Improving and Evaluating Contrastive Learning in Abstractive Summarization with a Better Baseline and More Variable Datasets

Zhuowen Shen
University of Michigan
mickshen@umich.edu

Yiwen Yao
University of Michigan
yywynn@umich.edu

Abstract

Summary generation is a useful application of Neural Language Processing (NLP) in practice. Most researches on this topic are applying the sequence-to-sequence (Seq2Seq) learning framework. Contrastive learning is a new powerful approach in self-supervised learning. SimCLS (Liu and Liu, 2021) is a simple framework for contrastive learning of abstractive summarization and was applied on CNNDM (Nallapati et al., 2016a) and XSum (Narayan et al., 2018) datasets. This project is to apply SimCLS framework on more variable datasets to compare their performances and further, we would like to improve the model structure using SimCSE (Gao et al., 2021) and training process to obtain a better performance. The codes are shown in https://github.com/MickShen7558/595-project.

1 Introduction

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks) (Maybury, 1999). Due to an over-abundance of data, and lack of manpower and time to interpret the data, automatic NLP methods for text summarization become crucial in today's world. Automatic text summarization has demonstrated its strength of more efficiency, less labor, and less biased compared to human summarizers (Stiennon et al., 2020).

Taking the advantage of Machine Learning and Deep Neuron Network, engineers now can construct neural networks to learn the hidden features in long sentences and accomplish the task of text summarization. Among all the models, the Seq2Seq neural models (Sutskever et al., 2014) have been widely used for text generation tasks, especially, the abstractive summarization (Nallapati et al., 2016b).

Contrastive learning aims to learn effective representation by pulling semantically close neighbors together and pushing apart non-neighbors (Hadsell et al., 2006). In recent years, the successful application of Contrastive Learning in computer vision (Chen et al., 2020) attracted the attention of the NLP scientists. Then the idea of contrastive learning was proven to work pretty well in text summarization tasks (Gao et al., 2021; Liu and Liu, 2021).

In this work, we propose to further improve the training process of contrastive learning in abstractive summarization task by applying a more holistic sentence embedding model SimCSE to do the contrastive learning. The core methodology of creating the contrastive task in this work is inspired by SimCLS. We incorporate the unsupervised version of SimCSE as a substitution to RoBERTa (Liu et al., 2019) in SimCLS, which offers more sentence-level features while preserving the token-level knowledge. Also, we would test the generality of the contrastive model on other various datasets.

2 Related Works

Many researches and methods are applied to the field of abstractive summarization.

2.1 Current Methods and Challenges

In (Lewis et al., 2020; Zhang et al., 2020), they have demonstrated promising potentials of Seq2Seq models in a wide range of tasks: abstractive dialogue, question answering, and summarization tasks. However, in text summarization, Seq2Seq models are still sharing some challenges in the training process. For example, Seq2Seq model is usually trained using Maximum Likelihood Estimation, while in practice they are trained with the teacher-forcing (Williams and Zipser, 1989) algorithm. This results in the incoherence between the objective function and the evaluation metrics (so called *exposure bias* in previous work (Bengio

et al., 2015)), where the objective function is based on local, token-level predictions while the evaluation metrics would compare the holistic similarity on the sentence-level (Liu and Liu, 2021).

2.2 Contrastive Learning

Due to the characteristic of contrastive learning and the success of SimCLR (Chen et al., 2020) in computer vision, NLP scientists start to use contrastive learning as an objective function to fine tune a pre-trained model. SimCSE (Gao et al., 2021) is a successful example. SimCSE uses the positive and negative pairs to form a contrastive learning so that the network can have a better sentence embeddings on semantic textual similarity tasks. SimCSE uses BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) as baselines and uses two independent random dropout layers to generate two sentence embeddings from one original embedding to have the positive pair and use the embedding from other sentences to create negative pair. Then SimCSE achieves unsupervised (self-supervised) learning. SimCSE can also be applied to supervised dataset by setting the embedding and the entailment as pairs. The unsupervised and supervised SimCSE using $\text{BERT}_{\text{base}}$ are shown to have 4.2% and 2.2% improvement compared to previous best results (Gao et al., 2021).

