基于组合线索和核心扩展方阵匹配的中日句对齐

胡海鹏[1] 闫永明[1] 吴宏林[1] 张俐[1] 刘绍明[2]
[1]东北大学自然语言处理实验室 沈阳 110004
[2]日本富士施乐公司 日本 神奈川
Email:huhp@ics.neu.edu.cn

摘要：该文提出了一种基于组合线索和核心扩展方阵匹配的中日句对齐算法。该方法利用字典、字形、长度和特殊字符相结合的组合线索来计算句子相似度，并利用核心扩展方阵匹配实现中日句对齐。该方法在一定程度上解决了传统的基于长度的方法的错误蔓延问题，而且充分挖掘了中日双语之间潜在的联系，增强了相似度计算的可信度。实验表明，在中日句对齐任务中该方法取得了比较满意的结果。

关键词：机器翻译 句对齐 组合线索 核心扩展方阵匹配

Sentence Alignment between Chinese and Japanese Based on Combined Clues and Kernel Extensional Matrix Matching

HU Haipeng[1], YAN Yongming[1], WU Honglin[1], ZHANG Li[1], LIU Shaoming[2]
[1] Natural Language Processing Laboratory, Northeastern University, Shenyang, 110004
[2] FujiXerox Co., Ltd. Kanagawa, Japan
Email:huhp@ics.neu.edu.cn

Abstract: This paper proposes a Chinese-Japanese sentence alignment algorithm based on combined clues and kernel extensional matrix matching. In this approach, the similarity of sentences is calculated by the combined clues, such as lexicon, morphology, length and special symbols, and the sentences are aligned by the kernel extensional matrix matching. This approach solves some wrong spread in previous length-based method to a certain extent, and adequately utilizes the potential relationship between Chinese and Japanese, and enhances the credibility of similarity caculation. Experimental results illustrate that this approach has better performance in Chinese-Japanese sentence alignment task.

Keywords: Machine Translation; Sentence Alignment; Combined Clues; Kernel Extensional Matrix Matching

1 引言

自然语言处理需要大量的不同对齐层次上的双语资源。随着互联网的迅速发展，人们可以很方便的获得篇章级对齐[1]的原始双语资源，但其它级别对齐的双语资源(如：段落级[2]，句子级[3]，短语级[4]和词汇级[5] 等)相对而言就要少得多，故如何把篇章级对齐的双语资源进行更细层次的对齐就变得极其重要。

句对齐语料库建设的直接应用目标是为基于实例的翻译引擎提供翻译实例，也为挖掘各种机器翻译知识提供一个基础资源[6]。因此，句对齐的准确性将会大大影响到机器翻译的质量。本文探讨了如何从篇章级对齐的双语文本中自动获取句对齐资源。

现有的句对齐方法主要有以下几类：(1)基于长度的方法[3, 7]：主要是与两种语言的长度有关，依靠双语参数的设计[8]，无需语言学知识，完全独立于语种。(2)基于词汇的方法[9-10]：通常利用双语词典和词汇信息来对齐句子，寻找两种语言间的同源词和关键词。(3)混合方法[11-12]：将词汇和长度方法结合起来，利用二者的优越性，既提

高了鲁棒性，又降低了计算复杂度[8]。

目前常用的句对齐方法对源语与目标语之间的联系挖掘不够充分，且对齐结果存在错误蔓延。针对这些问题，在东北大学与日本富士施乐合作科研项目过程中，共同提出了相应的改进方法，该方法采用组合线索和核心扩展方阵匹配实现对齐，其中引入组合线索计算中日文句子之间的相似度，以充分挖掘中日双语之间的潜在联系；引入核心扩展方阵进行句子匹配，以避免错误蔓延。

2 句对齐模型

在此模型中，采用基于组合线索的双语句子相似度计算，用以建立句对齐相似度矩阵；利用句对齐相似度矩阵建立句对齐选择矩阵；在句对齐选择矩阵中利用核心扩展方阵实现句子之间的匹配，获取句对齐结果。图 1 为句对齐模型

