Evaluating the Efficacy of Summarization Evaluation across Languages

Fajri Koto **Jey Han Lau** **Timothy Baldwin**
School of Computing and Information Systems
The University of Melbourne
ffajri@student.unimelb.edu.au, jeyhan.lau@gmail.com, tbaldwin@unimelb.edu.au

Abstract

While automatic summarization evaluation methods developed for English are routinely applied to other languages, this is the first attempt to systematically quantify their panlinguistic efficacy. We take a summarization corpus for eight different languages, and manually annotate generated summaries for focus (precision) and coverage (recall). Based on this, we evaluate 19 summarization evaluation metrics, and find that using multilingual BERT within BERTScore performs well across all languages, at a level above that for English.

1 Introduction

Although manual evaluation (Nenkova and Passonneau, 2004; Hardy et al., 2019) of text summarization is more reliable and interpretable, most research on text summarization employs automatic evaluations such as ROUGE (Lin, 2004), METEOR (Lavie and Agarwal, 2007), MoverScore (Zhao et al., 2019), and BERTScore (Zhang et al., 2020b) because they are time- and cost-efficient.

In proposing these metrics, the authors measured correlation with human judgments based on English datasets that are not representative of modern summarization systems. For instance, Lin (2004) use DUC[1] 2001–2003 for ROUGE (meaning summaries were generated with largely outdated extractive summarization systems); Zhao et al. (2019) use the TAC[2] dataset for MoverScore (again, featuring summaries from largely defunct systems; see Peyrard (2019) and Rankel et al. (2013)); and Zhang et al. (2020b) developed BERTScore based on a machine translation corpus (WMT). In contemporaneous work, Bhandari et al. (2020) address this issue by annotating English CNN/DailyMail summaries produced by recent summarization models, and found disparities over results from TAC.

[1] https://duc.nist.gov/data.html
[2] https://tac.nist.gov/data/

Gold summary : Info-A; Info-B; Info-C
System summary:
Good *focus,* and Good *coverage* : Info-A; Info-B; Info-C
Good *focus,* and Bad *coverage* : Info-A; Info-A
Bad *focus,* and Good *coverage* : Info-A; Info-B; Info-C; Info-D; Info-E
Bad *focus,* and Bad *coverage* : Info-D; Info-E; Info-F

Figure 1: Illustration of focus and coverage.

Equally troublingly, ROUGE has become the default summarization evaluation metric for languages other than English (Hu et al., 2015; Scialom et al., 2020; Ladhak et al., 2020; Koto et al., 2020b), despite there being no systematic validation of its efficacy across other languages. The questions we ask in this study, therefore, are twofold: (1) *How well do existing automatic metrics perform over languages other than English?* and (2) *What automatic metric works best across different languages?*

In this paper, we examine content-based summarization evaluation from the aspects of precision and recall, in the form of *focus* and *coverage* to compare system-generated summaries to ground-truth summaries (see Figure 1). As advocated by Koto et al. (2020a), focus and coverage are more interpretable and fine grained than the harmonic mean (F1 score) of ROUGE. This is also in line with the review of Hardy et al. (2019) on linguistic properties that have been manually evaluated in recent summarization research, who found precision and recall to be commonly used to complement ROUGE F1.

While it may seem more natural and reliable to evaluate focus and coverage based on the source document than the ground-truth summary, we use the ground-truth summary in this research for the following reasons. First, historically, validation of automatic evaluation metrics for summarization has been based primarily on ground-truth summaries (not source documents). Second, ROUGE (Lin,

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 801–812

2004) was initially motivated and assessed based on coverage over the DUC datasets[3] (Lin and Hovy, 2002) using annotations based on reference summaries (not source documents). Third, although it is certainly possible to generate different summaries for the same source document, we argue that the variance in content is actually not that great, especially for single-document summarization. Lastly, basing human evaluation (of focus and coverage) on the source article leads to more complicated annotation schemes, and has been shown to yield poor annotations (Nenkova and Passonneau, 2004; Fabbri et al., 2020).

In summary, this paper makes three contributions: (1) we carry out the first systematic attempt to quantify the efficacy of automatic summarization metrics over 8 linguistically-diverse languages, namely English (EN), Indonesian (ID), French (FR), Turkish (TR), Mandarin Chinese (ZH), Russian (RU), German (DE), and Spanish (ES); (2) we evaluate an extensive range of traditional and model-based metrics, and find BERTScore to be the best metric for evaluating both focus and coverage; and (3) we release a manually-annotated multilingual resource for summarization evaluation comprising 4,320 annotations. Data and code used in this paper is available at: https://github.com/fajri91/Multi_SummEval.

2 Related Work

As with much of NLP, research on automatic summarization metrics has been highly English-centric. Graham (2015) comprehensively evaluated 192 ROUGE variations based on the DUC-2004 (English) dataset. Bhandari et al. (2020) released a new (English) evaluation dataset by annotating CNN/DailyMail using simplified Pyramid (Nenkova and Passonneau, 2004). First, semantic content units (SCUs) were manually extracted from the reference, and crowd-workers were then asked to count the number of SCUs in the system summary. Their annotation procedure does not specifically consider focus, but is closely related to the coverage aspect of our work. Similarly, Fabbri et al. (2020) annotated the (English) CNN/DailyMail dataset for the four aspects of coherence, consistency, fluency, and relevance. While their work does not specifically study focus and coverage, relevance in their work can be interpreted as the harmonic mean of focus and coverage.