SimCLS (Liu and Liu, 2021) takes the advantage of contrastive learning to reduce the exposure bias in MLE structure. SimCLS uses BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2020) as base systems to generate summarization candidates, where the candidates of the same article vary a little from each other. SimCLS introduces a scoring model for evaluation using contrastive learning. Differing from the positive pairs and negative pairs in traditional contrastive learning models (Chen et al., 2020; Wu et al., 2020), SimCLS uses cosine similarity over all candidates. Then SimCLS selects the best candidate as the text summarization. In SimCLS, it's also proven that seleting the best summarization among candidates outperforms generating one summarization on average.

2.3 Datasets

Lots of datasets have been created to train and evaluate the text summarizaion.

- CNNDM CNN / Daily Mail (Nallapati et al., 2016a) dataset is a large dataset containing long online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average).

- XSum (Narayan et al., 2018) dataset is a highly abstractive dataset where in average the length of article is 431 words (20 sentences) and the length of summary is 23 words.

- Gigaword (Rush et al., 2015) dataset contains sentence summarization (8.3 tokens) with very short input documents (31.4 tokens).

- Reddit Webis-TLDR-17 Corpus (Völske et al., 2017), content and self-written summaries mined from Reddit, which is the first dataset in social media domain.

Many works have been working on these listed datasets, including the BART, Pegasus, GSum (Dou et al., 2020), Prophet (Qi et al., 2020), Control-Copying (Song et al., 2020), and Unified VAE + PGN (Choi et al., 2019).

3 Approach

We used the same framework as SimCLS, as shown in Fig. 1.

Given the article D and the reference summary $\hat{S}$, the goal of the summarization model f is to generate the candidate summary $S = f(D)$ such that it receives the highest score $m = M(S, \hat{S})$, where M is the evaluation metric and in our work we choose as ROUGE.

3.1 Candidate Loading

For the first stage to generate candidates, we will use the same Seq2Seq generator (BART (Lewis et al., 2020)) as SimCLS to generate candidate summaries $S_1, \cdots, S_n$ with a sampling strategy such as Beam Search to select those candidates.

SimCLS has already provided the preprocessed datasets for XSum and CNNDM. We constructed scripts to preprocess Reddit and Gigaword datasets. We loaded the complete Gigaword and Reddit datasets from Tensorflow datasets. Then we generated 16 candidates for each example via the bart model pretained on CNNDM by Facebook. The inputs to the next stage include the tokenized and untokenized articles, the tokenized and untokenized summaries, and the tokenized and untokenized candidates.

