (Devlin et al., 2019) has received a surge of attention, as they correlate strongly with human judgment of text quality in both reference-based and reference-free scenarios (Zhao et al., 2019; Zhang et al., 2020; Sellam et al., 2020; Rei et al., 2020; Gao et al., 2020; Thompson and Post, 2020; Zhao et al., 2020; Pu et al., 2021; Chen et al., 2021). While strong at sentence-level, these metrics are weak in recognizing coherence in inter-sentence contexts (just like BLEU and ROUGE), as BERT and the majority of BERT variants[2] that these metrics build on only capture discourse phenomena to a certain extent (Koto et al., 2021; Laban et al., 2021; Beyer et al., 2021). Thus, they are not suitable for evaluating long texts as in document-level MT evaluation. Works that either (i) average sentence-level evaluation scores as document score or (ii) assign a score to the concatenation of sentences within a document (Xiong et al., 2019; Liu et al., 2020; Saunders et al., 2020) do not factor interdependence between sentences into a document score, e.g., do not explicitly punish incoherent elements, thus are also inadequate.

Several attempts have been made towards discourse metrics in MT evaluation. Wong and Kit (2012); Gong et al. (2015); Cartoni et al. (2018) use the frequency of lexical cohesion devices (e.g., word repetition) over sentences to predict coherence of hypothesis translations, while Guzmán et al. (2014) and Joty et al. (2017) suggest to compare the difference of rhetorical structures between hypothesis and reference translations. Recently, Jiang et al. (2021) measure the inconsistency between hypothesis and reference translations in several aspects such as verb tense and named entities. However, these metrics do not leverage strong contextualized encoders, as has been shown to be a key ingredient for recent success of BERT-based metrics. Most recently, BARTScore (Yuan et al., 2021) uses sequence-to-sequence pretrained language models such as BART (Lewis et al., 2020) to measure how likely hypothesis and reference are paraphrased according to the probability of one given the other. While BARTScore constitutes the recent state-of-the-art in sentence-level correlation with human ratings in several aspects (incl. discourse), we find that (i) it performs still poorly at system level—which is particularly problematic as systems are typically compared in this manner. (ii) As based on a 'blackbox' language model, it cannot offer insights towards how it models coherence and what discourse phenomena it does (not) capture.

Coherence Models. In discourse, there have been many computational models (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013; Pitler and Nenkova, 2008; Lin et al., 2011) for text coherence assessment, the majority of which differ in *regularities* that they use to distinguish coherent from incoherent text, driven by different linguistic theories, *v.i.z.*, a pattern of (i) focus transitions in adjacent sentences (Grosz et al., 1995) and (ii) text organization regarding discourse relations over sentences (Mann and Thompson, 1988). For instance, Barzilay and Lapata (2008) and Guinaudeau and Strube (2013) use the distribution of entity transitions over sentences to predict text coherence, while Pitler and Nenkova (2008) and Lin et al. (2011) suggest to produce discourse relations over sentences with a discourse parser, showing that the relations are indicative of text coherence. In the last few years, neural coherence models have been explored. Popular examples are Tien Nguyen and Joty (2017), Mesgar and Strube (2018) and Moon et al. (2019). As they and the recent state-of-the-art (Mesgar et al., 2021) all have been trained on text readability datasets, with readability labels as supervision, they may suffer issues of domain shift when applied to MT and summarization evaluation. More importantly, they judge hypothesis coherence in the absence of reference, thus are not sufficient for reference-based evaluation. Our experiments involve two popular, unsupervised coherence models, entity graph (Guinaudeau and Strube, 2013) and lexical graph (Mesgar and Strube, 2016) treated as discourse metrics with the advantages on robustness (Lai and Tetreault, 2018).

Discourse Test Sets. Apart from evaluation metrics, there have been several discourse-focused test sets proposed to compare NLG systems, most of which have been studied in MT evaluation. For instance, the DiscoMT15 shared task (Hardmeier et al., 2015) compares MT systems, not based on translation adequacy but on the accuracy of pronoun translation for English-to-French, i.e., counting the number of correctly translated pronouns, given the annotated ones in reference. Bawden

[2]Recently, several discourse BERT variants such as Conpono (Iter et al., 2020) have been proposed, but they are not always helpful for evaluation metrics—see Table 2 (appendix).

et al. (2018) extend this by labeling both anaphoric pronouns and lexical cohesion devices on test sets, while Voita et al. (2018) construct English-to-Russian test sets focusing on deixis, ellipsis and lexical cohesion. Guillou et al. (2018); Lopes et al. (2020) construct English-to-German and English-to-French test sets targeting pronouns. While reliable, these test sets involve costly manual annotation, thus are limited to few language pairs.

In this work, we introduce DiscoScore to judge system outputs, which uses BERT to model readers' focus within hypothesis and reference, and thus clearly outlines the discourse phenomena being captured, serving as low-cost alternatives to discourse test sets for comparing discourse based NLG systems. More prominently, we derive discourse features from DiscoScore, which we use to understand the importance of discourse for evaluation metrics, and explain why one metric is superior to another. This parallels recent effort towards explainability for non-discourse evaluation metrics (Kaster et al., 2021; Fomicheva et al., 2021). Finally, we show that simple features can be indicative of the superiority of a metric, which fosters research towards finding insightful features with domain expertise and building upon these insights to design high-quality metrics.

3 Our Approach

In the following, we elaborate on the two variants of DiscoScore, FocusDiff and SentGraph, which we refer to as DS-FOCUS and DS-SENT.

Focus Difference. In discourse, there have been many corpus-based studies towards modeling focus transitions over sentences, showing that focus transition patterns are indicative of text coherence (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013). When reading a document, readers may have multiple *focus of attention*,

each associated to a group of expressions: (i) referring expressions such as pronouns and (ii) semantically related elements such as [*Berlin, capital*].

Here, we assume two focus based conditions that a coherent hypothesis should meet in reference-based evaluation scenarios:

- A large number of focus overlaps between a hypothesis and a reference.
- Each focus overlap is nearly identical in terms of semantics and frequency, where frequency shows how often a focus is mentioned in a hypothesis or in a reference.

In the following, we present focus modeling towards semantics and frequency, according to which we compare hypothesis with reference.

