Get To The Point: Summarization with Pointer-Generator Networks

Abigail See
Stanford University
abisee@stanford.edu

Peter J. Liu
Google Brain
peterjliu@google.com

Christopher D. Manning
Stanford University
manning@stanford.edu

Motivation

- Neural sequence-to-sequence models have provided a viable new approach for abstractive text summarization (meaning they are not restricted to simply selecting and rearranging passages from the original text). However, these models have two shortcomings: they are liable to reproduce factual details inaccurately, and they tend to repeat themselves. In this work we propose a novel architecture that augments the standard sequence-to-sequence attentional model in two orthogonal ways.

- First, we use a hybrid pointer-generator network that can copy words from the source text via pointing, which aids accurate reproduction of information, while retaining the ability to produce novel words through the generator. Second, we use coverage to keep track of what has been summarized,which discourages repetition. We apply our model to the CNN / Daily Mail summarization task, outperforming the current abstractive state-of-the-art by at least 2 ROUGE points.

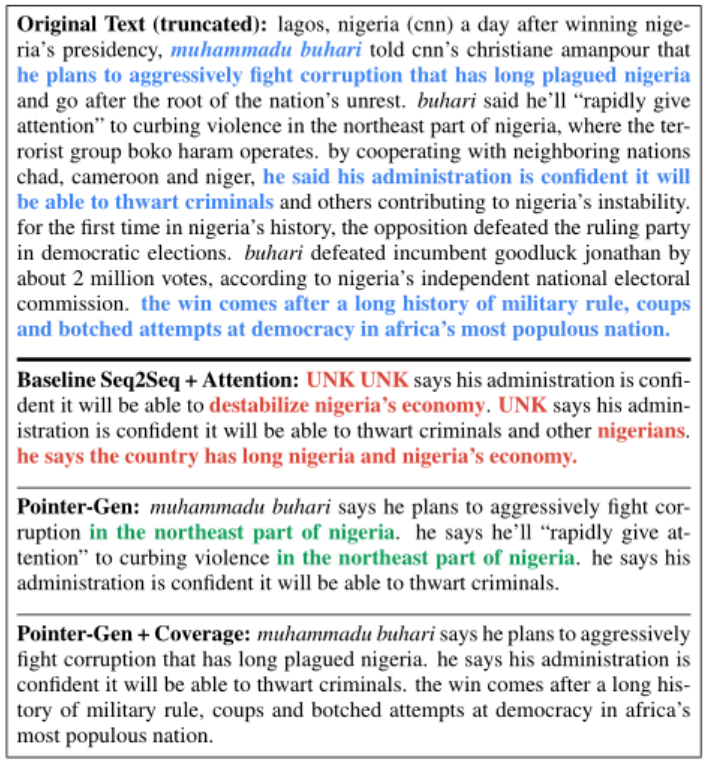

Figure 1: Comparison of output of 3 abstractive summarization models on a news article. The baseline model makes **factual errors**, a **nonsensical sentence** and struggles with OOV words *muhammadu buhari*. The pointer-generator model is accurate but **repeats itself**. Coverage eliminates repetition. The final summary is composed from **several fragments**.

Introduction

- Summarization is the task of condensing a piece of text to a shorter version that contains the main information from the original. There are two broad approaches to summarization: extractive and abstractive. Extractive methods assemble summaries exclusively from passages (usually whole sentences) taken directly from the source text, while abstractive methods may generate novel words and phrases not featured in the source text – as a human-written abstract usually does. The extractive approach is easier, because copying large chunks of text from the source document ensures baseline levels of grammaticality and accuracy. On the other hand, sophisticated abilities that are crucial to high quality summarization, such as paraphrasing, generalization, or the incorporation of real-world knowledge, are possible only in an abstractive framework.

