跨语言话题分析关键技术研究

学　　生：唐 国 瑜
学　　号：2009310437
导　　师：郑　方　研究员

2014年12月16日

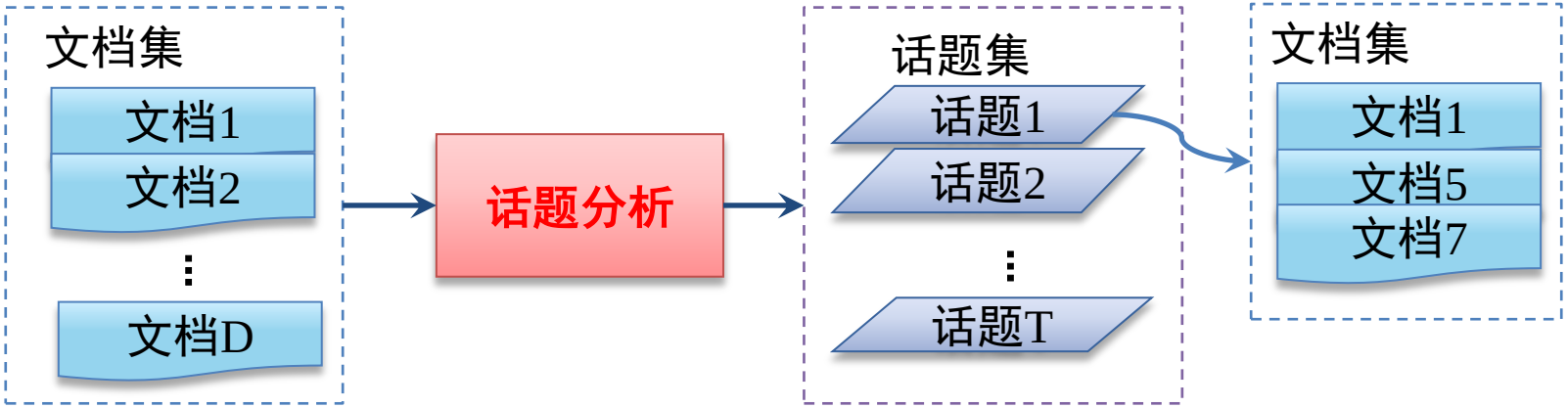
提纲

- **研究背景**
- **研究现状**
- **研究工作**
  - 研究内容及思路
  - 基于全局词义的跨语言文档建模
  - 基于统计词义的主题模型
  - 基于统计词义的跨语言主题模型
- **论文创新性及贡献**
- **研究展望**

- **Part 1: 研究背景**

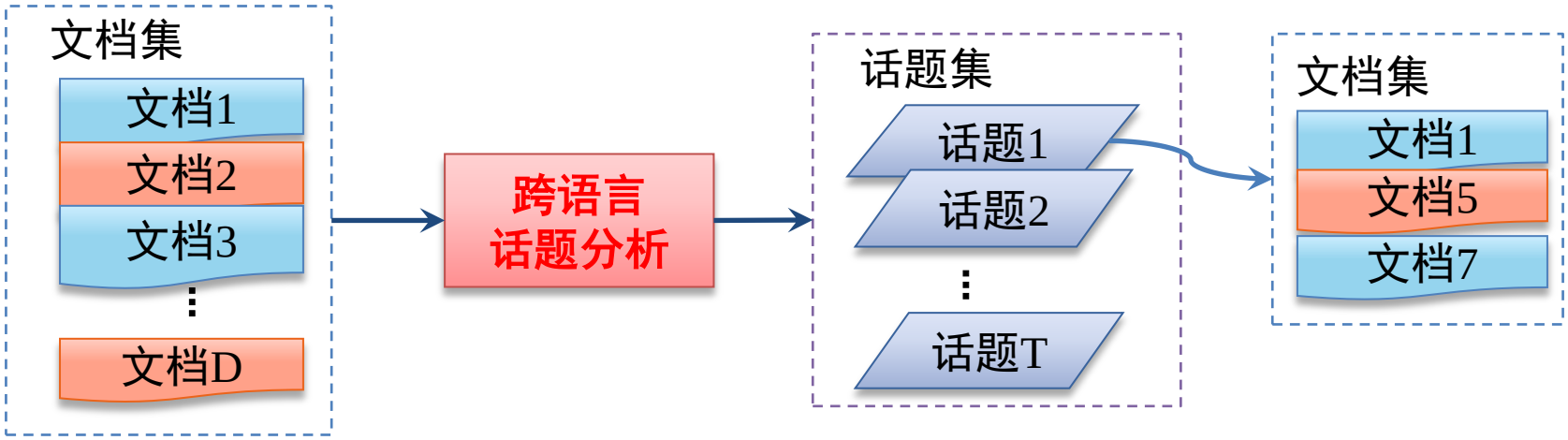
研究背景

- **话题分析问题定义**
  - 给定一个文档集，识别出该文档集中每个文档的话题
- **起源：话题检测与跟踪（TDT）国际评测**
  - 1996年，DARPA，NIST评测
  - 处理对象：传统新闻媒体信息流数据
- **发展：**网页新闻、博客新闻、微博新闻、科技论文
- **应用背景：**网络舆情分析系统的核心模块



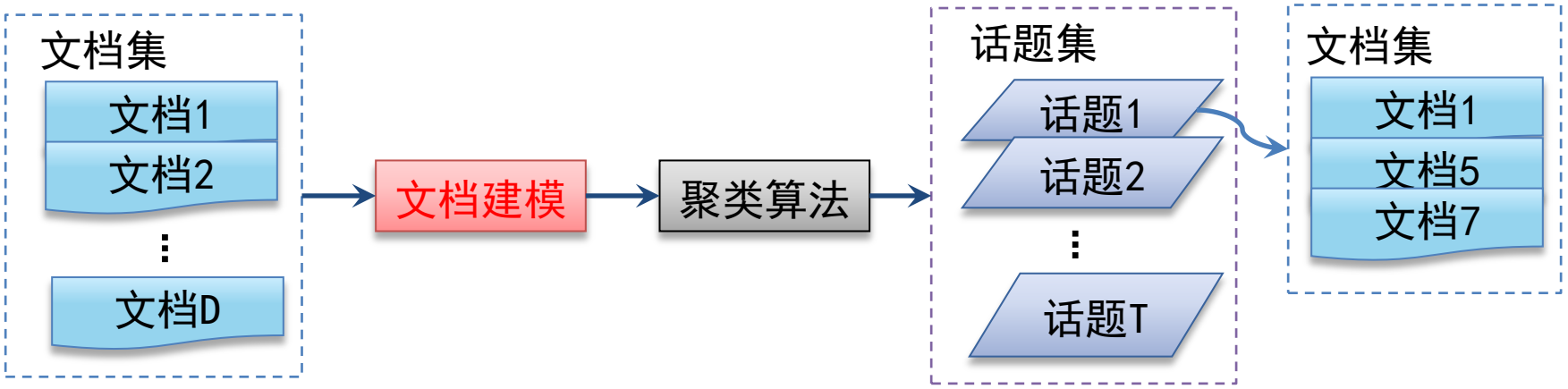
跨语言问题

- **现实意义**
  - 从非母语的互联网内容中获取信息
  - 了解其他国家或地区的报道或评论
- **科学问题**
  - 如何处理多语言文档集（每个文档仅采用一种语言）
  - 如何识别跨语言话题（一个话题可以含有不同语言的文档）



- **Part 2: 研究现状**

单语言话题分析



- 传统文档建模
  - 向量空间模型
  - 语言模型
  - 缺点：无法解决同义词和多义词问题

➔基于语义空间的文档建模

基于语义空间的文档建模

- **基于显语义空间的文档建模**
  - Wikipedia(Gabrilovich et al., 2007)；WordNet(Hotho et al., 2003)；HowNet(Huang et al., 2010)
  - **缺点**：概念需要人工定义，受到语义资源规模和更新速度的影响；粒度不一，难以形成符合实际的概念划分

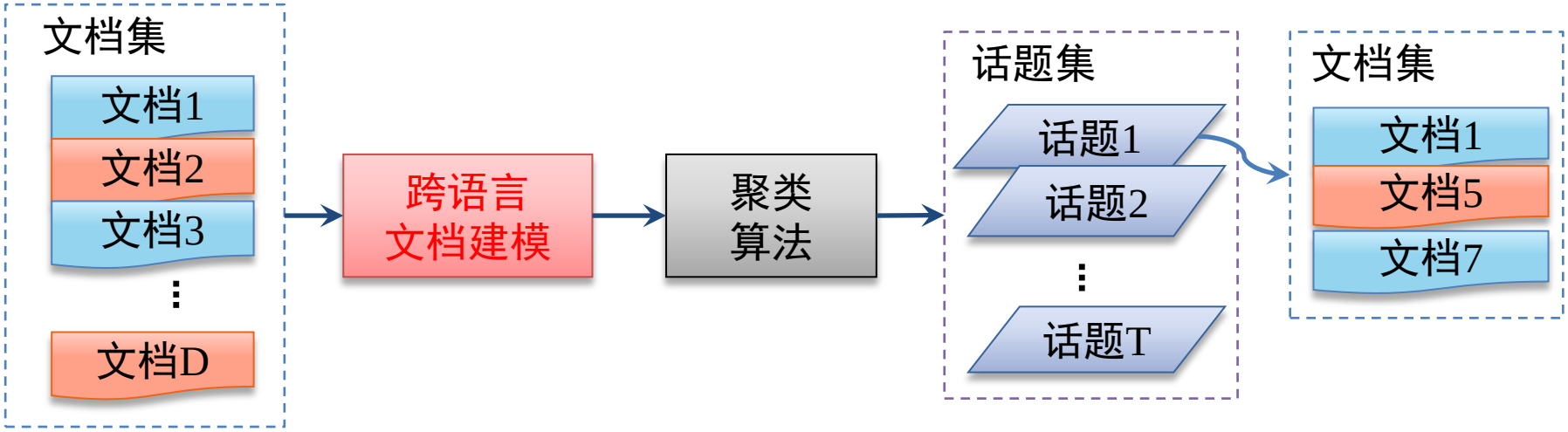
- **基于统计语义空间的文档建模**
  - （见下一页）

基于统计语义空间的文档建模

- **主题模型**
  - LSA (Landauer and Dumais, 1997); PLSA (Hofmann,1999); LDA(Blei et al.,2003)
  - **缺点**：从文档层面解决同义词和多义词问题，效果并不理想
- **词关系**
  - 广义向量空间模型（Wong et al.,1985; Farahat and Kamel, 2011）；基于词类簇（Pessiot et al.,2010）
  - **缺点**：无法解决多义词问题

- 基于统计词义
  - （Navigli and Grisafulli,2010）
  - **缺点**：只考虑到一个词的多义现象，没有考虑到词与词之间的关系。

跨语言话题分析



- 基于直接对应的文档建模
  - 机器翻译(Leek et al.,1999)
  - 缺点：受到机器翻译性能和效率的影响
- 词对应
  - (Chen and Lin, 2000; Pouliquen et al.,2004)
  - 缺点：翻译歧义

➔ 基于语义空间的跨语言文档建模

基于语义空间的跨语言文档建模

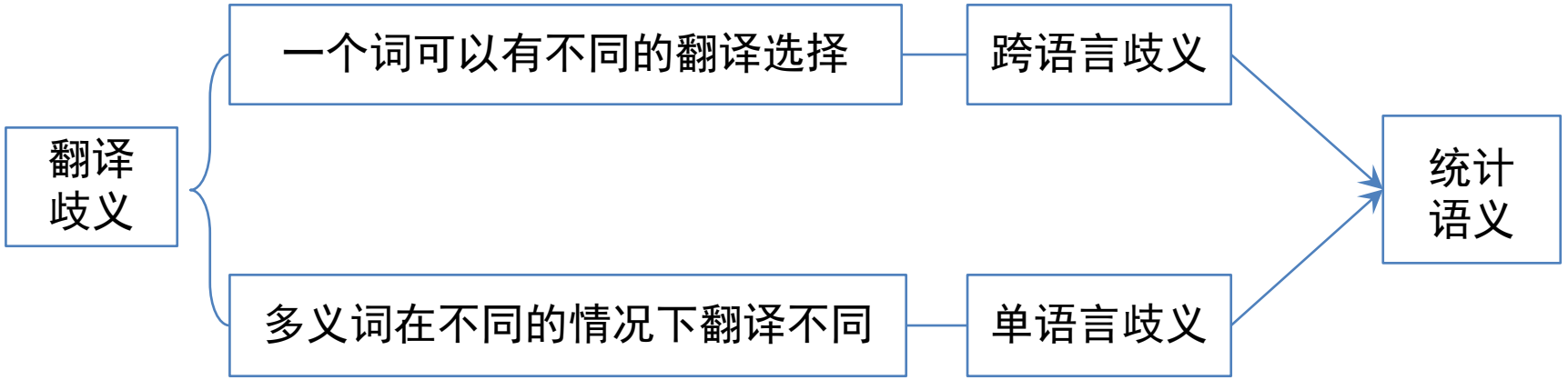
- **基于显语义空间的跨语言文档建模**
  - (Cimiano et al., 2009; Kumar et al., 2011)
  - **缺点**：受到语义资源规模和更新速度的影响

- **基于统计语义空间的跨语言文档建模**
  - 主题对齐
    - (Wei et al. , 2008; Muramatsu and Mori, 2004; Mimno et al., 2009)
    - **缺点**：文档集主题偏差
  - 词对齐
    - (Zhang et al.,2010; Boyd-Graber and Resnik, 2010; Boyd-Graber and Resnik, 2009; Jagarlamudi and Daum 2010)
    - **缺点**：文档层面消歧, 效果不理想

