北京航空航天大學
BEIHANG UNIVERSITY

自然语言处理

人工智能研究院

主讲教师 沙磊

平行文本的自动对齐

概要

- 平行文本自动对齐概述
- 双语句子级对齐简介
- 双语词语级对齐简介

什么是平行文本(parallel text)？

- 按照语料库所涉语种，语料库可区分为：
  - 单语语料库(monolingual corpora)

    涉及一种语言

    由一种语言的文本组成，例如，汉语文本
  - 多语语料库(multilingual corpora)

    涉及多种语言

    由多语平行文本组成，例如，汉-英-日三语文本
- 单语信息处理要用到单语语料库
- 多语信息处理要用到多语语料库，例如：机器翻译
- 双语语料库是涉及两种语言的语料库，由双语平行文 本组成

什么是平行文本?

- 多语平行文本由多个单语文本组成，这些文本之间具有翻译关系
- 双语平行文本由两个单语文本组成，这两个单语文本互为译文。
- 例如一个汉语文本和它的英文译文文本构成了一个汉英双语平行文本。
- 多语平行语料库又称作翻译语料库(translation corpora)。
- 多语平行语料库包含原文及其译文，是机器翻译等多语信息处理的重要资源。

什么是平行文本？

①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。②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。

①China advocates regional-security dialogue and cooperation at different levels, through various channels and in different forms. ②Such dialogue and cooperation should follow these principles: participation on an equal footing, reaching unanimity through consultation, seeking common ground while reserving differences, and proceeding in an orderly way and step by step. ③China 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.

双语对齐处理(Bilingual Alignment)

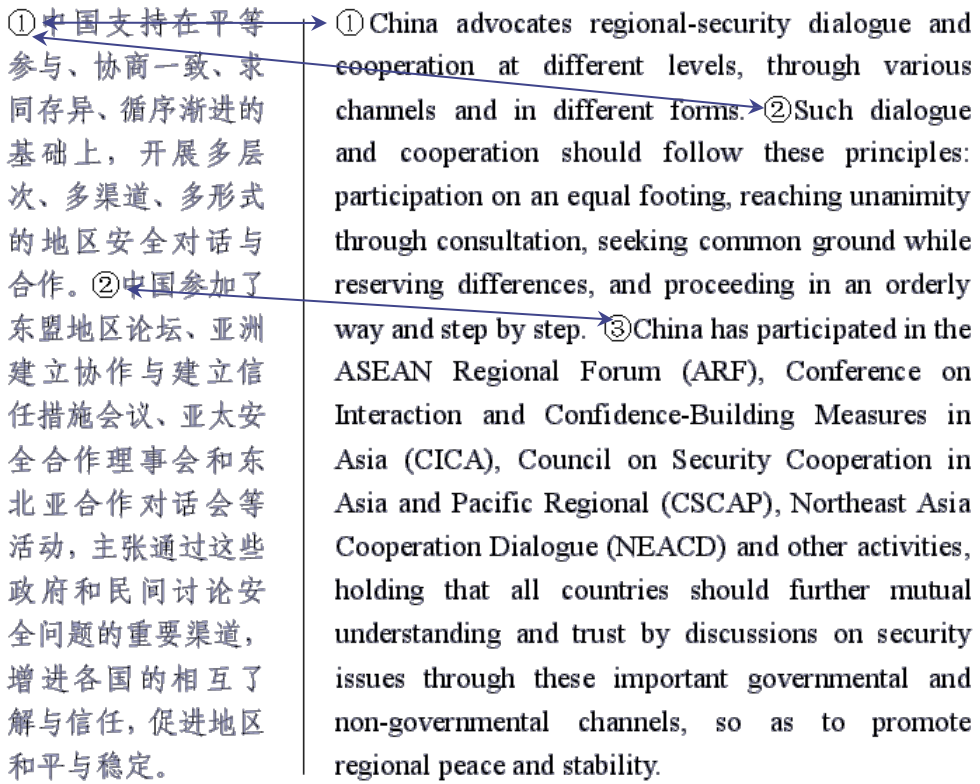
- 所谓**双语对齐处理**就是在两种语言文本的不同语言单位之间建立对应关系，也就是确定源语言文本中哪个(些)语言单位和目标语言文本中哪个(些)语言单位互为 翻译关系。
- 所谓自动双语对齐处理指的是通过一定的算法，由计算机完成在双语文本间建立对齐关系。
- 对齐可以在各种语言单位间进行，例如：文本级、段落级、句子级、短语级、词汇级
- 句子级对齐是最基本的对齐(段落对齐可视为一种特殊的句子的对齐)。也是目前解决的最好的一个对齐问题， 已经比较成熟。

概要

- 平行文本自动对齐概述
- **双语句子级对齐简介**
- 双语词语级对齐简介

双语句子级对齐

- 在双语文本间建立句子一级的对齐关系，就是要确定源语言文本中哪个(些)句子和目标语言文本中哪个(些) 句子互为译文。



双语句子级对齐

- 形式定义：令S为原文文本、T为译文文本，则：

  $S = s_1 s_2 \ldots s_l \quad T = t_1 t_2 \ldots t_m$

  寻求$A = a_1 a_2 \ldots a_n$，其中：　$a_i = (s_j..s_k, t_p..t_q)$，
  且原文片断$s_j..s_k$与译文片断 $t_p..t_q$ 互为译文，二者间也
  不存在更进一步的句子级对齐。(对齐句对、句珠)