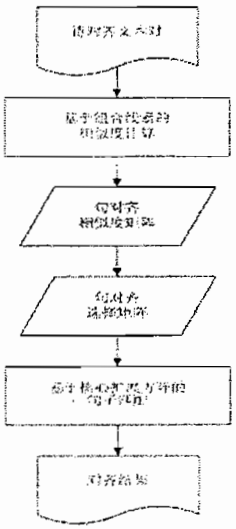

图 1 句对齐模型

2.1 基于组合线索的句子相似度计算

对于需要进行句对齐的互为译文的中、日文文本 C 和 J，假设 C 由 m 个句子构成，即 $C=CS_1CS_2...CSm$； J 由 n 个句子构成，即 $J=JS_1JS_2...JSn$。

基于组合线索的句子相似度计算利用字典相似度(SimDict)、字形相似度(SimMorph)和句长相似度(SimLength)计算中、日文句的相似度，将分别计算的结果乘以相应的权重加和，所得的结果再与特殊字符相似度(SValue)相加，作为中日句对间的相似度。

中文句子 CSm 与日文句子 JSn 的相似度计算公式如下：

$$\begin{aligned}\mathrm{Weight}(CS_m, JS_n) = {} & \alpha \times \mathrm{SimDict}(CS_m, JS_n) + \beta \times \mathrm{SimMorph}(CS_m, JS_n) + \\ & \gamma \times \mathrm{SimLength}(CS_m, JS_n) - \mathrm{SValue}(CS_m, JS_n)\end{aligned} \tag{1}$$

2.1.1 基于双语字典的相似度(SimDict)计算

在基于双语字典的相似度计算中，句子用单词的集合来表示。中、日文句子分别表示成集合 $CS_h=\{c_1,c_2,\cdots,c_m\}$ 和 $JS_k=\{j_1,j_2,\cdots,j_n\}$。则 SimDict 计算方法如下：

$$\mathrm{SimDict}(CS_h, JS_k) = \frac{|\mathrm{TransSetC}| + |\mathrm{TransSetJ}|}{|CS_h| + |JS_k|} \tag{2}$$

上式中，$|CS_h|$和$|JS_k|$分别为中、日文句中实词个数，|TransSetC|是在日文句中有译词的中文句中的实词个数，|TransSetJ|是在中文句中有译词的日文句中的实词个数：

$$TransSetC=\{c_p \mid c_p \in CS_h \wedge \exists j_q : j_q \in JS_k : SimDict(c_p, j_q)=1\} \tag{3}$$

$$TransSetJ=\{j_q \mid j_q \in JS_k \wedge \exists c_p : c_p \in CS_h : SimDict(c_p, j_q)=1\} \tag{4}$$

其中 $\mathrm{SimDict}(c_p,j_q)$为利用字典计算 c_p 和 j_q 的相似度，当 c_p 与 j_q 在双语字典中互译时，$\mathrm{SimDict}(c_p,j_q)=1$，否则为 0。

2.1.2 基于字形的相似度(SimMorph)计算

在基于字形的相似度计算中，句子用字符集合表示。中、日文句子分别表示成集

令 $CS_m=\{cc_1,cc_2,\cdots,cc_r\}$ 和 $JS_n=\{jc_1,jc_2,\cdots,jc_s\}$。则 SimMorph 计算方法如下：

$$SimMorph(CS_m,JS_n)=\frac{|MorphSetC|+|MorphSetJ|}{|CS_m|+|JS_n|} \tag{5}$$

上式中，|CSm|和|JSn|分别为中、日文句中字符的个数，|MorphSetC|是 CSm 中的中文汉字与 JSn 中日文汉字判定为相同的个数，|MorphSetJ|是 JSn 中的日文汉字与 CSm 中中文汉字判定为相同的个数，MorphSetC 和 MorphSetJ 的计算方法如下：

$$MorpSetC=\{cc_p \mid cc_p\in CS_m\wedge\exists jc_q : jc_q\in JS_n : IsSimMorph(cc_p,jc_q)=1\} \tag{6}$$

$$MorpSetJ=\{jc_q \mid jc_q\in JS_n\wedge\exists cc_p : cc_p\in CS_m : IsSimMorph(cc_p,jc_q)=1\} \tag{7}$$

其中 IsSimMorph(cc_p,jc_q)是判断 ccp 和 jcq 是否为字形相似的函数，当 cc_p 和 jc_q 相同或经简繁转换后的 ct_p 和 jc_q 相同时，IsSimMorph(cc_p,jc_q)=1，否则为 0。

