基于单语主题信息的翻译模型自适应研究

厦门大学
苏劲松

统计机器翻译 和 主题模型

- 统计机器翻译 (SMT)
  - word-based -> phrase-based -> syntax-based-> …
  - 融入更多的上下文信息成为SMT研究热点之一
- 主题模型
  - 挖掘文档-词语所蕴含的潜在主题语义关系
  - PLSA，LDA等在自然语言处理领域得到广泛应用
- 目前SMT研究所使用的上下文存在信息表层，数据稀疏，局限于单个句子等缺陷，而主题模型信息能够很好地克服上述缺陷。因此，引入主题模型信息有望改善SMT的性能。

- Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, Qun Liu. Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. In *Proc.* of ACL 2012.

统计机器翻译自适应研究

- SMT自适应问题的定义
  - 当翻译文本和训练语料来自不同领域时，翻译系统性能严重下降。例如：用新闻语料训练的模型来翻译博客语料。

- 翻译模型自适应 *vs* 语言模型自适应
  - 我们主要关注翻译模型自适应
  - 方法也可以应用于语言模型自适应

相关工作

- 双语语料
  - 获取双语语料 [Hildebrand 2005; Munteanu 2005]
  - 挖掘双语语料深层次知识 [Foster 2007; Civera 2007; Lv 2007; Matsoukas 2009; Foster 2010]
  - 双语语料数量少
- 单语语料
  - 通过单语语料来生成平行语料 [Ueffing 2008; Wu 2008; Bertoldi 2009; Schwenk 2009]
  - 平行语料质量无法保证

我们的工作

- 利用单语主题信息来进行翻译模型自适应
  - 单语语料更容易获取
  - 主题信息适合表示上下文，可以克服原来上下文词，词性信息的一些缺陷

例子

Out-of-domain 平行语料

1. National bank/银行 reported earnings (经济主题)
2. The bank/银行 just closed down our overdraft (经济主题)
3. On the bank/河岸 of the Nile River (地理主题)
4. The river floods the bank/河岸 (地理主题)

例子

Out-of-domain 平行语料

1. National bank/银行 reported earnings (经济主题)
2. The bank/银行 just closed down our overdraft (经济主题)
3. On the bank/河岸 of the Nile River (地理主题)
4. The river floods the bank/河岸 (地理主题)

bank的译文和当前句子的主题密切相关

经济主题 ⇨ bank/银行

地理主题 ⇨ bank/河岸

例子

Out-of-domain 平行语料

1. National bank/银行 reported earnings (经济主题)
2. The bank/银行 just closed down our overdraft (经济主题)
3. On the bank/河岸 of the Nile River (地理主题)
4. The river floods the bank/河岸 (地理主题)

bank的译文和当前句子的主题密切相关

经济主题 ⇨ bank/银行

地理主题 ⇨ bank/河岸

最大似然估计

P(银行|bank) = P(河岸|bank) = 0.5

In-domain 单语语料

1. These figures check with the bank statement (经济主题)
2. Bank crisis is not over (经济主题)
3. The bank agreed credits to the company (经济主题)
4. Along the river bank there is a hedge (地理主题)

In-domain 单语语料

1. These figures check with the bank statement (经济主题)
2. Bank crisis is not over (经济主题)
3. The bank agreed credits to the company (经济主题)
4. Along the river bank there is a hedge (地理主题)

P(银行|bank) 和 P(河岸|bank) 仍然都是0.5吗 NO

In-domain 单语语料

1. These figures check with the bank statement (经济主题)
2. Bank crisis is not over (经济主题)
3. The bank agreed credits to the company (经济主题)
4. Along the river bank there is a hedge (地理主题)

P(银行|bank) 和 P(河岸|bank) 仍然都是0.5吗 NO

直觉上，in-domain单语语料中更多句子属于经济主题，P(银行|bank) = 3/4 P(河岸|bank) = 1/4更为合理。

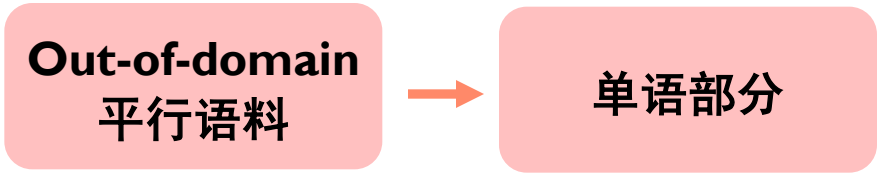
- 规则译文选择和规则所在文本的主题信息密切相关
- 同一条规则在不同领域语料中具有不同的主题概率分布
- 如果可以量化主题信息对规则翻译概率的影响，那么可以根据规则在目标翻译领域中的主题概率分布来调整规则翻译概率

