基于BERT-SUMOPN模型的抽取-生成式文本自动摘要
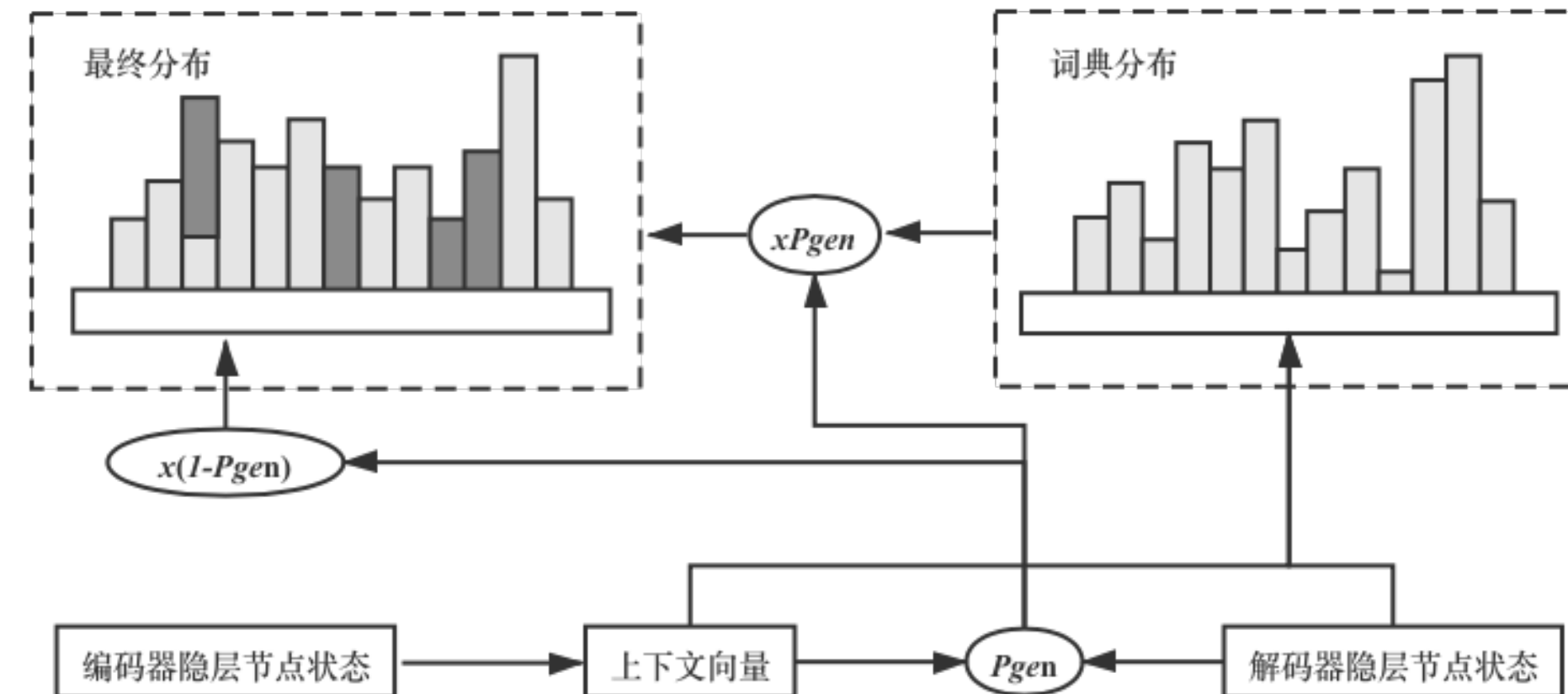
谭金源 刁宇峰 林鸿飞 （大连理工大学 信息检索实验室，辽宁省 大连市 116024）

论文摘要

抽取式摘要可读性、准确性较差，而生成式摘要也存在连贯性、逻辑性的不足。此外，两种摘要方法的传统模型对文本的向量表示往往不够充分、不够准确。针对以上问题，该文提出了一种基于BERT-SUMOPN模型的抽取-生成式摘要方法。模型通过BERT预训练语言模型获取文本向量，然后利用抽取式结构化摘要模型抽取文本中的关键句子，最后将得到的关键句子输入到生成式指针生成网络中，通过EAC损失函数对模型进行端到端训练，结合coverage机制减少生成重复，获取摘要。

