Paraphrase Generation Evaluation Powered by an LLM: A Semantic Metric, Not a Lexical One

Quentin Lemesle[1], Jonathan Chevelu[1], Philippe Martin[1], Damien Lolive[1,2], Arnaud Delhay[1], Nelly Barbot[1],

[1]Univ Rennes, CNRS, IRISA, Expression, [2]Univ of South Brittany, CNRS, IRISA, Expression,

Correspondence: quentin.lemesle@irisa.fr

Abstract

Evaluating automatic paraphrase production systems is a difficult task as it involves, among other things, assessing the semantic proximity between two sentences. Usual measures are based on lexical distances, or at least on semantic embedding alignments. The rise of Large Language Models (LLM) has provided tools to model relationships within a text thanks to the attention mechanism. In this article, we introduce ParaPLUIE (ParaPhrase, Llm Used for Improved Evaluation), a new measure based on a log likelihood ratio from an LLM, to assess the quality of a potential paraphrase. This measure is compared with usual measures on two known by the NLP community datasets prior to this study. Three new small datasets have been built to allow metrics to be compared in different scenario and to avoid data contamination bias. According to evaluations, the proposed measure is better for sorting pairs of sentences by semantic proximity. In particular, it is much more independent to lexical distance and provides an interpretable classification threshold between paraphrases and non-paraphrases.

1 Introduction

In the field of automatic generation of paraphrases, plenty of definitions of paraphrases have been proposed (Mel'čuk, 1997; Barzilay and McKeown, 2001; Sekine, 2005; Zhao et al., 2009; Fabre et al., 2021). All those definitions point the importance of meaning conservation, that is inherently an ambiguous concept.

Despite this, paraphrase generation systems need semantic measures to be trained or evaluated. Usually, metrics work either with lexical matching (Papineni et al., 2002) or embedding matching (Zhang et al., 2020). By design, lexical matching approaches struggle to reconcile simple transformations like synonym replacement (Banerjee and Lavie, 2005). Moreover, they have difficulties to reject sentences with an opposed meaning if they are lexically close. On the other hand, metrics that use semantic embedding matching, are laid on subphrasal alignments without taking into account a global view of sentences. These two points have been highlighted by Zhang et al. (2019) and have led to the construction of PAWS dataset.

The TRANSFORMER architecture (Vaswani et al., 2017) and the emergence of Large Language Models have brought many advances in the area of natural language processing. Specifically, the self-attention mechanism can capture semantic relations in a large context. Chen et al. (2023) have demonstrated that an LLM is capable of scoring the quality of reference-free sentences.

Our main contributions detailed in this paper are: (1) a new semantic metric for paraphrase classification, ParaPLUIE (ParaPhrase, Llm Used for Improved Evaluation), based on an LLM and its output perplexity, (2) a deep analysis of usual metrics performances on a semantic proximity task, (3) three new human labeled datasets of paraphrases and non-paraphrases to evaluate metrics.

The paper is organized as follows. First, metrics usually used to classify paraphrases are summed up in section 2. A novel automatic metric dedicated to semantic proximity, named ParaPLUIE, is proposed in section 3. The reference evaluation datasets are then described in section 4, including three new datasets of human labeled paraphrases. State of the art metrics are evaluated together with ParaPLUIE in section 5. It should be noted that, despite their variety, automatic metric scores seem correlated to the edit distance (i.e LEV. distance), which is not the case for ParaPLUIE.

2 Metrics

Metrics usually used to evaluate meaning conservation between two sentences can be split into two main groups: one that involves metrics measuring how much the lexical structure is similar between

Proceedings of the 31st International Conference on Computational Linguistics, pages 8057–8087
January 19–24, 2025.

two sentences, thanks to a lexical distance, whereas the other involves metrics estimating the semantic proximity between two sentences, thanks to embedding matching.

2.1 Lexical structure based ones

In the first group, we can include the Levenshtein distance (LEV.) (Levenshtein, 1965), BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).

LEV. gives a measurement of differences between two character strings. This metric counts the minimum number of deletions, insertions and replacements of characters, required to transform one string into another. As the considered strings become longer, the LEV. distance increases, it is then generally normalised by the longest string length.

BLEU has been designed to assess the translation quality. It consists in computing the n-gram overlap between a candidate and a reference sentence as well as a brevity penalty. Usually, n-grams up to 4 words long are considered. In this paper, we use the *Torchtext*[1] implementation of BLEU with the default settings.

METEOR echoes the design of BLEU, calculating a harmonic mean of the uni-gram precision and recall between the hypothesis and the source. Moreover, METEOR considers a synonym matching to compute its score. METEOR has shown a better correlation with human judgement than BLEU.

It might be argued that, if two sentences have a close lexical structure, they are more likely to be paraphrases. This is why, even if it does not seem adequate, lexical metrics can be used to assess if two sentences share a common meaning. The weakness of this argument is that, even if two sentences share a common structure, they can convey a different meaning like these two sentences: *"The cat is alive"* and *"The cat was alive"*.

2.2 Semantic proximity based ones

To address the issue of sentences with common structure but different meaning or the opposite, a research effort has been made to create another group of metrics. These metrics rely on semantic distances and use token embeddings to symbolize words inside an LLM. In this second group, we can include $\text{BERT}_{\text{score}}$ (Zhang et al., 2020) and ParaScore (Shen et al., 2022).

$\text{BERT}_{\text{score}}$ is a score of similarity between each token embeddings of a candidate (here named hypothesis) and a source. Its definition is based on the following assumption: if a pairing between two sentences exists such that all embeddings that form them are close, then their meaning is close. In the experiments, we use the BERT base uncased model (Devlin et al., 2019) from *Hugging Face*[2].

Shen et al. (2022) point out that, while lexical distance between two sentences increases, the performance of metrics decreases. To deal with this issue, they propose ParaScore, a metric that extends $\text{BERT}_{\text{score}}$ by including the normalized LEV. distance to determine a similarity score.

It is important to note that semantic similarity metrics take into account a word to word matching, without considering higher level semantic relations. This carries a risk concerning the quality of classification of paraphrases.

3 ParaPLUIE

Usual metrics are focused on lexical proximity, or at best on token embedding alignments. As a result, their capacity to catch complex relations between sentences is limited. Recently, the TRANSFORMER architecture, thanks to the self-attention mechanism (Vaswani et al., 2017), has demonstrated that, it is possible to more effectively consider the internal relationships within a text.

LLMs are intended to model the probabilities associated to a token, knowing the previous ones. It is thus possible to compare two similar sequences, to calculate a class belonging degree, while considering intricate and subtle relations inside sentences.

We propose ParaPLUIE, a novel semantic proximity metric, relying on a learnt probabilistic model of an LLM. ParaPLUIE is defined as the log likelihood ratio of "*yes*" versus "*no*" knowing a template (Tpl, see section 3.1) filled with the source **S** sentence to paraphrase and the evaluated hypothesis **H**, i.e:

$$\text{ParaPLUIE}(S, H) = \log\left(\frac{p\,(\text{yes}|\text{Tpl}(S, H))}{p\,(\text{no}|\text{Tpl}(S, H))}\right)$$

The intuition behind ParaPLUIE comes from the fact that LLM are able to criticize sentences while generating. In that case, their surprise on the appearance of a token can be used as a metric.

[1] https://pytorch.org/text/stable/data_metrics.html

[2] https://huggingface.co/spaces/evaluate-metric/bertscore

A positive score is given to a couple of sentences if the system estimates that they are likely to be paraphrases. On the opposite, the system gives a negative score when it estimates that they are not paraphrases. This property helps the interpretation of results unlike other scoring metrics because it creates a natural threshold decision at zero. This score is a real value whose range depends on the learnt probabilistic model.

3.1 Templates

A template is a prompt that is filled with sentences to evaluate. In the following, we note **S** as the source sentence and **H** as a candidate paraphrase of **S**. The template mimics a dialog with a user and an assistant. This is because the model used in these experiments is a fine-tuned LLM, learnt to work as a conversational agent. We consider three different templates in our experiments.

3.1.1 Template: DIRECT

This naive template directly explains the intended task and the expected output format to the model.

$\textbf{Tpl}_{\text{Direct}}(\textbf{S}, \textbf{H})$:
($user$): *You will receive two sentences A and B. Do these two sentences mean the same thing? Answer with only one word "yes" or "no".*
($assistant$): *Please provide the sentences for me to evaluate.*
($user$): *A: "***S***"; B: "***H***"*

3.1.2 Template: INDIRECT

(Qiao et al., 2023) points out that using a chain of thoughts may help the LLM to answer correctly. In other words, letting an LLM generate context or explanations about a question makes it more likely to be right in its answer. Inspired from this, this template involves a generation step, which we denote **E**. First, the model generates its answer, then we request a one word summary.

$\textbf{Tpl}_{\text{Indirect}}(\textbf{S}, \textbf{H})$:
($user$): *You will receive two sentences A and B. Do these two sentences mean the same thing?*
($assistant$): *Please provide the sentences for me to evaluate.*
($user$): *A: "***S***"; B: "***H***"*
$Generation \longrightarrow E$
($assistant$): **E**
($user$): *Summarize your answer with only one word "yes" or "no".*

3.1.3 Template: FS-DIRECT

Numerous studies have shown that a few-shots approach helps LLMs to give an accurate answer (Rios and Kavuluru, 2018; Brown et al., 2020; Chung et al., 2024). We use an improved version of the DIRECT template which contains few examples of the task. These examples were generated using an LLM and were labeled by three experts. We have intentionally picked examples where ParaPLUIE with the DIRECT template made scoring errors. More precisely, we have chosen examples for which the associated score was likely to classify them as paraphrase, while they are non-paraphrases and reciprocally. We have also picked some examples where the model was right with it prediction. The complete template is available in A.4.

3.2 Practical computation

To compute the prediction score with ParaPLUIE, we evaluate the ratio between the probability that the template is followed by the "yes" token and the probability that the template is followed by the "no" token. As the two templates differ by only one token ("yes" or "no"), we can reformulate the equation using perplexities. This is convenient as the perplexity reflects the model "surprise" for the token prediction.

$$\text{ParaPLUIE}(S,H) = \log\left(\frac{p\,(\text{yes}|\text{Tpl}(S,H))}{p\,(\text{no}|\text{Tpl}(S,H))}\right)$$
$$= \log\left(\frac{ppl\,(\text{Tpl}(S,H)\circ\text{no})^{T+1}}{ppl\,(\text{Tpl}(S,H)\circ\text{yes})^{T+1}}\right)$$

where T is the number of tokens that made up the template and "$\circ$" a text concatenation operator.

Moreover, as LLMs are trained by using the perplexity as a loss function, we can use it directly. Then, the metric equation becomes:

$$\text{ParaPLUIE}(S,H) = (T+1)\times$$
$$(loss_{LLM}(\text{Tpl}(S,H)\circ\text{no}) - loss_{LLM}(\text{Tpl}(S,H)\circ\text{yes}))$$

The figure 1 illustrates the workflow of ParaPLUIE.

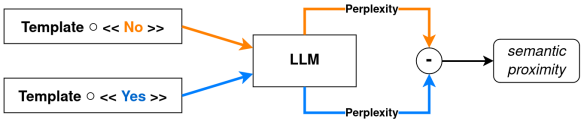


Figure 1: Illustration of ParaPLUIE workflow.

	MRPC		PAWS		MCPG		LLM		HC	
	yes	no	yes	no	yes	no	yes	no	yes	no
LEV. ↓	0.38 ±0.16	0.51 ±0.13	0.20 ±0.15	0.32 ±0.15	0.22 ±0.19	0.50 ±0.22	0.49 ±0.17	0.55 ±0.16	0.45 ±0.20	0.36 ±0.24
BLEU ↑	0.40 ±0.21	0.28 ±0.18	0.62 ±0.18	0.49 ±0.19	0.56 ±0.35	0.20 ±0.28	0.24 ±0.20	0.18 ±0.18	0.18 ±0.26	0.24 ±0.31
METEOR ↑	0.69 ±0.14	0.56 ±0.15	0.91 ±0.06	0.88 ±0.07	0.84 ±0.20	0.47 ±0.29	0.64 ±0.18	0.54 ±0.20	0.59 ±0.24	0.60 ±0.28
$\text{BERT}_{\text{score}}$ ↑	0.82 ±0.07	0.74 ±0.08	0.94 ±0.04	0.91 ±0.04	0.92 ±0.11	0.71 ±0.16	0.82 ±0.08	0.76 ±0.11	0.80 ±0.11	0.77 ±0.13
ParaScore ↑	0.83 ±0.07	0.76 ±0.08	0.92 ±0.04	0.92 ±0.04	0.90 ±0.11	0.73 ±0.16	0.82 ±0.08	0.76 ±0.11	0.80 ±0.11	0.77 ±0.13
ParaPLUIE DIRECT ↑	20.02 ±8.94	4.41 ±15.43	22.04 ±6.65	12.80 ±13.46	21.83 ±6.47	−3.56 ±13.05	23.84 ±5.27	16.44 ±13.81	19.94 ±10.58	−10.10 ±10.93
ParaPLUIE INDIRECT ↑	14.71 ±12.79	−2.61 ±14.88	18.33 ±9.59	6.96 ±16.09	18.29 ±9.58	−9.86 ±8.78	19.07 ±8.07	10.91 ±14.84	13.53 ±14.33	−11.94 ±7.64
ParaPLUIE FS-DIRECT ↑	5.00 ±7.92	−4.89 ±8.30	9.10 ±7.36	−3.82 ±10.38	9.50 ±6.87	−10.48 ±6.67	10.16 ±7.02	4.05 ±11.19	6.87 ±9.82	−12.79 ±5.35

Table 1: Average scores and standard deviation of each measure on each corpus. Datasets have been split according to the hypothesis sentence label: yes, for a paraphrase or no. The ↑ is associated to a metric where the higher the score is, the closer the sentences are to each other. The ↓ sign means the opposite.

In our experiments, we use the MISTRAL *7B Instruct v0-2*[3] version of MISTRAL, in half-precision configuration. This one is a medium size language model with 7 billion parameters. It is based on the TRANSFORMER architecture and uses a sliding attention window to reduce computing costs. The dataset used for its training is not disclosed. With this configuration, the model needs approximately 15 GB of memory. We have conducted our experiments on a computer equipped with a Nvidia RTX 4090 GPU. The code is released as supplementary material.

	Size	Repartition	
		P.	Not P.
MRPC	5 801	3900 : 67%	1901 : 33%
PAWS	8 000	3539 : 44%	4461 : 56%
MCPG	146	101 : 69%	45 : 31%
LLM	578	457 : 79%	121 : 21%
HC	200	100 : 50%	100 : 50%

Table 2: Distribution of couple for each corpus according to their labels. Size denotes the number of couples in each dataset. Couples labeled as paraphrase are denote by P. and non-paraphrases by Not P.

4 Datasets

Evaluating automatic metrics on sentence to sentence semantic proximity involves using datasets of labeled paired sentences as paraphrases/not paraphrases. Optimally, for assessing the relevance of metrics in challenging cases, labeled pairs as non-paraphrases should be lexically or semantically close (without being considered as paraphrases by a human).