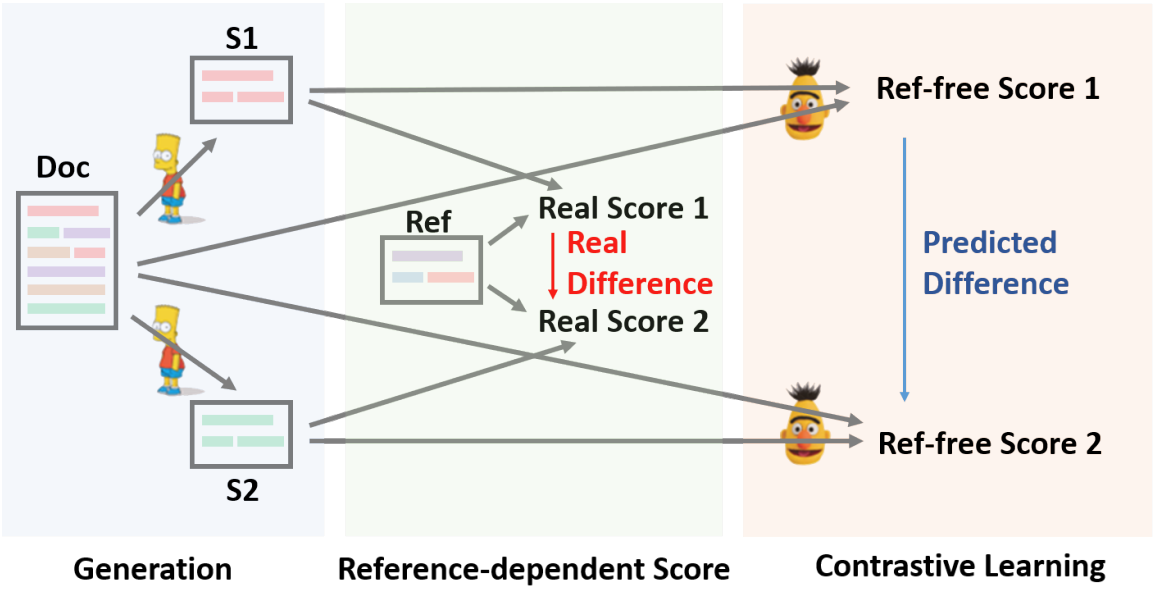

Figure 1: SimCLS framework for two-stage abstractive summarization, where Doc, S, Ref represent the original document, generated summary and reference respectively (Liu and Liu, 2021). The left part is the candidate loading network, where we generate various candidate for each article and select some best candidates. The middle part calculate the score, $i.e.$ the similarity between the candidates and the reference summary. The right part is the contrastive learning which ranks the summarization quality of the candidates and therefore trains the scoring model in the middle part.

3.2 Candidate Evaluation

In the candidate evaluation, the idea is that a better candidate summary S_i should receive a higher score with respect to the source article D, Therefore, we construct the scoring function $h(\cdot)$ using the pre-trained sentence embedding model $\varphi(\cdot)$:

$$h(S_i, D) = \textit{cosine similarity}(\varphi(S_i), \varphi(D)) \quad (1)$$

where the *cosine similarity* is calculated between the encoding of the first tokens, same as the SimCLS.

To take the advantage of the SimCSE on semantic textual similarity, we use the unsupervised SimCSE-RoBERTa$_{\text{base}}$ as our φ. Since SimCSE is also trained contrastively over different candidates, we believe that it would help improve the network performance on finding the best summarization.

Then we output the summary S with the highest score:

$$S = \underset{S_i}{\text{argmax}} h(S_i, D) \quad (2)$$

The last step is to construct a contrastive learning to train the scoring model give the best summary the highest score. We first rank the candidates $\widetilde{S}_1, \cdots, \widetilde{S}_n$ in descending order based on $M(\widetilde{S}_i, \hat{S})$. Then we adopt the ranking loss in SimCLS (Liu and Liu, 2021):

$$\begin{aligned} L = &\sum_i \max(0, h(D, \tilde{S}_i) - h(D, \hat{S})) \\ &+ \sum_i \sum_{j>i} \max(0, h(D, \tilde{S}_j) - h(D, \tilde{S}_i) + \lambda_{ij}) \end{aligned} \quad (3)$$

4 Evaluation

Similar to (Liu and Liu, 2021), we evaluate the sentence-level learning quality by ROUGE scores (Lin, 2004) as main evaluation metrics. We use ROUGE-1 and ROUGE-2, which evaluate the overlapping of unigram and bigrams between the system and reference summaries, and ROUGE-L, which takes into account sentence level structure similarity using the longest common sebsequence.

4.1 Experiment Details

In out experiment, we train and test our model and the original SimCLS model with the same amount of data to see the performance difference. The size of the training datasets, evaluation datasets and testing datasets are $10,000$, $1,000$ and $1,000$, which is shown in Tab. 1.

The generator that we used is BART (facebook/bart-large-cnn on hugging face).

	CNNDM	XSum	Reddit	Gigaword
Train	10,000	10,000	10,000	10,000
Test	1,000	1,000	1,000	1,000
Val	1,000	1,000	1,000	1,000

Table 1: Sizes of portions of the datasets we used to train, test and evaluate the networks.

We use the same learning rate scheduling formula as SimCLS (Liu and Liu, 2021), and the formula is

$$lr = 0.002 \cdot \min(\text{step\_num}^{-0.5}, \text{step\_num} \cdot \text{warmup\_steps}^{-1.5}), \quad (4)$$

where the warm-up steps is $2,000$. Some critical hyper parameters we used are shown in Tab. 2.

Parameter	Value
Batch Size	1
Epoch Number	5
Report Frequency	100
Accumulate Step	12
Margin	0.01
Gold Margin	0
Gradient Normalization	0
Gold Similarity	False
Max Learning Rate	2e-3
Scale	1
Data Type	diverse
Max Length	120
Min Length	16
Candidate Weight	1
Gold Weight	1

Table 2: Hyper Parameters

For the original SimCLS, the scoring model is RoBERTa (roberta-base on hugging face). For our model, we used unsupervised RoBERTa based SimCSE (princeton-nlp/unsup-simcse-roberta-base on hugging face).

