语料库加工：基于长度的双语句子自动对齐

统计：从描述到推断

杨沐昀

教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)

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….
- 共现 统计是不是还可以用？

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?

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.

……

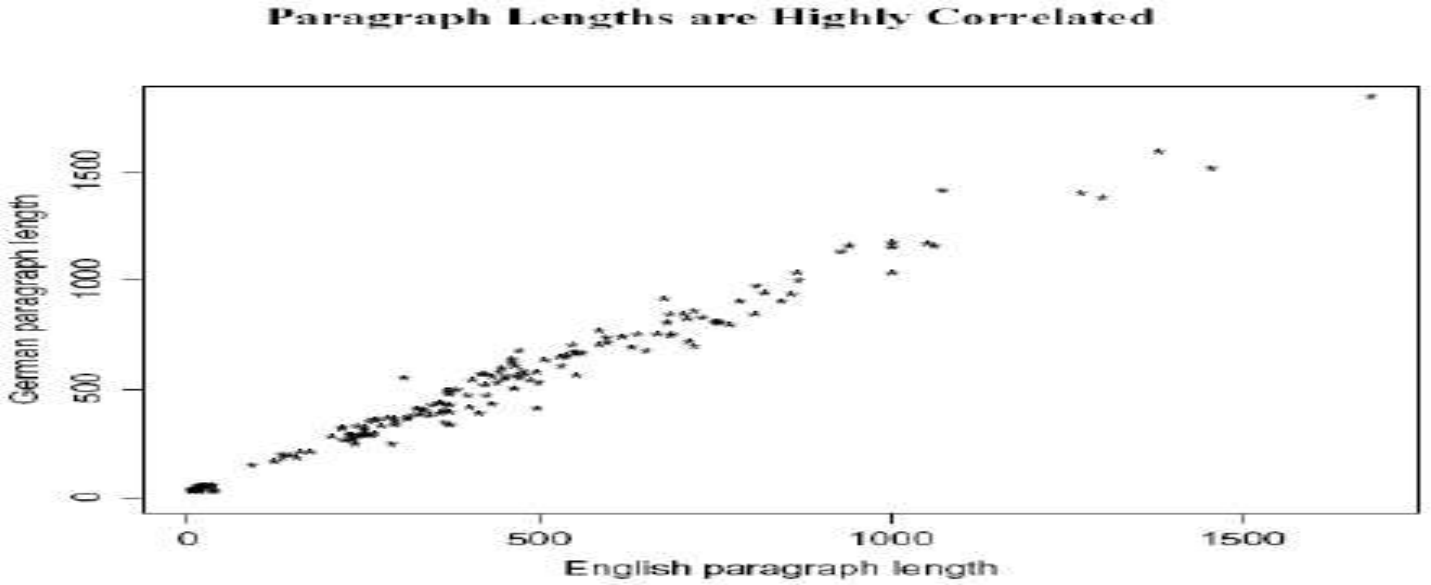
句子对齐问题描述

- **识别双语文本中句子之间的对应关系;**
- **将给定双语文本:**
  - $S=s_1, s_2, \ldots\ldots s_n \quad T=t_1, t_2, \ldots\ldots t_m$
  - **转换成一个句珠序列**:$B = b_1, b_2\ldots\ldots b_k$
  - **要求:最小、唯一、无交叉**
- **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 length at 0.991



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(L_1,L_2):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