翻译模型自适应

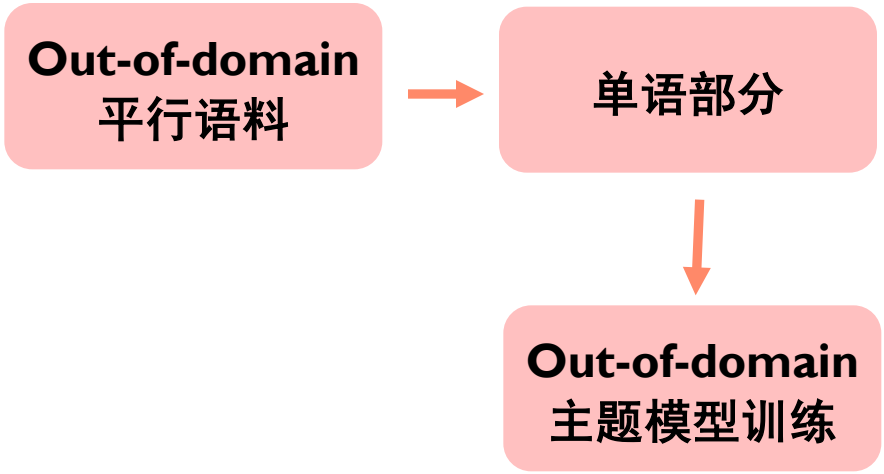
翻译模型自适应

Out-of-domain
平行语料

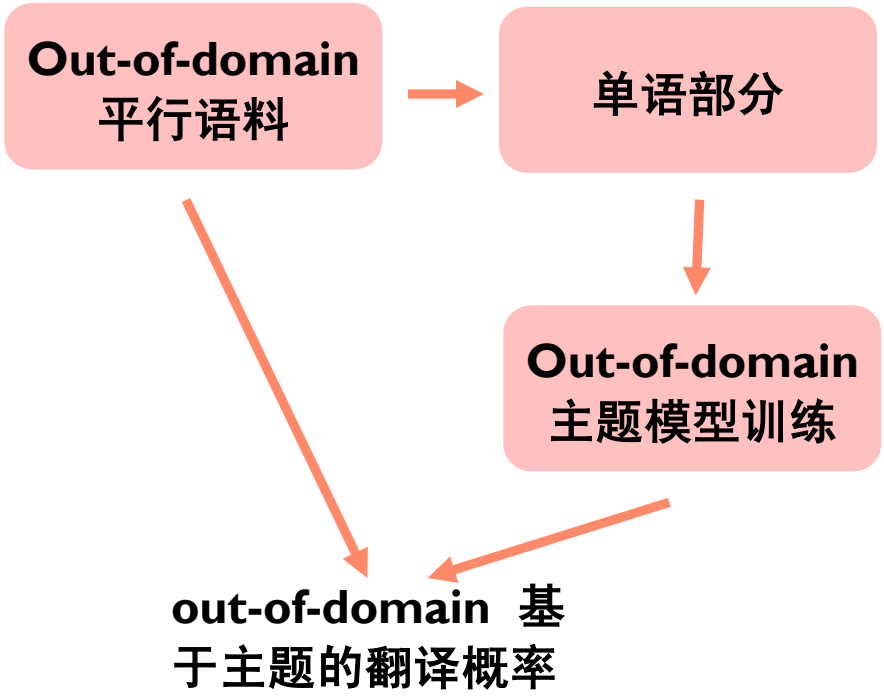
翻译模型自适应