[3]DUC 2001, 2002, 2003

There is little work on summarization evaluation for languages other than English, and what work exists is primarily based on summaries generated by unsupervised extractive models dating back more than a decade, for a small handful of languages. Two years prior to ROUGE, Saggion et al. (2002) proposed a summarization metric using similarity measures for English and Chinese, based on cosine similarity, unit overlap, and the longest common subsequence ("LCS") between reference and system summaries. In other work, Saggion et al. (2010) investigated coverage, responsiveness, and pyramids for several extractive models in English, French, and Spanish.

To the best of our knowledge, we are the first to systemically quantify the panlinguistic efficacy of evaluation metrics for modern summarization systems.

3 Evaluation Metrics

We assess a total of 19 different evaluation metrics that are commonly used in summarization research (noting that lesser-used metrics such as FRESA (Saggion et al., 2010) and RESA (Cohan and Goharian, 2016) are omitted from this study).

ROUGE (Lin, 2004) measures the lexical overlap between the system and reference summary; based on the findings of Graham (2015), we consider 7 variants in this paper: ROUGE-1 (unigram), ROUGE-2 (bigram), ROUGE-3 (trigram), ROUGE-L (LCS), ROUGE-S (skip-bigram), ROUGE-SU (skip-bigram plus unigram), and ROUGE-W (weighted LCS).[4]

METEOR (Lavie and Agarwal, 2007) performs word-to-word matching based on word-alignment, and was originally developed for MT but has recently been used for summarization evaluation (See et al., 2017; Chen and Bansal, 2018; Falke and Gurevych, 2019; Amplayo and Lapata, 2020).[5]

BLEU (Papineni et al., 2002) is a precision-based metric originally developed for MT, which measures the n-gram match between the reference and system summary. Based on the findings of Graham (2015), we use BLEU-4 according to the SacreBLEU implementation (Post, 2018).[6]

MoverScore (Zhao et al., 2019) measures the Euclidean distance between two contextualized BERT representations, and relies on soft align-

[4]https://github.com/bheinzerling/pyrouge
[5]http://www.cs.cmu.edu/~alavie/METEOR/
[6]https://github.com/mjpost/sacrebleu

ments of words learned by solving an optimisation problem.[7] We adapt use the default configuration (n-gram=1) over 5 different pre-trained models, as detailed below. Note that MoverScore is symmetric (i.e. MoverScore(x, y) = MoverScore(y, x)), and as such is not designed to separately evaluate precision and recall.

BERTScore (Zhang et al., 2020b) computes the similarity between BERT token embeddings of system and reference summaries based on soft overlap, in the form of precision, recall, and F1 scores.[8] Zhang et al. (2020b) found that layer selection (i.e. which layer to source the token embeddings from) is critical to performance. Since layer selection in the original paper was based on MT datasets, we perform our own layer selection using a similar methodology as the authors, specifically considering precision and recall for focus and coverage, respectively.

For both MoverScore and BERTScore, we experiment with two classes of BERT-style model: (1) multilingual models, in the form of cased and uncased multilingual BERT (Devlin et al., 2019), and base and large XLM-R (Conneau et al., 2020), for a total of 4 models;[9] and (2) a monolingually-trained BERT for the given language, as listed in the Appendix. While we expect monolingual BERT models to perform better, we also focus on multilingual models, both to confirm whether this is the case, and to be able to draw findings for languages without monolingual models.

4 Experimental Setup

For each language, we sample 135 documents from the test set of a pre-existing (single-document) summarization dataset: (1) CNN/DailyMail (English: Hermann et al. (2015)); (2) Liputan6 (Indonesian: Koto et al. (2020b)); (3) LCSTS (Chinese: Hu et al. (2015)); and (4) MLSUM (French, Turkish, Russian, German, Spanish: Scialom et al. (2020)). We source summaries based on two popular models: pointer generator network (See et al., 2017) and BERT (Liu and Lapata, 2019; Dong et al., 2019),[10] and have 3 annotators annotate focus and coverage for each reference–system summary pair.[11] The motivation for using BERT-based systems is that our study focuses on non-English summarization, where BERT-based models dominate.[12] The total number of resulting annotations is: 8 languages × 135 documents × 2 models × 2 criteria (= focus and coverage) × 3 annotators = 12,960.

Lang	Quality (%)	Pearson correlation (r)		
		Agreement		Focus–Coverage
		Focus	Coverage	
EN	90	0.47	0.46	0.58
ID	97	0.64	0.63	0.80
FR	98	0.63	0.65	0.71
TR	97	0.74	0.79	0.74
ZH	92	0.61	0.60	0.78
RU	98	0.60	0.64	0.78
DE	90	0.78	0.83	0.89
ES	95	0.60	0.61	0.76

Table 1: Analysis of the annotations for each language, in terms of: (1) average quality control score of approved HITs (%); (2) one-vs-rest human agreement (r); and (3) correlation (r) between focus and coverage.

For annotation, we used Amazon Mechanical Turk[13] with the customized Direct Assessment ("DA") method (Graham et al., 2015; Graham et al., 2017), which has become the de facto for MT evaluation in WMT. For each HIT (100 samples), DA incorporates 10 pre-annotated samples for quality control. Crowd-sourced workers are given two texts and asked the question (in the local language): *How much information contained in the second text can also be found in the first text?* We combine focus and coverage annotation into 1 task, as the only thing that differentiates them is the ordering of the system and reference summaries, which is opaque to the annotators.[14] Workers responded by scoring based via a slider button (continuous scale of 1–100).[15]

For each HIT, we create 10 samples for quality control: 5 samples are random pairs (should be

[7] https://github.com/AIPHES/emnlp19-moverscore

[8] https://github.com/Tiiiger/bert_score

[9] Note that both multilingual BERT and XLM were explicitly trained over all eight target languages used in this paper.