- 大部分情况下，一个原文句子对应一个译文句子。一 个句对由一个原文句子和译文句子构成。

- 人工对齐准确难以完成海量数据的对齐

双语句子级对齐

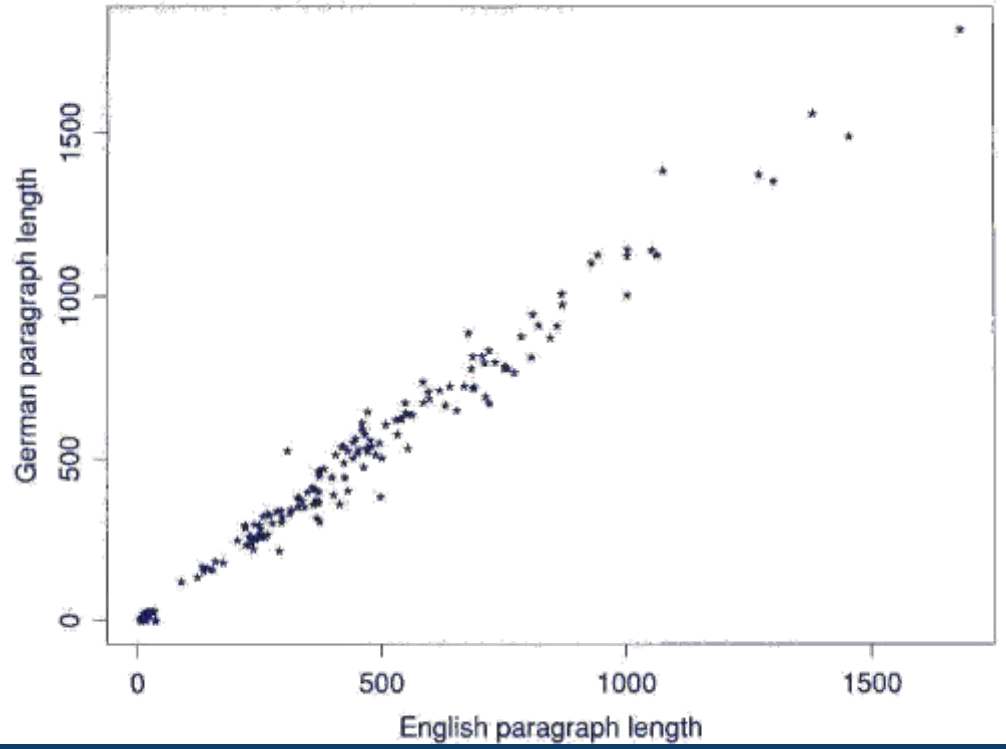
- 关键问题
  - 机器不能在理解的基础上进行对齐
  - 并非严格的一一对应
  - 译文可能涉及语序的调整
  - 可能出现省略不译的现象，反之译文中也可能增加原文中没有的内容
- 这些问题解决好了，对齐问题就好办了
- 这些问题严重吗？不太严重。
  - 从句子层级看，语序不会剧烈调整
  - 大部分情况是一一对应

双语句子级对齐

- 句子对齐的基本方法
  - 基于长度的对齐方法
    - Brown等人的工作(1991)
    - Gale等人的工作(1993)
  - 基于单词的对齐方法
    - Kay等人的工作(1993)
- 两种方法对齐准确率都较高，对一般文本，都在90% 以上。
- 基于长度的对齐方法效率优于基于单词的对齐方法。
- 基于单词的对齐方法：利用单词的对应关系，来决定句子的对齐关系。

双语句子级对齐

- 依据：
  - 互为翻译的两个句子在长度上高度相关。
  - 翻译时，句子顺序不做剧烈改变。（不考虑交叉）



双语句子级对齐

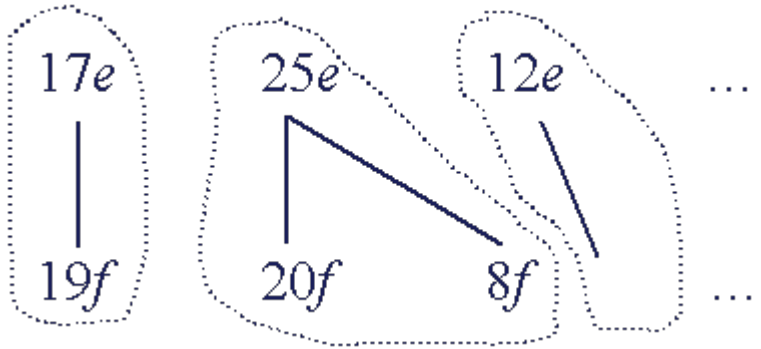
- Gale 考虑的对齐模式(英语-法语、英语-德语)
  (1) 1-0
  (2) 0-1
  (3) 1-1
  (4) 1-2
  (5) 2-1
  (6) 2-2

- Brown等考虑的对齐模式(英语-法语)
  (1) *e*
  (2) *f*
  (3) ef
  (4) eef
  (5) *eff*

- 其他语言呢？

双语句子级对齐

- 基于长度的对齐方法只利用了文本中句子的长度信息
- 待对齐的两个文本不过是两个数字(长度)序列



- 长度的衡量
  - 以词为单位
  - 以字符为单位
- 长度序列怎样转换为对齐模式序列?