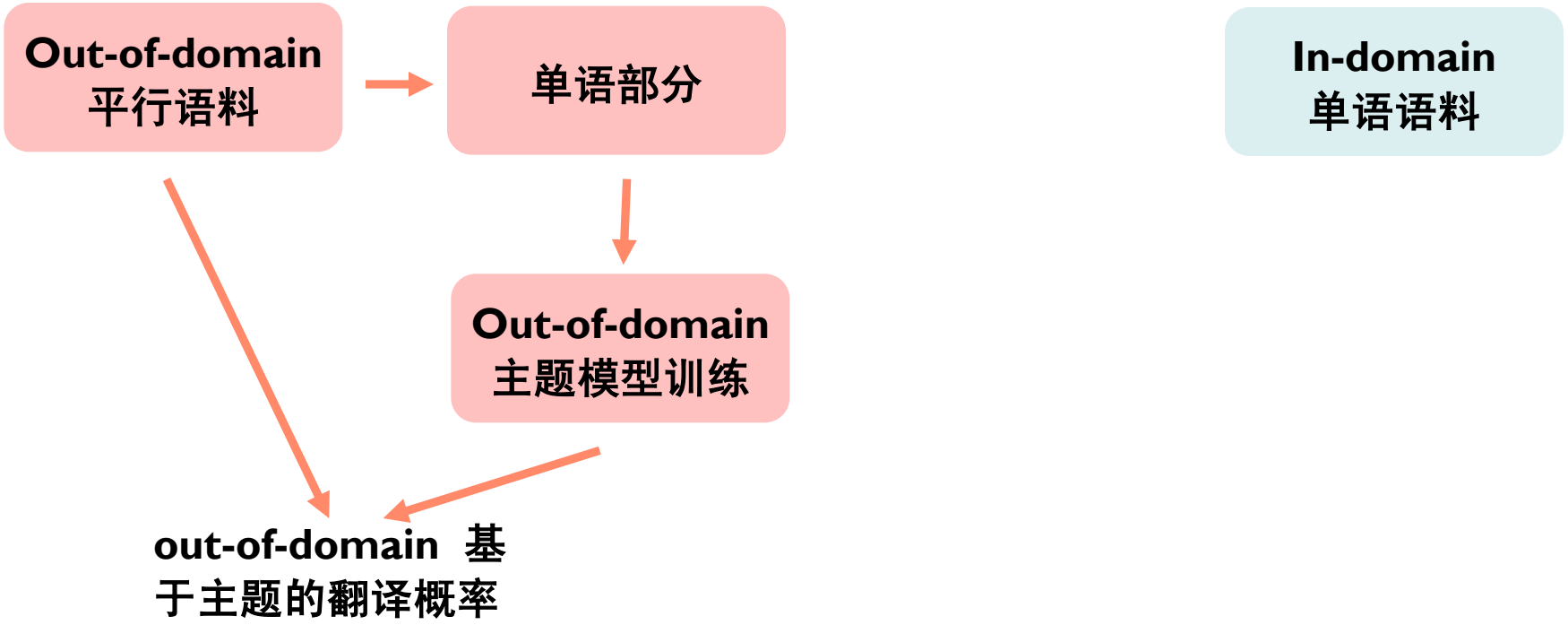
翻译模型自适应



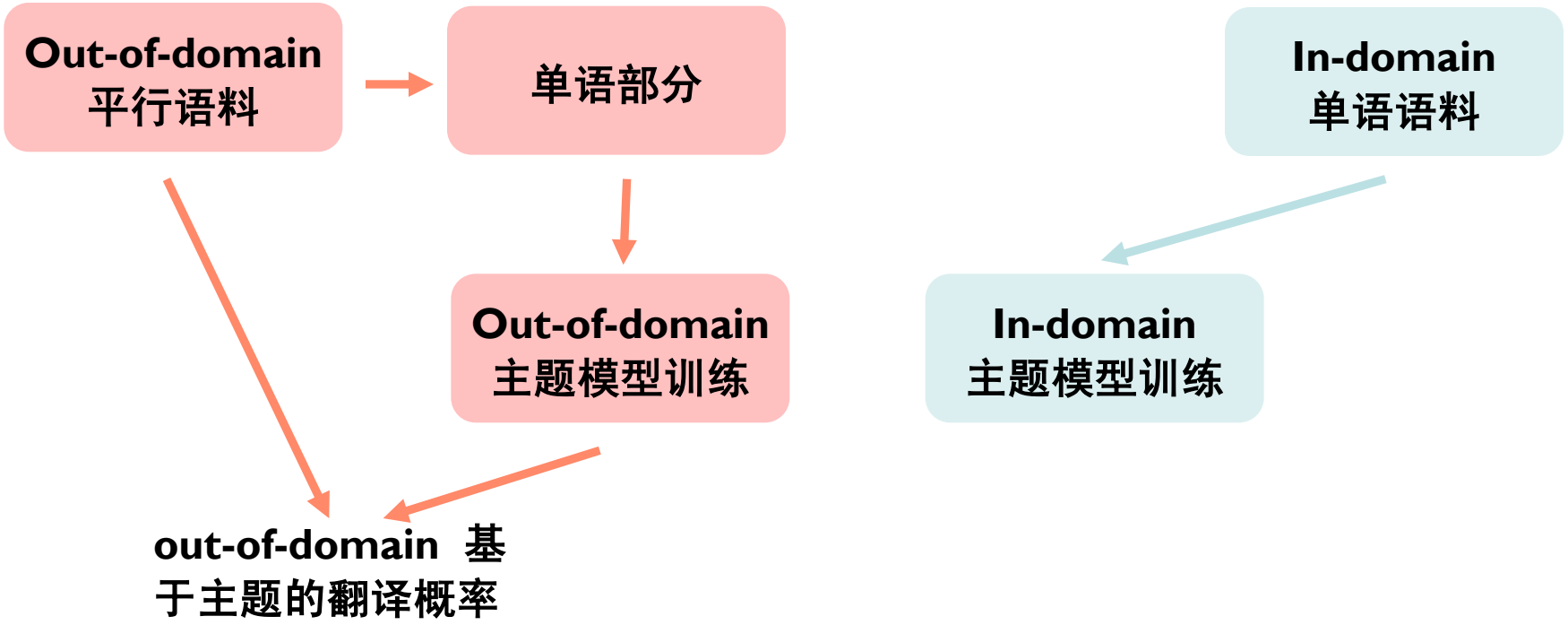
翻译模型自适应


Out-of-domain
平行语料
单语部分
Out-of-domain
主题模型训练
out-of-domain 基