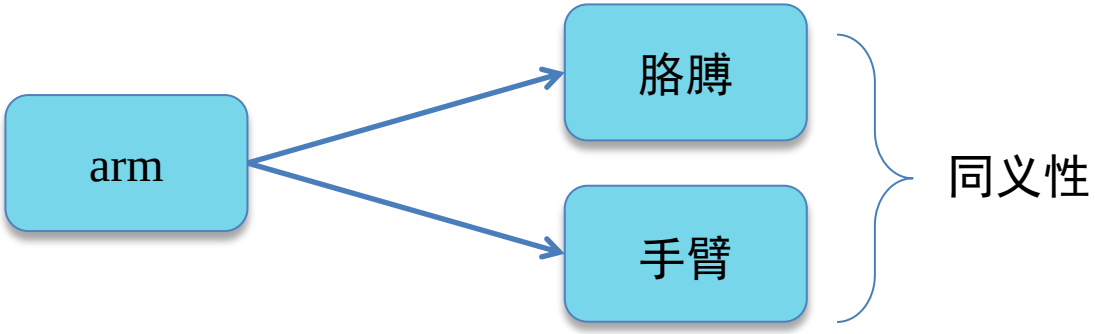
研究难点

- **核心问题**
  - 跨语言文档建模

- **难点**
  - 翻译歧义

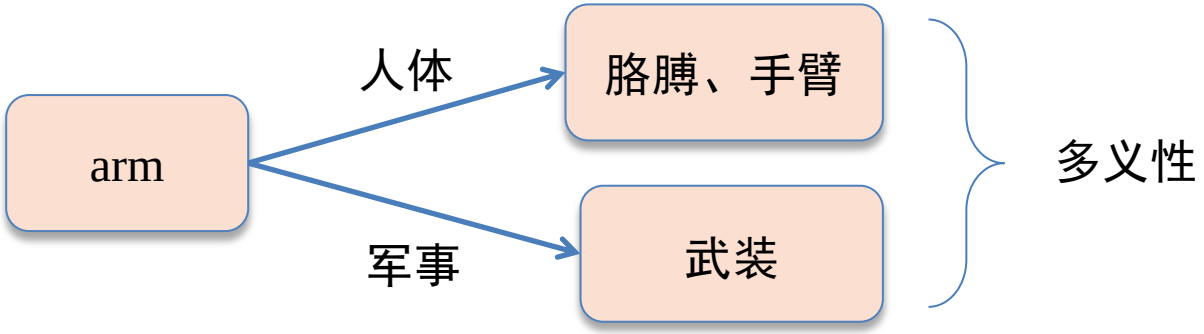


如何更好地解决跨语言歧义



- 语言的同义性
  - 需要对词和它的翻译有着相同或者相近的表示
- 现有基于统计语义的方法或者只能获取主题层面的对齐或者通过词对齐构造跨语言主题空间，效果不理想
- 首要问题：如何充分利用语义空间获取更好的对齐信息？

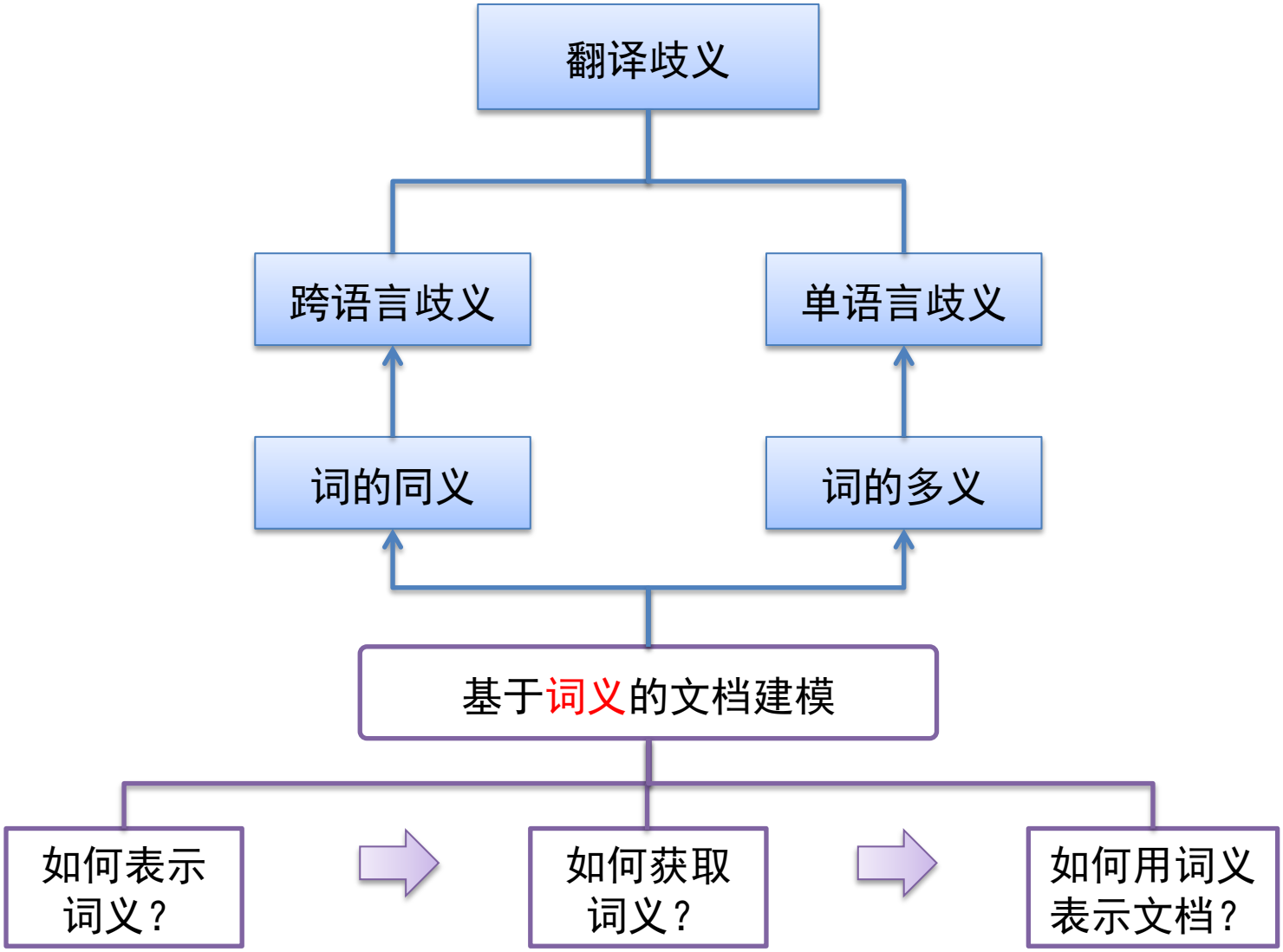
如何更好地解决单语言歧义



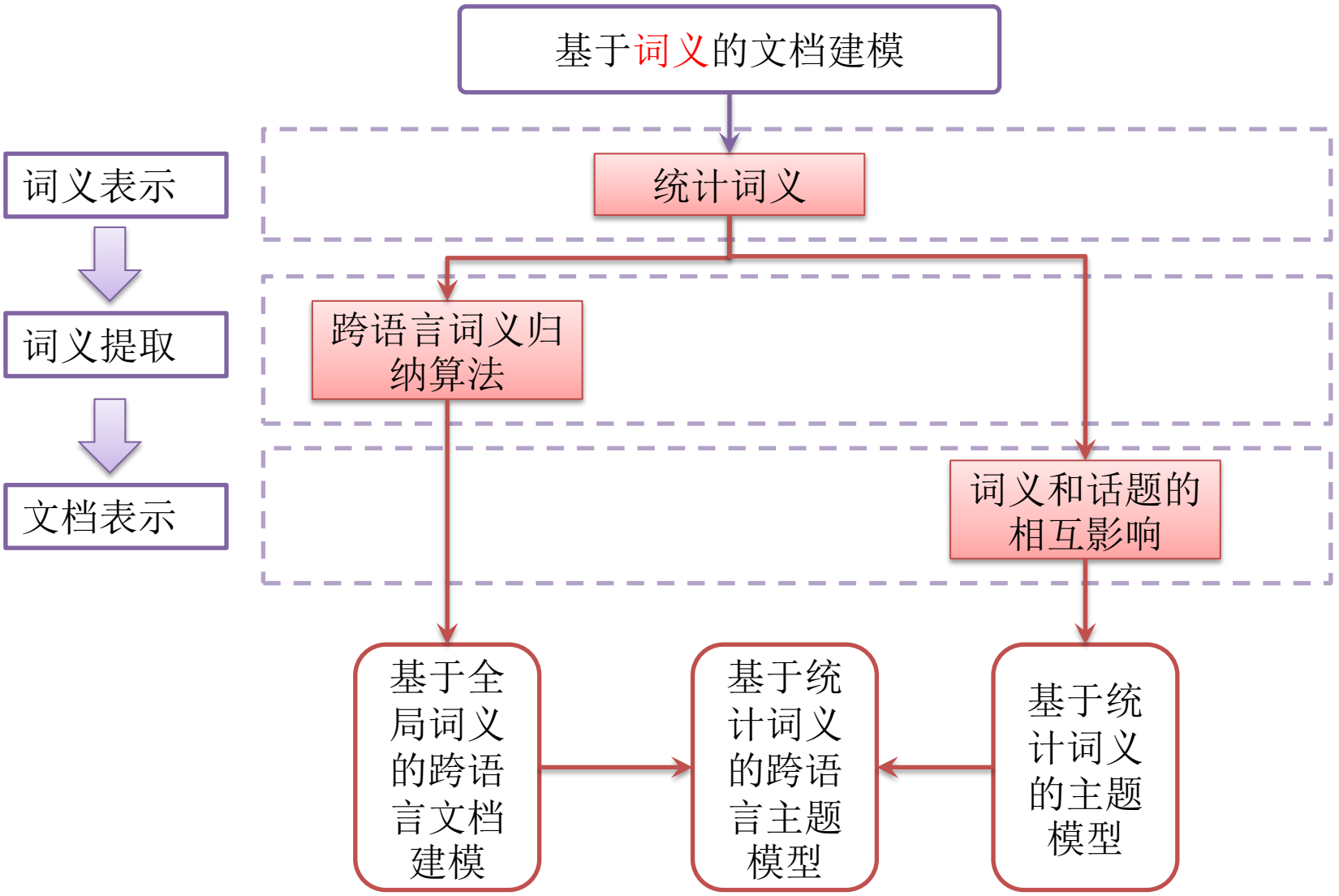
- **语言的多义性**
  - 需要根据不同的上下文判断词的语义，对于表达不同含义的同一个词给出不同的表示。
- **现有的基于统计语义的方法只能从文档层面进行消歧，不能充分利用近距离上下文信息。**
- **本研究认为：近距离的上下文信息更能反映词的含义。**
  - 句子级别的上下文

- **Part 3: 研究工作**
  - 研究思路及内容
  - 基于全局词义的跨语言文档建模
  - 基于统计词义的主题模型
  - 基于统计词义的跨语言主题模型
  - 结论

研究思路（1/2）


翻译歧义
跨语言歧义
单语言歧义
词的同义
词的多义
基于词义的文档建模
如何表示词义？
如何获取词义？
如何用词义表示文档？

研究思路（2/2）


基于词义的文档建模
词义表示
词义提取
文档表示
统计词义
跨语言词义归纳算法
词义和话题的相互影响
基于全局词义的跨语言文档建模
基于统计词义的跨语言主题模型
基于统计词义的主题模型

1、基于全局词义的文档建模方法

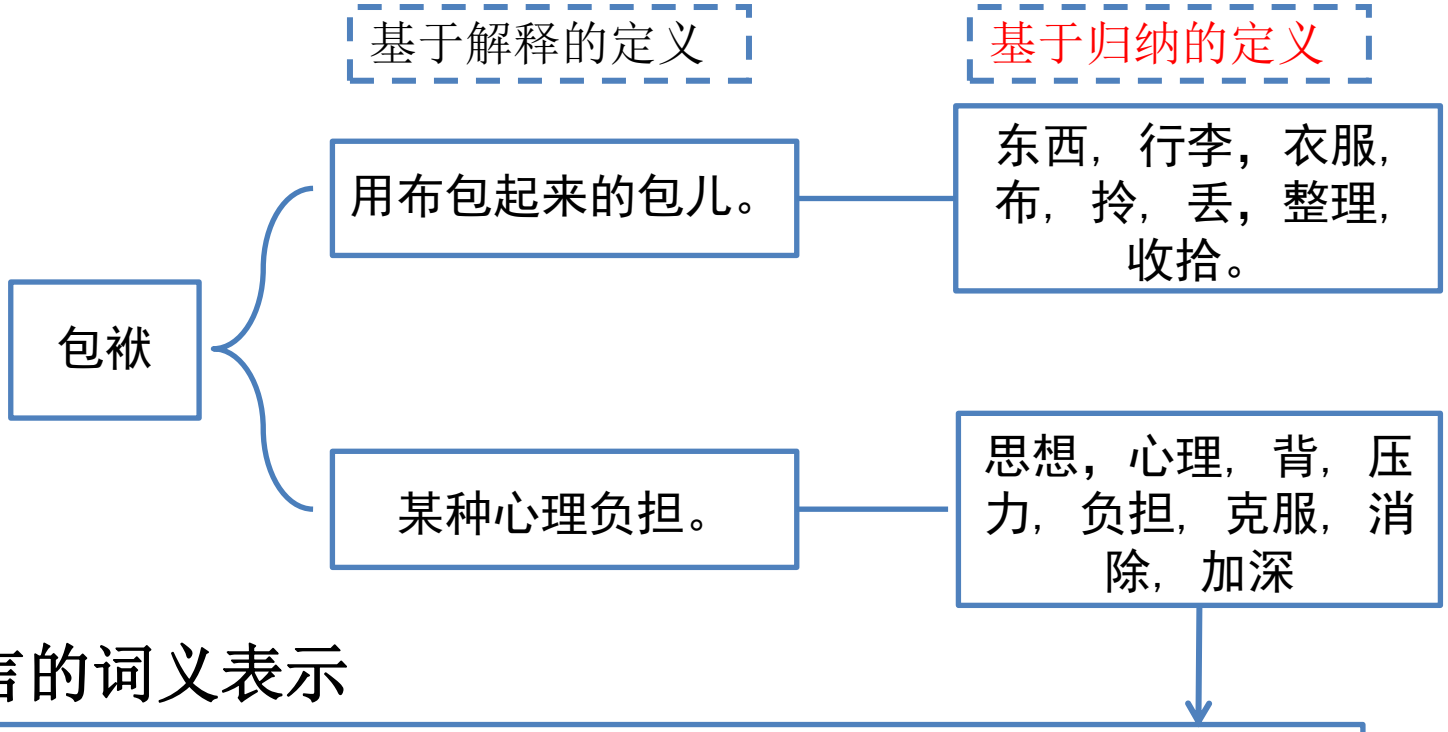
- **研究问题1：如何表示跨语言词义**
  - 基于归纳的定义
  - 跨语言局部词义和跨语言全局词义

