一种从可比数据中挖掘平行资源的有效框架

向露，周玉，宗成庆

中 国 科 学 院 自 动 化 研 究 所

模 式 识 别 国 家 重 点 实 验 室

提纲

- 双语平行资源的获取
- 传统处理方法与不足
- 两级平行资源获取框架
  - 平行句对的抽取方法
  - 平行片段的抽取方法
- 实验结果与实例分析
- 总结与展望

提纲

- **双语平行资源的获取**
- 传统处理方法与不足
- 两级平行资源获取框架
  - 平行句对的抽取方法
  - 平行片段的抽取方法
- 实验结果与实例分析
- 总结与展望

双语平行资源的获取

- 平行资源(Parallel Corpus)
  - 重要性的资源
    - 统计机器翻译

$$P(e) = \arg\max_{e'} P(e' \mid f)$$
$$= \arg\max_{e'} P(f \mid e') \cdot P(e')$$

双语平行语料
（f_1,e_1),(f_2,e_2)...

翻译模型
$P(f \mid e)$

决定了翻译
知识覆盖率

双语平行资源的获取

- 平行资源的现状
  - 资源稀缺
    - 规模、领域、语言对等受限，影响机器翻译性能
    - 手动构建耗费巨大的人力、物力
- 平行资源的获取
  - 自动获取？
    - 获取网络双语文本中的双语平行资源
      - 双语平行句对
      - 双语平行片段

双语平行资源的获取

- 以“机器翻译”为例

简单来说，机器翻译是通过将一个自然语言的字辞取代成另一个语言的字辞。借由使用语料库的技术，可达成更加复杂的自动翻译，包含可更佳的处理不同的文法结构、词汇辨识、惯用语的对应等。

目前的机器翻译软件通常可允许针对特定领域或是专业（例如天气预报） 来加以客制化，目的在于将词汇的取代缩小于该特定领域的专有名词上，以借此改进翻译的结果。这样的技术针对一些使用较正规或是较制式化陈述方式的领域来说 特别有效。

On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text because recognition of whole phrases and their closest counterparts in the target language is needed.

Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used.

双语平行资源的获取

- 以“机器翻译”为例

简单来说，机器翻译是通过将一个自然语言的字辞取代成另一个语言的字辞。借由使用语料库的技术，可达成更加复杂的自动翻译，包含可更佳的处理不同的文法结构、词汇辨识、惯用语的对应等。

目前的机器翻译软件通常可允许针对特定领域或是专业（例如天气预报） 来加以客制化，目的在于将词汇的取代缩小于该特定领域的专有名词上，以借此改进翻译的结果。这样的技术针对一些使用较正规或是较制式化陈述方式的领域来说 特别有效。

On a basic level, MT performs simple substitution of words in one natural language for words in another but that
alone usually can … good
translation of a text … n of
whole phrases … sest
counterparts in the target language is needed.

平行句对

Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used.

双语平行资源的获取

- 以“机器翻译”为例

简单来说，机器翻译是通过将一个自然语言的字辞取代成另一个语言的字辞。借由使用语料库的技术，可达成更加复杂的自动翻译，包含可更佳的处理不同的文法结构、词汇辨识、惯用语的对应等。

目前的机器翻译软件通常可允许针对特定领域或是专业（例如天气预报） 来加以客制化，目的在于将词汇的取代缩小于该特定领域的专有名词上，以借此改进翻译的结果。这样的技术针对一些使用较正规或是较制式化陈述方式的领域来说 特别有效。

On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text … n of whole phrases … sest counterparts in th… e is needed.

平行片段

Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used.