于主题的翻译概率

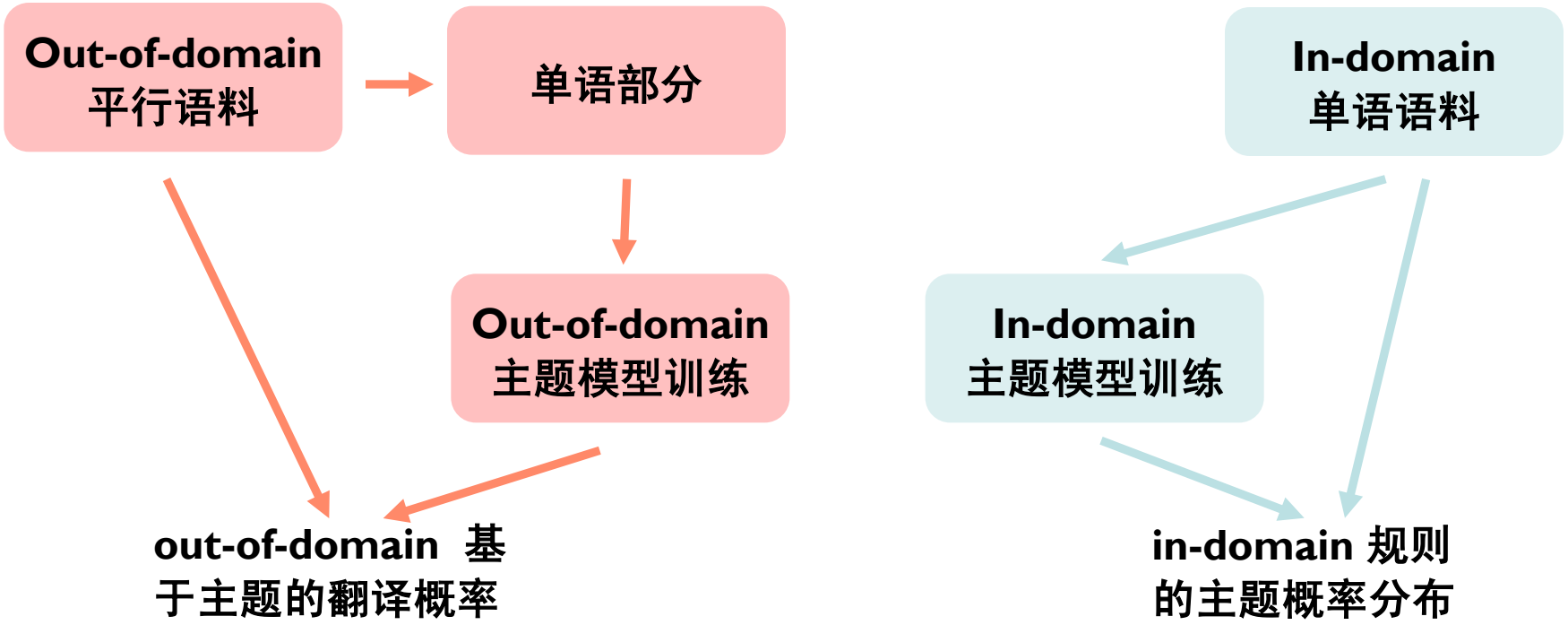
翻译模型自适应


Out-of-domain
平行语料
单语部分
In-domain
单语语料
Out-of-domain
主题模型训练
out-of-domain 基
于主题的翻译概率