Sequence-to-sequence attentional model

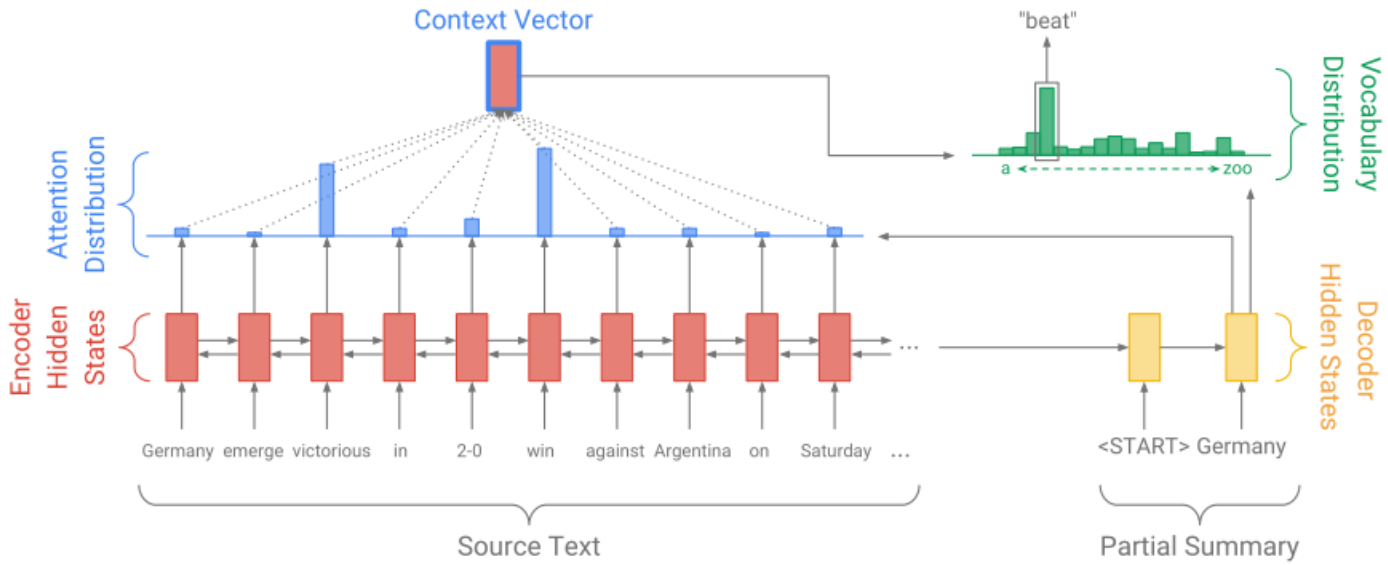

Figure 2: Baseline sequence-to-sequence model with attention. The model may attend to relevant words in the source text to generate novel words, e.g., to produce the novel word *beat* in the abstractive summary *Germany* ***beat*** *Argentina 2-0* the model may attend to the words *victorious* and *win* in the source text.