[10] English, Indonesian and Chinese summaries were generated with the Liu and Lapata (2019) model, and the Dong et al. (2019) model was used for the MLSUM-based languages.

[11] Summaries for all datasets except LCSTS were sourced from the authors of the dataset. For LCSTS, we trained the two models ourselves based on the training data.

[12] BERT-based summaries are representative of transformer-based model, and the ROUGE score gap over state-of-the-art models (Zhang et al., 2020a) for English is only ∼2 points.

[13] https://www.mturk.com

[14] For focus, the first text is the reference and the second text the system summary; for coverage, the order is reversed.

[15] See Appendix for the MTurk annotation interface for each language.

	Focus									Coverage								
	EN	**ID**	**FR**	**TR**	**ZH**	**RU**	**DE**	**ES**	**Avg**	**EN**	**ID**	**FR**	**TR**	**ZH**	**RU**	**DE**	**ES**	**Avg**
Traditional Metrics																		
ROUGE-1	0.61	0.69	0.68	0.81	0.80	0.47	0.88	0.53	0.68	0.62	0.72	0.67	0.83	0.79	0.58	0.89	0.67	0.72
ROUGE-2	0.57	0.63	0.67	0.80	0.76	0.48	0.87	0.61	0.67	0.56	0.66	0.71	0.79	0.75	0.59	0.89	0.67	0.70
ROUGE-3	0.46	0.53	0.59	0.76	0.67	0.31	0.85	0.54	0.59	0.48	0.57	0.63	0.74	0.66	0.46	0.88	0.58	0.62
ROUGE-L	0.60	0.69	0.68	0.81	0.79	0.46	0.87	0.54	0.68	0.61	0.72	0.67	0.83	0.79	0.59	0.89	0.67	0.72
ROUGE-S	0.59	0.65	0.60	0.78	0.70	0.46	0.85	0.51	0.64	0.60	0.69	0.67	0.78	0.73	0.53	0.89	0.64	0.69
ROUGE-SU	0.59	0.66	0.61	0.78	0.72	0.43	0.85	0.50	0.64	0.60	0.70	0.68	0.78	0.75	0.56	0.89	0.65	0.70
ROUGE-W.12	0.60	0.67	0.67	0.81	0.78	0.44	0.87	0.53	0.67	0.58	0.69	0.67	0.81	0.78	0.59	0.89	0.66	0.71
METEOR	0.47	0.67	0.64	0.74	0.81	0.55	0.83	0.60	0.66	0.63	0.71	0.64	0.80	0.78	0.58	0.89	0.69	0.72
BLEU-4	0.46	0.56	0.64	0.70	0.70	0.39	0.85	0.50	0.60	0.48	0.58	0.59	0.67	0.69	0.31	0.85	0.54	0.59
MoverScore																		
mono-BERT	0.58	0.65	0.71	0.82	0.77	0.49	0.89	0.59	0.69	0.59	0.62	0.67	0.78	0.77	0.41	0.88	0.61	0.67
mBERT (cased)	0.54	0.68	0.77	0.79	0.76	**0.60**	0.88	0.63	0.70	0.52	0.69	0.72	0.75	0.75	0.49	0.85	0.68	0.68
mBERT (uncased)	0.59	0.69	**0.78**	0.81	0.76	**0.60**	0.89	**0.67**	**0.72**	0.59	0.69	0.75	0.77	0.75	0.50	0.86	0.70	0.70
XLM (base)	0.53	0.64	0.69	0.80	0.71	0.35	0.87	0.56	0.64	0.58	0.62	0.63	0.74	0.69	0.22	0.85	0.64	0.62
XLM (large)	0.51	0.58	0.68	0.79	0.57	0.33	0.87	0.53	0.61	0.55	0.62	0.59	0.72	0.58	0.21	0.84	0.56	0.58
BERTScore																		
mono-BERT	**0.62**	**0.71**	0.73	**0.83**	**0.82**	0.51	**0.91**	**0.67**	**0.72**	0.66	**0.74**	**0.77**	**0.88**	**0.80**	0.65	**0.92**	**0.74**	**0.77**
mBERT (cased)	0.56	**0.71**	0.73	**0.83**	0.78	0.56	0.90	0.59	0.71	**0.67**	0.73	0.70	0.87	0.79	**0.72**	0.90	0.71	0.76
mBERT (uncased)	0.61	**0.71**	0.72	**0.83**	0.79	0.55	0.90	0.62	**0.72**	0.64	**0.74**	0.72	0.87	0.79	0.70	0.90	0.71	0.76
XLM (base)	0.59	0.65	0.67	**0.83**	0.79	0.34	0.89	0.58	0.67	0.64	0.71	0.66	0.86	0.73	0.67	0.90	0.70	0.74
XLM (large)	0.60	0.66	0.68	**0.83**	0.79	0.42	0.90	0.60	0.69	0.65	0.70	0.69	0.86	0.74	0.66	0.90	0.70	0.74
Human performance	0.47	0.64	0.63	0.74	0.61	0.60	0.78	0.60	0.63	0.46	0.63	0.65	0.79	0.60	0.64	0.83	0.61	0.65

Table 2: Pearson correlation (r) between automatic metrics and human judgments (for Pointer Generator and BERT models combined). We compute the precision and recall of ROUGE and BERTScore for focus and coverage, respectively. BERTScore uses the optimized layer, and other metrics are computed using the default configuration of the original implementation.