双语平行资源的获取

- 网络上大量存在的双语文本的特点：
  - 可比文本
    - Wikipedia 、双语新闻等
  - 在这些双语文本中，不仅含有双语平行句对，还含有大量的平行片段，而这些资源对于机器翻译都是有帮助的
- 如何判定两个句子是平行句对？
- 如何抽取平行片段？

提纲

- 双语平行资源的获取
- **传统处理方法与不足**
- 两级平行资源获取框架
  - 平行句对的抽取方法
  - 平行片段的抽取方法
- 实验结果与实例分析
- 总结与展望

传统处理方法与不足

- 已有工作
  - 篇章对齐
    - 跨语言检索 (Munteanu and Marcu, 2005)
    - “inter-wiki”对Wikipedia篇章进行对齐 (Smith et al., 2010)
  - 平行句对判定
    - 扩展传统的句子对齐算法 (Zhao and Vogel, 2002)
    - 计算cosine相似度 (Fung and Cheung, 2004)
    - 分类器 (Munteanu and Marcu, 2006)
    - ……

传统处理方法与不足

- 已有工作(cont.)
  - 平行片段抽取
    - 传统短语抽取方法
      - 与词对齐一致性特征抽取短语 (Koehn, 2003)
    - 基于词的信号过滤方法 (Munteanu and Marcu, 2006)
    - 层次对齐模型 (Riesa and Marcu, 2012)
    - ……

