语料库加工

双语句子自动对齐&
双语词典获取

杨沐昀
语言技术研究中心
哈工大计算学部

基于长度的双语句子自动对齐

Prelude

Decompose Task and Formulate It into Model

- 双语语料库加工中的一个基本问题：
  - 缘起
  - Given thousands pairs of paper abstract translations between Chinese and English, can we extract a Chinese-English bilingual dictionary for academic writing?
    - E.g. for computer domain, for electronic domain….

Prelude

- We need word correspondence info;
  - How to identify the word likely to correspond?
    - Frequency? Position? Length?....too complex
- Decompose this task:
  - Find the smaller unit with corresponding words
  - Statistics would do the rest…
- How? Sentence alignment

双语句子自动对齐

- 句子对齐问题描述
- 基于长度的句子对齐方法
- 基于长度的汉英句子对齐性能

句子对齐问题描述

汉语	英语	类型
1995年初我来成都的那天，没想到会是在一个冬季的漆黑的日子。	I little thought when I arrived in Chengdu in the dark, dark days of winter, early in 1995, that I would still be here more than five years later.	1：1
那时我也根本没有想到会在这儿呆上五年，也不知道我会遇到一位成都的女儿，并且后来还娶她为妻。 一个完全陌生的家庭接纳了我，我也因此成为成都的一部分。	I little knew that I would meet one of Chengdu's daughters, and later marry her, thus acquiring a whole new family who embraced me as one of them, and thus I became part of this place.	2：1

句子对齐问题描述

中国支持在平等参与、协商一致、求同存异、循序渐进的基础上，开展多层次、多渠道、多形式的地区安全对话与合作。

中国参加了东盟地区论坛、亚洲建立协作与建立信任措施会议、亚太安全合作理事会和东北亚合作对话会等活动，主张通过这些政府和民间讨论安全问题的重要渠道，增进各国的相互了解与信任，促进地区和平与稳定。

……

China advocates regional-security dialogue and cooperation at different levels, through various channels and in different forms.

Such dialogue and cooperation should follow these principles: participation on an equal footing, reaching unanmimity through consultation, seeking common ground while reserving differences, and proceeding in an orderly way and step by step.

China has participated in the ASEAN Regional Forum (ARF), Conference on Interaction and Confidence-Building Measures in Asia (CICA), Council on Security Cooperation in Asia and Pacific Regional (CSCAP), Northeast Asia Cooperation Dialogue (NEACD) and other activities, holding that all countries should further mutual understanding and trust by discussions on security issues through these important governmental and non-governmental channels, so as to promote regional peace and stability.

