PSentScore: Evaluating Sentiment Polarity in Dialogue Summarization

Yongxin Zhou, Fabien Ringeval, François Portet
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG
38000, Grenoble, France
{yongxin.zhou,fabien.ringeval, francois.portet}@univ-grenoble-alpes.fr

Abstract

Automatic dialogue summarization is a well-established task with the goal of distilling the most crucial information from human conversations into concise textual summaries. However, most existing research has predominantly focused on summarizing factual information, neglecting the affective content, which can hold valuable insights for analyzing, monitoring, or facilitating human interactions. In this paper, we introduce and assess a set of measures *PSentScore*, aimed at quantifying the preservation of affective content in dialogue summaries. Our findings indicate that state-of-the-art summarization models do not preserve well the affective content within their summaries. Moreover, we demonstrate that a careful selection of the training set for dialogue samples can lead to improved preservation of affective content in the generated summaries, albeit with a minor reduction in content-related metrics.

Keywords: Dialogue Summarization, Evaluation, Sentiment Polarity

1. Introduction

Automatic dialogue summarization has been widely studied and applied to various domains, including meeting (Carletta et al., 2005; Zhong et al., 2021), chat (Gliwa et al., 2019), email thread (Zhang et al., 2021), media interview (Zhu et al., 2021), customer service (Favre et al., 2015; Lin et al., 2021) and medical dialogue (Song et al., 2020). However, most research has focused on summarizing factual information, leaving aside affective content.

While it is essential to summarize the most pertinent factual information, the subjective content can also provide valuable insights. The integration of subjective content, such as affective aspects, into summaries could bring various benefits. These benefits encompass enhancing customer service, facilitating collaborative interactions, and offering improved support to healthcare patients. For instance, in the customer service sector, call center telephone conversations play a vital role in monitoring and enhancing service quality. Many calls contain emotional information that are deemed important to report (Roman et al., 2008).

Even though summarizing the affective part of dialogues could be highly valuable in many applications, it has been understudied, with only one study focusing on this topic (Roman et al., 2008). Affective content has been the target of a few summarization tasks such as opinion summarization (Wang and Ling, 2016). However, such tasks mainly focus on non-dialogue text reviews. In dialogues, factual information and subjective content are often intertwined. Therefore, when summarizing dialogues, it remains crucial to capture and synthesize not only the objective facts but also the subjective content.

One of the main limitations to incorporating sentiment into summarization lies in the human guidelines used to write the human references. Often, summarization tasks are crafted with a primary focus on facts and objectives, providing little guidance to human summarizers on how to handle affective content, as noted by previous research (Roman et al., 2008). A recent counter-example is the DialogSum dataset (Chen et al., 2021), where annotators were required to pay extra attention to several different aspects including *Emotions*. This shift in dataset design illustrates a growing recognition within the research community of the importance of incorporating affective content into summaries.

In the context of dialogue summarization, a pertinent question is how to reliably measure the ability of generative models to capture and convey affective information from the input dialogue in generated summaries. The current automatic evaluation methods for dialogue summarization mostly rely on n-gram comparisons and embedding distances between generated and reference summaries, while some studies proposed new metrics to evaluate faithfulness in dialogue summarization (Wang et al., 2022b). However, these metrics have been designed for factual correctness, and do not focus on evaluating the relevance of the affective content.

In this paper we make contributions to the field of dialogue summarization, which are outlined below:[1]

1. We emphasize the importance of affective content for dialogue summarization, especially in the context of customer service and health care.

[1] https://github.com/yongxin2020/PSentScore

LREC-COLING 2024, pages 13290–13302
20-25 May, 2024.

2. We propose *PSent*, a measure that calculates the proportion of affect charged words (positive/negative) in a given text.

3. We built several systems running sentiment analysis at the word-level in dialogues and evaluated how much affective content is preserved in summaries using *PSentScore* which is based on *PSent*, a reference-less measure.

4. We exploited the DialogSum Challenge framework to provide a reliable set of data and state-of-the-art methods for the automatic generation of summaries, and analyzed the affective content using *PSentScore*.

The results show that by filtering dialogues according to sentiment, we can significantly improve the preservation of both positive and negative sentiments in summarization, while preserving the performance of factual information.

2. Related Work

In the context of describing the human state of mind, several terms are usually employed, such as affect, feeling, emotion, sentiment, and opinion, which are sometimes used interchangeably despite existing differences between them (Munezero et al., 2014). In our study, even though the conversations are discourses, as we are performing a textual analysis, we consider affective content as sentiments expressed by the interlocutors and use the term *sentiment* throughout the paper.

2.1. Sentiment in Dialogue Summarization

Reporting the affective states of interlocutors in dialogue summarization is important in several cases. For instance, it can improve the customer experience by finding out whether the customer feedback is positive or negative (Zhou et al., 2022). In the context of health care, it is crucial to know how patients feel during human interactions such as clinical meetings, or human-machine interactions such as digital therapies (Tarpin-Bernard et al., 2021).

As we mentioned earlier, factual information and subjective content are often intertwined in dialogues, and while it is important to summarize the most relevant factual information, the subjective content can also provide key information. A study by Roman et al. (2008) revealed that whenever a dialogue contains an extreme emotion, this behavior is reported in human written dialogue summary. The study also shows that the emotional reporting varies considerably depending on the summarizer's viewpoint, and that size constraints have no impact on the emotional content reported in the summaries. In addition to this empirical evidence, there are some theoretical arguments in favor of the presence of emotions/sentiment in dialogue summarization. For instance, Tuggener et al. (2021) mapped dialogue types (categorization of dialogue types according to Walton and Krabbe, 1995) to summary items, and *Emotions* was explicitly mapped and emphasized as one of the summary items along with the following dialogue types: *Deliberation*, *Information seeking*, and *Eristics*.