For a hypothesis, we introduce a bipartite graph $\mathcal{G}^{\text{hyp}} = (\mathcal{V}, \mathcal{S}, \mathbf{A}^{\text{hyp}})$, where $\mathcal{V}$ and $\mathcal{S}$ are two sets of vertices corresponding to a set of foci and all tokens (per occurrence a word is a separate token) within a hypothesis. Let $\mathbf{A} = \{0,1\}^{n \times m}$ be an adjacency matrix where n and m are the number of foci and tokens respectively, and A_{ij} equals 1 if and only if the i-th focus associates to the j-th token. Let $\mathbf{F}^{\text{hyp}} \in \mathbb{R}^{n \times d}$ be a matrix of focus embeddings and $\mathbf{Z}^{\text{hyp}} \in \mathbb{R}^{m \times d}$ be a matrix of contextualized token embeddings with d as the embedding size. Similarly, we use notation $\mathcal{G}^{\text{ref}}$, $\mathbf{F}^{\text{ref}}$ and $\mathbf{Z}^{\text{ref}}$ for a human reference.

We use contextualized encoders such as BERT to produce token embeddings $\mathbf{Z}^{\text{hyp}}$ and $\mathbf{Z}^{\text{ref}}$. We use a simple approach to model both semantics and frequency of a focus. That is, we assign per focus v an embedding by summing token embeddings that a focus is associated to:

$$\mathbf{F}_v^{\text{hyp}} = \sum_{u \in \mathcal{N}(v)} \mathbf{Z}_u^{\text{hyp}}, \ \mathbf{F}_v^{\text{ref}} = \sum_{u \in \mathcal{N}(v)} \mathbf{Z}_u^{\text{ref}} \quad (1)$$

where $\mathcal{N}(v)$ is a set of tokens (e.g., a group of semantically related expressions) associated with a focus v. In matrix notation, we rewrite Eq. (1) to $\mathbf{F}^{\text{hyp}} = \mathbf{A}^{\text{hyp}}\mathbf{Z}^{\text{hyp}}$, similarly for $\mathbf{F}^{\text{ref}}$.

Next, we measure the distance between a common set of foci Ω in a hypothesis and reference pair based on their embeddings:

$$\text{DS-F}\textsc{ocus}(\text{hyp}, \text{ref}) = \frac{1}{\text{N}} \sum_{\text{u} \in \Omega} \|\mathbf{F}_{\text{u}}^{\text{hyp}} - \mathbf{F}_{\text{u}}^{\text{ref}}\| \quad (2)$$

where DS-FOCUS is scaled down by the factor of N, the number of foci in hypothesis.

Sentence Graph. Few contextualized encoders can produce high-quality sentence embeddings in the document context, as they do not model interdependence between sentences. According to Centering theory (Grosz et al., 1995), two sentences are marked continuous in meaning when they share at least one focus, on the one hand; one marks a meaning shift for two sentences when no focus appears in common, on the other hand. From this, one can aggregate sentence embeddings for which corresponding sentences are considered continuous. In the following, we present a graph-based approach to do so.

For a hypothesis[3], let $\mathbf{S}^{\text{hyp}} \in \mathbb{R}^{n \times d}$ be a matrix of sentence embeddings with n and d as the number of sentences and the embedding size. We introduce a graph $\mathcal{G}^{\text{hyp}} = (\mathcal{V}, \mathbf{A}^{\text{hyp}})$ where $\mathcal{V}$ is a set of sentences and $\mathbf{A}^{\text{hyp}}$ is an adjacency matrix weighted according to the number of foci shared between sentences and the distance between sentences as listed below to depict two variants of $\mathbf{A}^{\text{hyp}}$:

- unweighted: $\mathbf{A}^{\text{hyp}}_{ij} = 1/(j-i)$ if the i-th and the j-th sentences have at least one focus in common (otherwise 0), where $j-i$ denotes the distance between two sentences and $\mathbf{A}^{\text{hyp}}_{ij} = 0$ when $j \leq i$.
- weighted: $\mathbf{A}^{\text{hyp}}_{ij} = a/(j-i)$, where a is the number of foci shared in the i-th and the j-th sentences, with the same constraints on j and i as above.

Analyses by Guinaudeau and Strube (2013) indicate that global statistics (e.g., average) over such adjacency matrices can distinguish incoherent from coherent text to some degree. Here we depict adjacency matrices as a form of sentence connectivity derived from focus transitions over sentences. We use them to aggregate sentence embeddings from hypothesis and from reference:

$$\hat{\mathbf{S}}^{\text{hyp}} = (\mathbf{A}^{\text{hyp}} + \mathbf{I})\mathbf{S}^{\text{hyp}}, \ \hat{\mathbf{S}}^{\text{ref}} = (\mathbf{A}^{\text{ref}} + \mathbf{I})\mathbf{S}^{\text{ref}}$$

where $\mathbf{I}$ is an identity matrix that adds a self-loop to a graph so as to include self-embeddings when updating them.

Next, we derive per graph an embedding with simple statistics from $\hat{\mathbf{S}}^{\text{hyp}}$ and $\hat{\mathbf{S}}^{\text{ref}}$, i.e., the concatenation of mean-max-min-sum embeddings. Finally, we compute the cosine similarity between two graph-level embeddings:

$$\text{DS-SENT}(\text{hyp}, \text{ref}) = \text{cosine}(\mathcal{G}^{\text{hyp}}, \mathcal{G}^{\text{ref}}) \quad (3)$$

Choice of Focus. In discourse, often four popular choices are used to describe a focus: (i) a noun that heads a NP (Barzilay and Lapata, 2008), (ii) a noun (Elsner and Charniak, 2011), (iii) a coreferent entity associated with a set of referring expressions (Guinaudeau and Strube, 2013) and (iv) a semantic entity associated with a set of lexical related words (Mesgar and Strube, 2016).

In this work, we investigate two focus choices: noun (NN) and semantic entity (Entity). Linguistically speaking, the latter is a lexical cohesion device in the form of repetition. From this, NN as focus yields few useful coherence signals but a lot of noise, while Entity as focus uses 'signal compression' by means of aggregation to produce better signals. To produce entities, we first extract all nouns in hypothesis (or reference), and aggregate them into different semantic entities if their cosine similarities based on Dep2Vec word embeddings (Levy and Goldberg, 2014) is greater than a threshold—assuming that nouns with high similarity refer to the same semantic entity.

4 Experiments

4.1 Evaluation Metrics

In the following, we list all of the evaluation metrics, and elaborate on them in Appendix A.1.