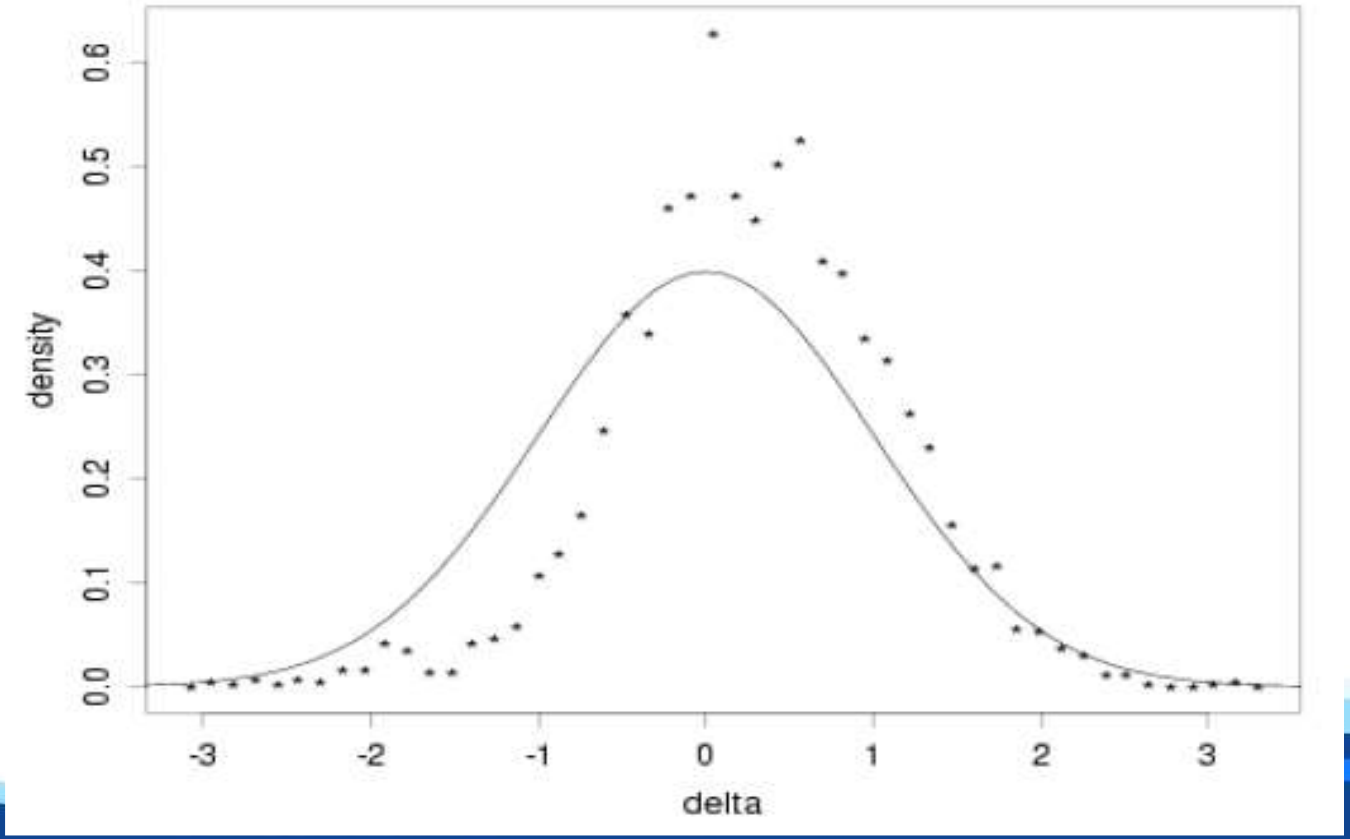
X呈正态分布，X的期望记做c，X的方差记做V^2

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 $c \approx 72302/68450 \approx 1.06$

  $V^2 \approx 5.6$

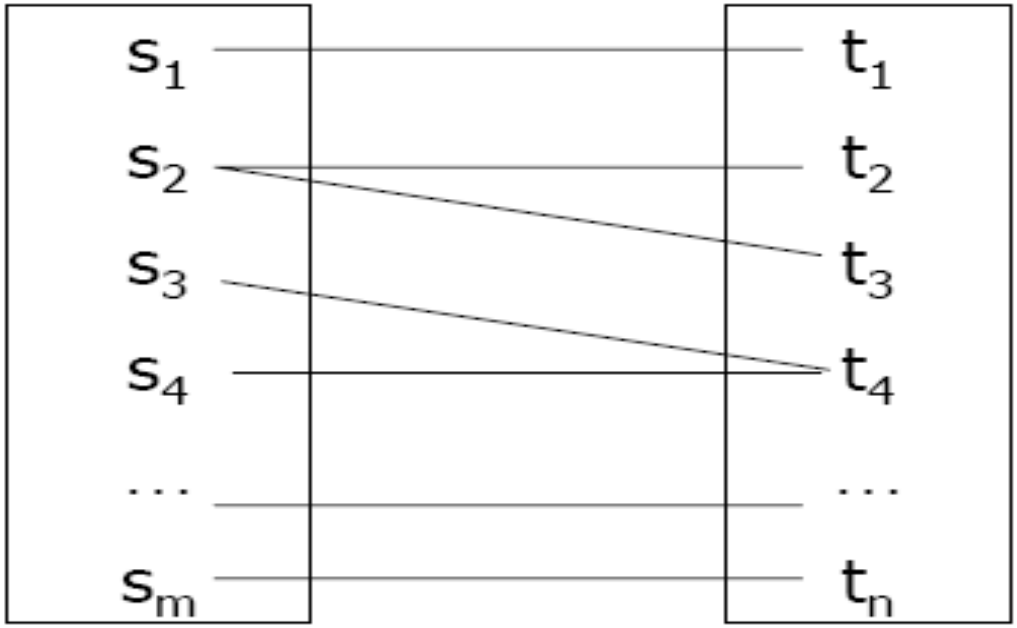
  Gale & Church (1993)