Model	**Universal layer**	
	Focus	**Coverage**
mBERT (cased)	12	5
mBERT (uncased)	12	6
XLM-R (base)	4	4
XLM-R (large)	10	9

Table 3: Recommended layers for multilingual models.

scored 0) and the remaining 5 samples are repetitions of the same summary with minor edits (should be scored 100). For each language, we asked a native speaker to translate all instructions and the annotation interface. For a single HIT, we paid USD$13, and set the HIT approval rate to 95%. For HITs to be included in the annotated data, a quality control score of at least 7 out of 10 needed to be achieved. HITs below this threshold were re-run (ensuring they were not completed by a worker who had already completed that HIT), until three above-threshold annotations were obtained.[16] For each language, the HIT approval rate is set to 95% (with the number of HITs approved varying across languages). The annotation for English was restricted to US-based workers, and for other languages except Chinese was based on countries where the language is an official language.[17]

To obtain focus and coverage values, we follow standard practice in DA in z-scoring the scores from each annotator, and then averaging.

5 Results

5.1 Annotation Results

In Table 1, we present the results of the human annotation. We first normalize the ratings from each HIT into a z-score, and one-vs-rest Pearson correlation (excluding quality control items) to provide an estimate of human agreement/performance.[18] For all languages, we observe that the average quality and human agreement is moderately high. However, the agreement does vary, and it affects the interpretation of the correlation scores when we assess the automatic metrics later. Note also that we get the lowest score for English, meaning the results for non-English languages are actually more

[16] We approved all HITs with at least 30 minutes working time and a minimum quality control score of 5, irrespective of whether they passed the higher quality-control threshold required for the ground truth.

[17] In MTurk, we did not set a specific location for Chinese because we found there are no workers in China.

[18] We follow Lau et al. (2020) in computing one-vs-rest correlation: we randomly isolate a worker's score (for each sample) and compare it against the mean score of the rest using Pearson's r, and repeat this for 1000 trials to get the mean correlation.

robust.[19]

Although focus and coverage are positively correlated in Table 1, the distribution of scores varies quite a bit between languages: English annotation variance is higher than the other languages, and has the lowest correlation between focus and coverage (r = 0.57); for French, Russian, and Spanish, summaries generally have low focus and coverage (for more details, see scatterplots of focus-coverage in Figure 2 of the Appendix).

5.2 Correlation with Automatic Evaluation

In Table 2 we present the Pearson correlation between the human annotations and various automatic metrics, broken down across language and focus vs. coverage, and (naively) aggregated across languages in the form of the average correlation. We also include the one-vs-rest annotator correlation (Section 5.1) in the last row, as it can be interpreted as the average performance of a single annotator. Recognizing the sensitivity of Pearson's correlation to outliers (Mathur et al., 2020), we manually examined the distribution of scores for all language–system combinations for outliers (and present all scatterplots in Figure 2 of the Appendix).

The general pattern is consistent across languages: BERTScore performs better than other metrics in terms of both focus and coverage. This finding is consistent with that of Fabbri et al. (2020) wrt expert annotations of relevance (interpreted as the harmonic mean of our focus and coverage). ROUGE-1 and ROUGE-L are overall the best versions of ROUGE, while BLEU-4 performs the worst. For coverage, METEOR tends to be competitive with ROUGE-1, especially for EN, FR, DE, and ES, in large part because these languages are supported by the METEOR lemmatization package.

For some pre-trained models, MoverScore is competitive with BERTScore, although the average correlation is lower, especially for coverage.

We perform layer selection for BERTScore by selecting the layer that produces the highest correlation. For monolingual BERT the selection is based on the average correlation across the two summarization models, while for the multilingual models it is based on overall result across the 8 languages $\times$ 2 models. Table 3 details the recommended layer for computing BERTScore for each of the multilingual models.[20]

We observe that BERTScore with monolingual BERT performs the best, at an average of 0.72 and 0.77 for focus and coverage, resp., but only marginally above the best of the multilingual models, namely mBERT uncased (0.72 and 0.76, resp.). Given that layer selection here was performed universally across all languages (to ensure generalizability to other languages), our overall recommendation for the best metric to use is BERTScore with mBERT uncased.

When we compare the metric results to the one-vs-rest single-annotator performance from Table 1, we see a positive correspondence between the relative scores for annotator agreement and metric performance, which we suspect is largely an artefact of data quality (i.e. the metrics are assessed to perform better for languages with high agreement because the quality of the ground-truth is higher), but further research is required to confirm this. Generally the best metrics tend to outperform single-annotator performance substantially ($>$0.10), suggesting these metrics are more reliable than a single annotator.

6 Conclusion

In this work, we developed a novel dataset for assessing automatic evaluation metrics for focus and coverage across a broad range of languages and datasets. We found that BERTScore is the best metric for the vast majority of languages, and advocate that this metric be used for summarization evaluation across different languages in the future, supplanting ROUGE.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful feedback and suggestions. The first author is supported by the Australia Awards Scholarship (AAS), funded by the Department of Foreign Affairs and Trade (DFAT), Australia. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at The University of Melbourne. This facility was established with the assistance of LIEF Grant LE170100200.

[19] The relative quality for different languages largely coincides with the findings of Pavlick et al. (2014).

[20] Recommended layers for monolingual BERT are detailed in the Appendix.

References

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, 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)*, pages 9347–9359, Online. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32, pages 13063–13075.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Tobias Falke and Iryna Gurevych. 2019. Fast concept mention grouping for concept map-based multi-document summarization. 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 695–700, Minneapolis, Minnesota. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Karl Moritz Hermann, Tom Koisk, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 1693–1701.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. FFCI: A framework for interpretable automatic evaluation of summarization. *arXiv preprint arXiv:2011.13662*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020b. Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020c. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

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.

