Topic-Aware Neural Keyphrase Generation for Social Media Language

Yue Wang

摘要

每天都有大量的用户生成的内容出现在社交媒体平台上。为研究关键词预测，促进语言的自动理解该文章从大量的帖子中提取突出的信息。与许多现有的从原文当中提取单词从而形成关键词的方法不同，该文章提出了一种基于 seq2seq 的神经网络关键词生成框架从而能够生成不在文本当中的关键词。此外该模型具有主题感知能力，可以对语料库级别的潜在主题表示进行联合建模从而缓解社交媒体语言中广泛存在的数据稀疏问题。该文章在中英文的的社交媒体上的三个数据集进行实验，实验结果显示文章模型的效果优于不适用潜在主题提取和生成的模型。

1 引言

社交媒体在全球范围内发展十分迅速，并给人们提供了更多丰富且新鲜的信息。与此同时，迅速发展的社交媒体也带来了每天数以百万的帖子的文本量在此情况下，开发一个能够自动吸收大量社交媒体文本并找出其中重要内容的系统十分迫切。

尽管社交媒体关键词识别方面已有了大量的努力，但大多数都是从源帖子中提取关键词，因此对于帖子的关键词并没有在帖子中出现的情况表现并不好。由于存在用户发帖子写作风格并不正式的情况，关键词在帖子中缺失的情况十分突出。本文的工作与以往的研究不同，采用了一个序列生成的框架对社交媒体的关键词进行预测，从而实现生成未出现在帖子文本当中的关键词。

2 相关工作

2.1 关键词预测

许多之前的工作是基于提取的监督或无监督的方法，即从原文档中提取关键词。其中监督的方法大多基于序列标注[1] 和使用各种特征的二元分类[2]。对于无监督的方法有如图排名[3]，文档聚类[4]和像 TF-IDF[5] 的统计模型

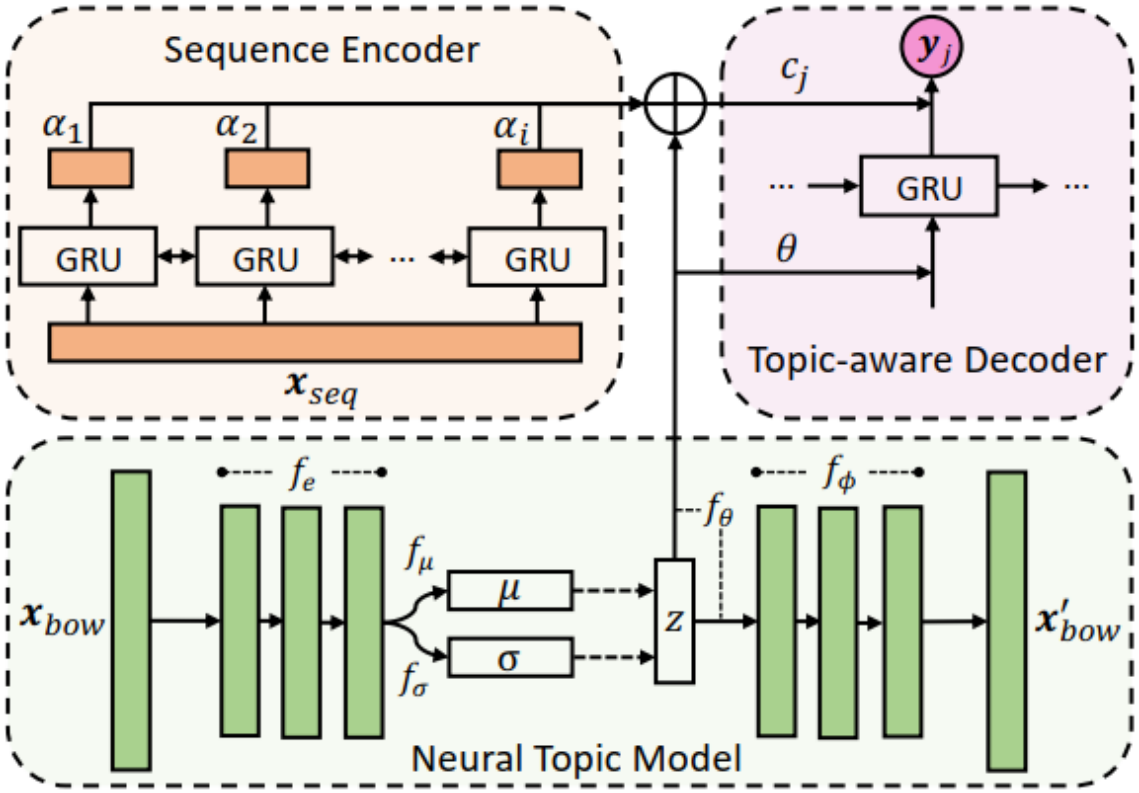
不同的是该文章的关键词生成受 seq2seq 关键词生成模型[6-9] 的启发。但这些方法最初是设计给科学文章且社交媒体语言存在稀疏性，因此模型的表现不可避免受到影响。