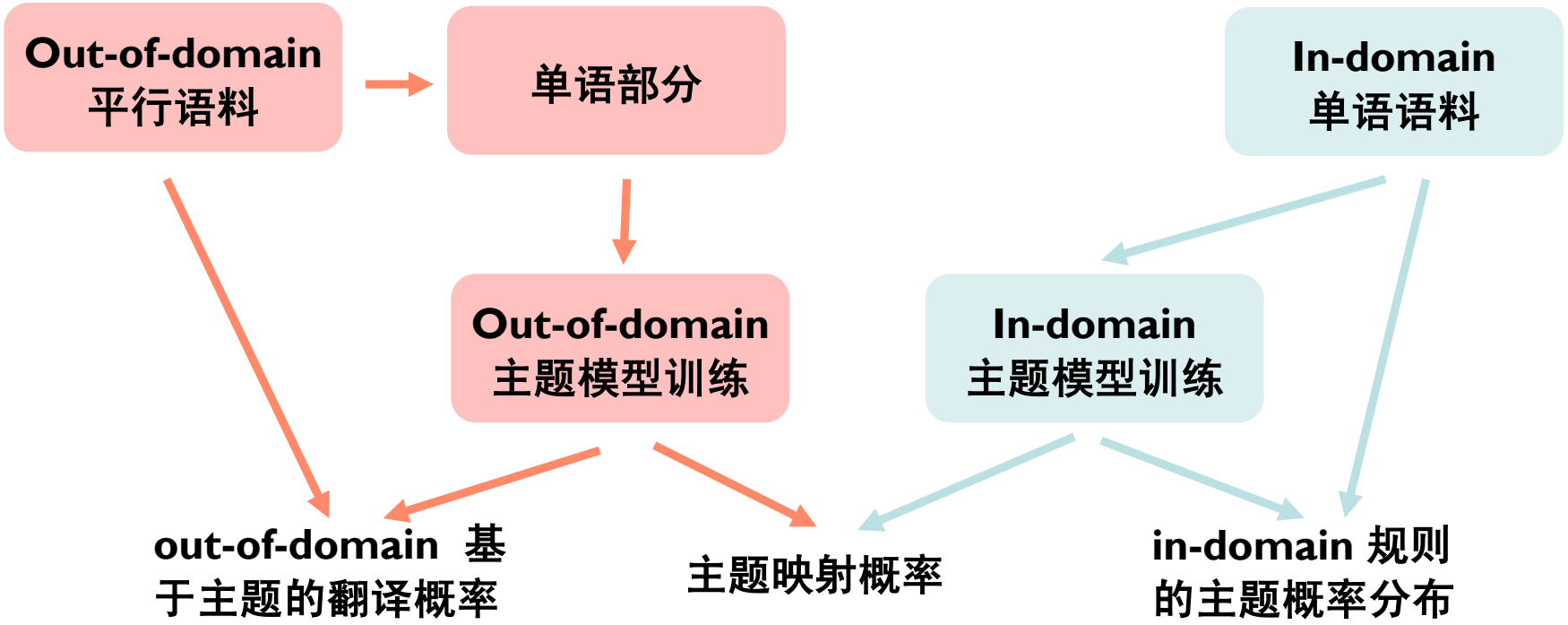
翻译模型自适应


Out-of-domain
平行语料
单语部分
In-domain
单语语料
Out-of-domain
主题模型训练
In-domain
主题模型训练
out-of-domain 基
于主题的翻译概率