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.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. 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.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Horacio Saggion, Dragomir Radev, Simone Teufel, and Wai Lam. 2002. Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-Morales. 2010. Multilingual summarization evaluation without human models. In *Coling 2010: Posters*, pages 1059–1067, Beijing, China. Coling 2010 Organizing Committee.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020.

MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML 2020: 37th International Conference on Machine Learning*.

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

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.

A Supplementary Materials

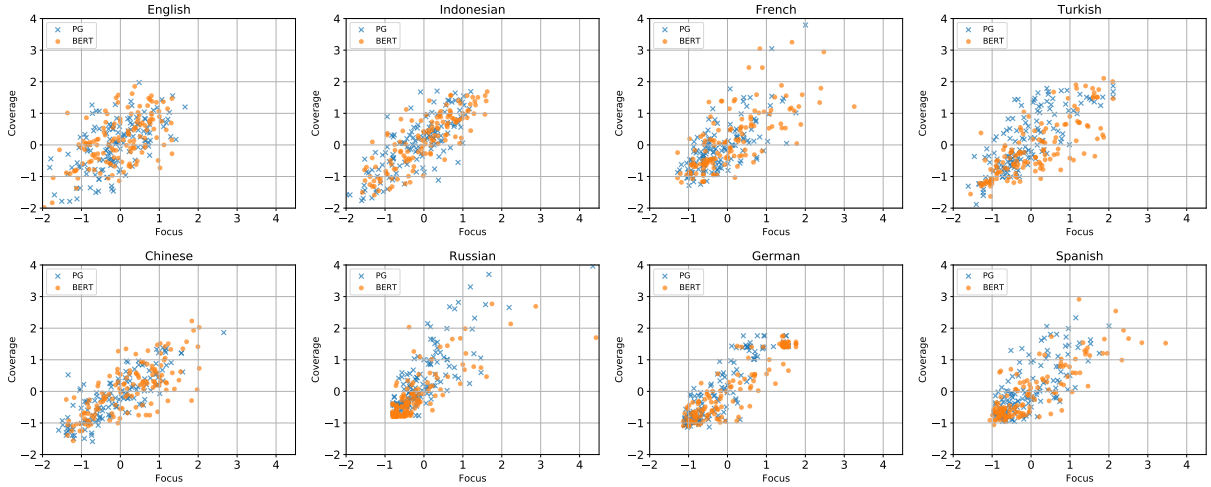

Figure 2: Annotation result (focus vs. coverage) after z-score normalization for each of the 8 languages.

Lang	Model	Recommended layer	
		Focus	Coverage
EN	`bert-base-uncased` (Devlin et al., 2019)	1	2
ID	`indolem/indobert-base-uncased` (Koto et al., 2020c)	2	2
ZH	`bert-base-chinese` (Devlin et al., 2019)	8	9
FR	`camembert-base` (Martin et al., 2020)	10	9
TR	`dbmdz/bert-base-turkish-uncased`	12	4
RU	`DeepPavlov/rubert-base-cased` (Kuratov and Arkhipov, 2019)	4	12
DE	`bert-base-german-dbmdz-uncased`	12	12
ES	`dccuchile/bert-base-spanish-wwm-uncased`	4	4

Table 4: Recommended layers for computing focus and coverage via BERTScore with monolingual model.

Language	ISO	Data	Data Split			Pointer Generator			BERT		
			Train	Dev	Test	R1	R2	RL	R1	R2	RL
English	EN	CNN/DailyMail	287,226	13,368	11,490	39.53	17.28	36.38	42.13	19.60	39.18
Indonesian	ID	Liputan6	193,883	10,972	10,792	36.10	19.19	33.56	41.08	22.85	38.01
Chinese	ZH	LCSTS	2,400,591	8,672	725	32.39	19.92	29.45	38.47	25.45	35.30
French	FR	MLSUM	392,902	16,059	15,828	26.50	9.49	20.30	28.52	11.73	22.51
Turkish	TR	MLSUM	249,277	11,565	12,775	39.77	26.45	36.12	41.28	28.16	37.79
Russian	RU	MLSUM	25,556	750	757	5.39	0.60	4.62	6.01	1.02	5.75
German	DE	MLSUM	220,887	11,394	10,701	36.86	27.06	35.04	44.11	33.99	42.10
Spanish	ES	MLSUM	266,367	10,358	13,920	25.05	7.44	19.53	26.48	9.59	21.69

Table 5: Details of datasets and ROUGE scores of summarization models used in this study. Other than for Chinese, we use summaries provided by the respective authors. For MLSUM, we report slightly different ROUGE-L scores because we use the original ROUGE package.