Non-discourse Metrics. We consider BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), SBERT (Reimers and Gurevych, 2019), S^3-pyr (Peyrard et al., 2017), BLEURT (Sellam et al., 2020), BARTScore (Yuan et al., 2021), PRISM (Thompson and Post, 2020).

Discourse Metrics. We consider RC and LC (Wong and Kit, 2012) and Lexical Chain (Gong et al., 2015). We consider two coherence models, EntityGraph (Guinaudeau and Strube, 2013) and LexicalGraph (Mesgar and Strube, 2016), and treat them as discourse metrics.

DiscoScore. DS-FOCUS can be parameterized with two focus choices: noun (NN) or semantic entity (Entity). DS-SENT can be parameterized not only with focus, but also with the choices of *unweighted* (-U) and *weighted* (-W). For DS-FOCUS, we use Conpono (Iter et al., 2020) that finetuned BERT with a novel discourse-level objective regarding sentence ordering. For DS-SENT, we use BERT-NLI. This is because we find this configuration performs best after initial trials—see Table 2 (appendix). Figure 5 (appendix) shows all variants of DiscoScore. Concerning the threshold of Dep2Vec to produce entities, after experimenting with several alternatives we set it to 0.8 for DS-FOCUS (Entity) in all setups, and to 0.8 in summarization and to 0.5 in MT for DS-SENT (Entity).

[3]For simplicity, we omit the notation $\mathbf{S}^{\text{ref}}$ and $\mathcal{G}^{\text{ref}}$ for a reference.

4.2 Datasets

We consider two datasets in summarization: SummEval (Fabbri et al., 2021) and NeR18 (Grusky et al., 2018), and one dataset in document-level MT: WMT20 (Mathur et al., 2020). Note that these datasets consist of hypotheses paired with human-written references, where hypotheses are machine-generated texts of varying qualities given by neural and non-neural, extractive and abstractive language models. We outline these datasets in Appendix A.2, and provide data statistics in Table 9 (appendix).

5 Results

We first examine the importance of discourse for evaluation metrics—which underpins the usefulness of discourse metrics, and then benchmark DiscoScore on summarization and MT datasets.

Importance of Discourse. DS-FOCUS and DS-SENT concern the modeling of discourse coherence on two different levels: (i) the occurrences of foci, and (ii) the interdependence between sentences driven by focus transitions, both reflecting the discourse characteristics of a text. In the following, we describe these discourse features, and examine their importance for assessing system outputs by contrasting the discourse patterns of hypothesis and reference.

- **Focus Frequency**, denoted by FREQ(x), equals the ratio between the total frequencies of foci and the number of foci in a text x, where x is hypothesis or reference. We exclude foci occurring only once.
- **Sentence Connectivity**, denoted by CONN(x), equals the average of all elements in adjacency matrix representing the interdependence between sentences in a text x (hypothesis/reference).
- As in DiscoScore, we consider two focus choices (NN and Entity) and the choices of *unweighted* (-U) and *weighted* (-W) for these discourse features. Figure 5 (appendix) shows the links between DiscoScore and the features.

Figure 2 shows that the scales on FREQ(ref) and FREQ(hyp) in summarization differ by a large amount, i.e., from 0.5 to 2.5 on y-axis and up to 6 on x-axis. This means that hypothesis and reference can be strongly distinguished by FREQ(x), which underpins the usefulness of including such

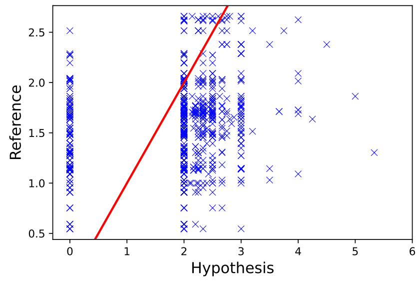


Figure 2: Scatter plot to display FREQ(hyp) (based on NN) on x-axis and FREQ(ref) on y-axis on SUMMEval. Each point contains two frequencies from a pair of hypothesis and reference. The points below the auxiliary line are the ones for which FREQ(hyp) > FREQ(ref).

discourse signals in the assessment of system outputs when references are available. Further, the larger scale on FREQ(hyp) indicates that foci in hypothesis are more repetitive than in reference, as a result of needless repetition in poor summaries—in line with previous studies on incoherent machine translations (Guillou, 2013; Voita et al., 2019). The results for other discourse features are similar, we provide them in Figure 6 (appendix).

Overall, these results show discourse features can separate hypothesis from reference.

5.1 Text Summarization

Correlation Results. Table 1 compares metrics on SUMMEval on system level. Most of non-discourse metrics have a lowest correlation with human rated coherence among four quality aspects. Even worse, ROUGE-L and SBERT do not correlate with coherence whatsoever. BARTScore, the recent state-of-the-art metric, is very weak when operated on system level, notwithstanding that it has been fine-tuned on "document-to-summary" parallel data from CNN/DailyMail—which SUMMEval is constructed from. We note that SUMMEval uses multiple references. BARTScore by default compares a hypothesis with one reference at a time, then takes the average of multiple evaluation scores as a final score. Table 8 (appendix) shows that we can improve system-level BARTScore to some degree by replacing 'average' with 'max' (i.e., taking the maximum score), but DS-FOCUS is still much better overall, i.e., surpassing BARTScore by ca. 10 points on average.

Table 7 (appendix) reports correlation results on NeR18 that uses single reference. We find that half of hypotheses do not contain 'good foci', and as such the foci-based discourse features outlined

Settings	Metrics	Coherence	Consistency	Fluency	Relevance	Average
	Non-discourse metrics					
$m(\text{hyp}, \text{ref})$	ROUGE-1	9.09	27.27	18.18	9.09	15.91
	ROUGE-L	0.00	36.36	21.21	18.18	18.94
	BERTScore	30.30	30.30	51.52	54.55	41.67
	MoverScore	36.36	42.42	63.64	60.61	50.76
	SBERT	3.03	33.33	30.30	27.27	23.48
	BLEURT	45.45	**51.52**	**72.73**	63.64	58.33
	BARTScore	60.61	36.36	45.45	48.48	47.73
	PRISM	51.52	39.39	**72.73**	69.70	58.33
	S^3-pyr	18.18	24.24	9.09	6.06	14.39
	Discourse metrics					
$m(\text{hyp})$	RC	45.45	**51.52**	**54.55**	57.58	**52.27**
	LC	**51.52**	45.45	48.48	57.58	50.76
	Entity Graph	42.42	12.12	15.15	18.18	21.97
	Lexical Graph	48.48	6.06	15.15	18.18	21.97
$m(\text{hyp}, \text{ref})$	Lexical Chain	42.42	6.06	9.09	18.18	18.94
	DS-FOCUS (NN)	**75.76**	**63.64**	**78.79**	**81.82**	**75.00**
	DS-FOCUS (Entity)	69.70	57.58	72.73	75.76	68.94
	DS-SENT-U (NN)	48.48	54.55	63.64	60.61	56.82
	DS-SENT-U (Entity)	54.55	60.61	75.76	66.67	64.39
	DS-SENT-W (NN)	51.52	51.52	66.67	63.64	58.33
	DS-SENT-W (Entity)	51.52	57.58	66.67	63.64	59.85

