HonestBait: Forward References for Attractive but Faithful Headline Generation

Chih-Yao Chen* UNC Chapel Hill cychen@cs.unc.edu

Dennis Wu* Northwestern University hibb@u.northwestern.edu

Lun-Wei Ku Academia Sinica lwku@iis.sinica.edu.tw

Abstract

Current methods for generating attractive headlines often learn directly from data, which bases attractiveness on the number of user clicks and views. Although clicks or views do reflect user interest, they can fail to reveal how much interest is raised by the writing style and how much is due to the event or topic itself. Also, such approaches can lead to harmful inventions by over-exaggerating the content, aggravating the spread of false information. In this work, we propose HonestBait, a novel framework for solving these issues from another aspect: generating headlines using forward references (FRs), a writing technique often used for clickbait. A self-verification process is included during training to avoid spurious inventions. We begin with a preliminary user study to understand how FRs affect user interest, after which we present PANCO[1], an innovative dataset containing pairs of fake news with verified news for attractive but faithful news headline generation. Automatic metrics and human evaluations show that our framework yields more attractive results (+11.25% compared to human-written verified news headlines) while maintaining high veracity, which helps promote real information to fight against fake news.

1 Introduction

Fake news has become a medium by which to spread misinformation (Oshikawa et al., 2020; Vicario et al., 2019). One common way to fight against fake news is to release verified news.[2] However, as the goal of news verification is to correct misinformation, verified news headlines are often bland, making it difficult to gain the attention of users, which works against the need to alleviate the harmful impact of fake news. Therefore, headlines for verified news articles should be rewritten to be more intriguing but still faithful, which is expected to pique reader interest in verified news. Many studies have been conducted on generating attractive headlines (Jin et al., 2020; Xu et al., 2019), among which *clickbait* represents the style that generates the most reads or clicks. Despite their success in attracting readers, there are several challenges in current models. First, clickbait datasets for training headline generators with sensational style transfer are commonly collected based on the amount of views or clicks, which assumes that headline popularity is always due to the writing style (Song et al., 2020). However, user reading preferences could also be motivated by trending topics or major events. For instance, "*Flights cancelled as typhoon nears*" was the most popular news on a day that a typhoon was coming. Although such headlines get many views and clicks, the writing style itself is not interesting, and could end up as noise in the dataset. Second, harmful "hallucinations" created by headlines exaggerated to be more sensational could distort the meaning of the original article. This is especially critical as we do not want our model itself to spread misinformation. However, as such sensational headline generation models often generate clickbait with more ambiguous words, it increases the difficulty of evaluating faithfulness by aligning title semantics with the news content.

In this work, we propose making real news intriguing by learning what fake news is good at. We seek to learn what makes fake news eye-catching instead of simply mimicking the titles of fake news. Quantity-wise, the many circulating fake news articles serve as learning materials by which we can learn to generate more attractive headlines; style-wise, fake news is deliberately written to attract attention. To learn such attractive writing styles, we adopt the forward-reference (FR) writing technique (Blom and Hansen, 2015), which draws from

* Equal contribution.

[1] Data is publicly available at: https://github.com/dinobby/HonestBait

[2] In this work, we define "verified news" as news written specifically to clarify false information; the term "real news" is defined as general news that does not contain misinformation.

Findings of the Association for Computational Linguistics: ACL 2023, pages 4810–4824

psychology and journalism, and is frequently used to create attractive headlines. Specifically, FR creates an information gap between readers and the news content with the headline, motivating the reader's curiosity (Loewenstein, 1994) to investigate the news content, and hence provoking the desire to click on the headline. One example is the headline *"Wanna be an enviable couple? 12 things a happy couple must do... It's that simple!"*, which drives readers to find out what those things are.

Here, to understand the relation between veracity, attractiveness, and FR types in news headlines, we conducted a preliminary user study to investigate the attractiveness of fake and real news, and analyzed the FR types used in headlines in terms of veracity. Given these results and observations, we propose HonestBait, a novel framework by which to generate attractive but faithful headlines. In this framework, we use FR to remove the need to learn directly from the click-based dataset. To ensure the faithfulness of the generated headlines, we design a lexical-bias-robust textual entailment component on the generated headline and its original content to confirm that the content infers the headline. In addition, we propose PANCO, an innovative dataset which consists of pairs of fake and verified news headlines, their content, and their FR types. We conduct experiments on PANCO and evaluate the results in terms of both automatic metrics and human evaluation. In sum, the contributions of our work are threefold:

- We conduct a thorough user study to understand the relation between reading preferences and FR types on fake news and verified news.
- We propose a novel framework for generating attractive but faithful headlines. In human evaluations, HonestBait largely outperforms baselines on attractiveness and faithfulness.
- We propose a new dataset containing pairs of fake and verified news, including their headlines, content, and FR types in headlines.

2 Related Work

2.1 Forward Referencing as a Lure

Loewenstein (1994) shows how the desire for information motivates human curiosity. Forward-referencing has been defined as a technique for creating curiosity gaps at a discourse level for use in headlines (Blom and Hansen, 2015; Yang, 2011). A similar concept is cataphora, in which information is forwarded as a teaser at the sentence level (Baicchi, 2004; Halliday and Hasan, 1976). Kuiken et al. (2017) investigate how editors rewrite headlines for digital platforms, and analyze the linguistic features of what makes for an attractive headline. Zhang et al. (2018) address attractive headline generation as question headline generation (QHG), which assumes that interrogative sentences are more popular. Although this modality is indeed a type of FR, we argue that the interrogative style may not be suitable for all kinds of headlines, especially verified news. Hence in our work, we fully consider all kinds of FR which are commonly used and seen in social media and on digital platforms. Sample headlines exhibiting FR techniques can be found in Fig. 4 in the appendix.