4.2 Results

4.2.1 Qualitative Results

In this part, we will show the references, the summaries generated by our model, the summaries generated by SimCLS, and the untrained model.

Tab. 3 shows the example of CNNDM dataset.

Tab. 4 shows the example of XSum dataset.

Tab. 5 shows the example of Reddit dataset.

Tab. 6 shows the example of Gigaword dataset.

4.2.2 Quantitative Results

The ROUGE scores on datasets CNNDM, XSum, Gigaword, and Reddit are shown separatedly in Tab. 7, 8, 9, 10.

5 Discussion

5.1 Framework

As demonstrated in Sec. 3, our framework can be divided into two parts. After generating the candidates and using beam search to find 16 best candidates, the generators would not be used or updated in the later learning process. Therefore, the upper-limit of the quality of the final output in our framework is determined by the existing generators, and the usage of vanilla beam search (which is not differentiable) wouldn't cause problem in our training.

5.2 Experiment Process

In the original SimCLS, they set warm-up steps as 10000 and trained for 40 hours on 4 GTX-1080-Ti GPUs on CNNDM dataset and 20 hours on XSum dataset. With very limited computational resources, clearly it's not possible for us to fully reproduce or fine-tune SimCLS, or to train our new model on any one of the four datasets that we are focusing on in our work. Therefore, to have a general idea on the performance of our network and SimCLS on four different dataset, we speed up the training by reducing the warm-up steps and reduced the size of dataset. Though we cannot say that which network performs better, we can find how fast each network is learning. By comparing to other SOTA networks, we can also verify the strong power of our network and SimCLS.

5.3 Qualitative Results

In this part, we will discuss the results shown in Tab. 3, 4, 5 and 6 to get a qualitative understanding of our model.

From the tables, we can see that the summaries generated our model contains equal or more information than in the summaries generated by SimCLS. Another observation is that the generated summaries are sometimes incomplete due to the limitation of generators, and in this case, our scoring model can compensate these issues and select the best candidate.

It is notable that when the article is long, the generated summaries tend to contain information appearing earlier in the article. For example, in the

System	Summary	Article
Ref.	gary gardner confirms he'll report to aston villa for pre-season training. the 22-year-old is out on loan at championship side nottingham forest. tim sherwood is keen to asses gardner ahead of next season. the midfielder would prefer a move back to forest if villa does n't wok out. click here for all the latest aston villa news.	tim sherwood will welcome gary gardner back into his first team squad during pre-season to assess closely whether the 22-year-old can cut it for aston villa in the premier league. gardner has enjoyed a successful loan spell at nottingham forest and scored a superb free-kick in front of the villa manager during the defeat to watford at the city ground. sherwood is keen to see gardner, a former england under-21 player, train day-to-day before making up his mind and he has spoken previously about his desire for homegrown players to make an impact at the club. gary gardner -lrb- left -rrb- will report to aston villa for pre-season training to be assessed by tim sherwood. the villa boss has inspired the club since being appointed and helped lead them to an fa cup final. gardner, born in solihull, joined villa's academy aged seven and last year was rewarded with a new contract that expires in june 2016 despite suffering serious knee injuries.'i will go back to villa in the summer and we will see what happens,'he said. 'when i am back there in training, it will be my chance to impress. tim sherwood will not have seen much of me in training, because i have not been there. 'villa are the main club, i have been there since i was seven and it is the team i support. but forest have been fantastic to me. it is definitely the second team in my heart. 'if i was made available for loan again next season, if it does not work out at villa, forest would be top of my list, definitely. it is the best loan move i have had.'gardner scored a stunning free-kick against watford with sherwood in attendance at the city ground.
SimCLS	gary gardner will report to aston villa for pre-season training. tim sherwood will assess the 22-year-old's fitness. gardner has enjoyed a successful loan spell at nottingham forest. the former england under-21 star scored a stunning free-kick against watford.	
Our Model	tim sherwood will welcome gary gardner back to aston villa for pre-season training. the villa boss will assess closely whether the 22-year-old can cut it in the premier league. gardner has enjoyed a successful loan spell at nottingham forest. the former england under-21 star scored a stunning free-kick against watford with sherwood in attendance.	
Origin.	gary gardner has enjoyed a successful loan spell at nottingham forest. the 22-year-old will report to aston villa for pre-season training. tim sherwood will assess closely whether gardner can cut it for villa. gardner scored a stunning free-kick against watford. the former england under-21 star joined villa's academy aged seven.	

