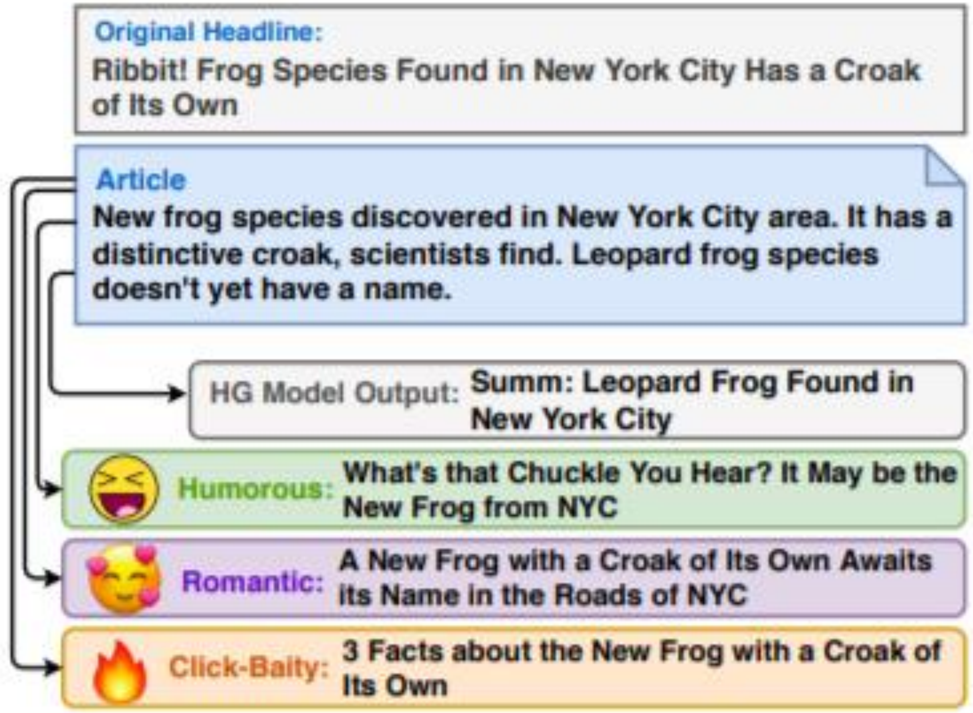
# Hooks in the Headline: Learning to Generate Headlines with Controlled Styles

**Di Jin,[1] Zhijing Jin,[2] Joey Tianyi Zhou,[3*] Lisa Orii,[4] Peter Szolovits[1]**
[1]CSAIL, MIT, [2]Amazon Web Services, [3]A*STAR, Singapore, [4]Wellesley College
{jindi15,psz}@mit.edu, zhijing.jin@connect.hku.hk
zhouty@ihpc.a-star.edu.sg, lorii@wellesley.edu

# Motivation

- Current headline generation (HG) models can only generate plain, factual headlines, failing to learn from the original human reference,they  can’t generate appealing headlines.
- We propose a new task, Stylistic Headline Generation (SHG), to enrich the headlines with three style options (humor, romance and clickbait), in order to attract more readers.

# Motivation



Original Headline:
Ribbit! Frog Species Found in New York City Has a Croak of Its Own

Article
New frog species discovered in New York City area. It has a distinctive croak, scientists find. Leopard frog species doesn't yet have a name.

HG Model Output: Summ: Leopard Frog Found in New York City

Humorous: What's that Chuckle You Hear? It May be the New Frog from NYC

Romantic: A New Frog with a Croak of Its Own Awaits its Name in the Roads of NYC

Click-Baity: 3 Facts about the New Frog with a Croak of Its Own

Figure 1: Given a news article, current HG models can only generate plain, factual headlines, failing to learn from the original human reference. It is also much less attractive than the headlines with humorous, romantic and click-baity styles.

# Contribution

- it is the first research on the generation of attractive news headlines with styles without any specific style-specific data
- Automatic and human evaluation shows out model **TitleStylist** can generate relevant, fluent headlines with three styles
- Our model can flexibly incorporate multiple styles,

# Model-Problem Formulation

- The model is trained on a source dataset $S$ and target dataset $T$,$S = \{(a_i, h_i)\}_{i=1}^{N}$ ,denote $A = \{a_i\}_{i=1}^{N}, H = \{h_i\}_{i=1}^{N}$. where $a_i$ is an article and $h_i$ is its plain headline.$T = \{t_i\}_{i=1}^{N}$ comprises of sentences written in a specific style.
- Our task is to learn the conditional distribution $P(T|A)$.