……

句子对齐（识别句珠）

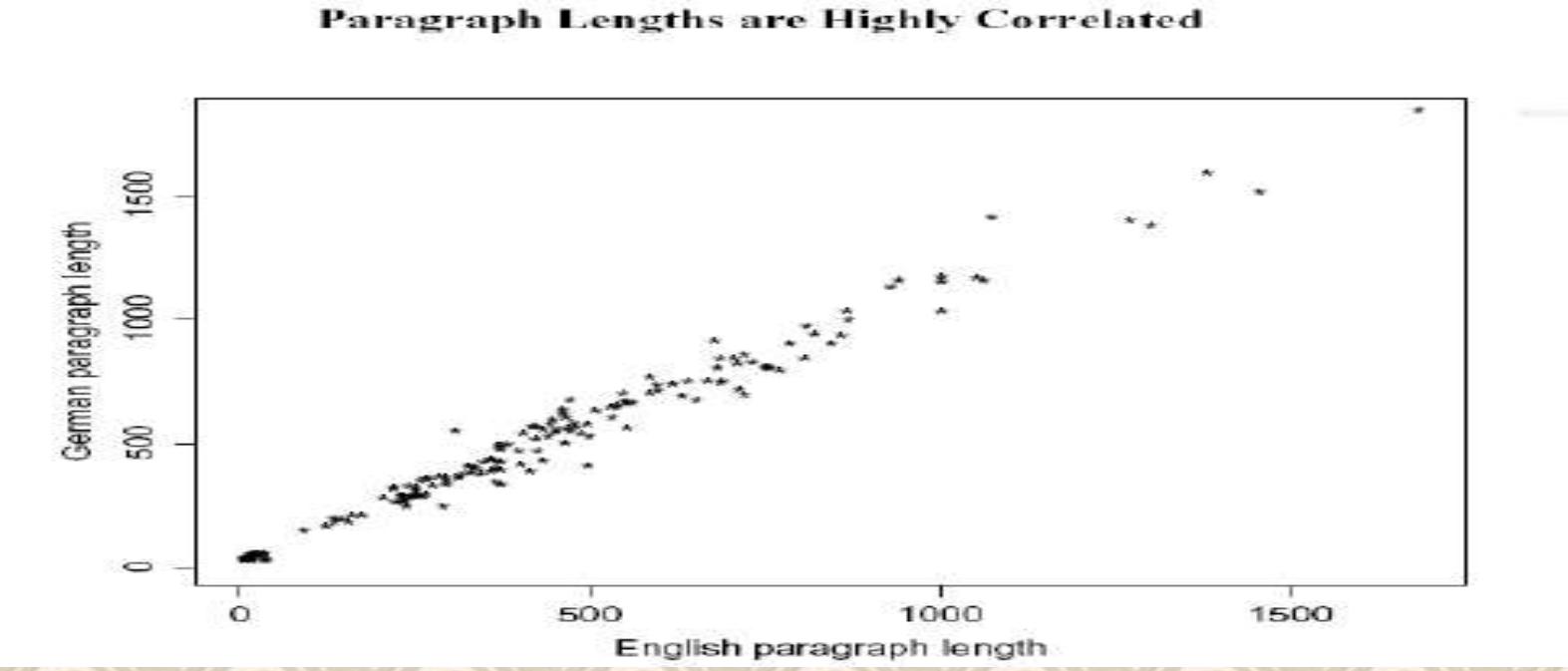
句子对齐问题描述

- 识别双语文本中句子之间的对应关系;
- 将给定双语文本:
  - $S=s_1, s_2, \ldots\ldots s_n$　　$T=t_1, t_2, \ldots\ldots t_m$
  - 转换成一个句珠序列:$B = b_1, b_2\ldots\ldots b_k$
  - 要求：最小（句珠内无句珠）；
    唯一（一个句子仅属于一个句珠）；
    无交叉（后句对齐一定在前句对齐位置之后）
- 该句子对齐概率为：

$$P(B \mid (S,T)) = P(B) = \prod_{i=1}^{k} p(b_i)$$

基于长度的句子对齐

- **Observation**
  - **Longer sentences in one language tend to be translated into longer sentences in other language;**
  - **Shorter sentences tend to be translated into shorter sentences;**
  - **E.g.: English paragraph length correlates German paragraph**



基于长度的句子对齐

- 基本思想：源语言和目标语言的句子长度存在一定的比例关系

$$
\begin{aligned}
p(b_i) &= p(match|(L_1, L_2)) \\
&= p((L_1, L_2)|match) \bullet p(match) \\
&= p(L_1, L_2) \bullet p(match)
\end{aligned}
$$

- 用两个因素来估计一个句珠的概率
  - p(L1,L2):对齐句珠中源语言和目标语言中句子的长度
  - p(match):对齐模式,源语和目标语中的句子数

基于长度的句子对齐

- Assuming $p(l_1,l_2)$ with normal distribution:

S中任意一个字符在T中所对应的字符数是个随机变量，记做X

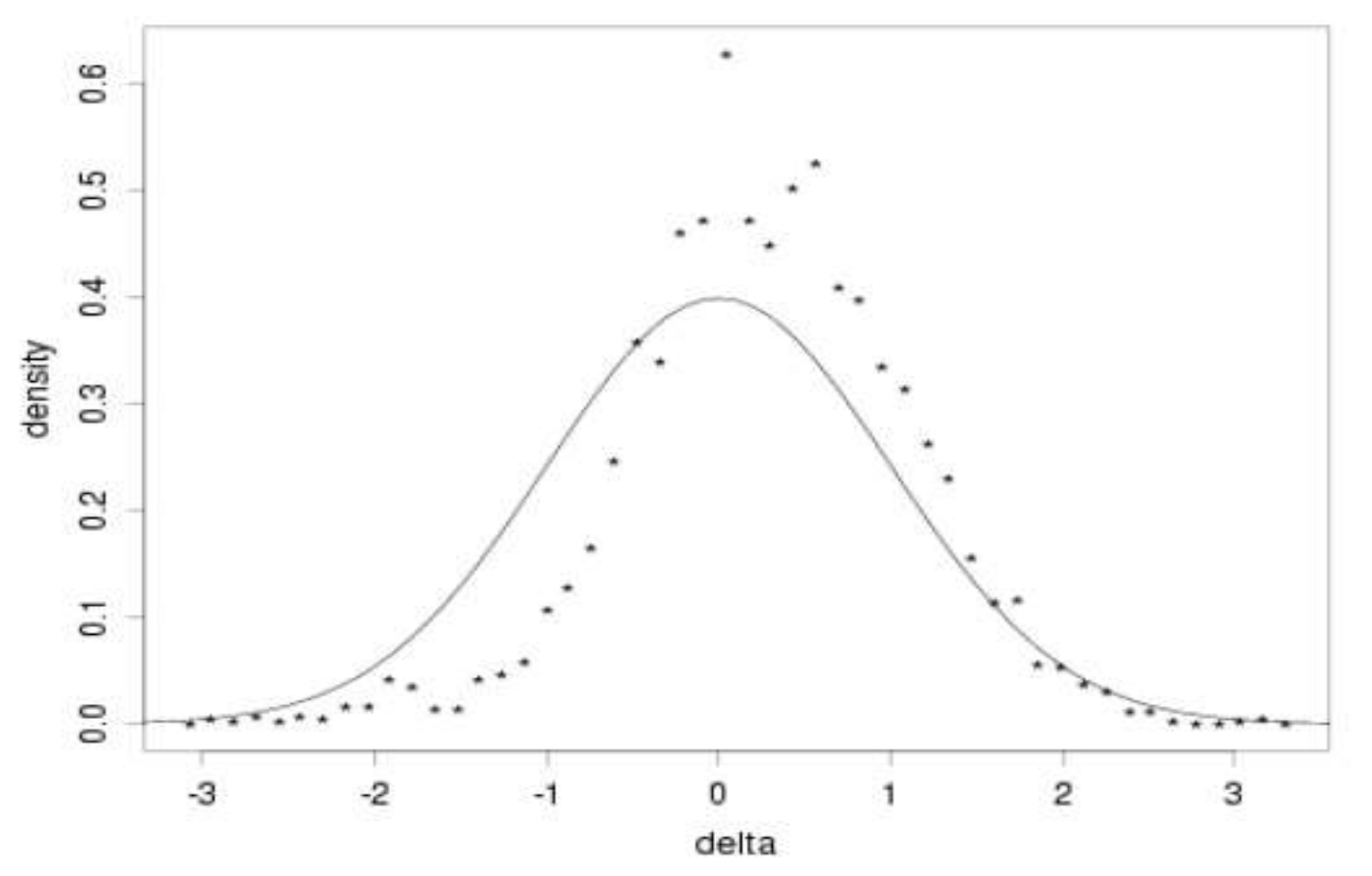
X呈正态分布，X的期望记做c，X的方差记做V^2

由此则可定义随机变量 δ 来度量两个句子之间的长度差距关系

$$\delta(l_i, l_j) = \frac{l_j - c \times l_i}{\sqrt{l_i \times V^2}}$$

基于长度的句子对齐

- δ 服从标准正态分布