2.2 Headline Generation

Headline generation can be viewed as a more specific summarization task. Qi et al. (2020) propose a Transformer-based, self-supervised n-gram prediction objective. Liu (2019) propose BERTSum, a variation of BERT (Devlin et al., 2019) for extractive summarization. See et al. (2017) propose an attention-based pointer generator with a copy mechanism, which has made great progress in summarization. Although its ability to copy text from the source context is powerful, using it directly for verified news often leads to bland titles. Hence we apply FRs and a sensationalism scorer to produce more satisfying results. Xu et al. (2019) propose auto-tuned reinforcement learning to generate sensational headlines using a pretrained sensationalism scorer; the resulting score is used as the reward to enhance the attractiveness. Although generating attractive headlines has been widely explored (Song et al., 2020; Jin et al., 2020), we focus more on fidelity to ensure that the semantics of the generated headline are faithful to the source content to avoid harmful hallucination.

2.3 Faithful Summarization

Recent work investigates how to improve the faithfulness of the generated summary or headline. Matsumaru et al. (2020) propose pretraining a textual entailment scorer to filter out noisy samples in the dataset, preventing hallucination or unfaithful generation. Maynez et al. (2020) analyze the faithfulness of current abstractive summarization systems, and discover that textual entailment is better correlated to faithfulness than standard metrics. Based

on such work, one major direction is to evaluate generated summaries in terms of textual entailment rather than raw metrics such as ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002). Accordingly, we propose a faithfulness scorer based on textual entailment to evaluate how well the generated headlines fit the semantics of the content.

3 Preliminary User Study

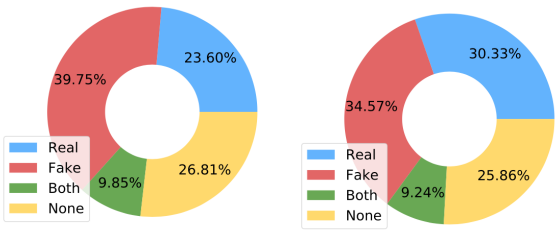
In this section, we investigate for a given topic which of the fake or real headlines users are more interested in, and how often forward references are found in interesting titles. Accordingly, we seek to test the following two hypotheses:

H1: *Fake news headlines motivate user reading interest more than real news headlines.*

H2: *Forward references are commonly seen/used in headlines which interest users.*

We conducted the user study on both Chinese and English news to determine whether forward references were used across languages. For English headlines, we adopted FakeNewsNet (Shu et al., 2018), which contains fake and real news headlines about gossip and political news from GossipCop and PolitiFact. Since the real and fake news in FakeNewsNet are not paired up, we performed topical clustering to alleviate topical bias. For Chinese headlines, we directly leveraged news pairs labeled as *disagreed* in the WSDM fake news challenge dataset,[3] which contains one fake news headline and its corresponding verified news headline.

We conducted the English user study using Amazon Mechanical Turk (Crowston, 2012). Each pair was labeled by three turkers, whereas each Chinese pair was annotated by five native speakers we recruited. To test H1, annotators chose which headline they wanted to read further, with four options: *first headline*, *second headline*, *both*, and *none*. News veracity was not revealed during the study. Results show that both Chinese and English readers prefer fake news headlines. For Chinese headlines, 39.75% of fake titles were judged to be more interesting than the real ones, whereas only 23.60% of real titles won. For English headlines, the percentages are 34.57% and 30.33%, respectively. Note that in English, we are comparing real news with fake news due to the scarcity of paired verified and fake news data, whereas in Chinese, we are comparing verified news with fake news. This could be why the preference for real and fake news in English is closer than in Chinese. Even so, both Chinese and English show with statistical significance (p-values far less than 0.05) that readers prefer fake headlines. We report the complete distribution including ties, as shown in Fig. 1. This result supports H1: fake news headlines motivate reading interest more than real news headlines.



(a) Chinese headlines (b) English headlines

Figure 1: Reading preferences w.r.t. real news and fake news including ties. The sample size is 8,424 / 6,497 for Chinese / English headlines.

To test H2, we randomly sampled 1,000 preferred and rejected headlines, respectively, from the previous user study, and asked another set of three annotators to label the FR type. Results show that 73.48% of Chinese and 85.32% of English preferred headlines utilizing FR techniques (at least one FR included in the headline), whereas in rejected headlines, the ratio is 22.35% / 17.72%. This further supports H2: FR is commonly used in interesting headlines. In conclusion, we found that fake news headlines draw more reader interest, and the use of FR techniques is a key part of what makes headlines intriguing.