1-1 1-2 …

双语句子级对齐

长对长、短对短

- 隐藏层：生成对齐模式系列
- 第二层：为每个对齐模式生成句子长度
  - $\hat{A} = \underset{A}{\operatorname{argmax}} P(A \mid LS, LT) = \underset{A}{\operatorname{argmax}} P(LS, LT \mid A) P(A)$
- 不同的对齐模式间互相独立

$$P(A) = \prod_{i=1}^{n} type_i$$

- 长度和对齐模式有关

$$P(LS, LT) = \prod_{i=1}^{n} P(lsg_i, ltg_i \mid type_i)$$

优先选择下面两个长度正相关的对齐结果

$ls_k + ls_{k+1} + \ldots + ls_{k+x-1}$

$lt_r + lt_{r+1} + \ldots + lt_{r+y-1}$

$$\hat{A} = \underset{A}{\operatorname{argmax}} P(A \mid LS, LT)$$

$$= \underset{type_1 type_2 \ldots type_n}{\arg\max} \prod_{i=1}^{m} P(ls_k ls_{k+1} \ldots ls_{k+x-1}, lt_r lt_{r+1} \ldots lt_{r+y-1} \mid type_i) P(type_i)$$

双语句子级对齐

$$\hat{A} = \underset{type_1 type_2 \ldots type_n}{\arg\max} \prod_{i=1}^{n} P(ls_k ls_{k+1} \ldots ls_{k+x-1}, lt_r lt_{r+1} \ldots lt_{r+y-1} \mid type_i) P(type_i)$$

$$\hat{A} = \underset{type_1 type_2 \ldots type_n}{\arg\max} \sum_{i=1}^{n} \log(P(ls_k ls_{k+1} \ldots ls_{k+x-1}, lt_r lt_{r+1} \ldots lt_{r+y-1} \mid type_i) P(type_i))$$

$$\hat{A} = \underset{type_1 type_2 \ldots type_n}{\arg\min} (-\sum_{i=1}^{n} \log(P(ls_k ls_{k+1} \ldots ls_{k+x-1}, lt_r lt_{r+1} \ldots lt_{r+y-1} \mid type_i) P(type_i)))$$

- 项 $-\log(P(ls_k\ ls_{k+1} \ldots ls_{k+x-1},\ lt_r\ lt_{r+1} \ldots lt_{r+y-1} | type_i) P(type_i))$ 可解释为对齐句对 $(ls_k\ ls_{k+1} \ldots ls_{k+x-1},\ lt_r\ lt_{r+1} \ldots lt_{r+y-1})$ 之间的一种距离度量
- 两个文本间的距离，句对距离之和
- 距离最小、对齐最好

双语句子级对齐

- 如何估算 -log($P(ls_k\ ls_{k+1}\ldots ls_{k+x-1}, lt_r\ lt_{r+1}\ldots lt_{r+y-1}|type_i)\ P(type_i)$)
- Gale等认为，语言L_1中的每一个字符在语言L_2中所对应的字符数C是一个随机变量，随机变量C服从正态分布$N(c,s^2)$，

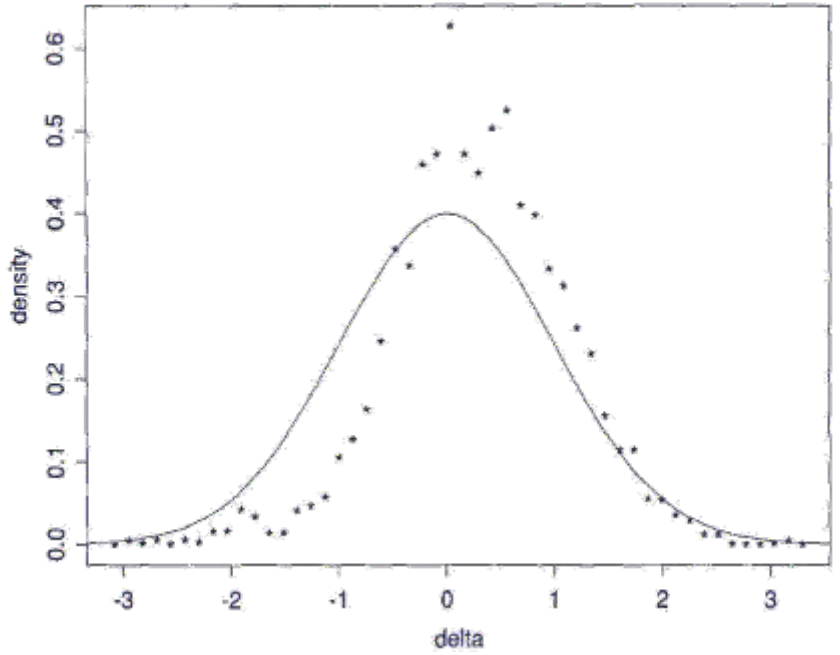
在此基础上，定义随机变量δ:

$$\delta = (l_2 - l_1 c) / \sqrt{l_1 s^2}$$

- δ服从标准正态分布。
- Gale等以字符为单位衡量句子长度。（可靠）
- 基于双语语料，估计c和s^2
  - 德英、法英数据基本一致，Gale认为c和s2独立于语言(至少对欧洲语言而言)
  - c = 1
  - s2= 6.8

双语句子级对齐

- 随机变量δ的分布(依据对齐语料绘制)



- Gale等人的距离定义: -log Prob(δ|*match*)Prob(*match*)

双语句子级对齐

- Prob(*match*)可通过人工对齐的平行语料统计

对齐模式	频率	Prob(*match*)
1-1	1167	0.89
1-0 与 0-1	13	0.0099
2-1 与 1-2	117	0.089
2-2	15	0.011
	1312	1.00

