低资源摘要生成中的查询建模

徐雨默
导师: Mirella Lapata

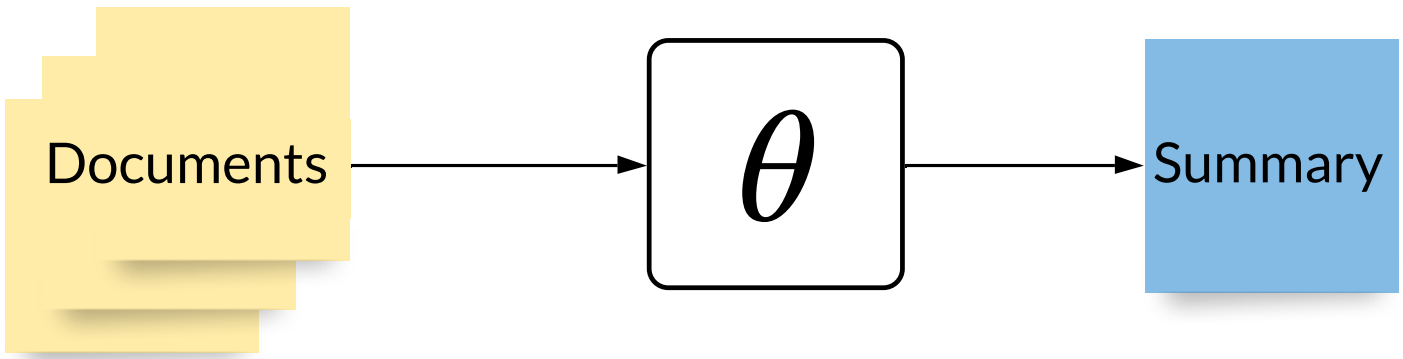
爱丁堡大学信息学院
语言、认知和计算研究所

大纲

1. 研究背景
2. 解决资源稀缺：基于掩码表示的查询建模
3. 面向高扩展性：基于隐变量的查询建模
4. 总结与展望

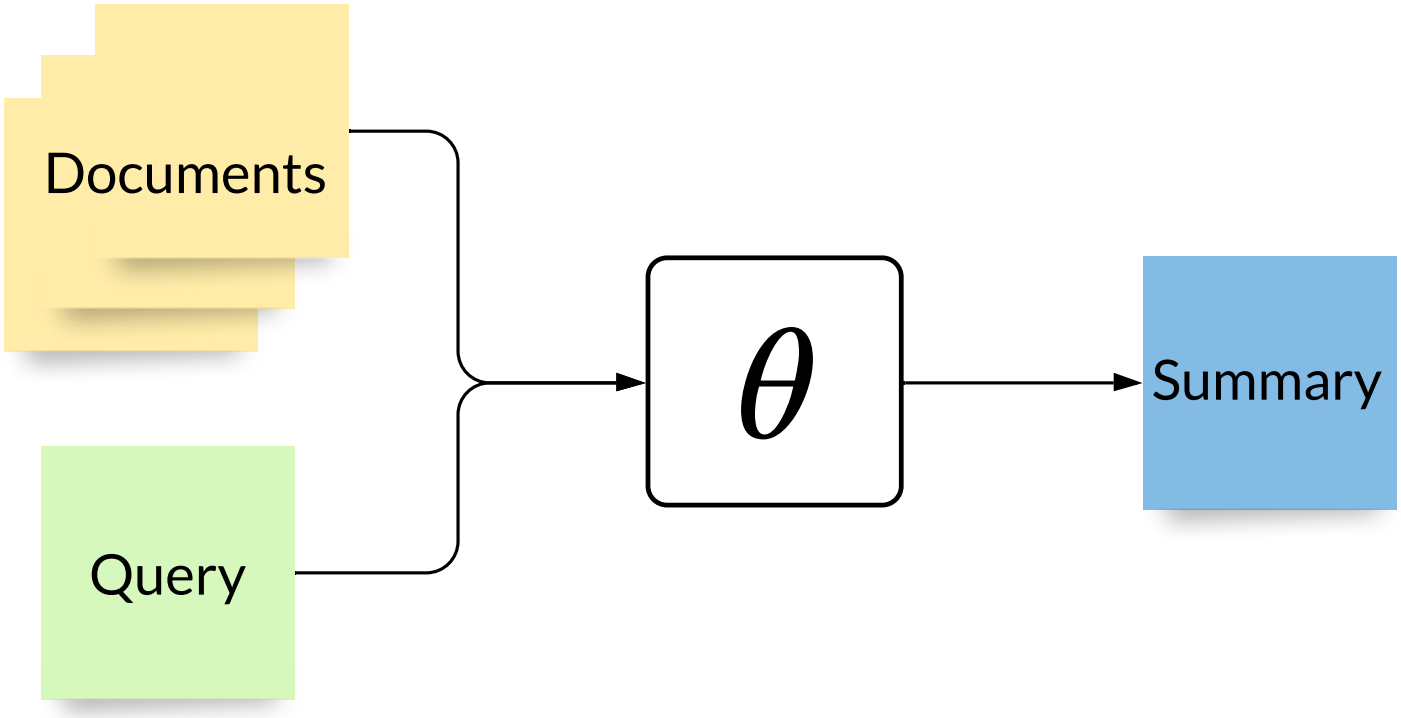
大纲

通用型文本摘要 Generic Summarization



- 抽取式: 摘要中的句子来自于原文
- 生成式: 可用原文不存在的词汇和短语

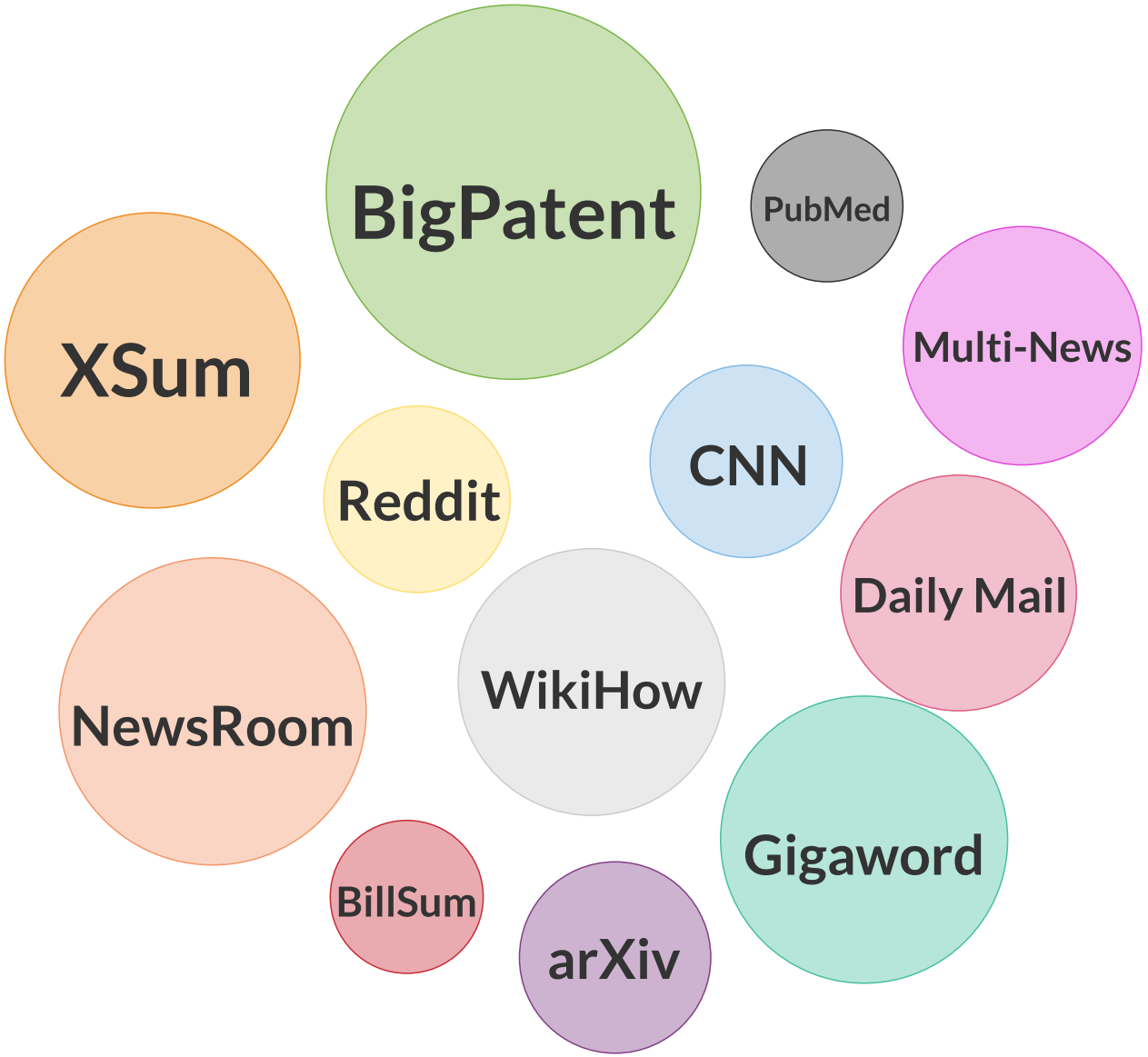
查询型文本摘要 Query Focused Summarization (QFS)



Example in DUC 2005

Title: Amnesty International
Narrative: What is the scope of operations of Amnesty International and what are the international reactions to its activities?