Pointer-generator network

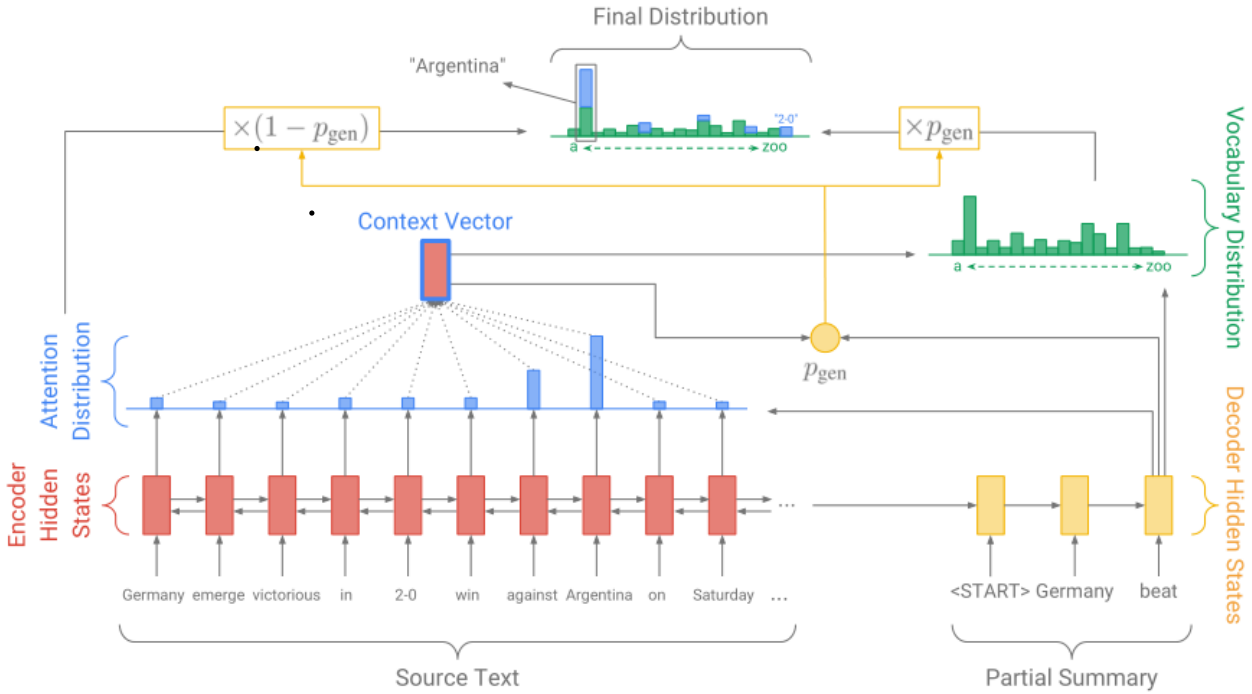

Figure 3: Pointer-generator model. For each decoder timestep a generation probability $p_{\text{gen}} \in [0, 1]$ is calculated, which weights the probability of *generating* words from the vocabulary, versus *copying* words from the source text. The vocabulary distribution and the attention distribution are weighted and summed to obtain the final distribution, from which we make our prediction. Note that out-of-vocabulary article words such as *2-0* are included in the final distribution. Best viewed in color.

Coverage mechanism

$$c^t = \sum_{t'=0}^{t-1} a^{t'} \tag{10}$$

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}}) \tag{11}$$

$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t) \tag{12}$$

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t) \tag{13}$$

以第一时刻为例，c1表示的是a0，那么a0跟a1当中选择最小的作为loss。如果a0和a1关注的都是同样的分布，那么loss就会比较大，如果他们关注的是不同的分布，因为选择的是两者之中最小的那一个，所以这样的loss会比较小。目的就是想让他每一个时刻关注的分布是不一样的，这样可以避免repeat。

Result

	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	**39.53**	**17.28**	**36.38**	17.32	18.72
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

Yixin Liu
Carnegie Mellon University
yixinl2@cs.cmu.edu

Pengfei Liu *
Carnegie Mellon University
pliu3@cs.cmu.edu

Motivation

- Seq2Seq models are usually trained under the framework of Maximum Likelihood Estimation (MLE) and in practice they are commonly trained with the teacher-forcing algorithm. This introduces a gap between the objective function and the evaluation metrics, as the objective function is based on local, token-level predictions while the evaluation metrics would compare the holistic similarity between the gold references and system outputs.

- Furthermore, during the test stage the model needs to generate outputs auto regressivelly, which means the errors made in the previous steps will accumulate. This gap between the training and test has been referred to as the exposure bias in the previous work.

- Specifically, inspired by the recent work of on text summarization, we propose to use a two-stage model for abstractive summarization, where a Seq2Seq model is first trained to generate candidate summaries with MLE loss, and then a parameterized evaluation model is trained to rank the generated candidates with contrastive learning. By optimizing the generation model and evaluation model at separate stages, we are able to train these two modules with supervised learning, bypassing the challenging and intricate optimization process of the RL-based methods.