The table 2 summarizes the size and distribution of paraphrases/non-paraphrases in each corpus.

4.1 Reference datasets

Our choice thus settled on two English corpora: MRPC (Dolan and Brockett, 2005) which includes examples of semantic inference (but asymmetric), PAWS (Zhang et al., 2019) designed to fool lexical metrics. Although these corpora are not without flaws, they are often used by the community (Devlin et al., 2019; Reimers and Gurevych, 2019; Fabre et al., 2021). Thus, they will serve as references.

4.1.1 MRPC

The MRPC (MS-SSLA licence) dataset used in this paper is available on HuggingFace[4]. This dataset has been generated automatically from a large corpus of newspapers organized by themes. During the labeling, the procedure was the following: for each couple of sentences, two evaluators have been asked if the pair can be considered as semantically equivalent. They were constrained to answer only by yes or no. In case of disagreement, a third evaluator answers with the same guideline. This dataset is mostly composed of entailments. Here is a characteristic example of non-paraphrase entailment

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[4] https://huggingface.co/docs/datasets/v1.13.0/about_dataset_features.html?highlight=mrpc

from MRPC: "*Last year, Bush appointed him to the Homeland Security Advisory Council.*" and "*He has also served on the president's Homeland Security Advisory Council.*".

4.1.2 PAWS

PAWS has been generated in a semi-automatic manner by word swapping and reverse translation. For each generated couple, 5 humans have labeled the couple as paraphrases or non-paraphrases. PAWS has been designed to be a challenge for automatic paraphrase classification systems. Indeed, generating sentences by word swapping often creates non-paraphrases, while maintaining a close lexical distance with the source sentence. Here is a typical example of non-paraphrase couple from PAWS: "*flights from New York to Florida*" and "*flights from Florida to New York*".

4.2 New datasets

MRPC and PAWS are known datasets and the training data of MISTRAL is not disclosed. Since they could have been used to train MISTRAL, to avoid poisoning bias, we use three unpublished datasets: MCPG dataset, LLM generated dataset and HC dataset. They are provided as supplementary material.

4.2.1 MCPG dataset

The MCPG dataset contains sentences generated by Monte-Carlo Tree Search algorithm using statistical paraphrase generation framework (paraphrase tables with pivot language, ...) (Fabre et al., 2021). Couples are evaluated by at least three judges on syntactic quality and semantic equivalence. The majority vote is used to defined a label. In case of equality, the couple is discarded. Not syntactically correct sentences are also discarded.

Here is a typical example of non-paraphrase couple from MCPG: "*a very old and rusted train parked on the tracks.*" and "*a very old and dirty train is parked in the grass.*".

4.2.2 LLM generated dataset

To build the LLM generated dataset we used MISTRAL (Jiang et al., 2023) and LLAMA2 (Touvron et al., 2023) to generate paraphrases. These two models have not been fine-tuned to generate paraphrases. Source sentences are randomly picked up from PAWS and MRPC sets. Two prompt templates are used for MISTRAL and one for LLAMA2. Moreover, to create diversity, and be more likely to generate non-paraphrases, a vulgar template has been designed. As LLAMA2 refuses to generate with this template, it was only used with MISTRAL.

Hypothesis paraphrases generated with this template contain a wider range of vocabulary. This is highlighted by their mean LEV. scores in table 1. Each sentence pairs created has been classified as paraphrase or non-paraphrase by at least one human judge. It is interesting to note that LLMs seem capable to generate paraphrases, and most of the times very good paraphrases.

Here is an example of a non-paraphrase: "*Trading volume was incredibly light at 500.22 million shares, below an already thin 611.45 million exchanged at the same point Thursday.*" and "*The trading volume was significantly lower than usual on this day, with only 500.22 million shares exchanged compared to 611.45 million shares traded at the same time the previous day.*"

Details for reproducibility – prompts used and evaluation guidelines – are provided in appendix A.2 and A.3. The annotated corpus is provided as supplementary material.

4.2.3 HC dataset

The HC dataset has been handmade by two experts. Source sentences have been randomly generated and often adapted to be possible to be paraphrased. All paraphrases and non-paraphrases have been written by one expert. The second one has read every couple to assess its correctness. This dataset has been designed to contain couples that are obvious for human but tricky for metrics. Half couples of the dataset are questions.

Here is a typical example of a non-paraphrase sentence couple: "*He can't take the joke.*" and "*He can't take the joker.*". This is a typical example of paraphrase question couple: "*How does it feel to be pregnant?*" and "*How does it look to have a bun in the oven?*".

5 Experiments & Results

The experiment aims to evaluate each metrics on all introduced datasets and to compare them. To do so, we compute a score with each metric for every couple (of paraphrases or non-paraphrases) from all corpora. Following the notation from 3.1, we note the source sentence **S** and the associated candidate paraphrase **H**. As there is no other reference to compare **H**, our evaluation takes place in a reference-free context. Couples are labeled as paraphrase or non-paraphrase by a human judge, so we consider these data as gold.

Results are then analysed to understand the strengths and the weaknesses of ParaPLUIE and other metrics. We first consider the score distribution, then look upon the metric accuracy (with F1, recall and precision) and the a posteriori decision threshold which maximizes the accuracy. To supplement, we discuss about the correlation of all metrics with the edit distance (i.e LEV. distance).

5.1 Score distribution

Our first analysis is on the score distribution. Table 1 presents the mean and standard deviation for each metric on all corpora. We can observe that the mean edit distance, is the lowest on PAWS dataset and the highest on the LLM generated. We can also observe that the mean edit distance between couples labeled as paraphrase and non-paraphrase, in the MCPG dataset, is important. Since the content of datasets seems different, according to the edit distance, this offers us a broad overview of different paraphrases/non-paraphrases.

We can point out that mean scores of every metrics, excluding ParaPLUIE, strongly overlap. This can be explained by the deliberately misleading nature of the corpora considered in this experiment. We can observe that, ParaPLUIE mean scores, on paraphrase pairs, overlap less on non-paraphrase scores. Moreover, the mean scores of non-paraphrases are always lower than the paraphrases' ones.

5.2 Metric accuracy and threshold

Following the distribution analysis, let us now look at the best accuracy of each metric on each dataset and their according information (F1, recall, precision and decision threshold).

We process an *a posteriori* analysis by considering a classification threshold. Sentence couples with a score under a given threshold are labeled non-paraphrases whereas others are labeled paraphrases. This classification is then compared with the ground truth. To determine the optimal decision threshold, we compute the accuracy on each possible threshold. The best *a posteriori* thresholds, with associated information, are reported table 3. We also provide results for the *a priori* ParaPLUIE's threshold of 0 to investigate its performances among the corpora.

In order to have a corpus with several types of sentence couples, the Mixed Balanced Dataset (MBD) is introduced in table 3. It is a balanced mix composed from all introduced corpora. It contains half paraphrases and half non-paraphrases: for every label, 45 couples are randomly picked from each dataset. The results are the means of 100 draws of this dataset and the 95% confidence score. Performances on MBD allow to evaluate the robustness of the metrics among different kind of corpus. Indeed, an *a posteriori* best threshold for a corpus could be a bad threshold for another one.

In table 3, we can notice, as expected, that a good threshold for a corpus is not applicable on another one, excepted for ParaPLUIE. Moreover, $\text{BERT}_{\text{score}}$ is not performing well on PAWS corpus, as explained by Zhang et al. (2020). We can also observe that choosing a threshold common to all datasets for a given metric significantly reduces its performance. This indicates that metrics struggle to correctly classify paraphrases and they are not resilient.

By focusing on the results of the different ParaPLUIE templates, we can see that their accuracy is significantly higher than others metrics on all datasets, except for the LLM corpus. Our assumption is that LLM for generation and LLM for evaluation share the sames flaws.

ParaPLUIE FS-DIRECT seems to be the best template since it provides the best accuracy and the best F1 overall. Its best *a posteriori* threshold on MBD is really close to zero. This is convenient because it follows the inherent natural threshold of ParaPLUIE.

Surprisingly, we can notice that the LEV. distance is not struggling much. Obviously LEV. is not a good metric for semantic evaluation. Since the scores obtained by others metrics and LEV. are close, ones can wondering if they are correlated. However ParaPLUIE may have a different behavior. This point is addressed in the following section.

5.3 Correlation with edit distance

To confirm our previous assumption, the Pearson correlation between metrics and edit distance is presented in table 4. It is focused on correlation inside each class – paraphrase/non-paraphrase. Indeed, the extent of belonging to a category should be related to semantic distance and not lexical distance. Undoubtedly, other metrics than ParaPLUIE are correlated with the edit distance. This is a concern because the semantic proximity estimation among two sentences should not be guided by the edit distance that separates them. The PAWS example presented in section 4.1.2 of non-paraphrase highlights this very well. We can observe that, all

		MRPC		PAWS		MCPG		LLM		HC		MBD	
Lev	*Threshold*	*0.52*		*0.12*		*0.60*		*0.87*		*0.80*		*0.39* $\pm$*0.01*	
	Acc.	0.69		0.70		0.79		0.79		0.51		0.61 $\pm$0.00	
	F1	0.78		0.55		0.86		0.88		0.67		0.61 $\pm$0.01	
	Recall	0.81		0.41		0.95		**1.00**		**1.00**		0.61 $\pm$0.02	
	Precision	0.75		**0.84**		0.79		0.79		0.50		0.62 $\pm$0.00	
Bleu	*Threshold*	*0.00*		*0.67*		*0.53*		*0.00*		*0.00*		*0.44* $\pm$*0.02*	
	Acc.	0.67		0.67		0.76		0.79		0.50		0.61 $\pm$0.00	
	F1	0.80		0.54		0.80		0.88		0.67		0.56 $\pm$0.01	
	Recall	**1.00**		0.47		0.71		**1.00**		**1.00**		0.50 $\pm$0.02	
	Precision	0.67		0.69		0.79		0.79		0.50		0.65 $\pm$0.01	
Meteor	*Threshold*	*0.52*		*0.92*		*0.72*		*0.26*		*0.38*		*0.66* $\pm$*0.01*	
	Acc.	0.73		0.59		0.83		0.80		0.54		0.62 $\pm$0.00	
	F1	0.81		0.53		0.88		0.88		0.64		0.65 $\pm$0.00	
	Recall	0.87		0.51		0.85		0.98		0.80		0.70 $\pm$0.01	
	Precision	0.76		0.54		0.90		0.81		0.53		0.61 $\pm$0.00	
BertScore	*Threshold*	*0.73*		*0.96*		*0.76*		*0.61*		*0.76*		*0.79* $\pm$*0.01*	
	Acc.	0.73		0.64		0.84		0.80		0.57		0.64 $\pm$0.00	
	F1	0.82		0.48		0.89		0.89		0.59		0.68 $\pm$0.00	
	Recall	0.88		0.38		0.92		0.99		0.63		0.77 $\pm$0.01	
	Precision	0.76		0.68		0.86		0.80		0.56		0.62 $\pm$0.00	
ParaScore	*Threshold*	*0.74*		*1.00*		*0.82*		*0.61*		*0.77*		*0.82* $\pm$*0.00*	
	Acc.	0.72		0.56		0.82		0.80		0.56		0.64 $\pm$0.00	
	F1	0.80		0.00		0.87		0.89		0.59		0.66 $\pm$0.00	
	Recall	0.86		0.00		0.87		0.99		0.62		0.72 $\pm$0.01	
	Precision	0.75		0.40		0.87		0.80		0.56		0.62 $\pm$0.00	
ParaPLUIE DIRECT	*Threshold*	*8.76*	*0*	*22.84*	*0*	*10.21*	*0*	*12.42*	*0*	*8.57*	*0*	*17.94* $\pm$*0.58*	*0*
	Acc.	**0.78**	**0.78**	0.69	0.56	0.90	0.90	**0.83**	0.82	**0.90**	**0.90**	0.77 $\pm$0.00	0.73 $\pm$0.00
	F1	**0.85**	**0.85**	0.69	0.66	0.93	0.93	**0.90**	0.89	**0.90**	**0.90**	0.79 $\pm$0.00	0.78 $\pm$0.00
	Recall	0.89	0.93	0.77	**0.97**	0.94	**0.97**	0.97	0.98	0.88	0.91	0.88 $\pm$0.01	**0.95** $\pm$0.00
	Precision	0.81	0.78	0.62	0.50	0.92	0.89	**0.84**	0.82	0.93	0.89	0.72 $\pm$0.00	0.66 $\pm$0.00
ParaPLUIE INDIRECT	*Threshold*	*−8.26*	*0*	*19.10*	*0*	*−7.76*	*0*	*−11.90*	*0*	*−9.12*	*0*	*5.52* $\pm$*1.85*	*0*
	Acc.	**0.78**	0.76	0.65	0.64	**0.92**	0.89	0.81	0.80	0.87	0.83	0.76 $\pm$0.00	0.75 $\pm$0.00
	F1	0.84	0.82	0.69	0.69	**0.95**	0.92	0.89	0.88	0.88	0.81	0.78 $\pm$0.00	0.77 $\pm$0.00
	Recall	0.88	0.79	0.86	0.90	**0.97**	0.88	0.98	0.93	0.89	0.72	0.85 $\pm$0.01	0.84 $\pm$0.00
	Precision	0.81	**0.85**	0.57	0.55	0.92	0.96	0.82	**0.84**	0.86	**0.94**	0.72 $\pm$0.00	0.72 $\pm$0.00
ParaPLUIE FS-DIRECT	*Threshold*	*−3.25*	*0*	*5.53*	*0*	*−2.02*	*0*	*−7.53*	*0*	*−5.70*	*0*	*1.29* $\pm$*0.51*	*0*
	Acc.	0.76	0.75	**0.77**	0.75	0.91	0.90	0.81	0.79	0.89	0.86	**0.79** $\pm$0.00	0.77 $\pm$0.00
	F1	0.82	0.80	0.75	**0.76**	0.93	0.92	0.88	0.87	0.89	0.85	**0.80** $\pm$0.00	0.79 $\pm$0.00
	Recall	0.83	0.77	0.79	0.88	0.93	0.88	0.94	0.91	0.86	0.78	0.84 $\pm$0.01	0.84 $\pm$0.00
	Precision	0.82	0.84	0.71	0.66	0.94	**0.97**	0.83	0.83	0.92	**0.94**	**0.76** $\pm$0.00	0.74 $\pm$0.00

Table 3: F1, recall, precision and associate threshold of each metrics on every datasets, according to their best classification accuracy. Best accuracy's is in **orange**, F1 in **blue**, recall in **black** and precision in **violet**.