# Model Architecture

- We adopt a seq2seq model based on the Transformer architecture.it consists of a 6-layer encoder $E(.,\theta_E)$, and a 6-layer decoder $G(.,\theta_G)$.

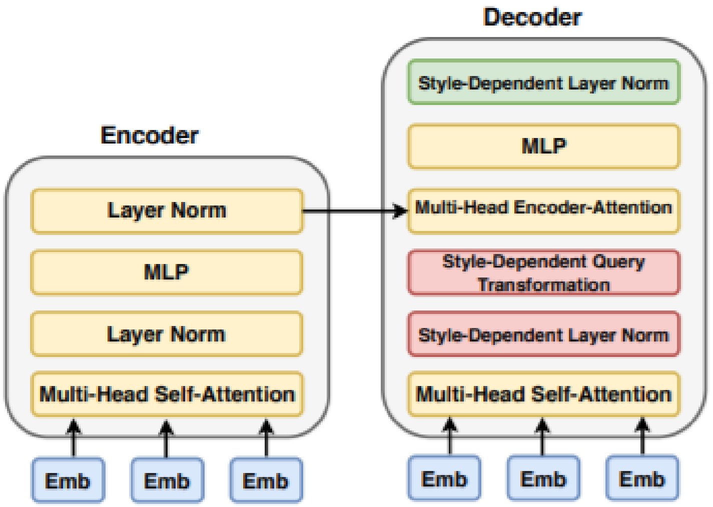

Decoder
Style-Dependent Layer Norm
MLP
Multi-Head Encoder-Attention
Style-Dependent Query Transformation
Style-Dependent Layer Norm
Multi-Head Self-Attention
Emb
Emb
Emb
Encoder
Layer Norm
MLP
Layer Norm
Multi-Head Self-Attention
Emb
Emb
Emb

Figure 2: The Transformer-based architecture of our model.

# Model- multitask learning

- To disentangle the latent style from the text,we adopt a multitask learning framework: training on on summarization and denoising auto-encoder(DAE)
- With the source domain dataset $S$, we can learn $z_s = E_S(A)$ and $H_s = G_s(z_s)$， where $z_s$ is the learned latent representation

$$\mathcal{L}_S(\boldsymbol{\theta}_{\boldsymbol{E}_S}, \boldsymbol{\theta}_{\boldsymbol{G}_S}) = \mathbb{E}_{(\boldsymbol{a},\boldsymbol{h})\sim S}[-\log p(\boldsymbol{h}|\boldsymbol{a}; \boldsymbol{\theta}_{\boldsymbol{E}_S}, \boldsymbol{\theta}_{\boldsymbol{G}_S})],$$

- Which can be furtherly expressed as:

$$p(\boldsymbol{h}|\boldsymbol{a}; \boldsymbol{\theta}_{\boldsymbol{E}_S}, \boldsymbol{\theta}_{\boldsymbol{G}_S}) = \prod^{L} p(h_t|\{h_1, ..., h_{t-1}\}, \boldsymbol{z}_S; \boldsymbol{\theta}_{\boldsymbol{G}_S}),$$

# Model- multitask learning

- **DAE training for $\boldsymbol{\theta}_{\boldsymbol{E_T}}$ and $\boldsymbol{\theta}_{\boldsymbol{G_T}}$**
- For target style source T, since we only have the sentence t without any paired articles, we train $z_T = E_T(\tilde{t})$ and $t = G_T(z_T)$ by solving an unsupervised re- construction learning task
- Where $z_T$ is the learned latent representation in the target domain, $\tilde{t}$ is obtained by randomly deleting or blanking some words and shuffling the word orders. We minimize the reconstruction loss:

$$\mathcal{L}_T(\boldsymbol{\theta}_{\boldsymbol{E_T}}, \boldsymbol{\theta}_{\boldsymbol{G_T}}) = \mathbb{E}_{\boldsymbol{t} \sim \boldsymbol{T}}[-\log p(\boldsymbol{t}|\tilde{\boldsymbol{t}})],$$