摘要资源：通用型

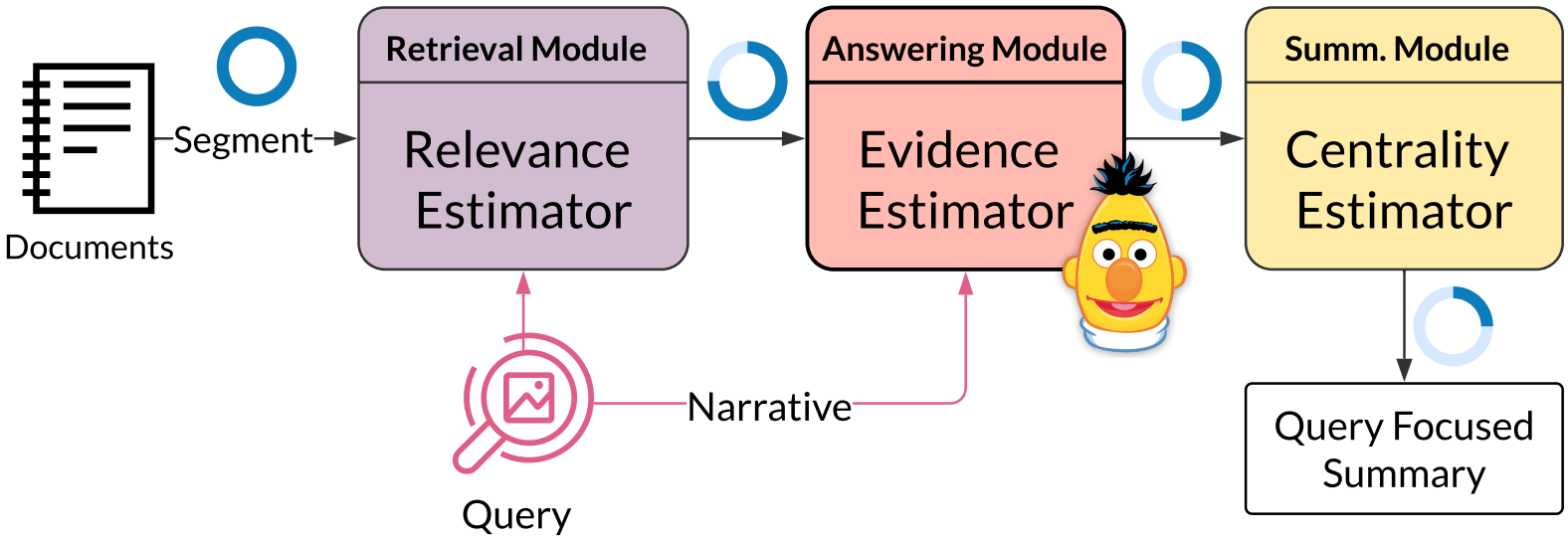

BigPatent
PubMed
XSum
Multi-News
CNN
Reddit
Daily Mail
NewsRoom
WikiHow
Gigaword
BillSum
arXiv

摘要资源：查询型

No large QFS training data!

大纲

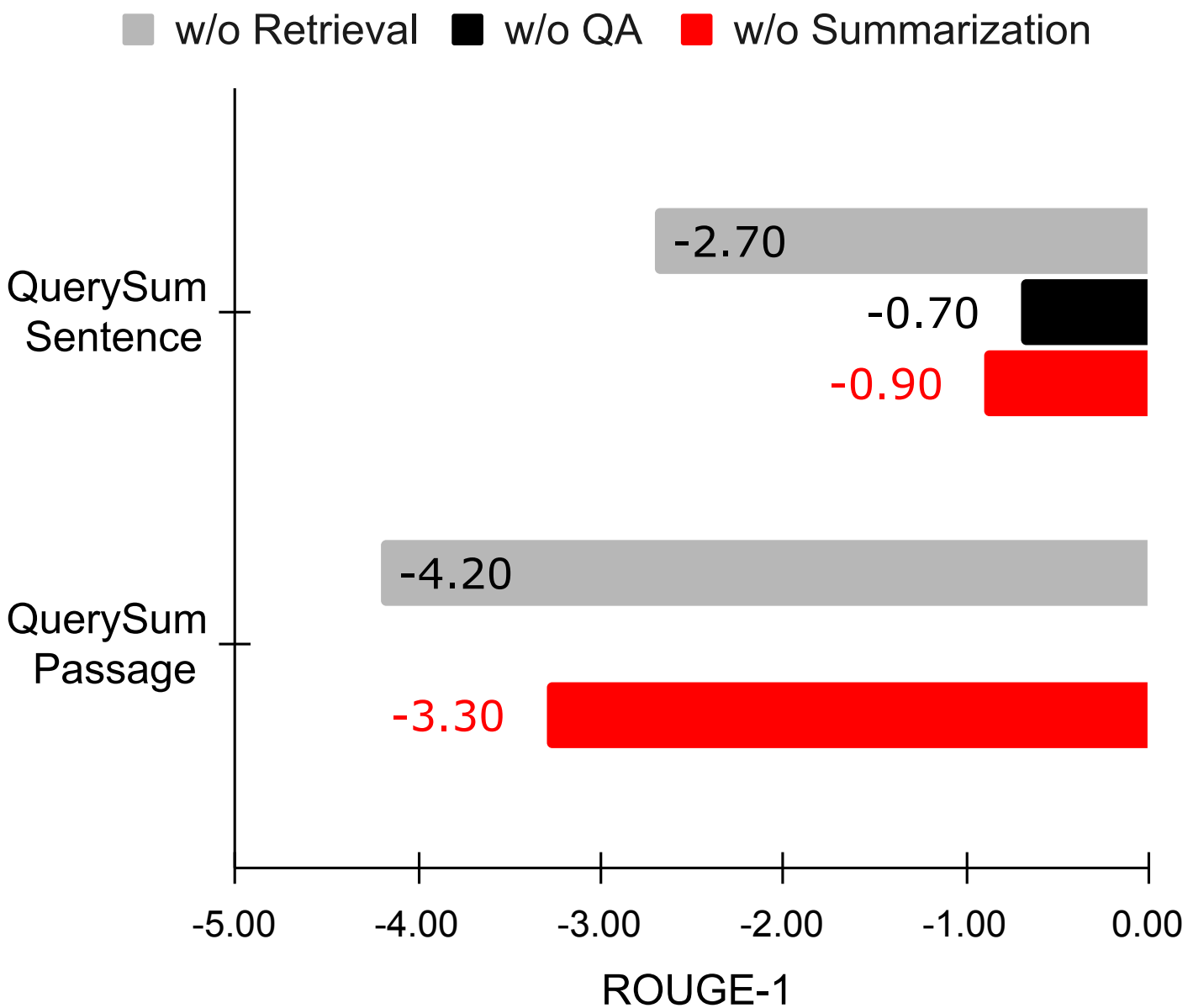
以问答资源作为远程监督 (Xu and Lapata, EMNLP'20)



QuerySum: 解耦相关性、证据性和中心性

- 用检索、问答和摘要三个相对独立的模块对输入逐层过滤
- 对输入的文档数目与大小不敏感
- 利用现存的问答资源完成更准确的语义匹配

QuerySum 消融实验



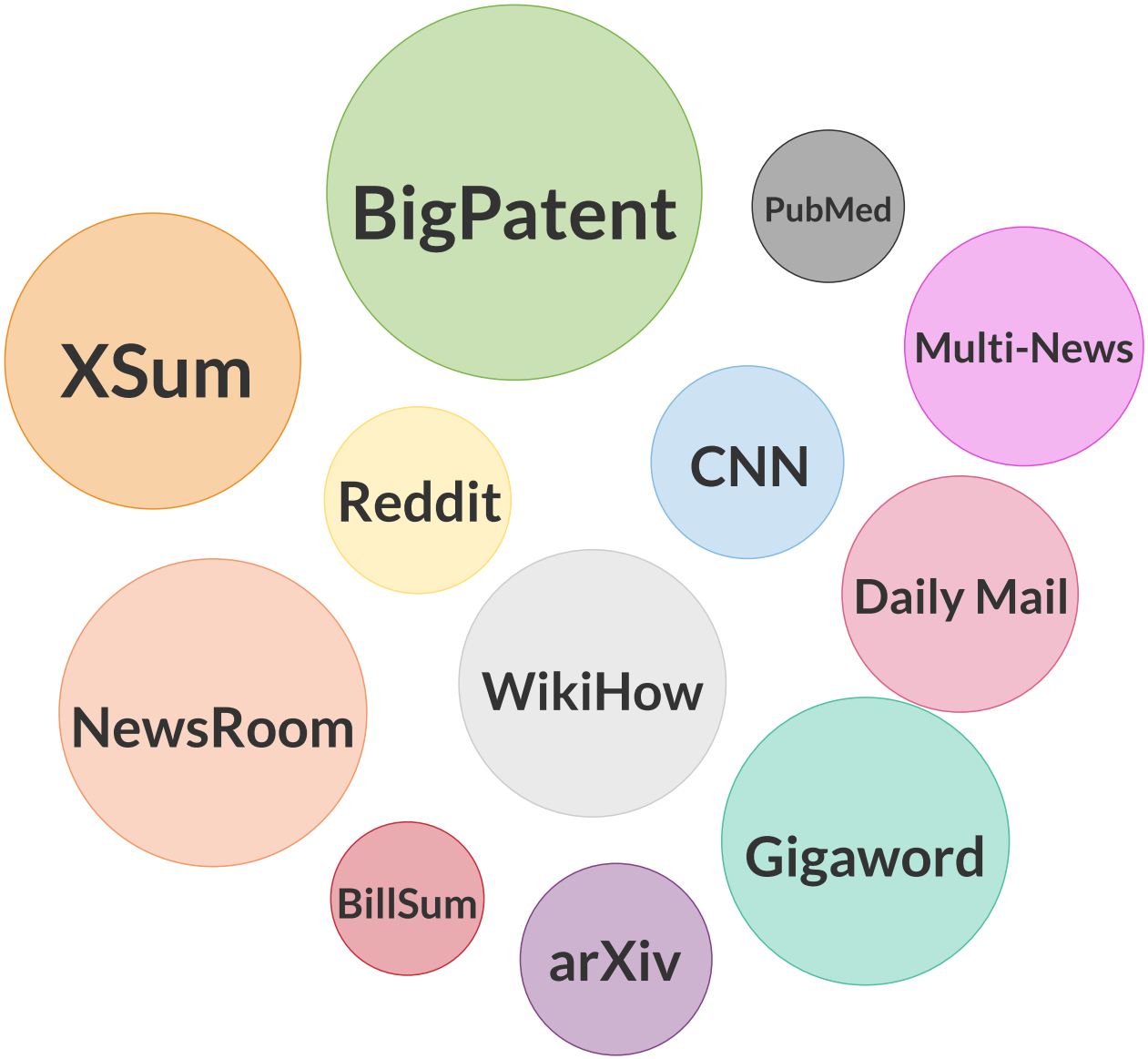
依赖问答资源的弊端

- 分布转移: 问答资源中的问题与查询型摘要中的查询，在分布上常常并不一致
- 标注成本: 问答资源的标注成本可以非常之高 (Bajaj et al., 2016; Kwiatkowski et al., 2019)
- 可获得性: 为任意领域或主题的查询去寻找合适的问答资源在实际上不可行

研究问题 (Xu and Lapata, ACL'21)

能否在不依赖查询相关的训练资源的条件下，
构建一个查询式摘要的生成系统?