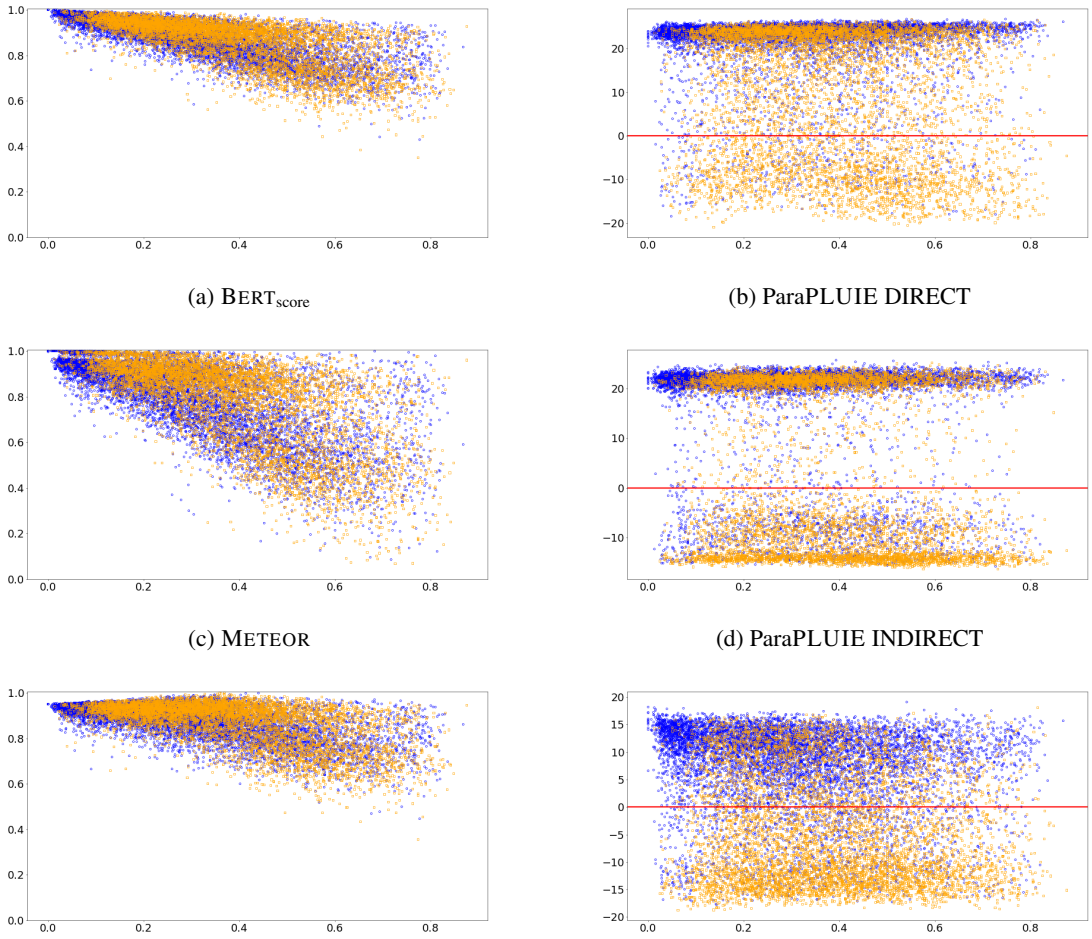
ParaPLUIE's templates are less correlated with the edit distance. This observation is also illustrated by figure 2. More precisely, $\text{BERT}_{\text{score}}$, ParaScore and METEOR scores are linked to the edit distance and they are not able to create clusters of paraphrases/non-paraphrases. On the opposite, different ParaPLUIE's templates are clearly less correlated and are able to cluster sentence pairs. We can regret that many false positives are in the paraphrase clusters of the templates DIRECT and INDIRECT. The ParaPLUIE FS-DIRECT makes less false positive and false negative errors. It is interesting to point out that, between the two clusters made by ParaPLUIE FS-DIRECT we can notice an area of uncertainty. The closer the score is to zero, the less the system is certain to classify the hypothesis sentence. This is a really attractive property as it enhances the natural dynamic of the measure. In other words, the higher the score of a hypothesis sentence is, the higher we can be confident in the classification and vice versa.

6 Conclusion

We propose ParaPLUIE, a new metric for evaluating semantic proximity between two sentences. ParaPLUIE is relying on a learnt probabilistic model of LLM. It is designed to return scores that can be directly interpreted thanks to its natural threshold. We have conducted experiments with various templates for ParaPLUIE on several English paraphrase corpora. We have created two new paraphrase corpora and use one not publicly

	MRPC		PAWS		MCPG		LLM		HC	
	yes	no	yes	no	yes	no	yes	no	yes	no
BLEU	−0.67	−0.60	−0.66	−0.57	−0.83	−0.84	−0.59	−0.64	−0.66	−0.71
METEOR	−0.63	−0.57	−0.45	−0.47	−0.75	−0.85	−0.53	−0.59	−0.66	−0.87
$\text{BERT}_{\text{score}}$	−0.72	−0.63	−0.61	−0.55	−0.77	−0.86	−0.63	−0.63	−0.80	−0.79
ParaScore	−0.56	−0.54	−0.10	−0.23	−0.64	−0.84	−0.56	−0.56	−0.74	−0.68
ParaPLUIE DIRECT	**−0.08**	**−0.14**	−0.07	**−0.02**	**−0.06**	−0.23	**−0.00**	**−0.13**	0.01	**0.04**
ParaPLUIE INDIRECT	−0.13	−0.15	**−0.04**	**−0.02**	−0.13	**−0.09**	−0.04	**−0.13**	−0.10	**0.04**
ParaPLUIE FS-DIRECT	−0.17	−0.15	−0.05	−0.13	−0.17	−0.25	−0.08	−0.22	**0.00**	−0.06

Table 4: Pearson correlation coefficients between evaluated metrics and the edit distance for each corpus and each class. Emphasis is placed on the weakest correlations.



(a) $\text{BERT}_{\text{score}}$ (b) ParaPLUIE DIRECT

(c) METEOR (d) ParaPLUIE INDIRECT

(e) ParaScore (f) ParaPLUIE FS-DIRECT

Figure 2: Score distribution of couples of all datasets in regards of the edit distance. Blue circles represent paraphrase and non-paraphrase are represented in orange squares. $\text{BERT}_{\text{score}}$, METEOR are ParaScore are between zero and one. For ParaPLUIE, the red line denotes the natural threshold at zero.

released to avoid poisoning bias. All of them are human annotated and are shared with this paper. Our experiments have shown that ParaPLUIE performs better than commonly used measures. Interesting properties of ParaPLUIE are the interpretable threshold between paraphrases and non paraphrases, and it is marginally correlated to the edit distance. In future works, we would like to create Small Language Models dedicated to the generation of paraphrases, thanks to knowledge distillation (Hsieh et al., 2023) and using ParaPLUIE as a loss.

7 Ethical considerations

It is important to keep in mind that, we do not trained or fine-tuned the LLM used for ParaPLUIE. Training could lead to better results, but that is uncertain, as that could remove some knowledge in the model. The best template ParaPLUIE FS-DIRECT does not need a generation step and MISTRAL is a medium size language model. For these reasons, scoring with ParaPLUIE is not much computing intensive.

Obviously, the use of a larger LLM could lead to better results. Nevertheless we appeal to not do that. As we live in a world of limited resources and energy, the research effort should be put into the adaptation and creation of small model dedicated to this task.

8 Limitations

This section aims to discuss about other limits than those already discussed.

Experiments in this study were lead on a limited quantity of data. The entire PAWS corpus was not used but only the *dev* subset. This is due to the high computational cost needed to use an LLM, specifically with the ParaPLUIE INDIRECT which needs a generation step. Results on the whole PAWS dataset may vary.

LLAMA2 and MISTRAL, although producing different results, are likely to be trained on very similar data. They both have TRANSFORMER-style architecture and have the same magnitude of weights. Other model architecture may produce different results.

Most sentence pairs inside the LLM dataset made for this experiment were labeled by only one human. Hence, inter-annotator agreement is not available. This corpus turns out to be highly unbalanced since the generation systems produce good paraphrase overall. Moreover it appears that none of the metrics are able to perform very well on it.

The LLM used for ParaPLUIE could have been trained on MRPC or PAWS. This potentially data contamination is set apart by datasets made for this experiment.

For ParaPLUIE FS-DIRECT, since sources sentences used to build the 6 examples are extracted from MRPC and PAWS, they shared the "style" as the considered corpora. Nonetheless evaluations on other dataset present good results event if the "style" differs.

Throughout this paper, three ParaPLUIE versions have been shown. We have tested other templates with different generation step strategies. The prompt plays a critical role here, as small changes in it can involve major differences in scoring. Templates versions proposed may not be optimal.

Acknowledgments

Research supported by Ministère des Armées - Agence de l'Innovation de la Défense. We would like to thank Betty FABRE for sharing the MCPG Dataset and Marjorie LATGER for helping to create the HC Dataset.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson,

Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Betty Fabre, Tanguy Urvoy, Jonathan Chevelu, and Damien Lolive. 2021. Neural-driven search-based paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2100–2111, Online. Association for Computational Linguistics.

Cédric Fayet, Alexis Blond, Grégoire Coulombel, Claude Simon, Damien Lolive, Gwénolé Lecorvé, Jonathan Chevelu, and Sébastien Le Maguer. 2020. FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, pages 22–25, Nancy, France. ATALA.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: Association for Computational Linguistics 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Vladimir Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk SSSR*, 163:845–848.

Igor M Mel'čuk. 1997. *Vers une linguistique sens-texte: leçon inaugurale faite le vendredi 10 janvier 1997*. Collège de France.

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.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. 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*. Association for Computational Linguistics.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, Brussels, Belgium. Association for Computational Linguistics.

Satoshi Sekine. 2005. Automatic paraphrase discovery based on context and keywords between NE pairs. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of North American Chapter of the Association for Computational Linguistics*.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842, Suntec, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Results with SBERT

As sentence-transformers *(SBERT)* (Reimers and Gurevych, 2019) all-MiniLM-L6-v2 [5] is the most famous tool on Hugging Face for sentence similarity with more than 45 Millions of downloads during August 2024 we wanted to include it in this work. Nevertheless, as it is a fine-tuned Bert which use a cosine similarity to compute a score and as the result with this metrics don't differ much than $\text{BERT}_{\text{score}}$ we choose to exclude it from the core of the paper. We report the result of sentence-transformers in the table 5. While results are, as expected, very similar to $\text{BERT}_{\text{score}}$, it performs better on the HC dataset. It is also less correlated than $\text{BERT}_{\text{score}}$ with the edit distance.

A.2 Templates used to generate paraphrases with LLM

We used two models to generate paraphrases, MISTRAL and LLAMA2. Both models are fine-tuned to work as conversational agents. The templates mimic a dialog with a user and the assistant. Here **S** denote the sentence which we want to paraphrase. To create diversity in generated paraphrases, we create a vulgar template. This one shows an example of paraphrase for a vulgar sentence. We empirically observe that, with this kind of example the model is more inventive and tends to use a wider range of vocabulary.

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

MISTRAL naive(S) :
(*user*): *You have to transform a sentence A into a paraphrase B. The purpose of A to B transformation is to maintain the original meaning of the sentence A in B.*
(*assistant*): *I've understood the instructions perfectly.*
(*user*): *A: "The little cat refreshes himself with water every morning."*
(*assistant*): *B: "Every morning, the little cat refreshes himself by drinking water."*
(*user*): *A: "**S**"*

LLAMA2 naive(S) :
(*system*): *You have to transform a sentence A into a paraphrase B. The purpose of A to B transformation is to maintain the original meaning of the sentence A in B.*
(*user*): *A: "The little cat refreshes himself with water every morning."*
(*assistant*): *B: "Every morning, the little cat refreshes himself by drinking water."*
(*user*): *A: "**S**"*

MISTRAL vulgar(S) :
(*user*): *You have to transform a sentence A into a paraphrase B. The purpose of A to B transformation is to maintain the original meaning of the sentence A in B. You can use offensive and disrespectful statements.*
(*assistant*): *I've understood the instructions perfectly.*
(*user*): *A: "Fuck both of you."*
(*assistant*): *B: "Go fuck yourselves."*
(*user*): *A: "**S**"*

A.3 Details about the LLM generated paraphrase dataset

Each judge had to self-assess his English level. They were able to select between, poor, good, advanced and native to self-assess their confidence in their English. To avoid bias in judgement, every button to choose was in grey except for the "Don't know" option which was colored in light-blue. To help judges, word differences between sentences were highlighted in yellow. Additionally, explanations about their task and examples of expected responses were available at any moment. Here are the explanations provided : *"You are going to see two sentences. You are asked to estimate the extent to which the two sentences share a com-*

	MRPC		PAWS		MCPG		LLM		HC		MBD
	According to max. accuracy 95% confidence interval on MBD										
Threshold	*0.68*		*0.99*		*0.83*		*0.75*		*0.65*		*0.79* $\pm$*0.01*
Acc.	0.73		0.62		0.83		0.80		0.71		0.66 $\pm$0.00
F1	0.81		0.43		0.88		0.88		0.75		0.71 $\pm$0.00
Recall	0.87		0.32		0.88		0.95		0.86		0.84 $\pm$0.01
Precision	0.75		0.64		0.87		0.82		0.67		0.62 $\pm$0.00
	yes: paraphrase no: non-paraphrase										
	yes	no	yes	no	yes	no	yes	no	yes	no	
Mean & STD.	0.83 $\pm$0.12	0.71 $\pm$0.15	0.97 $\pm$0.04	0.96 $\pm$0.05	0.91 $\pm$0.13	0.70 $\pm$0.19	0.89 $\pm$0.09	0.83 $\pm$0.12	0.80 $\pm$0.16	0.63 $\pm$0.21	
Pears Corr.	-0.41	-0.29	-0.20	-0.17	-0.66	-0.76	-0.35	-0.40	-0.40	-0.47	

Table 5: Results of sentence-transformers for each corpus. The scores have been calculated like the other metrics on the same datasets as in the main paper. Results are very similar to $\text{BERT}_{\text{score}}$. We can notice that it performs better than $\text{BERT}_{\text{score}}$ on the HC dataset. It is also less correlated than $\text{BERT}_{\text{score}}$ with the edit distance.

mon meaning. To help you, the differences between the sentences are highlighted in yellow of which several examples are shown bellow." Here are the examples provided:

- The cat drinks water.
- The cat eats kibble.
- Very different
- The associated actions have nothing in common even though the two sentences have the same subject.
- The cat drinks milk.
- The cat drinks water.
- Slightly similar
- The subjects and actions are similar, but water is not milk.
- The cat drinks water.
- The cat quenches its thirst.
- Mostly similar
- The only difference is that the first sentence specifies the type of liquid that is being drunk.
- The cat eats the mouse.
- The mouse is eaten by the cat.
- Same meaning
- Differences linked to context interpretation in these two sentences are too small to say that their meaning is different.
- The cat drinks tomato soup.
- Cat tomato soup.
- Don't know
- The second sentence doesn't make any sense. We can't draw any conclusions from it.

We have generated 605 hypothetical paraphrases with LLM from 605 source sentences. Each hypothesis paraphrase has been classified by at least one human judge. The evaluators were volunteers, non-experts in NLP domain. The evaluation protocol was as follows. Each judge was proposed to label up to 55 couples on a web-application (Fayet et al., 2020) in which 5 couples were reserved for the training trial. The training trial was the same for all judges.

Sentence pairs have been shown one by one, one sentence above the other. Presentation order of the sentence pair is chosen randomly. For each pair, judges had 5 possible answers: [Very different, Slightly similar, Mostly similar, Same meaning, Don't know], presented in this order. Evaluation guidelines with examples were also provided.