# Model- multitask learning

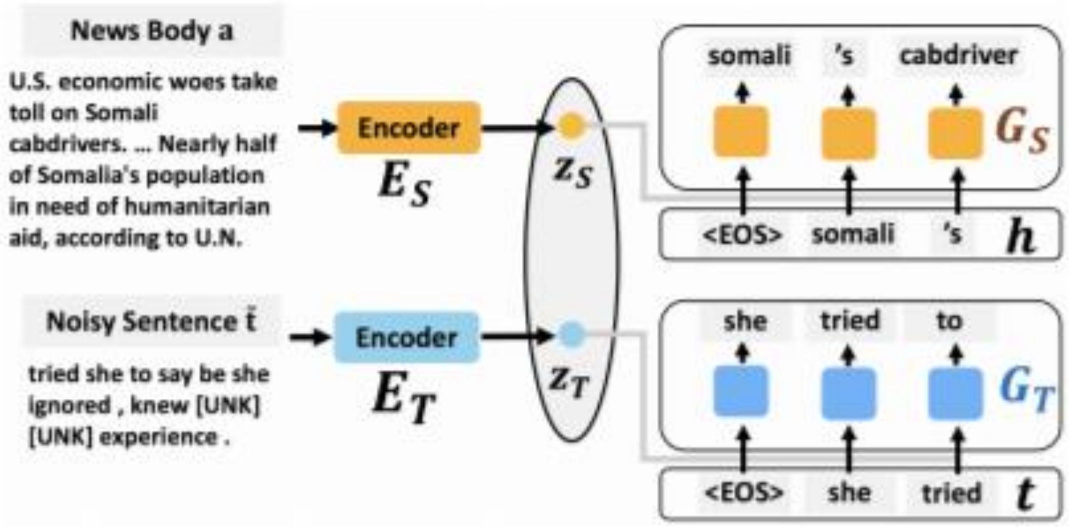

News Body a
U.S. economic woes take toll on Somali cabdrivers. ... Nearly half of Somalia's population in need of humanitarian aid, according to U.N.
Encoder
$E_S$
$z_S$
somali
's
cabdriver
$G_S$
<EOS>
somali
's
$h$
Noisy Sentence $\tilde{t}$
tried she to say be she ignored , knew [UNK] [UNK] experience .
Encoder
$E_T$
$z_T$
she
tried
to
$G_T$
<EOS>
she
tried
$t$

Figure 3: Training scheme. Multitask training is adopted to combine the summarization and DAE tasks.

We train the model by jointly minimize the supervised seq2seq training loss $L_S$ and the unsupervised denoised denoised auto-encoding loss $L_T$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_{\boldsymbol{E_S}}, \boldsymbol{\theta}_{\boldsymbol{G_S}}, \boldsymbol{\theta}_{\boldsymbol{E_T}}, \boldsymbol{\theta}_{\boldsymbol{G_T}}) &= \lambda \mathcal{L}_S(\boldsymbol{\theta}_{\boldsymbol{E_S}}, \boldsymbol{\theta}_{\boldsymbol{G_S}}) \\ &\quad + (1-\lambda)\mathcal{L}_T(\boldsymbol{\theta}_{\boldsymbol{E_T}}, \boldsymbol{\theta}_{\boldsymbol{G_T}}),\end{aligned} \tag{4}$$

# Model -Parameter-Sharing Scheme

- We perform more constraints in multitask training. We aim to infer the conditional distribution: $P(T|A) = G_T(E_S(A))$, However , without samples from $P(A,T)$, this challenge is even impossible if $E_S$ and $E_T$ , or $G_S$ and $G_T$ are completely independent of each other.
- by exposing the model to both summarization task and style-carrying text reconstruction task, the model will have some sense of target style while summarizing the article.
- we share all parameters of the encoder $E_S$ and $E_T$
- For $G_S$ and $G_T$ ,only style-independent parameters are shared , i.e., the parameters except Style Layer Normalization and Style-Guided Encoder Attention

# Model- style layer normalization

- The style layer normalization aims to transform a layers activation x into into a normalized activation z specific to the style s:

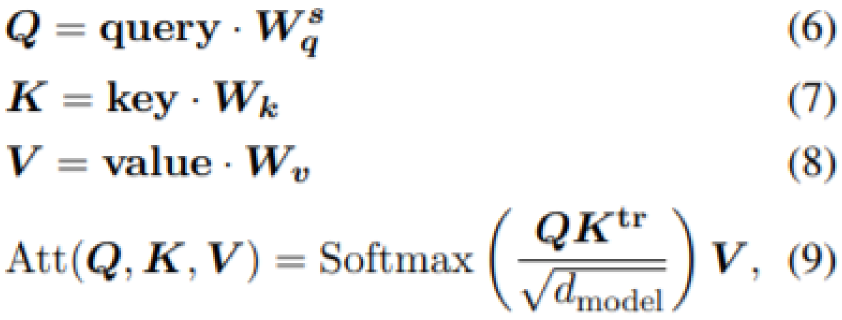
$$z = \gamma_s(\frac{x - \mu}{\sigma}) - \beta_s,$$

- Where $\mu$ and $\sigma$ are the mean and standard deviation of the batch of x, $\gamma_s$ and $\beta_s$ are the style-specific parameters learned from data.

we use style layer normalization and final layer normalization for all layers

# Model-Style-Guided Encoder Attention

- . The attention patterns should be different for the summarization and the reconstruction tasks due to their different inherent nature



$$\boldsymbol{Q} = \text{query} \cdot \boldsymbol{W}_q^s \quad (6)$$

$$\boldsymbol{K} = \text{key} \cdot \boldsymbol{W}_k \quad (7)$$

$$\boldsymbol{V} = \text{value} \cdot \boldsymbol{W}_v \quad (8)$$

$$\text{Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{tr}}}{\sqrt{d_{\text{model}}}}\right)\boldsymbol{V}, \quad (9)$$

- $\boldsymbol{W}_q^s, \boldsymbol{W}_k, \boldsymbol{W}_v$ denote the scaled dot-product matrix for affine transformation; we specialize e the dot-product matrix $\boldsymbol{W}_q^s$ for different styles

# Dataset

- --source data

we combine New York Times (56K) and CNN(90 K), resulting in 56,899 news abstracts-headlines pairs.

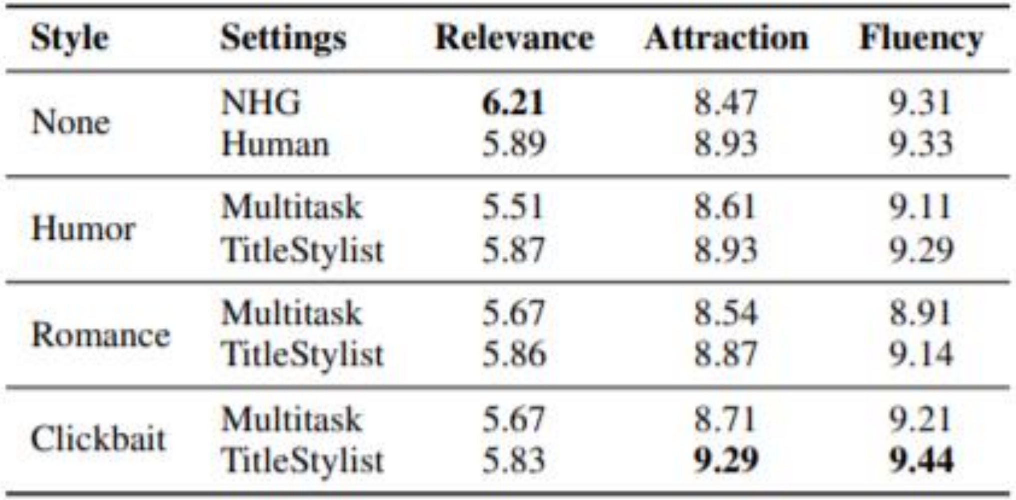
- --target data

➢ **Humor and Romance,** use humor and romance novel and split the document into sentences