- 对于Prob(δ|*match*)，令

  Prob(δ|*match*) = 2(1-Prob(|δ|)，其中：

$$\mathrm{Prob}(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta} e^{-z^2/2} dz$$

- 查表计算Prob(δ)或其他数值方法求解

双语句子级对齐

- 在已知参数c和s^2以及Prob($match$)后，即可计算最佳对齐枚举文本间所有可能的对齐

  分别计算距离

  选择最佳对齐

- 效率问题
- 引入动态规划，提高效率

  令$d(\ldots)$代表对齐句对的距离，则：

  $d(x_1,y_1;0,0)$、$d(x_1,0;0,0)$、$d(0,y_1;0,0)$
  $d(x_1,y_1;x_2,0)$、$d(x_1,y_1;0,y_2)$、$d(x_1,y_1;x_2,y_2)$

双语句子级对齐

- 设定二维数组D，令

  $D(i,j)$代表原文$s_1 s_2 \ldots s_i$及其译文$t_1 t_2 \ldots t_j$之间的最小距离，则

  $D(I,J)$代表两个文本间的距离

- 动态规划：
  - (1) 初始化 $D(0,0)=0$
  - (2) 递归计算$D(i,j)$
  - (3) 递归终止于$D(I,J)$
- 对齐信息的记录

$$D(i,j)=\min\begin{cases} D(i,j\text{-}1)+d(0,t_j;0,0) \\ D(i-1,j)+d(s_i,0;0,0) \\ D(i-1,j-1)+d(s_i,t_j;0,0) \\ D(i-1,j-2)+d(s_i,t_j;0,t_{j-1}) \\ D(i-2,j-1)+d(s_i,t_j;s_{i-1},0) \\ D(i-2,j-2)+d(s_i,t_j;s_{i-1},t_{i-1}) \end{cases}$$

双语句子级对齐

- Gale给出的一些结果
  - 对齐准确率在96%左右
  - 一多、多一以及多多对齐的错误率较高。2-2对齐的错误率是33%
  - 一一对齐准确率最高
  - 提取对齐距离最小的80%对齐句对，准确率可达99.3%

双语句子级对齐

- 基于长度的对齐方法没有用到任何词汇信息
- 基于词汇对齐信息可以推导出句子对齐信息
- 可利用词汇对应信息改善基于长度对齐的效果
- 锚点：双语文本中有明显对应关系的词汇

  数字、日期

  人名、地名

  同根词(network, Netzwerk international, international)

双语句子级对齐

①第三阶段为 1985 年至今。②1995 年，中国粮食总产量达到 4.666 亿吨，11 年间年均递增 1.2%。③这一时期，中国政府在继续发展粮食生产的同时，积极主动地进行农业生产结构调整，发展多种经营，食物多样化发展较快，猪牛羊肉、水产品、禽蛋、牛奶和水果产量分别达到 4254 万吨、2517 万吨、1676 万吨、562 万吨和 4211 万吨，比 1984 年分别增长 1.8 倍、3.1 倍、2.9 倍、1.6 倍和 3.3 倍。

①The third phase (1985-present): ②In 1995 the country's grain output totaled 466.6 million tons, increasing by an average of 1.2 percent a year over the previous 11 years. ③ While continuing to develop grain production in this period, the Chinese government has initiated measures to readjust the structure of agricultural production and develop a diversified agricultural economy. ④ At the same time rapid progress was achieved in the production of various other kinds of foodstuffs, with the output of meat (pork, beef and mutton), aquatic products, eggs, milk and fruit reaching **42.54** million tons, **25.17** million tons, **16.76** million tons, **5.62** million tons and **42.11** million tons respectively, or 2.8, 4.1, 3.9, 2.6 and 4.3 times the **1984** figures, respectively.

概要

- 平行文本自动对齐概述
- 双语句子级对齐简介
- **双语词语级对齐简介**

什么是词语对齐？

- 在互为译文的两个句子间寻找词语对译关系。
- 形式定义：

  令 $S=s_1 s_2 \ldots s_J$ 代表原文句子

  令 $T=t_1 t_2 \ldots t_I$ 代表译文句子，

  则 二者间词汇级对齐A可定义为

  $A \subseteq \{s_1, s_2, \ldots, s_J\} \times \{t_1, t_2, \ldots t_I\}$

  或者

  $A \subseteq \{(j,i) \mid j \in \{1,2,\ldots,J\}, i \in \{1,2,\ldots I\}\}$

- 过于一般化

什么是词语对齐？

- 限制条件：(原文到译文)不允许一对多的对应关系
- 原文中未译的词对应一个特殊的空词t_0
- 词汇对齐A是从集合$\{1,2,\ldots,J\}$到$\{0,1,2,\ldots,I\}$的映射