Despite theoretical and empirical support for the inclusion of affective information in dialogue summaries, the inclusion of emotions/sentiment as a summary item is not a common practice when designing datasets. In a recent survey (Tuggener et al., 2021), of the datasets that were listed, only one dataset – Call Centre Conversation Summarization (CCCS) (Favre et al., 2015) – was found to have exploited *Emotions/Sentiment* as a summary item, while there is no mention of *Emotions/Sentiment* in the guideline. It seems thus that research on sentiment in dialogue summarization suffers a lack of resources. We argue that this lack of development is mainly due to two reasons: 1) the fact that current corpora did not consider affect in their guideline for writing reference summary; and 2) that there is, to the best of our knowledge, no automatic measure to assess the affective aspect of a summary with respect to its original source.

2.2. Dialogue Summarization Corpora and Sentiment

Despite dialogue summarization being a well-established task, the formulation of summarization tasks has not reached a consensus in the linguistic and the Natural Language Processing (NLP) communities, which has prevented from reaching a mutually agreed-upon definition of what a dialogue summary should look like (Guo et al., 2022). In order to evaluate to which extent corpora used for dialogue summarization considered affective information in their summary, we performed a short overview of the guideline of several major dialogue summarization datasets, which have been widely used in NLP research. This is summarized in Table 1. For each corpus, we can see that the summary criteria used are different. For some corpora, only the data has been made available, while annotation guidelines are rarely accessible.

As it can be seen, some corpora do not reveal the criteria used for reference summaries. However, most of the corpora disclose their objectives to write reference summaries as for the AMI meeting corpus (Carletta et al., 2005), SAMSum (Gliwa et al., 2019) (written online conversation), TWEETSUMM (Feigenblat et al., 2021), which is focused on customer service, QMSum (Zhong et al., 2021),

Name	Domain	Language	Guideline Available	Guideline criteria for writing the reference summaries
AMI (Carletta et al., 2005)	Meeting	English	Yes	Abstractive summaries should have the following structure: abstract, decisions, problems/issues, actions. Extractive summaries: identify extracts from the transcript which jointly convey the correct kind of information about the meeting to fit the required purpose. The instructions do not mention emotion/sentiment.
RATP-DECODA (Favre et al., 2015)	Telephone Customer Service	French	No	We contacted the authors and obtained their summary definition, there is no mention of emotion/sentiment.
SAMSum (Gliwa et al., 2019)	Chat	English	Yes	(1) Be rather short, (2) extract important pieces of information, (3) include names of interlocutors, (4) be written in the third person. The instructions do not mention emotion/sentiment.
MEDIASUM (Zhu et al., 2021)	Media Interview	English	No	The reference summaries were downloaded from text descriptions of the input documents (interviews) available on the web.
TWEETSUMM (Feigenblat et al., 2021)	Customer Service	English	Yes	Extractive summary: highlight the most salient sentences in the dialog. Abstractive summaries: one sentence summarizing what the customer conveyed and a second sentence summarizing what the agent responded. The instructions do not mention emotion/sentiment.
QMSum (Zhong et al., 2021)	Multi-domain Meeting	English	Yes	The annotation process consists of three stages: topic segmentation, query generation, and query-based summarization. The instructions do not mention emotion/sentiment.
CSDS (Lin et al., 2021)	Customer Service	Chinese	No	There are three different summaries for each dialogue: an overall summary and two role-oriented summaries (user and agent). Emotion/sentiment is not mentioned.
DIALOGSUM (Chen et al., 2021)	Spoken	English	Yes	Convey the most salient information; Be brief; Preserve important named entities within the conversation; Be written from an observer perspective; Be written in formal language. Pay extra attention to the following aspects: Tense Consistency, Discourse Relation, **Emotion** and Intent Identification.

Table 1: Major datasets for dialogue summarization with their summaries criteria. DialogSum is the only one to include *Emotion* in the guideline.

a query-based multi-domain meeting summarization dataset, and DialogSum (Chen et al., 2021), which is a real-life scenario dialogue summarization dataset. We can notice that amongst all those corpora, only in the DialogSum dataset, annotators were explicitly instructed to describe important emotions related to events in the reference summary.

The fact that the annotation is not explicitly tasked with processing sentiment does not prevent the reference summaries from containing it, to a certain extent. We checked this by manually analyzing the 212 annotated synopses of 100 dialogues taken from the RATP DECODA corpus test set (Favre et al., 2015). We found that, although annotators were not explicitly instructed to indicate customer satisfaction in the synopsis, some annotators did mention customer feelings, but this only occurred in a few cases, namely 4% of the synopses.

2.3. Evaluation of Affective Content in Dialogue Summarization

Most evaluations of summarization tasks still rely on n-gram base measure such as ROUGE (Lin, 2004). The F1 scores of ROUGE-1, ROUGE-2 and ROUGE-L are mainly reported, which measure word overlap, bi-gram overlap and longest common sequence between generated summaries and references.

Other measures such as BERTScore embeddings (Zhang* et al., 2020) have also emerged to provide a more subtle evaluation of similarities by

taking context and semantic proximity into account.

Recent works (Huang et al., 2020; Fabbri et al., 2021) have already pointed out that these metrics do not correlate equally with all kinds of human judgments. However, we are not aware of any automatic metric measuring sentiment adequacy with sources in dialogue summarization or other linguistic summarization task.