Table 1: System-level Kendall correlations between metrics and human ratings of summary quality on SUMMEval. We bold numbers that significantly outperform others according to paired t-test (Fisher et al., 1937). m is a metric.

previously are less discriminative on NeR18 than on SUMMEval—see Table 9 (appendix). However, DS-FOCUS is still strong, ca. 20 points better than BARTScore in all aspects, despite that DS-FOCUS uses a much smaller contextualized encoder[4]. We note that the 'F-score' version of DS-FOCUS seems extremely strong on NeR18, but it is not robust across datasets, e.g., much worse than the original, precision-based DS-FOCUS on SUMMEval.

On a side note, coherence (mostly) strongly correlates with the other rating aspects on both SUMMEval and NeR18—see Figure 3. Thus, it is not surprising that both DS-FOCUS and DS-SENT correlate well with these aspects, despite that we have not targeted them. While strong on system level, DiscoScore could not show advantages on summary level—see Table 5 (appendix), but we argue that system-level correlation deserves the highest priority as systems are compared in this manner.

Overall, these results show that BERT-based non-discourse metrics correlate weakly with human ratings on system level. BARTScore also does so, though we improve it to some degree in multi-references settings. DiscoScore, particularly DS-FOCUS, performs consistently best in both single- and multi-references settings, and it is equally strong in all aspects.

As for discourse metrics, RC and LC that use discourse features are strong baselines as they outperform most of non-discourse metrics and coherence models (i.e., Entity and Lexical Graph) without the access to source texts and references. However, they are worse than both DS-FOCUS and DS-SENT. This confirms the inadequacy of RC and LC in that they do not leverage strong contextualized encoders and judge hypothesis in the absence of references. Moreover, we compare DiscoScore to a combination of two strong, complementary baselines, BARTScore and RC—a simple solution to address text coherence of non-discourse metrics. To combine them, we simply average their scores. We see the gains are additive in all aspects but coherence. DS-FOCUS wins all the time by a large margin—see Table 10 (appendix).

Taken together, these results show that any of the three—(i) leveraging contextualized encoders as in BERT-based metrics and BARTScore; (ii) leveraging discourse features as in RC and (iii) the ensemble of (i) and (ii) by averaging—is not sufficient, suggesting to combine (i) and (ii) as DiscoScore does.

Understanding DiscoScore. As for all variants of DiscoScore, we provide understanding on why

[4] DS-FOCUS uses Conpono on the same size of BERTBase. BARTScore uses BARTLarge finetuned on CNN/DailyMail.

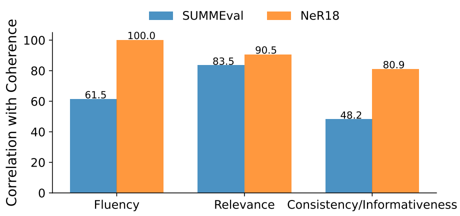


Figure 3: Pearson Correlation between coherence and other aspects on system level. SUMMEval and NeR18 use Consistency and Informativeness respectively.

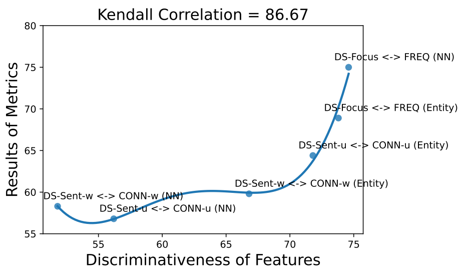


Figure 4: Correlations between the results of metrics and the discriminativeness of features on SUMMEval. Metric results are averaged across four rating aspects.

one variant is superior to another with the discourse features outlined in Figure 5 (appendix). To this end, we begin with defining the *discriminativeness* of these features as the magnitude of separating hypothesis from reference:

$$\mathcal{D}_{\mathcal{R}}(\text{hyp}, \text{ref}) := \frac{|\{(\text{hyp}, \text{ref})|\mathcal{R}(\text{ref}) < \mathcal{R}(\text{hyp})\}|}{N} \tag{4}$$

where N is a normalization term, $\mathcal{R}$ is any one of the discourse features in Figure 5 (appendix).

Figure 4 shows that the discriminativeness of these features strongly correlate with the results of the DiscoScore variants, i.e., that the more discriminative the features are, the better the metrics perform. This attributes the superiority of a metric to the fact that the discourse feature can better separate hypothesis and reference.

From this, we can interpret the performance gaps between the DiscoScore variants, namely (i) DS-FOCUS over DS-SENT: given *Focus Frequency* is more discriminative than *Sentence Connectivity*, it is not surprising that DS-FOCUS modeling discourse coherence with the former outperforms DS-SENT modeling with the latter, and (ii) DS-Focus (NN) outperforms DS-Focus (Entity) because *Frequency (NN)* can better separate hypothesis from reference than *Frequency (Entity)*.

Analyses. We provide analyses on the configuration of DiscoScore from three perspectives—see Appendix A.3: (i) the choice of BERT variants towards discourse- versus non-discourse BERT; (ii) the impact of adjacency matrices accounting for the interdependence between sentences and (iii) that we compare statistics- and alignment-based approaches to examine the best configuration for DS-SENT. Our results show the advantages of adjacency matrices and statistics based approach, and that discourse BERT only helps for DS-FOCUS.