摘要资源：通用型

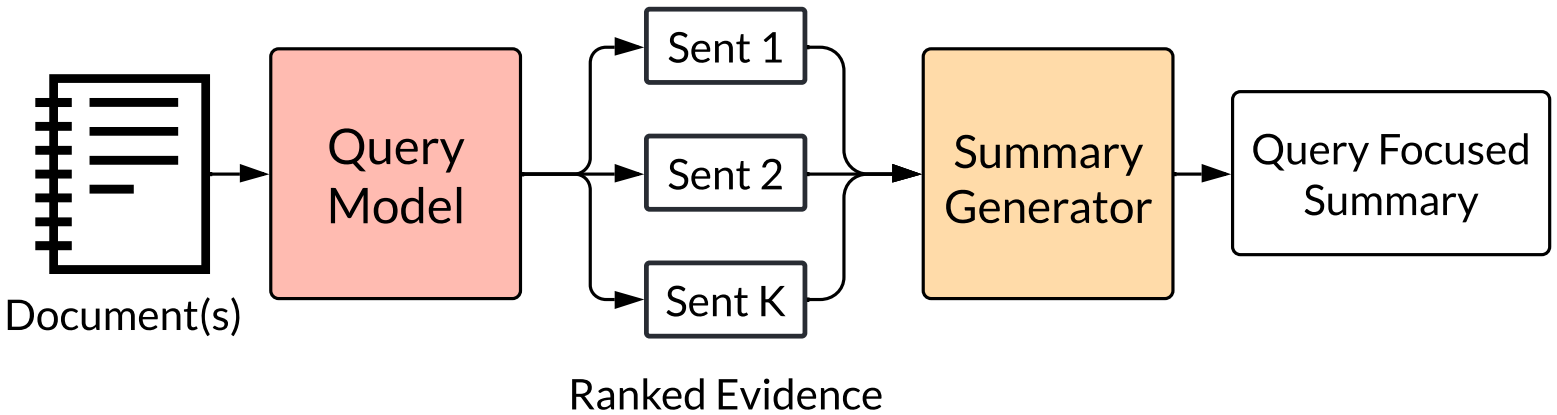

BigPatent
PubMed
XSum
Multi-News
CNN
Reddit
Daily Mail
NewsRoom
WikiHow
Gigaword
BillSum
arXiv

研究问题 (Xu and Lapata, ACL'21)

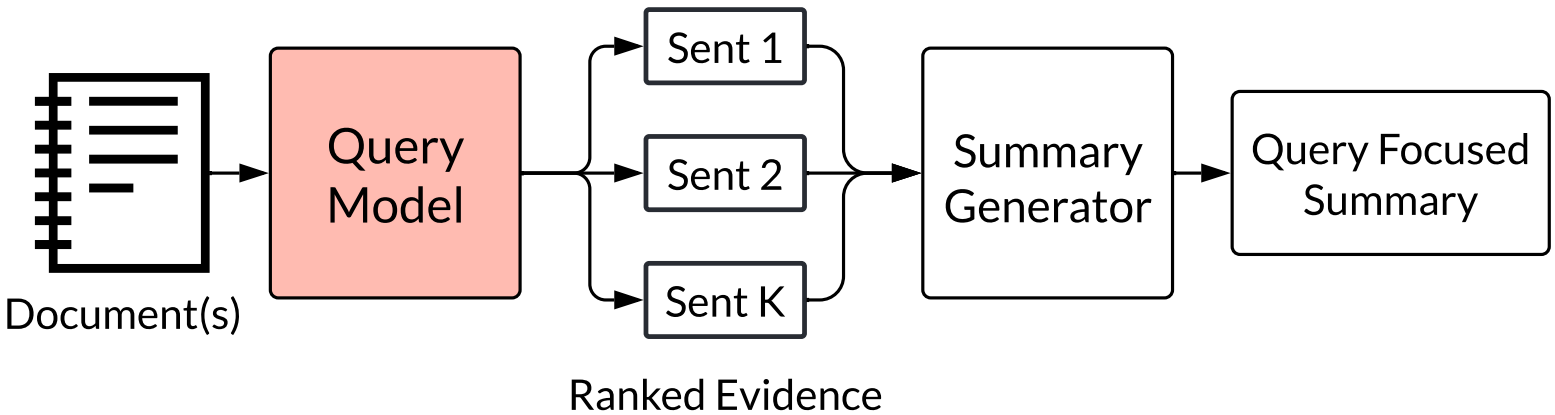
能否在不依赖查询相关的训练资源的条件下，
构建一个查询式摘要的生成系统?

使用 通用型摘要 的数据!

任务解构 (Xu and Lapata, ACL'21)



查询模型 (Xu and Lapata, ACL'21)



- 对原文句子进行查询相关性的排序学习
- 训练集中没有查询，且没有排序的监督信号

查询模型 (Xu and Lapata, ACL'21)

- 假设:

$$\text{(潜在的) 查询} \xrightarrow{\text{生成}} \text{(可观测的) 通用型摘要}$$

- 如何从通用型摘要中 逆向工程 出查询?

$$\text{(潜在的) 查询} \xleftarrow{\text{逆向工程}} \text{(可观测的) 通用型摘要}$$

- 基于掩码 (Masking) 的逆向工程

训练集：掩码处理摘要（生成代理查询）

- The Da Vinci Code was published in 2003, and within six years Brown had booted John Grisham from the No. 1 slot on the list of writers whose books were most often donated to Oxfam's 700 shops.
- The Independent in 2012 reported Brown's best-seller was the most-donated book for the fourth year running.

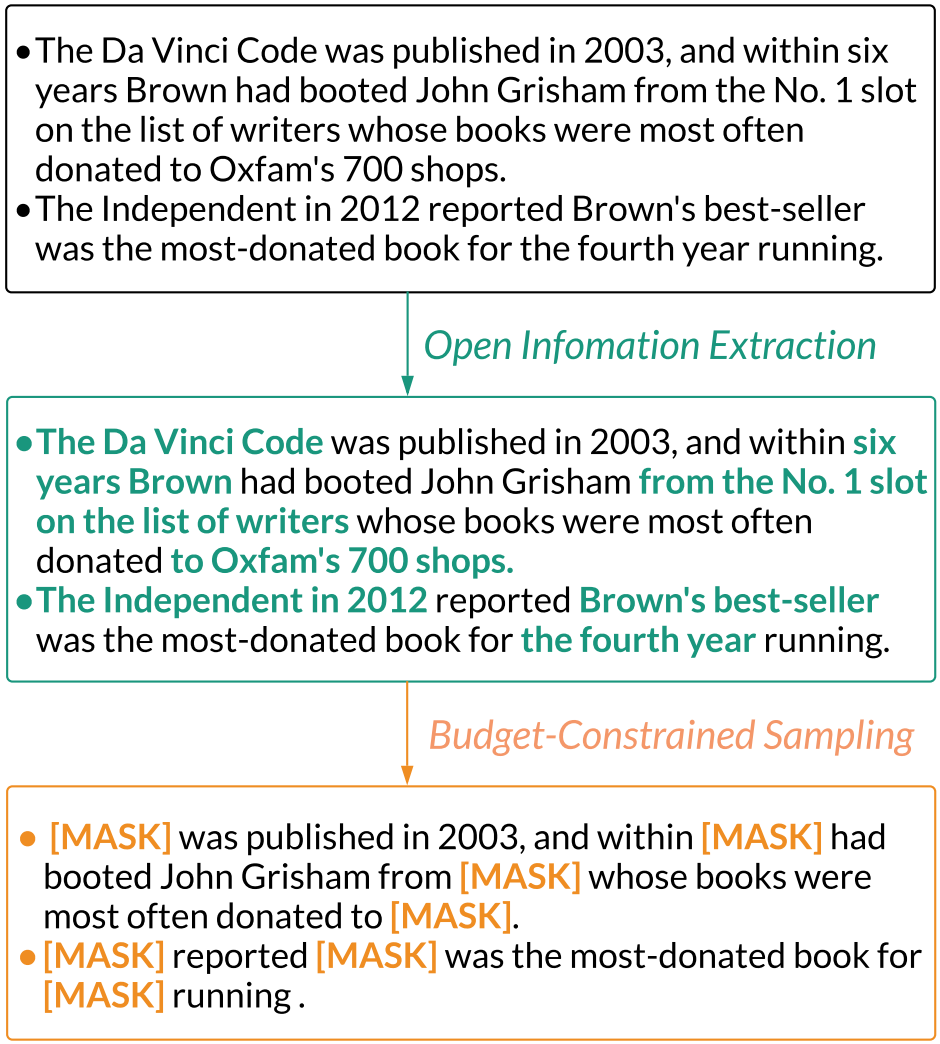
训练集：掩码处理摘要（生成代理查询）

- The Da Vinci Code was published in 2003, and within six years Brown had booted John Grisham from the No. 1 slot on the list of writers whose books were most often donated to Oxfam's 700 shops.
- The Independent in 2012 reported Brown's best-seller was the most-donated book for the fourth year running.