Metrics	POINTER GENERATOR									BERT								
	EN	ID	FR	TR	ZH	RU	DE	ES	Avg	EN	ID	FR	TR	ZH	RU	DE	ES	Avg
Traditional Metrics																		
ROUGE-1	0.59	0.68	0.54	0.82	0.81	0.52	0.85	0.50	0.66	0.61	0.70	0.73	**0.81**	0.78	0.59	0.89	0.54	0.71
ROUGE-2	0.60	0.59	0.56	0.83	0.78	0.54	0.86	0.61	0.67	0.53	0.65	0.71	0.77	0.73	0.56	0.87	0.61	0.68
ROUGE-3	0.49	0.52	0.49	0.80	0.68	0.39	0.85	0.56	0.60	0.43	0.53	0.63	0.73	0.63	0.29	0.85	0.54	0.58
ROUGE-L	0.59	**0.69**	0.55	0.83	0.81	0.51	0.85	0.50	0.67	0.60	0.68	0.73	0.80	0.75	0.58	0.88	0.57	0.70
ROUGE-S	0.60	0.62	0.48	0.79	0.70	0.55	0.83	0.50	0.63	0.58	0.67	0.64	0.76	0.69	0.59	0.86	0.51	0.66
ROUGE-SU	0.59	0.63	0.50	0.79	0.72	0.55	0.83	0.50	0.64	0.58	0.68	0.66	0.77	0.70	0.60	0.86	0.50	0.67
ROUGE-W.12	0.60	0.67	0.56	0.83	0.81	0.52	0.86	0.49	0.66	0.60	0.66	0.72	0.79	0.74	0.55	0.88	0.57	0.69
METEOR	0.49	0.65	0.51	0.82	0.85	0.52	0.86	0.68	0.67	0.45	0.68	0.70	0.71	0.77	0.52	0.85	0.58	0.66
BLEU-4	0.51	0.57	0.60	0.78	0.75	0.46	0.86	0.59	0.64	0.43	0.54	0.66	0.64	0.65	0.51	0.85	0.46	0.59
MoverScore																		
mono-BERT	0.62	0.67	0.63	**0.88**	0.80	**0.65**	**0.90**	0.60	0.72	0.54	0.63	0.74	0.77	0.73	0.61	0.89	0.58	0.69
mBERT (cased)	0.57	0.62	0.71	0.84	0.79	0.60	0.88	0.71	0.72	0.51	0.71	**0.78**	0.75	0.73	0.68	0.88	0.59	0.70
mBERT (uncased)	0.63	0.65	**0.76**	**0.88**	0.79	0.57	0.89	**0.74**	**0.74**	0.54	0.71	**0.78**	0.76	0.73	0.68	0.88	0.63	0.71
XLM (base)	0.56	0.61	0.60	0.86	0.73	0.32	0.86	0.71	0.66	0.49	0.63	0.71	0.77	0.68	0.52	0.89	0.51	0.65
XLM (large)	0.53	0.57	0.60	0.84	0.62	0.30	0.87	0.63	0.62	0.48	0.58	0.69	0.76	0.52	0.42	0.87	0.46	0.60
BERTScore																		
mono-BERT	0.62	0.68	0.71	0.86	0.82	0.42	**0.90**	0.69	0.71	0.62	0.73	0.72	0.80	**0.81**	**0.71**	**0.91**	**0.66**	**0.75**
mBERT (cased)	0.61	0.67	0.64	0.84	0.77	0.54	0.89	0.69	0.71	0.58	**0.74**	0.76	**0.81**	0.77	**0.71**	**0.91**	0.54	0.73
mBERT (uncased)	**0.64**	0.68	0.67	0.86	0.79	0.48	0.89	0.70	0.72	0.63	0.72	0.73	**0.81**	0.77	0.69	0.90	0.57	0.73
XLM (base)	0.62	0.65	0.59	0.86	0.80	0.39	0.87	0.63	0.68	0.61	0.64	0.69	**0.81**	0.78	0.64	0.90	0.56	0.70
XLM (large)	0.63	0.65	0.64	0.87	0.80	0.35	0.88	0.67	0.68	**0.64**	0.66	0.69	0.80	0.77	0.66	0.90	0.57	0.71

Table 6: Pearson correlation (r) between automatic metrics and human judgments for **focus**. We compute the precision for ROUGE and BERTScore. BERTScore uses the optimized layer, and other metrics are computed by using default configuration of the original implementation.

Metrics	POINTER GENERATOR									BERT								
	EN	ID	FR	TR	ZH	RU	DE	ES	Avg	EN	ID	FR	TR	ZH	RU	DE	ES	Avg
Traditional Metrics																		
ROUGE-1	0.59	0.73	0.66	0.79	**0.82**	0.52	0.90	0.67	0.71	0.64	0.71	0.70	0.86	0.76	0.57	0.89	0.70	0.73
ROUGE-2	0.55	0.65	0.68	0.76	0.78	0.63	0.89	0.64	0.70	0.57	0.66	0.72	0.83	0.72	0.52	0.90	0.73	0.71
ROUGE-3	0.49	0.59	0.58	0.69	0.68	0.50	0.88	0.57	0.62	0.48	0.55	0.66	0.79	0.63	0.41	0.89	0.64	0.63
ROUGE-L	0.58	0.74	0.64	0.79	**0.82**	0.54	0.90	0.66	0.71	0.63	0.70	0.71	0.86	0.77	0.56	0.90	0.72	0.73
ROUGE-S	0.60	0.69	0.63	0.74	0.77	0.52	0.89	0.66	0.69	0.61	0.69	0.69	0.81	0.70	0.51	0.89	0.71	0.70
ROUGE-SU	0.59	0.70	0.64	0.75	0.79	0.52	0.89	0.67	0.69	0.61	0.69	0.70	0.82	0.71	0.57	0.89	0.71	0.71
ROUGE-W.12	0.54	0.71	0.64	0.77	0.81	0.55	0.90	0.65	0.69	0.61	0.68	0.69	0.85	0.75	0.56	0.90	0.68	0.71
METEOR	0.60	0.72	0.65	0.77	0.81	0.55	0.89	0.63	0.70	0.66	0.69	0.69	0.83	0.75	0.59	0.89	0.75	0.73
BLEU-4	0.48	0.61	0.63	0.61	0.70	0.49	0.84	0.50	0.61	0.49	0.56	0.59	0.75	0.67	0.54	0.87	0.59	0.63
MoverScore																		
mono-BERT	0.57	0.65	0.63	0.73	0.79	0.68	0.86	0.55	0.68	0.61	0.60	0.69	0.86	0.75	0.66	0.91	0.68	0.72
mBERT (cased)	0.53	0.67	0.68	0.71	0.77	0.60	0.82	0.63	0.68	0.53	0.71	0.75	0.82	0.73	0.63	0.89	0.74	0.73
mBERT (uncased)	0.58	0.68	0.74	0.72	0.76	0.58	0.84	0.64	0.69	0.59	0.70	0.76	0.85	0.73	0.65	0.90	0.76	0.74
XLM (base)	0.56	0.61	0.52	0.68	0.71	0.31	0.82	0.62	0.60	0.58	0.64	0.68	0.83	0.65	0.52	0.90	0.68	0.68
XLM (large)	0.52	0.62	0.50	0.66	0.59	0.31	0.82	0.49	0.56	0.57	0.61	0.63	0.82	0.56	0.48	0.88	0.63	0.65
BERTScore																		
mono-BERT	0.63	0.74	0.76	**0.87**	0.81	**0.72**	**0.92**	**0.73**	**0.77**	0.67	**0.74**	**0.78**	**0.89**	**0.78**	0.63	**0.92**	**0.78**	**0.77**
mBERT (cased)	**0.67**	**0.75**	0.67	0.85	**0.82**	0.70	0.91	0.72	0.76	**0.68**	0.71	0.74	**0.89**	0.76	**0.69**	0.90	0.72	0.76
mBERT (uncased)	0.63	**0.75**	0.70	0.85	0.81	0.67	0.91	0.71	0.76	0.64	0.73	0.76	**0.89**	0.77	0.68	0.90	0.73	0.76
XLM (base)	0.66	0.72	0.68	0.84	0.77	0.63	0.91	0.70	0.74	0.64	0.70	0.67	0.88	0.69	0.67	0.89	0.71	0.73
XLM (large)	0.66	0.70	0.68	0.84	0.77	0.59	0.91	0.70	0.73	0.66	0.69	0.70	0.88	0.70	**0.69**	0.90	0.72	0.74