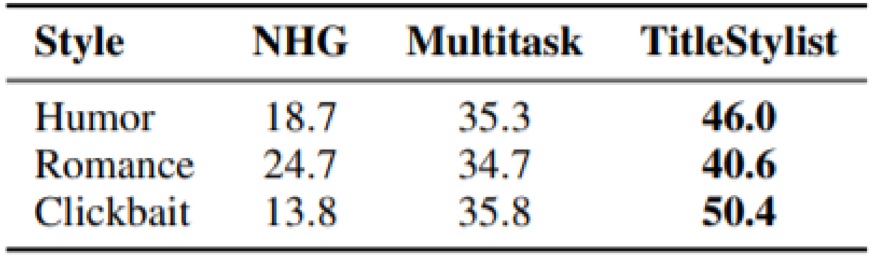
➢ **Clickbait,** The Examiner SpamClickBait News dataset

# Experiments-human evaluation

- We randomly select 50 articles and evaluate in four criteria:: (1) relevance, (2) attractiveness, (3) language fluency, and (4) style strength



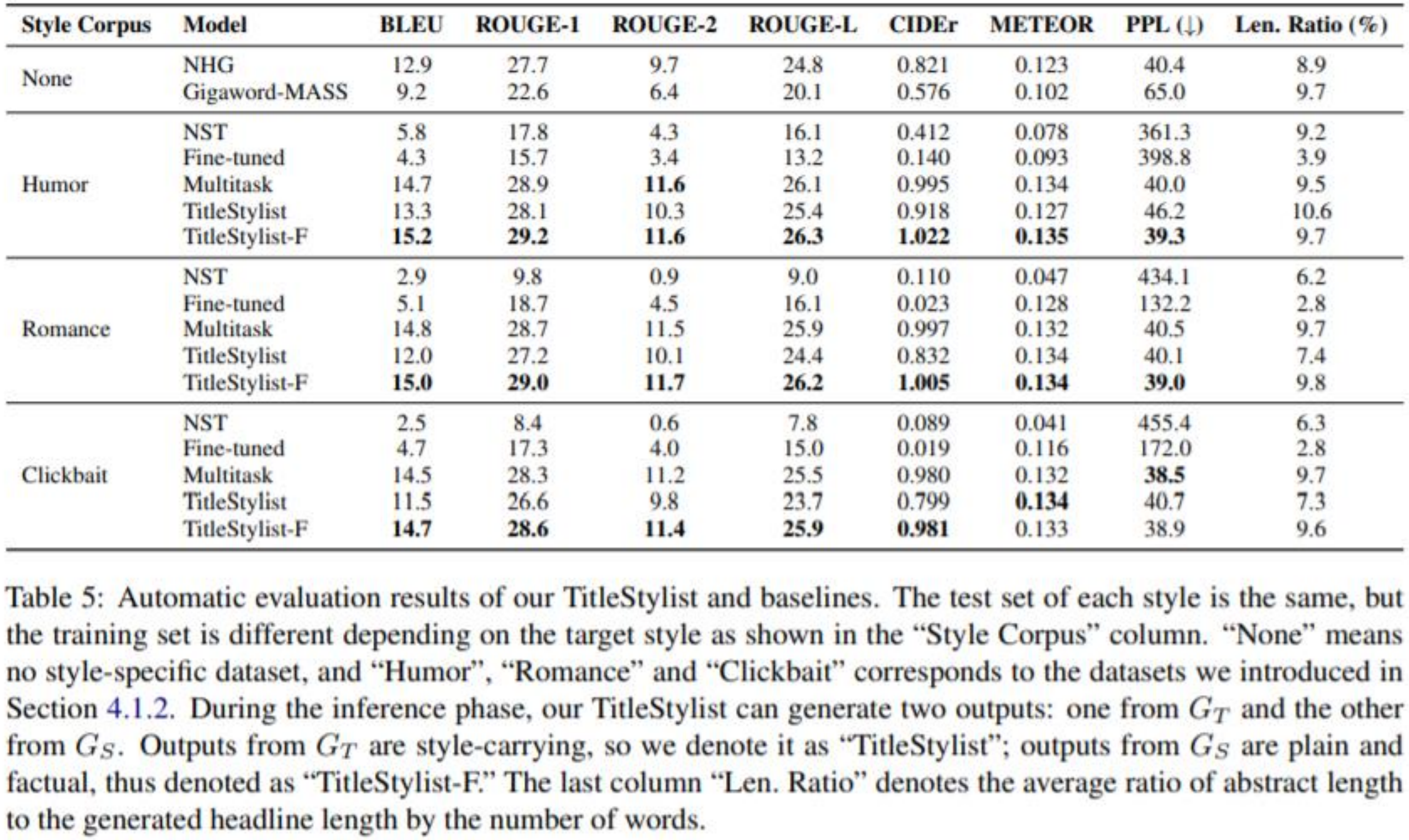
| Style | Settings | Relevance | Attraction | Fluency |
|---|---|---|---|---|
| None | NHG | **6.21** | 8.47 | 9.31 |
| | Human | 5.89 | 8.93 | 9.33 |
| Humor | Multitask | 5.51 | 8.61 | 9.11 |
| | TitleStylist | 5.87 | 8.93 | 9.29 |
| Romance | Multitask | 5.67 | 8.54 | 8.91 |
| | TitleStylist | 5.86 | 8.87 | 9.14 |
| Clickbait | Multitask | 5.67 | 8.71 | 9.21 |
| | TitleStylist | 5.83 | **9.29** | **9.44** |