Open Infomation Extraction

- **The Da Vinci Code** was published in 2003, and within **six years Brown** had booted John Grisham **from the No. 1 slot on the list of writers** whose books were most often donated **to Oxfam's 700 shops.**
- **The Independent in 2012** reported **Brown's best-seller** was the most-donated book for **the fourth year** running.

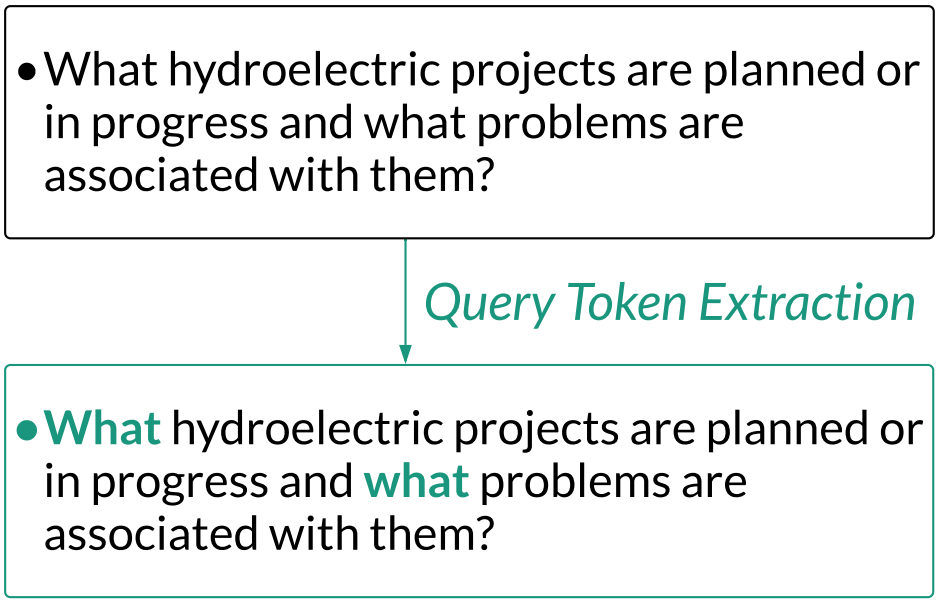
训练集：掩码处理摘要（生成代理查询）



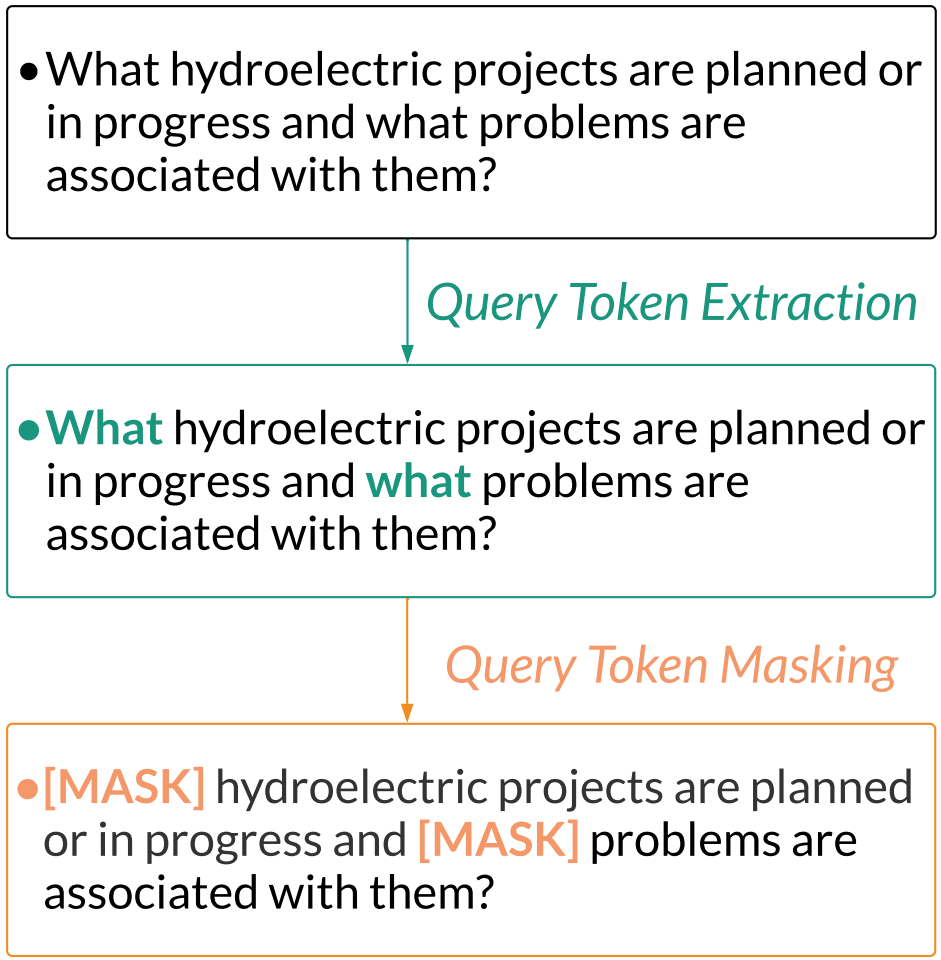
测试集：掩码处理真实查询

- What hydroelectric projects are planned or in progress and what problems are associated with them?

测试集：掩码处理真实查询



测试集：掩码处理真实查询



查询模型：MaGRE

- MaRGE: **Ma**sked **R**OU**GE** Regression model
- 训练:

$$\{\text{掩码摘要（代理查询）}, \text{句子}\} \xrightarrow{\theta} \text{ROUGE}$$

- 测试:

$$\{\text{掩码查询}, \text{句子}\} \xrightarrow{\theta} \text{相关性估计}$$

取前k个高相关的句子作为摘要生成的输入

实验配置

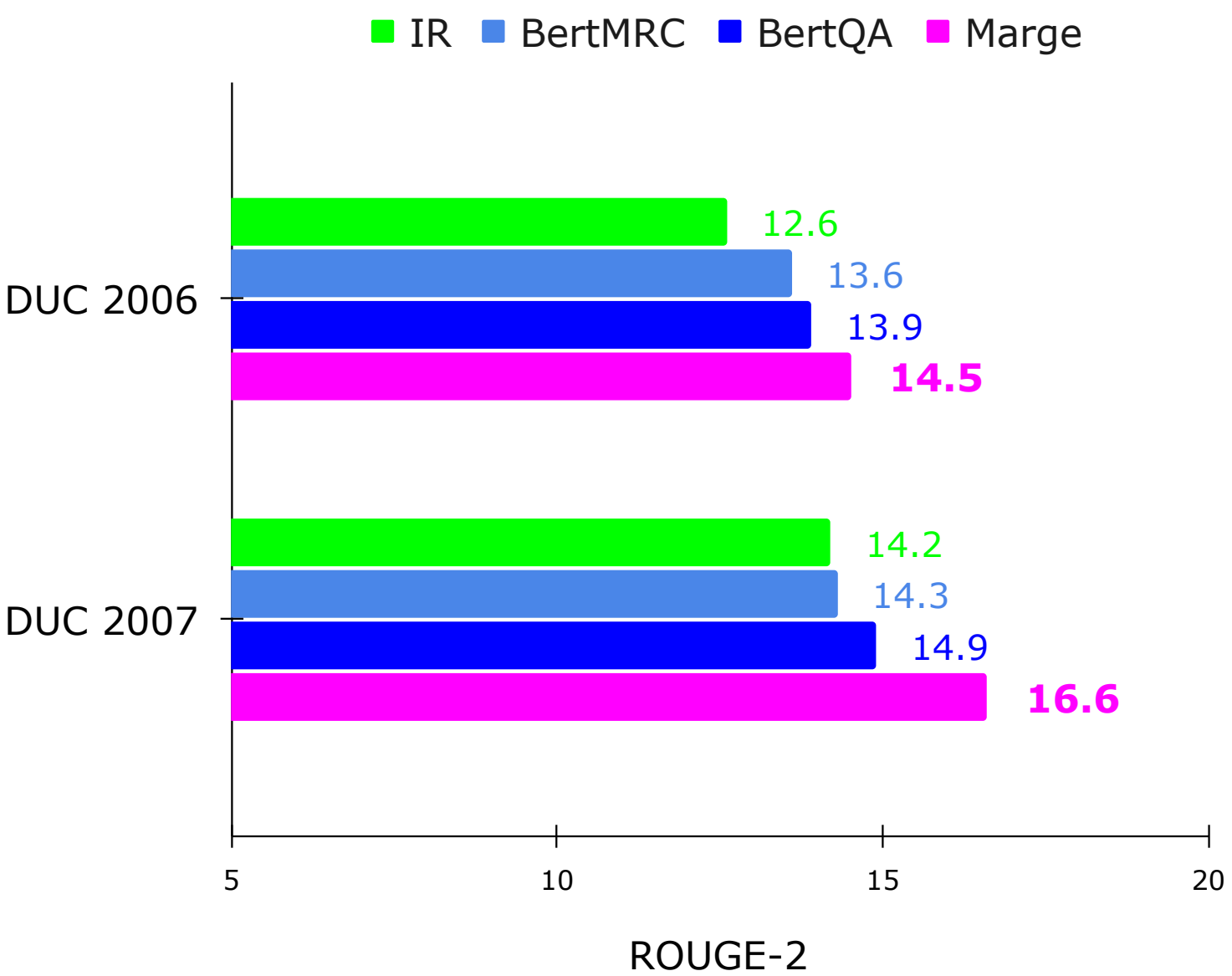
模型训练: 通用型摘要数据集

- 多文档: Multi-News (Fabbri et al., 2019)
- 单文档: CNN/DM (Hermann et al., 2015)

