

跨语言信息检索 ——CLIR

<http://hi.baidu.com/drkevinzhang/>

小组成员：李雪霞 赵晨晓
朱崇燕 李 雯
指导老师：赵燕平 张华平

跨语言检索简介

CLIR (Cross-Language Information Retrieval)

目的

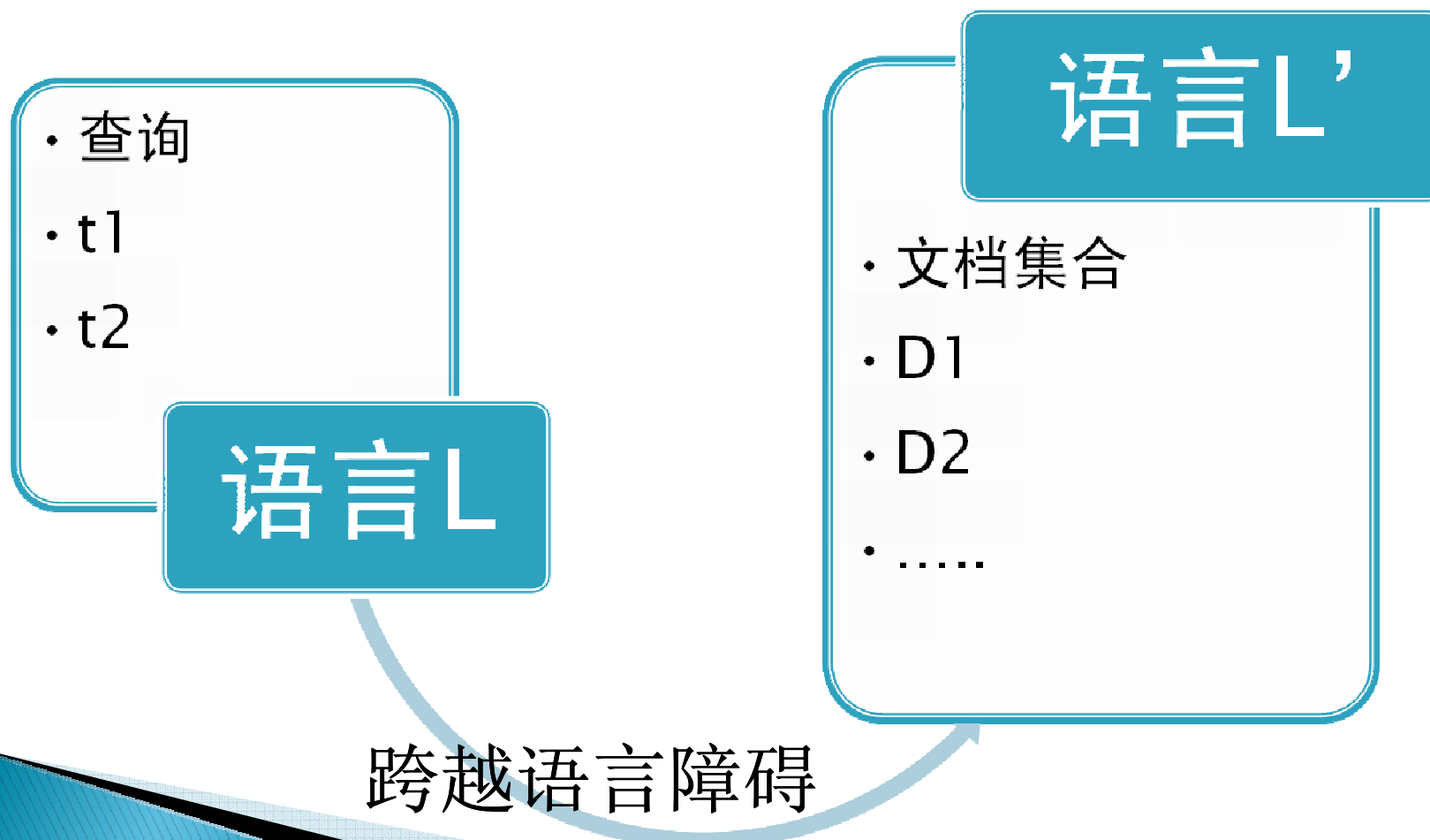
- 允许用户使用语言L表达出查询意图，但可以检索到采用另一个语言L'撰写出的文档。

思想

- 用户希望根据一种语言的查询来获得多种语言的文档集合。

CLIR与单语言搜索主要区别

- ▶ CLIR中的查询与文档集合不能够直接匹配



- ▶ 用户在该页面的搜索框中键入关键词后，相应关键词将被转换成用户希望搜索材料的目标语言（可通过该网页中提供的下拉菜单具体加以选择），谷歌搜索引擎将查找相应语种的网页材料，并把搜索结果再转换成用户本土语言返回给用户。

跨语言检索

Google 搜索

获得约 53,900,000 条结果（用时 0.55 秒）

[高级搜索](#)

翻译的外文网页



跨语言检索 的跨语言搜索结果 - 我的语言: [中文\(简体\)](#) ▼

语言

所查询字词的译文

英语 ☒

Cross-language retrieval - [修改](#)

[53,800,000 个结果](#)

中文(繁体) ☒

跨語言檢索 - [修改](#)

[76,300 个结果](#)

[添加语言](#) ▼

[跨语言信息检索- 维基百科，自由的...](#)

源语言: 英语

跨语言信息检索（**跨语言信息检索**）是一个子域的信息**检索**与撷取不同的**语言**信息处理的书面...

☒ [隐藏原始文本](#)

[Cross-language information retrieval - Wikipedia, the free ...](#)

Cross-language information **retrieval** (CLIR) is a subfield of information **retrieval** dealing with **retrieving** information written in a language different from ...

en.wikipedia.org/wiki/Cross-language_information_retrieval

4.1.1 跨语言搜索系统需要的资源

多语言

- 双语词项列表
- 词项字典
- 可比语料库
- 并行语料库



可比语料库

用语言L写的文档，以及主题相同但采用L'撰写的文档所组成的集合

主题相同
但不是相互译文

并行语料库

提供语言L写的文档，该文档是用语言L'写的文档的直接译本

用语言L写的每篇文档都是对应语言L'的直译

哈利·波特

《哈利·波特与死亡圣器》(《Harry Potter and the Deathly Hallows》), 是英国女作家J.K.罗琳创作的哈利·波特系列小说的第7部, 也是该系列的终结篇。根据小说内容, 在与伏地魔的大战中, 将有五十个以上的正反派角色死亡, 故事一开始, 黑魔王伏地魔踌躇满志, 召集爪牙食死徒在马尔福家中开会, 筹划如何彻底解决哈利·波特以绝后患。伏地魔当场处决被他俘虏的霍格沃茨学院麻瓜学教授, 喂食他的巨蛇纳吉尼, 为他大举入侵魔法部与霍格沃茨学院揭开序幕

In Harry Potter and the Deathly Hallows, the seventh and final book in the epic tale of Harry Potter, Harry and Lord Voldemort each prepare for their ultimate encounter. Voldemort takes control of the Ministry of Magic, installs Severus Snape as headmaster at Hogwarts, and sends his Death Eaters across the country to wreak havoc and find Harry. Meanwhile, Harry, Ron, and Hermione embark on a desperate quest the length and breadth of Britain, trying to locate and destroy Voldemort's four remaining Horcruxes, the magical objects in which he has hidden parts of his broken soul. They visit the Burrow, Grimmauld Place, the Ministry, Godric's Hollow, Malfoy Manor, Diagon Alley...

美国宪法

THE CONSTITUTION OF THE
UNITED STATES OF AMERICA

March 4, 1789

Preamble

We the people of the
United States, in order to
form a more perfect union,
establish justice, insure
domestic tranquility, provide
for the common defense,
promote the general welfare,
and secure the blessings of
liberty to ourselves and our
posterity, do ordain and
establish this Constitution for
the United States of America.

... ..

美国宪法（中文）

序言

我们美利坚合众国
的人民，为了组织一个
更完善的联邦，树立正
义，保障国内的安宁，
建立共同的国防，增进
全民福利和确保我们自
己及我们後代能安享自
由带来的幸福，乃为美
利坚合众国制定和确立
这一部宪法。

