

汉英双语库词汇对齐研究¹

王斌 刘群 张祥

中国科学院计算技术研究所, 北京 100080

摘要: 对齐的双语语料库能够为许多的自然语言应用提供重要的基础。其中, 对于许多基于双语语料库的应用来说, 双语语料库必须做到词汇级别的对齐。本文考虑在句子对齐基础上的汉英词汇自动对齐。本文依次提出了基于双语词典、基于语义相似、基于位置变形概率的汉英词汇对齐方法, 充分利用现有的有限资源, 提高汉英词汇对齐的正确率和召回率。

关键词: 自然语言处理, 双语语料库, 对齐, 词汇对齐

Word Alignment on Chinese-English Bilingual Corpora

Wang Bin Liu Qun Zhang Xiang

Institute of Computing Technology, Chinese Academy of Sciences, Beijing

Abstract: Aligned bilingual corpora can support many NLP applications. For some applications, the corpora must be aligned at the word level. This paper tries to align words within aligned Chinese-English sentences. It proposes bilingual lexicon based, sense similarity based and location distortion probability based methods orderly. We uses existing limited resources to improve the precision and recall of the Chinese-English word alignment.

Key words: NLP, Bilingual Corpora, Alignment, Word Alignment

1、引言

近年来, 语料库语言学的兴起是计算语言学中的重要事件。语料库以其覆盖面广、语料真实、信息丰富而为计算机自然语言处理提供了强有力的支持。双语库是同时含有两种语言对译信息的语料库, 它可以广泛地用于双语研究的许多领域(如机器翻译、词典编纂等等), 具有很高的研究和实用价值。

目前基于双语库的工作主要是对齐, 即找出双语文本片断之间的对译关系。语料库的对齐单位由大到小包括篇章、段落、句子、短语、单词等。基于双语库的应用一般都要求双语库做到句子级的对齐。在此基础上, 许多应用要求进一步做到词汇级别的对齐, 即在源文和对应的译文中找到词汇级的对译关系。图 1 给出了一个汉英词汇对齐的结果。

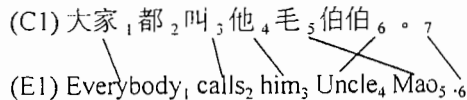


图 1 一个词汇对齐的结果

形式地, 词汇对齐结果的每个匹配对都可以用六元组 $\langle C_i, i, l_C, E_j, j, l_E \rangle$ 来表示², C_i 表示源语言词串, E_j 表示目标语言词串。 i, j 分别表示 C_i, E_j 在各自句子的起始词序号。 l_C, l_E 分

¹ 本课题得到 863-306 资助, 合同号为 863-306-03-06-2

² 本文不考虑不相邻词串组成的匹配单位

别表示词的个数。例如，图 1 中的一个匹配对就可以表示成<他, 4,1,him, 3,1>。显然， C_i 、 E_j 不全为空，当其中之一为空时，我们称该匹配对为空匹配对。

当前词汇对齐主要采用概率统计的方法。这种方法通常首先建立一个词汇对齐模型，然后利用概率统计得到的词汇匹配概率进行词汇对齐。

Brown^[2]为了进行基于统计的英法机器翻译，建立了多个词汇对齐模型，然后利用已经对齐的词汇对词汇间的直译概率、繁殖概率、位置变形概率等模型参数进行训练。Dagan^[3]等人使用改进的 Brown 的模型，首先通过字符串匹配获得部分对齐，然后使用词汇对齐模型中的各项概率参数进行英法词汇的对齐。Gale^[4]使用一种类 x^2 的概率分布，只选择部分相关的英法词汇对进行匹配。Ker^[5]使用一种基于语义类的方法对汉英句子进行词汇对齐。该方法通过大规模语料的训练，来获得汉英词翻译的语义类匹配概率，然后利用这些概率对汉英句子进行词汇对齐。Chang^[1]对经过词性标注的汉英句子进行汉英词之间词性翻译概率的训练，然后利用词性翻译概率进行词的对齐过程。