3. Measuring the Affective Content of Summaries

In this section, we present our method for measuring the affective content of summaries. Our hypotheses rely on the following:

- It is possible to measure automatically the presence of sentiment in texts with sufficient reliability, whether measured by dimensional polarities or categories, and at word or sentence level.
- The distribution of affective content in the summaries should be similar to that of the dialogue.
- Since affective and factual content might be interleaved, the measure might correlate with other context-based measures so it cannot replace them.

3.1. The PSent measure

In order to investigate whether the sentiment polarities of the input dialogues are preserved in the corresponding summaries, we assess the variability of sentiments in both the input dialogue and their corresponding summaries. Since our hypothesis is that the affective content of the input should be preserved in the output summary to a comparable extent, we propose to compute a ratio for the input and output affective content.

There are different resolutions in which sentiment can be calculated (document, paragraph, sentences, words). As a summary can be very short we assume that the word-level is the most adequate.

After labeling positive and negative words in each sentence using word-level Sentiment Analysis (SA) models, the number of positive and negative words and the total number of words in the input dialogue and corresponding summaries can be counted. We then calculate the proportion of affect charged words in the whole dialogue and in the summary, respectively. The formula used for this calculation is as follows:

$$PSent = (PosN + NegN)/TotalN \quad (1)$$

In eq. 1, *PSent* represents the proportion of sentimentally charged words in the given texts. We use *PSentDial*, and *PSentSumm* to represent the *PSent* in the input dialogue, and reference summaries (or generated summaries), respectively. Furthermore, we can also compute $PSent_P$ (resp. $PSent_N$) to denote the proportion of affect charged positive words – $PosN$ – only (resp. negative words – $NegN$ – only) in the given texts.

3.2. The PSentScore measure

Ideally, the summaries should mirror the sentiment proportion or polarity of the input texts. Therefore, the evaluation of a summarization system can be performed by quantifying the disparity between the sentiment proportion or polarity in the generated summary and that in the input text. For multiple documents, a unified measure can be derived.

To examine whether the sentiment polarities presented in the input dialogues and in the output summaries are equivalent, we first calculate *PSentDial* and *PSentSumm* for each dialogue-summary pair.

Then to explore the relationship strength between *PSentDial* and *PSentSumm* in various splits of dialogue summarization datasets, we compute and present *PSentScore* using the following measures: 1) Spearman's rank correlation coefficient (r_s) – eq. 2, which assesses the monotonic relationships between two variables (Zar, 2005); 2) Concordance Correlation Coefficient (CCC) – eq. 3, which quantifies the similarity between two sets of data, i.e. the trends between two variables; 3) Mean Absolute Error (MAE) – eq. 4, calculates errors between two sets of values, it is also known as scale-dependent accuracy as it calculates error in observations taken on the same scale.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2)$$

Where: r_s represents the Spearman's rank correlation coefficient. d_i represents the differences between the ranks of corresponding data points in the two variables being compared. n is the number of data points.

$$CCC(x, y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (3)$$

Where: ρ represents the Pearson correlation coefficient between x and y values. σ_x and σ_y represent the standard deviation of the x and y values. μ_x and μ_y represent the mean of the x and y values.

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (4)$$

Where: n is the number of data points. y_i and $\hat{y}_i$ represent the values of the two variables for the i-th data point, respectively.

3.3. Experimental design

We intend to show the effect of the measure empirically. Hence the first step of the method is to select a corpus with reference texts containing some affective items. Out of all the corpora mentioned in Section 2.2, we selected DialogSum (Chen et al., 2021), which is composed of social conversations that are often affect charged. As a reminder, annotators were explicitly instructed to describe important emotions related to events in the summary. In addition, DialogSum was used in a challenge, in which several teams participated and presented their results (Chen et al., 2022).

We then computed $PosN$ and $NegN$ at the word-by-word level, using state-of-the-art models based on BERT (Devlin et al., 2019). These models were evaluated on a separate corpus and used to evaluate to which extent DialogSum contains sentiment in its documents.

Finally, using *PSentScore* and standard measures, we evaluated to which extent the state-of-the-art models handle sentiment with DialogSum. We then proposed a method to select the training target by eliminating documents without affective content in the input dialogue and/or summary, and trained models on this filtered data to evaluate whether sentiment handling can be improved.

4. Measuring Affective Content of Reference Summaries

Initially, we adopted the *opinion_lexicon* (Hu and Liu, 2004) dictionary as the simplest approach for Sentiment Analysis (SA). This dictionary consists of two lists of positive and negative words; any word that is not positive or negative is thus labeled as neutral. However, this dictionary-based approach has some limitations, as the polarity of some words may vary depending on their context (e.g., the word "kind"), and such differences cannot be distinguished by this dictionary-based approach. To overcome this limitation, we then explored contextual SA at the word level and considered training a SA model for this purpose.

4.1. Training Word-level Sentiment Analysis Models

4.1.1. Corpus: Stanford Sentiment Treebank (SST)

The SST dataset (Socher et al., 2013) is the first corpus that provides fully labeled parse trees, enabling a complete analysis of the compositional effects of sentiment in language. This dataset has been extensively studied for binary single sentence sentiment classification (positive/negative) and fine-grained sentiment classification (five classes). Given the complete parse tree annotations, it presents an opportunity to adapt it for word level SA. To the best of our knowledge, this is the only dataset available for word level sentiment classification.