5.2 Document-level Machine Translation

Correlation Results. Table 12 (appendix) compares metrics on WMT20. We see that non-discourse metrics seem much better, but these results are not consistent to the discriminativeness of the discourse features—see Table 11 (appendix). For instance, in cs-en, the discourse features (Frequency and Connectivity) corresponding to DS-FOCUS and DS-SENT clearly separate hypothesis from reference due to the probability of $\mathcal{D} > 0$ being over 70%. However, both DS-FOCUS and DS-SENT correlate weakly with human rated adequacy. Recently, Freitag et al. (2021a) provide justification to the inadequacy of the 'adequacy' ratings, as 'adequacy' sometimes cannot distinguish human from system translations and correlates weakly with multiple aspects (e.g., fluency and accuracy). Thus, they re-annotate WMT20 with the MQM and pSQM rating schemes, which has been subsumed into the annotation guideline of the most recent WMT evaluation campaign (Freitag et al., 2021b). Here, we perform an extra study on these ratings on both document- and system-levels. Note that system-level ratings are said to be the average of document-level ones in our setting. Table 6 (appendix) shows that DS-SENT is much better than BARTScore on system level, surpassing it by 25 points in terms of MQM and 14 points in pSQM.

Overall, these results in MT are consistent with those in summarization, i.e., DiscoScore is strong on system levels for both tasks, but it cannot show gains on fine-grained levels. Section A.4 (appendix) show inter-correlations between metrics.

6 Conclusions

Given the recent growth in discourse based NLG systems, evaluation metrics targeting the assessment of text coherence are essential next steps for properly tracking the progress of these systems.

Although there have been several attempts made towards discourse metrics, they all do not leverage strong contextualized encoders which have been held responsible for the recent success story of NLP. In this work, we introduced DiscoScore that uses BERT to model discourse coherence from two perspectives of readers' focus: (i) frequencies and semantics of foci and (ii) focus transitions over sentences used to predict interdependence between sentences. We find that BERT-based non-discourse metrics cannot address text coherence, even much worse than early feature-based discourse metrics invented a decade ago. We also find that the recent state-of-the-art BARTScore correlates weakly with human ratings on system level. DiscoScore, on the other hand, performs consistently best in both single- and multi-reference settings, equally strong in coherence and several other aspects such as factual consistency, despite that we have not targeted them. More prominently, we provide understanding on the importance of discourse for evaluation metrics, and explain the superiority of one metric over another with simple features, in line with recent work on explainability for evaluation metrics (Kaster et al., 2021; Fomicheva et al., 2021).

Scope for future research is huge, e.g., developing reference-free discourse metrics comparing source text to hypothesis, improving discourse metrics on fine-grained levels[5], and ranking NLG systems via discourse metrics and rigorous approaches (Peyrard et al., 2021; Kocmi et al., 2021).

7 Impact and Limitations

To our knowledge, we, for the first time, combine the elements of discourse and BERT representations to design an evaluation metric (DiscoScore) for text quality assessment in summarization and MT. While our experiments are conducted on English datasets, DiscoScore could adapt to many other languages in which references and foci are available. We believe that this work fosters future research on text generation systems endowed with the ability to produce well-formed texts in discourse.

However, we acknowledge several limitations of this work, which require further investigation in future. We now discuss them in the following:

Entity as Focus. We follow the idea of Mesgar and Strube (2016) in the discourse community, which clusters nouns into entities based on their static word embeddings. Although simple, it sometimes helps for DiscoScore. However, alternatives aiming to produce better entities have not been explored in this work, e.g., replacing static with contextualized embeddings, and weighting entities by their occurrences in hypothesis/reference.

Weakness on Fine-Grained Assessment. In summarization and MT, we show that our novel DiscoScore largely outperforms the current state-of-the-art BARTScore on system levels for both tasks, while it cannot show advantages on finer-grained levels such as document- and summary-levels. This might be because modeling focus alone is insufficient to perform much more challenging, finer-grained assessment of text quality. Future work could also factor other discourse phenomena (e.g., discourse connectives and coreference) into the assessment of text coherence.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments that greatly improved the texts. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1 and the Klaus Tschira Foundation, Heidelberg, Germany. Steffen Eger is funded by DFG Heisenberg grant EG 375/5-1.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

[5] Recently, Steen and Markert (2022) introduce a fine-grained evaluation setup to compute summary-level correlation, which performs computing over summaries not produced by multiple systems, but rather by a single system. This is because systems sometimes substantially differ in quality, which implies that involving multiple systems could result in inaccurate evaluation outcomes in the presence of system-level confounders.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bruno Cartoni, Jindřich Libovický, and Thomas Brovelli (Meyer), editors. 2018. *Machine Translation Evaluation beyond the Sentence Level*. Alicante, Spain.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662, Hong Kong, China. Association for Computational Linguistics.

Wang Chen, Piji Li, and Irwin King. 2021. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 404–414, Online. Association for Computational Linguistics.

Elisabet Comelles, Jesús Giménez, Lluís Màrquez, Irene Castellón, and Victoria Arranz. 2010. Document-level automatic MT evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden. Association for Computational Linguistics.

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.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Ronald Aylmer Fisher et al. 1937. The design of experiments. *The design of experiments.*, (2nd Ed).

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J Grosz et al. 1977. The representation and use of focus in a system for understanding dialogs. In *IJCAI*, volume 67, page 76. Citeseer.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the*

2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18, Sofia, Bulgaria. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870, Online. Association for Computational Linguistics.

Yuchen Jiang, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, and Ming Zhou. 2021. Blond: An automatic evaluation metric for document-level machinetranslation. *CoRR*, abs/2103.11878.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2017. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *CoRR*, abs/2107.10821.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 302–308.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas Mcdonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online.

Mohsen Mesgar, Leonardo F. R. Ribeiro, and Iryna Gurevych. 2021. A neural graph-based local coherence model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2316–2321, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423, San Diego, California. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu. 2019. A unified neural coherence model. 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 2262–2272, Hong Kong, China. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. 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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315, Online. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2020. Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2022. How to find strong summary coherence measures? a toolbox and a comparative study for summary coherence measure evaluation. *arXiv preprint arXiv:2209.06517*.

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.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020,* *Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

A Appendix

A.1 Evaluation Metrics

Non-discourse Metrics. We consider the following non-discourse metrics.

- BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are precision- and recall-oriented metrics respectively, both of which measure n-gram overlap between a hypothesis and a reference.

- BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) are set-based metrics used to measure the semantic similarity between hypothesis and reference. BERTScore uses greedy alignment to compute the similarity between two sets of BERT-based word embeddings from hypothesis and from reference, while MoverScore uses optimal alignments based on Word Mover's Distance (Kusner et al., 2015) to do so.