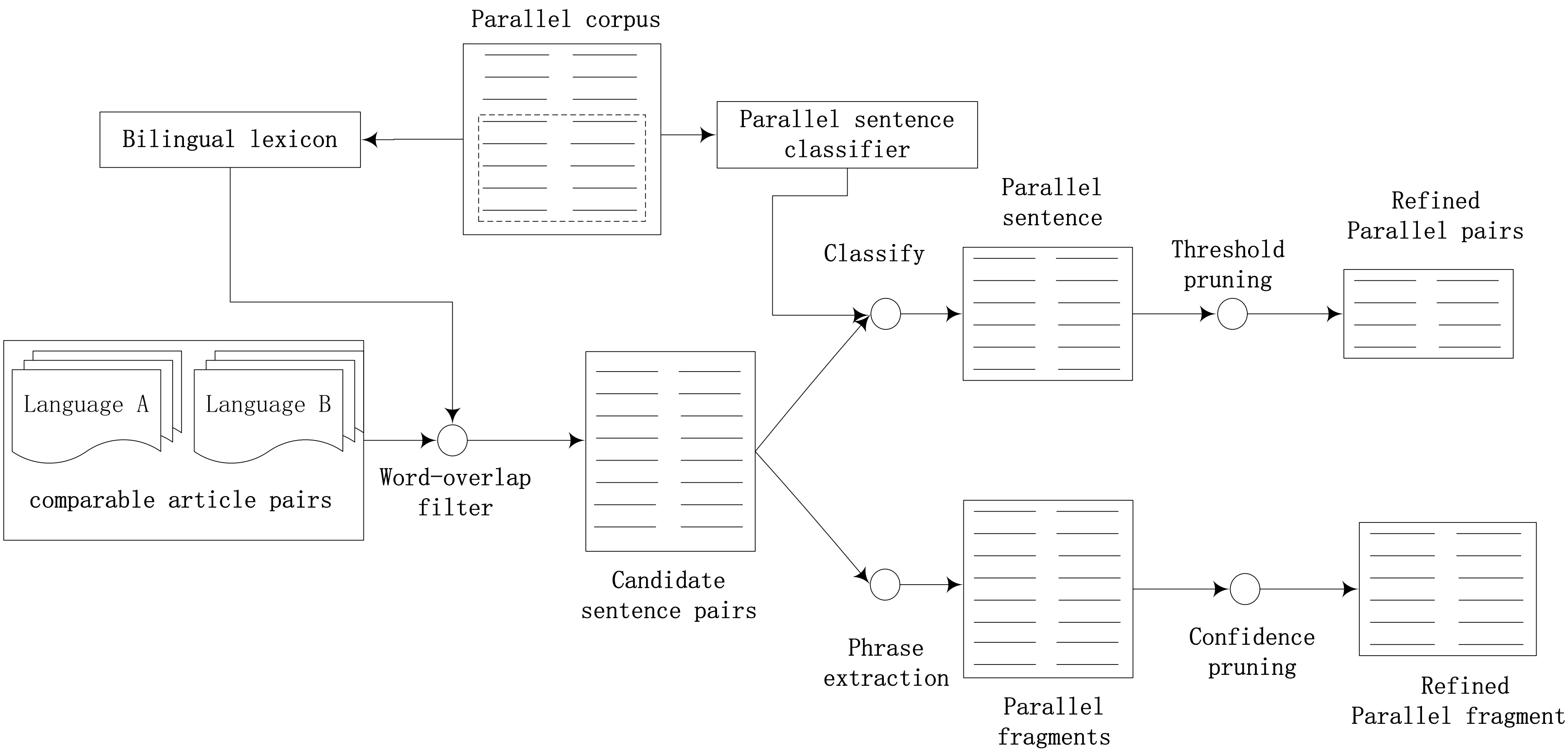
传统处理方法与不足

- 缺陷
  - 只是单一的抽取平行句对或者平行片段
  - 平行句对的抽取
    - 对于特征的选择没有进行深入的分析
  - 平行片段的抽取
    - 句对中存在很多不是互译的部分，传统的词对齐方法并不适用
  - 我们需要找到统一的框架来进行平行句对和片段的抽取