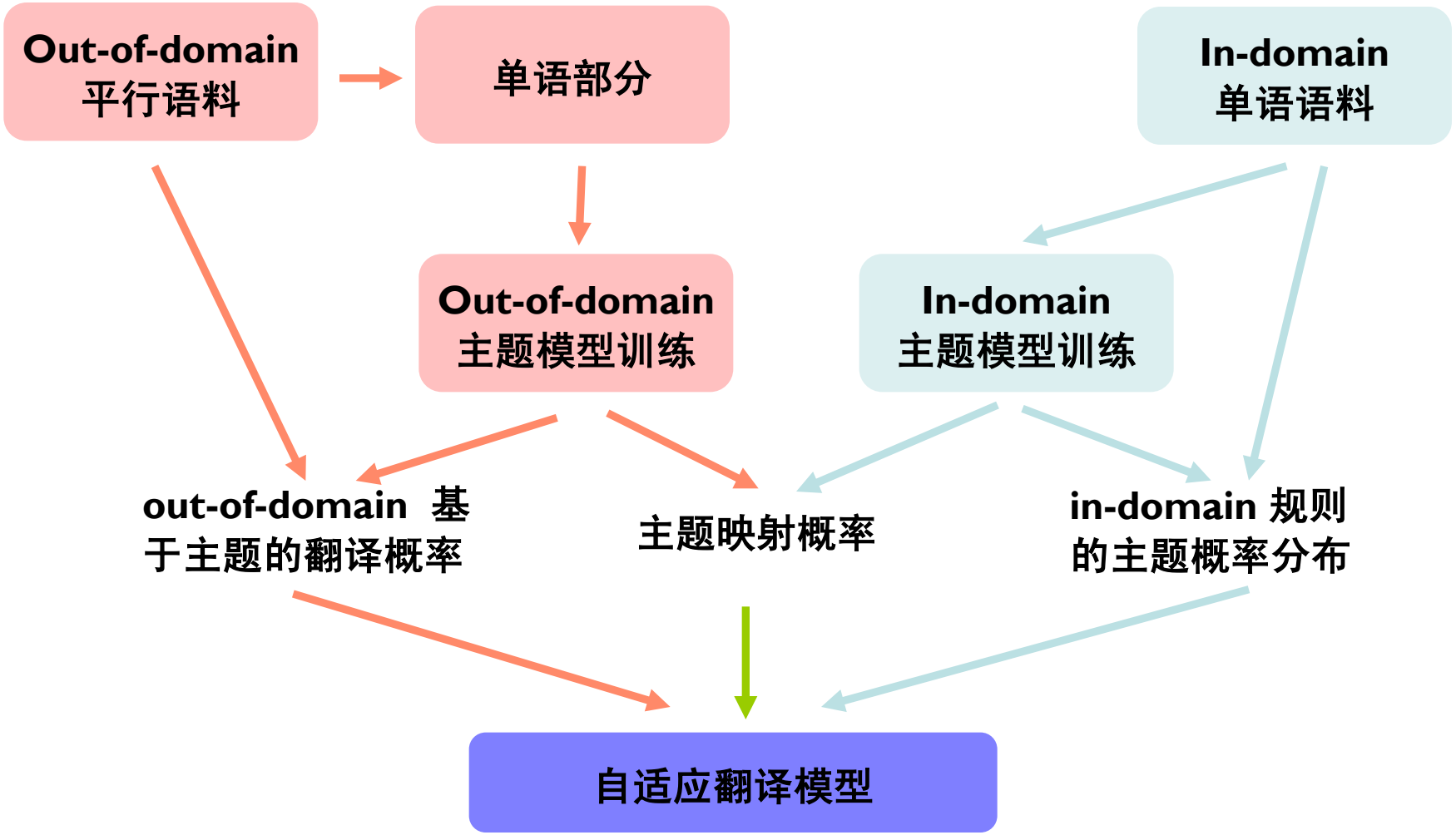
翻译模型自适应


Out-of-domain
平行语料
单语部分
In-domain
单语语料
Out-of-domain
主题模型训练
In-domain
主题模型训练
out-of-domain 基
于主题的翻译概率
in-domain 规则
的主题概率分布

翻译模型自适应


Out-of-domain
平行语料
单语部分
In-domain
单语语料
Out-of-domain
主题模型训练
In-domain
主题模型训练
out-of-domain 基
于主题的翻译概率
主题映射概率
in-domain 规则
的主题概率分布

翻译模型自适应


Out-of-domain
平行语料
单语部分
In-domain
单语语料
Out-of-domain
主题模型训练
In-domain
主题模型训练
out-of-domain 基
于主题的翻译概率
主题映射概率
in-domain 规则
的主题概率分布
自适应翻译模型

规则翻译概率

规则翻译概率

$$\phi(\tilde{e}|\tilde{f}) = \sum_{t_{f\_out}} \sum_{t_{f\_in}} \phi(\tilde{e}|\tilde{f}, t_{f\_out}) \cdot p(t_{f\_out}|t_{f\_in}) \cdot p(t_{f\_in}|\tilde{f})$$