The SST dataset includes fine-grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences. The partition statistics are presented in Table 3. In the following, we only focus on studying word level polarity. Hence, we preprocessed the original SST that was annotated with a 5-point Likert scale ("very negative", "negative", "neutral", "positive" and "very positive") into a 3-point Likert scale by simply merging "very negative" into "negative" and "very positive" into "positive"; this will be referred to as **SST3** in this paper.

4.1.2. Word-level SA Models

We built three specific models to perform word-level SA. To train BERT-based word-level SA models, we adapted the training code provided by Hugging Face for token classification tasks.[2]

token-dict. The dictionary-based classifier based on *opinion_lexicon* (Hu and Liu, 2004), it relies on a list of positive and negative opinion or sentiment words in English (about 6,800 words).

BERT-SST3 For the second model, we fine-tuned BERT (Devlin et al., 2019) using the preprocessed SST3 dataset with word-level annotation, where each word receives a label. We used a learning rate of $5\text{e}{-05}$ for 3 epochs. The model with the lowest validation loss was selected for reporting results on the test set and for further use.

BERT-DS-SST3 As the SST dataset is composed of movie reviews and is not specific to conversational setting, we used domain adaptation to familiarize our model with dialogue-specific characteristics. To do so, we automatically annotated the DialogSum training partition using the `token-dict.` (dictionary-based classifier, each word gets a label) because it is independent of the domain. We then fine-tuned BERT on this annotated corpus with a learning rate of $2\text{e}{-05}$ for 5 epochs, selecting the model with the lowest validation loss.

Next, we proceeded to further fine-tune the selected model on the training set of SST3, which was annotated by human annotators. We used a learning rate of $5\text{e}{-05}$ for 3 epochs. The aim was to obtain a model adapted to the DialogSum dataset but trained with reliable annotation from

[2] https://github.com/huggingface/transformers/tree/main/examples/pytorch/token-classification

	overall_accuracy	precision	recall	f1
token-dict.	88.82	73.61	60.96	65.64
BERT-SST3	97.87 (±0.06)	94.43 (±0.43)	94.07 (±0.38)	94.24 (±0.15)
BERT-DS-SST3	97.96 (±0.04)	94.53 (±0.17)	94.39 (±0.20)	94.46 (±0.10)

Table 2: Performances in terms of accuracy, precision, recall, f1 (%) on the test set of the SST-3 dataset, for different models: `token-dict.`, `BERT-SST3` and `BERT-DS-SST3`. Statistics are given in the following format: mean (standard deviation), based on three runs. Macro results for precision, recall and f1.

Split	# samples
Training	8544
Validation	1101
Test	2210

Table 3: Basic statistics for SST dataset.

SST dataset. We selected the model with the lowest validation loss for evaluation on the test set and for future use. Both the training and prediction were performed on a NVIDIA Quadro RTX 6000 GPU.

4.1.3. Word-level Sentiment Analysis Models' Performance

Table 2 shows the results of the models on the SST3 test set. We evaluate token classification performance using common metrics such as token-level precision, recall, and F1 score. Due to an imbalance in the number of neutral token labels in preprocessed SST3, we report "macro" results for these metrics when evaluating hypotheses.

The lexicon-based dictionary (`token-dict.`) shows poor performance on the SST token classification task, while `BERT-DS-SST3` performs similarly to `BERT-SST3`. It seems that domain adaptation has not decreased the performance on SST3 and, on the contrary, has stabilized its performance with a lower standard deviation than that of `BERT-SST3`.

4.2. Affective Representation of DialogSum

In what follows, we focus on comparing and evaluating the affective representation in input dialogues and reference summaries from the DialogSum dataset. We employ `BERT-DS-SST3` to calculate the *PSent* of each dialogue and its corresponding reference summaries.

In Figure 1, we present box plots of the distributions of *PSentDial* versus *PSentSumm* for the DialogSum training and validation sets. The figure includes two versions of the distributions: *Full*, which comprises all samples, and *Filtered*, which removes samples with *PSentDial* and/or *PSentSumm* values equal to zero (cf. Table 4 for the statistics). We use the *Filtered* version to avoid the potential impact of zero values on reported results.

Considering all samples, the median of *PSentSumm* is lower than that of *PSentDial*, indicating that there may be an under-representation of affective states in the reference summaries of the train and dev partitions, even though emotions are indicated in the corpus annotation guidelines. For the *Filtered* distribution on the training and validation sets, the median of *PSentSumm* is similar to that of *PSentDial*. However, for both versions, the distributions outside of the quartiles of the box plots are more varied for *PSentSumm* than for *PSentDial*.

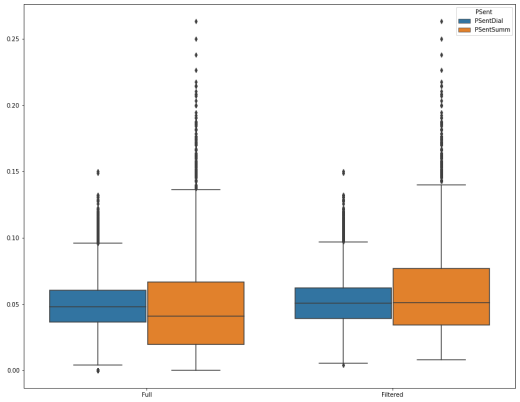

Figure 1: Box plots for *PSentDial* (left) vs. *PSentSumm* (right) distribution using `BERT-DS-SST3` on the full DialogSum training and validation sets. *Filtered* means that samples with *PSentDial* or *PSentSumm* values equal to zero have been removed.