2.1.3 基于句子长度的相似度(SimLength)计算

基于句子长度计算相似度的方法很多，本文采用文献[13]中的方法，计算如下：

$$SimLength(CS_h,JS_k)=\arg\max_{A}\prod_{P\in A}Pr(S(l_1,l_2)\mid match(P)) \tag{8}$$

2.1.4 基于特殊字符的相似度(SValue)计算

在本文的句对齐模型中，将使用四种特殊的字符来计算中、日文句的特殊字符相似度，分别为：数字、英文、引号和括号特殊字符。将 SValue 在公式(1)中的最大值设定为 0.2，其中每种特殊字符权重相同且为 0.05，即对于数字、英文、引号和括号特殊字符，如在中文句 CS_h 和日文句 JS_k 中有一种相同的特殊字符，则 SValue 的值增加 0.05，最大增加到 0.2，作为 SimDict、SimMorph 和 SimLength 计算对齐相似度后的补充。

2.2 基于核心扩展方阵的句子匹配

利用 2.1 节介绍的基于组合线索的方法，计算需要对齐的中、日文句子之间的相似度，根据相似度构造句对齐相似度矩阵，并根据句对齐相似度矩阵构造句对齐选择矩阵，然后利用核心扩展方阵在句对齐选择矩阵上进行匹配得到最终句对齐结果。

2.2.1 构造对齐相似度矩阵和对齐选择矩阵

根据中、日文句个数 m 和 n 构造 $m\times n$ 的二维句对齐相似度矩阵 SimMatrix，矩阵中元素的值为该元素所在行对应的中文句子与所在列对应的日文句子的相似度。

$$SimMatrix[h][k]=Sim(CS_h,JS_k) \tag{9}$$

然后构造 $m\times n$ 的对齐选择矩阵 SelMatrix，矩阵中的元素记录了每一行和每一列对齐相似度大小的排序信息。

2.2.2 句子匹配

句子匹配是根据 SelMatrix 中的信息来挑选作为句对齐结果的对齐。句子匹配主要分为三个部分，分别是：挑选有“1/1”的行中的对齐、挑选无“1/1”的行中的对齐和空对齐处理。

①挑选有“1/1”的行中的对齐

步骤 1：根据挑选行中“1/1”所在位置建立核心扩展方阵——对齐选择矩阵中以 1/1 位置为中心，大小为 3×3 的矩阵；

步骤 2：计算是否有多对齐可能，如有将多对齐情况加入候选集，否则转到步骤 5；

步骤 3：对于候选集中的多对齐，判断其对齐类型；

步骤 4：根据步骤 3 的判断结果，修正多对齐；

步骤 5：将已确定的对齐关系加入已完成对齐队列，并记录句子的对齐状态；

步骤 6：判断是否有未处理的“1/1”位置，如没有返回步骤 1，否则结束挑选。

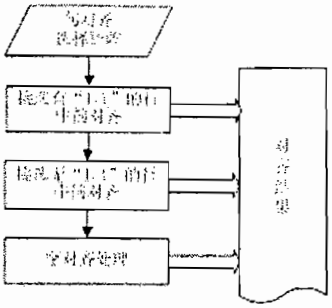

图 2 句子匹配流程

②挑选无“1/1”的行中的对齐

挑选无“1/1”的行中的对齐模块与上一模块类似，唯一不同是上一模块是选择行中“1/1”位置，以“1/1”位置为中心考虑对齐；而此模块是找到没有“1/1”位置且未对齐的行，在此行中寻找对齐相似度排行加和最小的位置，并判断此位置是否会产生交叉对齐，如不会则以此位置为中心考虑对齐；否则保留此行未对齐状态，待空对齐处理模块进行处理。