4.1.2 CLIR系统测评

测评标准

计算跨语言查询
的平均准确度

计算跨语言检索性能与单
语言检索性能的百分比

4.2 跨越语言障碍

跨越语言障碍的核心问题

应该翻译什么？



应该翻译那种级别的符号？



我们如何利用翻译？



我们如何除去不当翻译？

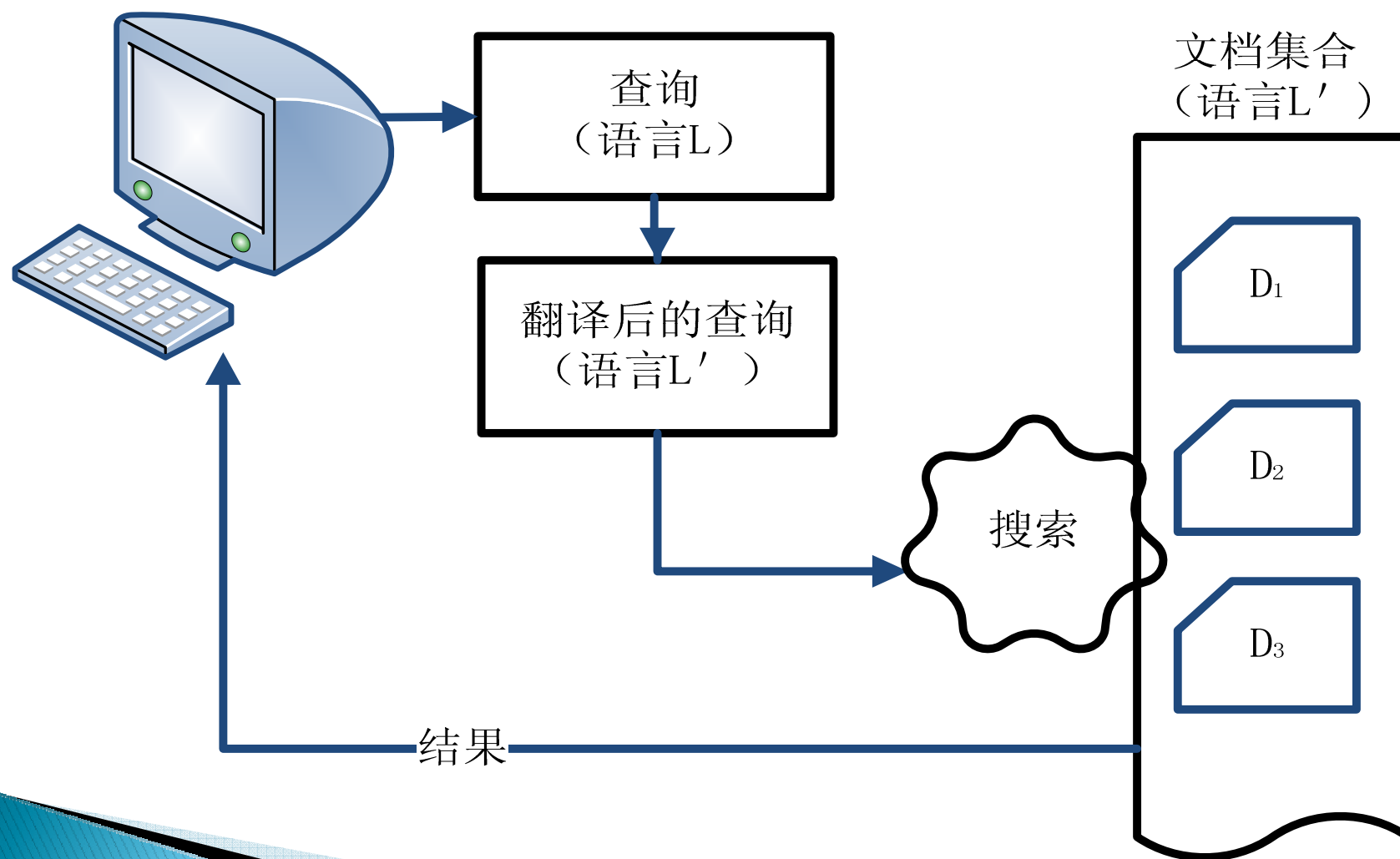


4.2 跨越语言障碍

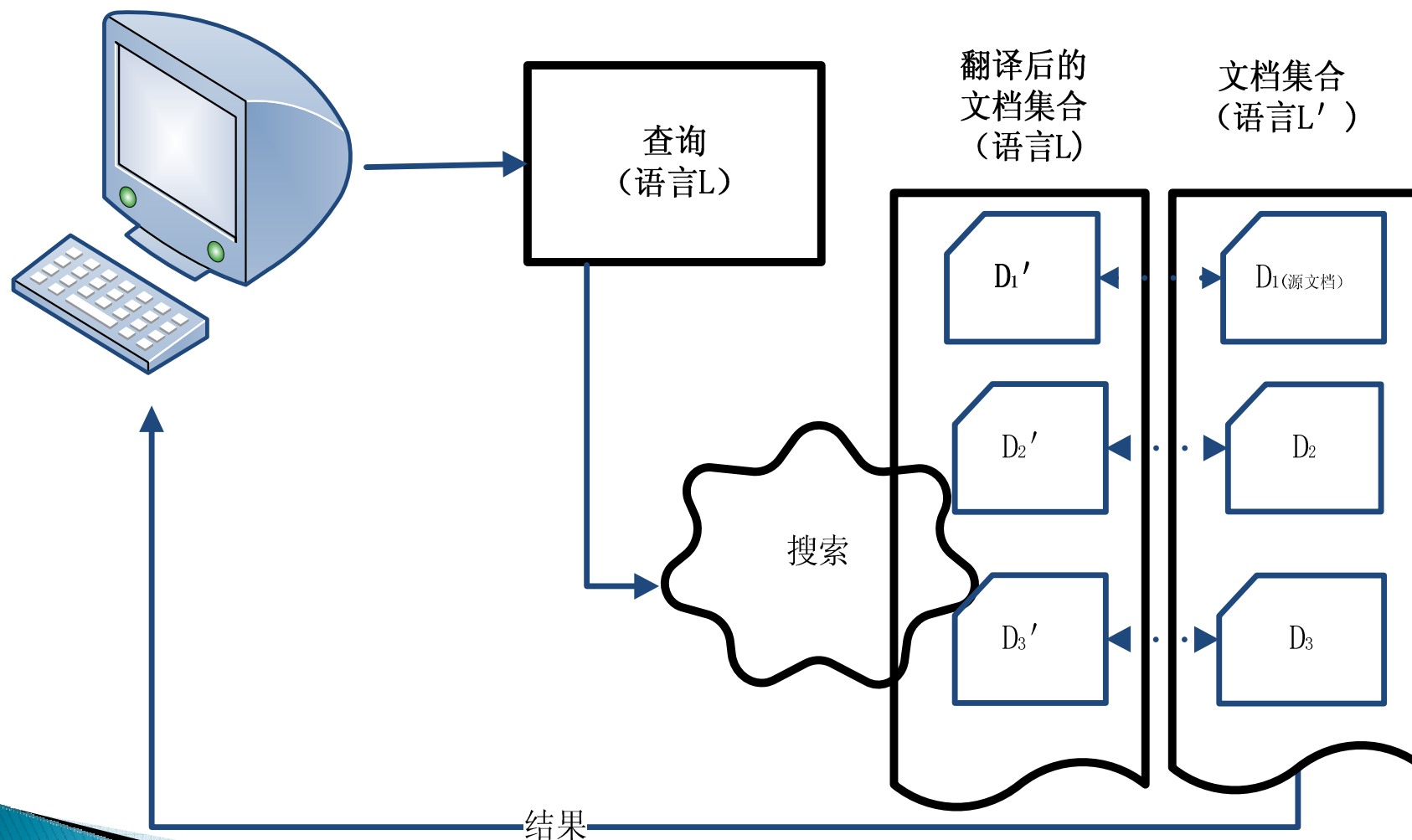
应该翻译什么

- ▶ 查询
- ▶ 文档
- ▶ 翻译查询与文档翻译为某种内部的中间表示形式

4.2.1 查询翻译

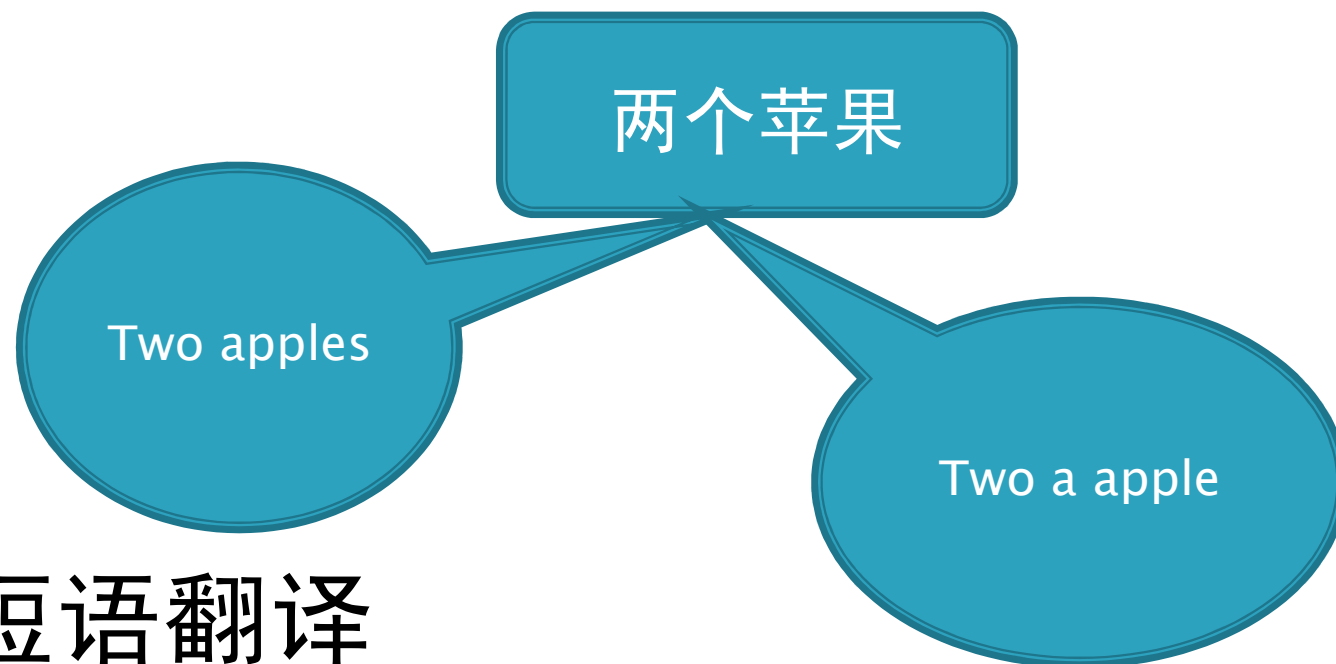


4.2.2 文档翻译



应该翻译那种级别的符号？

词干
词语
短语

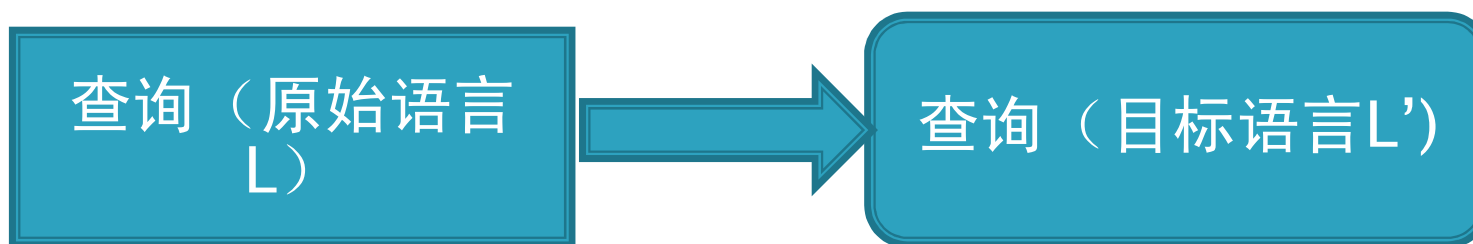


4.2.3 短语翻译

- ▶ 与简单实用词翻译，基于短语的翻译方法可以显著提高性能

我们该如何利用翻译？

- ▶ 如何让选择译文以及如何将译文组合为采用目标语言表达的新查询