| Style | NHG | Multitask | TitleStylist |
|---|---|---|---|
| Humor | 18.7 | 35.3 | **46.0** |
| Romance | 24.7 | 34.7 | **40.6** |
| Clickbait | 13.8 | 35.8 | **50.4** |

Table 4: Percentage of choices (%) for the most humorous or romantic headlines among TitleStylist and two baselines NHG and Multitask.

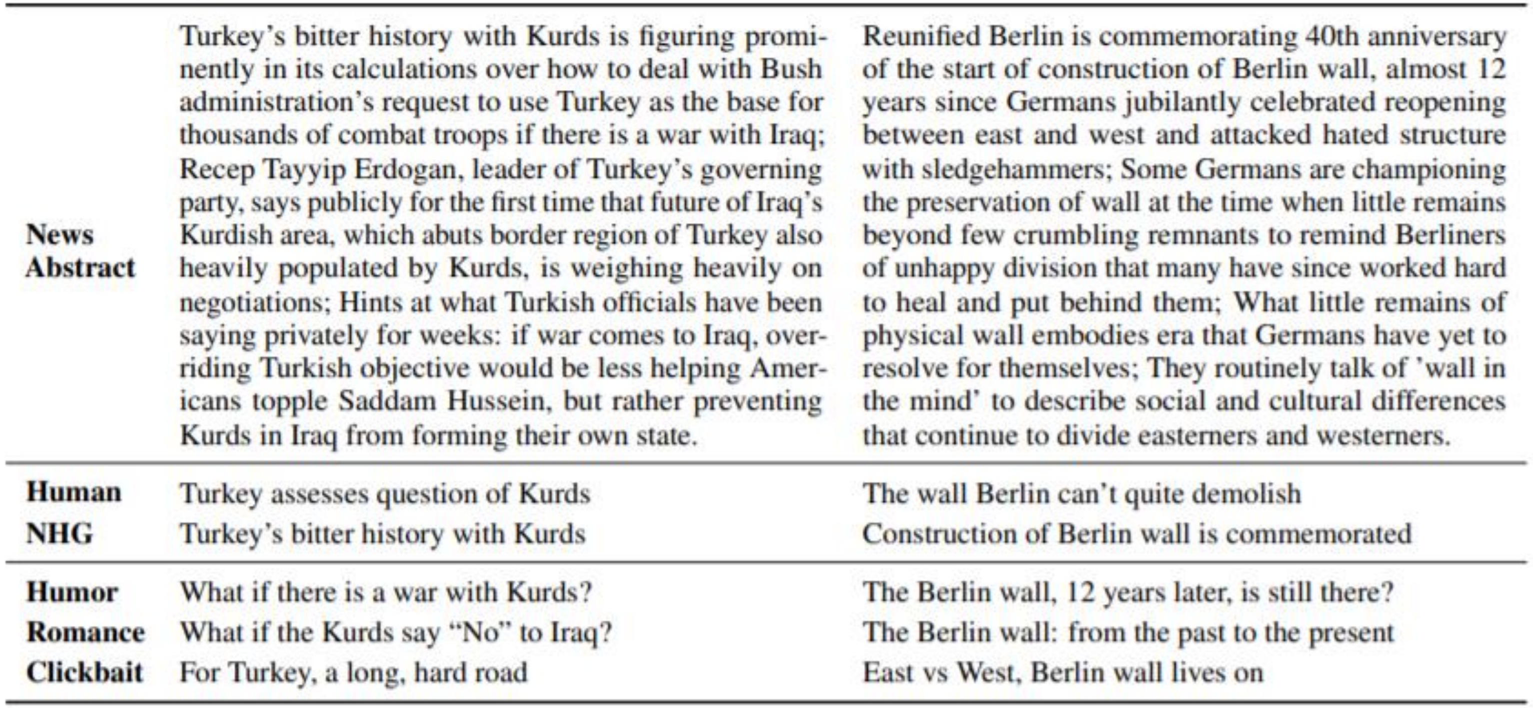
# Experiments-automatic evaluation



| Style Corpus | Model | BLEU | ROUGE-1 | ROUGE-2 | ROUGE-L | CIDEr | METEOR | PPL (↓) | Len. Ratio (%) |
|---|---|---|---|---|---|---|---|---|---|
| None | NHG | 12.9 | 27.7 | 9.7 | 24.8 | 0.821 | 0.123 | 40.4 | 8.9 |
| | Gigaword-MASS | 9.2 | 22.6 | 6.4 | 20.1 | 0.576 | 0.102 | 65.0 | 9.7 |
| Humor | NST | 5.8 | 17.8 | 4.3 | 16.1 | 0.412 | 0.078 | 361.3 | 9.2 |
| | Fine-tuned | 4.3 | 15.7 | 3.4 | 13.2 | 0.140 | 0.093 | 398.8 | 3.9 |
| | Multitask | 14.7 | 28.9 | **11.6** | 26.1 | 0.995 | 0.134 | 40.0 | 9.5 |
| | TitleStylist | 13.3 | 28.1 | 10.3 | 25.4 | 0.918 | 0.127 | 46.2 | 10.6 |
| | TitleStylist-F | **15.2** | **29.2** | **11.6** | **26.3** | **1.022** | **0.135** | **39.3** | 9.7 |
| Romance | NST | 2.9 | 9.8 | 0.9 | 9.0 | 0.110 | 0.047 | 434.1 | 6.2 |
| | Fine-tuned | 5.1 | 18.7 | 4.5 | 16.1 | 0.023 | 0.128 | 132.2 | 2.8 |
| | Multitask | 14.8 | 28.7 | 11.5 | 25.9 | 0.997 | 0.132 | 40.5 | 9.7 |