At the end of the evaluations, 276 couples have been labeled as "Same meaning", 181 as "Mostly similar", 93 "Slightly similar", 28 "Very different" and 22 "Don't know". We consider couples labeled as "Very different" and "Slightly similar" as non-paraphrases. "Mostly similar" and "Same meaning" labeled couples are considered as paraphrases.

Overall, 457 couples have been labeled as paraphrases, 121 as non-paraphrases, 22 as indeterminate and 5 have been used for the training trial and were not taken in account. We did not include couples labeled as indeterminate in the dataset.

A.4 Complete few shots template

Tpl$_{\text{FS-Direct}}$(**S**, **H**):
($user$): *You will receive two sentences A and B. Do these two sentences mean the same thing? Answer with only one word "yes" or "no".*
($assistant$): *Please provide the sentences for me to evaluate.*
($user$): *A: "Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence ."; B: "Amrozi accused his brother, whom he disparagingly referred to as 'the liar witness', of intentionally twisting his testimony."*
($assistant$): *No*
($user$): *A: "Pennmakkal is an Indian Malayalam film from 1966, produced by J. Sasikumar and directed by KP Kottarakkara."; B: "The Indian Malayalam film 'Pennmakkal', released in 1966, was produced by J. Sasikumar and directed by KP Kottarakkara."*
($assistant$): *Yes*
($user$): *A: "Sorkin , who faces charges of conspiracy to obstruct justice and lying to a grand jury , was to have been tried separately."; B: "Despite being accused of conspiring to obstruct justice and perjury, Sorkin was supposed to stand trial on his own."*
($assistant$): *No*
($user$): *A: "Gilroy police and FBI agents described Gehring as cooperative , but said Saturday that he had revealed nothing about what had happened to the children ."; B: "Although Gilroy police and FBI agents reported that Gehring was cooperative , he hadn't disclosed any information about the children's whereabouts or what had happened to them as of Saturday ."*
($assistant$): *No*
($user$): *A: "Whereas "e" the electric charge of the particle and A is the magnetic vector potential of the electromagnetic field."; B: "The electric charge of the particle is denoted by "e", and the magnetic vector potential of the electromagnetic field is denoted by 'A'."*
($assistant$): *Yes*
($user$): *A: "The Jidanul River is a tributary of the Jiul de Vest River in Romania."; B: "The Jidanul River is a mere insignificant stream that flows into the grand Jiul de Vest River in Romania."*
($assistant$): *No*
($user$): *A: "***S***"; B: "***H***"*

A.5 Additional results

This section aims to give a deeper look in the behavior of each metrics on each corpus.

A.5.1 Quality of classification - Score centered

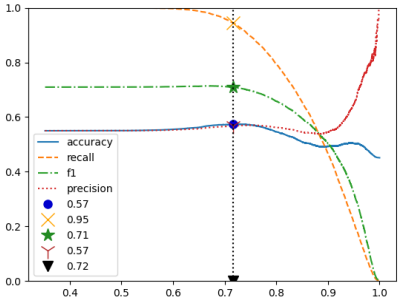
In this section, you will find the accuracy, F1, recall and precision score evolution of each metrics on each dataset. The y-axis is the score range of the metric and the x-axis is the score grant by a metric to couples of sentences composing a dataset. Each couples has been ranked according to it's score. The figures 3, 4, 5, 6, 7 and 8 respectively present accuracy, F1, recall and precision score for each metric for dataset Global, MRPC, PAWS, MCPG, LLM and HC. Global denotes all corpora concatenated.

A.5.2 Quality of classification - Rank centered

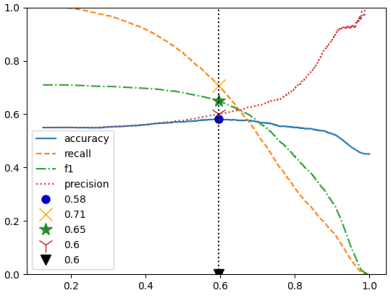
In this section, you will find accuracy, F1, recall and precision evolution of each metrics on each dataset. The y-axis is the score range of the metric and the x-axis is the rank of couples of sentences composing a dataset according to their score grant by metric. Each couples has been rank according to it's score. The figures 9, 10, 11, 12, 13 and 14 respectively present accuracy, F1, recall and precision score for each metric for dataset Global, MRPC, PAWS, MCPG, LLM and HC. Global denote all corpora concatenated.

A.5.3 Correlation with edit distance

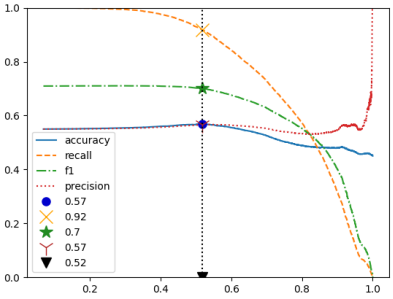
In this section, you will find the correlation of each metric with the edit distance on each dataset. The y-axis is the score distribution of couples of all datasets in regards of the edit distance in x-axis. Blue circles represent paraphrase and non-paraphrase are represented in orange squares. The figures 15, 16, 17, 18, 19 and 20 respectively present that for each metric for dataset Global, MRPC, PAWS, MCPG, LLM and HC. Global denote all corpora concatenated.



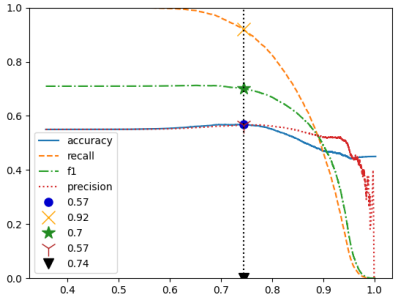
(a) BERT$_{\text{score}}$



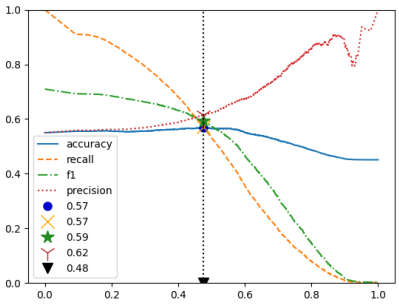
(b) LEV.



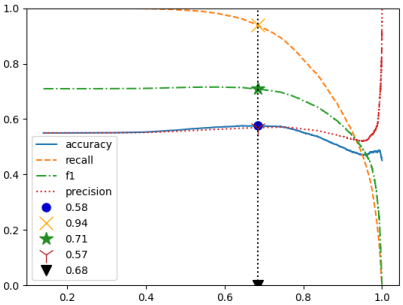
(c) METEOR



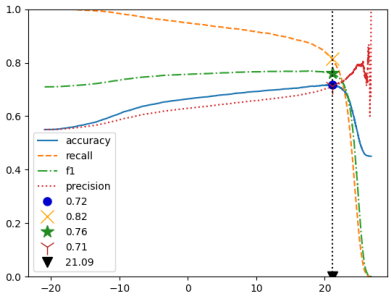
(d) ParaScore



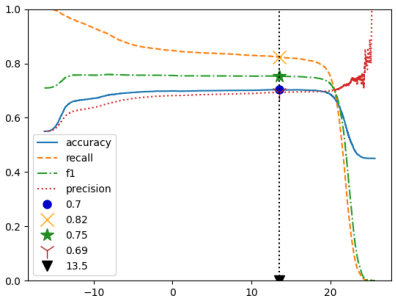
(e) BLEU



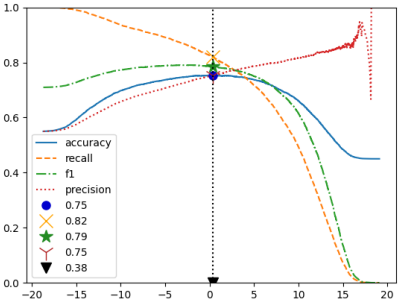
(f) sentence-transformers



(g) ParaPLUIE DIRECT

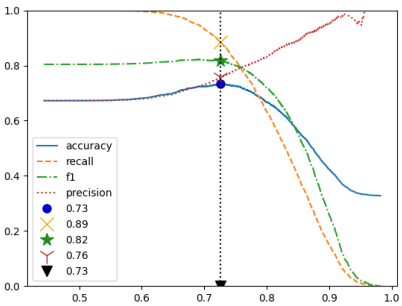


(h) ParaPLUIE INDIRECT

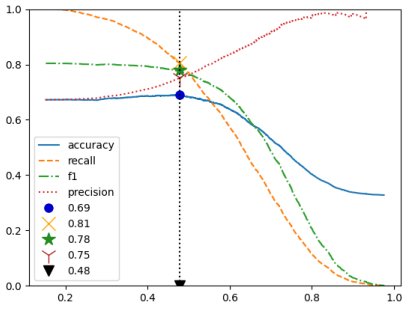


(i) ParaPLUIE FS-DIRECT

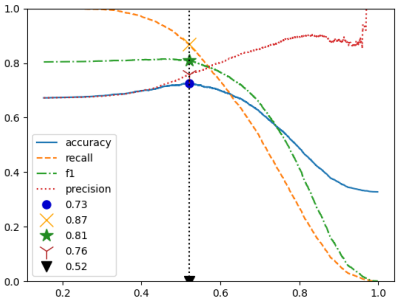
Figure 3: Global corpus



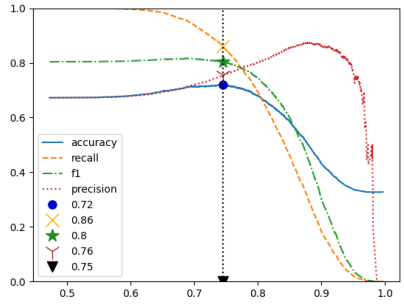
(a) BERT$_{\text{score}}$



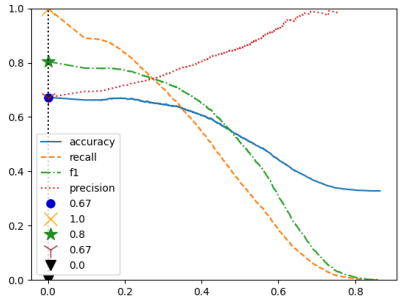
(b) LEV.



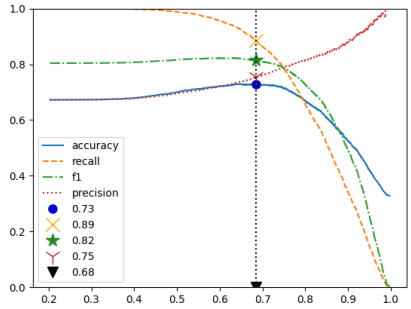
(c) METEOR



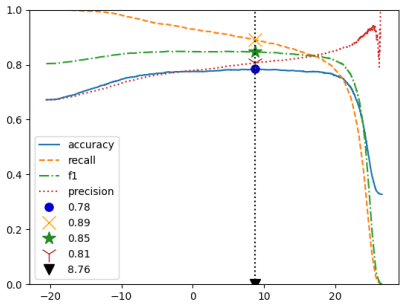
(d) ParaScore



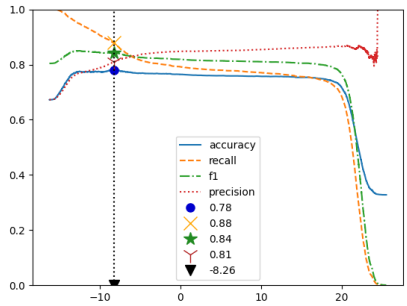
(e) BLEU



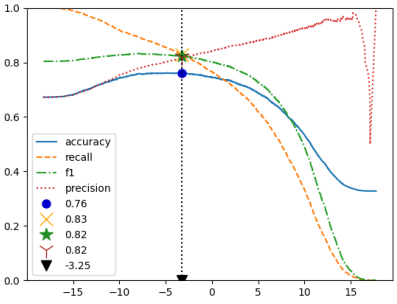
(f) sentence-transformers



(g) ParaPLUIE DIRECT

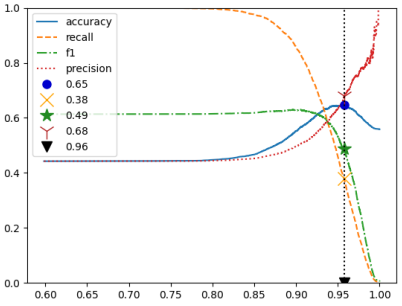


(h) ParaPLUIE INDIRECT

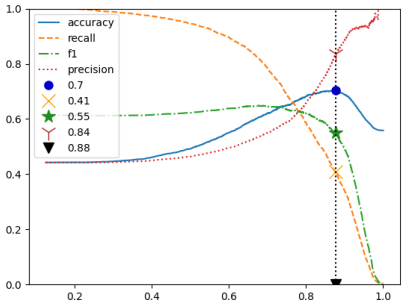


(i) ParaPLUIE FS-DIRECT

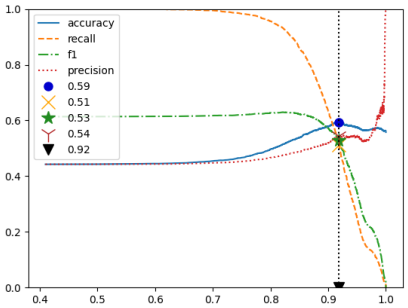
Figure 4: MRPC Corpus



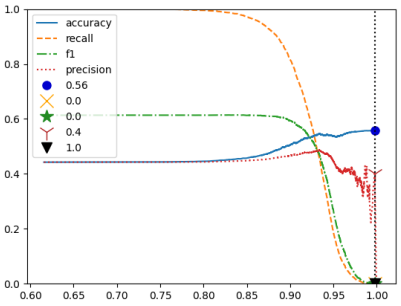
(a) BERT$_{\text{score}}$



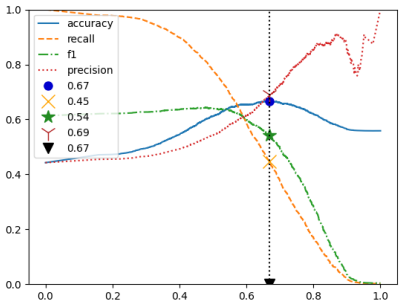
(b) LEV.