5. Assessing Sentiment Handling of Summarization Models

5.1. Filtering Methodology

In order to investigate the sentiment handling by the state-of-the-art models and to carefully select the training target by eliminating the pairs without affective content in the input dialogue or the summary, we filtered the DialogSum dataset using the word-level SA method mentioned earlier: `BERT-DS-SST3`.

Detailed statistics for the DialogSum dataset are provided in Table 4, and *Full set* represents its raw statistics: a total of 13,460 dialogues are divided

Part.	Full	**Filtered** by **BERT-DS-SST3**	**w/ zero value** Dial.	Sum.
train	12460	9687 (77.5%)	43	2757
dev	500	391 (78.2%)	2	108
test	500	499 (99.8%)	1	0

Table 4: Statistics for DialogSum dataset. For the Filtered corpus both input dialogue and reference summaries without affective content according to BERT-DS-SST3 were removed. However, for the test partition only the input dialogues were filtered out. Percentage of data kept is shown in parentheses (%).

into training (12,460), validation (500) and test (500) sets (Chen et al., 2021). The *Filtered* set contains 9687 training and 391 validation samples.

5.2. Experimental Setup

Following the state-of-the-art models (Chen et al., 2021), we fine-tuned the BART-Large model (Lewis et al., 2020) on the full set of DialogSum and on the filtered dataset.[3] We also trained the model on a corpus of the same size as the filtered dataset, but whose instances were randomly sampled from the full dataset. The hyperparameters setting was learning rate of $5e{-}05$ for 15 epochs. Experiments were performed on the NVIDIA Quadro RTX 6000 GPU and took about 2.5 hours for each run.

5.3. Evaluation Metrics

In addition to ROUGE (Lin, 2004)[4] and BERTScore (Zhang* et al., 2020)[5], we propose a new set of measures to assess the relevance of a summary with respect to the affective charge (proportion - PSent), and its polarity (PSent_P / PSent_N). We use BERT-DS-SST3 as the backbone to calculate them. The evaluation methods from sentiment perspectives are as follows:

PSentScore is our proposed *PSent* evaluation method from a proportional perspective. We compute the relationship strength between *PSentDial* and *PSentSumm* and provide Spearman/CCC/MAE scores. These scores indicate the monotonic relationships, trends, and errors between two sets of values, respectively (as mentioned earlier in 3.2).

PSentScore$_P$ and **PSentScore**$_N$ are proposed from the polarity perspective. We examine whether the positive (resp. negative) affective aspects presented in the input dialogues are also present in the output summaries.

For the above measures from affective perspectives, samples with zero *PSentDial* (or *PSent*$_P$*Dial* / *PSent*$_N$*Dial*) values are removed, to account for the potential impact of zero values on the reported correlation.

5.4. Quantitative Results

The results of the fine-tuned BART-Large model on different versions of the DialogSum dataset are presented in Table 5. We compare our results with previous results reported in DialogSum dataset paper (Chen et al., 2021), and with results from different teams in the challenge (Chen et al., 2022; Lundberg et al., 2022; Bhattacharjee et al., 2022; Chauhan et al., 2022). The predictions of the various systems were obtained from the corresponding authors, based on which we evaluated the *PSentScore* results. *Human* results are those computed by (Chen et al., 2022) obtained by averaging each human annotator scores against others.

As reported in Table 5, the GoodBai model (Chen et al., 2022) provides the highest ROUGE and BERT scores very close to the other teams and slightly better than the BART_{Large} model (2^{nd} line) provided as reference to the challenge. Our **baseline-BART**$_{Large}$ model shows similar performances as the BART_{Large} model (2^{nd} line). When trained on the Filtered dataset, the **baseline_Filtered** model exhibits a decrease of almost 1.5 points on all the ROUGE and BERT scores. However, when looking at the *PSentScore* measures, the **baseline_Filtered** model provides the best correlation of affective content between dialogues and summaries (.435/.348/.027), far from the state-of-the-art models (.364/.297 in Spearman and CCC at most, while lowest MAE value – .027 is similar).

Looking at *PSentScore*$_P$ and *PSentScore*$_N$, the **baseline_Filtered** model also provides the best correlation of affective content in terms of polarity between dialogues and summaries (.370/.352/.023 and .449/.373/.015), while TCS_WITM (Chauhan et al., 2022) reached .375 on *PSentScore*$_P$ in Spearman and GoodBai (Chen et al., 2022) reached .014 on *PSentScore*$_N$ in MAE, which are better than **baseline_Filtered**.

The **baseline_sub-sampled** showed a decrease in ROUGE and BERTScore compared to the **baseline-BART**$_{Large}$ with a reduced number of training samples, while the *PSentScore* measures were similar. The results show that by filter-

[3]We adapted the training code from Wang et al. (2022a) to reproduce the results of the baseline model (BART-Large) on the DialogSum dataset.

[4]We used the Hugging Face script of the ROUGE metric which uses the Google Research implementation https://github.com/huggingface/datasets/blob/main/metrics/rouge/rouge.py, which is also the one used in Chen et al. (2021).

[5]Following Chen et al. (2022), we use RoBERTa (Liu et al., 2019) large as the backbone to compute BERTScore and the precision scores are reported.