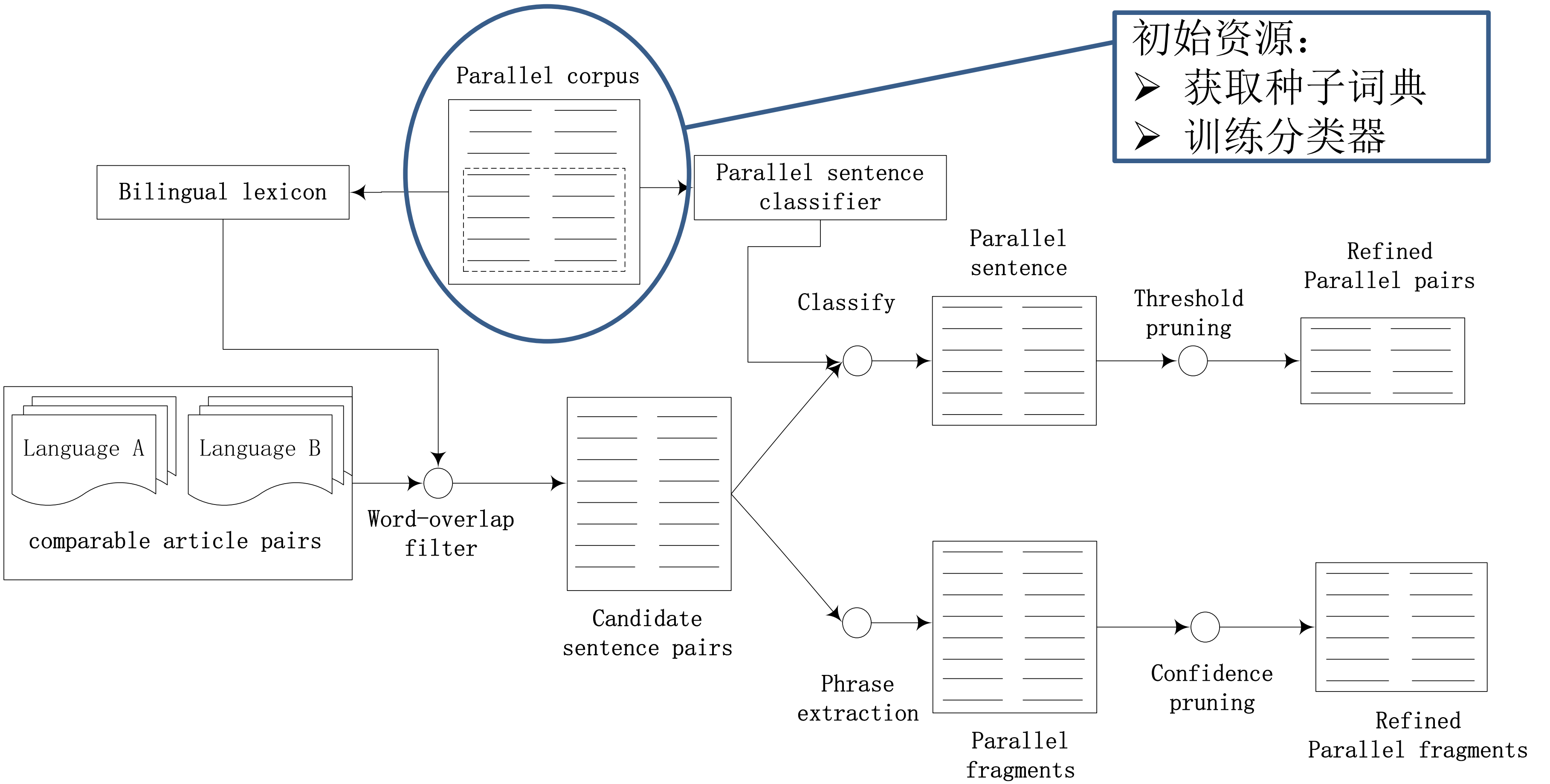
提纲

- 双语平行资源的获取
- 传统处理方法与不足
- **两级平行资源获取框架**
  - 平行句对的抽取方法
  - 平行片段的抽取方法
- 实验结果与实例分析
- 总结与展望

两级平行资源获取框架


Parallel corpus
Bilingual lexicon
Parallel sentence classifier
Parallel sentence
Refined Parallel pairs
Classify
Threshold pruning
Language A
Language B
comparable article pairs
Word-overlap filter
Candidate sentence pairs
Phrase extraction
Parallel fragments
Confidence pruning
Refined Parallel fragments

两级平行资源获取框架



初始资源：
➢ 获取种子词典
➢ 训练分类器
Parallel corpus
Bilingual lexicon
Parallel sentence classifier
Parallel sentence
Refined Parallel pairs
Classify
Threshold pruning
Language A
Language B
comparable article pairs
Word-overlap filter
Candidate sentence pairs
Phrase extraction
Parallel fragments
Confidence pruning
Refined Parallel fragments

两级平行资源获取框架

- 候选句对选择
  - 思想
    - 如果两个句子是平行的或者是含有平行片段的，那么这两个句子至少有一部分是互译的
  - 使用word-overlap来进行过滤
    - 根据种子词典进行过滤
    - 互译单词覆盖率大于阈值的句对作为候选句对

提纲

- 双语平行资源的获取
- 传统处理方法与不足
- 两级平行资源获取框架
  - 平行句对的抽取方法
  - 平行片段的抽取方法
- 实验结果与实例分析
- 总结与展望

平行句对的获取方法

- 将平行句对的获取看作一个分类问题
  - 二分类（平行|非平行）

- 训练数据
  - 从种子语料中构建训练语料
    - 从种子语料中选取部分句对作为正例（平行）
    - 通过这部分正例产生负例（非平行）

平行句对的获取方法

- 选取的特征
  - Munteanu&Marcu(2005)中的特征：
    - 一般特征
      - 源端和目标端句子长度，长度差值以及长度比；
      - 互译单词所占的比例；
    - 词对齐特征
      - 没有链接的词的个数以及其所占比例；
      - 三个最大的繁衍率的值；
      - 最长的连续对齐跨度的长度；
      - 最长的未对齐跨度的长度

平行句对的获取方法

- 选取的特征(cont.)
  - 新的特征
    - 词对齐打分
    - 句对的余弦相似度
    - “翻译哨兵”特征

平行句对的获取方法

- 一个例子
  - “翻译哨兵”特征

希腊人 期许 着 仔细 的 论点 ， 但 在 牛顿 的 时代 ， 所 使用 的 方法 则 较 不 严谨 。
the greeks expected detailed arguments , but at the time of isaac newton the methods employed were less rigorous .

数学家 也 研究 纯 数学 ， 也 就 是 数学 本身 ， 而 不 以 任何 实际 应用 为 目标 。
a distinction is often made between pure mathematics and applied mathematics .