Contrastive Learning Framework for Abstractive Summarization

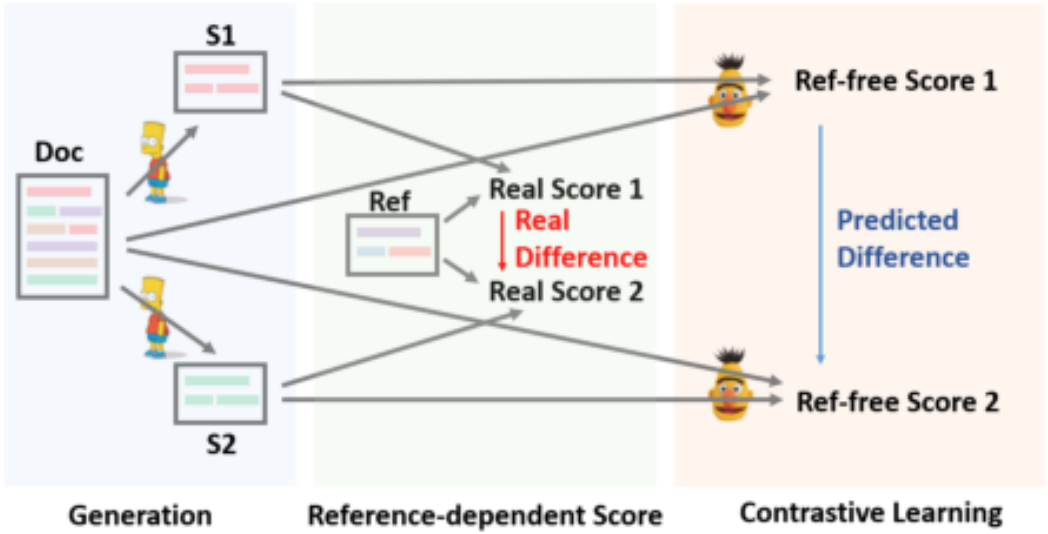

Figure 1: SimCLS framework for two-stage abstractive summarization, where Doc, S, Ref represent the document, generated summary and reference respectively. At the first stage, a Seq2Seq generator (BART) is used to generate candidate summaries. At the second stage, a scoring model (RoBERTa) is used to predict the performance of the candidate summaries based on the source document. The scoring model is trained with contrastive learning, where the training examples are provided by the Seq2Seq model.

Stage I: Candidate Generation The generation model $g(\cdot)$ is a Seq2Seq model trained to maximize the likelihood of reference summary $\hat{S}$ given the source document D. The pre-trained $g(\cdot)$ is then used to produce multiple candidate summaries $S_1, \cdots, S_n$ with a sampling strategy such as Beam Search, where n is the number of sampled candidates.

Stage II: Reference-free Evaluation The high-level idea is that a better candidate summary S_i should obtain a higher quality score w.r.t the source document D. We approach the above idea by contrastive learning and define an *evaluation function* $h(\cdot)$ that aims to assign different scores $r_1, \cdots, r_n$ to the generated candidates solely based on the similarity between the source document and the candidate S_i, i.e., $r_i = h(S_i, D)$. The final output summary S is the candidate with the highest score:

$$S = \underset{S_i}{\operatorname{argmax}}\, h(S_i, D). \tag{1}$$

Here, we instantiate $h(\cdot)$ as a large pre-trained self-attention model, RoBERTa (Liu et al., 2019). It is used to encode S_i and D separately, and the cosine similarity between the encoding of the first tokens is used as the similarity score r_i.

Contrastive Training Instead of explicitly constructing a positive or negative example as most existing work with contrastive learning have adopted (Chen et al., 2020; Wu et al., 2020), here the "*contrastiveness*" is reflect in the diverse qualities of naturally generated summaries evaluated by a parameterized model $h(\cdot)$. Specifically, we introduce a ranking loss to $h(\cdot)$:

$$\begin{aligned} L = & \sum_i \max(0, h(D, \tilde{S}_i) - h(D, \hat{S})) \\ & + \sum_i \sum_{j>i} \max(0, h(D, \tilde{S}_j) - h(D, \tilde{S}_i) + \lambda_{ij}), \end{aligned} \tag{2}$$

where $\tilde{S}_1, \cdots, \tilde{S}_n$ is descendingly sorted by $M(\tilde{S}_i, \hat{S})$. Here, $\lambda_{ij} = (j-i) * \lambda$ is the corresponding margin that we defined following Zhong et al. (2020), and λ is a hyper-parameter.[1] M can be any automated evaluation metrics or human judgments and here we use ROUGE (Lin, 2004).

大体上说，在训练过程中，$h(\cdot)$会学习真实评价指标的排序模式，即真实的评价指标负责提供希望模型学习的排序结果，而模型需要在没有参考摘要的条件下依靠原文档为候选摘要排序。

Results on CNNDM dataset

System	R-1	R-2	R-L	BS	MS
BART*	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	-	-
Origin	44.39	21.21	41.28	64.67	58.67
Min	33.17	11.67	30.77	58.09	55.75
Max	54.36	28.73	50.77	70.77	61.67
Random	43.98	20.06	40.94	64.65	58.60
SimCLS	**46.67**†	22.15†	**43.54**†	**66.14**†	**59.31**†

Table 1: Results on CNNDM. **BS** denotes BERTScore, **MS** denotes MoverScore. **Origin** denotes the original performance of the baseline model. **Min**, **Max**, **Random** are the oracles that select candidates based on their ROUGE scores. †: significantly better than the baseline model (Origin) ($p < 0.01$). *: results reported in the original papers.