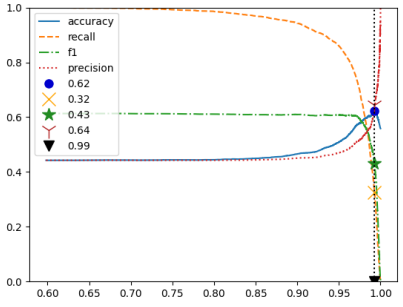
(c) METEOR



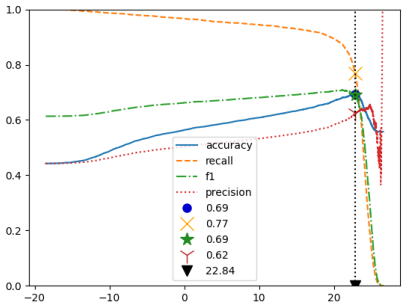
(d) ParaScore



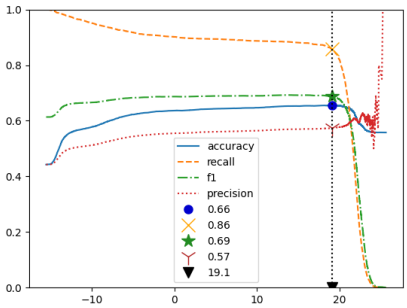
(e) BLEU



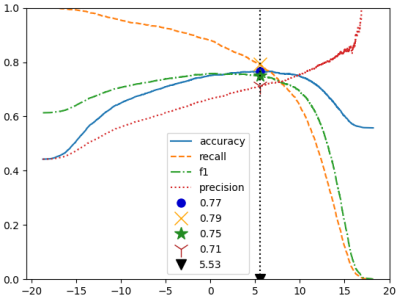
(f) sentence-transformers



(g) ParaPLUIE DIRECT

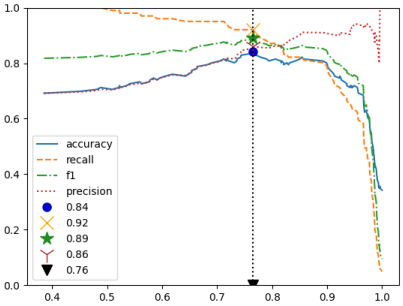


(h) ParaPLUIE INDIRECT

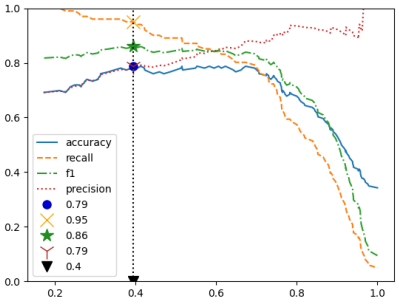


(i) ParaPLUIE FS-DIRECT

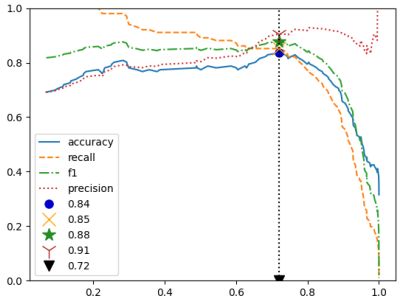
Figure 5: PAWS Corpus



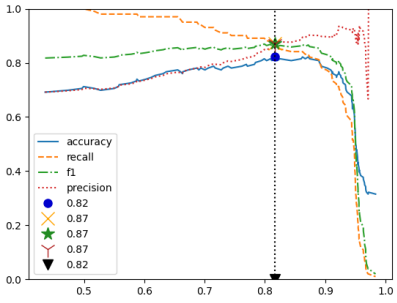
(a) BERT$_{\text{score}}$



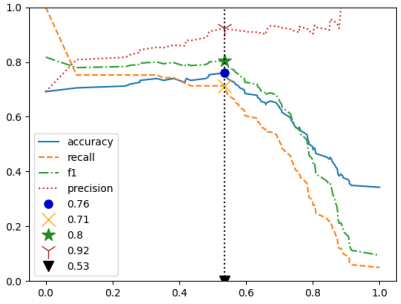
(b) LEV.



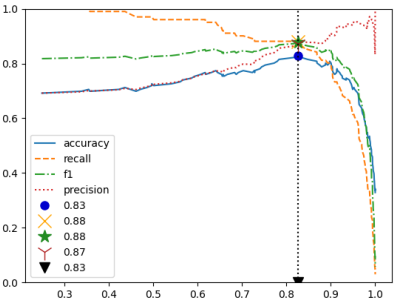
(c) METEOR



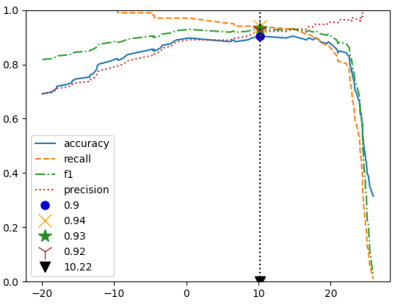
(d) ParaScore



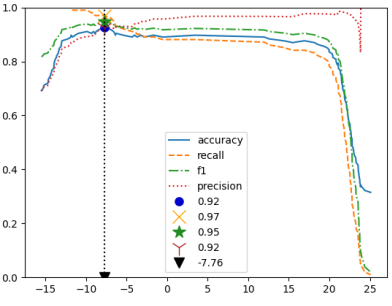
(e) BLEU



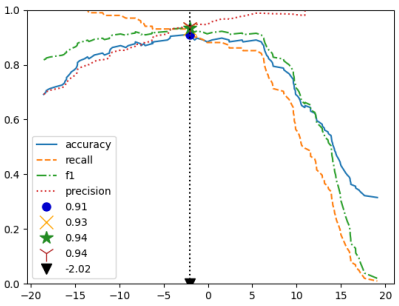
(f) sentence-transformers



(g) ParaPLUIE DIRECT

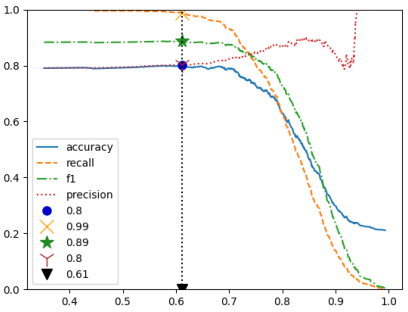


(h) ParaPLUIE INDIRECT

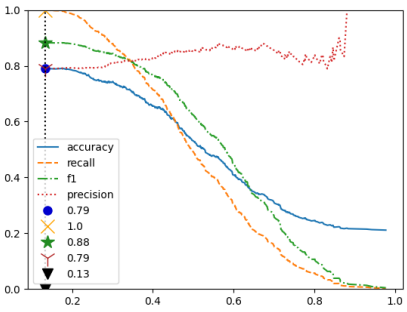


(i) ParaPLUIE FS-DIRECT

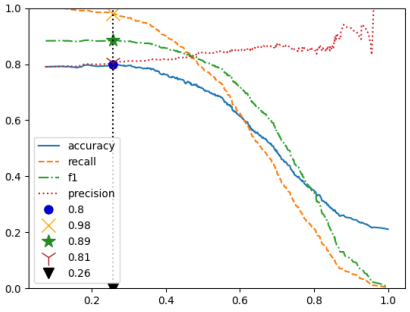
Figure 6: MCPG Corpus



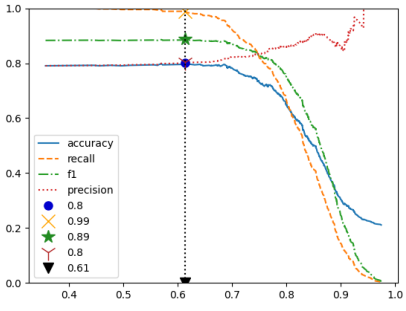
(a) BERT$_{\text{score}}$



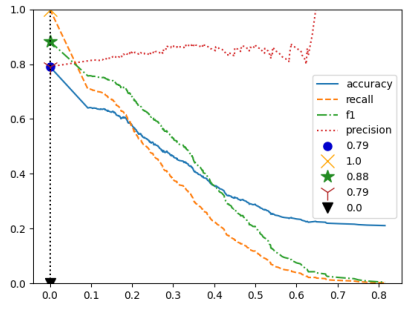
(b) LEV.



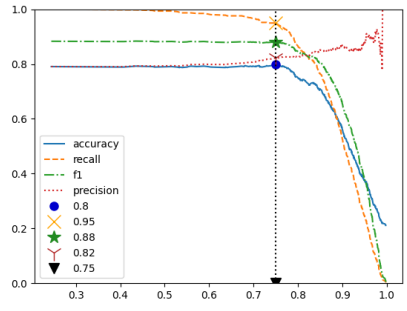
(c) METEOR



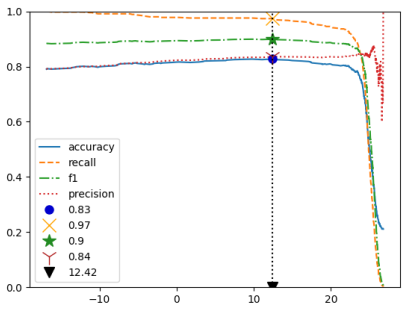
(d) ParaScore



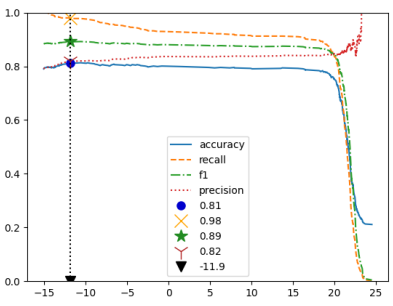
(e) BLEU



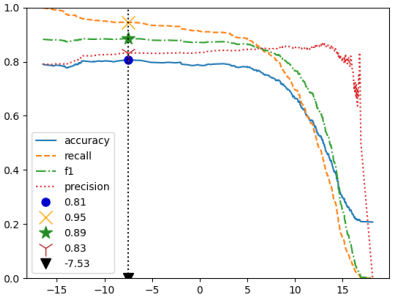
(f) sentence-transformers



(g) ParaPLUIE DIRECT

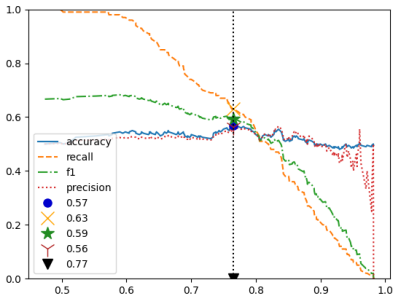


(h) ParaPLUIE INDIRECT

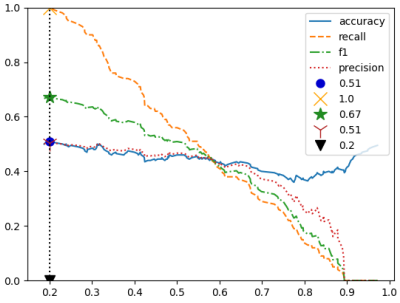


(i) ParaPLUIE FS-DIRECT

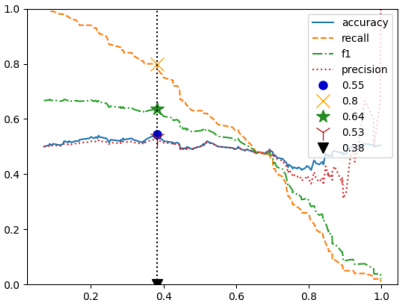
Figure 7: LLM Corpus



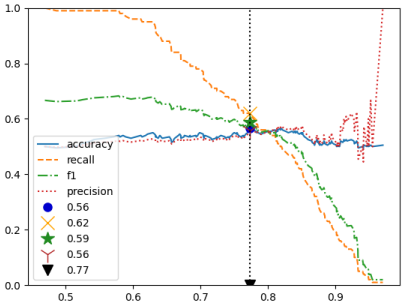
(a) BERT$_{\text{score}}$



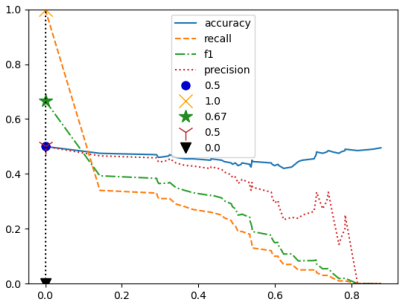
(b) LEV.



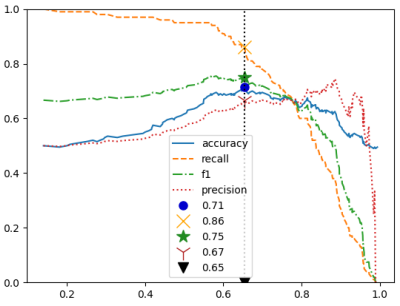
(c) METEOR



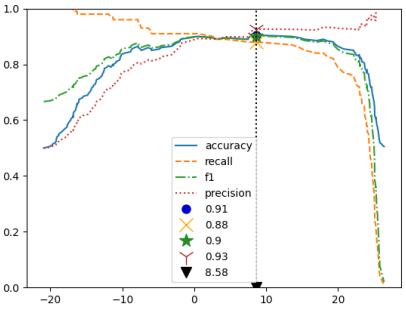
(d) ParaScore



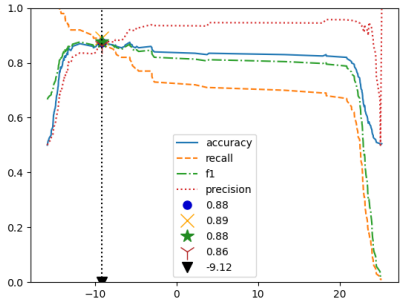
(e) BLEU



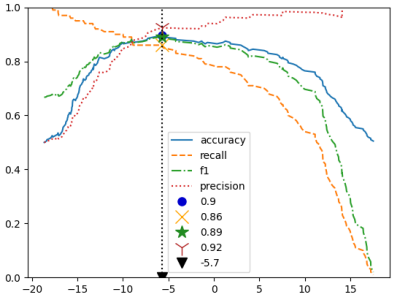
(f) sentence-transformers



(g) ParaPLUIE DIRECT

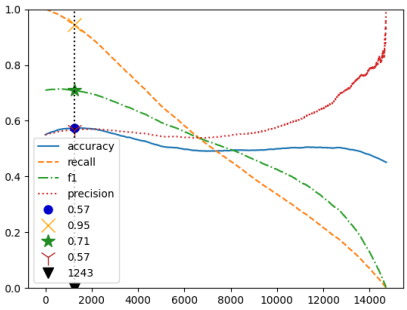


(h) ParaPLUIE INDIRECT

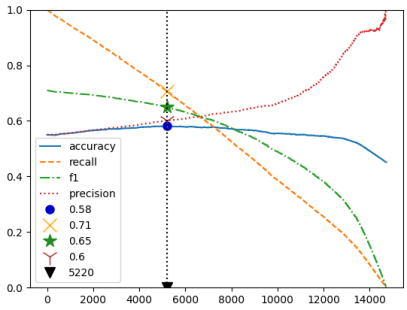


(i) ParaPLUIE FS-DIRECT

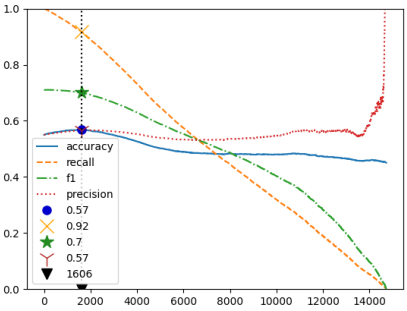
Figure 8: HC Corpus



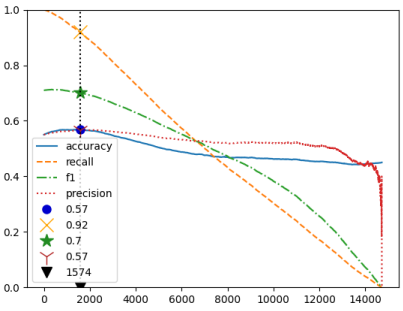
(a) BERT$_{\text{score}}$



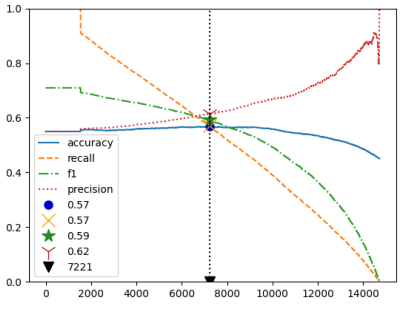
(b) LEV.