- ▶ 语言L中的一个词可能对应语言L'中的多个词
 - 一个词，部分词还是全部词
 - 信任一个翻译结果而不是另外一个

4.2.4 译文的选择

▶ 双语词表的质量

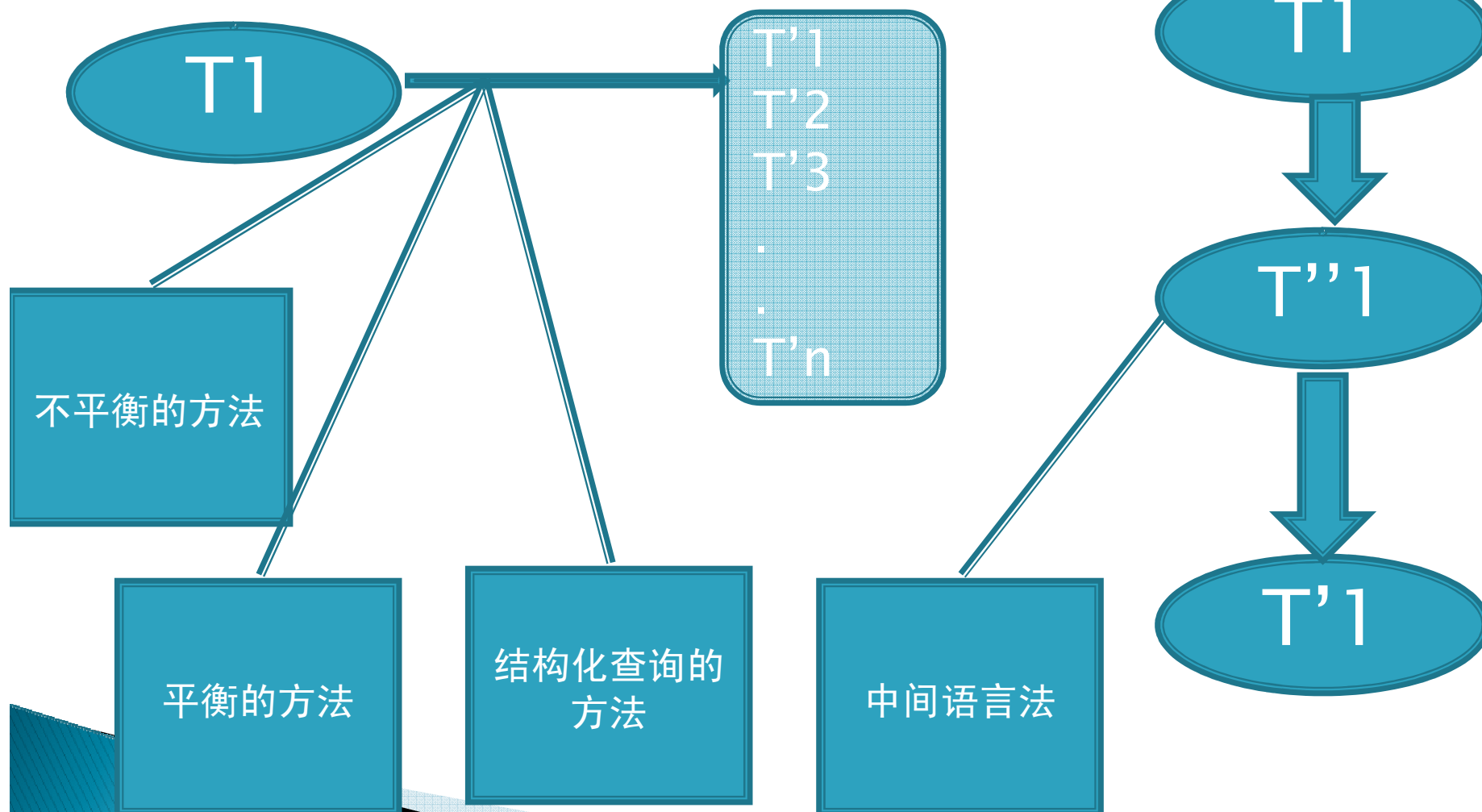
跨语言搜索系统都会使用多种技术提高精度

- 这些系统的核心： 双语词典
- 双语词典： 将语言L中的一个词映射到L'中的多个词
- 使用词典的主要问题： 歧义问题、词语覆盖不全、机器可读资源匮乏

查询（原始语言
L）



新查询（目标语言L'）

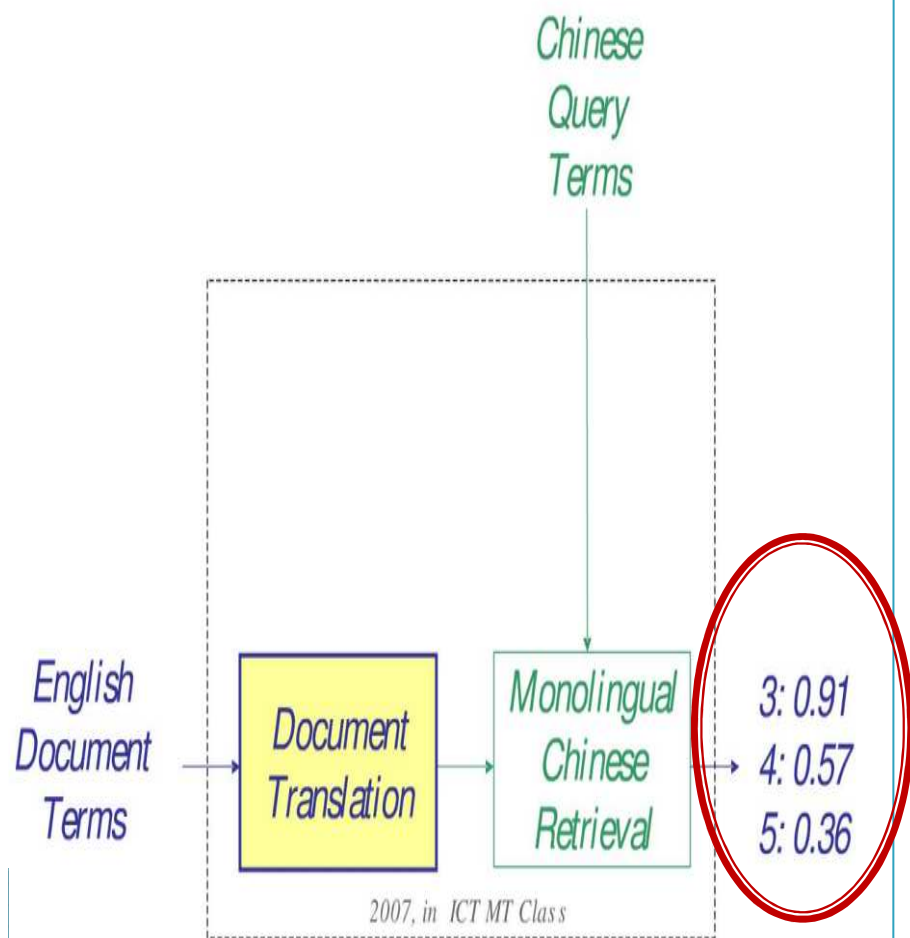


4.2.5 翻译删减技术

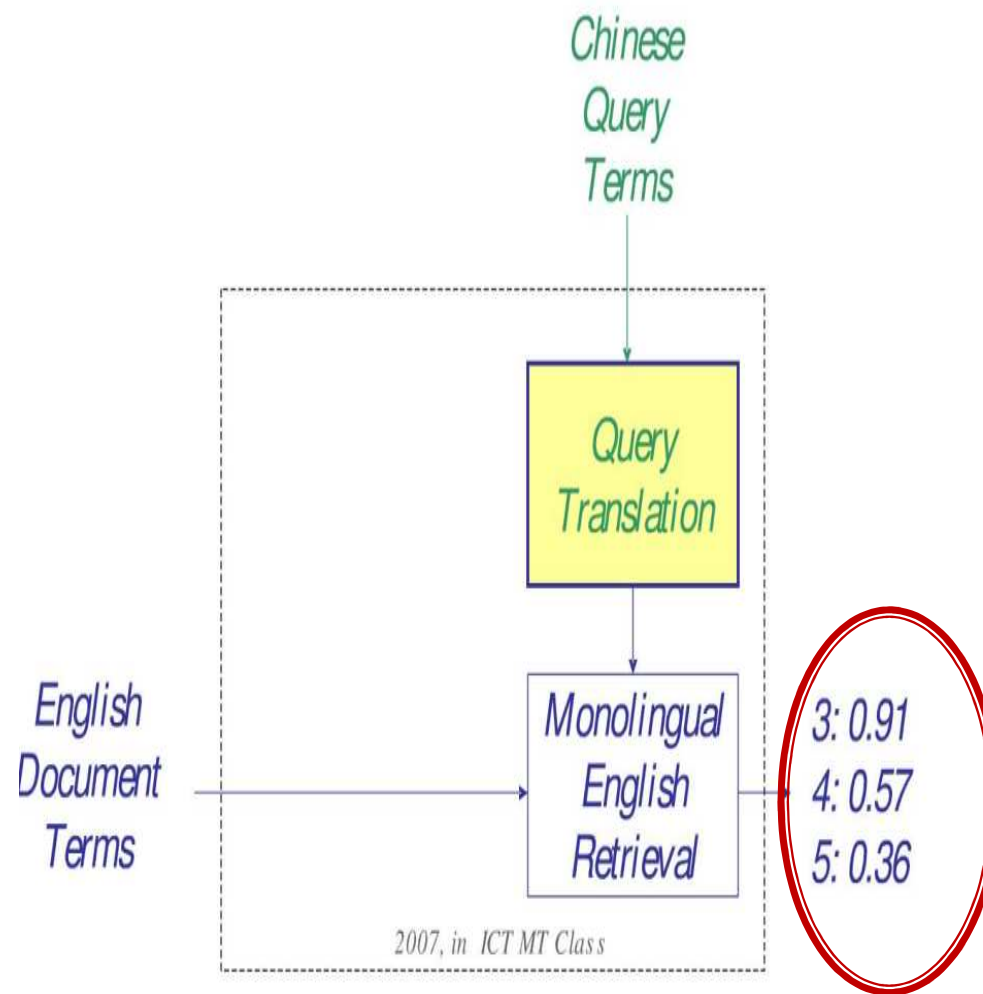
- ▶ 确定了译文，需要针对给定短语的译文列表进行删减处理
- ▶ 可以使用HMM发来计算最可能的查询翻译
 - HMM：计算最有可能的查询翻译，并且使用传统的语言模型对检测结果进行合并
 - 我 是 一个 学生
 - I am a student
 - ~~Me~~ ~~are/is~~ ~~an/one~~ pupil