- **研究问题2：如何获取跨语言词义**
  - 跨语言词义归纳算法

- **研究问题3：如何用跨语言词义表示文档**
  - 跨语言词义作为特征
  - 现有的文档建模方法

词义的表示

- 单语言词义的表示

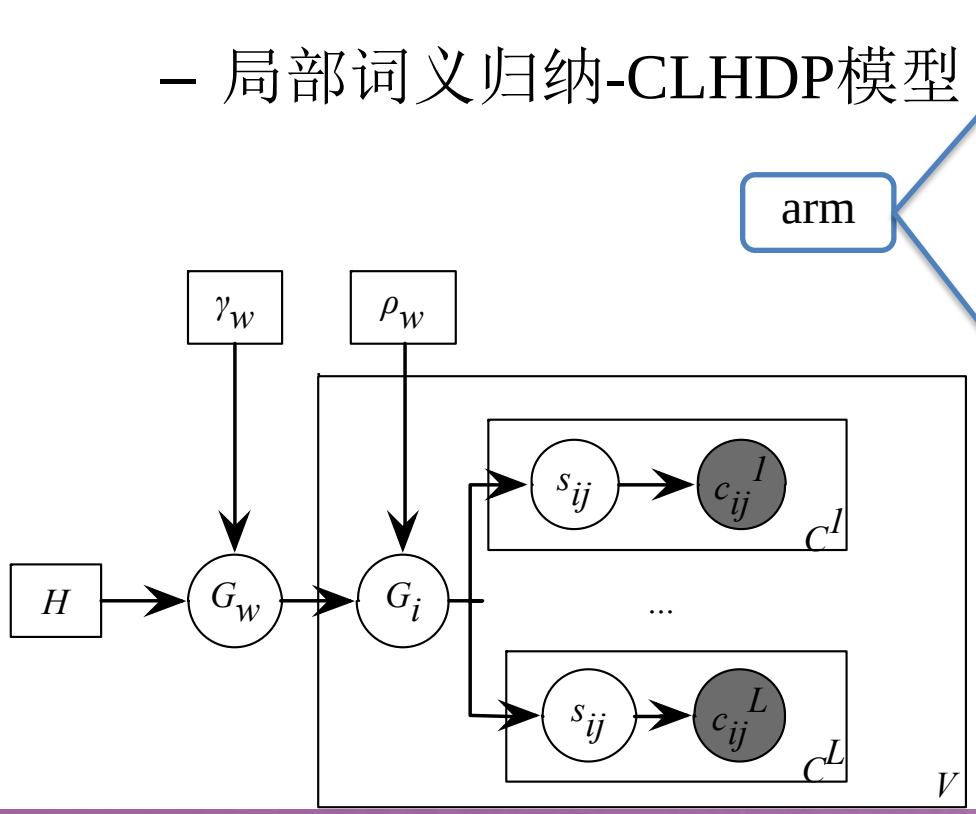


- 跨语言的词义表示

思想, 心理, 背, 压力，负担, 甩掉，消除，加深,
burden, psychological,suffer, pressure, endure, eliminate, alleviate, add

跨语言词义的表示和获取（1/2）

- **跨语言局部词义**
  - 多种语言中的一组上下文的词的概率分布
  - 局部词义归纳-CLHDP模型



arm

arm#1={limb: 0.159, forelimb: 0.069, sleeve: 0.019; 手臂: 0.137, 上肢: 0.079, 衣袖: 0.017}

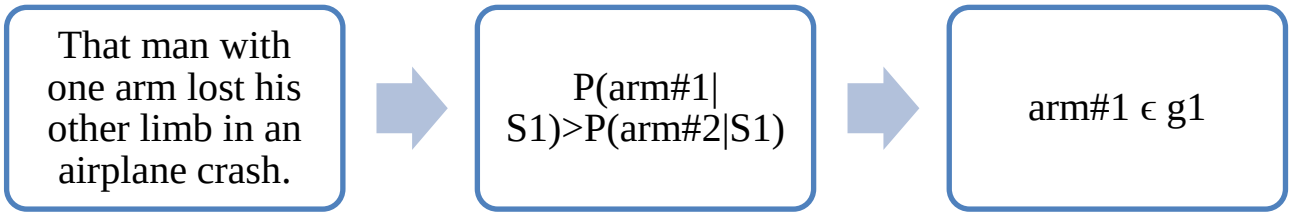
arm#2={weapon: 0.116, war: 0.039, battle: 0.026; 装备: 0.153, 武器: 0.027; 战争: 0.026}

跨语言词义的表示和获取（2/2）

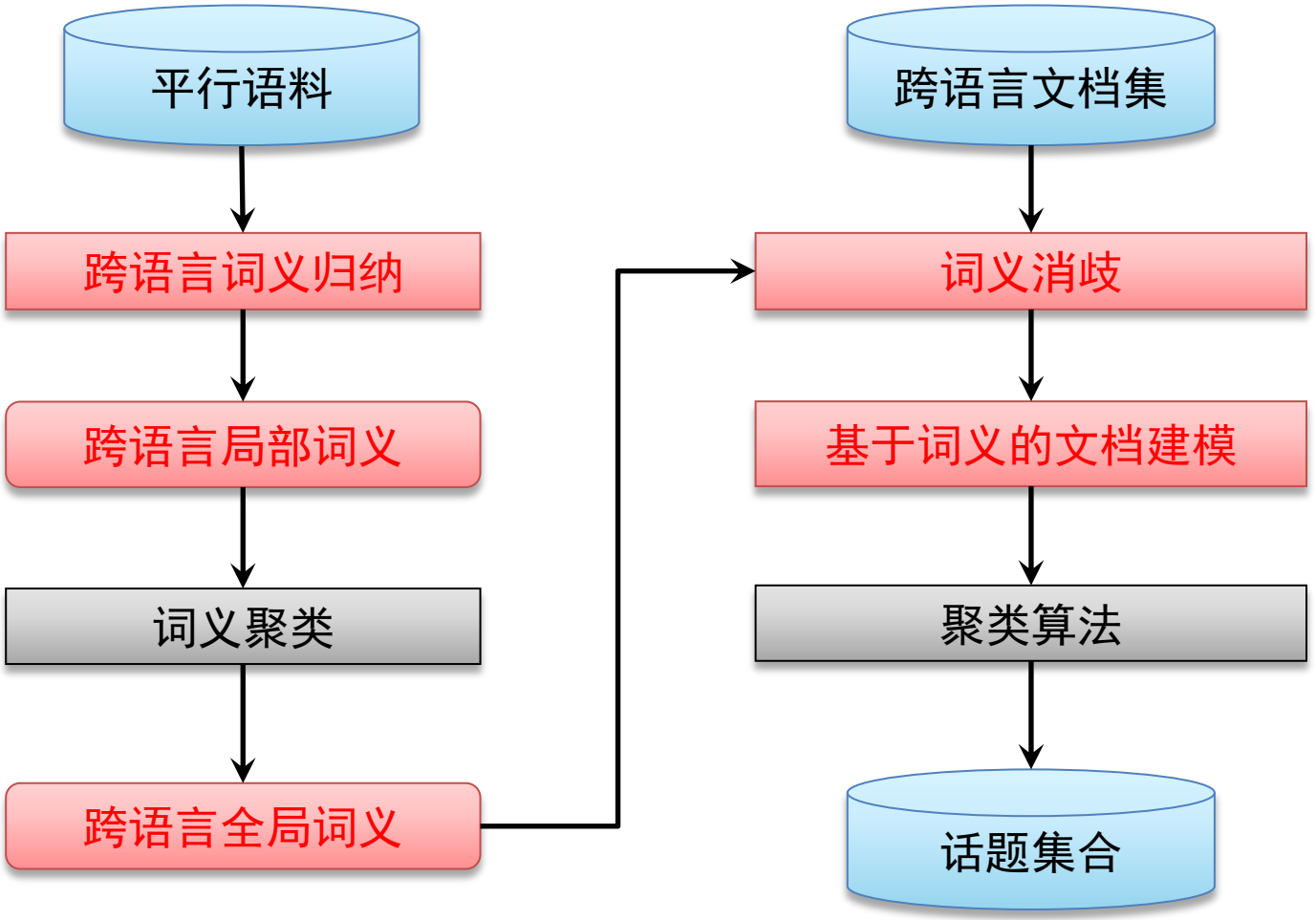
- **跨语言全局词义**
  - 一组同义的局部词义

g#1={arm#1, 手臂#1}
g#2={arm#2,weapon#1, 装备#1}

  - 上下文词和概率分布作为特征和权重
  - 聚类算法
- **跨语言词义消歧**



算法流程


平行语料
跨语言词义归纳
跨语言局部词义
词义聚类
跨语言全局词义
跨语言文档集
词义消歧
基于词义的文档建模
聚类算法
话题集合

实验评测

- **实验设置**
  - 开发集：1M 平行句对（LDC数据集）
  - 翻译概率：在开发集中使用Giza++获得
  - 评测集

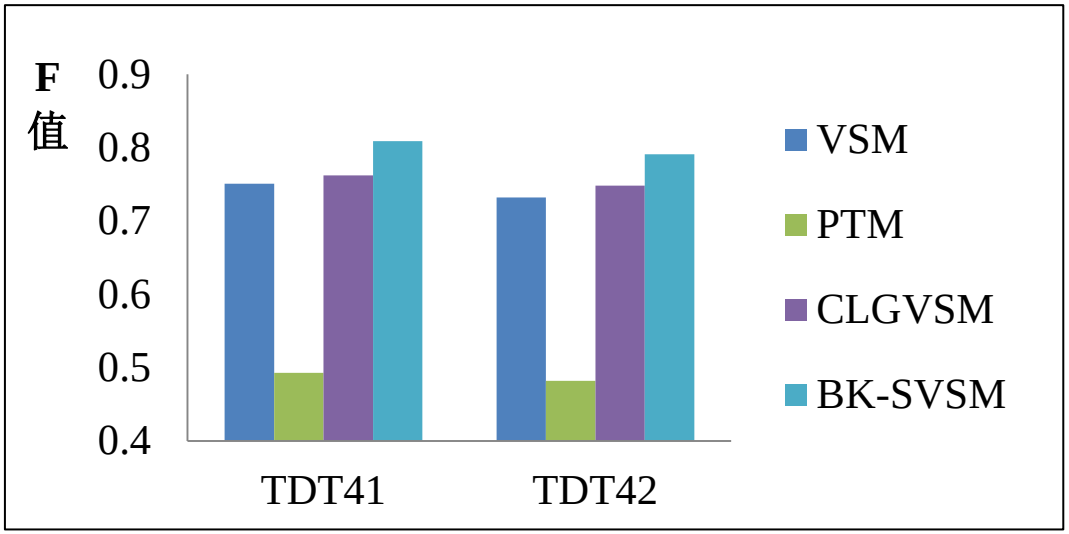
语料	TDT41 (2002)	TDT42 (2003)
英文(话题数/文档数)	38/1270	33/617
中文(话题数/文档数)	37/657	32/560
总计(话题数/文档数)	40/1927	37/1177

- **评测指标**
  - F-值（准确率、召回率）

基线系统

- **向量空间模型（VSM）**
  - 采用VSM表示文档,余弦相似度计算文档相似度。并从翻译概率获取翻译信息
- **多语言主题模型（PTM）(Mimno et al., 2009)**
  - 在平行语料中训练跨语言主题,然后在测试集中用跨语言主题表示文档
- **跨语言广义向量空间模型（CLGVSM）**
  - 跨语言词相似度
  - 拓展广义向量空间模型
  - 基于“软匹配”的特征选择算法

实验结果



- **多语言主题模型（PTM）**
  - 文档集话题偏差
- **跨语言广义向量空间模型（CLGVSM）**
  - 针对跨语言歧义，没有考虑单语言歧义
- **基于全局词义的方法（BK-SVSM）**
  - 有利于解决跨语言歧义和单语言歧义

小结

基于全局词义的跨语言文档建模

- 定义了跨语言局部词义和全局词义
- 提出了跨语言词义归纳方法
- 提出了基于词义的文档建模方法

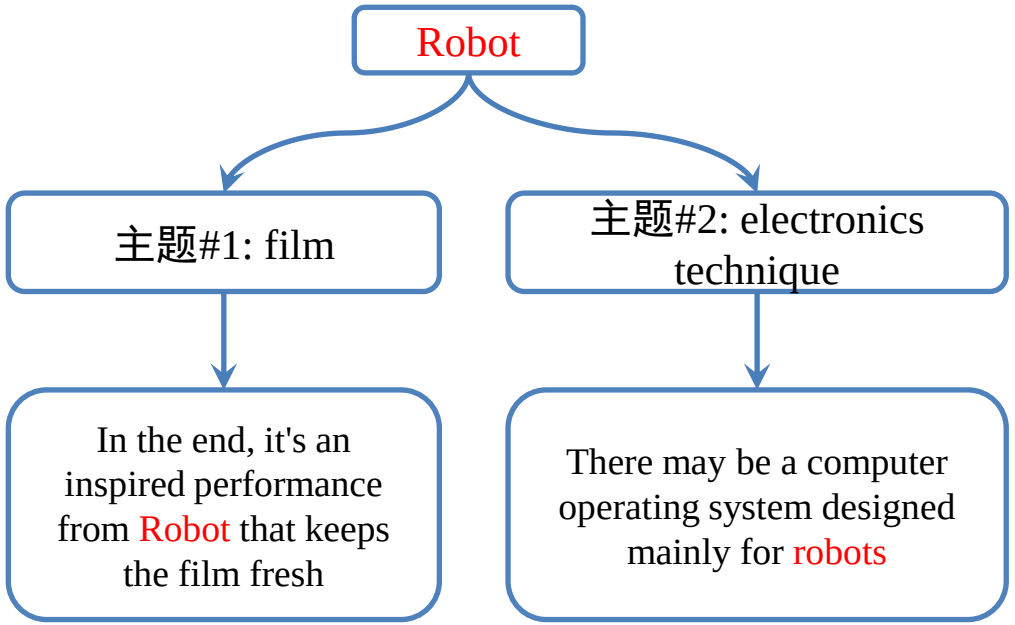
更好的文档建模方法？

基于统计词义的主题模型

- 文档中主题对词义的影响

2、基于统计词义的主题模型

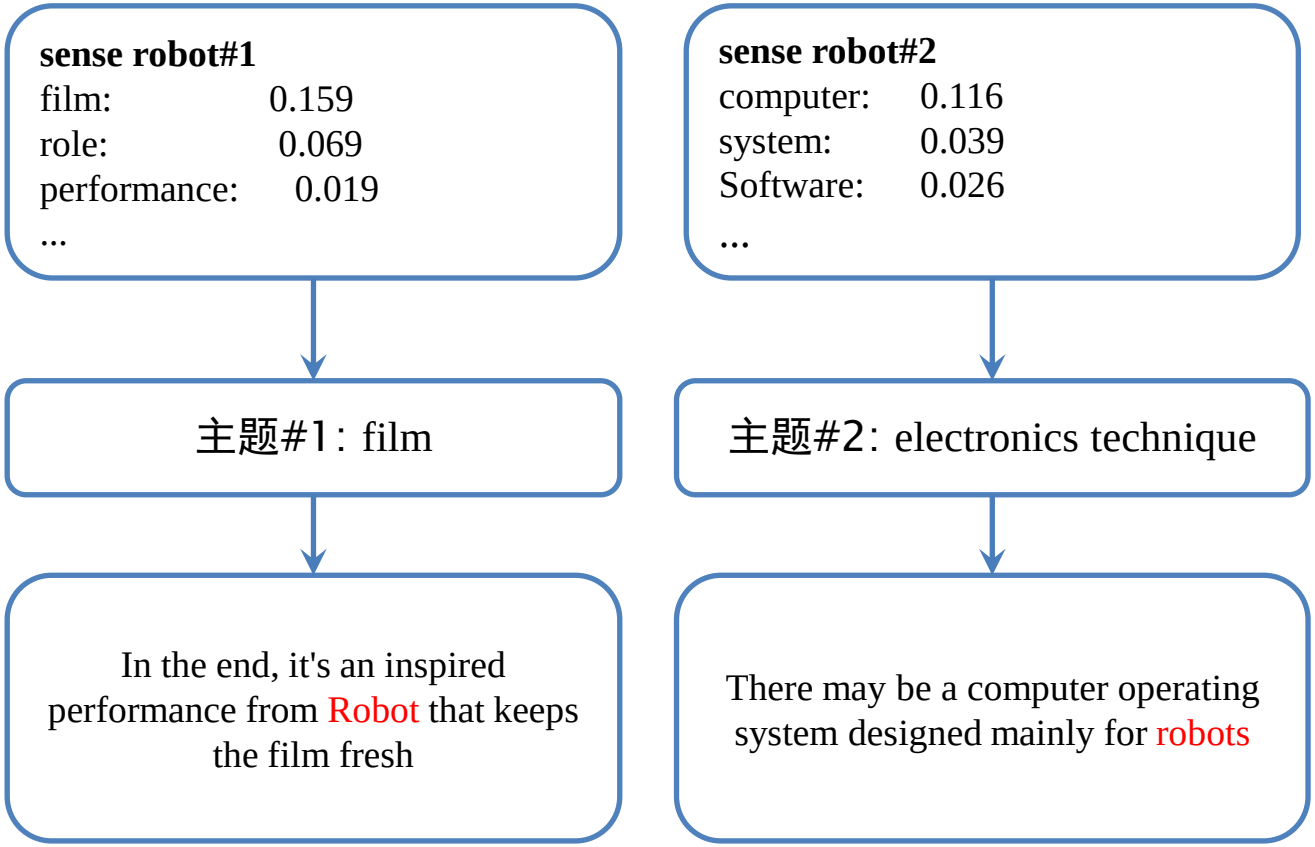
- 研究问题：如何用统计词义表示文档。
- 问题分析
  1. 传统主题模型（LDA）
     - 依赖于词的共现来挖掘语义信息。