规则翻译概率

$$\phi(\tilde{e}|\tilde{f}) = \sum_{t_{f\_out}} \sum_{t_{f\_in}} \phi(\tilde{e}|\tilde{f}, t_{f\_out}) \cdot p(t_{f\_out}|t_{f\_in}) \cdot p(t_{f\_in}|\tilde{f})$$

out-of-domain 基于主题的规则翻译概率

规则翻译概率

主题映射概率

$$\phi(\tilde{e}|\tilde{f}) = \sum_{t_{f\_out}} \sum_{t_{f\_in}} \phi(\tilde{e}|\tilde{f}, t_{f\_out}) \cdot p(t_{f\_out}|t_{f\_in}) \cdot p(t_{f\_in}|\tilde{f})$$

out-of-domain 基于主题的规则翻译概率

规则翻译概率

主题映射概率

$$\phi(\tilde{e}|\tilde{f}) = \sum_{t_{f\_out}} \sum_{t_{f\_in}} \phi(\tilde{e}|\tilde{f}, t_{f\_out}) \cdot p(t_{f\_out}|t_{f\_in}) \cdot p(t_{f\_in}|\tilde{f})$$

out-of-domain 基于主题的规则翻译概率

in-domain 规则的主题概率分布

Out-of-domain 基于主题的规则翻译概率 $\phi(\tilde{e}|\tilde{f},t_{f_{out}})$

- 最大似然估计
- 考虑规则的主题概率分布

$$\phi(\tilde{e}|\tilde{f},t_{f\_out})=\frac{\sum_{\langle \mathrm{f},\mathrm{e}\rangle\in C_{out}} count_{\langle \mathrm{f},\mathrm{e}\rangle}(\tilde{f},\tilde{e})\cdot P(t_{f\_out}|\mathrm{f})}{\sum_{\tilde{e}'}\sum_{\langle \mathrm{f},\mathrm{e}\rangle\in C_{out}} count_{\langle \mathrm{f},\mathrm{e}\rangle}(\tilde{f},\tilde{e}')\cdot P(t_{f\_out}|\mathrm{f})}$$

主题映射概率 $p(t_{f_{out}} | t_{f_{in}})$

- 主题模型通过词和主题到词的生成概率来反映主题
- 把词作为中间变量来进行主题映射

$$p(t_{f_{out}} | t_{f_{in}}) = \sum_{w_f \in C_{f\_out} \cap C_{f\_in}} p(t_{f_{out}} | w_f) \cdot p(w_f | t_{f_{in}})$$

$p(t_{f_{out}} | w_f)$：最大似然估计 **out-of-domain** 语料

$p(w_f | t_{f_{in}})$：**in-domain** 主题模型

In-domain 规则主题概率分布 $p(t_{f_{in}}|\tilde{f})$

- Out-of-domain平行语料中的规则并没有都在in-domain单语语料库中出现
- 使用词主题概率分布来进行平滑

$$p(t_{f_{in}}|\tilde{f}) \approx \theta \cdot p_{mle}(t_{f_{in}}|\tilde{f}) + (1-\theta) \cdot p_{word}(t_{f_{in}}|\tilde{f})$$

最大似然估计

基于词的主题概率分布
考察两种平滑方法
“Noise-OR方法”和“Averaging方法”

两种平滑方法

- Noise-OR 方法
  - 类似于Zens et al.(2004)提出的翻译概率计算方法

$$p_{word}(t_{f_{in}}|\tilde{f}) = 1 - p_{word}(\overline{t}_{f_{in}}|\tilde{f}) \approx 1 - \prod_{w_f \in \tilde{f}} p(\overline{t}_{f_{in}}|w_f)$$

- Averaging 方法

$$p_{word}(t_{f_{in}}|\tilde{f}) \approx \sum_{w_f \in \tilde{f}} p(t_{f_{in}}|w_f) / |\tilde{f}|$$

词汇化概率计算

- 把词看为长度为1的短语
- 采用相类似的方法来计算词汇化概率

$$\mathrm{w}(e|f) = \sum_{t_{f\_out}} \sum_{t_{f\_in}} \mathrm{w}(e|f, t_{f\_out}) \cdot p(t_{f\_out}|t_{f\_in}) \cdot p(t_{f\_in}|f)$$

实验设置

- Out-of-domain 平行语料
  - FBIS + Hansards part of LDC2004T07（54.6K 文档 / 1M 句对）
- In-domain 单语语料
  - 2009搜狐博客语料
  - 英文博客语料(Schler 2006)
- 开发测试集
  - dev：NIST06 测试集的网络部分（27 文档/1048 句子）
  - test：NIST08 测试集的网络博客部分（33 文档/666 句子）

- 语言模型
  - 英文博客语料 4元
  - Gigaword 新华语料 4元
- 解码器
  - MOSES (Koehn 2007)
- 主题模型工具
  - HTMM (Gruber 2007)
  - 句子中词具有相同主题概率分布
  - 使用默认参数，主题个数为50