基于长度的句子对齐

- 随机变量X的期望c和方差V^2可以从已经对齐好的双语平行语料库中估算得到

比如：英语-法语 $c \approx 72302/68450 \approx 1.06$

$V^2 \approx 5.6$

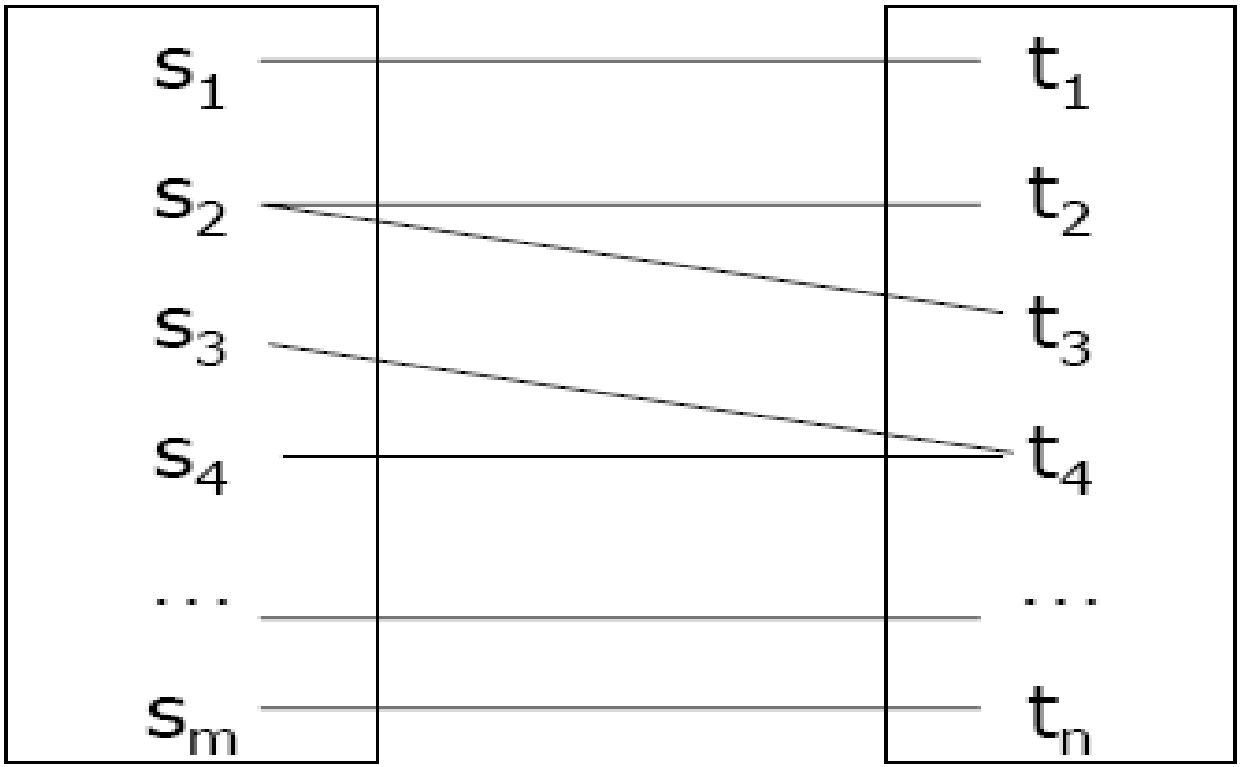
Gale & Church (1993)

英语-汉语 $c \approx 1.46$ $V^2 \approx 2.9$

刘昕 等(1995)