Model # samples	**R1** 500*	**R2** 500*	**RL** 500*	**BERTScore** 500*	**PSentScore** 499†	**PSentScore**$_P$ 491†	**PSentScore**$_N$ 419†
Human (Chen et al., 2022)	53.35	26.72	50.84	92.63	-	-	-
BART$_{Large}$ (Chen et al., 2021)	47.28	21.18	44.83	-	-	-	-
GoodBai (Chen et al., 2022)	**47.61**	**21.66**	45.48	**92.72**	.357/.289/.027	.341/.307/.024	.397/.358/**.014**
UoT (Lundberg et al., 2022)	47.29	21.65	**45.92**	92.26	.356/.297/.027	.364/.325/.023	.383/.338/.014
IITP-CUNI (Bhattacharjee et al., 2022)	47.26	21.18	45.17	92.70	.348/.289/.031	.311/.280/.027	.397/.295/.018
TCS_WITM (Chauhan et al., 2022)	47.02	21.20	44.90	90.13	.364/.294/.028	**.375**/.331/.024	.431/.333/.014
baseline-BART$_{Large}$	47.36	21.23	44.88	91.42	.353/.292/.029	.318/.295/.025	.395/.322/.016
baseline_sub-sampled$\diamond$	46.94	20.52	44.43	91.29	.351/.294/.028	.351/.319/.024	.410/.352/.016
baseline_Filtered$\diamond$	45.78	19.69	43.21	90.83	**.435**/**.348**/**.027**	.370/**.352**/**.023**	**.449**/**.373**/.015

Table 5: Comparison of results from the DialogSum challenge teams and our BART-Large models fine-tuned on the full (`baseline-BART`$_{Large}$) and filtered corpus (`baseline_Filtered`) of the DialogSum dataset. The `baseline_sub-sampled` model has been trained on a corpus of the same size as the Filtered dataset but whose instances have been randomly sampled from the the full DialogSum dataset. $\diamond$ indicates training on partial corpora: training set 9687 (77.5%), dev set 391 (78.2%). *PSentScore* values indicate evaluation results: Spearman (↑) / CCC (↑) / MAE (↓). * refers to the Full DialogSum test set, and † to the Filtered test set.

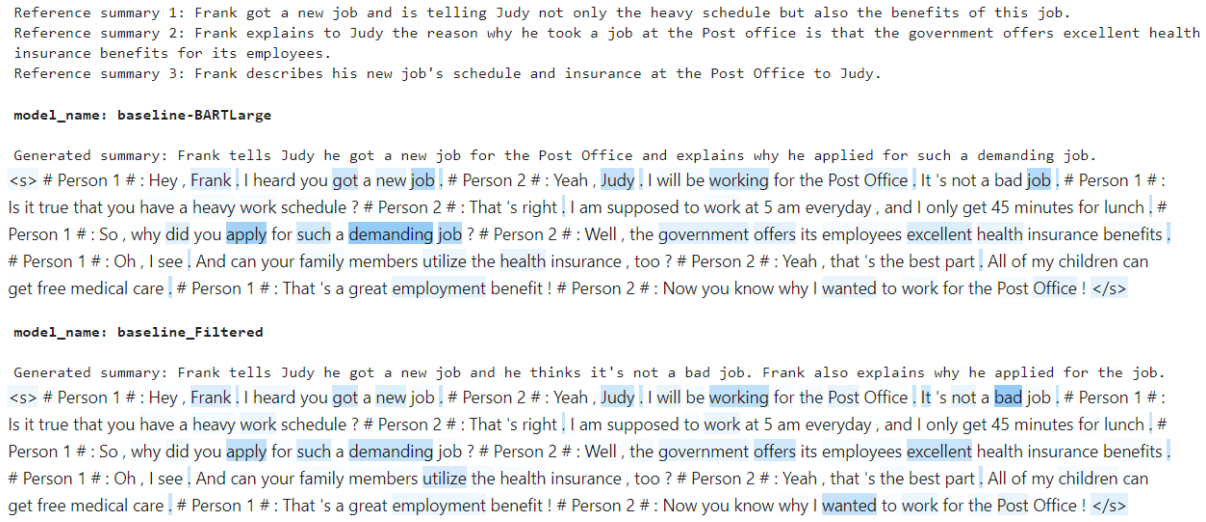

Figure 2: Example of test_20, with three references, and predictions as well as visualization of the attention of two models: `baseline-BART`$_{Large}$ and `baseline_Filtered`.

ing dialogue samples according to affective content (**`baseline_Filtered`**), we can significantly increase the preservation of both positive and negative sentiment in summaries, while preserving factual information. It's worth noting that certain decreases in ROUGE and BERTScore may be due to the smaller amount of training data.

5.5. Example Analysis

To determine whether attention focuses more on affective words for our models trained with the filtered dataset, we visualized the distribution of attention weights for several examples. Examples are chosen with maximum *PSentDial* using the word-level sentiment analysis models: `BERT-DS-SST3` and `token-dict.`.

We then visualized the distribution of their attention in the input dialogue for the following two models: `baseline-BART`$_{Large}$ and `baseline_Filtered`. In detail, we have calculated the encoder-decoder cross-attention, calculated the attention weights of the last layer (12^{th}) and the last head (16^{th}), the calculated attention weights are the dialogue versus the summary.

In what follows, we present one example with the predictions and visualization of attention from the two models mentioned previously. In Figure 2, the predicted summary of `baseline_Filtered` includes "he thinks it's not a bad job", in addition to the factual information "Frank tells Judy he got a new job" presented in both predictions, we can see that the word "bad" in the dialogue is particularly emphasized, and the word "excellent" receives more attention in the last model.

6. Discussion and Conclusion

In dialogue summarization, the most important content almost always focuses on factual information,

leaving aside the affective content of the interaction. We argue that affective information is important content to report in dialogue summaries. In order to measure affective content in dialogues and in summaries, we trained SA models at the word level. We conducted a corpus-based analysis on the DialogSum corpus, in which dataset annotators were explicitly instructed to include affective content when writing reference summaries, we show that affective content is omitted to some extent in reference summaries in dialogue datasets.

