# MIND THE GAP!:

## Injecting Commonsense Knowledge for Abstractive Dialogue Summarization

Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seung-won Hwang, Jinyoung Yeo

COLING 2022

## MOTIVATION

Unlike conversational document-to-document summarization, dialogue-to-document summarization suffers from the discrepancy between input and output forms, which makes learning their mapping patterns more challenging. There are two key challenges that make summarizing dialogues harder than documents.

First, detecting unspoken intention is crucial for understanding an utterance.

Second, there exists information that can only be understood when its hidden meaning is revealed.

**Alyssa**: What do you think about it?
**Derek**: I can fart bright stripes and bright stars better then she sings. ⇒ xWant, to make fun of someone
**Alyssa**: The best part is that she acts like she nailed it. But at least it's funny in a good way.
**Derek**: It is 😂

**Golden summary**
Derek and Alyssa make fun of Fergie's performance of the national anthem.

**Melody**: youre probably due for a new one anyway, no?
**Peggy**: you're right. 5 years is a long time to own one.
**Melody**: yes, thats ancient by laptop standards
⇒ HinderedBy, the laptop is too old
**Peggy**: ok. i might just not bother getting it repaired after all.
**Melody**: sounds like a good idea

**Golden summary**
Melody's 5-year-old laptop is broken. Tomorrow she'll know what's wrong. She won't be repairing it, because her laptop is too old. Instead, she'll buy a new one.

## FILL IN THE GAP

| | |
|---|---|
| **Utterance** | Charlie : Do you really believe that dreams can mean something? |
| HINDEREDBY | Charlie doesn't believe in dreams. |
| XWANT | to talk to someone about dreams. |
| XINTENT | to believe in something. |
| XNEED | to have a dream. |
| XREASON | Charlie is a skeptic. |

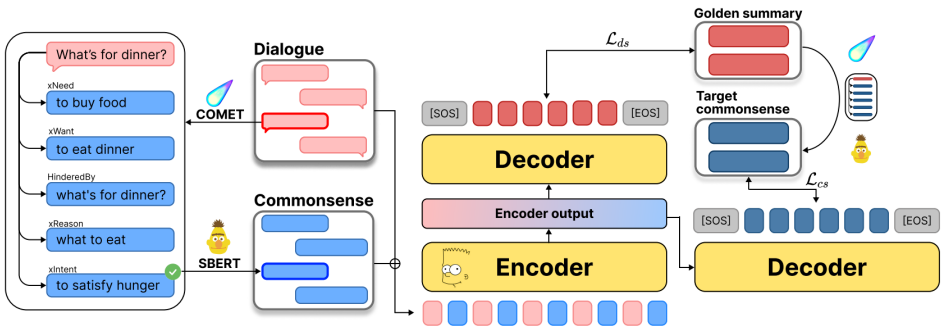
Table 1: Example of commonsense knowledge generated by COMET given a dialogue.

| | |
|---|---|
| **Prev-Utterances** | Jane : google maps says it is at least 3h<br>Steven : I used to make it in 2, trust me<br>Jane : but it's almost 300km<br>Steven : the road is new , we will make it |
| **Utterance** | Jane : I don't want to stress out, let's meet at 4:30 instead of 5, ok? |
| XINTENT | to avoid stress. |
| XWANT | **to not be late.** |
| XREACT | annoyed |
| XEFFECT | PersonX sweats from nervousness. |
| XATTR | nervous. |

Table 2: Example of commonsense knowledge generated by PARA-COMET given a dialogue.

Commonsense knowledge models such as COMET can generate a set of event-centered (e.g.,HINDEREDBY, XREASON, XNEED) and social interaction(e.g., XINTENT, XWANT) commonsense inferences.

## METHOD




Our task definition follows a sequence-to-sequence learning problem setting.

Our goal is to learn a mapping function M : D → Y where D = {u1, u2, ..., un} is a dialogue with n utterances, and Y = {y1, y2, ..., ym} is a corresponding summary of m utterances.

(STEP 1) We generate and filter to acquire a set of commonsense knowledge C = {c1, c2, ..., cn} based on D.

$$c_i = \underset{c_i^r}{\text{argmax}}(\text{score}(u_i, c_i^r)) \quad (r \in \mathcal{R})$$

(STEP 2) Then, we adjust the mapping function as M : X → Y, where X is a cross concatenation of D and C.

$$\mathcal{X} = \mathcal{D} \oplus \mathcal{C} = \cdots \parallel u_i \parallel \text{<I>} c_i \text{</I>} \parallel \cdots \qquad \mathcal{L}_{ds} = -\sum_{i=1}^{|\mathcal{Y}|}\sum_{j=1}^{|y_i|} \log P(w_{i,j}|w_{i<j}, x; \theta_{ds})$$

(STEP 3) we add an auxiliary task commonsense supervision, M*: X → Z, where the target commonsense Z = {z1, z2, ..., zm} is acquired based on Y.