2.2 主题建模

主题模型通常基于贝叶斯图形模型的 LDA[10], 然而这种模型需要专业的知识去定制推理算法。本文章的框架采用了神经主题模型[11-12] 用以推断潜在得主题，且有助于与其他神经网络进行端到端训练。这种方法在引用推荐[13] 和对话理解[14]上很有帮助。还有一种同时训练主题模型和短文本分类得方法[15]由于关键词的巨大多样性，并不适合该文章的场景[16]。

3 本文方法

3.1 本文方法概述



本文的模型包括两个部分一个是基于 seq2seq 的模型，和一个 NTM 神经主题模型。

NTM 模型包括 Encoder 和 Decoder 两个部分且是一个类似于数据重建的过程。该模型的输入输出都采用词袋模型的方式。其中 Encoder 会利用输入的 x_{bow} 通过神经网络学习到用于表示主题 z 的变量 μ 和 σ。而 Decoder 会通过 μ 和 σ 生成主题变量 $z \sim \mathcal{N}(\mu, \sigma^2)$，再通过 z 生成代表 k 个主题的 k 维向量 $\theta = softmax(f_\theta(z))$，最后利用感知机将 θ 重建为 x'_{bow}，在此过程中学习到的 θ 将会作为主题加入到关键词生成的过程中。

而 seq2seq 的模型的 Encoder 会将文本序列 x_{seq} 输入到双向 GRU 中生成隐状态 M，而 Decoder 会将 M 和 θ 输入到单向 GRU 中使用 Attention 和 Copy 机制输出关键词短语, 这允许模型能够直接从文本当中提取关键词。

3.2 损失函数定义

对于 NTM 模型而言其损失函数包括两个部分一个是文本重建的损失，另一个则是 KL 散度

$$\mathcal{L}_{NTM} = D_{KL}(p(z) \parallel q(z \mid x)) - \mathbb{E}_{q(z|x)}[p(z \mid x)]$$

而关键词生成模型的损失函数则是交叉熵损失

$$\mathcal{L}_{KG} = -\sum_{n=1}^{N} \log(Pr(y_n \mid x_n, \theta_n))$$

因此整个模型框架的损失函数可以定义为

$$\mathcal{L} = \mathcal{L}_{NTM} + \gamma \cdot \mathcal{L}_{KG}$$

4 复现细节

4.1 与已有开源代码对比

本次实验的复现基于文章提供的开源代码，在 NTM 部分添加了一种不同的神经主题模型 ETM[17] (Embedded Topic Model)。ETM 模型会将词汇映射到 L 维的低维空间，并用 K 个主题代表每一篇文档，通过将词语嵌入低位空间用向量表示能够表示词语之间的相关性。

加入 ETM 模型后，关键词生成模型在进行联合训练时可以选择不同的主题模型提供的主题表示，增加选择的多样性。同时在复现时为缓解 ETM 训练时可能出现的 KL 消失问题，我在 ETM 的损失函数中将 KL 散度添加了 0-1 的权重循环训练[18]，这在本次复现中对 ETM 的训练有帮助。