以上的方法都需要大规模的复杂的训练过程，字符串的匹配显然不易用于汉英词汇对齐，企图完全通过统计的方法来进行词汇对齐也显然没有充分利用现有的资源并且很难同时获得较高的召回率和正确率。显然地，双语词典含有丰富的词汇互译信息，因此它是进行词汇级别对齐的有效的可靠的工具。汉英词汇对齐的第一种方法是基于双语词典的对齐方法。

2、基于双语词典的词汇对齐

形式地，假设汉语句子 $C=C_1C_2\dots C_m$ 和英语句子 $E=E_1E_2\dots E_n$ 互为翻译，每个 C_i 或者 E_j 为句中的词或者标点符号。基于双语词典的汉英词汇对齐(BiDictAlign)算法如下：

- (1) 初始化：集合 $Set_C = \{ \langle C_1, 1 \rangle, \langle C_2, 2 \rangle, \dots, \langle C_m, m \rangle \}$ 。集合 $Set_E = \{ \langle E_1, 1 \rangle, \langle E_2, 2 \rangle, \dots, \langle E_n, n \rangle \}$ ，集合 $Set_A = \phi$ ，剔除集合 Set_C 中的助词和语气词{的, 得, 地, 了, ...}；
- (2) 对于任意 $\langle C_i, i \rangle \in Set_C$ ， $1 \leq i \leq m$ ， $\langle E_j, j \rangle \in Set_E$ ， $1 \leq j \leq n$ ，如果 $CEDictSim(C_i, E_j) > h_1$ (h_1 为给定的阈值)，则将六元组 $\langle C_i, i, 1, E_j, j, 1 \rangle$ 加入集合 Set_A ；
- (3) 重复(2)，直到 Set_A 不再变化为止；
- (4) $Set_C = Set_C - \{ \langle C_i, i \rangle \mid \text{存在 } E_j, j, \text{ 使得 } \langle C_i, i, 1, E_j, j, 1 \rangle \in Set_A \}$ ， $Set_E = Set_E - \{ \langle E_j, j \rangle \mid \text{存在 } C_i, i, \text{ 使得 } \langle C_i, i, 1, E_j, j, 1 \rangle \in Set_A \}$ ；
- (5) 输出 Set_A 的匹配词对， Set_C 和 Set_E 中的词汇以空匹配输出。

算法结束时， Set_A 中存放的是匹配的词汇对， Set_C 和 Set_E 中分别是未找到相应匹配的汉语词汇和英语词汇。剔除部分相互冲突的匹配对，可以得到高可靠性的匹配集合：

$CONN = \{ \langle C_i, i, 1, E_j, j, 1 \rangle \mid \text{每个 } C_i, E_j \text{ 在所有的输出匹配对中仅出现一次} \}$

算法中， $CEDictSim(c,e)$ 表示汉语词 c 和 e 的匹配相似度， $EESim(x,y)$ 表示两个英语词 x 和 y 的匹配相似度，它们的定义如下：

$$CEDictSim(c, e) = \begin{cases} 1 & c = e \\ \max_{x \in Dict(c)} EESim(x, e) & c \neq e \end{cases}$$

$$EESim(x, y) = \begin{cases} 0.9 & \text{if } x = y \\ 0.8 & \text{if } x \in DelFix(y) \\ 0.7 & \text{if } x \text{与} y \text{的前5个字节相等} \\ 0 & \text{else} \end{cases}$$

其中， $Dict(c)$ 函数表示在双语词典中查到 c 的译词的集合。 $DelFix$ 为英语单词形态分析

函数，它以某个可能已经变形的英语单词为输入，输出其可能的原形集合。

实验中发现，使用 BiDictAlign 虽然可以得到比较可靠的非空匹配词对，但由于双语词典²的局限性，使用 BiDictAlign 达到的正确率和召回率都有限。鉴于同义词替代在汉英翻译中的普遍性，本文提出了一种基于语义相似度的词汇对齐方法。