Table 7: Pearson correlation (r) between automatic metrics and human judgments for **coverage**. We compute the recall for ROUGE and BERTScore. BERTScore uses the optimized layer, and other metrics are computed by using default configuration of the original implementation.

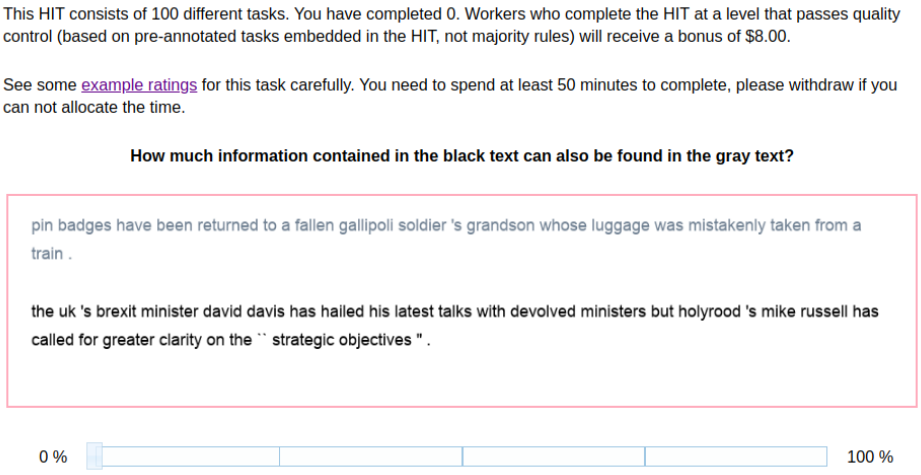

Figure 3: MTurk annotation interface for **English**.

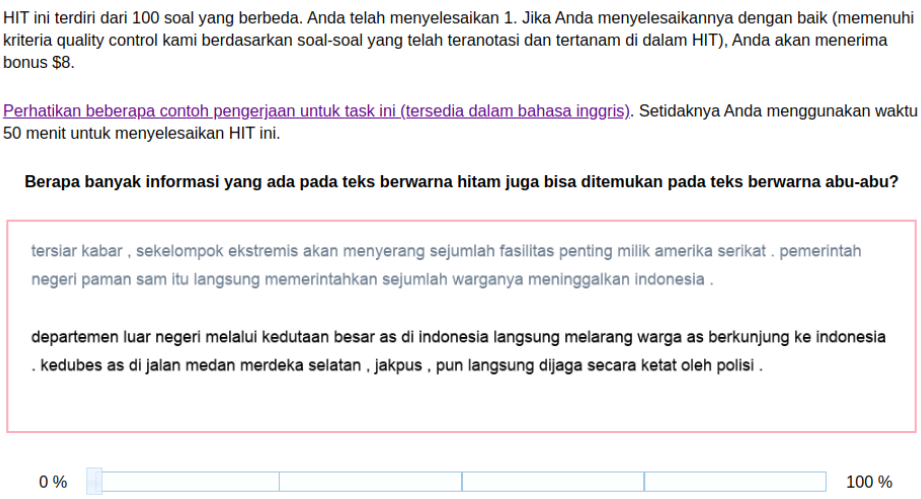

Figure 4: MTurk annotation interface for **Indonesian**.