  $A:\{1,2,...,J\} \rightarrow \{0,1,2,...,I\}$

  令$a_j=A(j)$，则：

  $A = a_1 a_2 \ldots a_J$

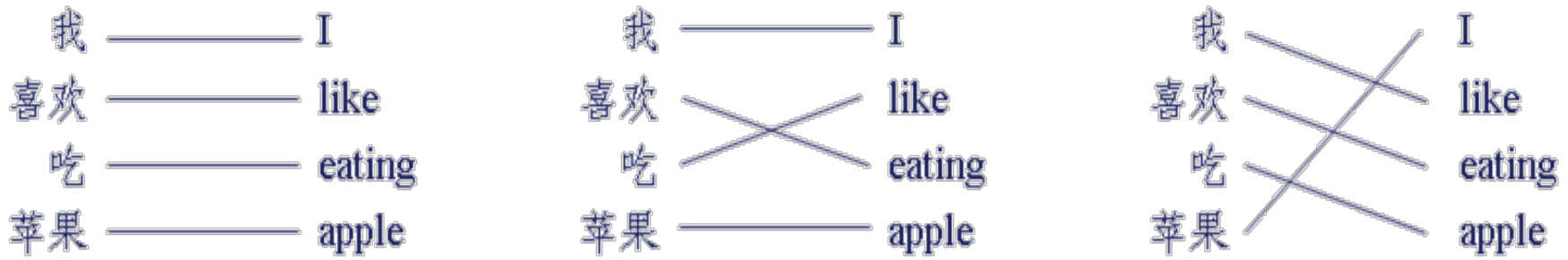
- 词语对齐举例

词语对齐

- 词语对齐较句子对齐困难
  - 词及其译词的长度间不存在相关关系
  - 翻译时，词序发生剧烈变化
  - 对应模式情况复杂
  - 对应关系难以确定（虚词）
- 词语对齐的基本方法
  - 统计模型法
    - 建立统计对齐的数学模型
  - 启发式方法
    - 不一定建立对齐模型，运用假设-检验等技术
- 简要介绍统计模型法

词语对齐

- 从统计角度看，所有的对齐都是可能的，只不过概率大小不同



- 原文句子、译文句子长度分别是J、I，共有多少可能的对齐？

词语对齐

- 统计对齐的任务，就是从众多的对齐中找出概率最大的对齐，即韦特比对齐。

$$\hat{a_1^J} = \underset{a_i^J}{\operatorname{argmax}} P(a_1^J \mid s_1^J, t_1^I)$$

- 另有

$$P(a_1^J \mid s_1^J, t_1^I) = \frac{P(a_1^J, s_1^J, t_1^I)}{P(s_1^J, t_1^I)} = \frac{P(a_1^J, s_1^J \mid t_1^I) P(t_1^I)}{P(s_1^J \mid t_1^I) P(t_1^I)} = \frac{P(a_1^J, s_1^J \mid t_1^I)}{P(s_1^J \mid t_1^I)}$$

- 故有：

$$\hat{a_1^J} = \underset{a_i^J}{\operatorname{argmax}} P(a_1^J \mid s_1^J, t_1^I) = \underset{a_i^J}{\operatorname{argmax}} P(s_1^J, a_1^J \mid t_1^I)$$

求解韦特比对齐

- 可以通过下面的过程计算韦特比对齐
  1) 罗列出原文句子和译文句子间所有可能的对齐
  2) 对每一种对齐，计算$P(S, A|T)$
  3) 寻找能使$P(S, A|T)$取得最大值的A作为韦特比对齐
- 问题一：如何计算$P(S, A|T)$?
- 问题二：罗列所有对齐效率如何？

- 给定S、T，$P(S|T)$与$P(S, A|T)$之间的关系如下：

$$P(s_1^J | t_1^I) = \sum_{a_1^J} P(s_1^J, a_1^J | t_1^I)$$

统计模型参数化及参数训练

- 与其它统计建模问题类似，精确计算$P(S, A|T)$并不现实的，需要对$P(S, A|T)$进行近似与化简，把$P(S, A|T)$的计算问题归结为一组统计参数的函数形式，即：

$$P_\theta(S, A \mid T)$$

- 不同形式的统计建模，对应不同的参数θ。
- 利用训练语料，估算模型参数θ 。令训练语料为：

$$\{ (S_k, T_k) \mid k = 1,2,\ldots K \}$$

则可通过最大似然估计法，估计模型参数。

$$\hat{\theta} = \arg\max_\theta \prod_{k=1}^{K} P_\theta(S_k \mid T_k) = \arg\max_\theta \prod_{k=1}^{K} \sum_{A_k} P_\theta(S_k, A_K \mid T_k)$$

词语对齐问题的统计建模问题

- 如何对$P(S, A|T)$进行统计建模是解决词语对齐问题的关键(参数化)。
  - 对词语对齐问题有足够的描述能力
  - 存在快速有效的计算方法，可以快速训练，并计算最佳对齐。
- 是统计机器翻译的一个子问题。
- IBM公司的Brown提出了五种建模方法，描述能力依次增强，但计算复杂性依次提高。分别称之为IBM模型 一、二、三、四和五。(1993)
- Vogel 提出一种类HMM模型，描述能力介于IBM模型二和三之间。(1996)
- Och提出组合模型，将不同模型加权组合。(2003)

对齐故事一

- $P(S,A|T)$可以(精确)写作

$$P(s_1^J, a_1^J \mid t_1^I) = P(J \mid t_1^I) \prod_{j=1}^{J} P(a_j \mid a_1^{j-1}, s_1^{j-1}, J, t_1^I) P(s_j \mid a_1^j, s_1^{j-1}, J, t_1^I)$$

- 上述公式实际蕴含了一种由译文句子T出发，生成原文句子S及(S,T)之间一种词语对齐的方法。

(1) 基于译文句子 t_1^I，按照概率分布 $P(J \mid t_1^I)$选择原文句子的长度 J；

(2) 基于译文句子 t_1^I、原文句子长度 J，按照概率分布 $P(a_1 \mid J, t_1^I)$从译文句子中选择一个词位 a_1 与原文中第 1 个词位对应。