4 Methodology

Having motivated the use of FR, we propose HonestBait, a novel framework which incorporates FR techniques and veracity verification. HonestBait consists of two stages. In the first stage, we pretrain an FR predictor and an FR proposer (§ 4.1). Both of them take verified news titles as input. The FR predictor is trained to predict which FRs a verified headline contains; hence the gold label is the FRs of the current input verified headline. The FR proposer, in turn, learns to predict which combination of FRs the corresponding fake news exhibits; the gold label is the FRs of the corresponding fake headline of the current input verified headline. The main concept in stage 1 is learning FRs from fake news to provide the direction best suited to rewrit-

[3] https://www.kaggle.com/c/fake-news-pair-classification-challenge

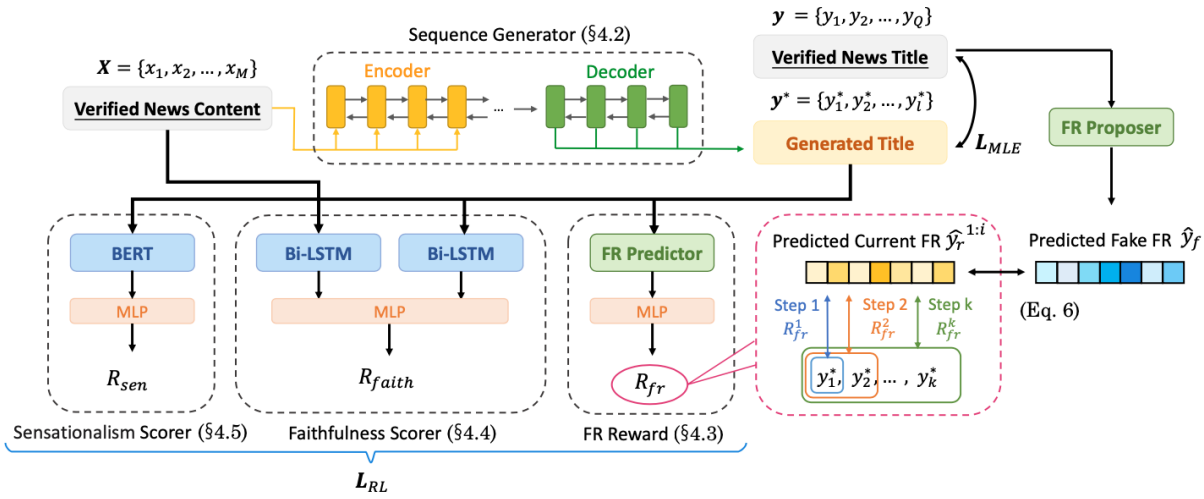


Figure 2: Stage 2: overall architecture of HonestBait. Required inputs are underlined.

ing a monotonic verified headline into an interesting headline.

When the FR predictor and FR proposer are ready, we proceed to stage 2 to generate attractive but faithful headlines. Figure 2 depicts the overall architecture of the second stage. The input during stage 2 consists only of verified news headlines and their content for learning headline generation, where the developed FR predictor and FR proposer together provide rewards to the learning model. First, we use a sequence generator (§ 4.2) to generate headlines from the input verified news content, and utilize the FR proposer to predict which combination of FR types is best suited to rewriting the input verified news headline. During each decoding step, we use the FR predictor to predict which FR types the currently generated headline contains, and we align the prediction from the FR proposer and the FR predictor to transform the original boring verified headlines into exciting ones. This is achieved by computing the FR type reward (§ 4.3). After decoding, we make use of a faithfulness scorer (§ 4.4) and a sensationalism scorer (§ 4.5) to compute the faithfulness and sensationalism rewards by which to evaluate the generated headline; all three rewards are then combined to make the generated results attractive but faithful. During inference, given verified news headlines and their content, HonestBait then generates attractive but faithful headlines using the above-mentioned components. Below we describe each major component in detail.

4.1 FR Predictor & FR Proposer

To mimic different FR types on datasets without FR type labels, we pretrain two multi-label classifiers: (1) A FR predictor, which predicts which FR type the generated headline contains; this is pretrained by taking verified news headlines as input and classifying which FR type these headlines exhibit. (2) A FR proposer, which learns what specific combination of FRs is best suited to rewriting a given verified title. This is trained by taking the verified headline as input and predicting the FR type of the corresponding fake news. Note that this setting is achievable because we have paired news data with both real and fake FR labels (see preview sample in Fig. 3 in the appendix).

We implement these FR classifiers with a BERT-based encoder. Given a verified news headline, we obtain a sentence-level representation h_p with the hidden state of the [CLS] token. The FR type $\hat{y_{fr}}$ is predicted by a MLP classifier following a sigmoid function and a softmax operation: $\hat{y_{fr}} = \text{softmax}(\sigma(W_p h_p + b_p))$, where $\hat{y_{fr}} \in \{0,1\}^l$, l is the number of the FR type, and W_p, b_p are trainable parameters. We pretrain these models using binary cross entropy loss, yielding a 0.91 micro-F1 score for the FR predictor and 0.65 for the FR proposer on a pretraining test set. Below we denote the FR predictor's prediction as $\hat{y}_r$ and the FR proposer's prediction as $\hat{y}_f$.

Predicting the fake version of FR types from the verified news headline is more challenging, as the performance of the FR proposer is lower than of the FR predictor (0.65 vs. 0.91). In practice, we could directly use the FR label of the fake news ac-

quired from our user study to replace $\hat{y}_f$, and view this setting as an upper bound for the FR proposer accuracy to calculate $\mathcal{R}_{fr}$. However, when we are not provided with the FR labels of fake titles, we do not know which FR technique(s) should be applied to rewrite the given verified news headline. Hence, the FR proposer can be used as an auxiliary tool to help decide which FR type to use; this is especially useful when the dataset contains no FR-type labels. After pre-training the FR predictor and proposer, we proceed to the second stage.