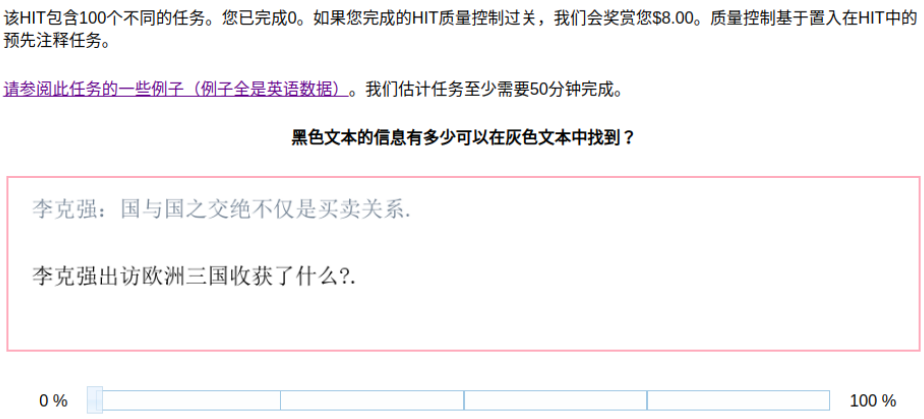

Figure 5: MTurk annotation interface for **Chinese**. Due to the page limit for the Appendix, the annotation interface for the other languages can be found at https://github.com/fajri91/Multi_SummEval

	Focus									Coverage								
	EN	**ID**	**FR**	**TR**	**ZH**	**RU**	**DE**	**ES**	**Avg**	**EN**	**ID**	**FR**	**TR**	**ZH**	**RU**	**DE**	**ES**	**Avg**
Traditional Metrics																		
ROUGE-1	0.59	0.69	0.62	0.75	0.80	0.33	0.77	0.45	0.63	0.58	0.71	0.56	0.79	0.78	0.47	0.75	0.59	0.65
ROUGE-2	0.58	0.63	0.64	0.75	0.78	0.37	0.78	0.56	0.64	0.54	0.65	0.62	0.77	0.75	0.39	0.77	0.63	0.64
ROUGE-3	0.49	0.55	0.62	0.70	0.68	0.27	0.77	0.51	0.58	0.47	0.58	0.59	0.69	0.65	0.28	0.75	0.58	0.57
ROUGE-L	0.60	0.69	0.62	0.75	0.80	0.32	0.76	0.46	0.62	0.57	0.70	0.55	0.79	0.78	0.47	0.74	0.58	0.65
ROUGE-S	0.61	0.71	0.61	0.75	0.81	0.36	0.77	0.45	0.63	0.60	0.72	0.56	0.79	0.79	0.39	0.75	0.58	0.65
ROUGE-SU	0.61	0.71	0.61	0.74	0.81	0.31	0.77	0.43	0.62	0.59	0.72	0.56	0.79	0.79	0.45	0.75	0.58	0.66
ROUGE-W.12	0.61	0.67	0.61	0.74	0.79	0.32	0.76	0.47	0.62	0.55	0.68	0.55	0.78	0.78	0.48	0.74	0.58	0.64
METEOR	0.47	0.68	0.60	0.74	0.82	0.40	0.75	0.54	0.62	0.61	0.70	0.59	0.81	0.78	0.44	0.74	0.62	0.66
BLEU-4	0.50	0.59	0.62	0.67	0.76	0.05	0.77	0.51	0.56	0.49	0.61	0.58	0.62	0.72	-0.04	0.73	0.55	0.53
MoverScore																		
mono-BERT	0.58	0.65	0.66	**0.81**	0.79	0.38	**0.85**	0.55	0.66	0.59	0.61	0.60	0.81	0.77	0.30	**0.83**	0.59	0.64
mBERT (cased)	0.53	0.68	0.73	0.80	0.77	0.44	0.83	0.60	0.67	0.50	0.68	0.65	0.77	0.74	0.37	0.79	0.65	0.64
mBERT (uncased)	**0.60**	0.69	**0.74**	0.80	0.77	**0.45**	0.83	**0.66**	**0.69**	0.58	0.68	0.68	0.77	0.74	0.42	0.79	0.67	0.67
XLM (base)	0.51	0.63	0.64	0.78	0.70	0.11	0.78	0.53	0.59	0.55	0.62	0.58	0.73	0.65	0.01	0.75	0.59	0.56
XLM (large)	0.51	0.59	0.63	0.72	0.54	0.10	0.76	0.44	0.54	0.54	0.61	0.52	0.68	0.52	0.07	0.72	0.50	0.52
BERTScore																		
mono-BERT	0.58	0.70	0.67	0.77	**0.81**	0.34	0.84	0.60	0.66	**0.64**	**0.73**	**0.70**	**0.86**	**0.78**	0.57	**0.83**	**0.70**	**0.73**
mBERT (cased)	0.53	**0.71**	0.68	0.78	0.78	0.41	0.82	0.55	0.66	**0.64**	0.72	0.58	0.85	0.77	**0.63**	0.77	0.67	0.70
mBERT (uncased)	0.58	0.70	0.66	0.79	0.79	0.40	0.83	0.58	0.67	0.60	0.73	0.60	0.85	0.77	**0.63**	0.77	0.66	0.70
XLM (base)	0.56	0.65	0.61	0.78	0.78	0.04	0.78	0.49	0.59	0.62	0.71	0.58	0.84	0.70	0.55	0.76	0.64	0.67
XLM (large)	0.57	0.66	0.62	0.78	0.77	0.17	0.80	0.52	0.61	0.62	0.69	0.60	0.84	0.72	0.50	0.78	0.65	0.68

Table 8: Spearman correlation (ρ) between automatic metrics and human judgments (for Pointer Generator and BERT models combined). We compute the precision and recall of ROUGE and BERTScore for focus and coverage, respectively. BERTScore uses the optimized layer, and other metrics are computed by using default configuration of the original implementation.