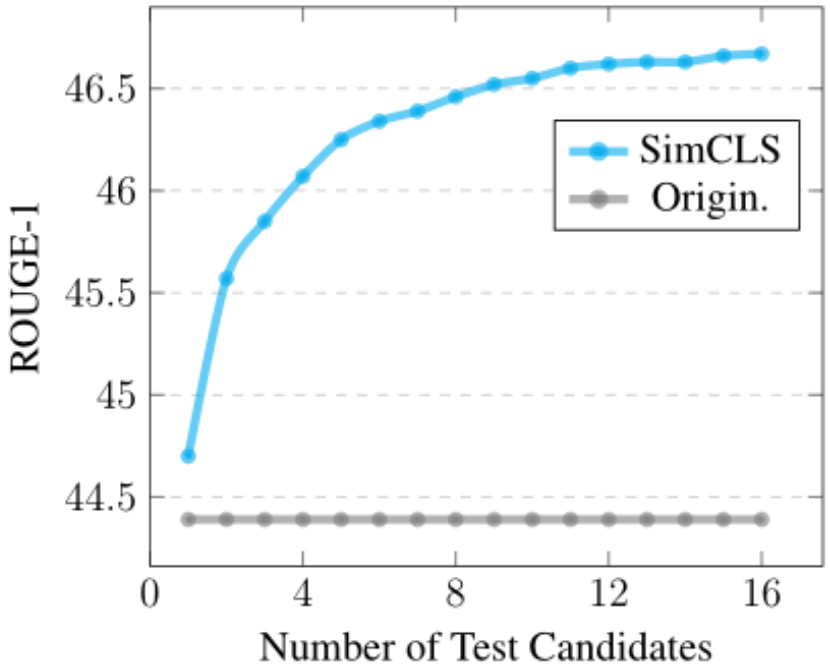

Figure 2: Test performance with different numbers of candidate summaries on CNNDM. **Origin** denotes the original performance of the baseline model.

Fine-grained Analysis

Entity-level

we compare the model performance w.r.t the salient entities, which are entities in source documents that appear in the reference summaries. Specifically, (1) we extract the entities from the source documents (2) select the salient entities based on the entities in reference summaries,

Sentence-level

Sentence Alignments Here we investigate if our method makes sentence-level differences compared to the baseline model. Specifically, (1) we match each sentence in the summaries to a sentence in the source documents based on their similarity(indicated by ROUGE scores), (2) compute the sentence-level similarity between the reference and system-generated summaries based on the overlaps of their matched sentences in the source documents.

Level	System	Precision	Recall	F-Score
Entity	Origin	40.70	59.13	48.22
	SimCLS	**43.36**	**59.79**	**50.27**
Sentence	Origin	38.11	38.65	37.18
	SimCLS	**42.58**	**40.22**	**40.12**

Table 3: Performance analysis on CNNDM dataset. **Origin** denotes the original performance of the baseline model.

首先抽取原文档和参考摘要中共有的实体，然后计算这些实体出现在候选摘要中的比例。将参考摘要和候选摘要中的句子与原文档的句子做语义对齐，然后计算参考摘要和候选摘要对应句子的重合度。

这表明 SimCLS 生成的摘要能够更好地捕捉实体级的语义信息，且在句子层面上与参考摘要的语义更相似。

System	Summary	Article
Ref.	chris ramsey says he has no problem shaking hands with john terry . queens park rangers host chelsea in the premier league on sunday . terry was once banned and fined for racist comments at loftus road . rio ferdinand , brother of anton , will not be fit to play against chelsea .	**queens park rangers manager chris ramsey has revealed he will have no problem shaking john terry's hand in light of the racist comments the former england captain directed at former rs defender anton ferdinand four years ago .** ***terry , who will line up against ramsey's side , was banned for four games and fined # 220,000 for the remarks made in october 2011 during chelsea's 1-0 defeat at loftus road .*** but ramsey , the premier league's only black manager , thinks the issue has been dealt with . ... ' i don't know what his feelings are towards me . as long as there wasn't anything on the field that was unprofessional by him , i would shake his hand . . **queens park rangers manager chris ramsey speaks to the media on friday ahead of the chelsea match .** chelsea captain john terry controls the ball during last weekend's premier league match against stoke . ramsey arrives for friday's pre-match press conference as qpr prepare to host chelsea at loftus road . ' the whole episode for british society sat uncomfortably . it's not something we want to highlight in football . it happened and it's being dealt with . we have to move on . and hopefully everyone has learned something from it . ' . *ramsey revealed that rio ferdinand , who labelled terry an idiot for the abuse aimed at his brother , won't be fit in time for a reunion with the chelsea skipper this weekend .* but the 52-year-old suspects his player's one-time england colleague will be on the receiving end of a hostile welcome from the home fans on his return the scene of the unsavoury incident . ... ferdinand and terry argue during qpr's 1-0 victory against chelsea at loftus road in october 2011 . **rio ferdinand , brother of anton , will not be fit for sunday's match against chelsea .**
SimCLS	queens park rangers host chelsea in the premier league on sunday . qpr boss chris ramsey says he will have no problem shaking john terry's hand . terry was banned for four games and fined # 220,000 for racist comments . rio ferdinand , brother of anton , will not be fit for the match at loftus road .	
Origin.	john terry was banned for four games and fined # 220,000 for the remarks made in october 2011 during chelsea's 1-0 defeat at loftus road . terry will line up against chris ramsey's side on sunday . rio ferdinand , who labelled terry an idiot for the abuse aimed at his brother , won't be fit in time for a reunion with the chelsea skipper this weekend .	