问题分析（1/2）

2. 使用词义作为附加特征

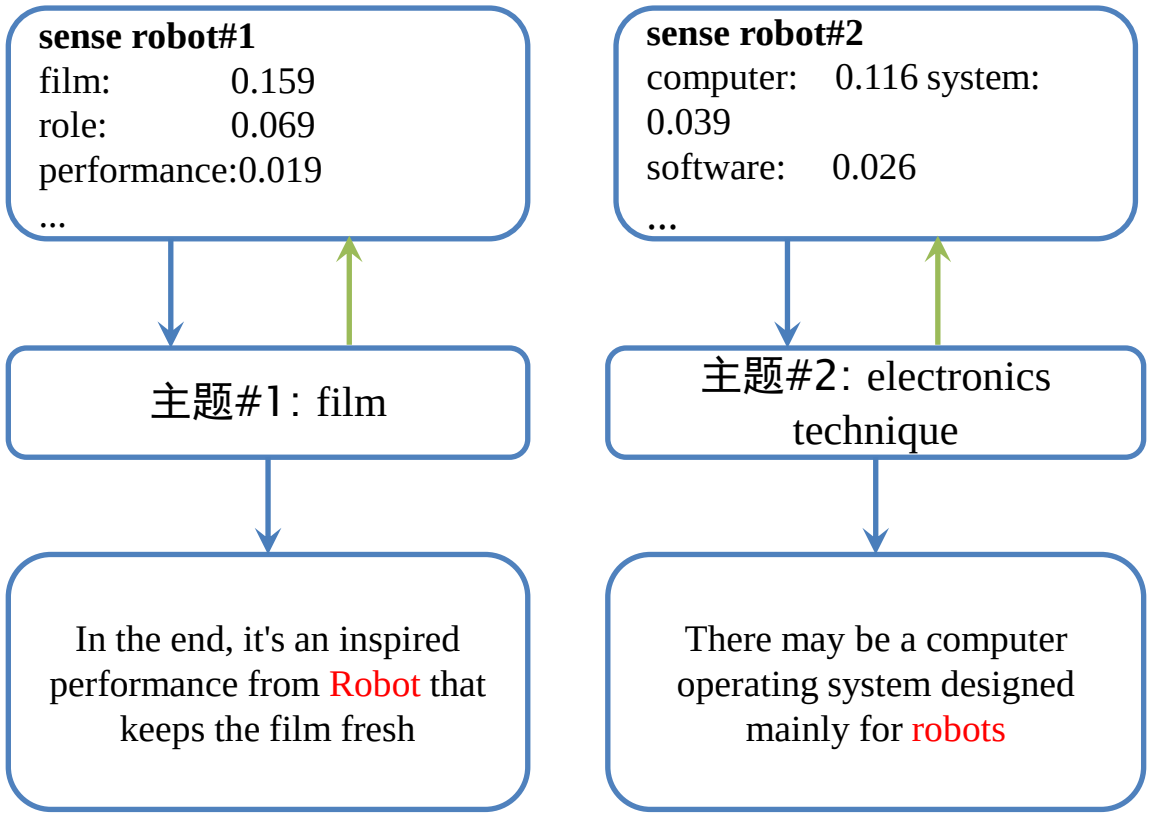
– 增强主题模型的区分性。



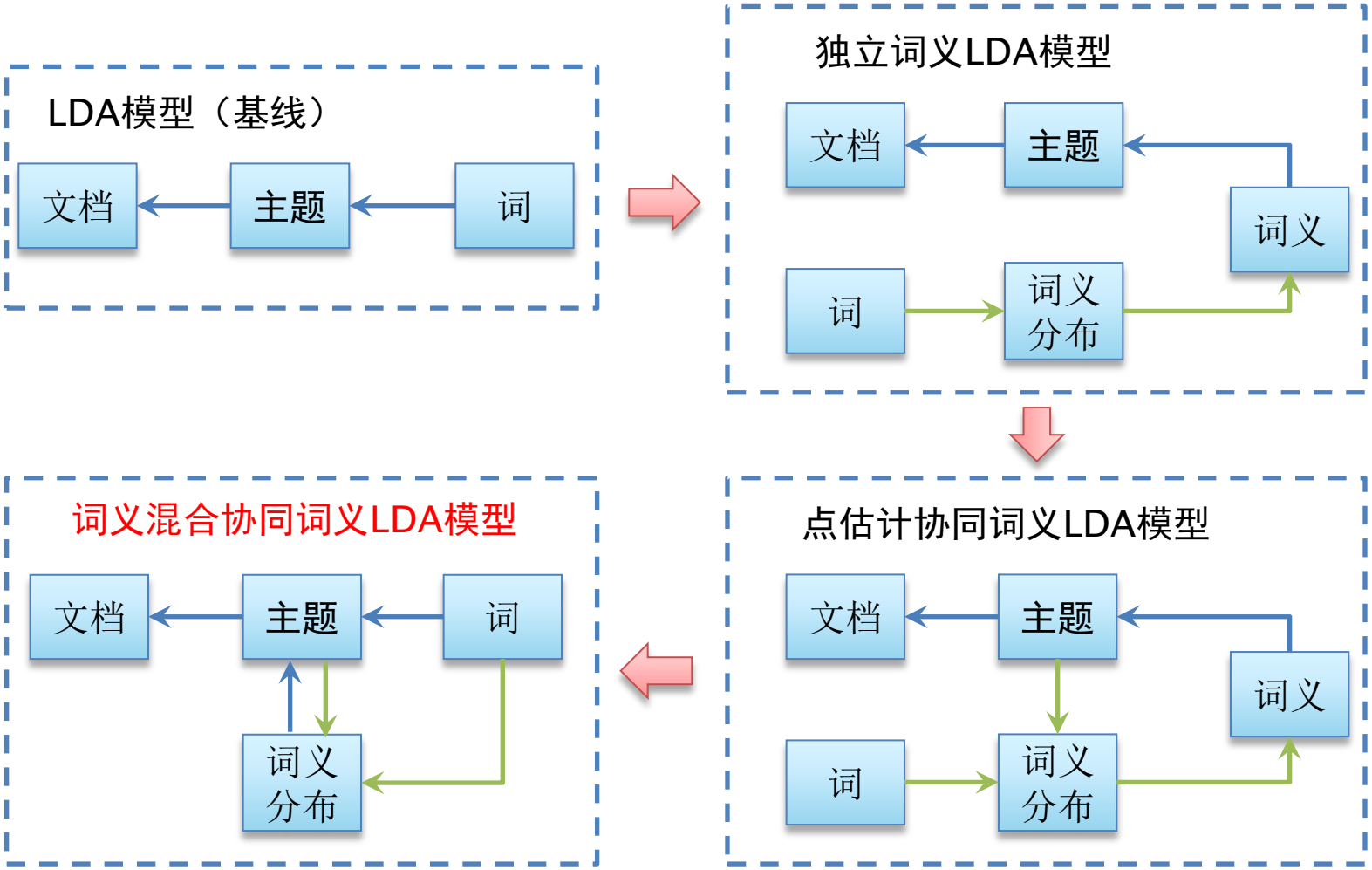
问题分析（2/2）

3. 主题是否对词义有影响？

– 主题“film”中“robot”的词义是“robot#1”的概率较大

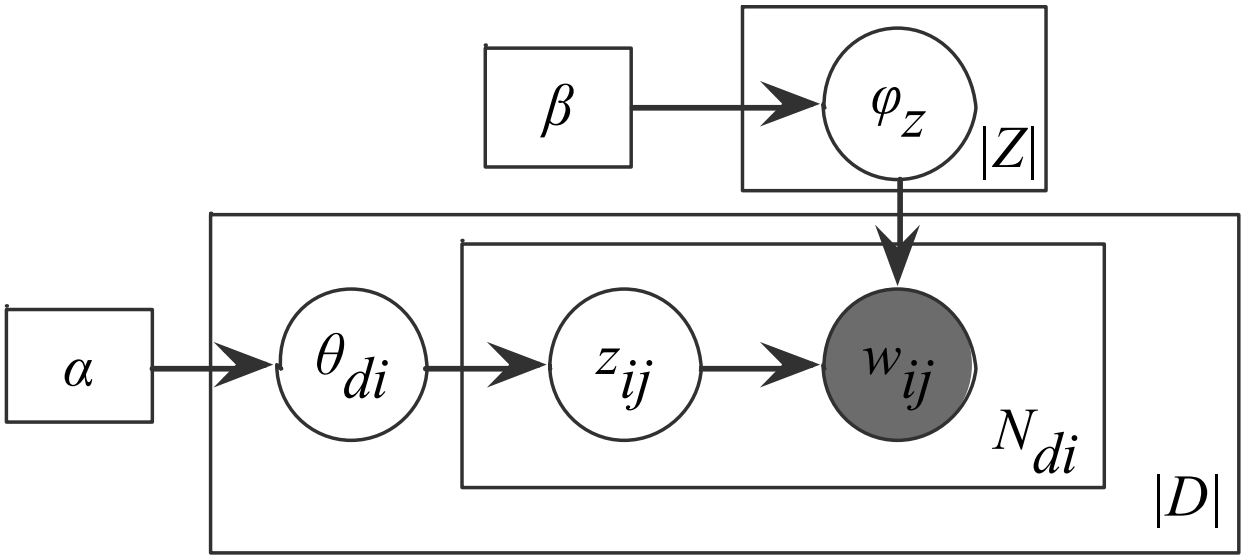


解决思路


LDA模型（基线）
文档
主题
词
独立词义LDA模型
文档
主题
词义
词
词义分布
点估计协同词义LDA模型
文档
主题
词义
词
词义分布
词义混合协同词义LDA模型
文档
主题
词
词义分布