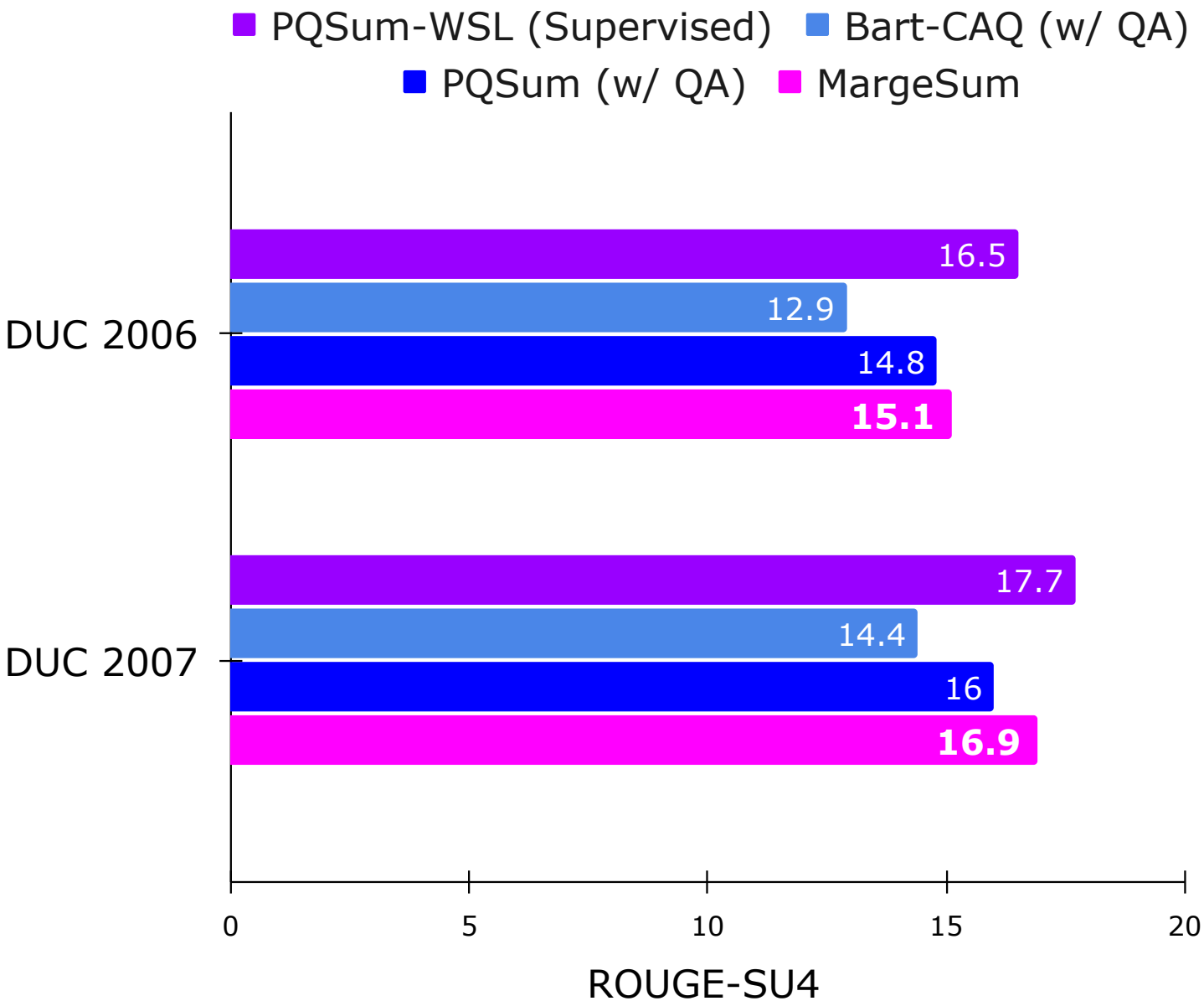
模型评估: 查询型摘要数据集

- 开发: DUC 2005
- 测试: DUC 2006-2007, TD-QFS

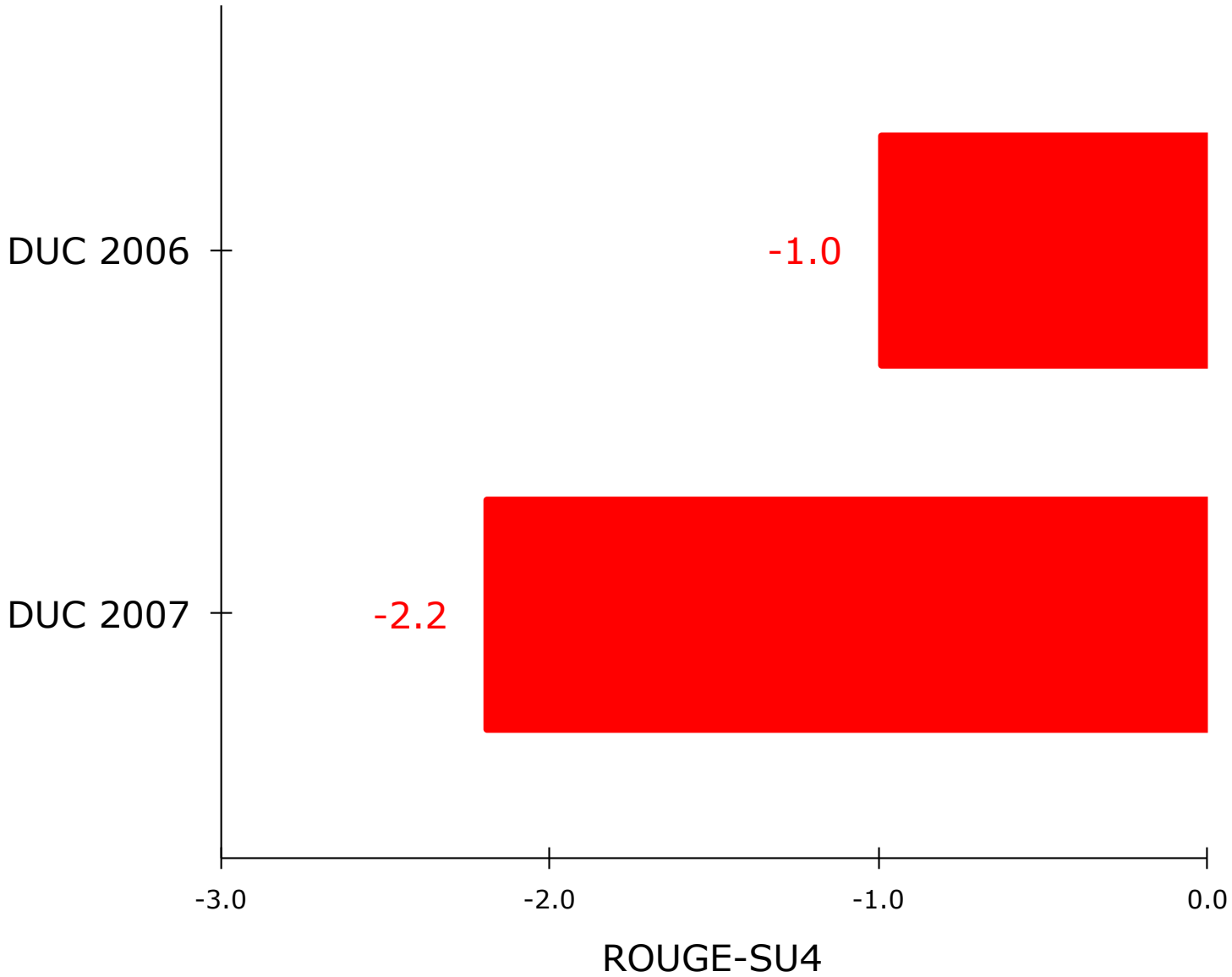
比较实验：查询建模


IR
BertMRC
BertQA
Marge
DUC 2006
12.6
13.6
13.9
14.5
DUC 2007
14.2
14.3
14.9
16.6
5
10
15
20
ROUGE-2

比较实验: 摘要生成


PQSum-WSL (Supervised)
Bart-CAQ (w/ QA)
PQSum (w/ QA)
MargeSum
DUC 2006
16.5
12.9
14.8
15.1
DUC 2007
17.7
14.4
16
16.9
0
5
10
15
20
ROUGE-SU4