CLIR中的语言模型及方法

文档翻译



查询翻译



语言模型

Query in English: <input type="text" value="Swiss bank"/>		<input type="button" value="Search"/>
<i>English</i>	<i>German</i>	
	<div>(Swiss) (Bankgebäude, bankverbindung, bank)</div>	
1 (0.72) <i>Swiss Bankers Criticized</i> AP / June 14, 1997	1 (0.91) <i>U.S. Senator Warpathing</i> NZZ / June 14, 1997	
2 (0.48) <i>Bank Director Resigns</i> AP / July 24, 1997	2 (0.57) <i>[Bankensecret] Law Change</i> SDA / August 22, 1997	
	3 (0.36) <i>Banks Pressure Existent</i> NZZ / May 3, 1997	

CLIR语言模型

$$P(q_L | D_{L'i}) = \prod_{j=1}^Q \alpha P(t_{Lj} | C_L) + (1 - \alpha) P(t_{Lj} | D_{L'i})$$

$$\prod_{k=1}^{|D_{L'}|} P(t_{L'k} | D_{L'i}) P(t_{Lj} | t_{L'k})$$

翻译后的查询词词频占检索文档大小的概率

翻译概率：对于一个用语言L'表示的词语，对应语言L的词语的概率

Gold silver truck

De1: "Shipment of gold samaged in a fire"

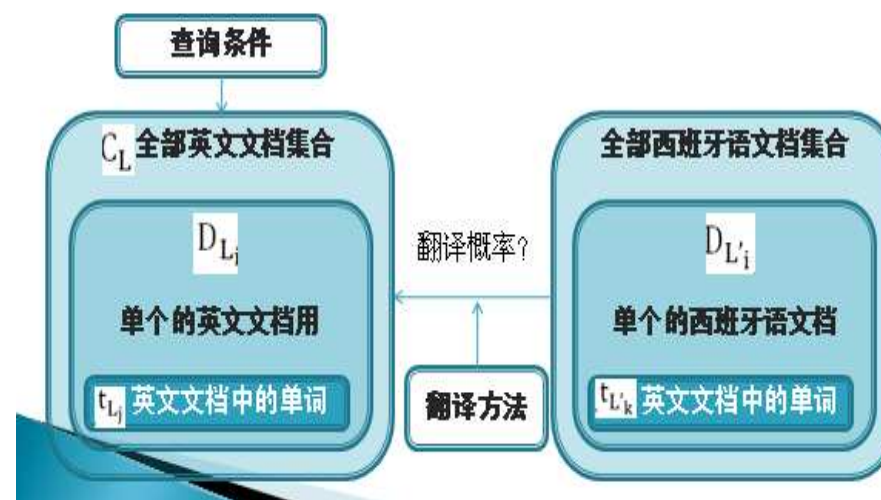
De2: "Delivery of silver arribed in a silver truck"

De3: "Shipment of gold arrived in a truck"

Ds1: El envoi del oro dano en un fuego

Ds2: La entrega de la plata lleo en un camion de palta

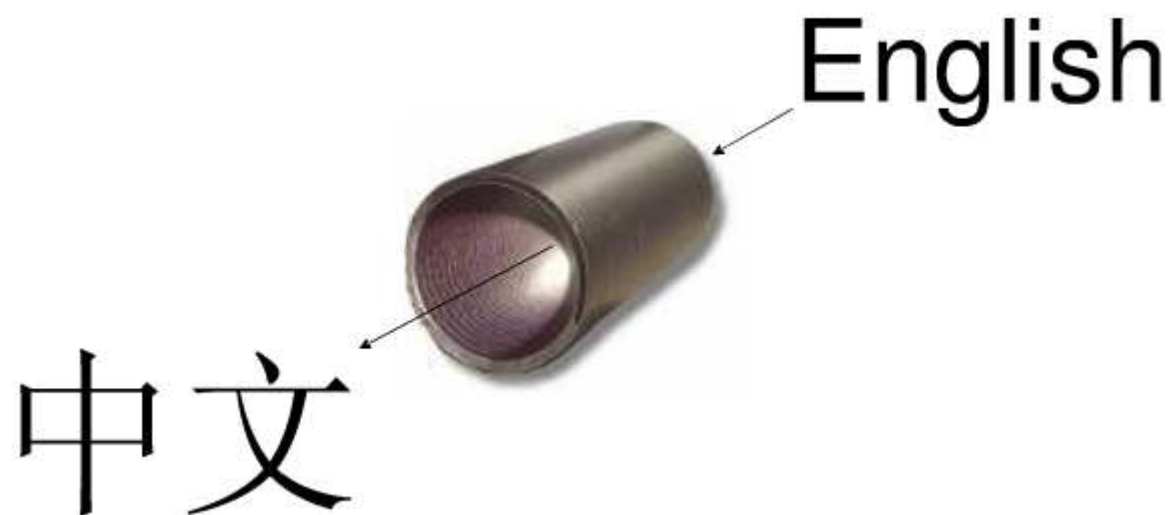
Ds3: El envoi del oro lleo en un camion



- ▶ 翻译方法
- ▶ 双语词典
- ▶ 双语语料库
- ▶ 机器翻译系统

机器翻译

- 机器翻译(machine translation)是使用电子计算机把一种自然语言(源语言,source language)翻译成另外一种自然语言(目标语言,target language)的一门学科
- 这门新学科同时也是一种新技术.它涉及到语言学、计算机科学、数学等许多部门,是非常典型的多边缘的交叉学科



机器翻译实例

- ▶ 如美国大型机器翻译系统SYSTRAN， 可进行俄英、 英俄、 德英、 汉法、 汉英机器翻译， 每小时可译30万~35万个词
- ▶ 开源统计机器翻译系统“摩西”
 - 目前最有影响的开源统计机器翻译系统
 - <http://www.statmt.org.moses>

基于词典的翻译

基于词典的查询条件翻译

Original query: El Nino and infectious diseases

Term selection: “El Nino” infectious diseases

Term translation: (传 传染病 易传染的) (变质 弊病 病)

(Dictionary coverage: “El Nino” is not found)

Translation selection: (传染病 易传染的) 病

Query formulation:

Structure:

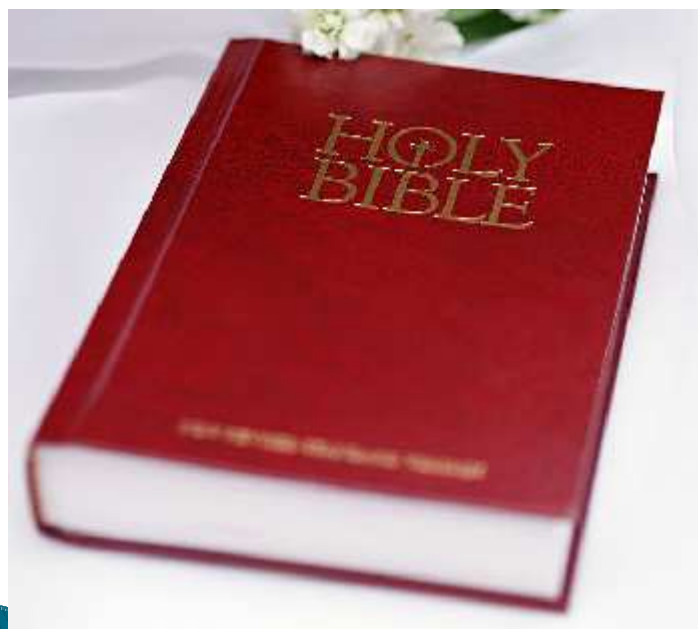
OP1 (OP2 (传染病 易传染的) 病))

双语语料库

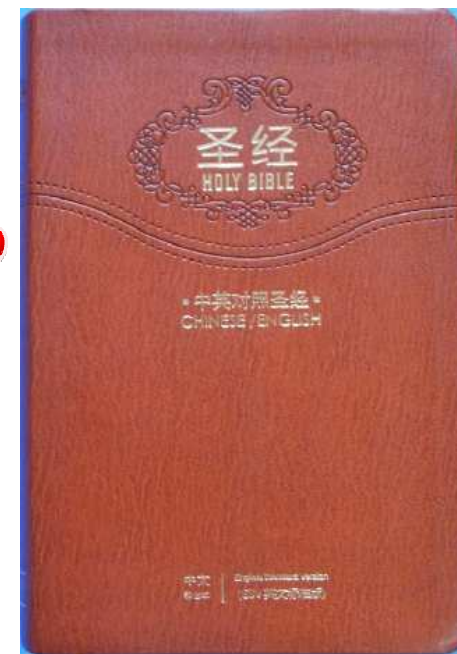
- ▶ 大型的应用系统手工构建词典或者复杂的多语种叙词表是不切实际的
- ▶ 基于语料库的方法利用现有的大型文献集合，对其进行自动分析，进而收取构建自动翻译技术所需的信息。

并行语料库

- ▶ 一本同时翻译为语言L和语言L'的书（平行翻译）



如何匹配？



对齐

- ▶ 双语语料库的对齐方式
- ▶ 段落对齐
- ▶ 句子对齐
- ▶ 短语对齐
- ▶ 单词对齐

汉语	英语
汉语	英语
<p>FDDI 最强处之一是其容错性。FDDI 不是单一环，而是两个反向旋转的环，每台机器能够向左右邻发送和接收数据。这个聪明的方案让网络即使在光缆出现故障时仍能继续工作。当一根光缆被切断时(FDDI 1.25 英里的传输距离能力很可能出现这种切断)，环两端的设备就起到了回送的作用，把反向旋转的双环改变成单环，每台设备通过两次。</p>	<p>One of FDDI's greatest strengths is its fault tolerance. Rather than being a single ring, FDDI is two counter-rotating rings. Each machine can both send to and receive from each of its two neighbors. This clever scheme allows the network to continue operating even in the event of a cable failure. When a cable is cut (and FDDI's 1.25-mile distance capability makes that more likely), the devices at both ends of the ring act as loopbacks, changing the cabling configuration from dual and counter-rotating rings to a single ring that passes through each device twice.</p>

IBM翻译模型

$$\text{Max } P(F|E) = \prod (g(t(f|e)))$$

约束条件: $\sum t(f|e)=1$

- 也就是说，给定训练语料库E和F，我们要求解一个概率分布 $t(f|e)$ ，使得翻译概率 $P_r(F|E)$ 最大。

□ 英语文本

Most of us, however, take life for granted. We know that one day we must die, but usually we picture that day as far in the future. When we are in buoyant health, death is all but unimaginable. We seldom think of it. The days stretch out in an endless vista. So we go about our petty tasks, hardly aware of our listless attitude toward life.