4.2 Sequence Generator

In the second stage, we adopt a pointer network (See et al., 2017) as the sequence generator because of its ability to copy words from the source text. Given verified news content with M tokens $X = \{x_1, x_2, \ldots, x_M\}$ and its corresponding real headline consisting of Q tokens $y = \{y_1, y_2, \ldots, y_Q\}$, the encoder encodes each token with a bidirectional LSTM (Hochreiter and Schmidhuber, 1997). We adopt Chinese word-level embeddings pretrained on the Weibo corpus (Li et al., 2018). The final distribution is combined with the probability computed by the copy mechanism, making words from source content available for generation. For the objective we use the negative log likelihood as

$$\mathcal{L}_{MLE} = -\frac{1}{T}\sum_{i}^{T} \log P_{final}(y_i). \qquad (1)$$

4.3 Forward Reference Reward

For each decoding time step, we calculate the FR reward once: tokens generated up to the current time step $y^*_{1:t}$ are sent to the FR predictor to derive $\hat{y}_r^{1:t}$ and calculate how well the generated text fits the FR prediction using the FR Proposer $\hat{y}_f$. After T steps of decoding, and after the headline is generated, we calculate the average FR reward as

$$\mathcal{R}_{fr} = \frac{1}{T}\sum_{i}^{T}(1 - \mathcal{D}(\hat{y}_f, \hat{y}_r^{1:i})), \qquad (2)$$

where $\mathcal{D}$ denotes a distance function—in our case the mean squared error—and $\mathcal{R}_{fr} \in [0, 1]$ is the average FR reward. Here $\hat{y}_r$ is the FR types exhibited by the current generated headline, which should align with the prediction from the FR proposer $\hat{y}_f$, which is pretrained to learn which specific combination of FRs are best suited to rewrite the given title. The closer they get, the higher $\mathcal{R}_{fr}$ is.

4.4 Faithfulness Scorer

Inspired by research which shows that textual entailment correlates better with faithfulness than raw metrics (Falke et al., 2019), we use a pretrained faithfulness scorer to evaluate whether the generated headline distorts or contradicts the corresponding content. When pretraining, we use a verified news headline and its content as a positive example, and use a fake news headline with the corresponding real news content as a negative example. We pretrain this as a natural language inference (NLI) task (classifying entailment and contradiction). The headline and content sentence embeddings of are denoted as x_f and w_f. We apply a popular method to encode sentences for the NLI model (Conneau et al., 2017):

$$h = [x_f; w_f; x_f - w_f; x_f \odot w_f], \qquad (3)$$

where ";" denotes concatenation, and "$\odot$" denotes the element-wise product. The faithfulness scorer achieves an accuracy of 0.83 on the testing set.

4.5 Sensationalism Scorer

Apart from the FR type reward, we make use of another BERT-based binary classifier to obtain the sensationalism score, since there are headlines that are still interesting without the use of FRs (around 27% according to our collected data). We first manually reviewed 100 news for each categories in seven different new sources, selected the news categories that were consistently sensational (more than two-thirds of the articles in such a category were sensational, e.g., fashion, gossip, headlines) and collected the news headlines along with the content in these categories. We reviewed 5,000 headlines in total and collected 50,000 sensational news headlines. For non-sensational headlines, we utilized a pointer generator to obtain a summary headline, and treated this as a non-sensational title since summarization models retain only the semantics of the content. In this way we ensured a 50/50 split for sensational and non-sensational headlines for training. We trained the sensation scorer using binary cross entropy along with a softmax layer to produce a sensationalism score $\in [0, 1]$: $\mathcal{R}_{sen} = \sigma(W_s h_s + b_s)$, where h_s is the aggregated representation of the [CLS] token produced by BERT, and W_s and b_s are learnable weights. The accuracy on the test set is 0.86, indicating its ability to discriminate sensational headlines.

4.6 Hybrid Training

We adopted reinforcement learning (RL) (Williams, 1992) to train our model with the weighted sum of $\mathcal{R}_{fr}$, $\mathcal{R}_{faith}$ and $\mathcal{R}_{sen}$ as the reward $\mathcal{R}$. Following Xu et al. (2019); Ranzato et al. (2015), we used the baseline reward $\hat{\mathcal{R}}_t$ to reduce variance, where $\hat{\mathcal{R}}_t$ is the mean reward estimated by a linear layer for each time step t during training. The final reward and the objective are

$$\begin{aligned} \mathcal{R} &= \mathcal{R}_{fr} + \alpha\mathcal{R}_{faith} + (1-\alpha)\mathcal{R}_{sen} \\ \mathcal{L}_{RL} &= -\frac{1}{T}\sum_{i}^{T}(\mathcal{R} - \hat{\mathcal{R}}_t)\log P_{final}(y_t). \end{aligned} \tag{4}$$

Similar to Xu et al. (2019), we computed the final loss as the combination of $\mathcal{L}_{MLE}$ and $\mathcal{L}_{RL}$:

$$\mathcal{L} = \lambda\mathcal{L}_{MLE} + (1-\lambda)\mathcal{L}_{RL}, \tag{5}$$

where α and $\lambda \in [0, 1]$ are hyperparameters that balance the weight of each component; the composite design here ensures that we produce headlines that satisfy all objectives. In sum, we use the FR reward to estimate whether the generated headline matches the FR type of its fake version, the faithfulness scorer to evaluate the textual entailment between the generated headline and the verified news content, and the sensationalism scorer to measure the sensationalism of the generated headline.

5 Experiment

In this section, we describe experiments conducted to evaluate HonestBait. We first describe the experimental dataset and then describe the result of human evaluation, automatic metrics, a case study, and hyperparameter analyses to further demonstrate the superiority of the proposed model.