消融实验：用 BertQA 取代 MaRGE



大纲

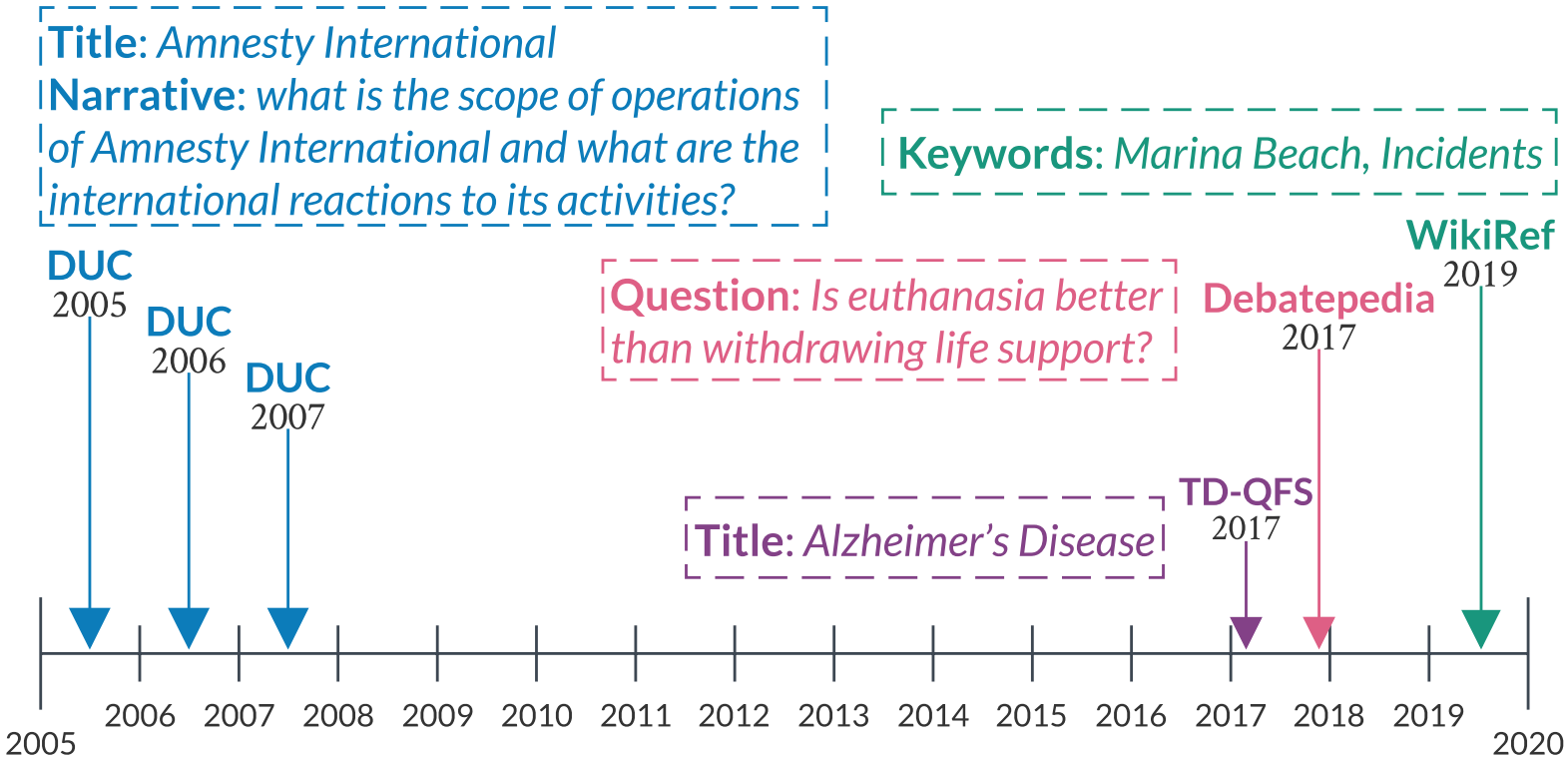
对查询语言的重新思考

Example in DUC 2005

Title: Amnesty International
Narrative: What is the scope of operations of Amnesty International and what are the international reactions to its activities?

- 目前为止，我们讨论的是由 DUC 基准下的查询语言
- 现有查询模型的表现强依赖于查询语言的形式
- 这是不是我们需要关心的唯一查询语言的形式呢?

查询语言的历史变化



通用型摘要可看做查询型摘要的特例：查询为空

研究问题 (Xu and Lapata, TACL'22)

如何构造一种摘要系统：
在不用重新训练的情况下，
即可处理多种查询语言的形式，
包括空查询？

解决思路

现有的工作：在词汇空间中建模查询（生成或掩码）

- 假设先验知识：我们需要有查询形式的先验知识（如一个开发集），才能去生成同分布的查询
- 缺乏扩展性：对特定的查询形式形成依赖

另一种思路：在表示空间中建模查询（掩码）

- 直觉：带着问题去读文本时，会形成一种存在在表示空间中的视图
- 将查询看做一种输入文档的视图

生成模型 (Xu and Lapata, TACL'22)

- LQSum: Document **Sum**marization with **L**atent **Q**ueries
- 定义 潜在查询 (latent query) 为 离散隐变量 $\mathbf{z}$:

$$\mathbf{x}_i \rightarrow \mathbf{z}_i$$

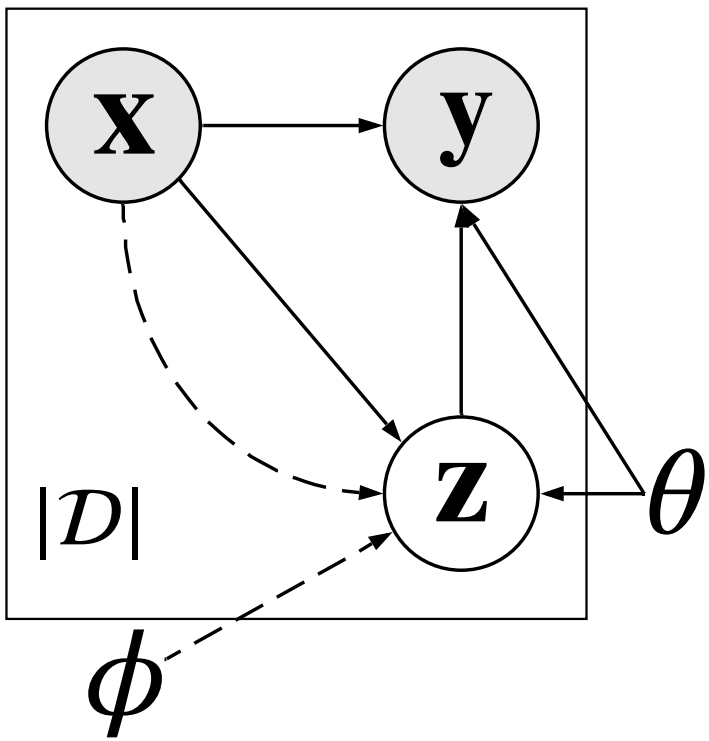
$p(\mathbf{z}_i|\mathbf{x}_i)$: 文档词 $\mathbf{x}_i$ 为查询词的概率

- 测试: 建模给定的可观测查询为离散观测变量 $\tilde{\mathbf{z}}$:

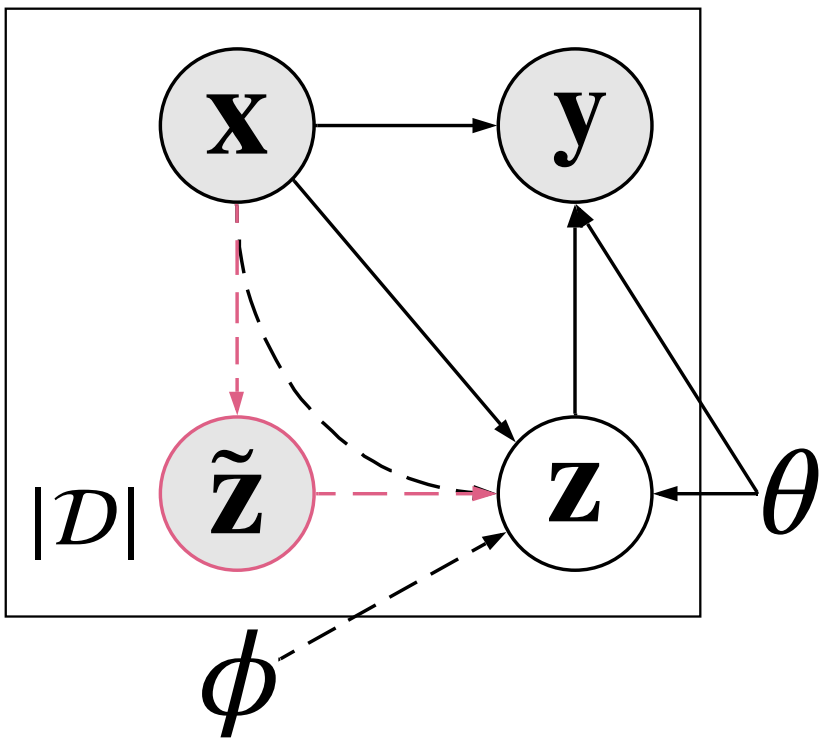
$$\{\mathbf{x}_i, \tilde{\mathbf{z}}_i\} \rightarrow \mathbf{z}_i$$

查询相关的隐变量和观测变量均定义在词级别

生成模型 (Xu and Lapata, TACL'22)



(a) Generative Process: Training



(b) Generative Process: Testing

推断模型 (Xu and Lapata, TACL'22)

- 变分推断：用变分后验 $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ 估计真实后验 $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y})$
- 后验正则：强制模型学习到有意义的隐变量

$$q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) \rightarrow o(\hat{\mathbf{z}}|\mathbf{x},\mathbf{y}) \tag{1}$$