- SBERT (Reimers and Gurevych, 2019) fine-tunes BERT on the NLI datasets and uses pooling operations to produce sentence embeddings. We compute the cosine similarity between two sentence representations from hypothesis and from reference.

- S^3-pyr and S^3-resp (Peyrard et al., 2017) are supervised metrics that linearly combine ROUGE, JS-divergence and ROUGE-WE scores, trained on the TAC datasets with human annotated pyramid and responsiveness scores as supervision.

- BLEURT (Sellam et al., 2020) is another supervised metric that fine-tunes BERT on the concatenation of WMT datasets and synthetic data in the MT domain, with human judgment of translation quality as supervision.

- BARTScore (Yuan et al., 2021) and PRISM (Thompson and Post, 2020) depict sequence-to-sequence language models as metrics to compare hypothesis with reference. In reference-based settings, they both measure the likelihood that hypothesis and reference are paraphrases, but differ in the language models they rely on. PRISM has been based on a neural MT system trained from scratch on parallel data in MT, while BARTScore uses BART (Yuan et al., 2021) that has been fine-tuned on CNN/DailyMail (Hermann et al., 2015)—which is parallel data in summarization. We use the 'F-score' version of BARTScore as recommended in Yuan et al. (2021).

Discourse Metrics. We consider the following discourse metrics (including ours and coherence models).

- RC and LC (Wong and Kit, 2012) require neither source texts nor references and use lexical cohesion devices (e.g., repetition) within a hypothesis to predict text coherence. LC computes the proportion of words within hypothesis that are lexical cohesion devices, while RC computes the proportion of times that lexical cohesion devices appear in hypothesis.

- Entity Graph (Guinaudeau and Strube, 2013) and Lexical Graph (Mesgar and Strube, 2016) are popular coherence models used to perform discourse tasks such as essay scoring, both of which introduce a graph with nodes as sentences and adjacency matrices as the connectivity between sentences. Here, we use the average of adjacency matrices from the hypothesis as the proxy of hypothesis coherence. While Entity Graph draws an edge between two sentences if both sentences have at least one noun in common, Lexical Graph draws an edge if two sentences have a pair of similar words in common, i.e., the cosine similarity between their embeddings greater than a threshold.

- Lexical Chain (Gong et al., 2015) extracts multiple lexical chains from hypothesis and from reference. Each word is associated to a lexical chain if a word appears in more than one sentence. A lexical chain contains a set of sentence positions in which a word appears. Finally, the metric performs soft matching to measure lexical chain overlap between hypothesis and reference.

- FocusDiff and SentGraph are the two variants of DiscoScore, which use BERT to model semantics and coherence of readers' focus in hypothesis and reference. In particular, FocusDiff measures the difference between a common set of foci in hypothesis and reference in terms of semantics and frequency, while SentGraph measures the semantic similarity between two sets of sentence embeddings from hypothesis and reference—which are aggregated according to the number of foci shared across sentences and the distance between sentences.

A.2 Datasets

We outline two datasets in summarization, and one in document-level MT.

Text Summarization. While DUC[6] and TAC[7] datasets with human rated summaries, constructed one decade ago, were the standard benchmarks for comparing evaluation metrics in summmarization, they collect summaries only from extractive summarization systems. In the last few years, abstractive systems have become popular; however, little is known how well metrics judge them. Recently, several datasets based on CNN/DailyMail have been constructed to address this. For instance, SummEval (Fabbri et al., 2021), REALSumm (Bhandari et al., 2020), XSum (Maynez et al., 2020) and FEQA (Durmus et al., 2020) all collect summaries from both extractive and abstractive systems, but differ in the aspects human experts rate summaries. In this work, we consider the following two complementary summarization datasets.

- SummEval has been constructed in multiple-references settings, i.e., that each hypothesis is associated to multiple references. It contains human judgments of summary coherence, factual consistency, fluency and relevance. We only consider abstractive summaries as they have little lexical overlap with references.

- NeR18 (Grusky et al., 2018), in contrast, has been constructed in single-reference settings. It contains human judgments of summary coherence, fluency, informativeness and relevance. As majority of summaries are extractive, we include both extractive and abstractive for the inclusive picture.

Document-level Machine Translation. As document-level human ratings in MT are particularly laborious, hardly ever have there been MT datasets directly addressing them. First attempts suggested to use the average of much cheaper sentence-level ratings as a document score for comparing document-level metrics (Comelles et al., 2010; Wong and Kit, 2012; Gong et al., 2015). However, human experts were asked to rate sentences in isolation within a document. Thus, human ratings at both sentence and document levels cannot reflect inter-sentence coherence. Recently, the WMT20 workshop (Mathur et al., 2020) asks humans to rate each sentence translation in the document context, and follows the previous idea of 'average' to yield document scores.

In this work, we use the WMT20 dataset with 'artificial' document-level ratings. Note that WMT20 comes with two issues: (i) though sentences are rated in the document context, averaging sentence-level ratings may zero out negative effects of incoherent elements on document level and (ii) unlike SummEval and NeR18, WMT20 only contains human judgment of translation *adequacy* (which may subsume multiple aspects), not *coherence*.

For simplicity, we exclude system and reference translations with lengths greater than 512—the number of tokens at maximum allowed by BERT, as only a small portion of instances is over the token limit. Note that it is effortless to replace BERT with Longformer (Beltagy et al., 2020) to deal with longer documents for DiscoScore.

Metrics	Encoders	Average
DS-Focus (NN)	+ BERT	71.97
	+ BERT-NLI	70.45
	+ Conpono	75.00
DS-Sent-u (NN)	+ BERT	35.61
	+ BERT-NLI	56.82
	+ Conpono	23.48

Table 2: Results of three contextualized encoders on SUMMEval. Results are averaged across four aspects.

Metrics	Average
DS-Sent-u (NN)	56.82
w/o sentence aggregation	46.21

Table 3: Ablation study on the use of adjacency matrix to aggregate sentence embeddings on SUMMEval.