Table 2: Sentence alignments between source articles and summaries on CNNDM dataset. The aligned sentences for reference and our summaries are **bolded** (they are the same in this example). The aligned sentences for baseline summaries are *italicized*. **Origin** denotes the original performance of the baseline model.

展示了一例摘要和原文的句子级匹配结果，可以看到SimCLS对齐的句子和参考摘要更为相近，而baseline关注了不够相关的句子。这里的参考摘要匹配到了原文的最后一句，而SimCLS很好地捕捉到了这一模式。

Positional Bias

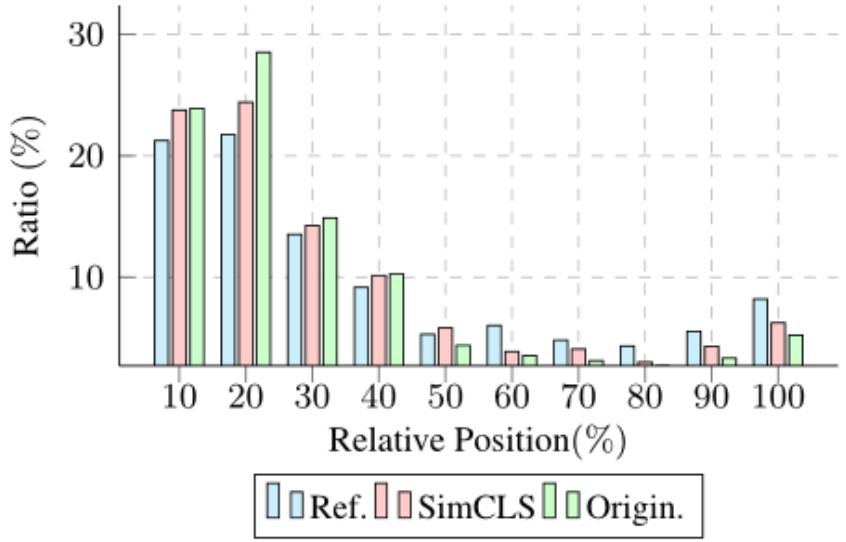

Figure 3: Positional Bias. X-asis: the relative position of the matched sentence in source documents. Y-axis: the ratio of the matched sentences. For fair comparison, articles are first truncated to the generator's maximum input length. **Origin** denotes the original performance of the baseline model.

进一步地，作者发现生成式摘要在处理长文档(30句以上)时存在**位置偏差(position bias)**，如下图所示，可以发现baseline会倾向于关注靠前的句子，这可能是由Seq2Seq模型自回归的生成方式导致的，而SimCLS能在一定程度上缓解位置偏差，这得益于diverse beam search和评价模型的引入。

本文的出发点是希望解决训练和测试的不一致的问题，这个问题可以分为两个方面，一个是自回归式的MLE本身存在的曝光偏差问题，另一个是目标函数和评价指标的不一致问题，而本文主要致力于解决后一个问题。本文的思路并不复杂，就是利用对比学习训练了一个能够在没有参考摘要的条件下打分的评价模型，该评价模型选择出的摘要在真实的评价指标上的表现比随机选择更好。这一思路其实可以推广到所有目标函数和评价指标不一致的场景下。