BERT-SUMOPN模型

提出一种基于BERT-SUMOPN模型的抽取-生成式文本自动摘要方法，结合抽取式摘要模型和生成式摘要模型的优点，有效提升摘要结果可读性、准确性、逻辑性。

抽取式模型

抽取式模型使用结构化注意力对基础树的根进行建模，通过迭代的方法根据过去的内容推断复杂结构，得到句子标签列表作为关键句抽取依据，不需要得到摘要结果。

生成式模型

生成式模型部分采用指针生成网络模型，对指针网络和基于注意力机制的序列到序列模型进行结合，允许通过指针直接指向生成的单词，也可以从固定的词汇表中生成单词。

为了实现模型的端到端训练，使用EAC损失函数(Extractive-Abstractive-Coverage loss function)作为BERT-SUMOPN模型的损失函数，通过最小化EAC损失函数同时训练抽取、生成式模型。
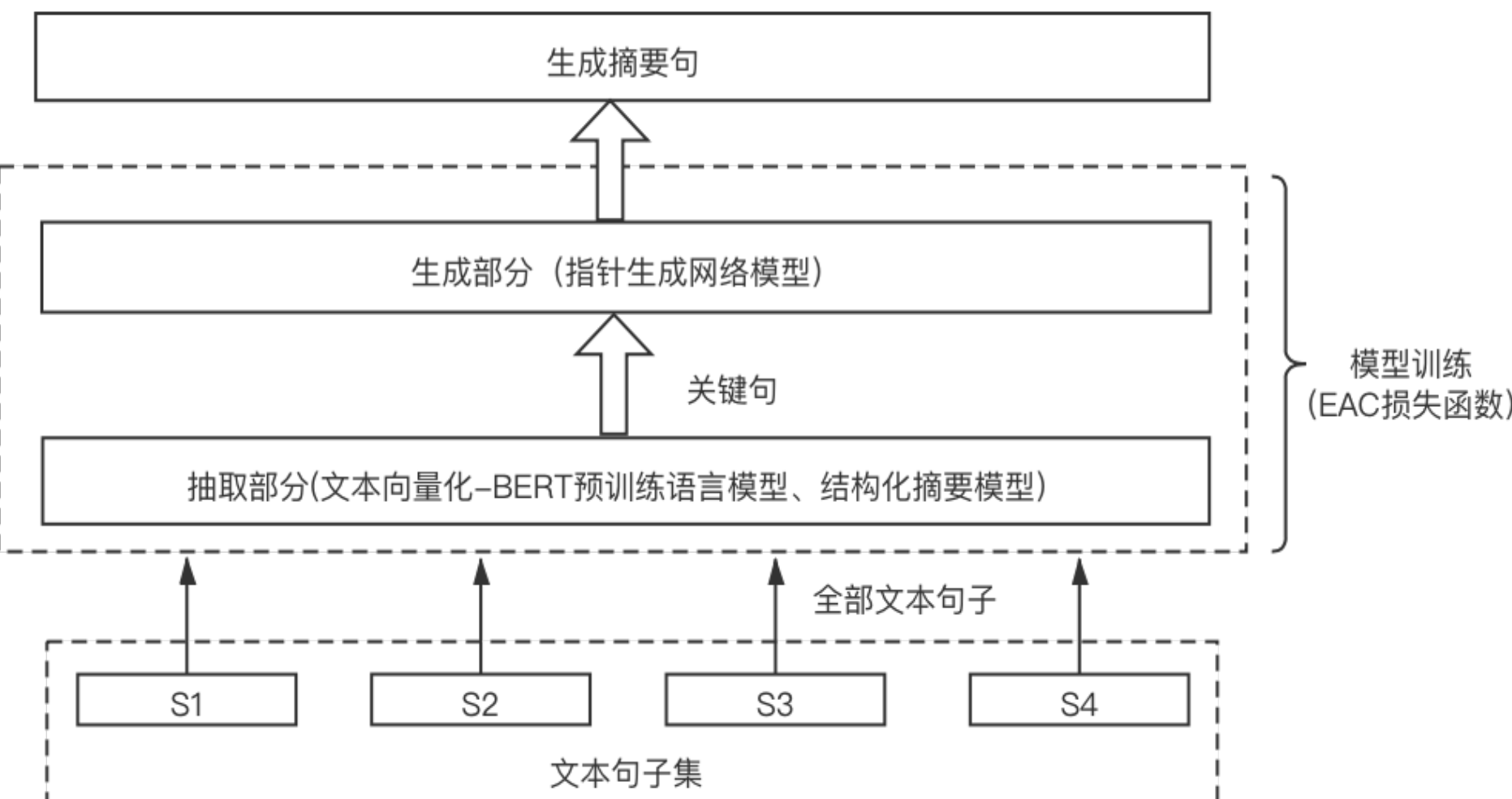
$$L_{eac} = \mu_1 L_{ex} + \mu_2 L_{ab} + \mu_3 L_{cov}$$

摘要模型对比

本文提出的BERT-SUMOPN模型（结合coverage机制）在BIGPATENT数据集摘要的ROUGE评分上效果最好，ROUGE-1和ROUGE-2得分分别提升了3.3%和2.5%。

	Method	ROUGE-1	ROUGE-2	ROUGE-L
	Seq2Seq	27.98	7.64	24.15
	PGNet[*]	30.59	10.01	25.65
	PGNet-coverage[*]	37.12	11.87	32.45
Abs	Point-Less	35.76	11.46	26.36
	Bottom-up	36.49	11.68	27.89
	S2S-ELMo	37.35	11.95	28.75
	TLM+E[*]	38.65	12.31	24.71
	end2end w/o inconsistency loss	38.26	12.15	32.69
Ext+Abs	rnn-ext-abs-RL-rerank	37.86	12.03	29.75
	BERT-SUMOPN(ours)	38.72	12.43	24.58
	BERT-SUMOPN+coverage(ours)	41.99	14.85	33.17

从摘要结果可以看出，BERT-SUMOPN模型生成的摘要包含更多关键信息，对原文的理解更加充分，生成的摘要内容更加准确、流畅，具有较好的逻辑性。

论文结论

本文提出了一种基于BERT-SUMOPN模型的抽取-生成式摘要方法，使用BERT预训练模型进行文本向量化，将抽取式模型抽取的关键句输入到生成式模型中获取摘要，并结合coverage机制减少生成重复。实验表明，该模型在BIGPATENT数据集上取得了很好的效果。

在未来的工作中，我们将继续尝试利用其他方法（强化学习等）对抽取式模型和生成式模型进行端到端训练，提升摘要结果的准确性、流畅性。