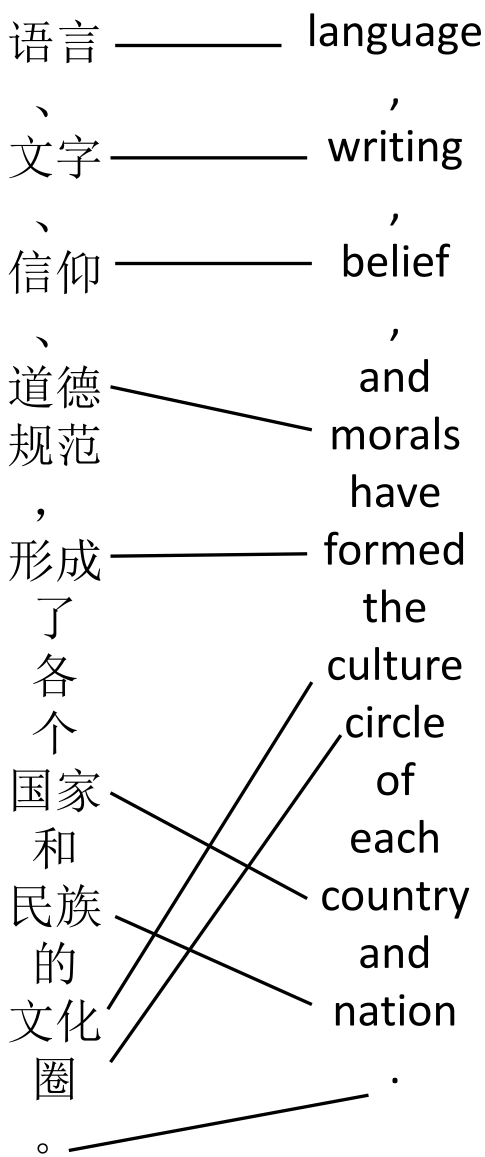
提纲

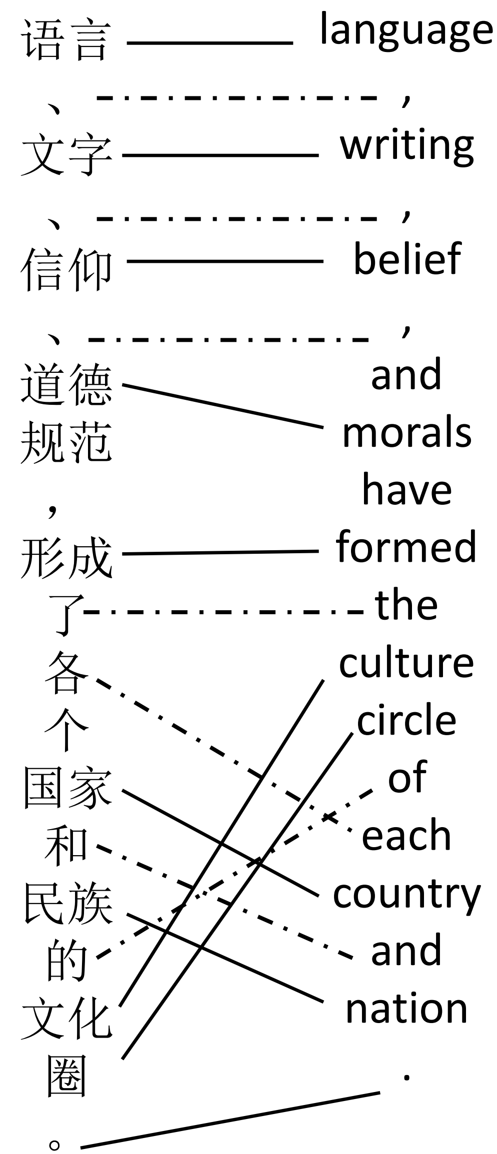
- 双语平行资源的获取
- 传统处理方法与不足
- 两级平行资源获取框架
  - 平行句对的抽取方法
  - **平行片段的抽取方法**
- 实验结果与实例分析
- 总结与展望