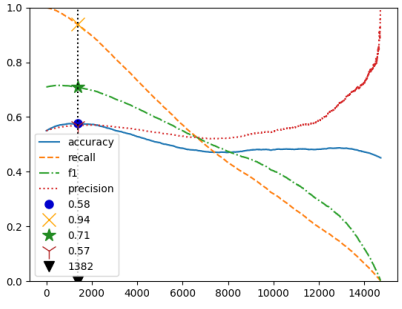
(c) METEOR



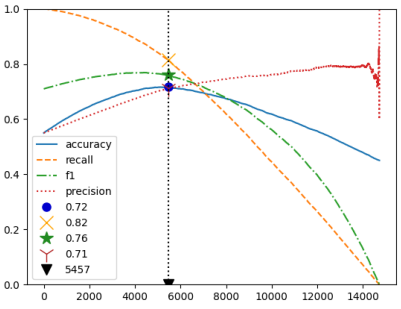
(d) ParaScore



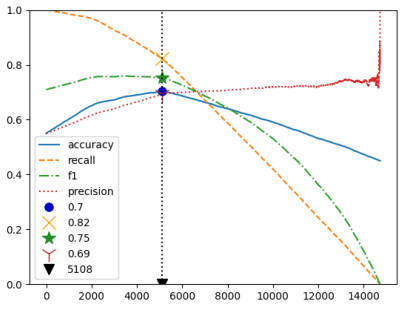
(e) BLEU



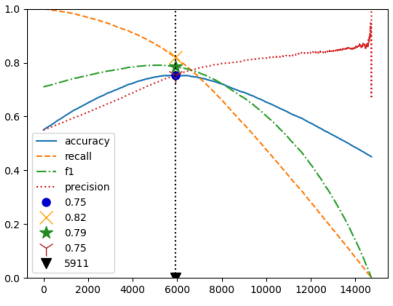
(f) sentence-transformers



(g) ParaPLUIE DIRECT

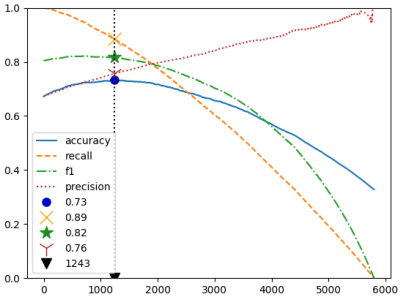


(h) ParaPLUIE INDIRECT

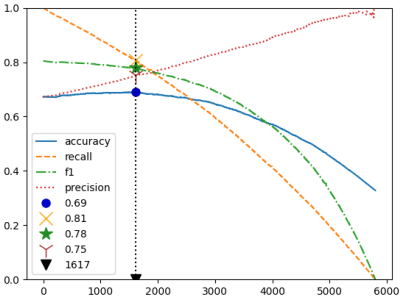


(i) ParaPLUIE FS-DIRECT

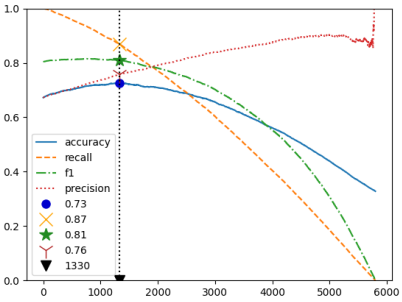
Figure 9: Global Corpus



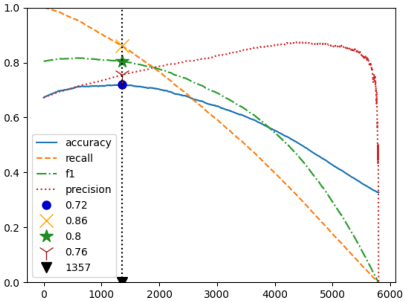
(a) BERT$_{\text{score}}$



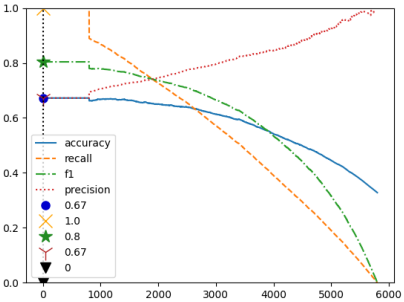
(b) LEV.



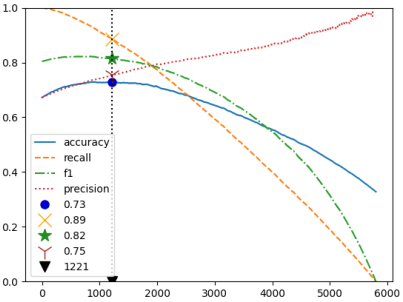
(c) METEOR



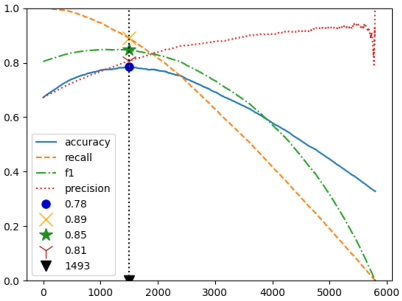
(d) ParaScore



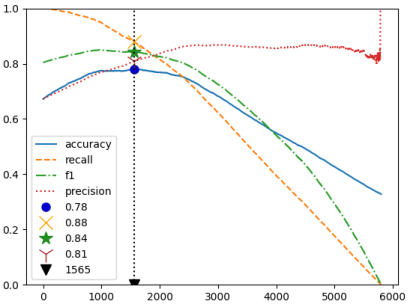
(e) BLEU



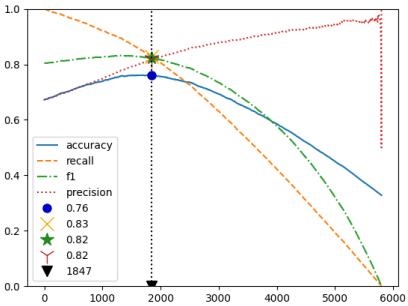
(f) sentence-transformers



(g) ParaPLUIE DIRECT

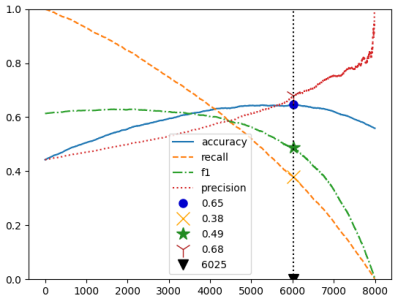


(h) ParaPLUIE INDIRECT

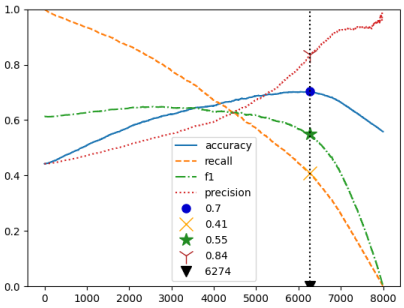


(i) ParaPLUIE FS-DIRECT

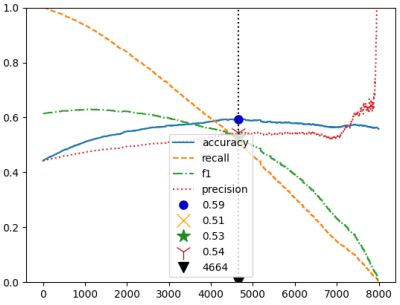
Figure 10: MRPC Corpus



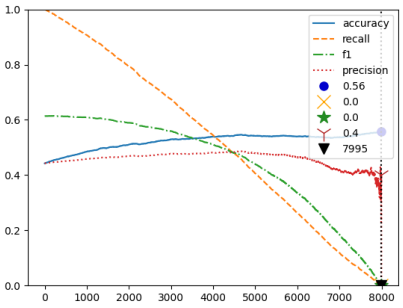
(a) BERT$_{\text{score}}$



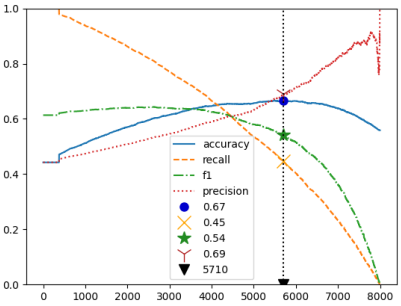
(b) LEV.



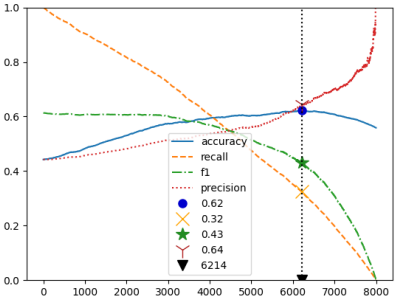
(c) METEOR



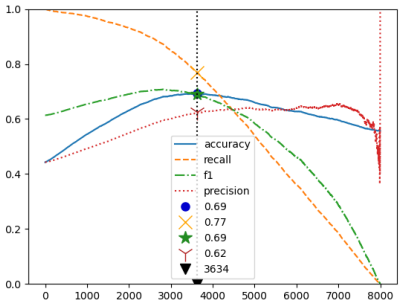
(d) ParaScore



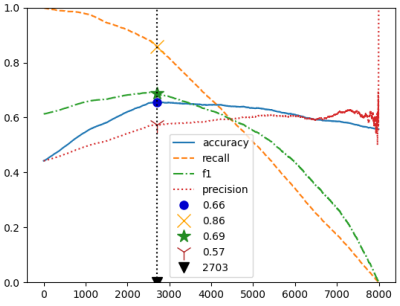
(e) BLEU



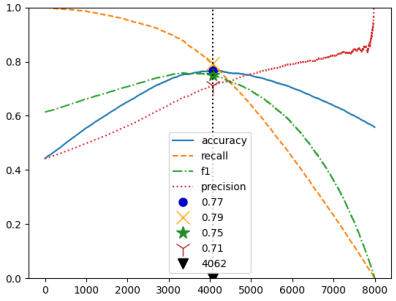
(f) sentence-transformers



(g) ParaPLUIE DIRECT

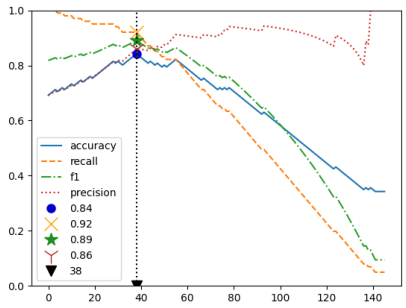


(h) ParaPLUIE INDIRECT

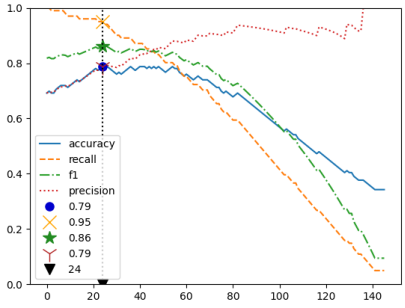


(i) ParaPLUIE FS-DIRECT

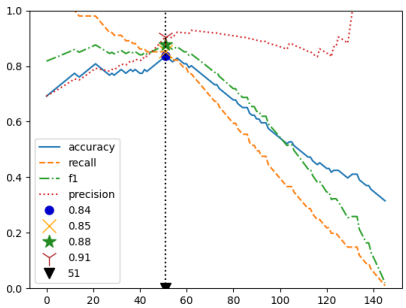
Figure 11: PAWS Corpus



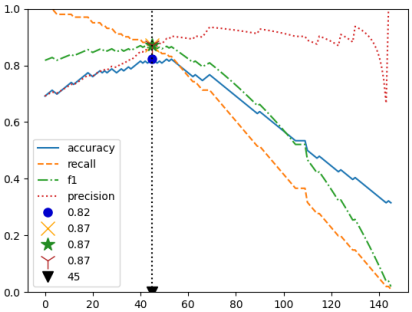
(a) BERT$_{score}$



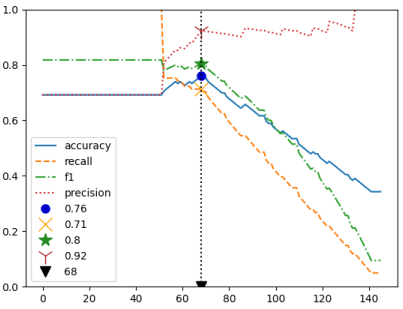
(b) LEV.



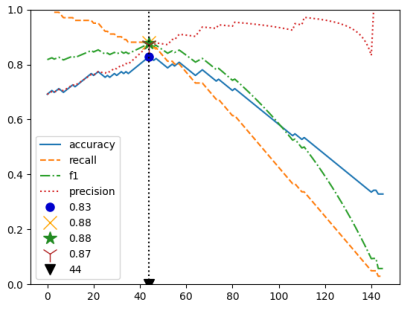
(c) METEOR



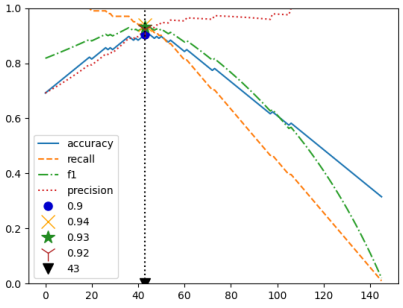
(d) ParaScore



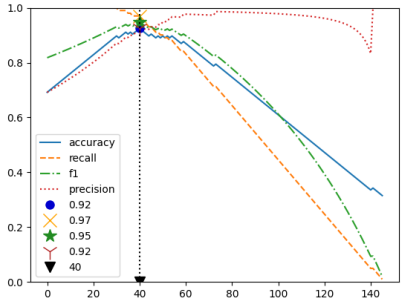
(e) BLEU



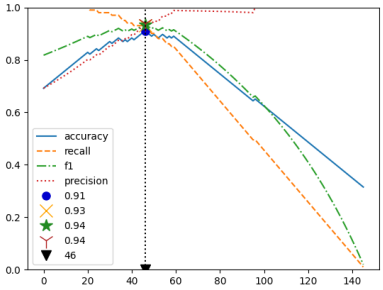
(f) sentence-transformers



(g) ParaPLUIE DIRECT

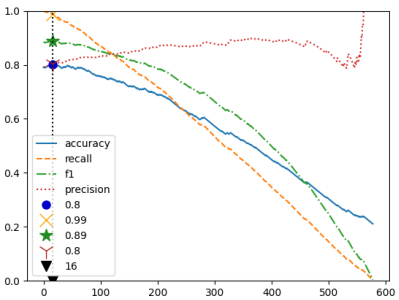


(h) ParaPLUIE INDIRECT

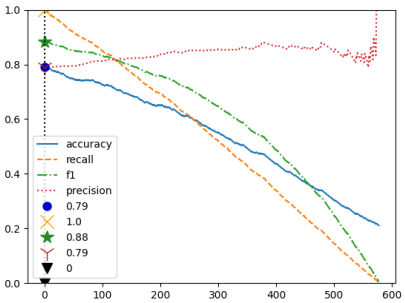


(i) ParaPLUIE FS-DIRECT

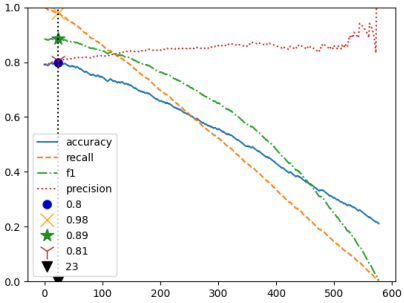
Figure 12: MCPG Corpus



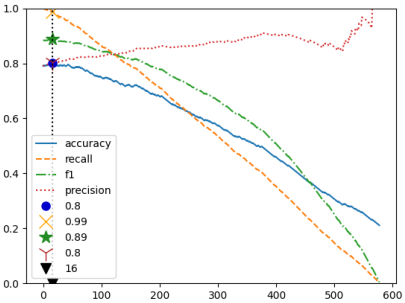
(a) BERT$_{\text{score}}$



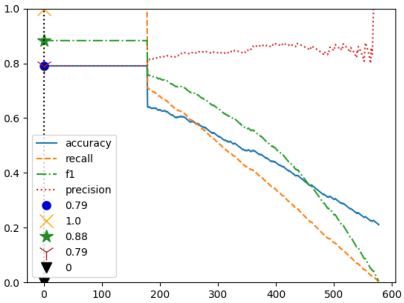
(b) LEV.