3、基于语义相似度的词汇对齐

在叙述基于语义相似度的词汇对齐之前，我们首先引入基于《同义词词林》的汉语语义相似度计算方法。作为词义赋值依据的《同义词词林》是现代汉语比较常用的一部类义词典。它所收词语全部按词义分类编排，描述了一个由上到下，由宽泛概念到具体词义的语义分类体系，《同义词词林》的整个语义分类体系可以用树形图表示如下：

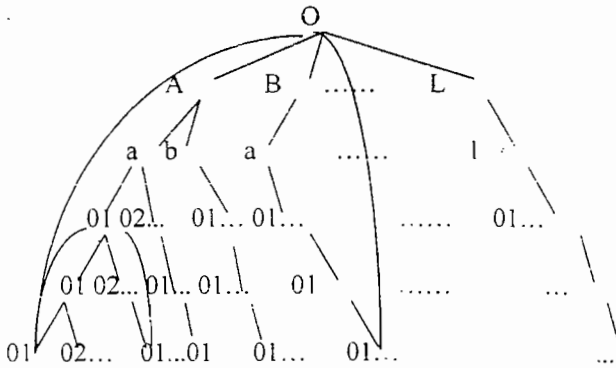


图 2 语义分类体系树形图

图 2 中从根结点 O 到每个叶结点的最短路径代表一个词义，其词义代码为根结点 O 到该叶结点的最短路径上所有代码依次组成的字符串(不包括 O)。于是两词义 S_1 与 S_2 之间的语义距离 $SenseDist(S_1, S_2)$ 可以定义为在上图中从结点 S_1 到结点 S_2 的最短路径的长度。

比如： $SenseDist(Aa010101.Aa010201)=4$ ， $SenseDist(Aa010101, Ba010101)=10$ 。

显然，按照《同义词词林》的分类原则， $SenseDist(S_1, S_2)$ 越小， S_1 与 S_2 就更可能同处于小的类中，因此可以认为它们在语义上越相似。于是可以定义词义 S_1 与 S_2 的语义相似度为：

$$SenseSim(S_1, S_2) = \begin{cases} 1/SenseDist(S_1, S_2) & S_1 \neq S_2 \\ 1 & S_1 = S_2 \end{cases}$$

在此基础上，定义两个汉语词 C_i, C_j 的语义相似度为：

$$CCClassSim(C_i, C_j) = \max_{\substack{S_m \in Senseof(C_i) \\ S_n \in Senseof(C_j)}} SenseSim(S_m, S_n)$$

其中， $Senseof(S)$ 函数返回词语 S 的词义代码集合。

定义英语词 E_j 和汉语词 C_i 的语义相似度为：

$$ECClassSim(E_j, C_i) = \max_{E_j \in DelFix(E_j)} \max_{C_k \in Dict(E_j)} CCClassSim(C_k, C_i)$$

形式地，基于语义相似度的词汇对齐(ClassAlign)算法如下：

- (1) 初始化：集合 $Set_C = \{ \langle C_1, 1 \rangle, \langle C_2, 2 \rangle, \dots, \langle C_m, m \rangle \}$ ，集合 $Set_E = \{ \langle E_1, 1 \rangle, \langle E_2, 2 \rangle, \dots, \langle E_n, n \rangle \}$ ，集合 $Set_A = \phi$ ，剔除集合 Set_C 中的助词和语气词{的, 得, 地, 了, ...}；
- (2) 如果集合 Set_C ， Set_E 均不为空，则对任意 $E_j \in Set_E$ ，任意 $C_i \in Set_C$ ， $1 \leq i \leq m$ ， $1 \leq j \leq n$ ，如果 $ECClassSim(E_j, C_i) > h_2$ (给定的阈值)，将六元组 $\langle C_i, i, 1, E_j, j, 1 \rangle$ 加入集合 Set_A ；
- (3) 重复(2)，直到 Set_A 不再变化为止；
- (4) $Set_C = Set_C - \{ \langle C_i, i \rangle | \text{存在 } E_j, j, \text{ 使得 } \langle C_i, i, 1, E_j, j, 1 \rangle \in Set_A \}$ ， $Set_E = Set_E - \{ \langle E_j, j \rangle | \text{存在 } C_i, i, \text{ 使得 } \langle C_i, i, 1, E_j, j, 1 \rangle \in Set_A \}$ ；
- (5) 输出 Set_A ，同时将 Set_C 和 Set_E 中的词以空匹配输出。