We then propose a new set of measures to evaluate the relevance of a summary based on the affective load (proportion) and its polarity (positive/negative). Using this measure, we show that the summarization model often exhibits a mismatch between the affective content of the input dialogue and the summary. We also show that by carefully selecting the training target, we can decrease this mismatch. This method provides a more comprehensive measure of dialogue summarization performance.

In this study, we chose the DialogSum corpus for analysis because it explicitly considers emotions in its annotation guidelines. In the future, we will extend our method and conduct large-scale analyses on various dialogue summarization datasets and with more fine-grained affective categories. We also plan to extend the use of *PSentScore* to other NLP tasks, such as summarizing reviews/opinions, generating emotional dialogues, etc. We can also consider maximizing the *PSentScore* for dialogue summarization as a functional goal, but this first requires an appropriate affect-oriented dataset.

Limitations

Our measure is still gross and focuses on proportion and polarity perspectives, tested only on one data set. It does not distinguish, for example, whether we are reporting anger or sadness with the same distribution as in the dialogue. For this reason, we will look at other measures that might account for this. Furthermore, the method currently only works for English.

We should also emphasize that the *PSent* measure depends on a word-level sentiment analysis model which might not be available or biased if trained on a dataset different from the one *PSent* is applied to. While our experiments focused on increasing the similarity of proportion of sentiment in the input and output texts, we did not perform a human evaluation of the outputs that might have provided more fine grained analyses. The standard automatic measures suggest that the summaries generated by the different models are somewhat similar but we recognize that the model learned on the Filtered corpus may generate degraded outputs due to less training data. Furthermore, the focus of the measure on the polarity is a crude evaluation of affective content that cannot account for subtle difference between the input text and the generated summaries as most of the other automatic measures.

While the `BERT-DS-SST3` model demonstrated promising performance on the SST corpus, it has not been evaluated on dialogue corpora. Annotating sentiment at the word level poses challenges, as annotators often lack consensus in their annotations. Our subsequent investigation will focus on sentiment analysis at the expression level. However, since *PSent* measures the conservation and proportion of sentimental words, it remains a suitable metric for our purposes.

Ethics Statement

The DialogSum corpus we used in this study is composed of resources freely available online without copyright constraint for academic use. According to the authors, the annotators had degrees in English Linguistics or Applied Linguistics. They received a salary of around 9.5 dollars per hour and took this annotation as a part-time job. We chose this corpus because it is the only dialogue summarization corpus we found that mentions emotions in its annotation guidelines, but we also acknowledge that the corpus consisting of social conversations may differ from other kinds of conversations such as in the medical domain or customer service that may have specific characteristics.

7. Acknowledgements

This research was supported by the Banque Publique d'Investissement (BPI) under grant agreement THERADIA and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). We would also like to thank the anonymous reviewers for their insightful comments.

8. Bibliographical References

Saprativa Bhattacharjee, Kartik Shinde, Tirthankar Ghosal, and Asif Ekbal. 2022. A multi-task learning approach for summarization of dialogues. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 110–120, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel

Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, page 28–39, Berlin, Heidelberg. Springer-Verlag.

Vipul Chauhan, Prasenjeet Roy, Lipika Dey, and Tushar Goel. 2022. TCS_WITM_2022 @ DialogSum : Topic oriented summarization using transformer based encoder decoder model. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 104–109, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Yulong Chen, Naihao Deng, Yang Liu, and Yue Zhang. 2022. DialogSum challenge: Results of the dialogue summarization shared task. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 94–103, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

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

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

Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. 2022. Questioning the validity of summarization datasets and improving their factual consistency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5716–5727, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Conrad Lundberg, Leyre Sánchez Viñuela, and Siena Biales. 2022. Dialogue summarization using BART. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 121–125, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111.

Norton Roman, Paul Piwek, and Ariadne Carvalho. 2008. Emotion and behaviour in automatic dialogue summarisation.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Franck Tarpin-Bernard, Joan Fruitet, Jean-Philippe Vigne, Patrick Constant, Hanna Chainay, Olivier Koenig, Fabien Ringeval, Béatrice Bouchot, Gérard Bailly, François Portet, Sina Alisamir, Yongxin Zhou, Jean Serre, Vincent Delerue,

Hippolyte Fournier, Kévin Berenger, Isabella Zsoldos, Olivier Perrotin, Frédéric Elisei, Martin Lenglet, Charles Puaux, Léo Pacheco, Mélodie Fouillen, and Didier Ghenassia. 2021. Theradia: Digital therapies augmented by artificial intelligence. In *Advances in Neuroergonomics and Cognitive Engineering*, pages 478–485, Cham. Springer International Publishing.

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.

Douglas Neil Walton and Erik C. W. Krabbe. 1995. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. Albany, NY, USA: State University of New York Press.

Bin Wang, Chen Zhang, Chengwei Wei, and Haizhou Li. 2022a. A focused study on sequence length for dialogue summarization.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022b. Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4897–4908, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Jerrold H. Zar. 2005. Spearman rank correlation.

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*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Yongxin Zhou, François Portet, and Fabien Ringeval. 2022. Effectiveness of French language models on abstractive dialogue summarization task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3571–3581, Marseille, France. European Language Resources Association.