基于长度的句子对齐

- p(match)对齐模式



基于长度的句子对齐

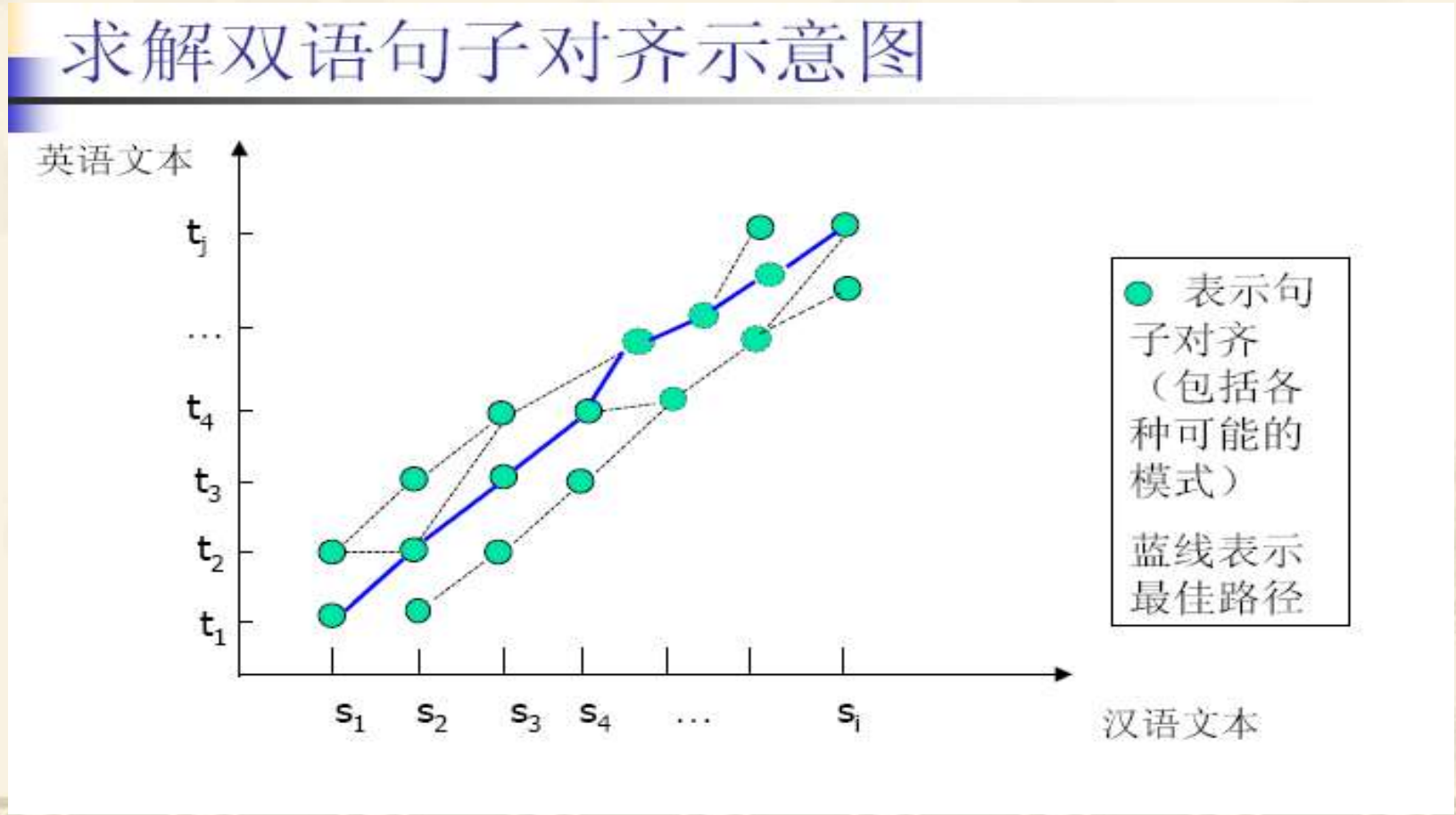
- Gale & Church(1993) 定义了六种配对模式，在实际语料[1]中的分布频度为：

句子配对模式 (Match)	出现次数	概率 P(Match)
1-0 或 0-1	13	0.0099
1-1	1167	0.89
1-2 或 2-1	117	0.089
2-2	15	0.011
	1312	1.00

Note1: UBS/Union Bank of Switzerland出版的经济报告，同时使用英、法、德三种语言

基于长度的句子对齐

- 最优路径的搜索：采用动态规划算法



基于长度的英汉句子对齐性能

- 汉英句子长度关系： $c=1.703$，$\sigma^2=2.88$
- 汉英句子对齐模式
- 实验语料
  - 计算机专业文献
    - 汉语1727句/英语1866句
  - 汉英对照《越女剑》
    - 汉语700句/英语898句
  - 杂类：新闻、议论文、说明文
    - 汉语2715句/英语3251句

汉英句子 翻译匹配模式	匹配模式 出现频率
1：1	0.813
1：2或2：1	0.134
1：3或3：1	0.031
2：2	0.015
1：0或0：1	0.007

基于长度的英汉句子对齐性能

- 实验结果

	计算机文献		小说		杂类	
	召回率	正确率	召回率	正确率	召回率	正确率
长度方法	**93.3%**	**94.5%**	**73.1%**	**76.2%**	**84.0%**	**82.6%**

【中文】今天,高可用性在中档计算机市场上占了主导地位,甚至进入了PC机王国。

【英文】Today high availability dominates the midrange computing markets and is even entering the PC realm.

【中文】文种皱眉道：“范贤弟，吴国剑士剑利术精。固是大患，而他们在群斗之时，善用孙武子遗法，更是难破难当。”

【英文】Wen Chung frowned "Brother Feng, the sharpness of their swords is a major problem, also the way their swordmen worked together in groups in accordance to Sun Tzu's Art of War.”