平行片段的获取方法

- 词对齐模型
  - IBM Model-1词对齐模型的不足
    - 对于高频词，比如“ the”、“的”等无法处理
    - 一端有多个译项时，不能确定对齐点
  - 提出两步词对齐模型
    - 阶段一：对非停用词进行对齐
    - 阶段二：对停用词进行对齐

改进的词对齐模型


语言
、
文字
、
信仰
、
道德
规范
，
形成
了
各
个
国家
和
民族
的
文化
圈
。
language
,
writing
,
belief
,
and
morals
have
formed
the
culture
circle
of
each
country
and
nation
.


语言
、
文字
、
信仰
、
道德
规范
，
形成
了
各
个
国家
和
民族
的
文化
圈
。
language
,
writing
,
belief
,
and
morals
have
formed
the
culture
circle
of
each
country
and
nation
.

平行片段的获取方法

以“一 国 两 制”的 办法 解决 台湾 问题， 才 能 最大 限度 地 寻求 两 岸 利益 的 公分母 。
we have hoped for many years to use the formula of “ one country , two systems ” to peacefully resolve the taiwan issue .

- 词对齐：源端->目标端，目标端->源端
- 词对齐取交集

平行片段的抽取方法

- 遍历词对齐，抽取满足下列条件的短语：
  - 源端和目标端的短语长度不小于3
  - 源端短语内的词只能对齐到目标端短语中的词；源端短语之外的词只能对齐到目标端短语之外的词；
  - 短语片段能够包含少量未对齐的词
  - 边界上未对齐的词只能为停用词

提纲

- 双语平行资源的获取
- 传统处理方法与不足
- 两级平行资源获取框架
  - 平行句对的抽取方法
  - 平行片段的抽取方法
- 实验结果与实验分析
- 总结与展望

实验设置

- 语料
  - 抽取对象：中英Wikipedia，199,984篇章对
  - 初始语料
    - LDC，208.5万句对
    - Wikipedia标题，23.6万
- 词典
  - 使用初始语料学习词典
    - 32.5万LLR lexicon

平行句对的抽取实验

- 实验设置
  - 分类器：最大熵模型
  - 训练集：
    - 5800 平行句对 作为正例
    - 5800 非平行句对作为负例
  - 测试集：
    - CWMT 2011：1000 正例 和 1000负例

平行句对的抽取实验

Features	Precision	Recall	F-score
F1:General features	0.7021	0.608	0.6516
F1+F2: no connection	**0.7414**	**0.889**	**0.8085**
F1~F2+F3: fertility	0.7395	0.9	0.8119
F1~F3+F4:contiguous connected span	0.7300	0.933	0.8191
F1~F4+F5: unconnected substring	0.7288	0.941	0.8214
F1~F5+F6: intersection	0.7344	0.91	0.8128
F1~F6+F7: union	0.7435	0.928	0.8256
F1~F7+F8: refined	0.7606	0.896	0.8227

平行句对的抽取实验

Features	Precision	Recall	F-score
F1:General features	0.7021	0.608	0.6516
F1+F2: no connection	0.7414	0.889	0.8085
F1+F2+F9:log probability	0.7429	0.893	0.8110
F1+F2+F9+F10:cosine similarity	0.7439	0.895	0.8125
F1+F2+F9+F10+F11:sentinels	**0.7466**	**0.899**	**0.8157**

平行句对的抽取实验

- 对100,000个句对抽取特征：

Features	F-score	Time (s)
F1~F5	0.8214	2259.26
F1~F8	0.8227	4097.63
F1+F2+F9+F10+F11	0.8157	**403.45**

- 平行句对抽取结果：

	Chinese-English
Candidate sentence pairs	2,101,770
Parallel sentence pairs	201,588