(3) 基于译文句子 t_1^I、原文句子长度 J、原文第 1 个词位对应的译文词位 a_1，按照概率分布 $P(s_1 \mid a_1, J, t_1^I)$，选择 s_1 作为原文句子的第 1 个词。

对齐故事一

(4) 基于译文句子 t_1^l、原文句子长度 J，原文第 1 个词位对应的译文词位 a_1、原文中第 1 个词 s_1，按照概率分布 $P(a_2 \mid a_1, s_1, J, t_1^l)$从译文句子中选择一个词位 a_2 与原文中第 2 个词位对应。

(5) 基于译文句子 t_1^l、原文句子长度 J，原文中前两个词位对应的译文词位 a_1a_2、原文中第 1 个词 s_1，按照概率分布 $P(s_2|a_1a_2, s_1, J, t_1^l)$，选择 s_2 作为原文句子的第二个词。

(6) 依次类推，循环往复，直到最终按照概率分布 $P(a_J \mid a_1 \ldots a_{J-1}, s_1 \ldots s_{J-1}, J, t_1^l)$和 $P(s_J \mid a_1 \ldots a_J, s_1 \ldots s_{J-1}, J, t_1^l)$分别选择出与原文中第 J 个词位对应的译文词位 a_J 以及原文中第 J 个词 s_J。

- 按照这个过程考虑下面的汉英句对。

  我 喜欢 吃 苹果

  I like eating apple

- 参数数量问题、数据稀疏问题

IBM模型一

- IBM模型一　作了如下的独立性假设

(1) 概率分布 $P(J|t_1^I)$与原文句子长度 J 以及译文句子 t_1^I无关，恒为常数$\varepsilon(0<\varepsilon<1)$，即 $P(J|t_1^I)=\varepsilon$。

(2) 概率分布 $P(a_j|a_1^{j-1},s_1^{j-1},J,t_1^I)$仅依赖译文句子长度 I，译文中 $I+1$ 个词位均以相等的概率与原文词位 j 对应，即 $P(a_j|a_1^{j-1},s_1^{j-1},J,t_1^I)=(I+1)^{-1}$。

(3) 概率分布 $P(s_j|a_1^j,s_1^{j-1},J,t_1^I)$仅依赖位于译文中与原文第 j 个词位对应的译文词 t_{a_j}，即 $P(s_j|a_1^j,s_1^{j-1},J,t_1^I)=t(s_j|t_{a_j})^1$，概率 $t(s_j|t_{a_j})$一般称作译文单词 t_{a_j}到原文单词 s_j 的翻译概率。

- IBM模型一

$$P(s_1^J,a_1^J|t_1^I)=\frac{\varepsilon}{(I+1)^J}\prod_{j=1}^{J}t(s_j|t_{a_j})$$

IBM模型一

- 在IBM模型一中，只有一种类型的参数，即单词翻译概率$t(s|t)$。该参数满足下面的约束条件：

$$\sum t(s|t)=1$$

- 如果模型参数已知，对给定的S、A、T，即可计算出

$$P(S,A|T)$$

- 根据模型一，再次考虑下面的例子
  - 我 喜欢 吃 苹果
  - I like eating apple

IBM模型二

- IBM模型一非常简单，这是以过度简化为代价的，描述能力差。单纯依靠IBM模型一，对齐效果有限。
- IBM模型二改进了IBM模型一中的词位对齐概率

(1) IBM 模型一假定译文中 $I+1$ 个词位以相等的概率与原文词位 j 对应。

(2) IBM 模型二则假定概率分布 $P(a_j \mid a_1^{j-1}, s_1^{j-1}, J, t_1^I)$依赖原文词位 j、原文句长 J 以及译文句长 I，即 $P(a_j \mid a_1^{j-1}, s_1^{j-1}, J, t_1^I) = a(a_j \mid j, J, I)$。

- IBM模型二的一般形式

$$P(s_1^J, a_1^J \mid t_1^I) = \varepsilon \prod_{j=1}^{J} a(a_j \mid j, J, I) t(s_j \mid t_{a_j})$$

- 原文词位和译文词位之间有相关性，模型二予以考虑。

（对比句子对齐问题中的顺序相关性问题）

IBM模型二

- IBM模型二有两类参数，即词位对齐概率$a(a_j \mid j, J, I)$和单词翻译概率$t(s \mid t)$。
- 词位对齐概率$a\ (a_j \mid j, J, I)$满足下面的约束条件：

$$\sum_{i=0}^{I} a(i \mid j, J, I) = 1$$

- 按照模型二，再次考虑下面的例子
  - 我 喜欢 吃 苹果
  - I like eating apple

Vogel的类HMM模型

- Vogel提出了下面的独立性假设

(1) 概率分布 $P(J \mid t_1^I)$仅依赖译文句子长度 I，即 $P(J \mid t_1^I) = P(J \mid I)$。

(2) 概率分布 $P(a_j \mid a_1^{j-1}, s_1^{j-1}, J, t_1^I)$依赖译文句子长度 I、原文第 j-1 个词位所对应的译文词位 a_{j-1}，即 $P(a_j \mid a_1^{j-1}, s_1^{j-1}, J, t_1^I) = P(a_j \mid a_{j-1}, I)$。

(3) 类HMM模型对概率分布 $P(s_j \mid a_1^j, s_1^{j-1}, J, t_1^I)$的处理与IBM模型一、IBM模型二相同，即 $P(s_j \mid a_1^j, s_1^{j-1}, J, t_1^I) = P(s_j \mid t_{a_j})$。