| | TitleStylist | 12.0 | 27.2 | 10.1 | 24.4 | 0.832 | 0.134 | 40.1 | 7.4 |
| | TitleStylist-F | **15.0** | **29.0** | **11.7** | **26.2** | **1.005** | **0.134** | **39.0** | 9.8 |
| Clickbait | NST | 2.5 | 8.4 | 0.6 | 7.8 | 0.089 | 0.041 | 455.4 | 6.3 |
| | Fine-tuned | 4.7 | 17.3 | 4.0 | 15.0 | 0.019 | 0.116 | 172.0 | 2.8 |
| | Multitask | 14.5 | 28.3 | 11.2 | 25.5 | 0.980 | 0.132 | **38.5** | 9.7 |
| | TitleStylist | 11.5 | 26.6 | 9.8 | 23.7 | 0.799 | **0.134** | 40.7 | 7.3 |
| | TitleStylist-F | **14.7** | **28.6** | **11.4** | **25.9** | **0.981** | 0.133 | 38.9 | 9.6 |

Table 5: Automatic evaluation results of our TitleStylist and baselines. The test set of each style is the same, but the training set is different depending on the target style as shown in the "Style Corpus" column. "None" means no style-specific dataset, and "Humor", "Romance" and "Clickbait" corresponds to the datasets we introduced in Section 4.1.2. During the inference phase, our TitleStylist can generate two outputs: one from $G_T$ and the other from $G_S$. Outputs from $G_T$ are style-carrying, so we denote it as "TitleStylist"; outputs from $G_S$ are plain and factual, thus denoted as "TitleStylist-F." The last column "Len. Ratio" denotes the average ratio of abstract length to the generated headline length by the number of words.