由于 ClassAlign 利用同义词进行词汇对齐，因此可以大大弥补双语词典的覆盖面的不足，找到更多的正确的非空匹配。又由于其算法相对 BiDictAlign 而言复杂度更高，所以它常常用作 BiDictAlign 算法的一个补充。下图中给出了 ClassAlign 对 BiDictAlign 的补充匹配的例子。

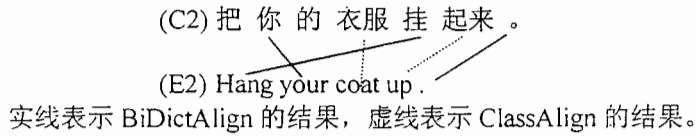


图 2 BiDictAlign&ClassAlign

4、基于位置变形距离的词汇对齐方法

我们发现，在以上 BiDictAlign 或者 ClassAlign 的对齐结果中，都可能具有相互冲突的词匹配结果，也就是说，有多个不同的词匹配对，它们却包含有同一个汉语词或者英语词。

形式地，在词汇对齐的结果中，如果存在两个不同的词匹配 $M_1 = \langle C_1, i_1, l_1, E_1, j_1, l_2 \rangle$ ， $M_2 = \langle C_2, i_2, l_3, E_2, j_2, l_4 \rangle$ ，它们满足以下条件之一：

- (i) $i_1 \neq 0, i_2 \neq 0, i_1 \leq i_2 \leq i_1 + l_1$ ；
- (ii) $i_1 \neq 0, i_2 \neq 0, i_2 \leq i_1 \leq i_2 + l_3$ ；
- (iii) $j_1 \neq 0, j_2 \neq 0, j_1 \leq j_2 \leq j_1 + l_2$ ；
- (iv) $j_1 \neq 0, j_2 \neq 0, j_2 \leq j_1 \leq j_2 + l_4$ ；

则称 M_1 和 M_2 相互冲突。对齐结果中所有的相互冲突词匹配组成冲突匹配集合。

我们对匹配词对的位置进行了考察。在考察中我们发现，如果把每个词对在汉英句子中的位置分别看成平面坐标系中的点的话，那么即使在发生短语移位的情况下，某个词对与其相邻词对(尤其当它们处于同一最小短语结构中)的左“斜率”或者右“斜率”将趋近于 1。于是可以通过对冲突匹配集合中每个匹配位置左右斜率的计算来获得可能性最大的匹配，剔除与之冲突的其他匹配。形式地，源文位置 i 到目标文本位置 j 的位置变形距离可以定义为：

$$LocDist(i, j) = \min(|Slope_L - 1|, |Slope_R - 1|)$$

$$\text{其中 } Slope_L = (j - j_L) / (i - i_L), \quad Slope_R = (j - j_R) / (i - i_R)$$

$$(i_L, j_L) = \underset{(C_i, i, l, E_j, j, l) \in CONN_{e_i}}{\operatorname{argmax}} i', \quad (i_R, j_R) = \underset{(C_i, i, l, E_j, j, l) \in CONN_{s_i}}{\operatorname{argmax}} i'$$

(i_L, j_L) 表示左边距离 i 最近的可靠联接的位置对

³ 由于英语单词之间比较的容易性，在 BiDictAlign 算法中只使用了汉英词典

(i_R, j_R) 表示右边距离 i 最近的可靠联接的位置对

$Slope_L$ 表示 (i, j) 相对于 (i_L, j_L) 的匹配值(类似于左斜率)