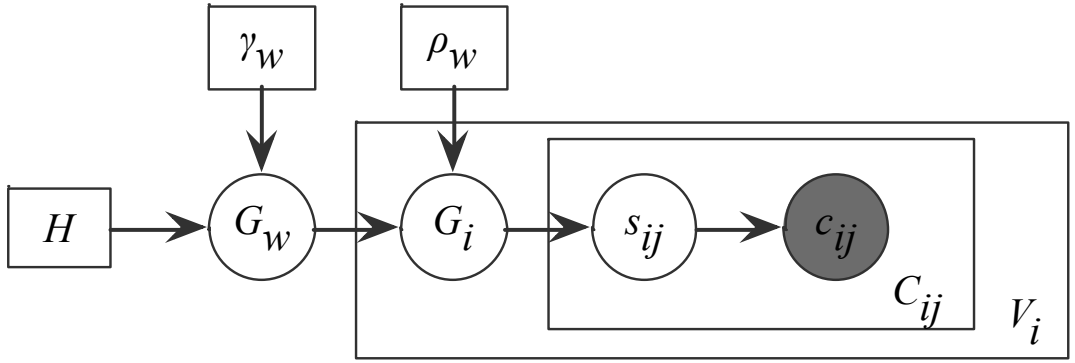
LDA模型（基线）

（Blei et al., 2003）

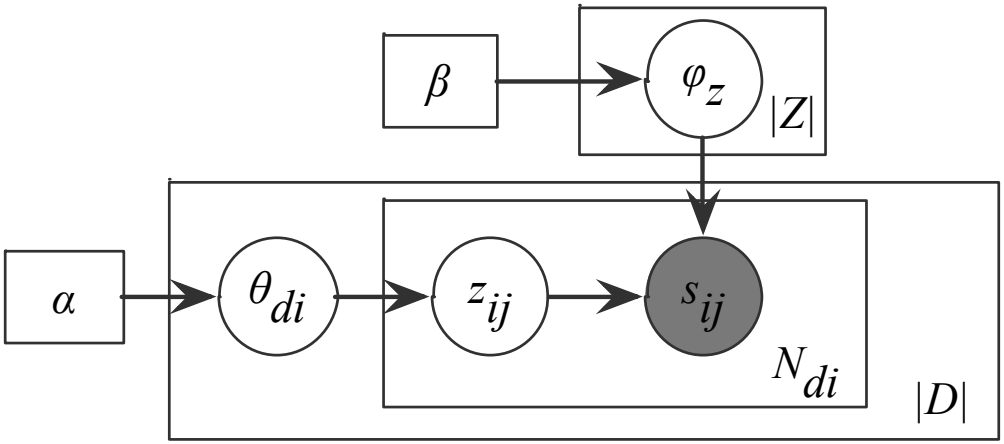


2.1 独立词义LDA模型（SA-SLDA）

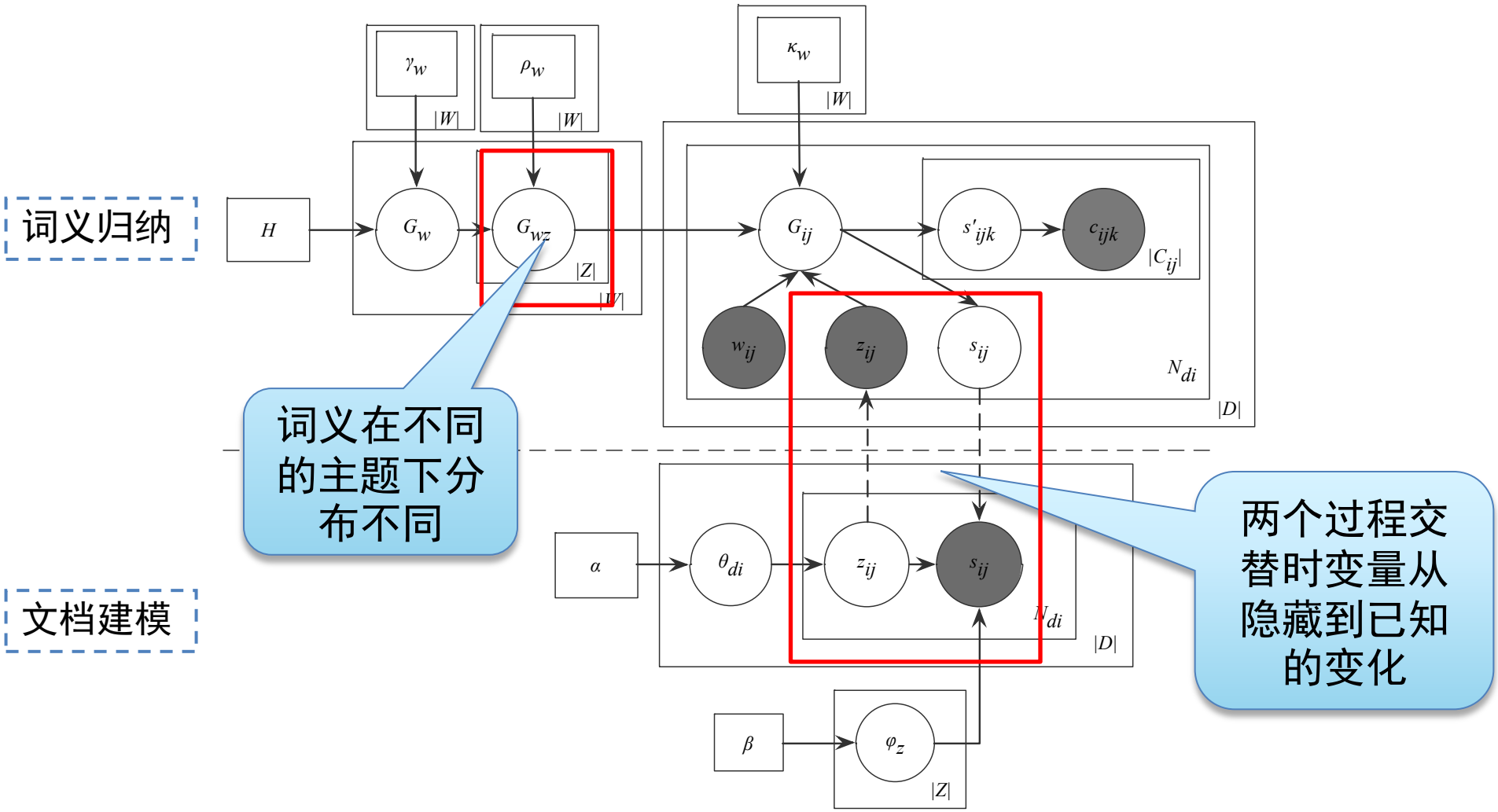
词义归纳



文档建模



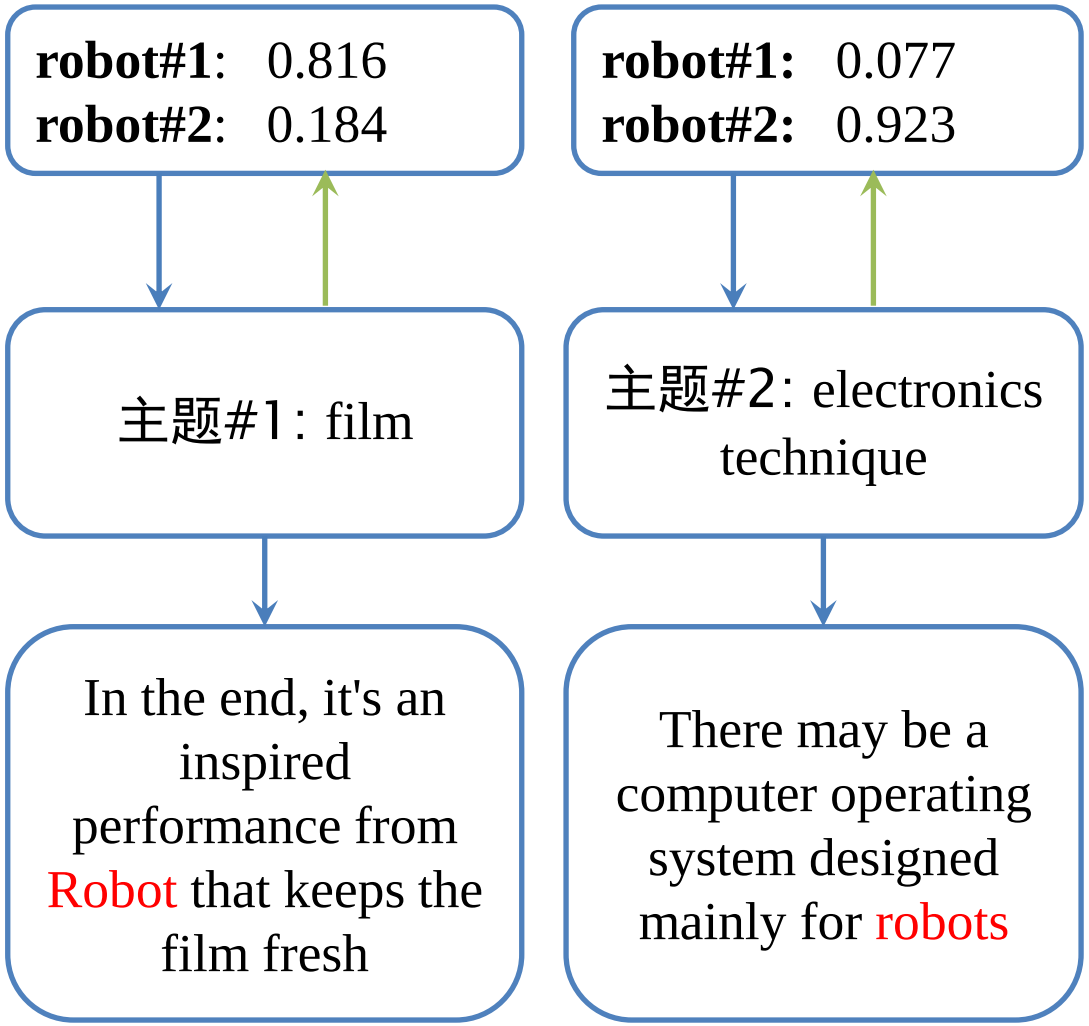
2.2 点估计协同词义LDA模型（PCo-SLDA）



2.3 词义混合协同词义LDA模型（SCo-SLDA）

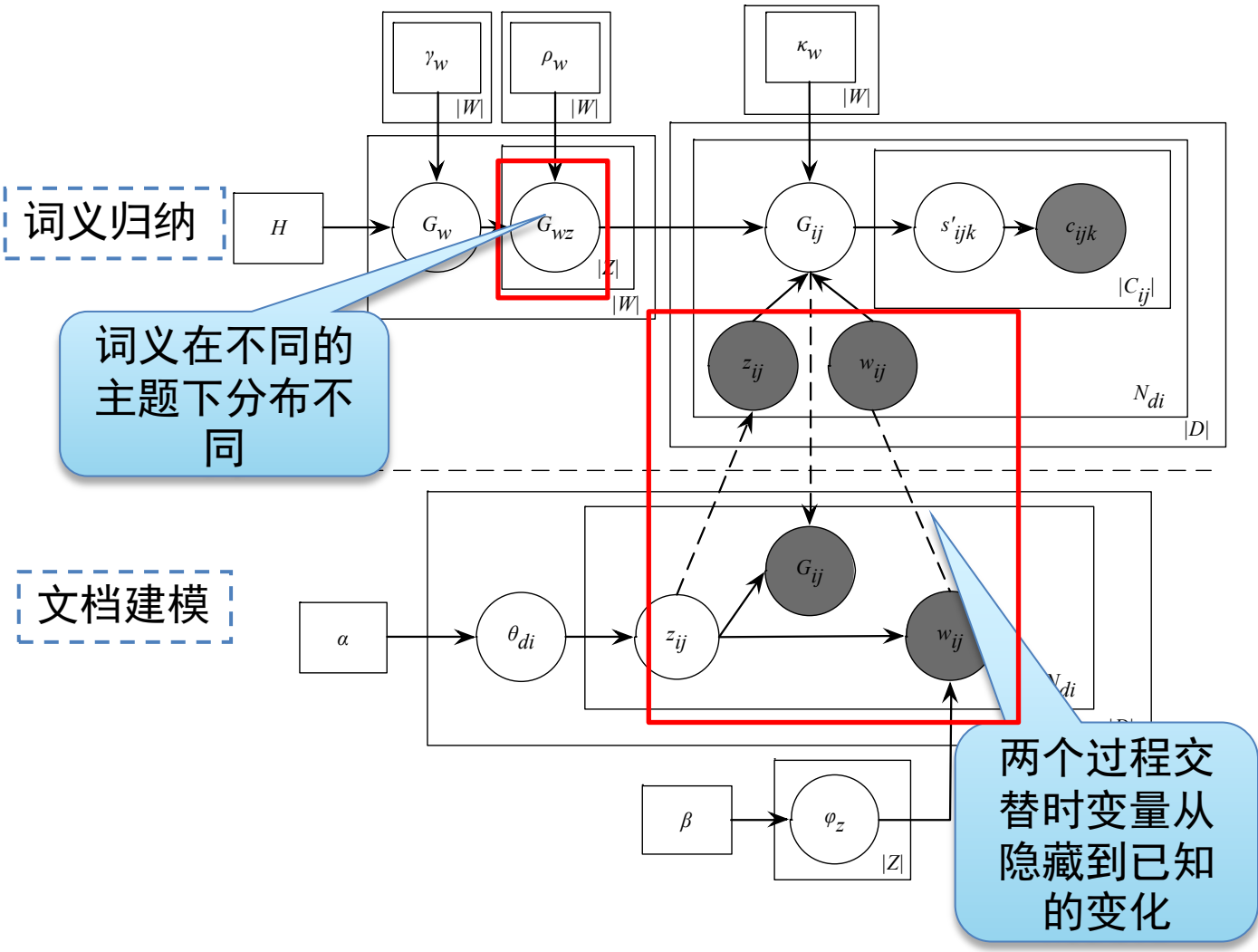
- 点估计确定词义

- 词义整体分布



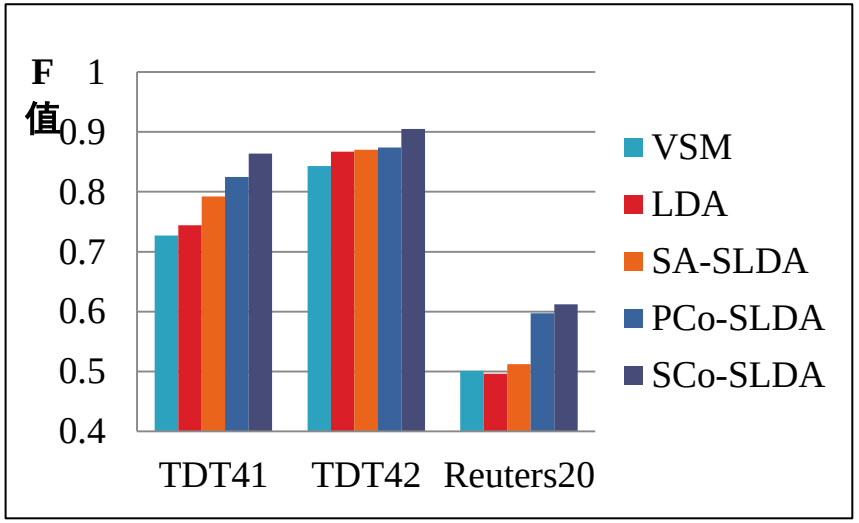
SCo-SLDA的概率图

- 主题决定于三个因素
  - 文档的主题分布
  - 给定主题生成词的概率
  - 给定词和主题，生成词义分布的概率



实验结果

- 使用词义作为特征可增强主题的区分性，改进话题分析的结果
- 考虑主题和词义的相互影响可以进一步改进性能
  - 同一个词在不同的主题下可能具有不同的词义。
  - 用主题作为词义的伪回馈可以生成主题相关的词义。
- 考虑多个可能的词义可以降低词义消歧的风险。



小结

基于统计词义的主题模型

- 将词义看作主题模型的一个隐藏变量
- 提出三个基于词义的主题模型
- 实验结果表明基于词义的主题模型可以改进单语言话题分析的性能。
  - 协同估计词义和话题可以进一步改进性能

跨语言

基于统计词义的跨语言主题模型

- 词对齐
- 词义对齐

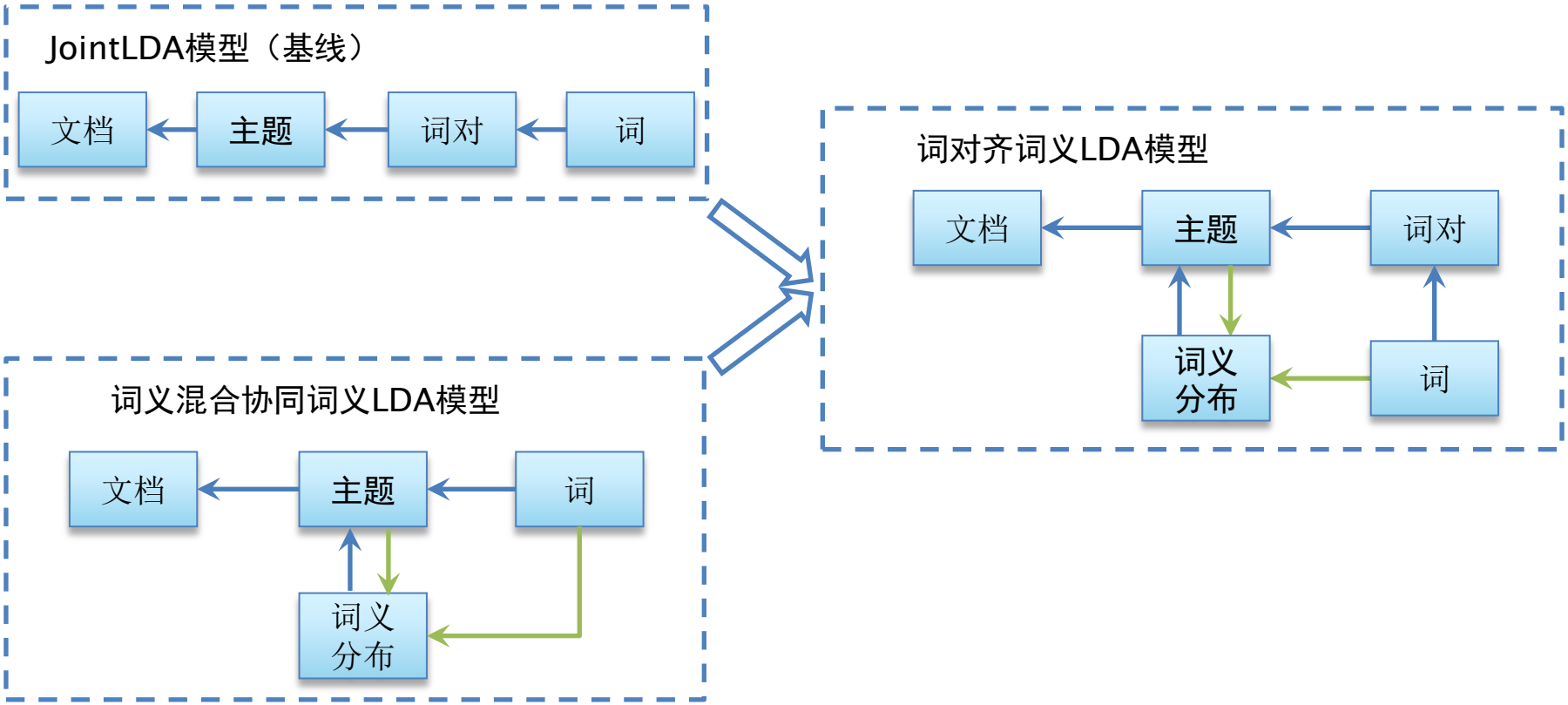
3、基于统计词义的跨语言主题模型

- **研究问题**
  - 基于统计词义的主题模型
    - 用词义改进单语言主题模型
  - 跨语言问题

- **问题分析**
  - 现有研究
    - 主题对齐：不同数据集上的偏差
    - 词对齐：消歧是在文档集全局层面上进行的
  - 基于全局词义的跨语言文档建模方法：没有考虑到主题信息对词义生成的影响
  - 基于统计词义的主题模型：无法解决跨语言问题