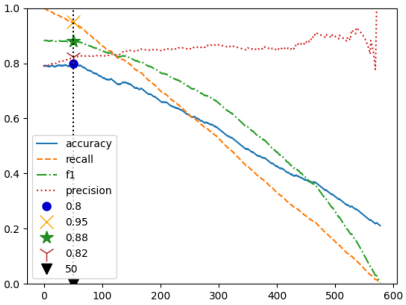
(c) METEOR



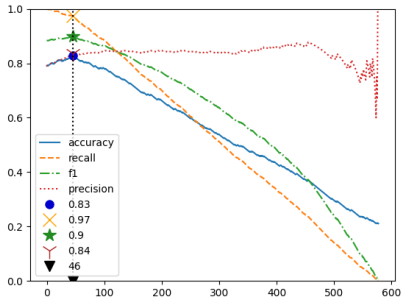
(d) ParaScore



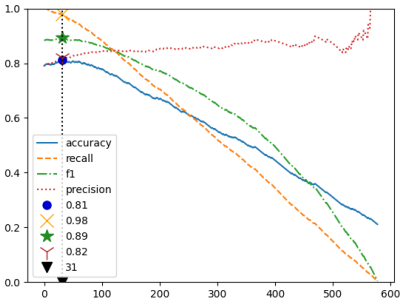
(e) BLEU



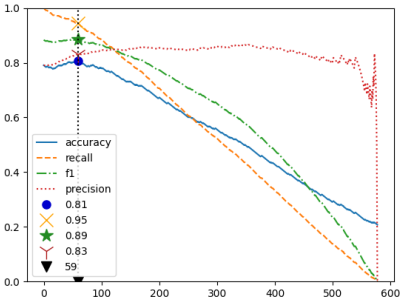
(f) sentence-transformers



(g) ParaPLUIE DIRECT

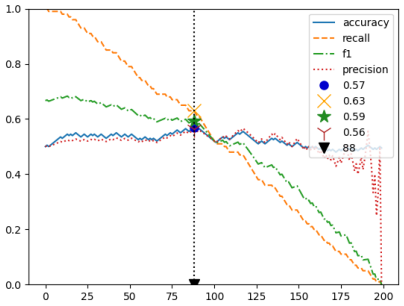


(h) ParaPLUIE INDIRECT

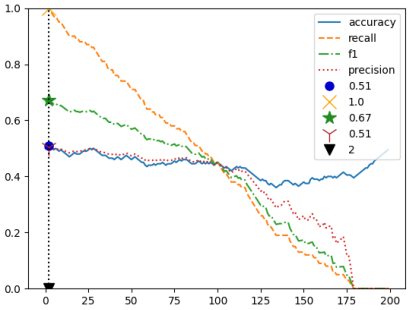


(i) ParaPLUIE FS-DIRECT

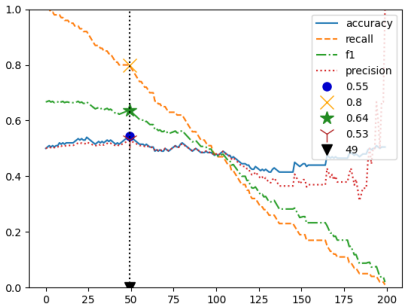
Figure 13: LLM Corpus



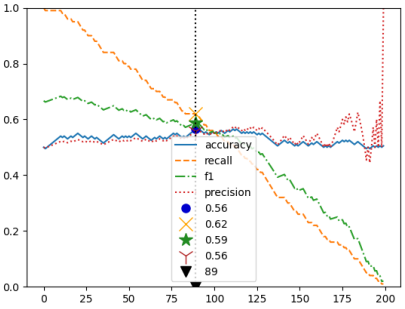
(a) BERT$_{\text{score}}$



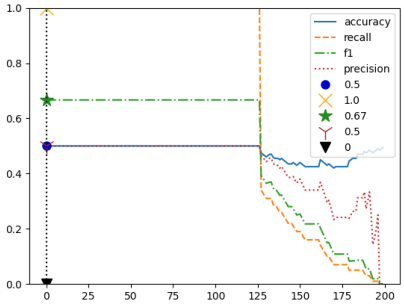
(b) LEV.