## RESULT

| Model | SAMSum R-1 | SAMSum R-2 | SAMSum R-L | SAMSum B-S | DialogSum R-1 | DialogSum R-2 | DialogSum R-L | DialogSum B-S |
|---|---|---|---|---|---|---|---|---|
| PointerGenerator (See et al., 2017)* | 32.27 | 14.42 | 34.36 | / | / | / | / | / |
| DynamicConv (Wu et al., 2019)* | 41.07 | 17.11 | 37.27 | / | / | / | / | / |
| Transformer (Vaswani et al., 2017)* | 42.37 | 18.44 | 39.27 | / | / | / | / | / |
| DialoGPT (Zhang et al., 2020c)† | 39.77 | 16.58 | 38.42 | / | / | / | / | / |
| BART-xsum (Lewis et al., 2020)† | 51.74 | 26.46 | 48.72 | 53.87 | / | / | / | / |
| UniLM (Dong et al., 2019)† | 47.85 | 24.23 | 46.67 | / | 42.38 | 16.88 | 34.36 | 69.40 |
| PEGASUS (Zhang et al., 2020a)† | 50.50 | 27.23 | 49.32 | 53.35 | 38.40 | 13.84 | 33.41 | 68.20 |
| BART-xsum (Lewis et al., 2020)‡ | 52.50 | 27.67 | 48.75 | 68.16 | 45.15 | 19.78 | 36.57 | 71.09 |
| D-HGN (Feng et al., 2021) | 42.03 | 18.07 | 39.57 | 64.20 | / | / | / | / |
| S-BART (Chen and Yang, 2021) | 50.70 | 25.50 | 48.08 | 70.07 | / | / | / | / |
| CODS (Wu et al., 2021) | 52.65 | 27.84 | **50.79** | 66.55 | 44.27 | 17.90 | 36.98 | 70.49 |
| SICK w/ COMET (Ours) | 53.04 | 27.60 | 48.49 | 71.61 | 45.70 | 20.08 | 40.26 | 71.08 |
| SICK++ w/ COMET (Ours) | 53.24 | 28.10 | 48.90 | 71.71 | **46.26** | **20.95** | **41.05** | 71.30 |
| SICK w/ PARA-COMET (Ours) | 53.39 | 28.42 | 49.12 | 71.83 | 46.01 | 20.30 | 40.75 | **71.57** |
| SICK++ w/ PARA-COMET (Ours) | **53.73** | **28.81** | 49.50 | **71.92** | 46.20 | 20.39 | 40.83 | 71.32 |

In SAMSum(Gilwa et al., 2019), SICK++ shows better performance with PARACOMET(Gabriel et al., 2021) than with COMET, however it shows opposite result in DialogSum(Chen et al., 2021).

We conjecture this due to the characteristic of datasets and commonsense models hold. Since SAMSum has shorter length of dialogues than DialogSum, the recurrent memory component of PARA-COMET is less likely to forget the previous sentences.

We expect to get better performance with the help of commonsense-models that maintains longer memories of sentences/ dialogues and leave this as future research.

## ANALYSIS

### Commonsense Applicability

| Model | SAMSum R-1 | SAMSum R-2 | SAMSum R-L | SAMSum B-S | DialogSum R-1 | DialogSum R-2 | DialogSum R-L | DialogSum B-S |
|---|---|---|---|---|---|---|---|---|
| BART-xsum | 20.83 | 4.28 | 15.28 | 46.59 | 17.40 | **4.16** | 13.80 | 42.97 |
| SICK | **23.12** | **5.09** | **17.45** | **47.69** | **18.32** | 3.80 | **14.98** | **43.97** |

Table 6: Zero-shot evaluation on SAMSum and DialogSum test set.

We experiment in a zero-shot setting to examine how commonsense knowledge solely affects dialogue summarization. We find that SICK outperforms BART-xsum, where the performance gain comes from additional commonsense. This also supports the idea that commonsense is essential in resolving the discrepancy between dialogues and documents.

### Data Efficiency & Effect of commonsense supervision

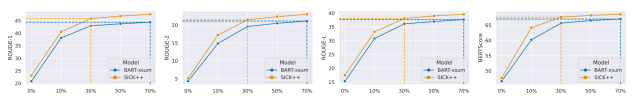
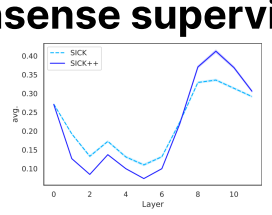


Figure 3: Performance of BART-xsum and SICK++ on SAMSum by varying the size of training data. We use $\text{BART}_{\text{BASE}}$ for both of them. Details are shown in Appendix.

Our method is helpful in situations where data is insufficient, meaning there is a trade-off (time vs data efficiency).

With only 30% of training data, SICK++ shows better performance than BART-xsum(base architecture) trained with 70% of training data.




Attention weights can be viewed as governing how "important" every other token is when producing the next representation for the current token (Clark et al., 2019).

Rogers et al. (2020) mentioned that final layers of language models are most task-specific, and we observe that SICK++ has marginally higher attention values.
SICK++ has marginally higher attention values than SICK in the upper layers, and we conjecture this due to the supervision provided by generating Z.

## CONCLUSION

In this work, we propose SICK and SICK++ framework to resolve the two key challenges:

i) filling in the gap in dialogues
ii) injecting commonsense knowledge into a model.

We show that the difficulties in dialogues are resolved with commonsense knowledge and demonstrated that our framework can successfully inject commonsense knowledge.

As a result of injected commonsense knowledge, we obtain competitive results on SAMSum and DialogSum benchmarks.

## QR CODE