# Experiments cases



| | | |
|---|---|---|
| **News Abstract** | Turkey's bitter history with Kurds is figuring prominently in its calculations over how to deal with Bush administration's request to use Turkey as the base for thousands of combat troops if there is a war with Iraq; Recep Tayyip Erdogan, leader of Turkey's governing party, says publicly for the first time that future of Iraq's Kurdish area, which abuts border region of Turkey also heavily populated by Kurds, is weighing heavily on negotiations; Hints at what Turkish officials have been saying privately for weeks: if war comes to Iraq, overriding Turkish objective would be less helping Americans topple Saddam Hussein, but rather preventing Kurds in Iraq from forming their own state. | Reunified Berlin is commemorating 40th anniversary of the start of construction of Berlin wall, almost 12 years since Germans jubilantly celebrated reopening between east and west and attacked hated structure with sledgehammers; Some Germans are championing the preservation of wall at the time when little remains beyond few crumbling remnants to remind Berliners of unhappy division that many have since worked hard to heal and put behind them; What little remains of physical wall embodies era that Germans have yet to resolve for themselves; They routinely talk of 'wall in the mind' to describe social and cultural differences that continue to divide easterners and westerners. |
| **Human** | Turkey assesses question of Kurds | The wall Berlin can't quite demolish |
| **NHG** | Turkey's bitter history with Kurds | Construction of Berlin wall is commemorated |
| **Humor** | What if there is a war with Kurds? | The Berlin wall, 12 years later, is still there? |
| **Romance** | What if the Kurds say "No" to Iraq? | The Berlin wall: from the past to the present |
| **Clickbait** | For Turkey, a long, hard road | East vs West, Berlin wall lives on |

Table 3: Examples of style-carrying headlines generated by TitleStylist.

# Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog

**Libo Qin[†], Xiao Xu[†], Wanxiang Che[†]*, Yue Zhang[‡], Ting Liu[†]**

[†]Research Center for Social Computing and Information Retrieval
[†]Harbin Institute of Technology, China
[‡]School of Engineering, Westlake University, China
[‡]Institute of Advanced Technology, Westlake Institute for Advanced Study
[†]{lbqin, xxu, car,tliu}@ir.hit.edu.cn, yue.zhang@wias.org.cn

# Model framework

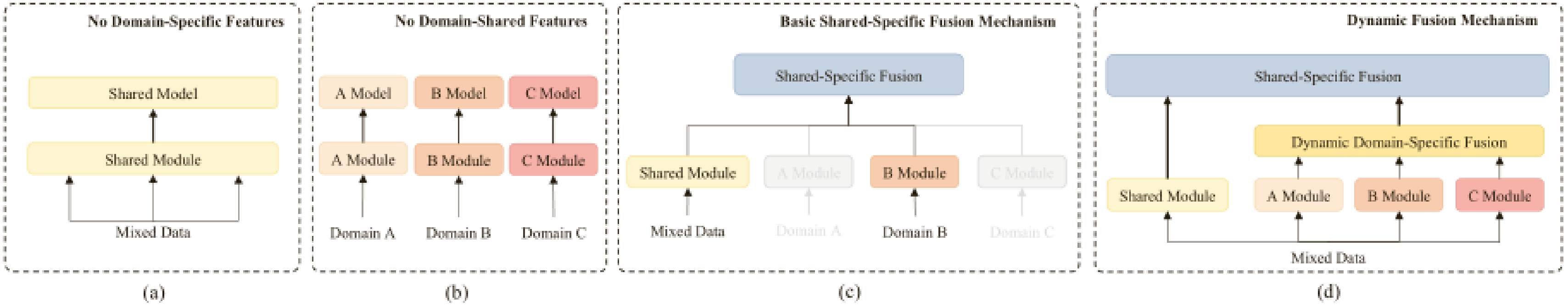


Figure 2: Methods for multi-domain dialogue. Previous work either trains a general model on mixed multi-domain mixed datasets (a), or on each domain separately (b). The basic shared-private framework is shown (c). Our proposed extension with dynamic fusion mechanism is shown (d).