A.3 Analyses on Text Summarization

Choice of BERT Variants. Table 2 compares the impact of three BERT variants on DiscoScore. Conpono, referred to as a discourse BERT, has fine-tuned BERT with a novel discourse-level objective regarding sentence ordering. While strong on discourse evaluation benchmarks (Chen et al., 2019),

[6] https://duc.nist.gov/data.html
[7] https://tac.nist.gov/data/

Metrics	Mechanisms	Average
DS-SENT-U (NN)	+ greedy align	21.97
	+ optimal align	26.52
	+ mean-max-min-sum	56.82

Table 4: Averaged results of SentGraph variants based on three mechanisms on SUMMEval.

Metrics	SUMMEval	NeR18
BARTScore	14.13	24.78
PRISM	14.92	18.89
DS-FOCUS (NN)	10.81	10.42
DS-SENT-U (NN)	15.71	3.81

Table 5: Summary-level averaged Kendall correlations across all rating aspects.

Conpono is not always helpful, e.g., BERT-NLI is better for DS-SENT. These results suggest the best configuration for DiscoScore.

Impact of Sentence Connectivity. Table 3 shows an ablation study on the use of sentence connectivity. Aggregating sentence embeddings with our adjacency matrices (see Eq.3) helps considerably. This confirms the usefulness of aggregation from which we include coherence signals in sentence embeddings.

SentGraph Variants. Table 4 compares three DS-SENT variants as to how we measure the distance between two sets of sentence embeddings from hypothesis and reference. In particular, we refer to BERTScore (Zhang et al., 2020) as a 'greedy align' mechanism used to compute the similarity between two sets of sentence embeddings. As for 'optimal align', we use MoverScore (Zhao et al., 2019) to do so. While the two alignments directly measure the distance between the two sets, the simple statistics, i.e., mean-max-min-sum, derives a graph embedding from each set and computes the cosine similarity between two graph embeddings. We see that the 'statistics' wins by a big margin, and thus adopt this DS-SENT variant in all setups.

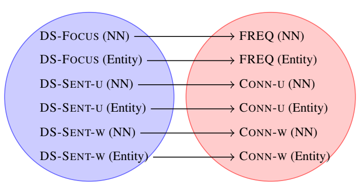


Figure 5: Links between the DiscoScore variants and discourse features.

Metrics	Sys-level MQM	Sys-level pSQM	Doc-level MQM	Doc-level pSQM
BARTScore	45.57	55.50	**34.90**	**28.96**
*DS-FOCUS (NN)	42.12	40.89	19.10	9.98
DS-SENT-U (NN)	**70.77**	**69.74**	19.98	14.49

Table 6: Document-level Kendall and system-level Pearson correlations between metrics and MQM/pSQM ratings on WMT20 in Chinese-to-English—which is the only language pair with such ratings in reference-based settings. *DS-FOCUS (NN) excludes focus that occurs only once in hypothesis/reference.

A.4 Analyses on MT

Correlation between Metrics. Figure 7 shows inter-correlations between metrics on WMT20 across languages. Overall, correlations are mostly high between non-discourse metrics, much weaker between discourse and non-discourse metrics—which confirms the orthogonality of them in that they rate translations in different aspects. We note that DS-FOCUS has the lowest correlations with all other metrics. For instance, DS-FOCUS is almost orthogonal to BERTScore and MoverScore. We investigated whether combining them receives additive gains. We find that a combination of DS-FOCUS and BERTScore (or MoverScore) provides little help in correlation with adequacy.

Settings	Metrics	Coherence	Fluency	Informative	Relevance	Average
	BARTScore	42.58	42.58	23.80	33.33	35.57
	PRISM	51.52	42.58	42.86	52.38	47.33
$m(\text{hyp}, \text{ref})$	DS-FOCUS (NN)	61.90	61.90	42.86	52.38	54.76
	DS-FOCUS* (NN)	**80.95**	**80.95**	**100.00**	**90.47**	**88.09**
	DS-SENT-U (NN)	14.29	14.29	14.29	23.81	16.67

Table 7: System-level Kendall correlations between metrics and human ratings on NeR18. DS-FOCUS* is the 'F-score' version of DS-FOCUS.

Settings	Metrics	Coherence	Consistency	Fluency	Relevance	Average
	BARTScore (max)	**78.79**	48.48	63.64	72.73	65.91
	BARTScore (original)	60.61	36.36	45.45	48.48	47.73
$m(\text{hyp}, \text{ref})$	FocusDiff (NN)	75.76	**63.64**	**78.79**	**81.82**	**75.00**
	FocusDiff (Entity)	69.70	57.58	72.73	75.76	68.94
	SentGraph-u (NN)	48.48	54.55	63.64	60.61	56.82
	SentGraph-u (Entity)	54.55	60.61	75.76	66.67	64.39

Table 8: System-level Kendall correlations between metrics and human ratings on SUMMEval in multi-reference settings. BARTScore (original) compares a hypothesis with one reference at a time, and takes the average of evaluation scores as a final score, while BARTScore (max) takes the maximum score.

			WMT20			
	SUMMEval	NeR18	cs-en	de-en	ja-en	ru-en
Number of references	11	1	1	1	1	1
Number of systems	12	7	13	14	11	13
Number of hypothesis per system	100	60	102	118	80	91
Number of sentences per hypothesis	3.13	1.90	15.21	13.84	11.29	9.46
Average number of foci in hypothesis	15.18	12.85	62.01	56.68	57.09	44.99
Average number of 'good foci' in hypothesis	2.47	2.56	13.16	13.37	15.07	9.95
Percent of hypotheses with 'good foci'	80.50%	43.80%	100%	98.60%	100%	100%

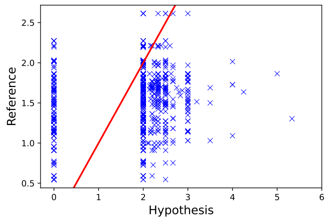
Table 9: Characteristics of summarization and MT datasets. 'good foci' denotes a focus appearing more than once in hypothesis. The more often a focus appears, the stronger the discourse signals are.