□ 汉语文本:

但是，我们大多数人把生活认为是理所当然的。我们知道，某一天我们一定会死，但通常我们把那天想象在遥远的将来。当我们心宽体健时，死亡几乎是不可想象的，我们很少想到它。时日在无穷的展望中延展着，于是我们干着琐碎的事情，几乎意识不到我们对生活的倦怠态度。

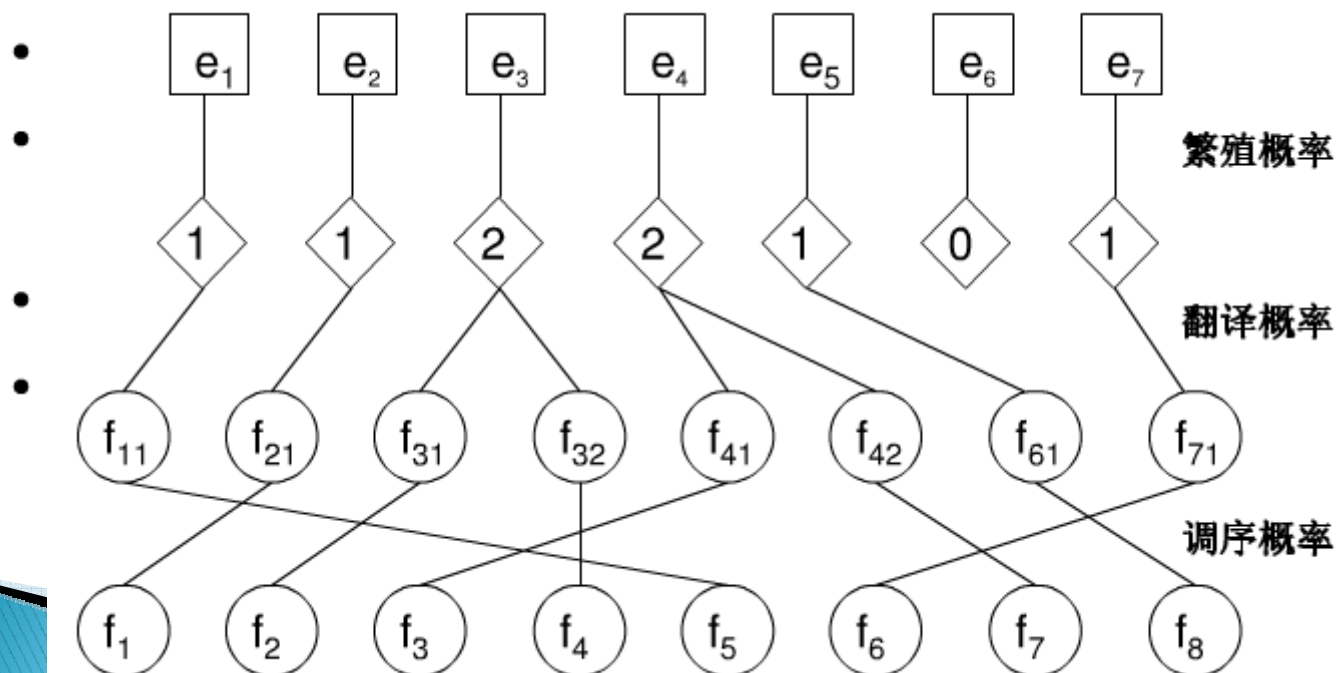
IBM模型1

源语言句子E:

I_1 am_2 a_3 $student_4$

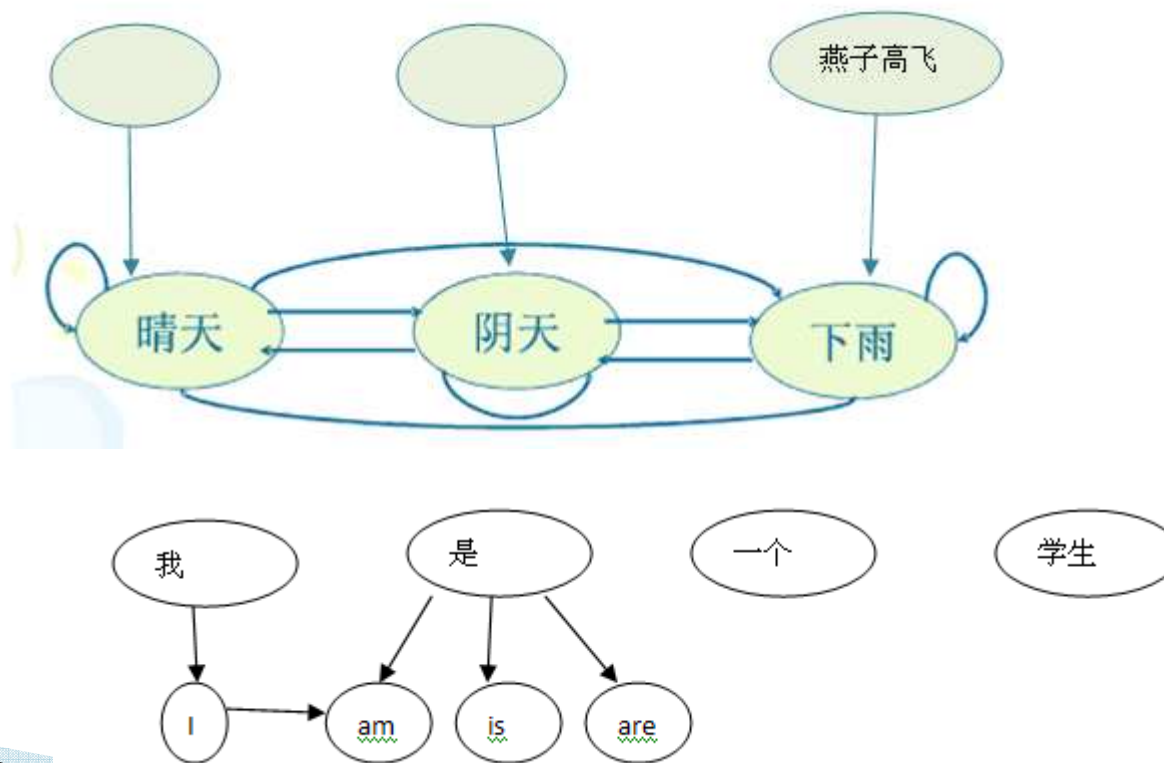
目标语言句子F:

(我) (是) (一) (个) (学生)



HMM

- ▶ 隐马尔可夫模型是一个双重随机过程-----具有一定状态数的隐马尔可夫链和显示随机函数集



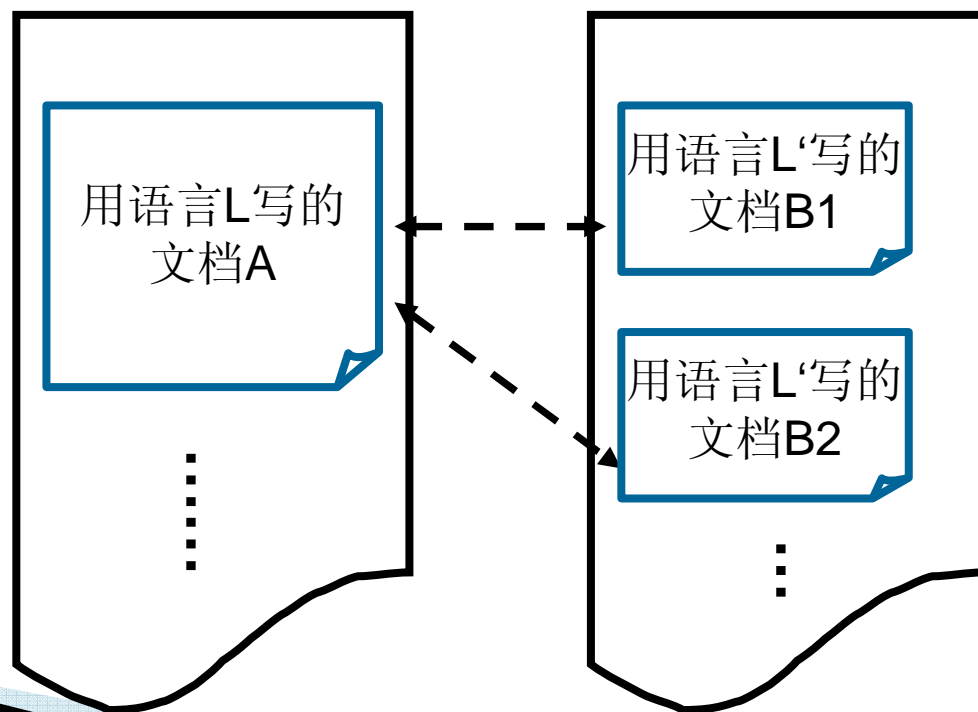
可比语料库方法

可比语料库方法的提出

- ▶ 并不是所有文章都有相对应的翻译
- ▶ 并行语料库获取并不容易，所适用的范围小
- ▶ 获取两种语言的可比语料库可能性比较大

可比语料库

- ▶ 每种语言写的文档都是关于同一个主题的，但相互之间不是直接翻译得来的



抽取双语词典

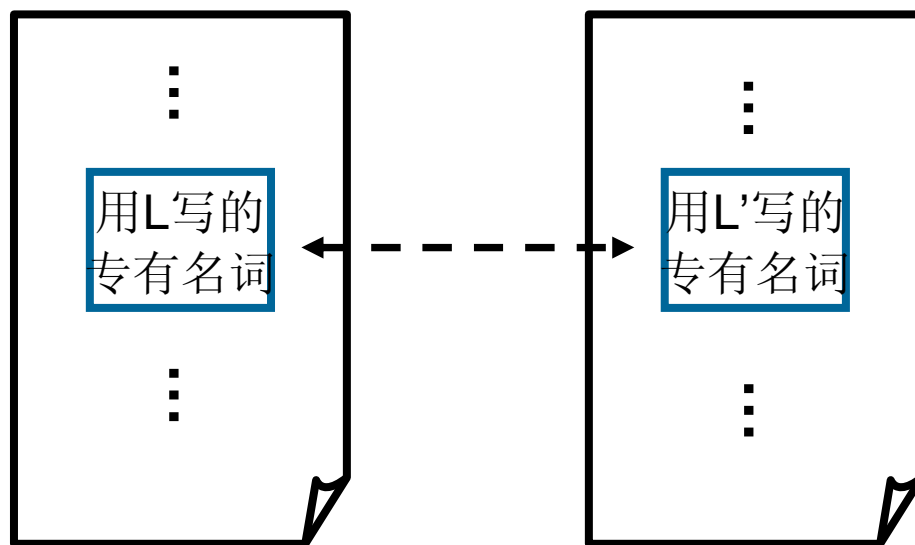
- ▶ 用对齐技术通过可比语料库中的文档建立双语词典
- ▶ 并且需要定义词语-词语的相似度函数
- ▶ 该词典是查询翻译的基础