  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

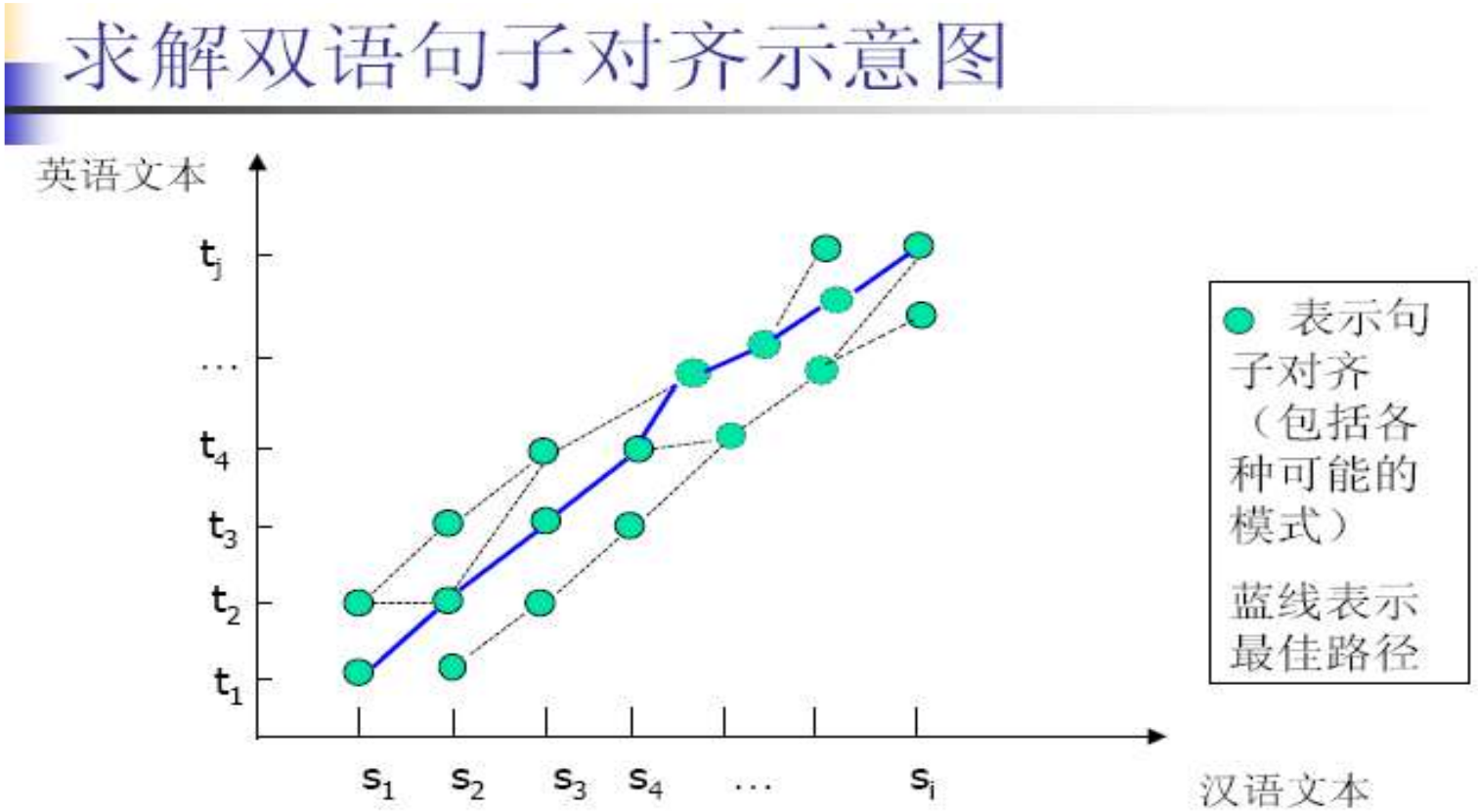
- Gale & Church(1993) 定义了六种配对模式，在实际语料[1]中的分布频度为：

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



基于长度的汉英句子对齐性能

- 汉英句子长度关系： c=1.703，σ^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

基于长度的汉英句子对齐性能

- 实验结果

【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.

	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】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

- 语言物理特征—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
- 问题
  - 由于没有考虑词语信息，有时会产生一些明显的错误
- 讨论
  - 长度计算可以采用词数或者字节数(which is better?)