- 类HMM模型的一般形式

$$P(s_1^J, a_1^J \mid t_1^I) = P(J \mid I) \prod_{j=1}^{J} P(a_j \mid a_{j-1}, I) P(s_j \mid t_{a_j})$$

Vogel的类HMM模型

- 翻译具有局部性，原文中邻近的词译成其它语言时，译词大多数情况下仍然保持较近的距离，类HMM模型对此作了考虑。
- 共有三类参数，分别是$P(J|I)$、$P(a_j|a_{j-1}, I)$以及$P(s|t)$
- 与HMM模型的关系

  译文词位a_j对应状态

  译文单词对应状态输出

- 按照类HMM模型，再次考虑例子

  我 喜欢 吃 苹果

  I like eating apple

基于繁殖率的模型

- 前述模型没有考虑一个译文单词对应多个原文单词的问题。
- 不同的单词，有不同的倾向，如：
  - University--大学
  - impolite --不 | 礼貌
- 一个译文单词*t*对应的原文单词的数量称为*t*的繁殖率(fertility)，译文单词的繁殖率实际上是一个随机变量

- IBM模型一、二及类HMM模型未能充分捕捉这种基于单词的倾向性，未能充分考虑译文单词繁殖率的问题。

对齐故事二

(1) 对 $1 \leqslant i \leqslant I$，按照概率分布 $P(\phi_i | \phi_1^{i-1}, t_1^I)$生成单词 t_i 的繁殖率ϕ_i。

(2) 按照概率分布 $P(\phi_0 | \phi_1^I, t_1^I)$生成空词 t_0 的繁殖率ϕ_0。

(3) 对 $0 \leqslant i \leqslant I$，为译文单词 t_i 生成一个由ϕ_i 个原文单词组成的词集τ_i 与之对应，其中 $\tau_i = \{\tau_{i1}, \tau_{i2}, \ldots, \tau_{i\phi_i}\}$[3]。在生成原文词集$\tau_i$ 中元素 τ_{ik} 时，服从概率分布 $P(\tau_{ik} | \tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^I, t_1^I)$。

(4) 对 $1 \leqslant i \leqslant I$，对每个词集$\tau_i$ 生成原文词位集π_i，其中$\pi_i = \{\pi_{i1}, \pi_{i2}, \ldots, \pi_{i\phi_i}\}$，其中$\pi_{ik}$代表 τ_i 中第 k 个词 τ_{ik} 在原文中的词位。在生成原文词集π_i 中元素 π_{ik} 时，服从概率分布 $P(\pi_{ik} | \pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, t_1^I)$。

(5) 对空词 t_0 对应的原文词集τ_0，生成原文词位集π_0。在生成原文词集π_0 中元素 π_{0k} 时，服从概率分布 $P(\pi_{0k} | \pi_{01}^{k-1}, \pi_1^I, \tau_0^I, \phi_0^I, t_1^I)$。

- 再看一个例子

调查 表明 没有 不 正当 的 做法

the investigation revealed no improper behavior

对齐故事二

(1) 首先为每个英文单词选择一个繁殖率，假定选择结果如下：

the – 0、 investigation – 1、 revealed – 1、 no – 1、 improper – 2、 behavior – 1

(2) 为空词 e_0 选择繁殖率 1。

(3) 为繁殖率大于 0 的每个英文单词生成一个对应的汉语词集，假定结果如下：

investigation – {调查}、 revealed – {表明}、 no – {没有}、

improper – {不，正当}、 behavior – {做法}、e_0 – {的}

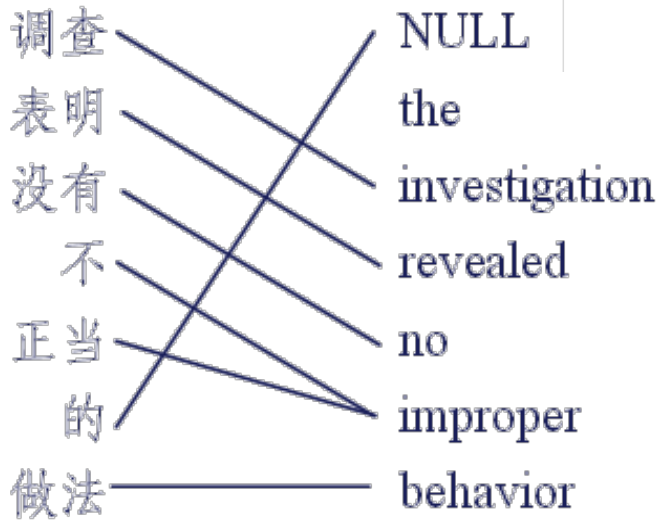
(4) 为上述除空词外的其它词所对应的词集选择原文词位集，假定结果如下：

{调查} – {1}、 {表明} – {2}、 {没有} – {3}、

{不，正当} – {4, 5}、 {做法} – {7}

(5) 为空词 e_0 所对应的词集{的}选择原文词位集{6}。

- 这个过程生成了右图的词对齐方案



对齐故事二

- 依据上述过程，有

$$P(\tau_0^I, \pi_0^I \mid t_1^I) = \prod_{i=1}^{I} P(\phi_i \mid \phi_1^{i-1}, t_1^I) P(\phi_0 \mid \phi_1^I, t_1^I) \times$$
$$\prod_{i=0}^{I} \prod_{k=1}^{\phi_i} P(\tau_{ik} \mid \tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^I, t_1^I) \times$$
$$\prod_{i=1}^{I} \prod_{k=1}^{\phi_i} P(\pi_{ik} \mid \pi_{i1}^{k-1}, \pi_0^{i-1}, \tau_0^I, \phi_0^I, t_1^I) \times$$
$$\prod_{k=1}^{\phi_0} P(\pi_{0k} \mid \pi_{01}^{k-1}, \pi_1^I, \tau_0^I, \phi_0^I, t_1^I)$$