对齐技术

- ▶ 按照时间以及独立于语言的描述符对齐
 - 如“.mil”表示该文档是关于军事的
- ▶ 用文档中的专有名词来帮助对齐

用语言L写的文档A

用语言L'写的文档B

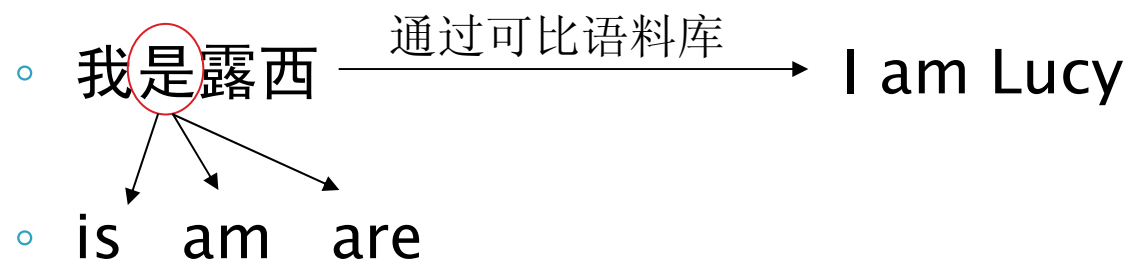


词语-词语之间的相似度函数

- ▶ 词共现度
- ▶ 词上下文

词共现度原理

- ▶ 好的翻译结果之间应该要在目标语言文档中共同出现，而错误的译文则不应该共现
- ▶ 采用按比例适应反复迭代过程，逐步计算给定词语是另外一个词语译文的概率



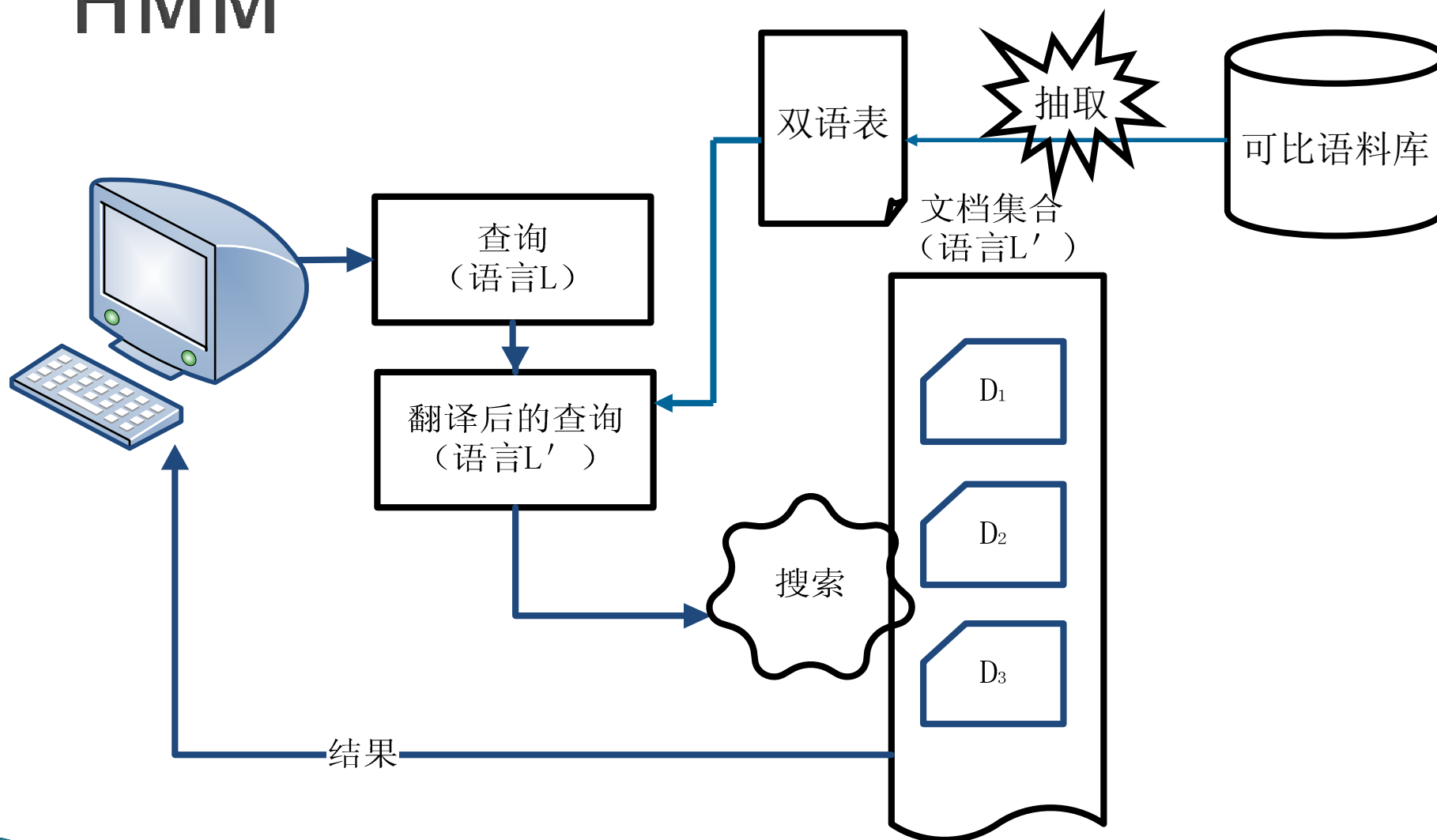
词上下文原理

- ▶ 根据周围词语得出目标词语的上下文向量
- ▶ 根据周围词语计算源语言的上下文向量
- ▶ 通过两个向量之间的余弦相似度最终来确定一个或多个译文

基于可比语料库的CLIR算法

- ▶ 统计语言模型
 - HMM（隐马尔科夫模型）
- ▶ 概率模型
 - 跨语言相关性模型
- ▶ 向量空间模型
 - GVSM（广义向量空间模型）
 - LSI（潜语义索引模型）