4.2 实验环境搭建

实验环境采用的是 Python 3.10.8 和 Pytorch 1.12.1

4.3 界面分析与使用说明

为适应 ETM 模型代码运行的需要，添加了一部分参数：

```
-ntm_type 选择主题模型为ntm还是etm
-rho_size 设置ETM模型中rho的维度
-enc_drop 设置ETM模型中encode的dropout rate
-theta_act 设置theta的激活函数
-klw_cycle 设置循环KL权重的轮数
-klw_start 设置KL权重从多少开始循环
-klw_stop 设置KL权重爬升到多少停止爬升
-klw_ratio 设置KL权重爬升时间在一轮循环中的占比

例如如训练ETM模型时应该使用：
python train.py -data_tag Weibo_s100_t10 -only_train_ntm -ntm_warm_up_epochs
    100 -ntm_type etm -rho_size 2750

而在联合训练时使用ETM应该使用：
python train.py -data_tag Weibo_s100_t10 -copy_attention -
   use_topic_represent -load_pretrain_ntm -joint_train -topic_attn -
   topic_dec -topic_copy -topic_attn_i -ntm_type etm -rho_size 2750 -
   check_pt_ntm_model_path [ntm_model_path]

使用联合训练的模型预测：
python predict.py -model [seq2seq model path] -ntm_model [ntm model path] -
   ntm_type etm -rho_size 2750

最后做出评分：
python pred_evaluate.py -src data/Weibo/test_src.txt -trg data/Weibo/
   test_trg.txt -pred [predictions file path]
```

4.4 创新点

添加了不同的主题模型进行对比，给使用者提供不同的选择。

5 实验结果分析

以下是原论文在提供的结果：

Model	Twitter			Weibo			StackExchange		
	F1@1	F1@3	MAP	F1@1	F1@3	MAP	F1@3	F1@5	MAP
Baselines									
MAJORITY	9.36	11.85	15.22	4.16	3.31	5.47	1.79	1.89	1.59
TF-IDF	1.16	1.14	1.89	1.90	1.51	2.46	13.50	12.74	12.61
TEXTRANK	1.73	1.94	1.89	0.18	0.49	0.57	6.03	8.28	4.76
KEA	0.50	0.56	0.50	0.20	0.20	0.20	15.80	15.23	14.25
State of the arts									
SEQ-TAG	22.79±0.3	12.27±0.2	22.44±0.3	16.34±0.2	8.99±0.1	16.53±0.3	17.58±1.6	12.82±1.2	19.03±1.3
SEQ2SEQ	34.10±0.5	26.01±0.3	41.11±0.3	28.17±1.7	20.59±0.9	34.19±1.7	22.99±0.3	20.65±0.2	23.95±0.3
SEQ2SEQ-COPY	36.60±1.1	26.79±0.5	43.12±1.2	32.01±0.3	22.69±0.2	38.01±0.1	31.53±0.1	27.41±0.2	33.45±0.1
SEQ2SEQ-CORR	34.97±0.8	26.13±0.4	41.64±0.5	31.64±0.7	22.24±0.5	37.47±0.8	30.89±0.3	26.97±0.2	32.87±0.6
TG-NET	-	-	-	-	-	-	32.02±0.3	27.84±0.3	34.05±0.4
Our model	**38.49**±0.3	**27.84**±0.0	**45.12**±0.2	**34.99**±0.3	**24.42**±0.2	**41.29**±0.4	**33.41**±0.2	**29.16**±0.1	**35.52**±0.1

而在复现中使用 ETM 跑出的结果是：

	Weibo			StackExchange		
	F1@1	F1@3	MAP@5	F1@3	F1@5	MAP@5
Micro	33.42	24.2	-	31.24	28.00	-
Macro	33.68	24.18	-	32.15	28.42	-
MAP@5	-	-	40.32	-	-	34.32

6 总结与展望

根据对比论文提供的结果可以发现，复现中使用 ETM 与关键词生成模型联合训练的结果并不如 NTM 好。但是相较于未使用主题模型的 Seq2seq 模型效果有提升，说明加入主题模型确实有效。而主题模型和关键词生成模型的合作并不是十分的紧密，将主题融入到关键词生成模型的方法仍然有开发的空间，

参考文献

[1] ZHANG Q, WANG Y, GONG Y, et al. Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter[C/OL]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016: 836-845. https://aclanthology.org/D16-1080. DOI: 10.18653/v1/D16-1080.