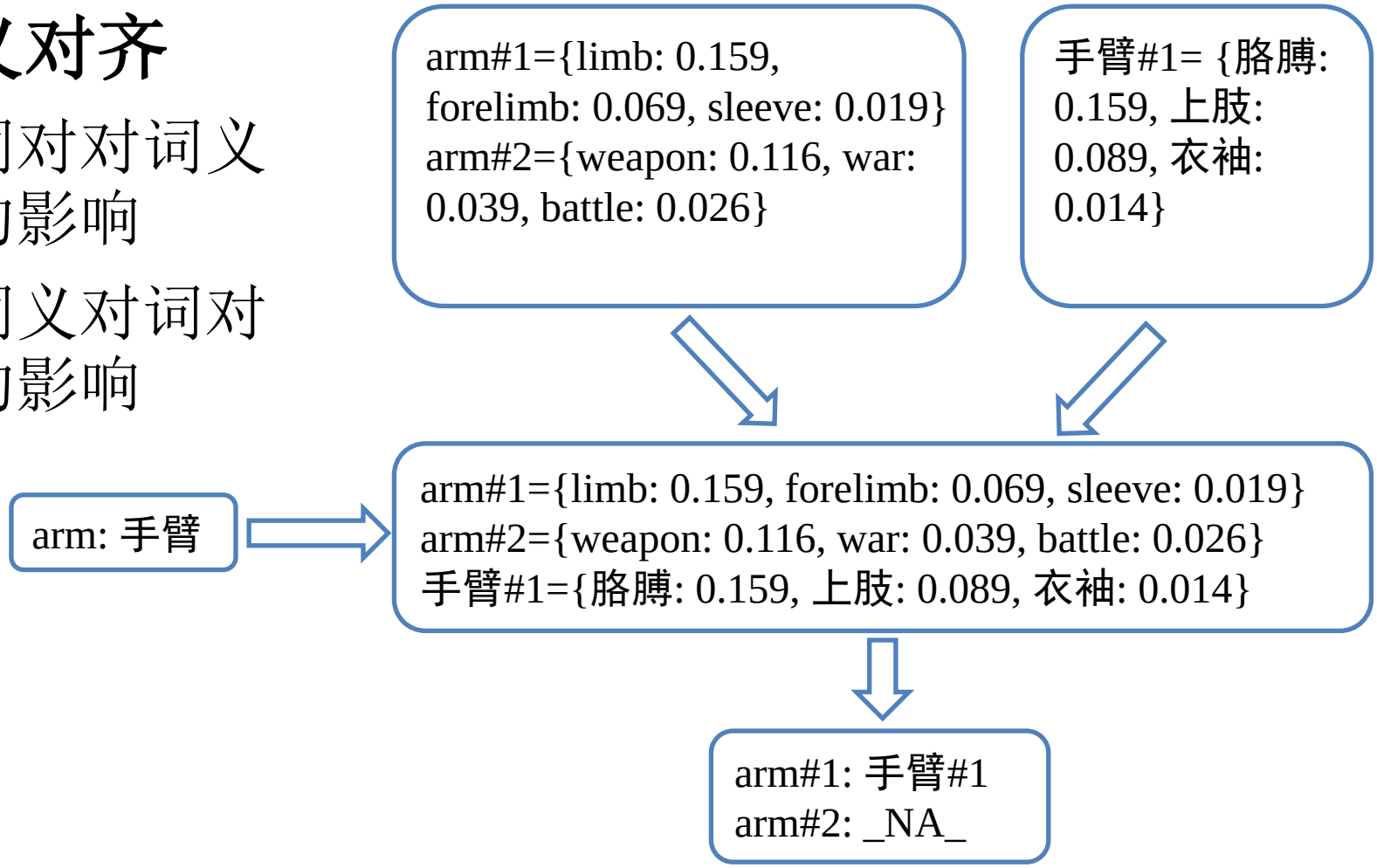
解决思路（1/3）

- 词对齐信息拓展词义主题模型

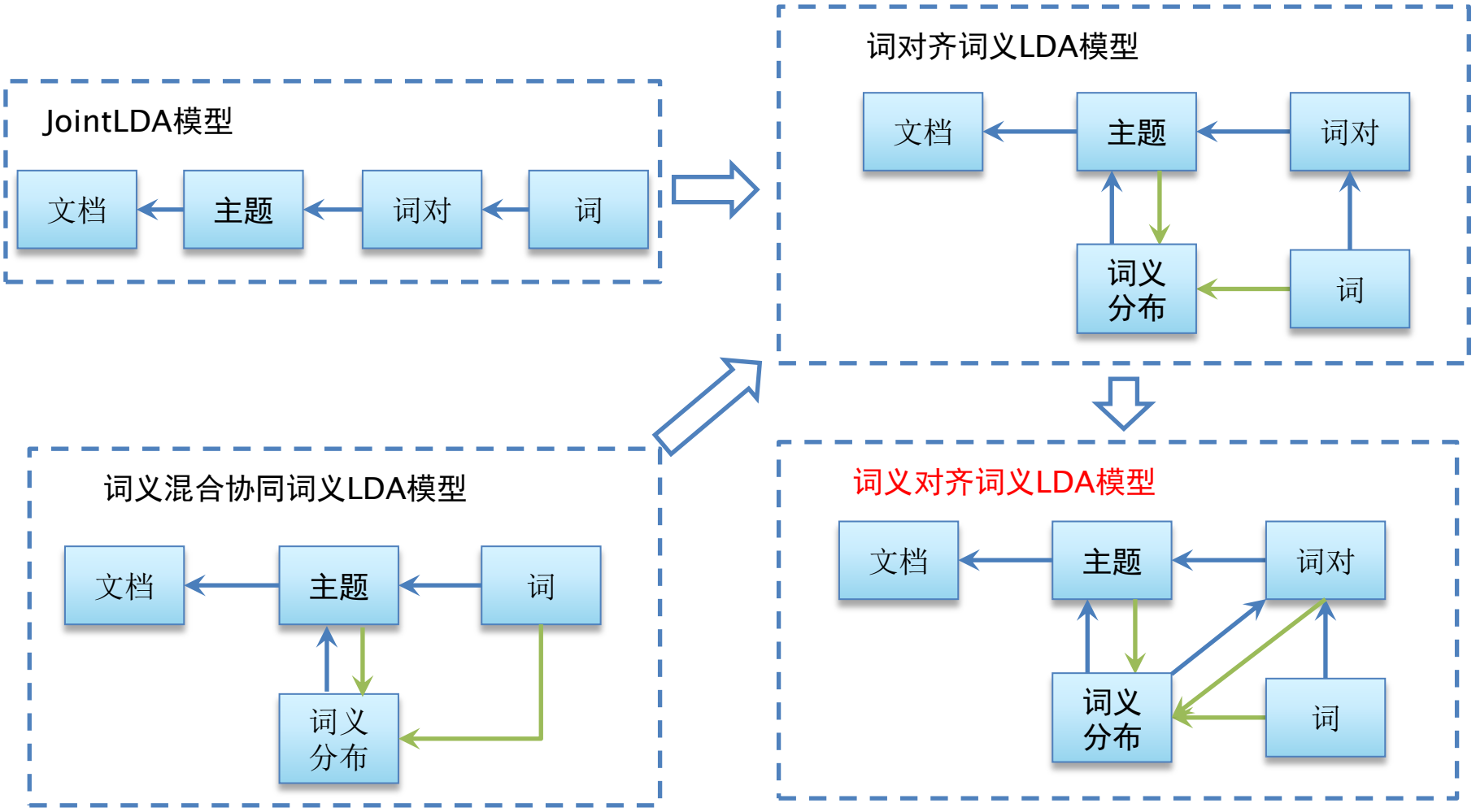


解决思路（2/3）

- 词义对齐
  - 词对对词义的影响
  - 词义对词对的影响

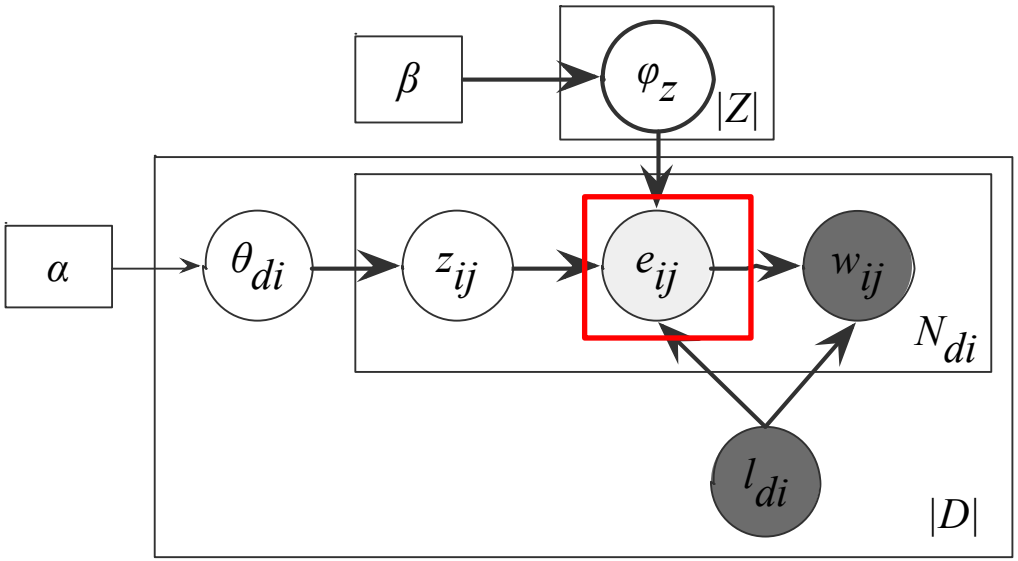


解决思路（3/3）


JointLDA模型
文档
主题
词对
词
词对齐词义LDA模型
文档
主题
词对
词义
分布
词
词义混合协同词义LDA模型
文档
主题
词
词义
分布
词义对齐词义LDA模型
文档
主题
词对
词义
分布
词