Metrics	Coherence	Consistency	Fluency	Relevance	Average
RC	45.45	51.52	54.55	57.58	52.27
BARTScore (max)	**78.79**	48.48	63.64	72.73	65.91
BARTScore (max) + RC	66.67	54.55	69.70	78.79	67.42
DS-FOCUS (NN)	75.76	**63.64**	**78.79**	**81.82**	**75.00**

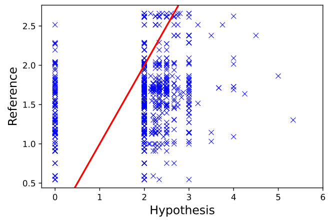
Table 10: Ensemble of non-discourse and discourse metrics (BARTScore + RC) vs DiscoScore.

	cs-en			de-en			ja-en			ru-en		
DiscoFeatures	$\mathcal{D} > 0$	$\mathcal{D} = 0$	$\mathcal{D} < 0$	$\mathcal{D} > 0$	$\mathcal{D} = 0$	$\mathcal{D} < 0$	$\mathcal{D} > 0$	$\mathcal{D} = 0$	$\mathcal{D} < 0$	$\mathcal{D} > 0$	$\mathcal{D} = 0$	$\mathcal{D} < 0$
Frequency (NN)	74.18	2.00	23.82	57.38	9.65	32.97	53.04	2.63	44.33	52.77	7.31	39.92
Frequency (Entity)	76.17	1.76	22.07	59.74	8.38	31.88	52.38	1.48	46.14	53.61	7.31	39.08
Connectivity-u (NN)	78.05	0.35	21.60	63.11	8.29	28.60	59.61	5.25	35.14	52.04	10.03	37.93
Connectivity-u (Entity)	79.46	0.35	20.19	62.02	8.20	29.78	59.44	5.09	35.47	52.87	9.40	37.72
Connectivity-w (NN)	77.93	0.24	21.83	64.85	4.64	30.51	59.12	0.49	40.39	59.98	5.12	34.90
Connectivity-w (Entity)	80.40	0.23	19.37	63.48	4.73	31.79	60.76	0.33	38.91	60.82	4.60	34.58

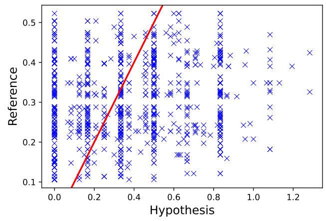
Table 11: Statistics of discourse features on WMT20. $\mathcal{D} > 0$ denotes the percent of 'reference-hypothesis' pairs for which $\mathcal{R}(\text{ref}) > \mathcal{R}(\text{hyp})$ with $\mathcal{R}$ as any one of these features, similarly for the definitions of $\mathcal{D} = 0$ and $\mathcal{D} < 0$. We exclude the pairs for which hypothesis and reference are the exact same.



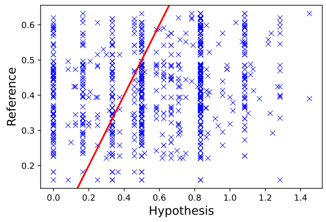
(a) Focus Frequency (NN)



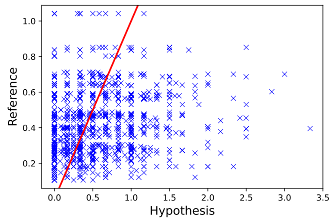
(b) Focus Frequency (Entity)



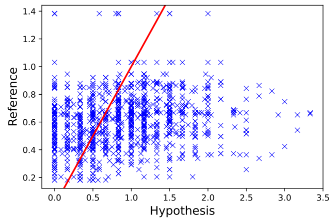
(c) Connectivity-u (NN)



(d) Connectivity-u (Entity)



(e) Connectivity-w (NN)



(f) Connectivity-w (Entity)

Figure 6: Distribution of discourse features over hypothesis and reference on SUMMEval.

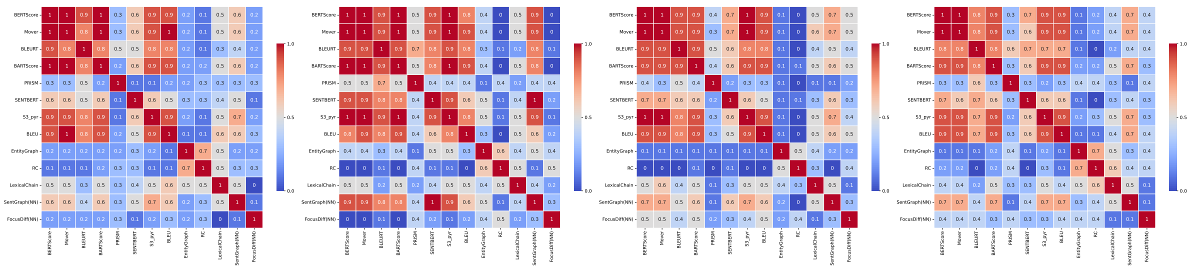


Figure 7: Pearson Correlations between metrics on WMT20 in cs-en, de-en, ja-en and ru-en (from left to right).

Settings	Metrics	Direct Assessment (Adequacy) cs-en	de-en	ja-en	ru-en	Average
	Non-discourse metrics					
$m(\text{hyp}, \text{ref})$	BLEU	7.44	57.52	41.48	10.74	29.30
	BERTScore	10.82	60.38	**46.95**	13.08	32.81
	MoverScore	**15.40**	**61.69**	42.12	13.78	**33.25**
	BARTScore	10.82	60.26	46.30	14.95	33.09
	PRISM	8.64	58.83	32.48	15.42	28.84
	SBERT	13.20	55.26	33.44	10.04	27.99
	BLEURT	12.01	58.83	37.94	**18.22**	31.75
	S^3-pyr	6.25	58.83	42.44	13.78	30.33
	S^3-resp	5.85	58.59	**47.26**	14.71	31.61
	Discourse metrics					
$m(\text{hyp})$	RC	5.85	7.19	**8.68**	9.34	7.77
	LC	**9.23**	1.72	3.53	6.07	5.14
	Entity Graph	5.06	**43.24**	3.53	10.51	15.59
	Lexical Graph	2.28	**43.60**	5.14	**13.55**	**16.15**
	Discourse metrics					
$m(\text{hyp}, \text{ref})$	Lexical Chain	21.54	35.15	15.11	16.12	21.99
	FocusDiff (NN)	7.64	33.13	19.29	2.57	15.66
	FocusDiff (Entity)	6.45	33.73	19.94	1.64	15.44
	SentGraph-u (NN)	7.64	57.16	39.22	18.22	**30.56**
	SentGraph-u (Entity)	7.65	57.17	39.23	18.22	**30.57**
	SentGraph-w (NN)	7.65	57.18	39.22	18.21	**30.57**
	SentGraph-w (Entity)	7.65	57.17	39.23	18.22	**30.57**

Table 12: Document-level Kendall correlations between metrics and human rated translation quality on WMT20.