【中文】书籍源源不断地问世，因此选定"名著"书目的工作似乎也无止境。

【英文】There is no end to the making of books. Nor does there seem to be any end to the making of lists of "great books”.

小结

- 统计方法（语言表层物理特征）—长度
- 基于长度的双语句子自动对齐优点
  - 不依赖于具体的语言；
  - 速度快；
  - 效果好
- 改进方向
  - 由于没有考虑词语信息，有时会产生一些明显的错误
- 思考题
  - 长度计算可以采用词数或者字节数 (which is better?)

基于共现的双语词典的获取

基于双语语料库的翻译词典获取

- 基本思想：如果汉语词出现在某个双语句对中，其译文也必定在这个句对中。

  I/ am/ not/ familiar/ with/ **water polo**/ ./

  我/ 不/ 懂/ **水球**/ 比赛/ 规则/ 。/

- Frequency would be helpful
  - freq (c-word, e-word)
  - If co-occurrence reliable?
    - f (c-word)& f(e-word) ?

基于共现的词汇对译模型

- How to balance f (c,e), f(c) & f(e)?
- 4种（6个）常用公式：
  - **Dice系数、**
  - **互信息（平方互信息、立方互信息）**
  - **联列表**
  - **对数相似性公式**

基于共现的词汇对译模型

- Wc = a word in the Chinese text of the parallel corpus
- We = a word in the English text of the parallel corpus
- freq(Wc,We)=frequency of Wc and We co-occur in a bead
- freq(Wc) = the frequency of Wc in the Chinese text
- freq(We) = the frequency of We in the English text
- N = number of sentence beads

$$h_{DICE}(Wc, We) = \frac{2 \times freq(Wc, We)}{freq(Wc) + freq(We)}$$

$$h_{MI}(Wc, We) = \log \frac{freq(Wc, We)}{freq(Wc) \times freq(We)}$$

基于共现的词汇对译模型

	Wc	~ Wc
We	a=freq(Wc,We)	b=freq(We)-freq(Wc,We)
~We	c=freq(Wc)-freq(Wc,We)	d=N-a-b-c

$$h_{CT}(Wc, We) = \phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Log Likelihood Ratio:

H(Wc,We)=2[logL(p1,a,a+b)+logL(p2,c,c+d)-logL(p,a,a+b)-logL(p,c,c+d)]

logL(p,n,k) = klog(p) + (n-k)log(1-p)

p1 = a/(a+b),

p2 = c/(c+d)

p = (a+c)/(a+b+c+d)