5.1 PANCO Dataset

We collected **Pa**ired **N**ews with **Co**ntent (PANCO), a subset of a fake news classification competition held by WSDM. The competition involved a textual entailment task in which two news headlines were given as input: the task was to predict the relationship between the headlines. Each sample in the original dataset included a fake news headline and a headline that was either *agreed* (two fake stories describing the same event), *unrelated* (two stories describing different events), or *disagreed* (two stories describing the same event, one of which was fake and the other was verified). We selected the *disagreed* pairs that contain a fake headline and its corresponding verified news headline, and augmented the dataset in the following way: (1) We used each title as a query which we submitted to Google Search to determine the source of each news story, and crawled the news content from sources which matched the title. (2) Five annotators labeled the FR type of each headline; the final label was decided by majority vote.

The proposed dataset consists of a total of 7,930 paired samples containing a fake news headline and the corresponding verified news headline along with their content and FR type. To better understand the dataset, we provide a preview sample in Fig. 3 in the appendix. The main novelty of PANCO is the collection of pairs (describing the same event) of fake and verified news with headlines and their content. In addition, we provide the FR type label for both verified and fake news as additional text features for further study. We provide a previewing sample from PANCO in Table 3.

5.2 Baseline and Settings

We compared the proposed model with the following strong baseline for headline generation. **Ptr-G** for pointer generator network (See et al., 2017), an LSTM-based model with attention and a copy mechanism. **Clickbait** (Xu et al., 2019), which uses a CNN-based sensationalism scorer to automatically balance MLE and reward loss, and also used as a reward to generate more sensational headlines. **ROUGE**, which uses the same architecture as Clickbait but with the ROUGE score as a reward. **BERTSum** (Liu, 2019), which utilizes BERT's architecture to encode source text and perform extractive summarization. **T5** (Raffel et al., 2020), a large Transformer-based model; we utilize T5 with PEGASUS (Zhang et al., 2020) pretraining to strengthen the baseline. **ProphetNet** (Qi et al., 2020, 2021), a Transformer-based model that utilizes future n-gram prediction as a self-supervision.

For human evaluation and the case study, we also include Gold, which represents human-written verified headlines as a strong baseline. Experimental settings are detailed as follows. We first pretrained all baselines on the LCSTS dataset (Hu et al., 2015) with 480,000 steps. LCSTS is a large-scale Chinese summarization dataset containing 2,400,591 samples with paired short text and summaries. We used the pretrained weights to fine-tune all baselines on the PANCO training set for another 20,000 steps.

Real headline	**Real content**	FR
辟谣：夏季曝晒后的瓶装水致癌 Rumor has it that bottled water exposed to the summer sun is carcinogenic.	夏天放在车里的瓶装水会致癌? 又是一条健康谣言... Bottled water is carcinogenic in the car when it's summer? Another health rumor...	No FR.
Fake Headline	**Fake Content**	FR
BBC紧急曝光: 这种水喝一口，就会致癌！ BBC urgent disclosure: Just one sip of this water can result in cancer!	今天，这个有害健康、甚至会夺人性命的巨大隐患终于被曝光了！世界卫生组织通报：9成以上瓶装水有毒... This enormous concern, which is dangerous and even fatal, has now been made public. WHO reports that over 90% of water in bottles is tainted...	FR1+FR5

Figure 3: A previewing sample from PANCO dataset, comprised of paired real and fake news headlines, the news content, and FR type labels for both real and fake headlines.

Reference	Verified news headline: The truth of using one drop of blood to test cancer. Fake news headline: Testing cancer using only one drop of blood! This is amazing.
News content	"The woman version of Jobs" Elizabeth Holmes became popular by proposing a revolutionary technique: using a single drop of blood to test cancer. But not for long: her lies were revealed, and she fell from favor. An expert said that liquid biopsies in clinics cannot yet be consisted the gold standard, and cannot completely replace tissue biopsy.
Ptr-G	Rumors! Jobs, really tells you the truth!
Clickbait	"Rumors" Jobs can heal the reagent box? Here's the truth!
Clickbait+ROUGE	Rumors! Can Jobs make people test for cancer?
BERTSum	A drop of blood can detect cancer?
T5	A drop of blood can detect cancer is a rumor, how can we do to prevent cancer?
ProphetNet	Clarification: Blood test cannot determine cancer.
HonestBait	A drop of blood can detect cancer? Experts clarified: it's a scam!

Table 1: Generated examples from different models. For brevity, we show part of the article and translated result.

We saved the checkpoints for all baselines every 2,000 steps, and compared them by selecting the best one on the validation set. The hyperparameters of HonestBait were also based on the validation results: $\lambda = 0.2$ and $\alpha = 0.4$.

Model	R_1	R_2	R_L	BS	FR
Ptr-G	41.86	28.18	37.30	69.61	55.32
Clickbait	41.02	28.03	36.64	69.52	69.11
ROUGE	43.75	27.65	35.65	71.56	58.91
ProphetNet	**46.82**	30.40	38.89	**73.57**	49.77
BertSum	28.09	16.15	18.86	63.22	16.83
T5	44.27	28.55	38.66	72.73	59.96
HonestBait	43.76	**31.45**	**40.42**	72.61	**80.42**

Table 2: Automatic metrics of proposed model against baselines. R_n is the n-gram ROUGE score, R_L is the ROUGE-L score, BS is the BERT score, and FR is the ratio of the generated headlines using FR.