9. Language Resource References

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236, Prague, Czech Republic. Association for Computational Linguistics.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013*

Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. EmailSum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909, Online. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

10. Appendix A. Example Analysis

In addition to the previous example, we present two other examples in the following. In Figure 3, attention is focused on words such as “enjoyed”, “weekend”, “marvelous”, “kind”, “invite”, “enjoyed having”, “stay”, “Chang sha”, which contain factual information or affective content. In the `baseline-BART`$_{Large}$ model, the words “enjoyed” and “enjoyed having” receive the most attention, and in `baseline_Filtered`, the word “marvelous” attracts the most attention.

In Figure 4, the expression “getting cold feet” is highlighted as a whole, indicating that the models have the ability to understand multi-word expressions. Furthermore, in `baseline_Filtered`, not only “freaking”, “marriage” and “being crazy” are highlighted, but also “jeopardize”.

Reference summary 1: #Person2# tells #Person1# #Person2# enjoyed the weekend organized by #Person1#.
Reference summary 2: #Person1# and #Person2# has enjoyed a nice weekend together.
Reference summary 3: #Person1# invited #Person2# to spend the weekend and #Person2# has enjoyed it.

model_name: baseline-BARTLarge

Generated summary: #Person1# invites #Person2# to stay for the weekend and #Person 2# appreciates it very much.
<s> # Person 1 # : Have you enjoyed your weekend ? # Person 2 # : Yes , it 's been marvelous . It really was very kind of you to invite me . I hope it hasn 't been too much trouble . # Person 1 # : Not at all . We 've really enjoyed having you . I hope you 'll come and stay again next time you 're in Chang sha . </s>

model_name: baseline_Filtered

Generated summary: #Person1# invites #Person2# to Changsha and #Person 2# thinks it's marvelous.
<s> # Person 1 # : Have you enjoyed your weekend ? # Person 2 # : Yes , it 's been marvelous . It really was very kind of you to invite me . I hope it hasn 't been too much trouble . # Person 1 # : Not at all . We 've really enjoyed having you . I hope you 'll come and stay again next time you 're in Chang sha . </s>

Figure 3: Example of test_151, with three references, and predictions as well as visualization of the attention of two models: baseline-BART$_{Large}$ and baseline_Filtered.

Reference summary 1: #Person1# is not ready for marriage. #Person2# reminds him of what he said a month ago and what will happen if he leaves Amy. After listening to #Person2#, #Person1# is ready for marriage.
Reference summary 2: #Person1# feels anxious about future marriage and sweats a lot. #Person2# helps #Person1# to think through the current situation and get ready for the marriage.
Reference summary 3: #Person1# is freaking out before his marriage with Amy and #Person2# helps him get over it by helping him think through all the details.

model_name: baseline-BARTLarge

Generated summary: #Person1# wants to break up with Amy. #Person2# tries to persuade #Person1 # not to do so and thinks he is just getting cold feet.
<s> # Person 1 # : Man , I 'm freaking out ! You gotta help me ! # Person 2 # : Who a , who a , take it easy , relax . G eez , you 're sweating like a pig ! What 's going on ? # Person 1 # : I can 't go through with this ! I just can 't ! I 'm not ready for marriage ! What was I thinking ? I 'm only thirty five years old ! I 've got my entire life ahead of me , adventures waiting ! I can 't settle down yet ! # Person 2 # : What are you talking about ? It wasn 't more than a month ago that you were r ambling on about how you are tired of living the life of a bachelor and how you envy your friends that have a family ! # Person 1 # : I know . I know ! # Person 2 # : Let 's think this through . First of all , you cannot leave Amy at the altar . Not only will she be humiliated and kill you , but she is the woman that you love and the woman of your dreams ! Why would you want to end that or jeopard ize it like this ? Second of all , you are just getting cold feet . You know deep down inside that you want to marry her , so cut the crap and do it ! # Person 1 # : You 're right . I 'm being crazy ! Ok , I 'm ready , let 's do this ! # Person 2 # : Great ! P hew ! That was a close one ! You had me going there for a minute I thought I was gonna have to slap some sense into you . </s>

model_name: baseline_Filtered

Generated summary: #Person1# is not ready for marriage. #Person2# tries to persuade #Person1 # not to end that or jeopardize it like this, but #Person 1# is still getting cold feet. Finally, he decides to marry Amy.
<s> # Person 1 # : Man , I 'm freaking out ! You gotta help me ! # Person 2 # : Who a , who a , take it easy , relax . G eez , you 're sweating like a pig ! What 's going on ? # Person 1 # : I can 't go through with this ! I just can 't ! I 'm not ready for marriage ! What was I thinking ? I 'm only thirty five years old ! I 've got my entire life ahead of me , adventures waiting ! I can 't settle down yet ! # Person 2 # : What are you talking about ? It wasn 't more than a month ago that you were r ambling on about how you are tired of living the life of a bachelor and how you envy your friends that have a family ! # Person 1 # : I know . I know ! # Person 2 # : Let 's think this through . First of all , you cannot leave Amy at the altar . Not only will she be humiliated and kill you , but she is the woman that you love and the woman of your dreams ! Why would you want to end that or jeopard ize it like this ? Second of all , you are just getting cold feet . You know deep down inside that you want to marry her , so cut the crap and do it ! # Person 1 # : You 're right . I 'm being crazy ! Ok , I 'm ready , let 's do this ! # Person 2 # : Great ! P hew ! That was a close one ! You had me going there for a minute I thought I was gonna have to slap some sense into you . </s>

Figure 4: Example of test_440, with three references, and predictions as well as visualization of the attention of two models: baseline-BART$_{Large}$ and baseline_Filtered.