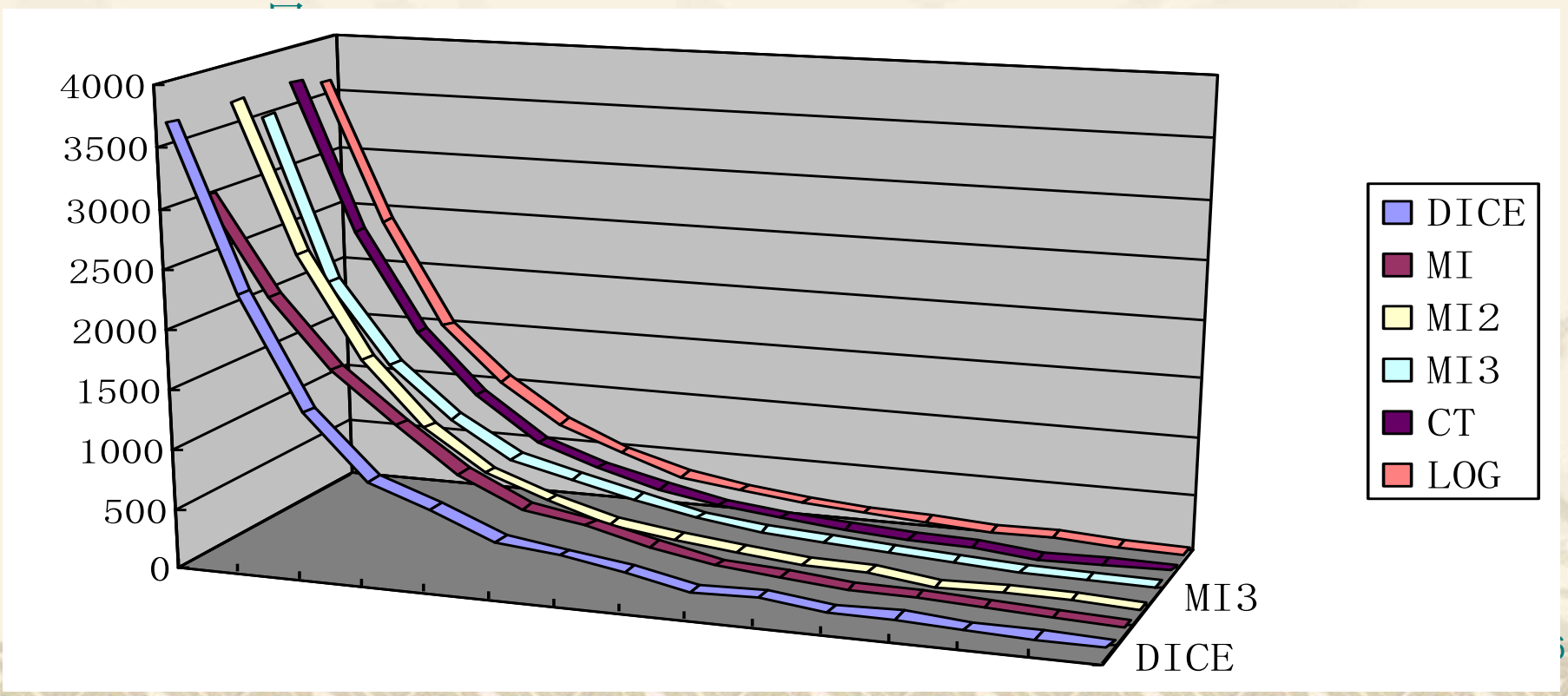
基于共现的词汇对译模型

- 性能对比实验
  - 实验语料：30094句对

	总词数(含标点)	单词数	频度>1 的单词数	词对共现>1 包含单词数
汉语	380,524	17,711	10,682	10195
英语	324,302	10,688	6,581	6115

基于共现的词汇对译模型

- **评价方式：** 专家独立于上下文进行判别
  - **评价1：每5000个**翻译词对候选中正确的译文数



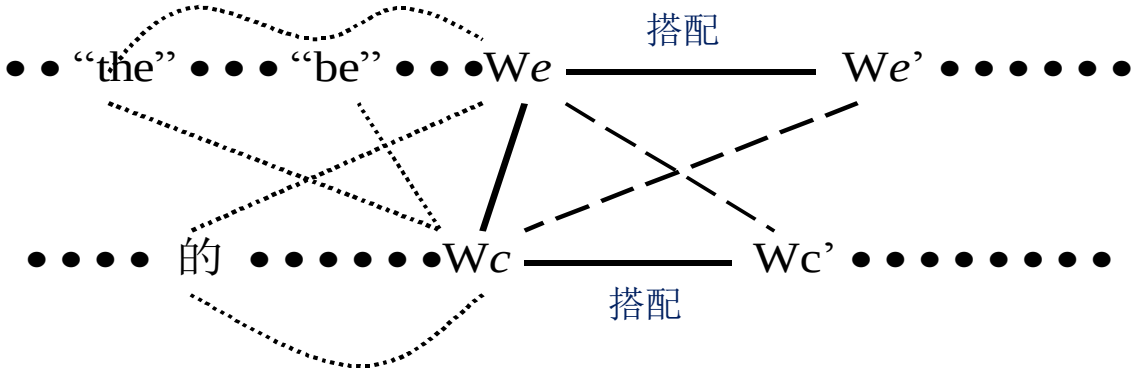
基于共现的词汇对译模型

- **评价2：综合考虑翻译词典的性能**
  - 评价指标的改进：译文类型得分×位置加权系数
- 10195个汉语单词的Top10译文(68144)