Table 3: The reference, SimCLS generated summary, our model generated summary, and untrained model generated summary example of CNNDM dataset

three generated summaries of SimCLS, our model and untrained model, they all miss the information "want to make sure she doesn't get hurt" in the reference, and this issue is probably caused by the generator. Comparing the rest of these summaries, the summaries generated by our model usually have a better quality than the ones generated by SimCLS and the untrained model. So we think under the same condition of candidate generator, our model is more likely to select the candidate with a greater similarity to the reference and the article.

When the article is short, there is no large difference between the three generated summaries, and most summaries are almost the same, so it more sense to look at the quantitative analysis on the Rouge scores.

5.4 Quantitative Results

From Tab. 7, 8, 9 and 10, we can see that in all four datasets that we tested on, with a tiny amount of training, our new model can achieve comparably good ROUGE scores. For XSum, our new model can already beat the SOTA. For CNNDM and Reddit, our new model has an analogous scores compared to the SOTAs.

Our network didn't reach the SOTA in Gigaword, and was beaten by a lot. The main reason is due to the low performance of the base generators. We can see that **Our**$_{raw}$ only have 31.40, 11.33, 28.19 for R-1, R-2, and R-L. This is a very low score compared to the SOTA. However, with tiny amount of training, the scorer can distinguish a much better candidate and improves the performance by a lot. This also help to prove that our framework has the ability to dig the potential of the existing networks.

We also notice that we beat SimCLS in CNNDM by a lot. This means that our network does improve the training process to some extent. In the other three datasets, out network also has similar scores compared to SimCLS. This shows that our network has a stronger learning ability on a tiny training dataset compared to SimCLS. With more training time and more warm-up runs, our network is highly expected to outperforms the SimCLS.

6 Conclusion

In this work, we used a new pre-trained model in SimCSE to the existing contrastive summarization network SimCLS. We take the advantage of the new pre-trained network in its high performance in text similarity, which plays a very important role as a score function in the contrastive

System	Summary	Article
Ref.	a book about the death of a british officer in afghanistan, once pulped by the ministry of defence, has won the orwell prize for political writing.	dead men risen, written by toby harnden and published by quercus, took the prize at a ceremony in westminster and was the judges'unanimous choice. it focuses on the death of lt colonel rupert thorneloe in 2009. it was published in amended form after the first print run was destroyed by the mod. the judges said the book "takes us into the hearts and minds of the welsh guards in a way that is both interesting and visceral". "it challenges every citizen of this country to examine exactly what we're asking soldiers to do in afghanistan,"the panel continued. "rather than offering easy answers it lets the soldiers speak for themselves."other awards presented included a posthumous honour for christopher hitchens, whose final book, arguably, was included on the long list for the top prize. hitchens'widow carol blue accepted the award on behalf of the vanity fair writer. the journalism prize was awarded to amelia gentleman for her work in the guardian, while the blog award went to rangers tax case - an online commentary on the ongoing financial problems at the historic scottish football club. the writers of the blog said they aimed to "provide the details of what rangers fc have done, why it was illegal and what the implications are for one of the largest football clubs in britain". the winners were chosen from shortlists of six books, six journalists and seven bloggers, whittled down from longlists of 17 books, 12 journalists and 18 bloggers. each of the winners received a \u00e2 3,000 prize.
SimCLS	a book about a welsh guards soldier killed in afghanistan has won the bollinger everyman book of the year award.	
Our Model	a book about a british soldier killed in afghanistan has won this year's man book international prize.	
Origin.	a book about the death of a british soldier in afghanistan has won the bollinger everyman book prize.	

Table 4: The reference, SimCLS generated summary, our model generated summary, and untrained model generated summary example of XSum dataset

task. Since this pre-trained network is naturally trained over contrastive task, it can help accelerating the training process in the task of selecting best summary. This also helps to mitigate the discrepancy between the training and test processes in the MLE framework in the original SimCLS. We evaluate our new model on the CNN / Daily Mail and XSum datasets within very limited computational resources and compared the learning progress with the original SimCLS. As we expected, the new model out-performs the SimCLS on the CNNDM dataset and receives approximate score on XSum, showing that our modification does improves the framework with the text summarization task.