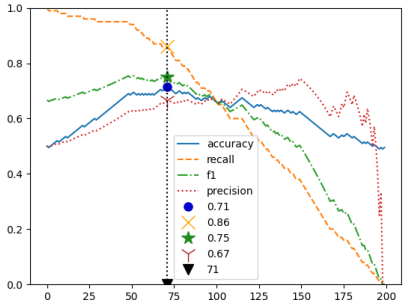
(c) METEOR



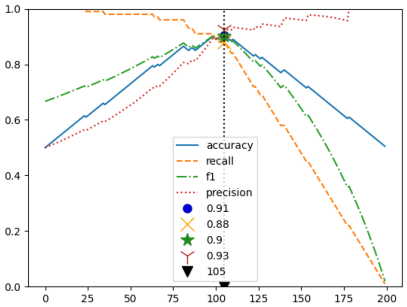
(d) ParaScore



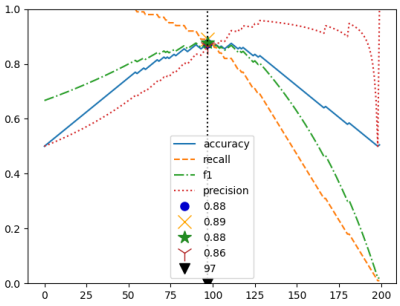
(e) BLEU



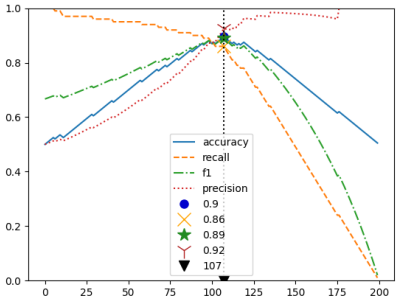
(f) sentence-transformers



(g) ParaPLUIE DIRECT

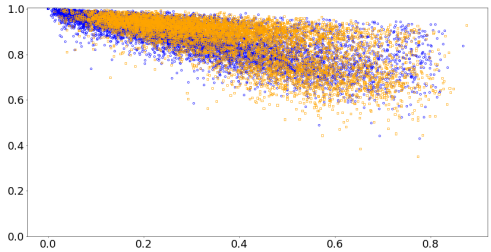


(h) ParaPLUIE INDIRECT

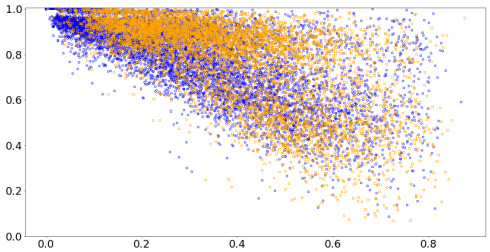


(i) ParaPLUIE FS-DIRECT

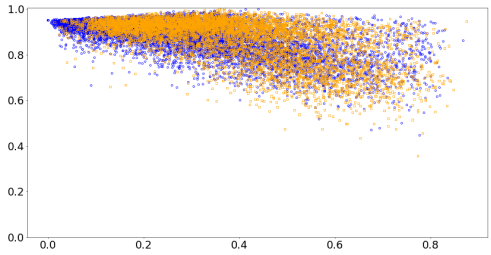
Figure 14: HC Corpus



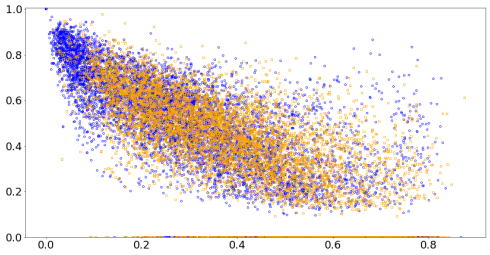
(a) BERT$_{\text{score}}$



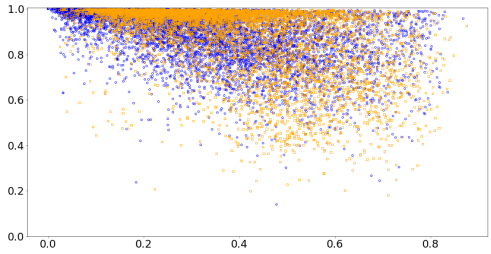
(b) METEOR



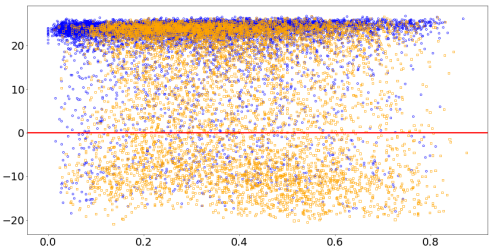
(c) ParaScore



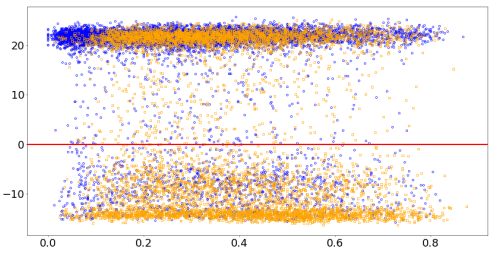
(d) BLEU



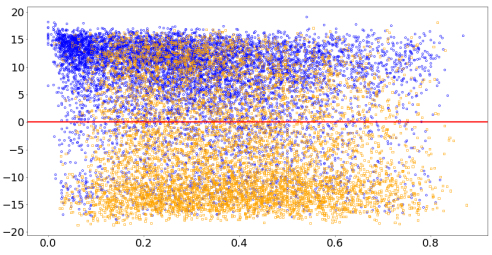
(e) sentence-transformers



(f) ParaPLUIE DIRECT

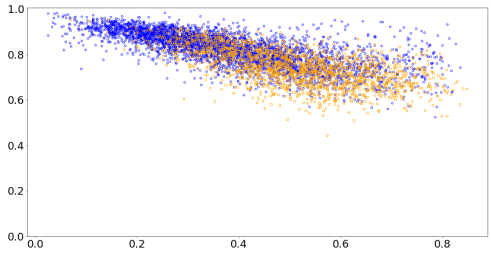


(g) ParaPLUIE INDIRECT

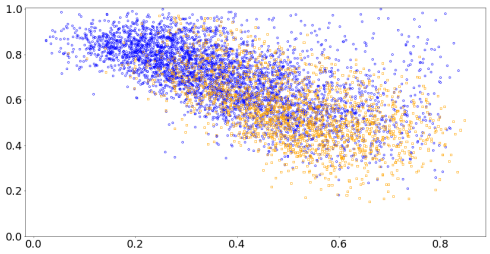


(h) ParaPLUIE FS-DIRECT

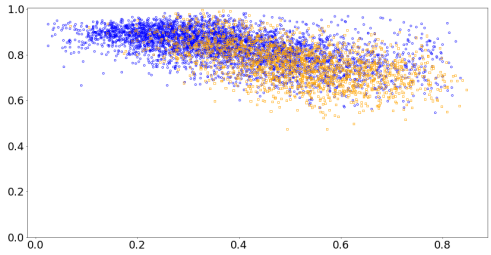
Figure 15: Global Corpus



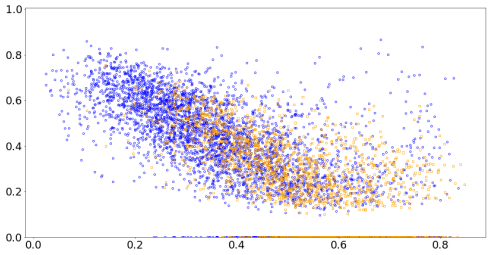
(a) BERT$_{score}$



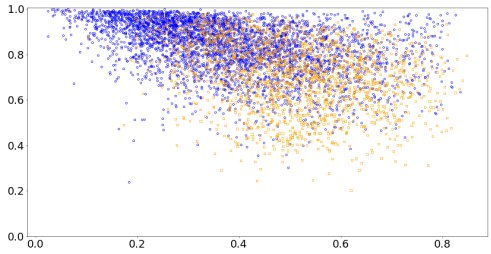
(b) METEOR



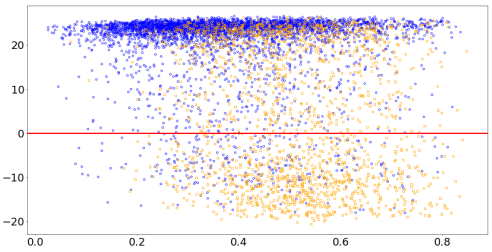
(c) ParaScore



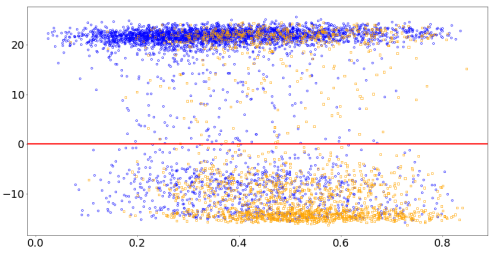
(d) BLEU



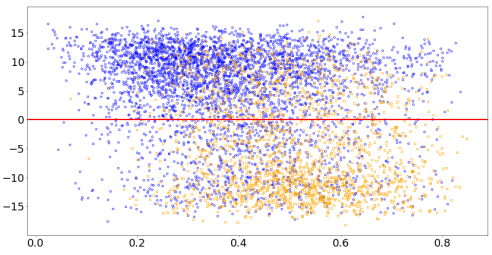
(e) sentence-transformers



(f) ParaPLUIE DIRECT

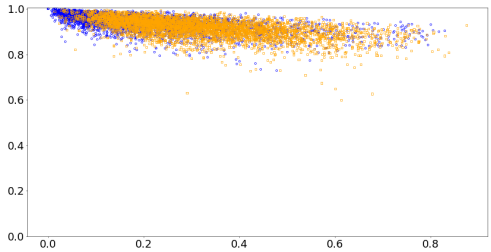


(g) ParaPLUIE INDIRECT

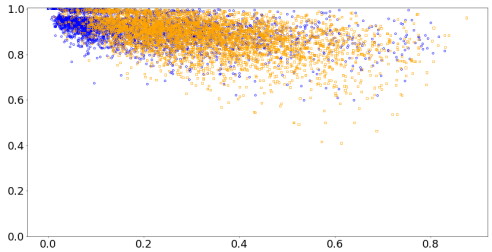


(h) ParaPLUIE FS-DIRECT

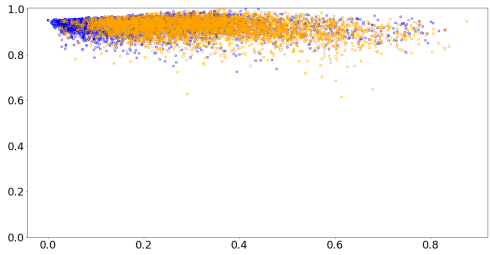
Figure 16: MRPC Corpus



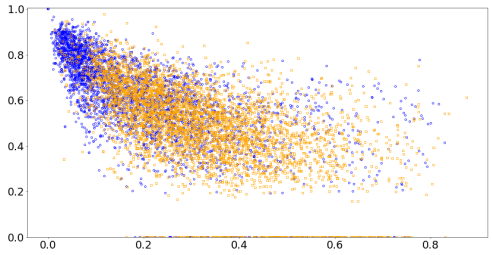
(a) BERT$_{score}$



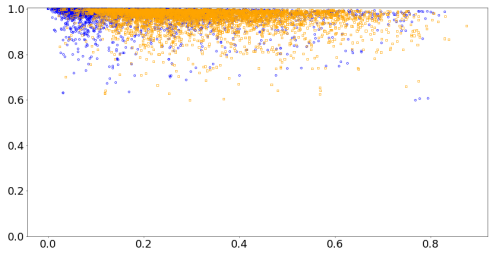
(b) METEOR



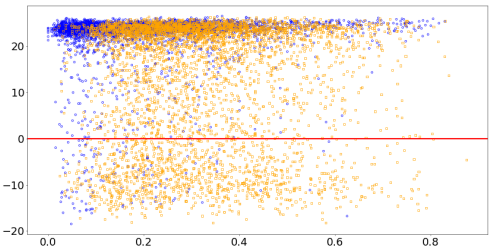
(c) ParaScore



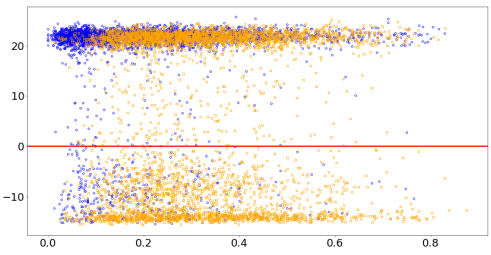
(d) BLEU



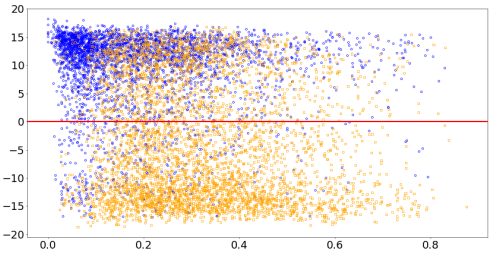
(e) sentence-transformers



(f) ParaPLUIE DIRECT

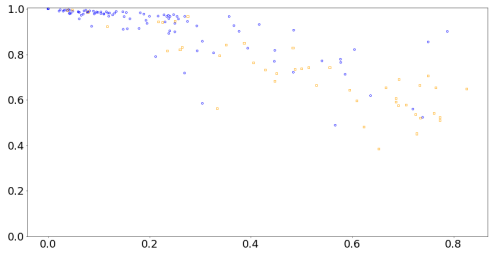


(g) ParaPLUIE INDIRECT

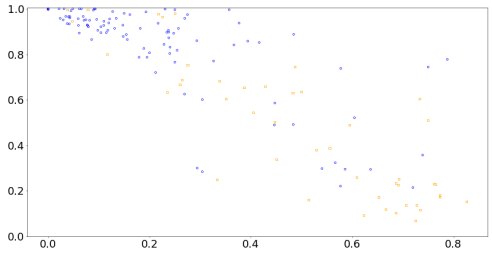


(h) ParaPLUIE FS-DIRECT

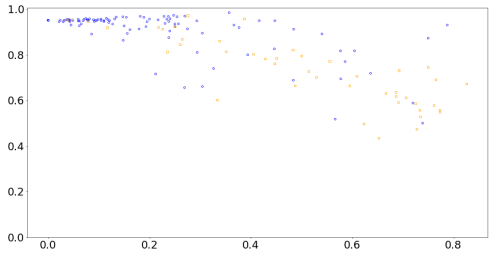
Figure 17: PAWS Corpus



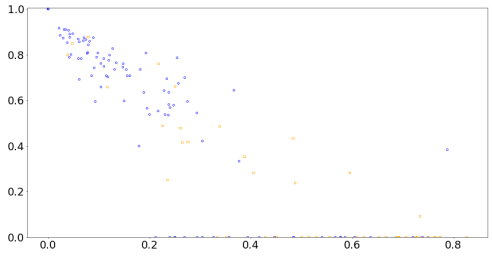
(a) BERT$_{\text{score}}$



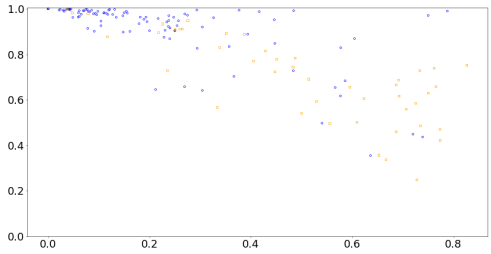
(b) METEOR



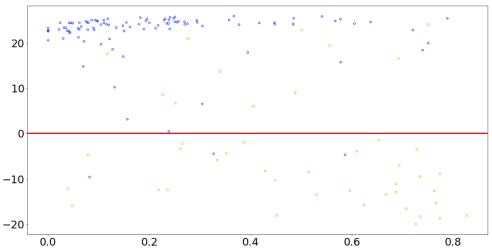
(c) ParaScore



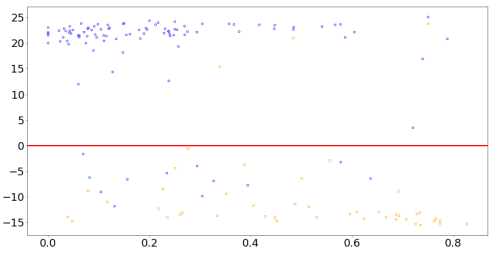
(d) BLEU



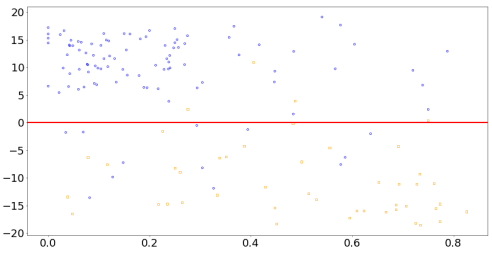
(e) sentence-transformers



(f) ParaPLUIE DIRECT

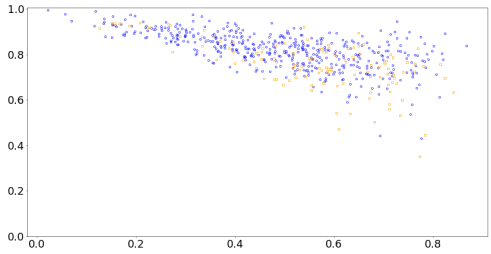


(g) ParaPLUIE INDIRECT

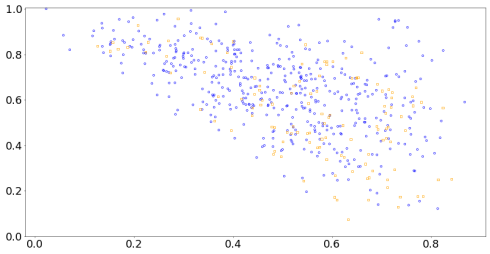


(h) ParaPLUIE FS-DIRECT

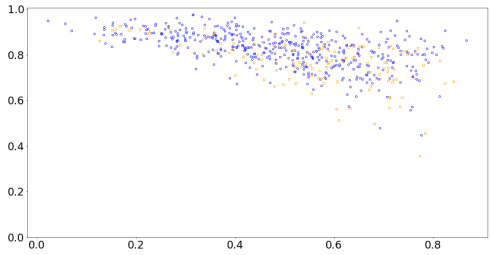
Figure 18: MCPG Corpus



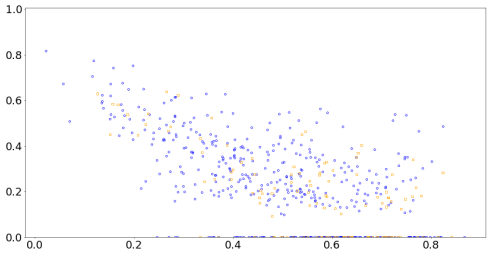
(a) BERT$_{\text{score}}$



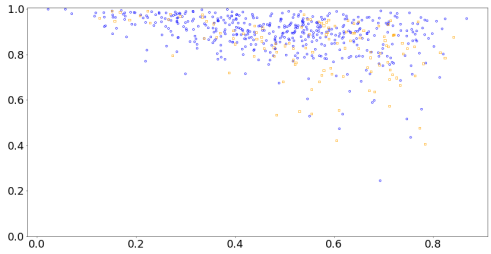
(b) METEOR



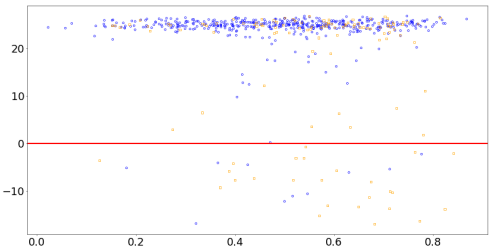
(c) ParaScore



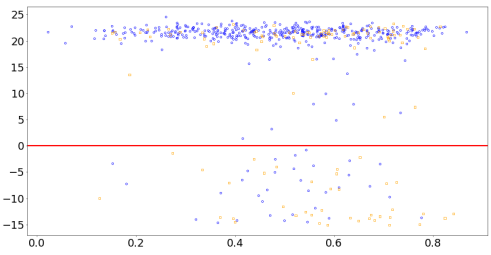
(d) BLEU



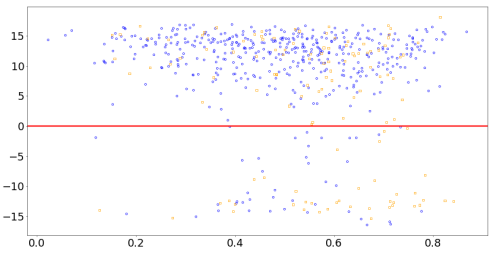
(e) sentence-transformers



(f) ParaPLUIE DIRECT

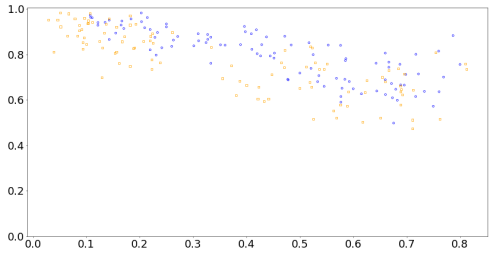


(g) ParaPLUIE INDIRECT

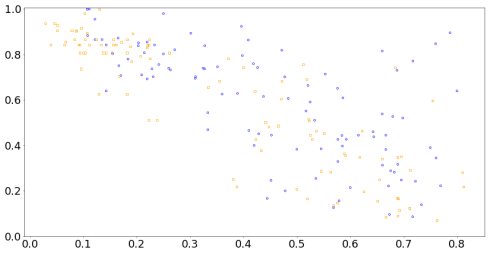


(h) ParaPLUIE FS-DIRECT

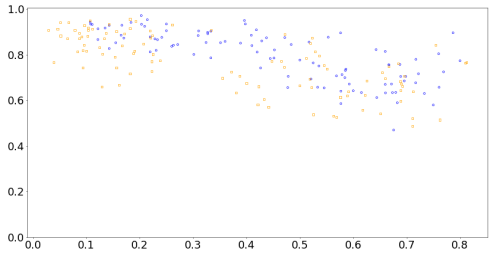
Figure 19: LLM Corpus



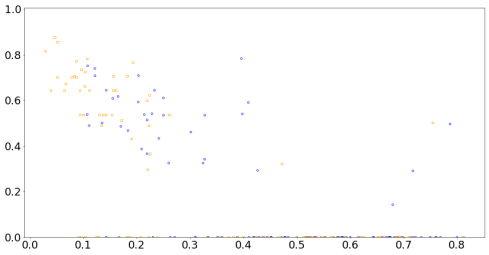
(a) BERT$_{\text{score}}$



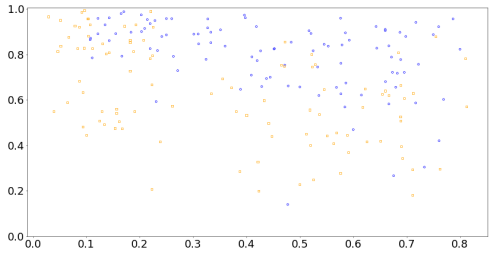
(b) METEOR



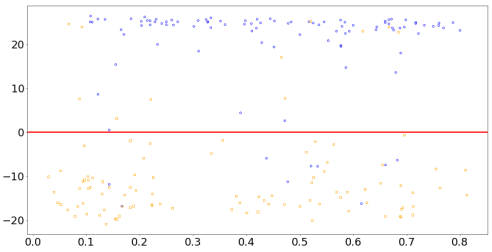
(c) ParaScore



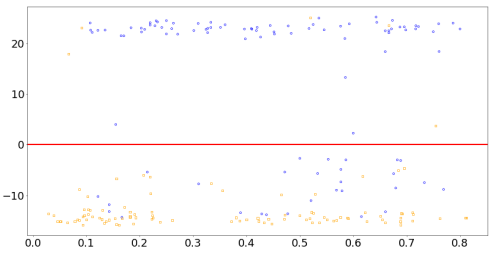
(d) BLEU



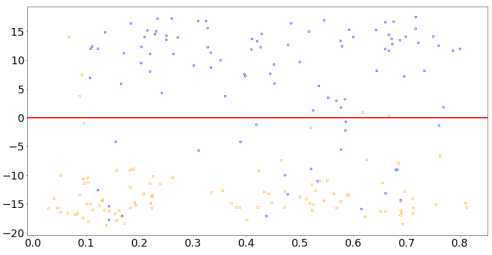
(e) sentence-transformers



(f) ParaPLUIE DIRECT



(g) ParaPLUIE INDIRECT



(h) ParaPLUIE FS-DIRECT

Figure 20: HC Corpus