HMM



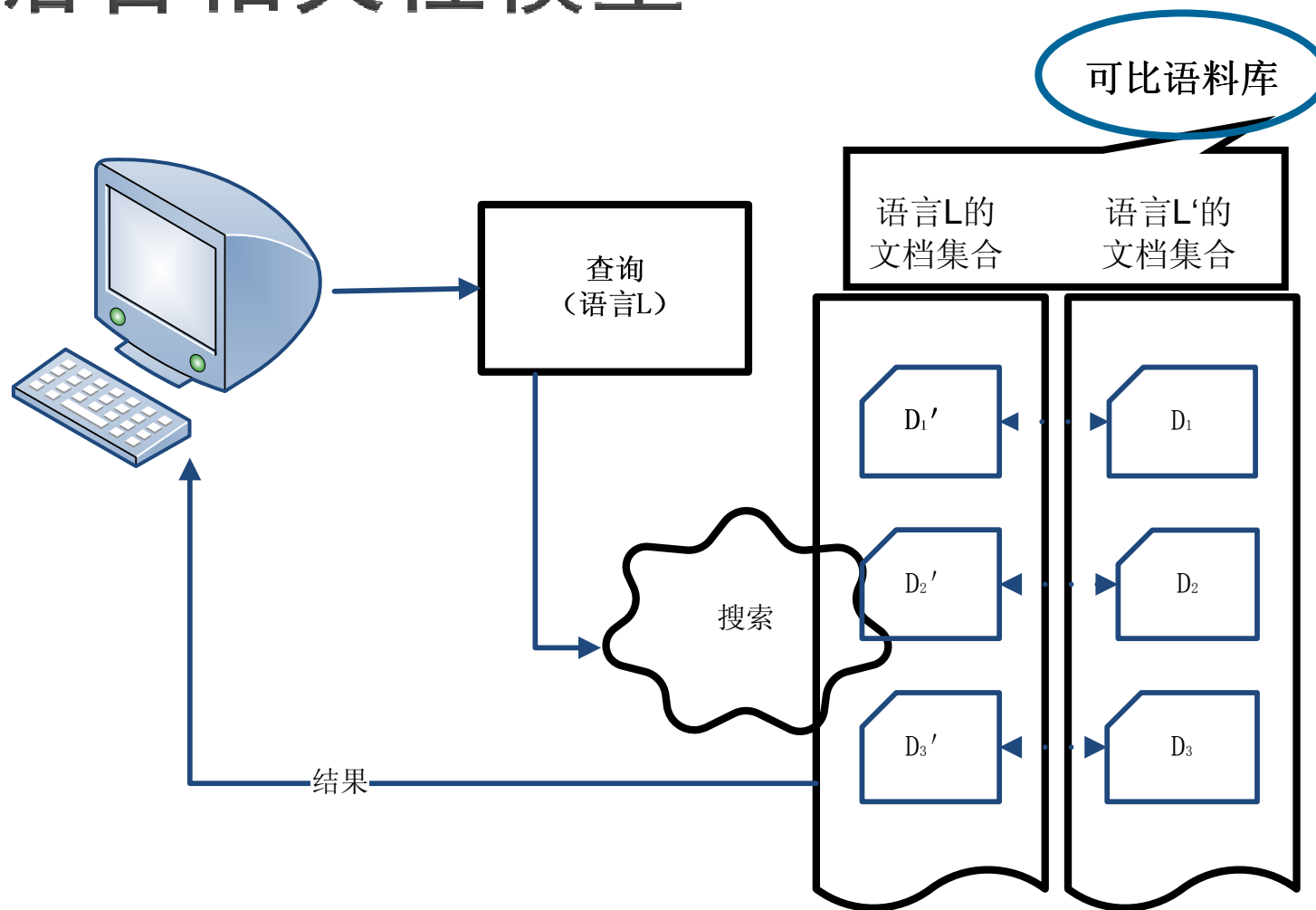
基于可比语料库的HMM原理

- ▶ 不使用统一的翻译概率，而使用语料库中事件发生的频率

- ▶ 如

▶	bark	树皮	0.5	→	0.3
▶		狗吠	0.5	→	0.7

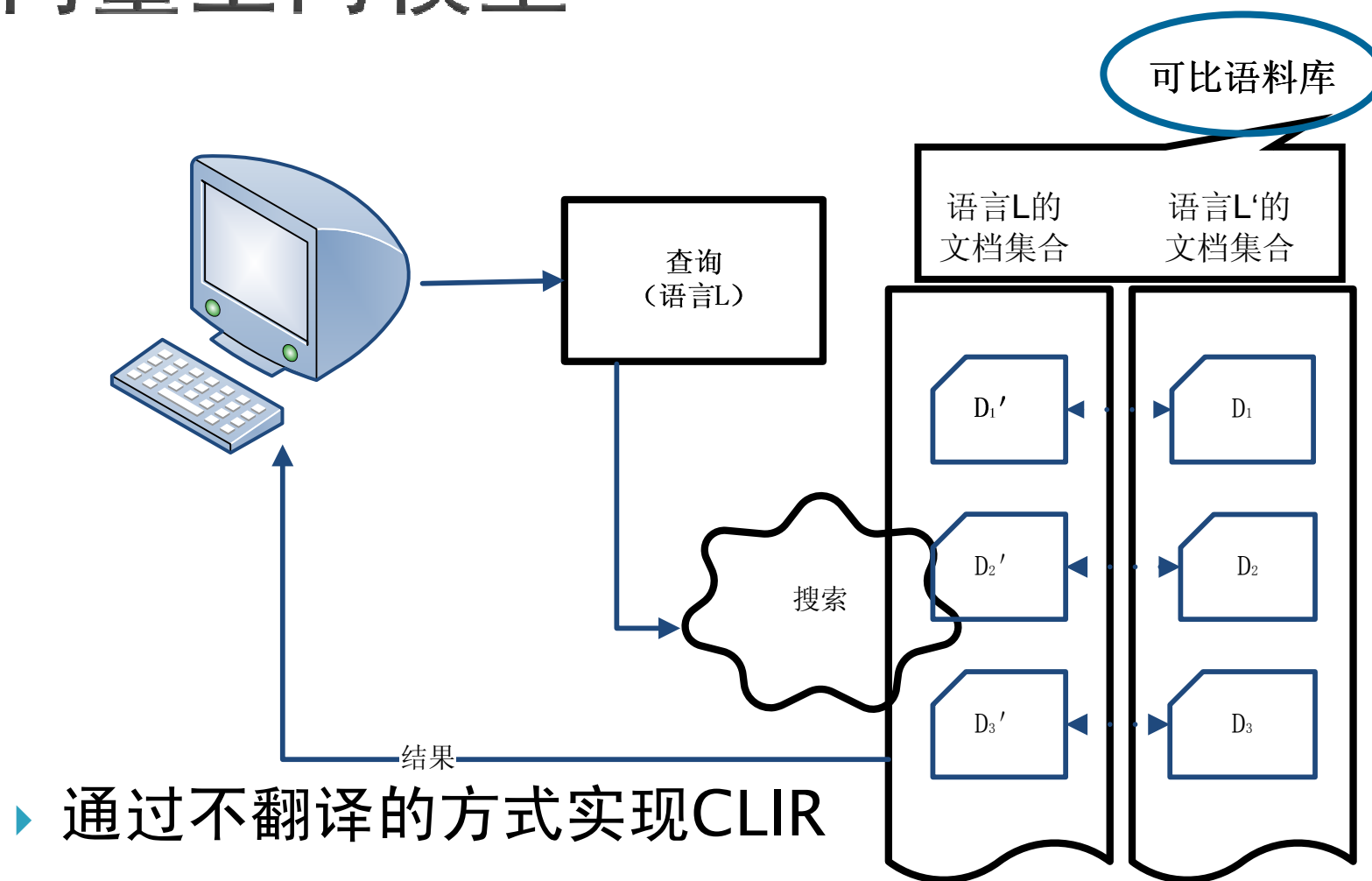
跨语言相关性模型



跨语言相关性模型原理

- ▶ 计算每一个词在一种语言中的概率
 - 计算语言L'中的词共现概率，要求该词必须与语言L所有查询词共现在一篇可比文档中
 - 可比语料库用来将用L语言写的共现文档映射到用L'语言写的文档
- ▶ 采用KL距离相对熵矩阵计算文档相关度

向量空间模型



- ▶ 通过不翻译的方式实现CLIR

GVSM原理

- ▶ 使用可比语料库为语言L建立词-文档矩阵 A^T 矩阵，为语言L'建立矩阵 B^T
- ▶ 得到两个向量：一个向量是其每个元素代表用语言L写的一篇文档，另一个向量是其每个元素代表用语言L'写的一篇文档
- ▶ 通过计算两个向量之间夹角的余弦计算跨语言的相似度

$$SC(Q, D_i) = \cos(A^T Q, B^T D_i)$$

GVSM 例子

- ▶ 用英语查询
- ▶ Gold Silver Truck

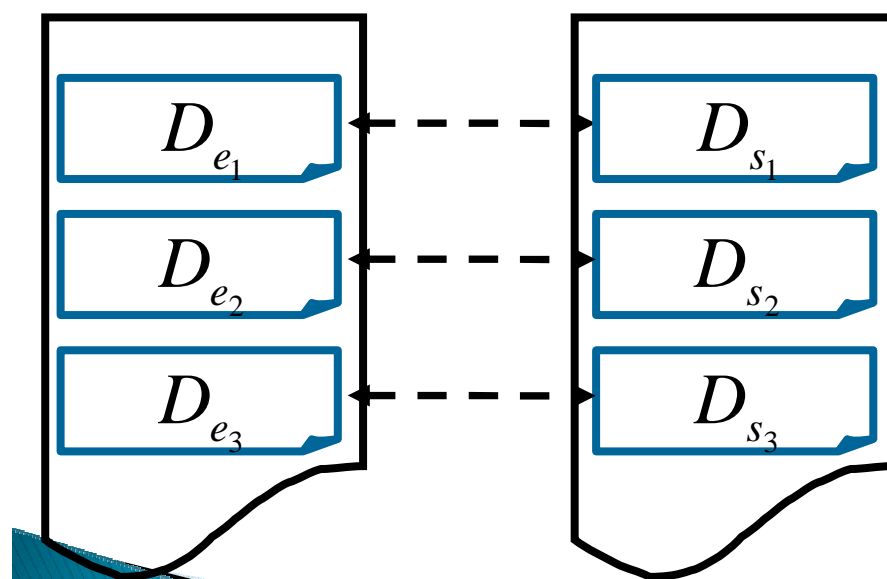
D_{e_1} : "Shipment of gold damaged in a fire. "

D_{e_2} : "Delivery of silver arrived in a silver truck. "

D_{e_3} : "Shipment of gold arrived in a truck. "

用英语写的文档集

用西班牙语写的文档集



D_{e_1} : El envío del oro dañó en un fuego

D_{e_2} : La entrega de la plata llegó en un camión de plata

D_{e_3} : El envío del oro llegó en un camión

GVSM 例子

表 4-1 英语集合中所有词的原始词频

$tf_{n,d}$	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
D_{c_1}	1	0	1	0	1	1	1	1	1	0	0
D_{c_2}	1	1	0	1	0	0	1	1	0	2	1
D_{c_3}	1	1	0	0	0	1	1	1	1	0	1

表 4-2 西班牙语集合中所有词的原始词频

$tf_{n,d}$	camión	daño	de	del	el	en	entrega	envío	fuego	la	llegó	oro	plata	un
D_{c_1}	0	1	0	1	1	1	0	1	1	0	0	1	0	1
D_{c_2}	1	0	2	0	0	1	1	0	0	2	1	0	2	1
D_{c_3}	1	0	0	1	1	1	0	1	0	0	1	1	0	1

□

GVSM 例子

$$Q' = A^T Q = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}^T \times \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = [1 \quad 3 \quad 2]$$

$$Q' = B^T D_{s_1} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix}^T \times \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = [8 \quad 2 \quad 6]$$

- 英语查询向量与第一个西班牙语文档向量的相似度:
- $SC(\text{Gold, Silver, Truck}, D_{s1}) = \cos (<1,3,2>, <8,2,6>) = 0.6814$

LSI原理

- 对融合了语言L中的词-文档矩阵A以及语言L'中的词-文档矩阵B的新的词-文档矩阵 $\begin{bmatrix} A \\ B \end{bmatrix}$ 应用奇异值

分解进行计算 $\begin{bmatrix} A \\ B \end{bmatrix} = U \sigma V^T$ 构造出不同语种的潜在语义空间

- 通过以下公式进行相似度计算

$$SC(Q, D_i) = \cos(U^T Q, U^T D_i)$$

跨语言检索实用策略

1

相关反馈

2

词干还原

3

N元语法模型

4

音译名

5

命名实体识别

6

检索融合

1、跨语言的相关反馈

- ▶ 翻译前

先检索，取检索结果集中排名靠前若干文档，在查询中加入的词排名靠前的若干词，然后翻译这些词语进行检索

例：英语-阿拉伯语

- ▶ 翻译后

- ▶ 将翻译后的查询词进行查询，取查询结果集中前20篇文档构建词项-词项贡献矩阵，对这些词的相似度求和，取前10个词项加入查询

1、跨语言的相关反馈

- ▶ 效果:
- ▶ 综合使用查询前和查询后查询扩展，效果提高
- ▶ 翻译资源词汇的覆盖率与查询扩展的效果成反比
- ▶ 对于亚洲语言，与逻辑回归合用，效果明显
- ▶ 效果与查询语言相关

2、词干还原

▶ 回退翻译

逐渐放宽约束进行匹配

例: jumping	-->	jumping
stem(jumping)	-->	jump
jump	-->	stem(jumping)
stem(jumping)	-->	stem(jumping)