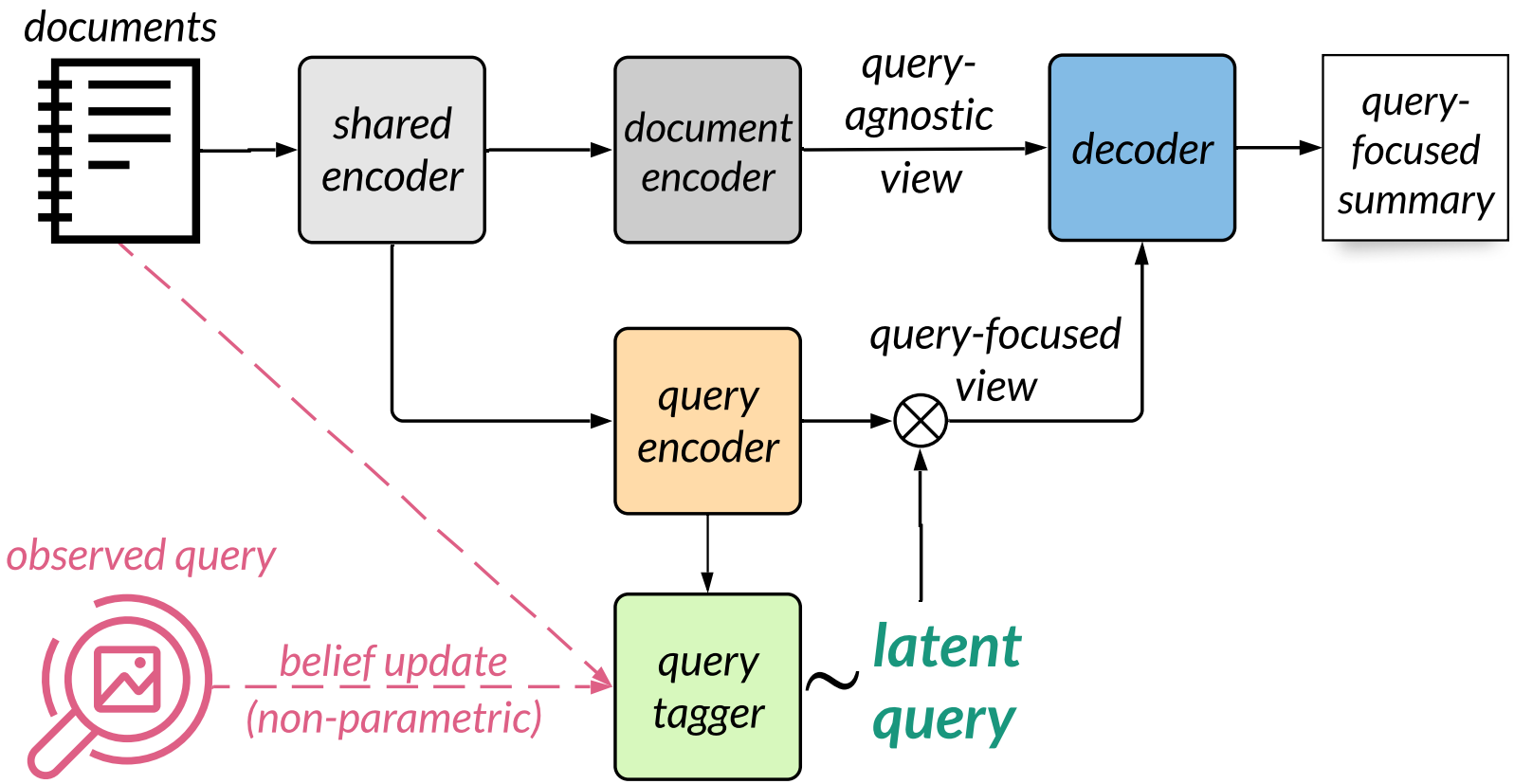
$o(\hat{\mathbf{z}}|\mathbf{x},\mathbf{y})$：序列标注；最长公共 BPE 字串的弱监督

- 重写训练目标：

$$\mathcal{L} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})\right]}_{\text{conditional language modeling}} + \underbrace{\beta\mathcal{H}\left(q_\phi(\mathbf{z}|\mathbf{x})\right) - \omega\mathcal{H}\left(o(\hat{\mathbf{z}}|\mathbf{x},\mathbf{y}), q_\phi(\mathbf{z}|\mathbf{x})\right)}_{\text{latent query modeling}} \tag{2}$$

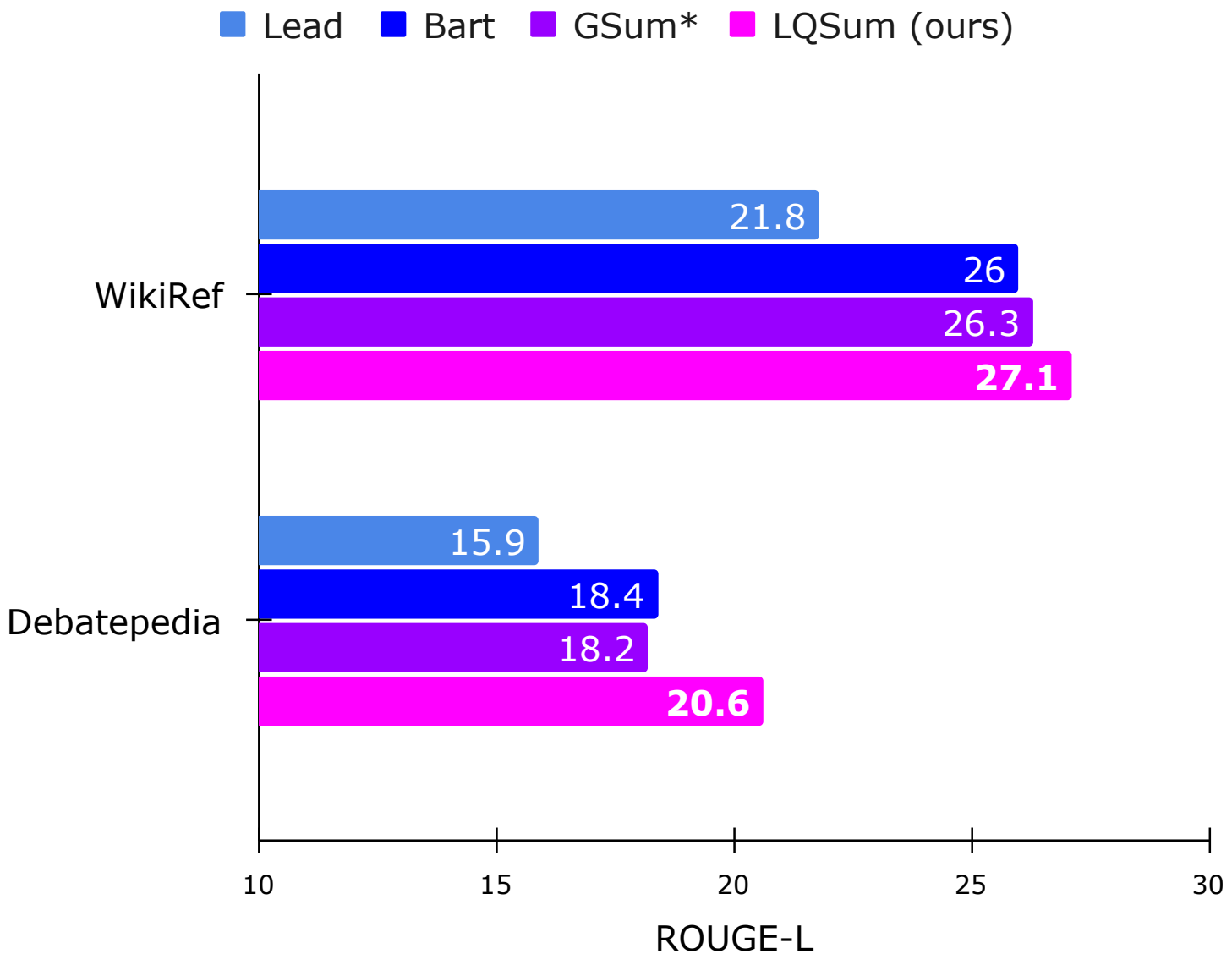
ω: controls the influence from the weak supervision $\hat{\mathbf{z}}$

参数化 (Xu and Lapata, TACL'22)