[2] WITTEN I H, PAYNTER G W, FRANK E, et al. KEA: Practical Automatic Keyphrase Extraction[J/OL]. CoRR, 1999, cs.DL/9902007. https://arxiv.org/abs/cs/9902007.

[3] MIHALCEA R, TARAU P. TextRank: Bringing Order into Text[C/OL]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004: 404-411. https://aclanthology.org/W04-3252.

[4] LIU Z, LI P, ZHENG Y, et al. Clustering to Find Exemplar Terms for Keyphrase Extraction[C/OL]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2009: 257-266. https://aclanthology.org/D09-1027.

[5] SALTON G, MICHAEL J. McGill[J]. Introduction to modern information retrieval, 1986, 1(4.1): 4-1.

[6] MENG R, ZHAO S, HAN S, et al. Deep Keyphrase Generation[C/OL]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 582-592. https://aclanthology.org/P17-1054. DOI: 10.18653/v1/P17-1054.

[7] CHEN J, ZHANG X, WU Y, et al. Keyphrase Generation with Correlation Constraints[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 4057-4066. https://aclanthology.org/D18-1439. DOI: 10.18653/v1/D18-1439.

[8] CHEN W, CHAN H P, LI P, et al. An Integrated Approach for Keyphrase Generation via Exploring the Power of Retrieval and Extraction[C/OL]//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). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 2846-2856. https://aclanthology.org/N19-1292. DOI: 10.18653/v1/N19-1292.

[9] CHEN W, GAO Y, ZHANG J, et al. Title-Guided Encoding for Keyphrase Generation[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 6268-6275. https://ojs.aaai.org/index.php/AAAI/article/view/4587. DOI: 10.1609/aaai.v33i01.33016268.

[10] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. J. Mach. Learn. Res., 2003, 3(null): 993-1022.

[11] MIAO Y, GREFENSTETTE E, BLUNSOM P. Discovering Discrete Latent Topics with Neural Variational Inference[C/OL]//PRECUP D, TEH Y W. Proceedings of Machine Learning Research: Proceedings of the 34th International Conference on Machine Learning: vol. 70. PMLR, 2017: 2410-2419. https://proceedings.mlr.press/v70/miao17a.html.

[12] SRIVASTAVA A, SUTTON C. Autoencoding variational inference for topic models[J]. ArXiv preprint arXiv:1703.01488, 2017.

[13] BAI H, CHEN Z, LYU M R, et al. Neural Relational Topic Models for Scientific Article Analysis[C/OL]//CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino, Italy: Association for Computing Machinery, 2018: 27-36. https://doi.org/10.1145/3269206.3271696. DOI: 10.1145/3269206.3271696.

[14] ZENG J, LI J, HE Y, et al. What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations[J/OL]. Transactions of the Association for Computational Linguistics, 2019, 7: 267-281. https://aclanthology.org/Q19-1017. DOI: 10.1162/tacl_a_00267.

[15] ZENG J, LI J, SONG Y, et al. Topic Memory Networks for Short Text Classification[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 3120-3131. https://aclanthology.org/D18-1351. DOI: 10.18653/v1/D18-1351.

[16] WANG Y, LI J, KING I, et al. Microblog Hashtag Generation via Encoding Conversation Contexts[C/OL]//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). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 1624-1633. https://aclanthology.org/N19-1164. DOI: 10.18653/v1/N19-1164.

[17] DIENG A B, RUIZ F J R, BLEI D M. Topic Modeling in Embedding Spaces[EB/OL]. arXiv. 2019. https://arxiv.org/abs/1907.04907.

[18] FU H, LI C, LIU X, et al. Cyclical annealing schedule: A simple approach to mitigating kl vanishing[J]. ArXiv preprint arXiv:1903.10145, 2019.