③空对齐处理

步骤 1：挑选未对齐的行，如没有则转到步骤 5；

步骤 2：选择未对齐行前后各一行中的“1/1”锚点，分别计算未对齐行与前一行和与后一行形成多对齐的相似度；

步骤 3：分别将多对齐相似度与原锚点相似度进行比较；

步骤 4：根据步骤 3 的比较结果修正对齐结果；

步骤 5：保存修正后的结果，并记录对齐状态；

步骤 6：判断是否还有未对齐的行，如没有则结束处理。

3 实验和分析

为了验证本文提出的中日句对齐模型，我们构造了两个句对齐系统(SenAlignSystem 简称 SAS)，其中 SAS_A 利用组合线索和动态规划实现对齐，SAS_B 利用组合线索和核心扩展方阵匹配实现对齐。在随机抽取的 8 篇篇章级对齐的文章(总计 522 句中文和 548 句日文，语料由富士施乐提供，领域为 IT 领域)上进行了测试，并与 Gale 的句对齐系统[3]进行比较，以下把 Gale 的句对齐系统作为 baseline 系统。实验设计如表 1 所示：

表 1 实验设计

	相似度计算方法	对齐匹配
SAS_A	组合线索	动态规划
SAS_B	组合线索	核心扩展方阵
Baseline	长度	动态规划

为了单独评价本文提出的基于组合线索的相似度计算方法，我们比较了 SAS_A 与 baseline。经过人工评测得到实验结果如表 2 所示：

表 2 SAS_A 与 Baseline 比较

	准确率	召回率	F1 值
SAS_A	**92.37%**	**92.37%**	**92.37%**
Baseline	81.73%	81.41%	81.57%

通过表 2，我们发现基于组合线索的相似度计算方法比仅仅基于长度的相似度计算方法在准确率和召回率上都有较大的提高。这是由于基于长度的方法只利用统计学的基本原理，没有充分挖掘中日双语之间潜在的联系。而 SAS_A 采用本文提出的基于字典、字形、长度和特殊字符的组合线索方法，较为充分的挖掘了中日双语之间的联系。实验结果充分验证了基于组合线索相似度计算方法的有效性。下面给出测试集中两个系统分别对齐结果的一个实例。

Baseline 对齐结果：　　(错误对齐)

C：还可以省略对原来硬件（PC AT 总线等）的检测等，实现高速启动。

J：EFI のファームウエアはプロセサのネイティブ・モードで機能するので、起動自体も高速になる。

SenAlignSystemA 对齐结果：　　(正确对齐)

C:还可以省略对原来硬件（PC AT 总线等）的检测等，实现高速启动。

J:古いハードウエア（PC AT バスなど）の検査などを省いて起動を高速にすることもできる。

上例中由于 Baseline 只利用长度来计算相似度，忽略了其它联系，导致对齐错误；而 SAS_A 由于采用本文提出的组合线索相似度计算方法，较全面的挖掘了中日双语之间的联系，因此得到了正确的对齐结果。

为了单独评价本文提出的核心扩展方阵匹配方法，我们比较了 SAS_A 与 SAS_B。经过人工评测得到实验结果如表 3 所示：

表 3 SAS_A 与 SAS_B 比较

	准确率	召回率	F1 值
SAS_A	92.37%	92.37%	92.37%
SAS_B	**97.85%**	**97.85%**	**97.85%**

通过表3,我们发现基于核心扩展方阵匹配方法比基于动态规划方法在准确率和召回率上都有较大的提高。这是由于SAS_B存在一定程度的错误蔓延现象；而基于核心扩展方阵匹配的核心是句对齐选择矩阵中值为“1/1”的元素。如果遇到错误对齐，可以保证在下一个“1/1”元素前结束错误蔓延。实验结果充分验证了基于核心扩展方阵匹配的有效性。下面给出测试集中快速终止蔓延的实例。

[Type:1-2; Weight=0.382736] (错误对齐开始)

Ci:因为在中国城市，新建的大规模集中公寓住宅占绝大多数，所以便于通信运营商利用光纤向配备 LAN 的大规模集中公寓住宅提供以太网接入服务。

Ji:中国の都市部では，新しい大規模だ集合アパートの数が圧倒的に多い，通信事業者は，光ファイバを通じてLAN を備える大規模だ集合アパートに Ethernet 接続を提供するている。

Ji+1:また，中国の Ethernet サービス・プロバイダは，欧米のベンダーの製品を多い利用するている。

[Type:1-1; Weight=0.240715] (错误对齐蔓延)

Ci+1:中国的以太网服务商多数采用欧美产品，ISP 的伙伴包括美国思科系统公司、加拿大的北电网络公司、美国 Juniper Networks 以及美国的 IBM 等。