实验一

- 考察平滑方法效果

Adaptation Method	(Dev) MT06 Web	(TST) MT08 Weblog
Baseline	30.98	20.22
Noisy-OR(5K)	31.16	20.45
Averaging(5K)	31.51	20.54
Noisy-OR(40K)	31.87	20.76
Averaging(40K)	31.89	21.11

Averaging 平滑方法更好。使用Noisy-OR 方法，规则概率分布更集中在某些主题上，可能导致更大概率估计偏差。

实验二

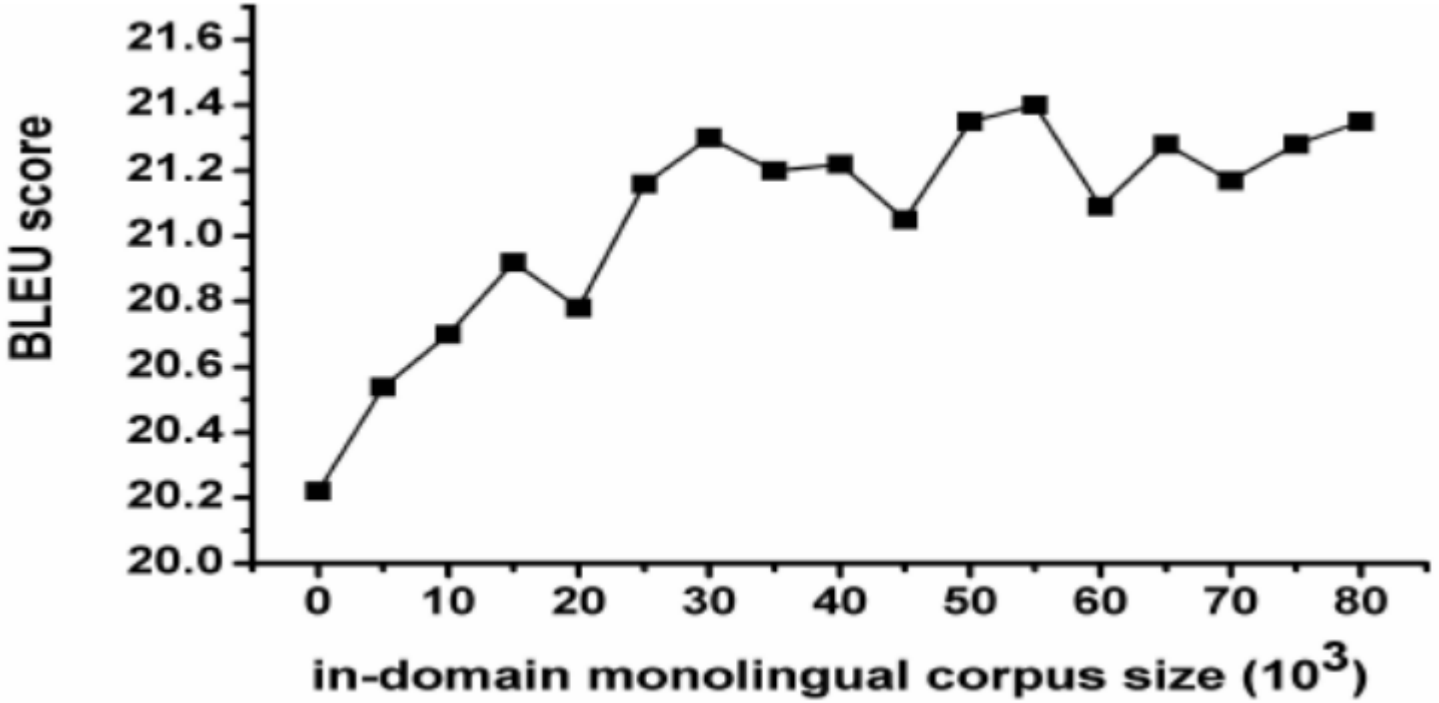
- 考察使用两个规则表的效果

Adaptation Method	(Dev) MT06 Web	(TST) MT08 Weblog
Baseline	30.98	20.22
AdapBp(5K)	31.51	20.54
+OutBp	31.84	20.70
AdapBp(40K)	31.89	21.11
+OutBp	32.05	21.20

同时使用两个规则表可以有效地较少自适应规则表的概率偏差。

实验三

- 考察in-domain单语语料规模对方法效果的影响



结论

- 单语语料的主题模型信息可以用于翻译模型自适应.

- 下一步工作
  - 与更多方法比较，例如, self-training.
  - 更好的平滑方法.
  - 合理的主题个数估计.

谢　谢