JointLDA（基线）

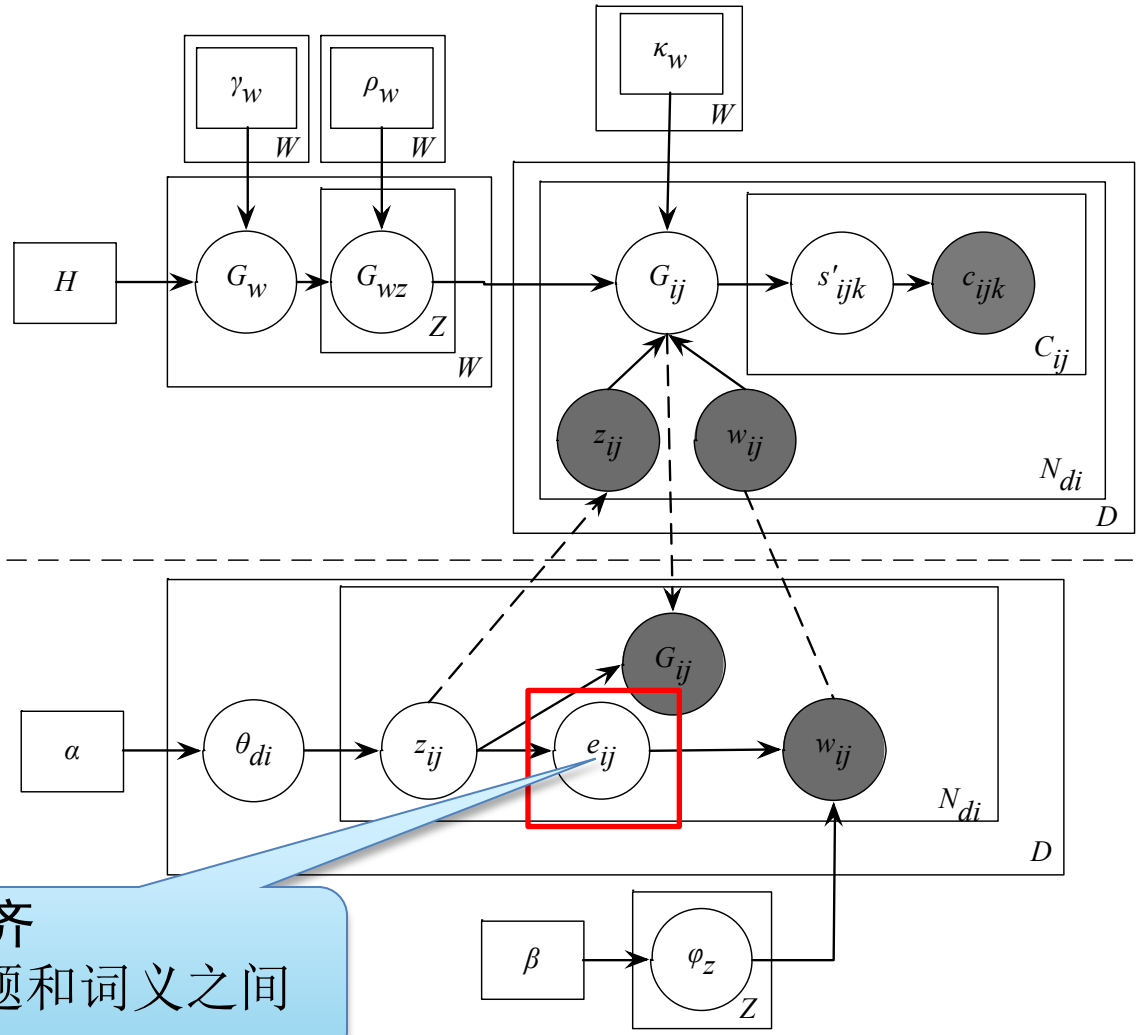
（Jagarlamudi and Daum 2010）



3.1 词对齐词义LDA模型（WA-SLDA模型）

词义归纳

文档建模

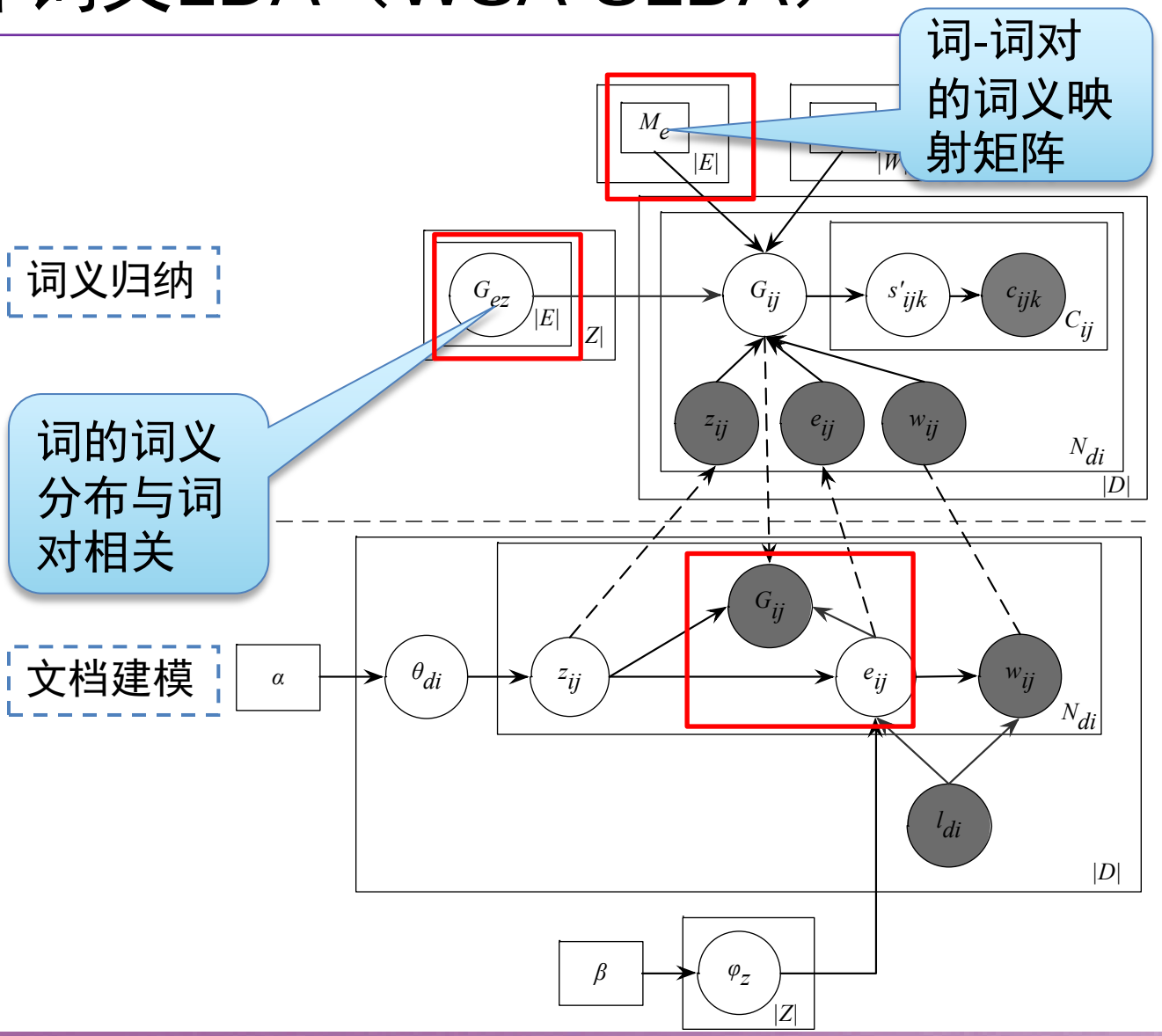


词对齐

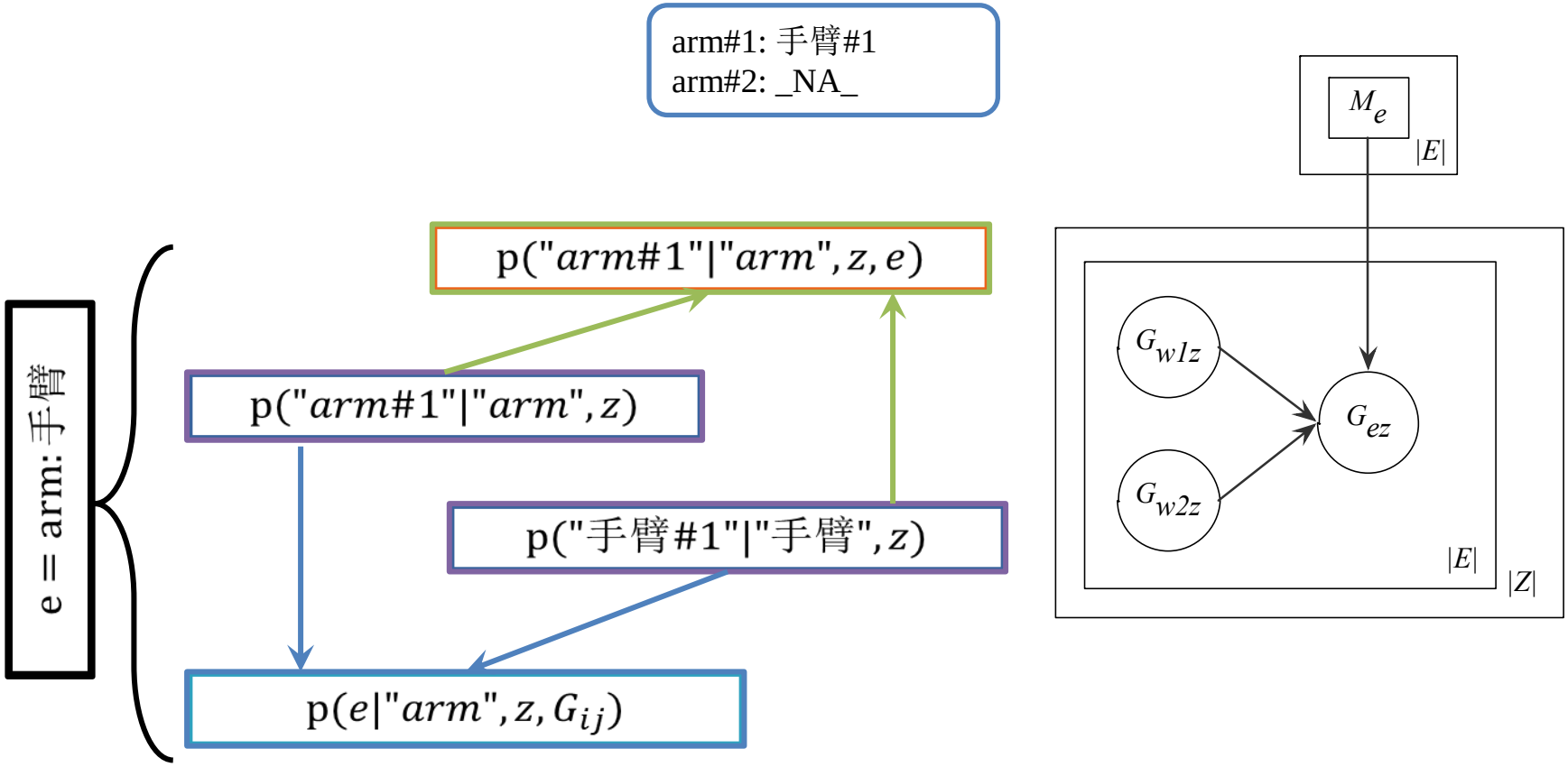
➢话题和词义之间

3.2 词义对齐词义LDA（WSA-SLDA）

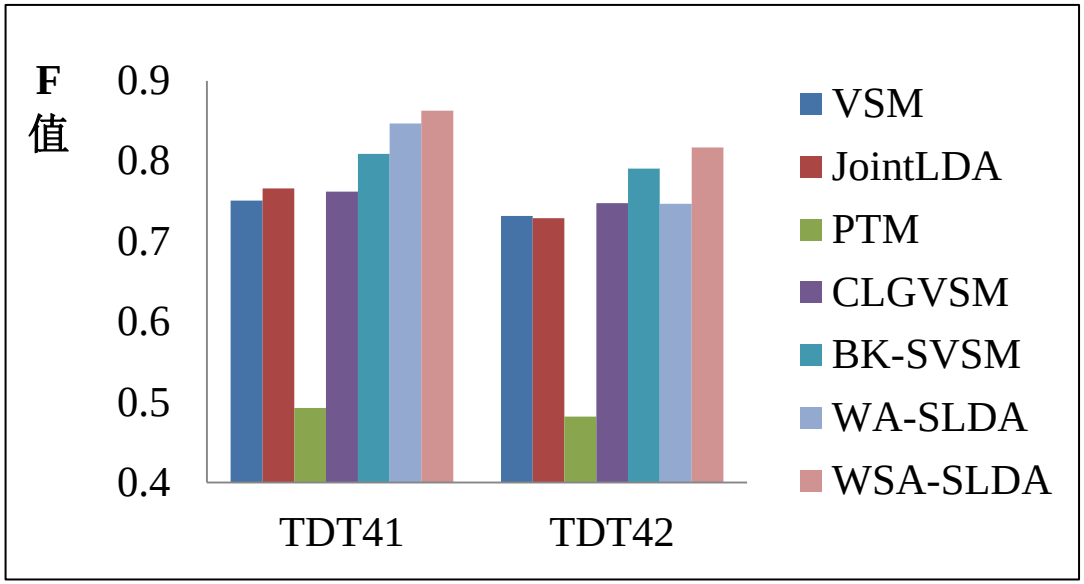
- 主题决定于三个因素
  - 文档的主题分布
  - 给定主题生成词对的概率
  - 给定词对和主题，生成词义分布的概率



词对词义分布


arm#1: 手臂#1
arm#2: _NA_
e = arm: 手臂
p("arm#1"|"arm", z, e)
p("arm#1"|"arm", z)
p("手臂#1"|"手臂", z)
p(e|"arm", z, G_ij)
M_e
|E|
G_w1z
G_w2z
G_ez
|E|
|Z|

实验结果



- **JointLDA**
  - 基于词
- **词对齐词义LDA**
  - 基于统计词义
- **基于全局词义的跨语言文档建模方法**
  - 平行语料和测试集的词义偏差
- **词义对齐词义LDA模型**
  - 词义和词对的相互影响

小结

- **在词义主题模型中引入词对齐信息**
- **设计了两个基于词对齐的跨语言词义主题模型**
  - 词对齐词义 LDA
  - 词义对齐词义 LDA
- **实验结果表明跨语言词义主题模型可以改进跨语言话题分析的性能**
  - 考虑词义对齐可以进一步改进性能

结论

- **语义表示：**
  - 与词相比，使用统计词义作为特征可以增强文档模型的区分性
- **语义获取**
  - 在跨语言词义归纳中考虑词义对齐有利于性能的改进（基于全局词义的跨语言文档建模方法，基于统计词义的跨语言主题模型）
- **文档表示**
  - 考虑词义和话题的相互影响有利于性能的改进（基于统计词义的主题模型，提出了基于统计词义的跨语言主题模型）

- **复杂度问题**
  - 高复杂性
    - 每个词的词义归纳模型即HDP模型的参数
  - 降低复杂度设想
    - 预先训练获得大多数词汇的词义分布
    - 分布式部署

论文创新性及贡献

1. **针对语言歧义问题，提出了基于词义的文档建模方法**
   I. 提出了基于全局词义的跨语言文档建模方法
   II. 提出了基于统计词义的主题模型：考虑词义和话题的相互影响
   III. 提出了基于统计词义的跨语言主题模型：针对跨语言问题

2. **针对跨语言歧义问题，提出了跨语言广义空间向量模型：将单语言广义向量空间模型扩展到跨语言问题**

3. **在话题分析任务中的全面评测**

研究展望

1. 在其他类型语料上进行评测

2. 将语义资源和统计方法结合

3. 将本文提出的文档建模方法应用到其他领域

4. 研究分布式算法，在大数据上进行应用

已发表的学术论文（1/2）

- 期刊
  1. Guoyu Tang, Yunqing Xia, Jun Sun, Min Zhang, Thomas Fang Zheng. Statistical word sense aware topic models. Soft Computing. (已被录用，SCI 源刊, IF=1.304)
  2. Guoyu Tang, Yunqing Xia, Erik Cambria, Peng Jin, Thomas Fang Zheng. Document representation with statistical word senses in cross-lingual document clustering. International Journal of Pattern Recognition and Artificial Intelligence. (已被录用，SCI源刊，IF=0.558)
  3. 唐国瑜, 夏云庆, 张民, 郑方. 基于跨语言广义向量空间模型的跨语言文档聚类方法. 中文信息学报, 2012, (02): 116-121.
  4. 唐国瑜, 夏云庆, 张民, 郑方. 基于词义类簇的文本聚类. 中文信息学报. 2013,(05): 113-119.