研究问题：词干自动还原

2、词干还原

▶ 自动构建词干还原器

使用带权重的互信息找到第一批候选的基本后缀

例：“-ing”多次出现在词尾，因此可以认为它是后缀

使用MDL(Minimum Description Length)纠正错误

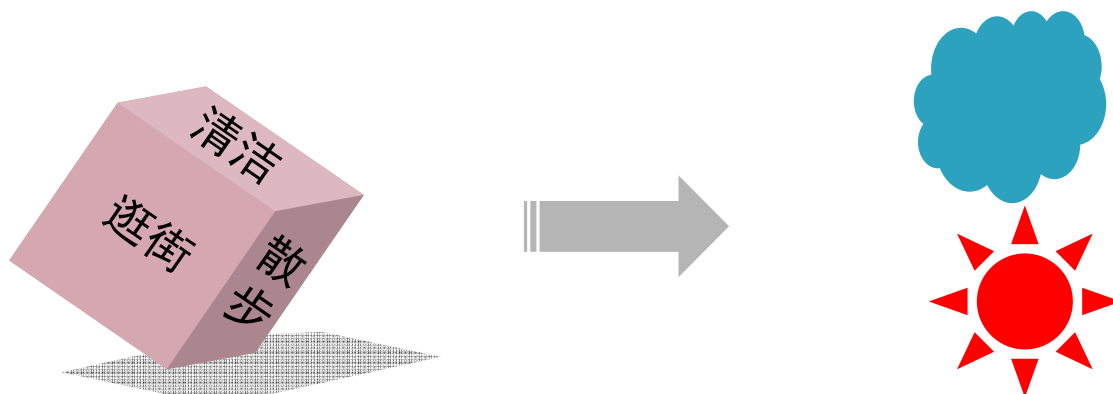
例：-news → new ❌ sing → s ❌

使用规则归纳法发现不常用的词

例：employ~~ee~~ → employ

2、词干还原

- ▶ 自动构建词干还原器
HMM（隐马尔科夫模型）



$\{S, O, P, A, B\}$, S 指状态的集合, O 指观察值的集合, P 指状态的初始概率, A 指状态间的转移概率, B 指状态与观察值之间的发射概率

3、n元语法模型

- N元语法模型 (N-gram Model)

$$\begin{aligned} P(\mathbf{w}) &= \prod_{i=1}^n P(\mathbf{w}_i | \mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_{i-1}) \\ &\approx \prod_{i=1}^n P(\mathbf{w}_i | \mathbf{w}_{i-N+1} \mathbf{w}_{i-N+2} \dots \mathbf{w}_{i-1}) \end{aligned}$$

- 假设：单词 w_i 出现的概率只与其前面的N-1个单词有关
 - ▶ 英语
n=4或5效果较好
 - ▶ 中文-日语
避免语言切分

4、音译名

- ▶ 英语-日语

语音方法

例: tomato

|tə'ma:təu|

to ma to

トマト

- ▶ 日语-英语

词切分算法

5、命名实体识别

- ▶ 笑话:
 - ▶ Are you Serious?
 - ▶ No,I'm joking.
-
- ▶ 翻译:
 - ▶ 你是塞雷拉斯吗?
 - ▶ 不,我是杰克。

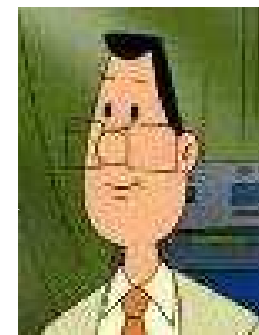
5、命名实体识别

- ▶ 任务：识别出待处理文本中三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体
- ▶ 过程通常包括两部分：（1）实体边界识别；（2）确定实体类别（人名、地名、机构名或其他）。
- ▶ 英语：大写字母、空格、前后缀等



爸爸，为什么俄国那么多人叫“斯基”呢？

因为俄国有很多车，所以有很多司机。



6、检索融合

- ▶ 多语言信息检索(MultiLingual Information Retrieval)
- ▶ 分布式
 - 将查询翻译为每种语言，并且在每种语言上都创建一个单独的索引。将**所有的相关性得分进行归一化**，并且给出单一的文档排序列表。
 - 将查询翻译为每种语言，并且在每种语言上都创建一个单独的索引。但是不使用得分来进行合并，而是**使用结果列表中的相对位置来合并**。
- ▶ 集中式
 - 将查询翻译为每种语言，并且将所有的查询结合在一起，转换为一个长查询。为每一种语言的所有文档创建一个大索引。在这个大索引上运行这个长查询

性能优于分布式

跨语言检索总结

跨语言检索的发展

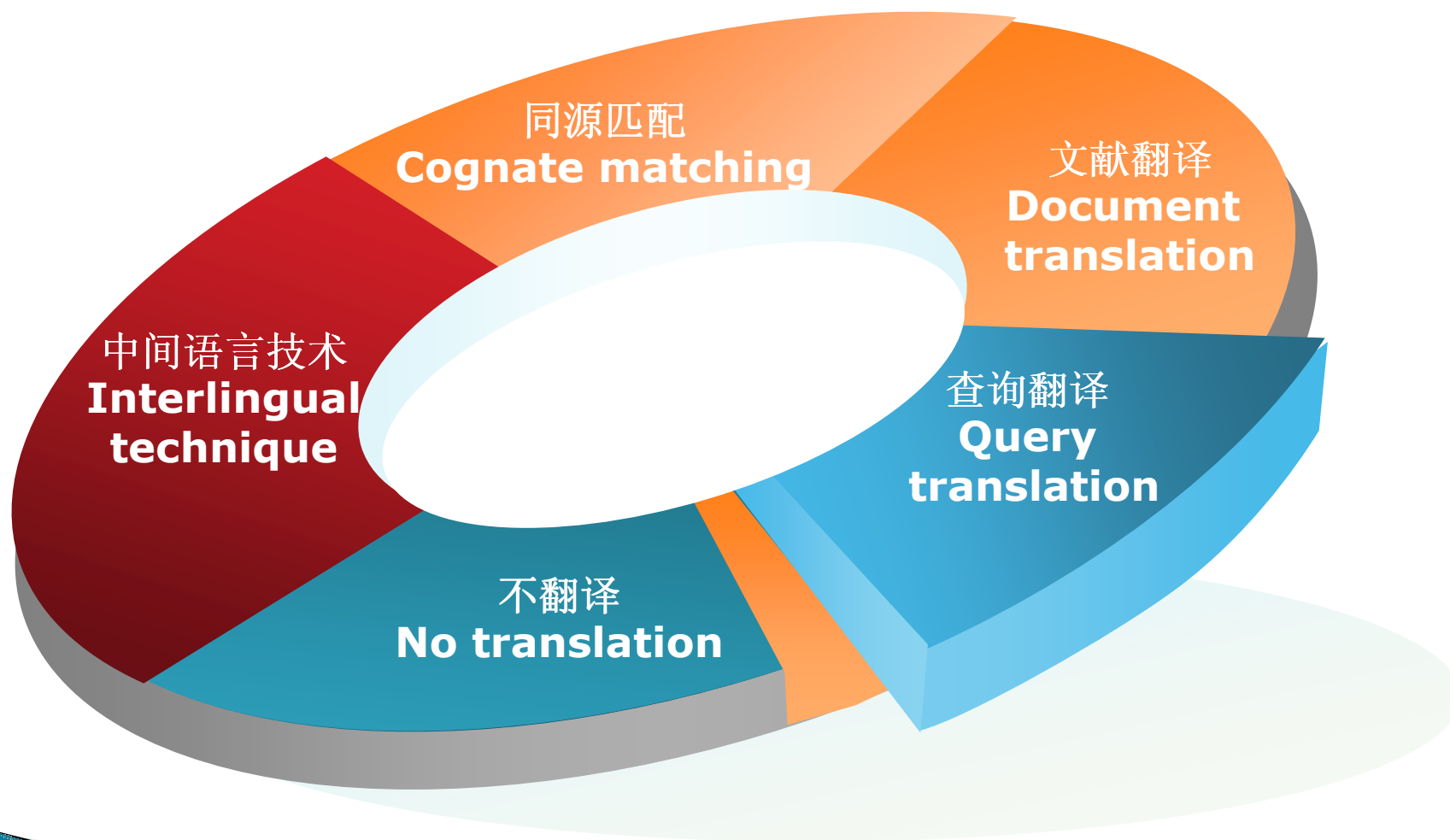


搜索引擎	跨语言检索	网页翻译	线上词典
Google(www.google.com)	有	有	无
Yahoo! (www.yahoo.com)	有	无	无
微软(search.msn.com)	有	有	无
AskJeeves(www.ask.com)	有	无	无
AltaVista(www.altavista.com)	有	有	无
AOL(search.aol.com)	有	无	无
MetaCrawler(www.metacrawler.com)	有	无	无
Dogpile(www.dogpile.com)	有	无	无
Teoma(www.teoma.com)	有	无	有
Vivisimo(vivisimo.com)	无	无	无
百度(www.baidu.com)	无	无	无
阿里巴巴(www.alibaba.com.cn)	无	无	无
雅虎中国搜索(cn.search.yahoo.com)	有	无	无
搜狐搜索——“搜狗”(www.sogou.com)	无	无	有
新浪搜索(cha.sina.com.cn)	无	无	无
网易搜索(so.163.com)	无	无	无
天网搜索(e.pku.edu.cn)	无	无	无
慧聪(www.hc360.com)	无	无	无
一搜(www.yisou.com)	有	无	无
中国搜索(www.zhongsou.com)	无	无	有

[图片来源]寿亦敏.跨语言信息检索的国内外比较研究[J].情报资料工作,2009(4):53-57

《网络信息内容安全》讲义/张华平/2010

五种技术路线



热点研究领域

- ▶ 1.翻译歧义

词性、并行语料库、互信息、Ontology技术

- ▶ 2.翻译资源构建

- ▶ 3.专有名词识别与音译

未登录词(out of vocabulary, OOV)

- ▶ 4.翻译技术

机器翻译系统、基于语料库、基于字典

- ▶ 5.系统评价

TREC、NTCIR、CLEF

- ▶ 6.多种方法的融合

- ▶ 7.多媒体跨语言检索

- ▶ 8.在专业数据库中的应用



谢谢
Thank you

Contact

Email: kevinzhang@bit.edu.cn

Welcome to visit my blog

<http://hi.baidu.com/drkevinzhang/>