$Slope_R$ 表示 (i, j) 相对于 (i_R, j_R) 的匹配值(类似于右斜率)

$CONN_{<i}$ 表示 CONN 中汉语位置值小于 i 的六元组集合

$CONN_{>i}$ 表示 CONN 中汉语位置值大于 i 的六元组集合

$LocDist(i, j)$ 越小, 可以认为位置 (i, j) 匹配的可能性越大, 从而实现冲突的消解。

使用位置变形概率的词汇对齐过程(LocAlign)为:

- (1) 使用 BiDictAlign(ClassAlign)进行词汇对齐, 计算 Set_A 、CONN(暂不计算 Set_C 和 Set_E);
- (2) 计算冲突匹配集合 $Set_D = Set_A - CONN$;
- (3) 对于每个 $\langle C_i, i, a, E_j, j, b \rangle \in Set_D$, $1 \leq i \leq m$, $1 \leq j \leq n$, 计算 $LocDist(i, j)$;
- (4) 按照 $LocDist$ 从小到大的顺序依次将 Set_D 中元素加入一个开始为空的集合 Set_U , 并且保证加入的元素与集合 Set_U 中已有的元素不冲突;
- (5) $Set_A = CONN + Set_U$, 重新计算集合 Set_C 、 Set_E 及 CONN,;
- (6) 输出 Set_A 、 Set_C 、 Set_E 。

以下是一个例子来说明 $LocDist$ 的计算以及 LocAlign 的过程。句子对

(C3) 从₁他₂参加₃革命₄以来₅, 已经₆有₇30₉年₁₀了₁₁。12

(E3) It₁ is₂ 30₃ years₄ since₅ he₆ joined₇ the₈ revolution₉ .10

经过 BiDictAlign 的结果为:

$Set_A = \{ \langle \text{从}, 1, 1, \text{joined}, 7, 1 \rangle, \langle \text{他}, 2, 1, \text{he}, 6, 1 \rangle, \langle \text{参加}, 3, 1, \text{joined}, 7, 1 \rangle, \langle \text{革命}, 4, 1, \text{revolution}, 9, 1 \rangle, \langle 30, 9, 1, 30, 3, 1 \rangle, \langle \text{年}, 10, 1, \text{years}, 4, 1 \rangle \}$

$CONN = \{ \langle \text{他}, 2, 1, \text{he}, 6, 1 \rangle, \langle \text{革命}, 4, 1, \text{revolution}, 9, 1 \rangle, \langle 30, 9, 1, 30, 3, 1 \rangle, \langle \text{年}, 10, 1, \text{years}, 4, 1 \rangle \}$

冲突匹配集合为:

$Set_D = Set_A - CONN = \{ \langle \text{从}, 1, 1, \text{joined}, 7, 1 \rangle, \langle \text{参加}, 3, 1, \text{joined}, 7, 1 \rangle \}$

为了剔除错误匹配, 以下对每个冲突匹配集合中的匹配, 计算它们的位置变形距离。

- (1) 求 $LocDist(1, 7)$

因为 (i_L, j_L) 不存在, $(i_R, j_R) = (2, 6)$, 所以 $LocDist(1, 7) = |(7-6)/(1-2)-1| = 2$

- (2) 求 $LocDist(3, 7)$

因为 $(i_L, j_L) = (2, 6)$, $(i_R, j_R) = (4, 9)$, 所以 $LocDist(3, 7) = \min(|(7-6)/(3-2)-1|, |(7-9)/(3-4)-1|) = 0$

(3) 因为 $LocDist(1, 7) < LocDist(3, 7)$, 所以可以认为“参加”与“joined”匹配的概率大于“从”与“joined”匹配的概率。于是, 通过计算, 可以将噪声 $\langle \text{从}, 1, 1, \text{joined}, 7, 1 \rangle$ 剔除。