已发表的学术论文（2/2）

- 会议
  1. Guoyu Tang, Yunqing Xia, Jun Sun, Min Zhang, Thomas Fang Zheng. Topic Models Incorporating Statistical Word Senses. 15th International Conference, CICLing 2014, p 151-162, Kathmandu, Nepal, April 6-12, 2014. (EI 会议，检索号:20142017719370)
  2. Guoyu Tang, Yunqing Xia, Weizhi Wang, Erik Cambria and Thomas Fang Zheng. Clustering tweets using Wikipedia concepts. The 9th edition of the Language Resources and Evaluation Conference, LREC2014 , p 2262-2267, Reykjavik, Iceland,26-31 May, 2014
  3. Guoyu Tang, Yunqing Xia, Erik Cambria and Peng Jin. Inducing Word Senses for Cross-lingual Document Clustering. Computational Intelligence and Security (CIS), 2013 9th International Conference on. IEEE, pp 409-414, Leshan, China, 14-15 Dec. 2013
  4. Guoyu Tang, Yunqing Xia, Min Zhang, Haizhou Li, Thomas Fang Zheng. CLGVSM: Adapting Generalized Vector Space Model to Cross-lingual Document Clustering. The 5th International Joint Conference on Natural Language Processing, IJCNLP 2011, p 580-588, Chiang Mai, Thailand, 8-13 Nov. 2011.
  5. Guoyu Tang, Yunqing Xia. Adaptive Topic Modeling with Probabilistic Pseudo Feedback in Online Topic Detection. 15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010, p 100-108, Cardiff, UK, June 23-25, 2010 (EI 会议，检索号: 20103313149645)

敬请各位老师指导！

谢谢！

1、跨语言广义空间向量模型

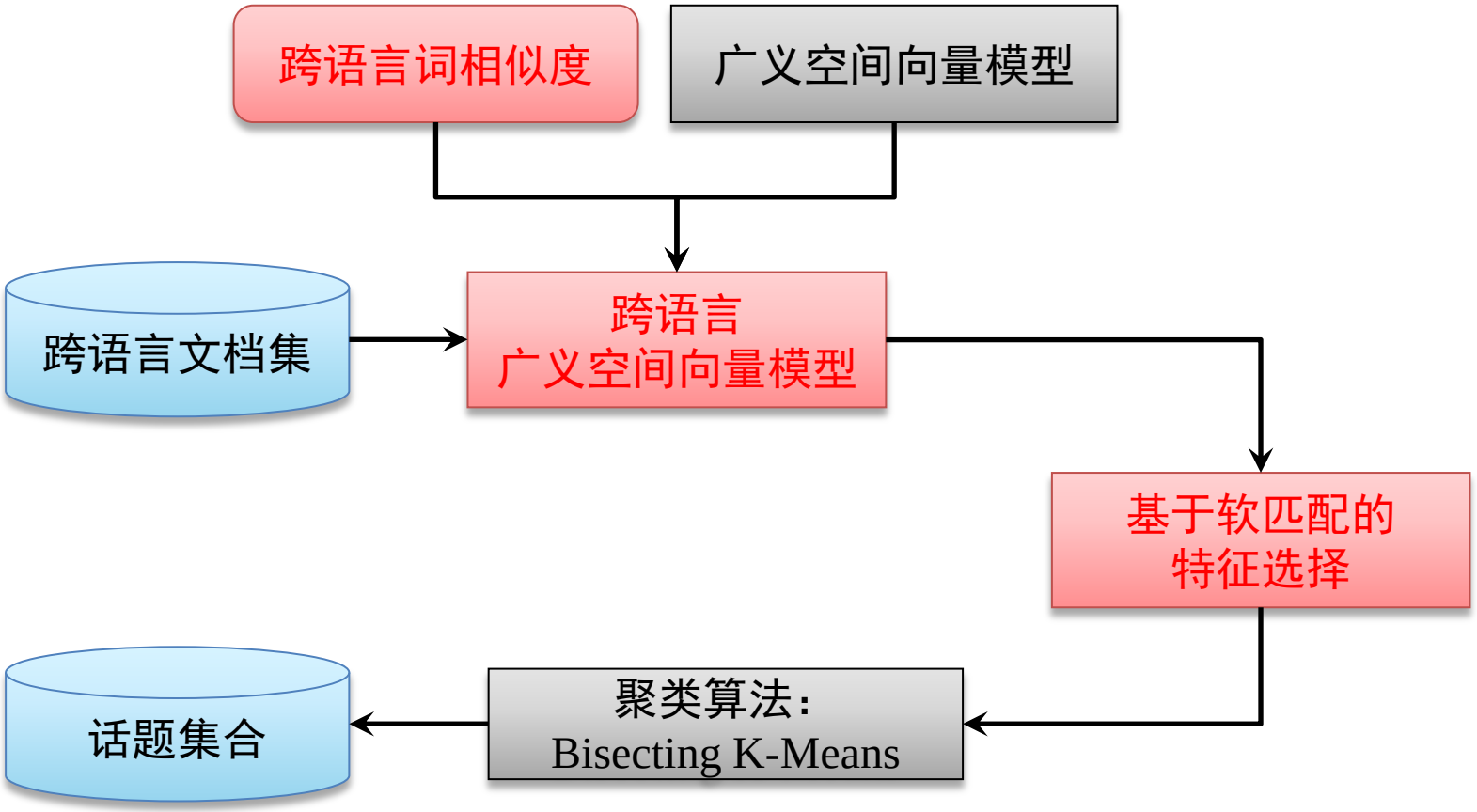
- **研究问题：跨语言文档建模**
  - 向量空间模型
    - 正交空间：“硬匹配”问题
  - 跨语言歧义加剧了“硬匹配”问题。

- **问题分析**
  - 针对跨语言歧义问题，词和它的所有翻译具有相同或者相近的表示
  - 在特征选择中需要避免“硬匹配”问题

解决思路

- 语义表示
  - 跨语言词相似度
    - 词和它的不同翻译有较大的相似度,有利于解决跨语言歧义问题
- 语义提取
  - 跨语言词相似度计算方法
- 文档表示
  - 非正交空间
    - 广义空间向量模型(Wong et al.,1985; Farahat and Kamel, 2011)
  - “软匹配”的特征选择方法
    - “软词频”，“软文档频”
    - 一个非特征的词可以影响它的近义词的权重,从而为跨语言文档建模做出贡献

跨语言广义向量空间模型（1/2）



跨语言广义向量空间模型（2/2）

- **广义空间向量模型(Wong et al.,1985; Farahat and Kamel, 2011)：非正交空间**

词的相关度
- ➢词向量的内积
- ➢词向量的长度代表词在文档上的重要性

跨语言词相似度
- ➢可以忽略相似度噪音，使相似度稀疏便于计算。

- **跨语言词相似度**
  - 基于语义资源：Hownet(Xia et al.,2011)
  - 基于统计: SOCPMI (Islam and Inkpen, 2006); COV (Farahat and Kamel, 2010).
  - 结合词典或者翻译概率

SOCPMI+翻译概率
性能最好

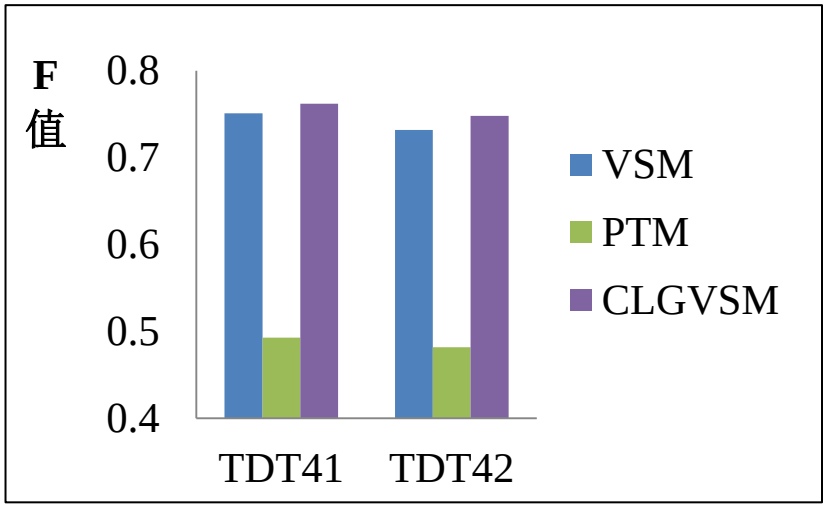
实验评测

- **实验设置**
  - 开发集：1M 平行句对
  - 词典：Hownet
  - 翻译概率：在开发集中使用Giza++获得
  - 评测集

语料	TDT41 (2002)	TDT42 (2003)
英文(话题数/文档数)	38/1270	33/617
中文(话题数/文档数)	37/657	32/560
总计(话题数/文档数)	40/1927	37/1177

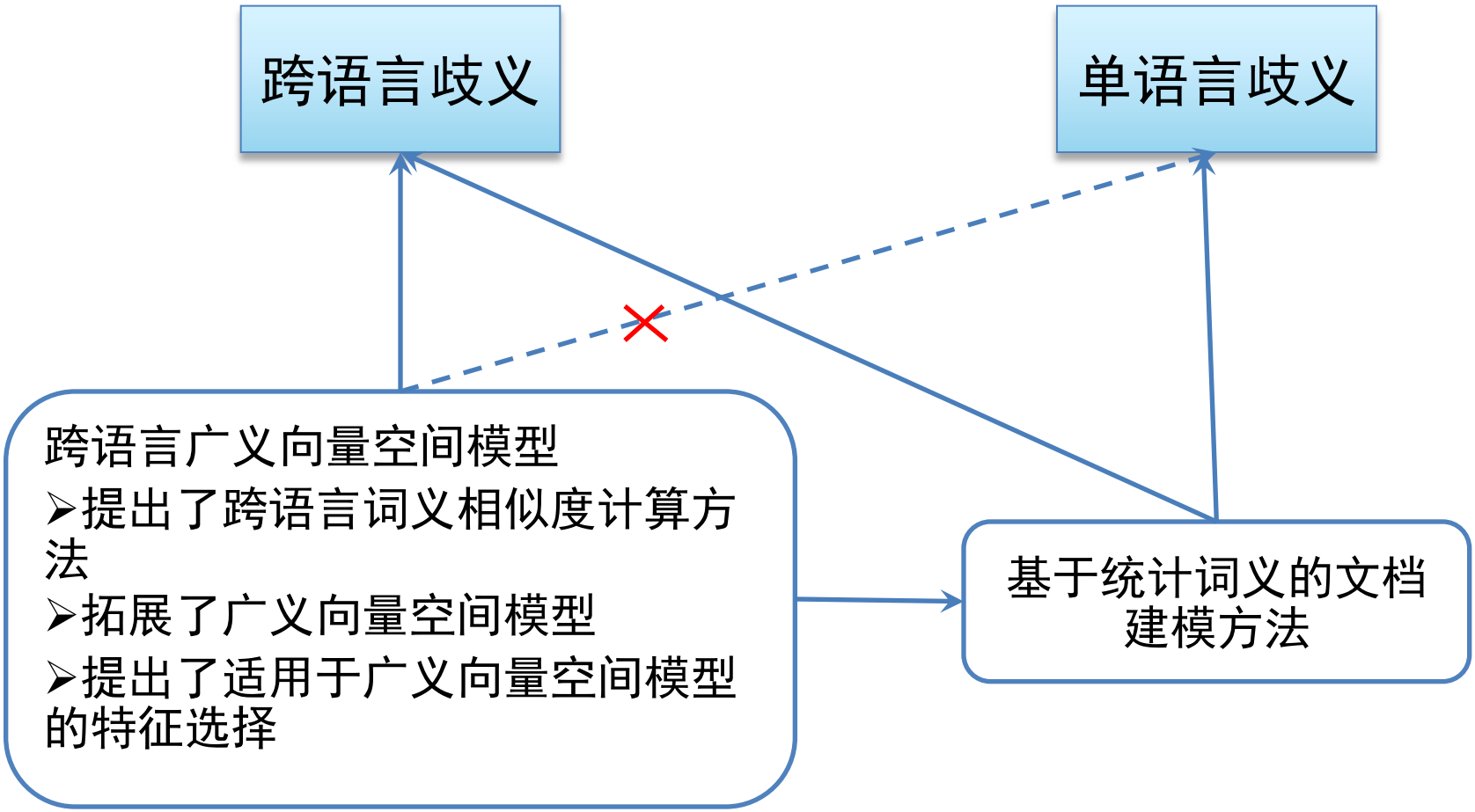
- **评测指标**
  - 准确率、召回率
  - F-值

实验结果

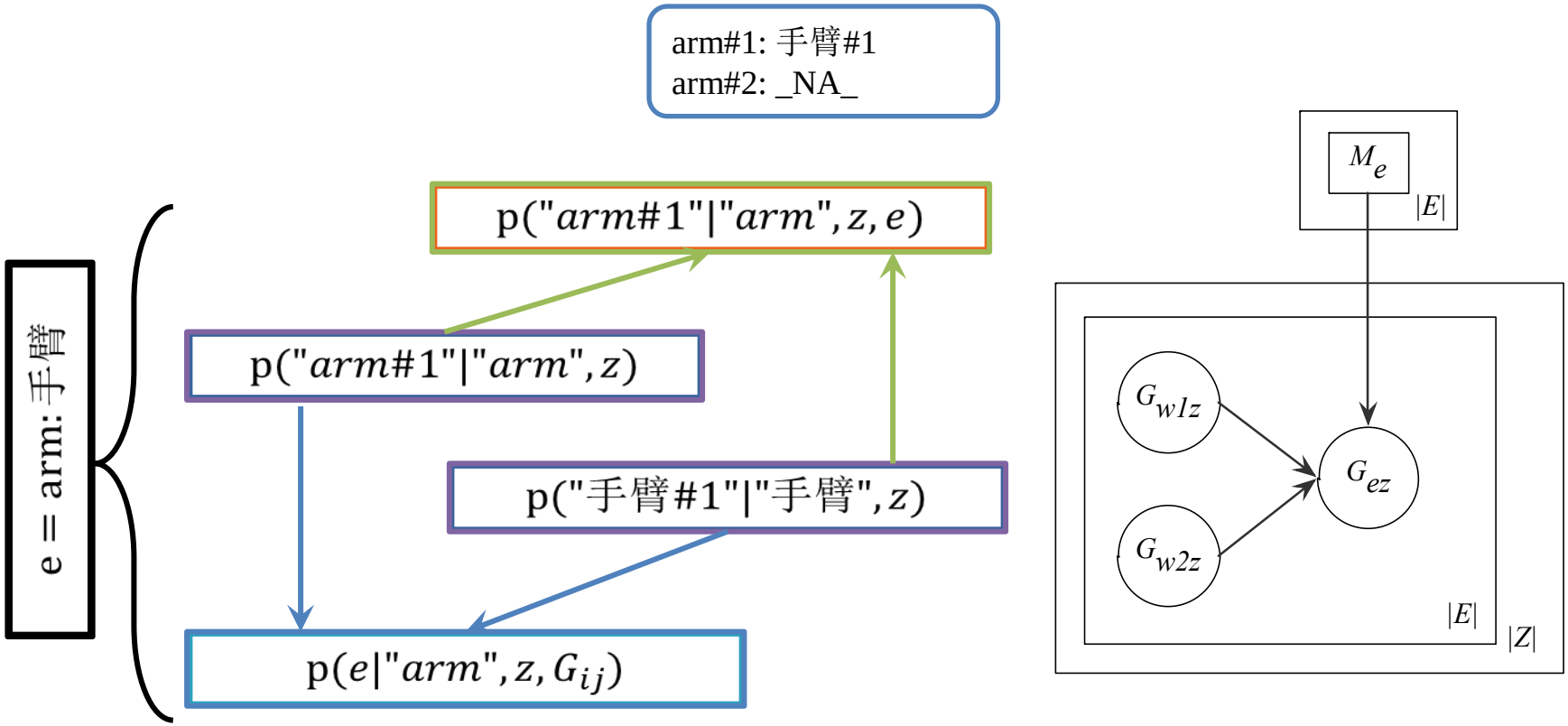


- **CLGVSM 的性能好于VSM**
  - 词相似度的贡献
- **CLGVSM 比PTM 的性能好。**
  - 文档集话题偏差

小结


跨语言歧义
单语言歧义
跨语言广义向量空间模型
➢提出了跨语言词义相似度计算方法
➢拓展了广义向量空间模型
➢提出了适用于广义向量空间模型的特征选择
基于统计词义的文档建模方法

词对词义分布



$$G_{ez} = \frac{\varphi(e, l_1) G_{w_{l_1} z} \cdot M_e^{l_1} + \varphi(e, l_2) G_{w_{l_2} z} \cdot M_e^{l_2}}{\varphi(e, l_1) + \varphi(e, l_2)}$$