To test the power of contrastive learning and generating summaries by selecting from candidates, we also evaluated the performance of the new model and the original SimCLS on two different datasets, Gigaword and Webis-TLDR-17 Corpus. After this little amount of training, both our network and the SimCLS performs well on those dataset, and the ROUGE scores are comparable to other SOTA papers.

All these demonstrated the huge power of the contrastive learning model and the framework of candidate generation plus scoring. This work expands SimCLS's direction in improving the training process and extending to other different styled datasets. This work also proves that existing abstractive systems have the potential of generating candidate summaries much better than the original outputs, and this scoring framework can help realize their potentials. This opens a new direction for future approaches on realizing the potentials of existing networks.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv preprint arXiv:1506.03099*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Hyungtak Choi, Lohith Ravuru, Tomasz Dryjański, Sunghan Rye, Donghyun Lee, Hojung Lee, and Inchul Hwang. 2019. VAE-PGN based abstractive model in multi-stage architecture for text summarization. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 510–515, Tokyo, Japan. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

System	Summary	Article
Ref.	my mil is involved with a guy who i think is a scam artist. i have little info to go off of and i want to make sure she doesn't get hurt.	so this is a long and complicated story. i'll do my best to summarize and keep it as succinct as possible. my mother-in-law is a widow of 2 years now. she's in her late 50's and was married at 19. she's extremely vulnerable, stubborn, and simply not very bright. after her husband passed, she had no interest in dating. she was seemingly content to live out the rest of her life alone. however, recently, she informed my wife that she was talking to someone. she was at the hospital for some health issues and met a guy. he was working in the hospital on the computer systems. he is a business owner from finland living in the us. he is also a widow with a daughter who normally lives with him but is currently overseas visiting her grandmother back in finland. as my mil hadn't dated since the 70s, she was very reluctant. he pursued and she eventually relented to sending emails back and forth. after about 2 months of emails, she finally gave him her phone number and they started texting. texting turned into phone calls. they are now at the point where they want to meet in person and start having a real relationship. at first, i was very happy for her. she's a pain in the ass and is essentially helpless as an adult so she relies heavily on my wife to handle a lot of things for her that she lacks the cognitive or physical capacity to handle herself. i felt this would be great for both her and my wife. she finally opened up to my wife and her sister about it and was willing to talk about it with about 5-6 people. my sil, in her true spoiled brat fashion, freaked out and refused to accept this. my wife was supportive so my mil opened up a lot to her about some things. the more i hear, the more i (and my wife) am fairly convinced that this is not kosher. now all of the following can be one big coincidence and he could be prince fuckin' charming. but, these are all the reasons i think this is a scam in some way, shape, or form. individually, they all seem superficial and inconsequential but when you add them all up, something is off. he was dying to spend time with her in the beginning but she wasn't ready. now that she is ready, he's really busy with work. he said that as soon as he's done with this big project, he's going to take 2 months off and they'll have time to spend together, yet they haven't spent one day together from the day they met in the hospital. he supposedly looks like j̈on snow's father.ï've never seen him before so i can't verify. but someone who looks like this, is several years younger than my mil, and is a wildly business owner is unlikely to go for someone like my mil...a frumpy half-assed housewife. he played the fellow widow card. he claimed he was getting a divorce as he moved to the us and she stayed in finland. then, in the process of the divorce, she passed away. details of her death are unclear to me and my wife. my mil won't give any information other than his first name. this is not his fault but i can't verify anything about this person without so much as a last name (re: alias). here's the kicker that set this wheel in my head in motion. he claimed that while his daughter was in finland and supposed to come back home to america this week, she was in a severe car accident. the doctors told his mother that it was unclear as to whether she was going to make it through or not. my mil was naturally very upset by this. she is about to go in for a breast reconstruction surgery this week (she's an inflammatory breast cancer survivor) and the timing is very convenient. now i could be reading way too much into this and i 1000% hope i am. but all of these things together make me 95% sure this guy is not on the level. call it intuition. so my question is, with such little information to go off of, and with my mil knowing as little as possible, how can i proceed to make sure she doesn't get hurt financially or otherwise?
SimCLS	my mother-in-law met a man in the hospital and they started texting and talking. he is a business owner from finland living in the us and a widow with a daughter. he claims he was getting a divorce as he moved to	
Our Model	the woman is a widow of 2 years and has no interest in dating. she met a guy at the hospital and they started texting and talking. they are now at the point where they want to meet in person and start having a real relationship.	
Origin.	my mother-in-law is a widow of 2 years now. she's extremely vulnerable and stubborn, and simply not very bright. she told her wife that she was talking to someone. he is a business owner from finland living in	

Table 5: The reference, SimCLS generated summary, our model generated summary, and untrained model generated summary example of Reddit dataset

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

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.