Ji+2:サービス・プロバイダのパートナとするて，米 Cisco Systems，カナダの Nortel Networks，米 Juniper Networks，米 IBM だどの名前が挙げるられるた。

[Type:1-2; Weight=0.354285] (错误对齐蔓延快速终止)

Ci+2:该调查是在 2004 年 10 月份实施的，通过从中国的主要通信 ISP 那里直接获取数据整理而成。

Ji+3:同調査は，2004 年 10 月に実施するれるたもの。

Ji+4:中国の主要通信事業サービス・プロバイダを通じて直接入手するたデータがまとめるられるた。

上例中，C_i 与 J_{i+1} 对齐错误，导致 C_{i+1} 与 J_{i+2} 错误对齐(蔓延)，但是由于本文采取的匹配方法，这个错误蔓延在下一个对齐快速终止，得到 C_{i+2} 与 J_{i+3}，J_{i+4} 的正确对齐。

为了综合评价本文提出的组合线索和核心扩展方阵匹配改进方法，我们比较了SAS_B 与 Baseline。经过人工评测得到实验结果如表 4 所示：

表 4 SAS_B 与 Baseline 比较

	准确率	召回率	F1 值
SAS_B	**97.85%**	**97.85%**	**97.85%**
Baseline	81.73%	81.41%	81.57%

通过表 4，我们发现在中日句对齐任务中使用本文提出的组合线索和核心扩展方阵匹配方法构造的 SAS_B 比 Gale 系统在 F1 值上提高了近 16.3%，充分验证了本文提出的组合线索和核心扩展方阵匹配的有效性。

4 结论

本文提出一种基于组合线索和核心扩展方阵匹配的句对齐模型：利用字典、字形、长度和特殊字符组合线索计算双语句子间的相似度，用以建立句对齐相似度矩阵；利用相似度矩阵建立句对齐选择矩阵；利用核心扩展方阵在句对齐选择矩阵上进行句子之间的匹配，获取句对齐结果。其中引入组合线索充分挖掘中日双语之间的潜在联系；引入核心扩展方阵以避免错误蔓延。实验表明，在中日句对齐任务中使用本文方法的句对齐系统比 Gale 系统在 F1 值上提高近 16.3%，充分验证了该方法的有效性。

参考文献

[1] Xu D. Tian C L. Aligning and matching of English-Chinese bilingual texts of CNS news[J].Machine Translation, 1994,14(1):1-33.

[2] Wang B, Liu Q, Zhang X. Automatic Chinese-English paragraph segmentation and alignment[J]. Joural of Softwares,2000,11(11):1547-1553.

[3] Gale, Church. A Program for Aligning Sentences in Bilingual Corpora[A] the proceeding of Annual meeting of ACL-29[C],Berkeley, CA., 1991.177-184.

[4] Imamura K. A hierarchical phrase alignment from English and Japanese bilingual text[A]. In: Proc of CICLing2001, Mexico City, Mexico: Springer, 206-207.

[5] Ker S J, Chang J S. A class-based approach to word alignment[J]. Computational Linguistics, 1997,23(2):313-344.

[6] 吴宏林. 面向机器翻译的汉日文本对齐研究[D]，2008,沈阳：东北大学.

[7] P.F. Brown, J. C.Lai&P.L. Mercer. Aligning Sentences in Parallel Corpora[A], In : the proceeding of Annual meeting of ACL-29[C], 1991.169-176.

[8] 张艳等. 基于长度的扩展方法的汉英句子对齐[J].中文信息学报.2005.Vol.19 No.5:31-36

[9] M.Kay & K. Roescheisen. Text-Translation Alignment[J],Computation Lingusitics 1993,19(1),121-142.

[10] S.F.Chen. Aligning Sentences in Bilingual Corpora Using Lexical Information[A], In :the proceeding of Annual meeting of ACL-31[C],1993

[11] Thomas C. Chuang & Jason S.Chang. Adaptive Sentence Alignment based on Length and Lexical Information [A]. In: the proceeding of ACL-40[C], 2002. 91-92.

[12] 刘听等. 基于自动抽取词汇信息的双语句对齐[J]. 计算机学报.1998, 21(8):51-158.

[13] 吕学强等. 基于统计的汉英句子对齐研究[J],小型微型计算机系统，2004,25(06): 990-992.