5.3 Human Evaluation

We first conducted a human evaluation to evaluate the attractiveness, faithfulness, and fluency of the generated headlines. We randomly selected 100 samples from the PANCO test data, and asked five native speakers to select headlines in response to the following questions: (1) which headline makes you want to read further? (2) which headline is more faithful to the content? (3) which headline is more fluent?

Model	ATRC	FAITH	FLCY
Ptr-G	-29.50%	-17.83%	-19.80%
Clickbait	-6.00%	-22.33%	-9.25%
ROUGE	-17.50%	-17.25%	-24.66%
BertSum	-30.50%	-21.99%	-9.70%
T5	-12.50%	-10.25%	-1.25%
ProphetNet	-5.60%	-5.50%	4.33%
Gold (human)	-11.25%	1.00%	8.34%
HonestBait	-	-	-

Table 3: Pairwise comparison in terms of attractiveness (ATRC), faithfulness (FAITH), and fluency (FLCY), shown as percentages. The larger the negative value, the more HonestBait outperforms.

The workers were given two generated titles and the story content, and were asked to select *first title*, *second title*, or *tie* in response to the questions.

Table 3 reports the pairwise comparison results as percentages. Each number in the table is the competing model compared to the proposed HonestBait, following Zhao et al. (2020). For example, the output of Ptr-G is 12.50%/45.50%/42.00% better/same/worse than HonestBait in terms of attractiveness, resulting in $12.50\% - 42.00\% = -29.50\%$ in the table. Results show that for both attractiveness and faithfulness, HonestBait outperforms all baselines by a large margin. We believe this is due to the use of forward referencing and the faithfulness check. Compared to the pure click-driven attractiveness-optimized Clickbait (Xu et al., 2019), HonestBait outperforms by directly learning writing skills to avoid other impact factors of attractiveness. In addition, boosting only attractiveness makes Clickbait relatively unfaithful (-22.33%). In terms of fluency, only ProphetNet and human-written headlines outperform our model. As we did nothing specifically to improve fluency such as ProphetNet's n-stream attention, this result indicates that HonestBait maintains reasonable fluency while increasing attractiveness and faithfulness. Note that compared to human-generated real headlines, HonestBait generates more attractive headlines (+11.25%) with only a modest drop in faithfulness (-1.00%). These results show the effectiveness of HonestBait for rewriting real news headlines to promote stories, as it maintains high faithfulness while being more attractive.

5.4 Automatic Metrics

We used three automatic metrics for evaluation: ROUGE-n (Lin, 2004), ROUGE-L, and the BERT score (Zhang* et al., 2020). Although in general, automatic metrics are shown to be not reliable for text generation (Sulem et al., 2018; Callison-Burch et al., 2006; Schluter, 2017; Wang et al., 2018), we still provide them here for reference. The results in Table 2 still show the good abstractive ability of HonestBait with the highest 40.42 R_L score. Among the baselines, ProphetNet is the strongest, with the highest R_1 and BERT scores, perhaps due to its n-stream self-attention mechanism. However, the extractive summarization model BERTSum performs worst here, as extracting a sentence from the article as its headline is not a common practice in general. In the last column of Table 2, we further use the FR predictor to detect which FR technique(s) the generated headlines are using, and report the percentage of generated headlines that use FR. The result shows that 80.42% of the headlines generated by HonestBait exploit FRs to make headlines more attractive, which is the highest among all models, indicating that HonestBait indeed learns to utilize FR techniques.

5.5 Ablation Study

To further investigate our framework, we conducted an ablation study. We compared each setting with the full framework using the evaluation protocol from § 5.3 by pairwise comparison, along with the automatic metrics for completeness. The results are shown in Table 4. Clearly, there is a significant drop in attractiveness when we remove the sensation scorer (-19.50%) or FR type reward (-16.00%), which indicates that even with the sensation scorer, attractiveness still decreases without the help of the FR reward (see setting *w/o FR*). That is, the FR reward indeed helps the model to learn attractive writing styles. In addition, removing the faithfulness scorer results in the largest decrease in faithfulness (-11.50%). This also shows that our faithfulness scorer prevents deviations in the generated headline. Interestingly, removing the sensation scorer increases the ROUGE score, perhaps because the sensation scorer helps to generate more diverse and interesting headlines, and thus can harm metrics which are based on word-level overlap. We also observe that removing the faithfulness scorer reduces the ROUGE score, which shows that the faithfulness scorer helps to produce headlines with more fidelity, and thus increases the word-level overlap between the generated headlines and the ground-truth. Note that as automatic metrics are still not the most important indicator of generation quality, thus we still keep sensation scorer for its improvements in terms of attractiveness and fluency even if removing the it leads to a higher ROUGE score.

	ATRC	FAITH	FLCY	R_2
W/o sen	-19.50%	-4.00%	-9.75%	**32.01**
W/o faith	-4.00%	-11.50%	-6.75%	28.81
W/o FR	-16.00%	- 5.50%	-6.25%	30.92
Full	-	-	-	31.45

Table 4: Ablation study result.