对齐故事二

- 在为译文词选择原文词集时，不同的顺序会导致相同的对齐
  如 improper {不，正当} {4, 5}
  improper {正当，不} {5, 4}

- 考虑所有能产生(*S*,*A*)的情况，有：

$$P(s_1^J, a_1^J \mid t_1^I) = \sum_{(\tau_0^I, \pi_0^I) \in <s_1^J, a_1^J>} P(\tau_0^I, \pi_0^I \mid t_1^I)$$

- IBM模型三、四、五均为上述模型的简化模型，均属于基于繁殖率的模型
- 对于基于繁殖率的模型而言，限于时间关系，不再做介绍。

Och的组合模型

- Och提出如下组合模型：

$$P_M(s_1^J, a_1^J \mid t_1^I) = \frac{\prod_{k=1}^{K} P_k(s_1^J, a_1^J \mid t_1^I)^{\alpha_k}}{\sum_{S,A} \prod_{k=1}^{K} P_k(S, A \mid t_1^I)^{\alpha_k}}$$

  - $P_k(.)$代表模型k
  - α_k是插值系数，代表相应模型的权重
  - $P_M(.)$是K个基本模型的组合模型

- Och将类HMM模型与IBM模型四进行了组合，并将组合模型称为模型六。

计算韦特比对齐

- 在模型建立以后，接下来还要考虑的一个问题是：
  在给定的模型中，怎样计算韦特比对齐
  (假定模型参数已知)

- 理论上当然可以枚举所有对齐方式，对每种对齐方式，
  计算$P(S, A|T)$，在寻求值最大的对齐

- 实际上不现实：
  假定原文句子、译文句子均由10个词组成，二者间可能的对齐方式有11^{10}种之多，也就是大约260亿

计算韦特比对齐(IBM 模型一)

- 幸运的是，对于IBM模型一、二及类HMM模型而言，韦特比对齐的计算并不困难。
- 对IBM模型三、四和五，无法精确计算出韦特比对齐， 要采用一定的近似策略。
- 对于IBM模型一而言，韦特比对齐可按照下面的方式计算

$$\hat{a}_1^J = \arg\max_{a_1^J} P(s_1^J, a_1^J \mid t_1^I) = \arg\max_{a_1^J} \left\{ \prod_{j=1}^{J} t(s_j \mid t_{a_j}) \right\} = \left[\arg\max_{a_j} \{ t(s_j \mid t_{a_j}) \} \right]_1^J$$

- 即顺次为每一个原文单词s_j选择一个能使$t(s_j|t_i)$取最大值的t_i与之对应。
- 举例

我 喜欢 吃 苹果　　I like eating apple

计算韦特比对齐(IBM 模型二)

- 对于IBM 模型二而言，存在类似的求解方法

$$\hat{a}_1^J = \arg\max_{a_1^J} P(s_1^J, a_1^J \mid t_1^I) = \arg\max_{a_1^J} \left\{ \prod_{j=1}^{J} a(a_j \mid j, I, J) t(s_j \mid t_{a_j}) \right\}$$

$$= \left[\arg\max_{a_j} \{ a(a_j \mid j, I, J) t(s_j \mid t_{a_j}) \} \right]_1^J$$

- 考虑例子

  我 喜欢 吃 苹果I like eating apple

- 对于类HMM而言，如何计算韦特比对齐？ 参考HMM中的韦特比算法

统计对齐模型的训练

- 可采用有指导方式，也可采用无指导方式。
- 采用有指导方式，需要事先标记词对齐语料。

  假定已经具有词对齐语料，单词翻译概率$t(s|t)$，可通过下面的方法估计。

$$t(s \mid t) = \frac{tc(s,t)}{tc(t)}$$

- 建立大规模词对齐语料库殊非易事，故而对齐模型训练，通常采用的仍是无指导方式：

  通常采用EM算法进行模型训练

  (迭代计算、逐步求精)

  隐藏信息是对齐信息

统计对齐模型的训练

- 由于没有对齐信息, 无法准确计算$tc(s, t)$
- 对于所有可能的对齐结果, 计算相应的期望频次

$$tc^*(s,t) = \sum_A P(A|S,T)tc(s,t)$$

$$tc^*(t) = \sum_i tc^*(s_i,t)$$

- 基于期望频次, 更新参数

$$t^*(s|t) = \frac{tc^*(s,t)}{tc^*(t)}$$

- 上述步骤分别对应EM算法中的E-Step和M-Step

统计对齐模型的训练

- 所有统计对齐模型，理论上都可采用上述方法进行参数训练。

- 这样的方法不现实，因为要考虑所有可能的对齐方案， 数量庞大
- 对于IBM模型一、二和类HMM模型而言，存在有效方法且能考虑到所有对齐方案。
- 对于其他模型，不存在考虑所有对齐方案的有效算法。（近似算法:只考虑部分对齐方案）
- 词对齐模型开源实现 GIZA++

Thank you!