基于位置变形距离的方法常常用于对前面两种方法的补充, 它能够消解前述方法中的匹配冲突, 从而达到剔除噪声获得正确匹配的目的。

5、实验结果及分析

采用上述方法, 我们对一组用于汉英机器翻译系统测试的约 500 对汉英对照句子进行了测试。实验中采用了一部约有 50000 多个汉语常用词的汉英电子词典, 一部约有 13000 个英语词条的英汉电子词典以及共有 61125 个汉语词的《同义词词林》电子词典。测试集中汉语句子的长度范围为 2~20 个词, 平均长度为 7.96 个词, 英语句子的长度范围为 2~19 个词, 平均长度为 8.92 个词。为了获得定量结果, 测试语料的例句已经预先进行了手工的词汇对齐, 并将手工对齐的结果作为测试的参照对齐。在 BiDictAlign 中, 本实验也将汉英句子

相邻的 3 个和 2 个词作为匹配单位进行了匹配，测试的结果如下表所示：

使用方法	召回率	正确率
BiDictAlign	63.2%	70.1%
BiDictAlign&LocAlign	63.2%	72.6%
BiDictAlign&ClassAlign	82.8%	89.7%
BiDictAlign&ClassAlign&LocDist	82.8%	92.4%

从上表可以看出，仅使用 BiDictAlign 达到的召回率和正确率都有限，主要原因是找出的匹配对中具有相当数量的空匹配对。在此基础上使用 LocAlign，虽然不会找到更多的匹配对，但可以剔除部分冲突的匹配对，从而提高正确率。结合使用 BiDictAlign 和 ClassAlign，ClassAlign 可以在找到更多非空匹配对的同时减少空匹配对的数量，因此召回率和正确率会得到较大程度的提高。该实验结果也说明了本文提出的三种方法在词汇对齐中的不同作用和不同使用方法。

6、结束语

本文提出了汉英词汇对齐的三种方法。其中基于双语词典的对齐方法可以获得可靠的非空匹配对，因此它常常作为初始的词汇对齐方法。但由于双语词典的覆盖面有限，基于双语词典的方法得到的对齐正确率和召回率也有限。而基于语义相似度的方法大大弥补了双语词典方法的不足，可以获得更多的正确非空匹配。基于位置变形距离的方法常常用于对前面两种方法的补充，它能消解前述方法中的匹配冲突，从而达到剔除噪声获得正确匹配的目的。

在以上方法的基础上，还可以采用规则匹配，共现概率等方法进一步提高汉英词汇对齐的召回率和正确率。相对于以往的词汇对齐方法，本文方法的优点在于，充分利用已有的有限资源，不需要大数据量的训练即可达到较高的召回率和正确率。

当然，由于汉英语言在语言习惯、表达方式等方面的固有差异（比如一个汉语词可能翻译成一个英语短语甚至子句），汉英词汇对齐的进一步工作存在相当的困难，是否将词汇对齐和亚句子对齐相融合、是否进行详尽的句法分析甚至语义分析、是否引入有效的预测机制都是下一步需要研究的问题。

参考文献

- [1] Chang, J. S. and Chen, M. H. C., "Using Partial Aligned Parallel Text and Part-of-speech Information in Word Alignment", Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA'94), pages 16-23, Columbia, MD., 1994
- [2] Brown, P. F., Cocke, J., Della Pietra, S. A., and Della Pietra, V. J., "A Statistical Approach To Machine Translation", Computational Linguistic, 16(2): 79-85, 1990
- [3] Dagan, I., Church, K. W., and Gale, W. A., "Robust Bilingual Word Alignment for Machine Aided Translation", Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pages 1-8, Columbus, OH., 1993
- [4] Gale, W. A., and Church, K. W., "Identifying Word Correspondences in Parallel Texts", Proceedings of the Fourth DARPA Speech and Natural Language Workshop, pages 152-157, Pacific Grove, CA., 1991
- [5] Ker, S. J., and Chang, J. S., "A Class-Based Approach to Word Alignment". Computational Linguistic, 23(2): 313-343, 1997