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 preprint arXiv:1907.11692*.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv:2106.01890*.

Mani Maybury. 1999. *Advances in automatic text summarization*. MIT press.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016a. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th*

System	Summary	Article
Ref.	zimbabwe opposition leader pessimistic about talks	opposition leader morgan tsvangirai expressed pessimism wednesday about talks on zimbabwe's proposed government of national unity.
SimCLS	zimbabwe opposition leader morgan tsvangirai expresses pessimism over proposed government of national unity.	
Our Model	opposition leader morgan tsvangirai expresses pessimism about talks on zimbabwe's proposed government of national unity.	
Origin.	opposition leader morgan tsvangirai expressed pessimism wednesday about talks on z zimbabwe's proposed government of national unity.	

Table 6: The reference, SimCLS generated summary, our model generated summary, and untrained model generated summary example of gigaword dataset

System	R-1	R-2	R-L
BART(SimCLS$_{raw}$)*	44.16	21.28	40.90
Pegasus*	44.17	21.47	41.11
Prophet*	44.20	21.17	41.30
GSum*	45.94	22.32	42.48
Our$_{raw}$	43.95	19.95	40.92
Our	45.49	21.33	42.44
SimCLS$_{little\ data}$	44.90	20.86	41.91
SimCLS	46.67	22.15	43.54

Table 7: Results on CNNDM. SimCSE$_{raw}$ are from the network never trained through this contrastive task. **Our** and SimCLS$_{little\ data}$ are from the networks trained on a small portion of the whole datasets in our experiment. SimCLS is from the trained model provided by (Liu and Liu, 2021). R stands for ROUGE. *: results reported in the original papers.

System	R-1	R-2	R-L
BART*	45.14	22.27	37.25
Pegasus(SimCLS$_{raw}$)*	47.21	24.56	39.25
GSum*	45.40	21.89	36.67
Our$_{raw}$	47.16	24.74	38.91
Our	47.31	25.07	39.34
SimCLS$_{little\ data}$	47.57	25.19	39.53
SimCLS	47.61	24.57	39.44

Table 8: Results on XSum. Systems' meanings are the same as in Tab. 7. R stands for ROUGE. *: results reported in the original papers.

System	R-1	R-2	R-L
Pegasus(SimCLS$_{raw}$)	18.69	3.70	16.15
Our$_{raw}$	18.51	3.73	15.87
Unified VAE + PGN*	20	4	14
Our	19.93	4.41	17.05
SimCLS$_{little\ data}$	20.09	4.34	17.17

Table 9: Results on Reddit. Systems' meanings are the same as in Tab. 7. R stands for ROUGE. *: results reported in the original papers.

System	R-1	R-2	R-L
BART(SimCLS$_{raw}$)	29.55	10.40	26.93
Our$_{raw}$	31.40	11.33	28.19
Pegasus*	39.12	19.86	36.24
Prophet*	39.51	20.42	36.69
ControlCopying*	39.08	20.47	36.69
Our	37.87	15.95	34.61
SimCLS$_{little\ data}$	37.53	15.74	34.38

Table 10: Results on Gigaword. Systems' meanings are the same as in Tab. 7. R stands for ROUGE. *: results reported in the original papers.

SIGNLL Conference on Computational Natural Language Learning, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016b. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *CoRR*, abs/1808.08745.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *CoRR*, abs/1509.00685.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8902–8909.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

A Work division

We discussed on the process and made decision together. The workloads are shared fairly within the team, and each team member has made equivalent contribution to this project. The whole division of our work is mainly shown as following:

Zhuowen Shen

- Reading papers
- Collecting datasets
- Reproducing SimCLS results
- Assembling SimCSE with SimCLS and vanity check
- Testing our network on various datasets
- Tuning hyperparameters and evaluating the results
- Writing final report

Yiwen Yao

- Reading paper
- Environment Setup
- Reproducing SimCLS results
- Testing SimCLS on various datasets
- Improve the algorithm
- Writing final report