	Dice	MI	MI^2	MI^3	CT	**Log**
完全正确译文	8677	8350	8685	8580	8685	**8703**
部分正确译文	2165	2097	2217	2178	2215	**2237**
加权得分	8713.85	8094.85	8759.7	8757.5	8752.4	**8888.7**
加权正确率	*20.37%*	*18.93%*	*20.48%*	*20.47%*	*20.46%*	***20.78%***
无正确译文的汉语词数	2859	2996	2842	2832	2847	2831

基于共现的词汇对译模型

- **错误分析：**
  - 低频词
  - 词形：　久 vs“ago” /“long”/ “Long” /“Longer”
  - 间接共现（Indirect Association）
    - 一个词的高频上下文同时和该词汇译文产生共现



间接共现问题的解决方法研究

- 1.高频干扰词(f>1000)--31个汉语/29个英语

汉语词10195 =>8587;平均正确译文1.12;平均精度

	Dice	MI	MI^2	MI^3	CT	**Log**
完全正确译文	8557	8325	8567	8579	8560	**8596**
部分正确译文	2030	1983	2049	**2065**	2047	2061
加权得分	8571.65	8038.65	8598.4	8704.2	8591.05	**8742.7**
加权正确率	29.83%	27.97%	29.92%	30.29	29.89%	**30.42%**
无正确译文的汉语词数	1357	1447	1355	1345	1359	**1343**

间接共现问题的解决方法研究

- 2.基于词性信息的搭配噪声过滤

序号	汉语词性	英语词性	互译概率
1	r (代词)	PRP (代词)	0.654
2	vz (助动词)	MD (情态动词)	0.563
3	c (连词)	CC (并列连词)	0.555
4	ng (普通名词)	NN (铺通名词)	0.449
5	d (副词)	RB (副词)	0.415
6	nm (人名)	NNP (专有名词)	0.409
7	p (介词)	IN (介词)	0.343
8	m (数词)	CD (数词)	0.339
9	a (形容词)	JJ (形容词)	0.312
10	vx(系动词)	VBZ(动词单三)	0.256

间接共现问题的解决方法研究

- 2.基于词性信息的搭配噪声过滤(续)
  - 错误分析：转译现象/复合词干扰/词性体系的影响

Log 模型	未考虑词性	考虑词性后
完全正确译文	8596	8144
部分正确译文	2061	1970
加权得分	8742.7	7920.65
加权正确率	30.42%	27.56%
无正确译文的汉语词数	1343	1446

汉英词典的迭代获取策略

- **迭代策略**
  - 1) 初始化；
  - 2) 使用对数相似性模型计算汉英翻译词对候选；
  - 3） 选取前n个汉英对译词对；
  - 4) 双语句对中剔除选定的翻译词对；
  - 5） 若不满足终止条件，重复步骤2；
- 几点说明： 复合词暂未考虑；可加入交互方式;

汉英词典获取的整体模型和试验

- 实验一： n的大小

Log 模型 （无高频词）	直接结果 n=5000	迭代结果 1 n=1000	迭代结果 2 n=500
含汉语词数	3298	3645	3713
完全正确译文	3412	3949	4049
部分正确译文	397	453	435
加权得分	3520.9	4061.65	4148.5
无正确译文的汉语词	157	111	99

汉英词典获取的整体模型和试验

- 是否迭代极限对比：n=1

Log 模型	没有迭代	迭代后(n=1)
汉语词总数	**8587**	**7812**
互译词对总数	43427	15424
完全正确译文	8596	8411
部分正确译文	2061	1601
加权得分	8742.7	8665.85
加权正确率	**30.42%**	**65.65%**
无正确译文的汉语词数	1343	849

汉英词典获取的整体模型和试验

- 意外发现：统计提取后存在众多词典译文

Log 模型（无高频词）	使用词典	无词典
互译词对总数	32401	33191
汉语词数	11628	7881
前 10 互译词对	24315	19194
完全正确译文	15336	8144
部分正确译文	1481	1546
加权得分	14885.6	8385.05
加权准确率	72.50%	54.14%
无正确译文的汉语词数	668	1024

汉英机器翻译词典的获取

- 小结
  - 以共现统计手段、结合迭代获取策略可以获取双语语料库中62%的汉语单词的译文，而且整体准确率达到72.5%
- 改进方向
  - **翻译单位识别;**
    - **引入结构信息（parsing）;**
    - **采用n-gram方法;**
  - **引入位置信息;**