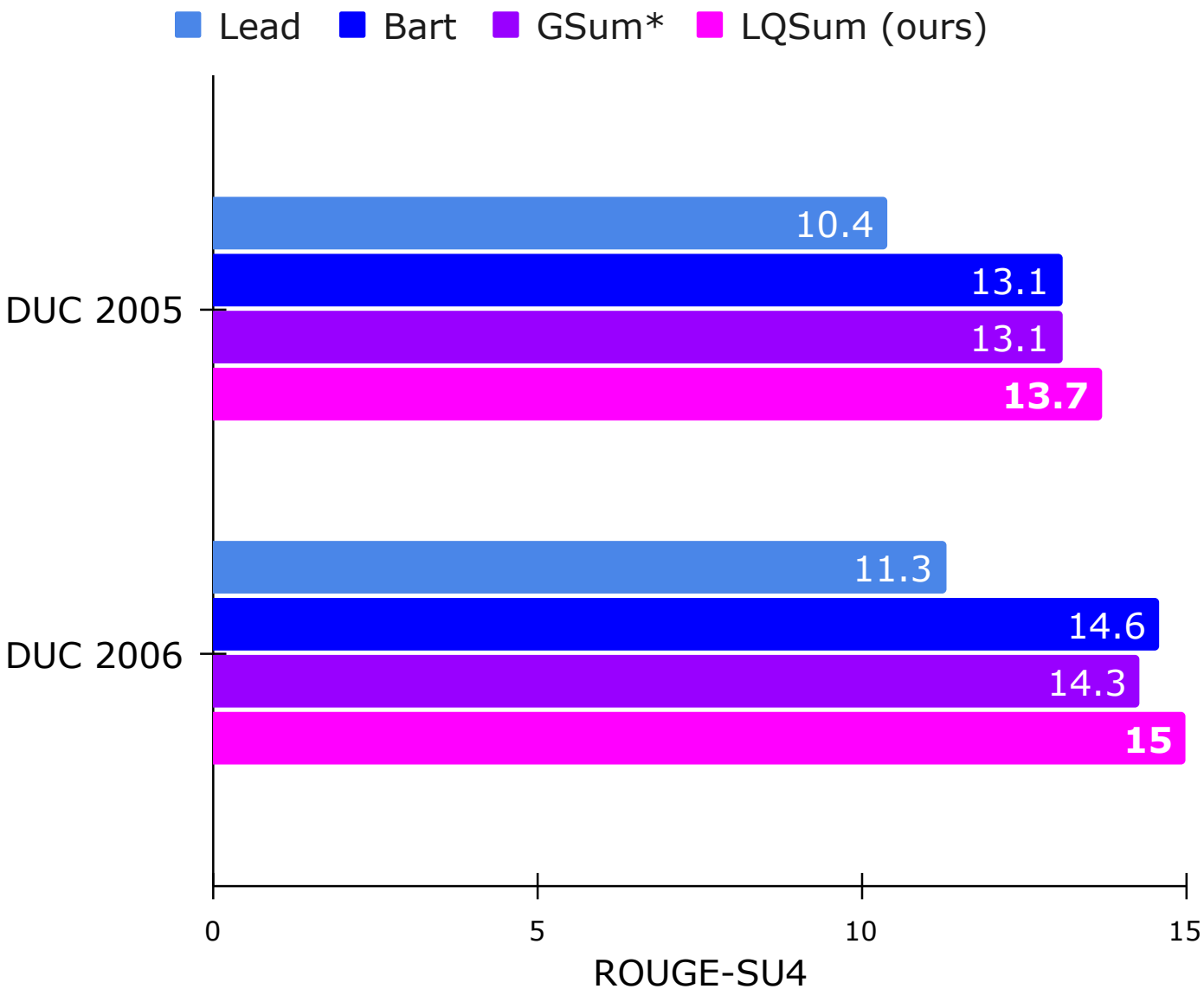


- 一个输入，两种视图
- 通用视图 (query-agnostic)：通用的篇章语义
- 查询视图 (query-focused)：考虑了潜在查询后的篇章语义

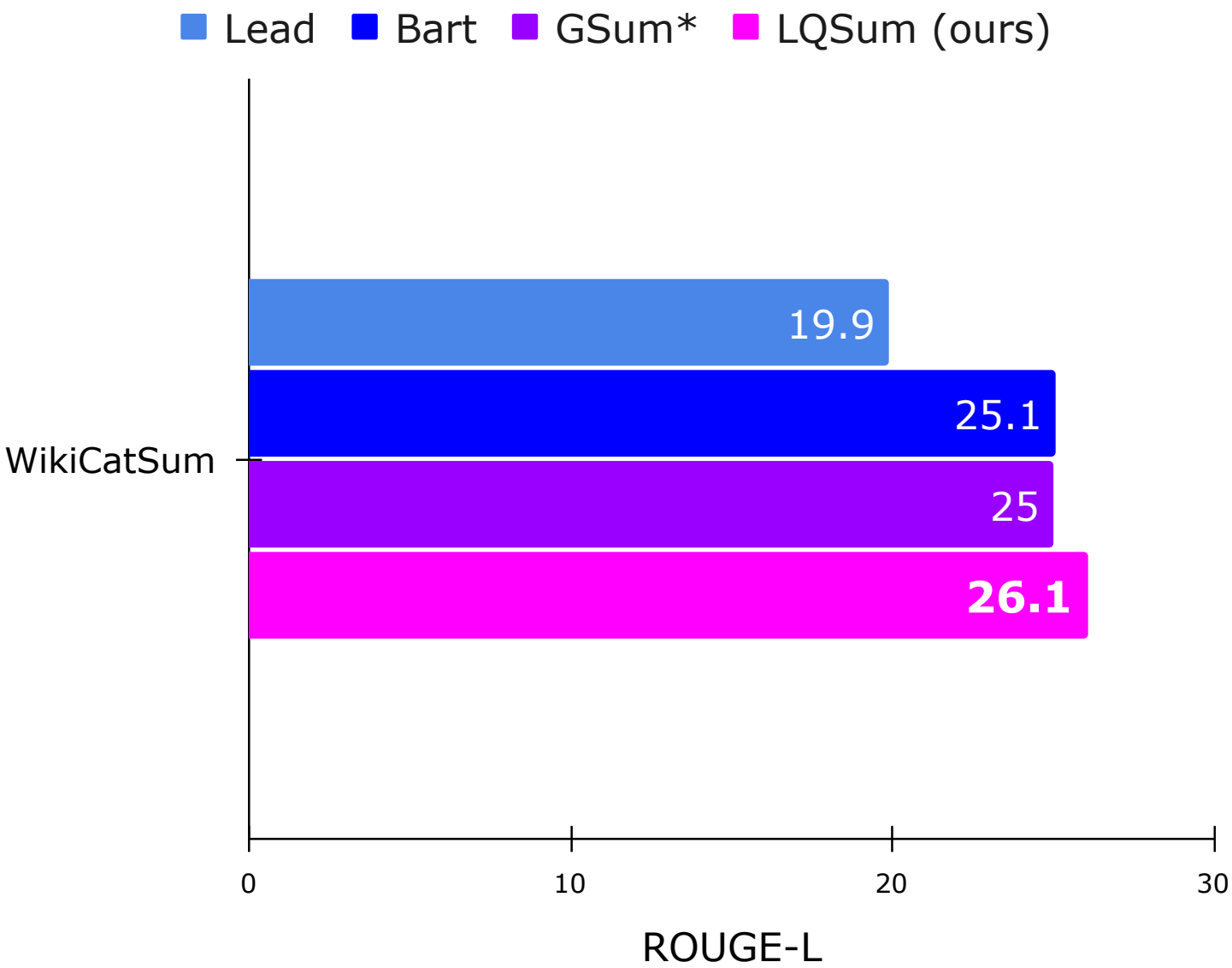
查询型摘要（零监督，单文档）



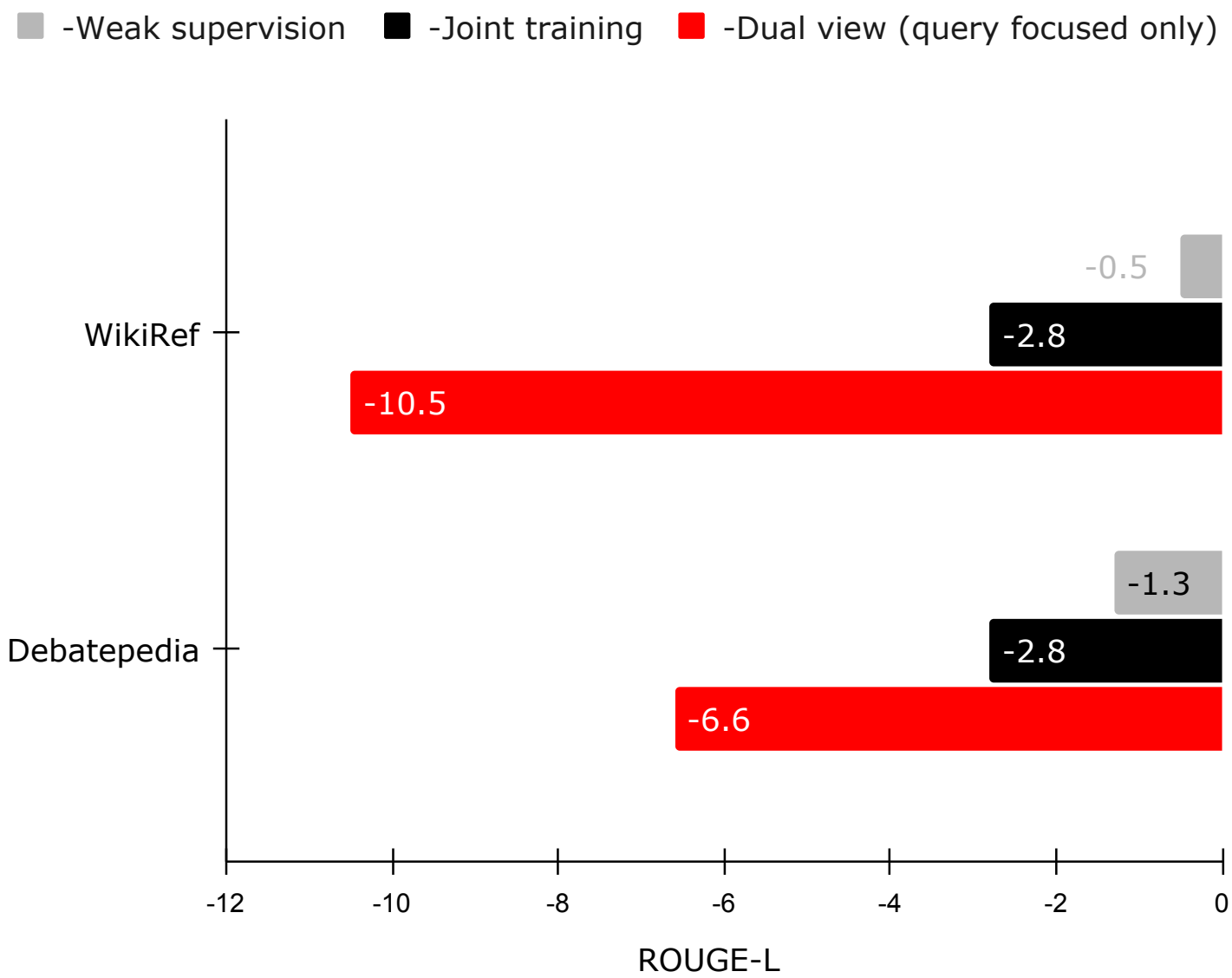
查询型摘要（零监督，多文档）



通用型摘要（零监督，多文档）



消融实验


-Weak supervision
-Joint training
-Dual view (query focused only)
WikiRef
Debatepedia
-0.5
-2.8
-10.5
-1.3
-2.8
-6.6
-12
-10
-8
-6
-4
-2
0
ROUGE-L

大纲

总结

1. 摘要任务最初提出的动机是为了提高人们信息获取的效率。作为面向用户的系统，其设计应当考虑用户的意图与交互，这使得查询建模成为一个重要的摘要子任务
2. 问答资源可以改善查询型摘要中的数据稀缺性，却也引入了额外的标注成本和依赖性
3. 掩码表示可以替代问答的角色，仅使用通用型摘要的数据即可完成查询型摘要的生成
4. 将潜在查询看做隐变量可以同时建模通用型和查询型的摘要，进一步提升摘要系统的健壮性和可扩展性

未来研究方向

- 跨语言 查询型摘要: 面向不同语言背景的用户
- 多模态 查询型摘要: 在文本、语音、视觉信息上查询
- 对话式 查询型摘要: 面向多轮交互的查询建模

感谢！

References:

1. *Coarse-to-Fine Query Focused Multi-Document Summarization*, Xu, Yumo, and Lapata, Mirella, EMNLP 2020
2. *Generating Query Focused Summaries with Query-Free Resources*, Xu, Yumo, and Lapata, Mirella, ACL 2021
3. *Document Summarization with Latent Queries*, Xu, Yumo, and Lapata, Mirella, TACL 2022