平行句对的抽取

- 好的例子

these archives were made up almost completely of the records of commercial transactions or inventories , with only a few documents touching theological matters , historical records or legends .
这些 档案 几乎 都 是 商业 事务 记录 或者 详细 目录 ， 仅 有 很少 的 记录 了 神的 事情 、 历史 记录 和 传说 。

their first design , using conventional layout and wooden furniture , proved to be too light .
他们 的 第一 个 设计 使用 传统 的 布局 以及 木材 ， 但是 被 证实 太 轻 了 。

平行句对的抽取

- 不好的例子

in addition to providing materials , libraries also provide the services of librarians who are experts at finding and organizing information and at interpreting information needs .
除了 提供 资源 ， 图书馆 还有 专家 和 图书 馆员 来 提供 服务 ， 他们 善于 寻找 和 组织 信息 ， 并 解释 信息 需求 。

a medical laboratory or clinical laboratory is a laboratory where tests are done on clinical specimens in order to get information about the health of a patient as pertaining to the diagnosis , treatment , and prevention of disease .
医学 实验室 ， 又 称为 临床 实验室 ， 是 指 在 其中 针对 临床 标本 进行 各种 试验 ， 或者 说 完成 形形色色 的 实验室 检验 项目 ， 以便 获得 病人 健康 状况 信息 的 一种 实验室 。

平行片段的抽取实验

- 例子

台中市 洲际 棒球场 . ||| taichung intercontinental baseball stadium .
物种 之间 基因 序列 ||| genes between species
寻找 和 组织 信息 ||| finding and organizing information
超越 了 建筑 的 围墙 ||| beyond the physical walls of a building
私人 档案 保存 在 ||| private archives were kept at
步兵 需要 装备 全自动 的 步枪 ||| infantry should be equipped with a fully-automatic rifle
选择 和 任何人 合作 ||| cooperate with whom they choose

在 1998 年 创建 ||| founded in 2005
自由 软体 并 不 使用 ||| free software came into use
的 特定 特征 。 ||| particular characteristics of an organism .
， 但 也 有 一些 ||| , although some

机器翻译实验

- 实验设置
  - Baseline训练语料
    - LDC，208.5万句对
    - Wikipedia标题，23.6万
  - 开发集和测试集：

A (Wikipedia)	DevA	390
	TestA	390
B (News)	NIST MT2003	919
	NIST MT2005	1082

  - 翻译系统：Moses
  - 语言模型：SRILM

机器翻译实验

	Test A	Test B
baseline	24.49	29.96
baseline +extracted sentence (201,588 sentence pairs)	**41.31 (+16↑)**	30.84 (+0.9↑)
baseline+ extracted fragment (7,708,424 fragments)	**45.20 (+20↑)**	30.21 (+0.25↑)
baseline + sentence + fragment	**50.52 (+26↑)**	30.23 (+0.27↑)

- 领域适应问题
- 加入语料的规模问题

机器翻译实验

	LLR positive lexicon size
LDC&WikiTitle corpus	325,001
LDC&WikiTitle corpus + extracted sentence	344,068 (+5.87)
LDC&WikiTitle corpus+ extracted fragment	**678,652 (+112.2)**
LDC&WikiTitle corpus + sentence + fragment	689,614

- 原始语料有可能被新加入的来自于不同领域的语料所淹没掉
- 进一步实验：
  - 研究如何使用挖掘到的资源

提纲

- 双语平行资源的获取
- 传统处理方法与不足
- 两级平行资源获取框架
  - 平行句对的抽取方法
  - 平行片段的抽取方法
- 实验结果与实例分析
- **总结与展望**

总结与展望

- 提出一个简单而有效的两级平行资源抽取框架
  - 句子级别的抽取
    - 分类器算法
    - 深入分析特征对分类效果的影响，提出一组有效的特征
  - 片段级别的抽取
    - 提出了一个两步词对齐算法
    - 基于该词对齐提出了有效的片段抽取方法
- 此框架能帮助抽取更多有助于机器翻译的资源

总结与展望

- 下一步工作：
  - 如何根据特定领域的机器翻译，使用我们抽取的资源？
  - 如何评价抽取的句对以及片段的可靠性？

THANKS
Q&A