5.6 Case Study

Table 1 shows an example illustrating headlines generated by different models. Results show that Ptr-G, Clickbait, and ROUGE extract the name "Jobs" from the article (highlighted in yellow),

which is a powerful ability of the copy mechanism to alleviate the generation of unknown tokens. However, in terms of being headlines, these texts are less satisfying in that they are not understandable. BERTSum and T5 make mistakes by generating open questions without answering them, which could motivate user interest but is not faithful enough for verified news headlines. Even more, T5 focuses on the wrong point borrowed from other articles as this article is not about cancer prevention, which could be harmful (highlighted in pink). In contrast, HonestBait generates interrogative sentences to attract readers, but with an explicit clarification of the fake information, and is aligned to the content (highlighted in green).

6 Conclusion

We present HonestBait, a novel framework for generating faithful but interesting headlines from a new aspect: forward references. Moreover, we construct PANCO, a novel dataset that includes the title and content of pairs of fake and verified news, along with their forward reference types for further research. Our user study shows that verified news headlines are relatively boring, and forward references are used in most headlines liked by readers. Experimental results show that HonestBait outperforms all baselines in both automatic and human evaluations, which demonstrates its effectiveness in generating attractive but faithful headlines. We expect HonestBait to help rewrite monotonous real news headlines to increase their exposure rate to help combat fake news.

Limitations

Although HonestBait shows promising results for generating attractive but faithful headlines, there are still some limitations: (1) HonestBait is a monolingual model that only supports Chinese. It requires three pre-trained scorers. Also, as the FR labels are specifically difficult to obtain, it is not easy to implement in other languages. (2) Running the whole framework with a batch size of 16 takes around 22 GB GPU memory, mostly because we must load all pre-trained models into the GPU. This can be alleviated by using a distilled pre-trained model. (3) On average, HonestBait generates more faithful headlines than other baselines, but it still occasionally produces false information or unwanted results. This work is only for academic purposes and is not ready for production.

Ethics Statement

Given that our dataset is in Chinese and requires a profound understanding of forward referencing for annotation and evaluation, we carefully selected annotators from our lab who specialize in NLP-related research and possess knowledge in linguistics. To ensure fairness, we provided all annotators with a payment of $6.66 per hour, which is 10% higher than the minimum hourly wage requirement in Taiwan.

During the data annotation process, we introduced the concept of forward referencing to the annotators, along with relevant examples. Only annotators who achieved an accuracy rate of over 80% were eligible to perform the actual annotation task. It's important to note that we solely asked annotators to label the "type of forward reference," which is well-defined, and not to assess the accuracy or truthfulness of the news articles. With five annotators who successfully passed the pretest, combined with the relatively objective nature of labeling forward reference types, we believe any potential bias during the data annotation process is minimal.

For the evaluation phase, an additional five annotators were tasked with determining the preferable title among two options, based on attractiveness, faithfulness, and fluency. These annotators are different from those who labeled the data to ensure a blind test. Although this task involves a greater level of subjectivity, we provided average statistics based on the assessments of the five annotators. Additionally, we maintained a blind test by recruiting separate evaluators and randomly shuffling the order of the two titles for each trial. This evaluation protocol aligns with standard practices employed in the research community, and we believe it effectively minimizes potential biases.

It is also important to note that we are not really *learning to mimic fake news*, by taking fake news headlines as the ground truth reference. Instead, we seek to learn the *writing techniques* that are often used in fake news to attract readers. As we are aware of the risk of producing misinformation, we want to again highlight the importance of the faithfulness check. HonestBait was designed only to assist journalists as a reference to write faithful headlines that users prefer for verified news. Even if we propose using a faithfulness scorer to increase fidelity, its nature, similar to attractive headline generation systems, still exhibits the risk

that HonestBait could be used by malicious users to generate sensational headlines for fake news. Additionally, HonestBait may misjudge offensive or unethical headlines to be a headline that users would prefer. Our goal is to fight fire with fire by leveraging fake news as learning material to fight against misinformation, by encouraging users to read verified news. We call on users not to abuse HonestBait to produce false information.

Acknowledgement

This work is supported by the National Science and Technology Council of Taiwan under grants 111-2221-E-001-021 and 111-2634-F-002-022.

References

Annalisa Baicchi. 2004. *The Cataphoric Indexicality of Titles*, pages 17–38.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, Online. Association for Computational Linguistics.

Jeffrey Kuiken, Anne Schuth, Martijn Spitters, and Maarten Marx. 2017. Effective headlines of newspaper articles in a digital environment. *Digital Journalism*, 5(10):1300–1314.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. 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.

Yang Liu. 2019. Fine-tune bert for extractive summarization.

George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116:75–98.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *ACL*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. pages 1906–1919.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, et al. 2021. Prophetnet-x: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation. *arXiv preprint arXiv:2104.08006*.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Yun Zhu Song, Hong Han Shuai, Sung Lin Yeh, Yi Lun Wu, Lun Wei Ku, and Wen Chih Peng. 2020. Attractive or faithful? popularity-reinforced learning for inspired headline generation. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, pages 8910–8917. AAAI press. Publisher Copyright: Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; null ; Conference date: 07-02-2020 Through 12-02-2020.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Trans. Web*, 13(2).

Haohan Wang, Zexue He, Zachary L. Lipton, and Eric P. Xing. 2019. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909, Melbourne, Australia. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8(3–4):229–256*.

Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? sensational headline generation with auto-tuned reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3065–3075, Hong Kong, China. Association for Computational Linguistics.

Youwen Yang. 2011. A cognitive interpretation of discourse deixis. *Theory and Practice in Language Studies*, 1:128–135.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 617–626.

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*.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. pages 8759–8771.

Appendix

A Debiasing Faithfulness Scorer

We observe that lexical bias affects our faithfulness scorer. In particular, many verified headlines entailed by their contents contain the words "verification" or "rumor" which results in a shortcut model, i.e., the NLI model tends to classify samples as entailment based simply on the existence of these certain words. In addition, the word-overlap bias (WOB)—classifying hypothesis and premise as entailment because of high word overlap (Naik et al., 2018)—also harms our entailment task to ensure faithfulness. Thus, we follow Zhou and Bansal (2020) in adopting a model-level debiasing module for pretraining entailment. A bag-of-words (BoW) sub-model is deployed to capture superficial features, since it has the least reasoning ability, and is more likely to use shortcuts to make predictions. The main NLI model, in turn, consists of two bi-LSTM networks that are capable of reasoning over deeper semantics. During training, HEX projection (Wang et al., 2019) is used to screen out superficial features by making the hidden state of the main NLI model and the BoW sub-model orthogonal, forcing the main classifier to focus on deeper semantic features.

B Analysis of Hyperparameters

λ	**Generated sample headlines**
0.2	Apples from Linyi county are unsalable? Linyi county government clarifies: **over-exaggerating**.
0.6	Are apples from the county of Linyi unsellable? **This story is a rumor!**
1.0	Apples from Linyi county are unsalable? e-commerce's customer service: the merchant may violate portrait rights.

Table 5: Headlines generated with different λ. Orange words are more sensational expressions.

Here we provide a qualitative analysis to examine the sensitivity of λ and α; recall that λ balances MLE loss and RL loss, and α influences the sensationalism. In a sense, a higher λ leads to robust yet boring generation, as a higher λ relies more on MLE, and MLE loss is calculated according to the gold title. Table 5 summarizes title generation with different λ. Note that $\lambda = 0.0$ is ignored, as it completely relies on RL loss, which often leads to broken generation results and is not practical in general. When $\lambda = 1.0$, the model relies completely on MLE loss and is identical to using only Ptr-G. A smaller λ creates more diversity, and $\lambda = 0.2$ balances diversity, attractiveness, and fluency. Also, in $\lambda = 0.2$ and $\lambda = 0.6$, more sensational or eye-catching words are used (highlighted in orange in Table 5), whereas $\lambda = 1.0$ shows a plain, ordinary tone. When $\lambda = 1.0$, the generated results are unrelated and unintelligible, which also shows that our faithfulness scorer helps align headline to content, since there is no faithfulness reward when $\lambda = 1.0$.

We also conducted an analysis of how different values of α affect the generated headline. In Table 6, a lower α indicates that a greater emphasis is put on sensationalism. A higher α yields a relatively simple and monotonous sentence structure. In Table 6, $\alpha = 1.0$ predominantly generates affirmative sentences including "can", "is" or "will", which are highlighted in red. On the other hand, a less dominant α provides more flexibility with respect to the sentence structure and adds diversity. When the reward is completely provided by the sensation scorer and the FR type reward ($\alpha = 0.0$), it seems that the model generates headlines from a different aspect and focuses on different keywords (highlighted in blue). However, such diversity comes at the risk of spurious invention. When $\alpha = 0.0$, the generated result is similar to the tone of fake news, which creates a clickbait without specifying the facts. When $\alpha = 0.4$, the generated headlines maintain high veracity while improving attractiveness. Accordingly, we use $\lambda = 0.2$ and $\alpha = 0.4$ as our default setting.

α	**Generated sample headlines**
0.0	**How much harm** will new clothes do to our body?!
0.4	Rumor has it that formaldehyde in new clothes causes cancer.
0.8	Formaldehyde **can** cause cancer.
1.0	Formaldehyde **is** a carcinogen.

Table 6: Generated headlines with different α. Blue words are more diversified expressions, and red words are monotonic affirmatives.

FR type	Example headlines
1. Demonstrative pronouns	这是今年最大的养生谣言！ **This** is the biggest regimen rumor in the year!
2. Personal pronouns	据说这是他最后一部参与表演的电影 It is said that this is the last movie **he** participated in.
3. Define articles	Apple Logo 来自图灵咬一口的「苹果」？ The logo of Apple Inc. comes from **"the apple"** that Alan Turing bit?
4. Ellipses	接听了陌生Facetime，结果... After answering the unfamiliar Facetime, **it turns out...**
5. Imperatives	我的天！雪糕二次冷冻会产生可溶性毒蛋白? **OMG!** Will a second freezing of ice cream produce **soluble toxic protein**?
6. Interrogatives	微波炉加热食物会致癌吗? Does microwaved food cause cancer**?**
7. General nouns	天再冷也不能吃辣的**5**种人，吃了等于慢性自杀！ **5 kinds of people** who can't eat spicy food no matter how cold it is outside: eating it is tantamount to slow suicide!
8. Location Adverbs	经常吃方便面真的会致癌吗？正确的解释在这里 Does eating instant noodles often really cause cancer? **Here** is the correct explanation.

Figure 4: Examples of different types of forward references. Words highlighted in orange are the main characteristics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
after the conclusion (section 6).

☑ A2. Did you discuss any potential risks of your work?
in the ethic statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1.

☑ A4. Have you used AI writing assistants when working on this paper?
correct grammar errors throughout the whole paper.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5.2.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.2.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5.2.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5.1, Section 5.4.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 5.4.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Due to the space limit, we can provide details here. We recruit 5 graduate students from Taiwan to conduct the human annotation and we pay $10 per hour for them.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 5.1.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Due to the space limit, we can provide details here. They are all graduate students from Taiwan.