基于多特征融合的先秦典籍汉英句子对齐研究*

梁继文[1]　江　川[2]　王东波[2,3]

[1]（南京大学信息管理学院 南京 210023）

[2]（南京农业大学信息科学技术学院 南京 210095）

[3]（鲁汶大学比利时政府研发监测中心（ECOOM） 鲁汶 B-3000）

摘要：【目的】实现先秦典籍古文-英文双语句子自动对齐，为构建典籍双语句级平行语料库、跨语言检索提供支持。【方法】将典籍汉英句子自动对齐问题视为候选句对分类问题，根据实验语料特点，结合已有研究选取对齐句对特征，基于“整体分类”与“序列标注”两种不同的理念，识别候选句对中的对齐句对。【结果】在序列标注实验中，LSTM-CRF 模型的句子对齐效果最佳 F 值为 92.67%；在整体分类实验中，SVM 识别效果最佳 F 值为 90.63%；在特征组合实验中，同时使用 4 种特征的 F 值为 91.01%，效果优于其他特征组合。【局限】有待补充类型更丰富的原始语料。【结论】融合 4 种特征的 LSTM-CRF 神经网络模型能够有效识别古文-英文对齐句对，实现典籍双语句子自动对齐。

关键词：句子对齐　多语言信息处理　汉英平行语料　先秦典籍　数字人文

分类号：G351

DOI：10.11925/infotech.2096-3467.2019.0268

引用本文：梁继文，江川，王东波．基于多特征融合的先秦典籍汉英句子对齐研究[J]．数据分析与知识发现，2020，4（9）：123-132.(Liang Jiwen, Jiang Chuan, Wang Dongbo. Chinese-English Sentence Alignment of Ancient Literature Based on Multi-feature Fusion[J]. Data Analysis and Knowledge Discovery, 2020, 4(9): 123-132.)

1　引　言

历史典籍英译是推动中华传统文化走向世界的主要途径，提供了古文与英文对照信息的双语平行语料库则是其重要载体。典籍双语平行语料库由典籍的古文原文及其对应的英文译文构成，在机器翻译、跨语言信息检索等领域中具有较高的研究与利用价值。同时，随着数字人文的兴起，典籍平行语料库的建设作为人文计算的基础，在进行跨语言文体分析、语言差异挖掘等研究时对提升中国文化的软实力具有重要意义。

现有典籍双语平行语料库以白话文-英文、白话文-古文对应为主，平行单位通常局限在篇章或段落。句子级别对齐的典籍平行语料库精度较高，能提供更为有效的序化信息，衍生出专业性较强的双语词典，为后续开展相关跨语言人文计算（如语义提取、领域本体探索等）提供数据支持[1]。构建精加工的典籍句级平行语料库势在必行，古文与英文的句子对齐则成为首要任务。双语句子对齐指源语言与目标翻译语言在句子级别上的语义匹配，因古文与英文文本间存在较大差异，缺乏同源词、共现词与双语词典，所以实现典籍汉英句子对齐难度较大。

本文使用先秦典籍古文-英文双语语料，在前人研究的基础上结合语料特点选取长度与词汇信息构造句子对齐的特征，基于整体分类与序列标注的方法实现先秦典籍古文-英文双语句子自动对齐，为后

通讯作者：王东波，ORCID：0000-0002-9894-9550，E-mail：db.wang@njau.edu.cn。

*本文系国家自然科学基金面上项目“基于典籍引得的句法级汉英平行语料库构建及人文计算研究”（项目编号：71673143）的研究成果之一。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

续基于典籍双语平行语料库的跨语言信息检索、机器翻译以及其他数字人文研究提供重要资源。

2 相关研究

对齐技术是平行语料库构建过程中的研究重点，对齐单位小到词汇、短语，大到段落、篇章，其中句子对齐实现了双语句间语义对应，在信息检索、机器翻译与文本挖掘等任务中更为适用。句子对齐技术利用双语句对间的特征找到最优的匹配模式。自20世纪80年代以来，关于句子对齐的研究通常使用基于长度的方法、基于词汇信息的方法以及基于长度与词汇信息相结合的方法。

2.1 基于长度的方法

Brown等和Gale等观察到源语言文本中句子长度与对应译文句子长度有较大正相关，提出基于长度的对齐算法，使用动态规划算法对加拿大议会会议英法语料进行实验[2-3]。张霞等分别以句子动词数、实词数及字节等数量作为句长计算单位，并证实以词作为句子长度计算单位时效果最好[4]。

但基于长度的方法更适用于印欧等同源语系的句子对齐，不适用于跨语系或是噪声较多的语料[5]，且忽略文本的词形、语义信息，鲁棒性较差，易发生错误蔓延。

2.2 基于词汇信息的方法

基于词汇信息的句子对齐有两种：基于同源词信息的对齐与基于词典互译信息的对齐。Simard等首次将同源词引入句子对齐任务[6]。Church和Melamed基于同源词提出句子对齐算法[7-8]。Kay等在英-德语料库中认为包含互译词汇最多的候选句对是最佳对齐句对[9]。Ma提出基于词典的句子对齐算法并开发工具Champollion[10]。此外，也可将双语中互为翻译的特殊符号、实体等信息作为“锚点”进行句子对齐，如使用人名、地名作为锚点信息进行典籍术语对齐[11]以及使用共现字及互信息特征进行古-白句子对齐优化[12]。

基于词汇的方法提升了句子对齐的精准度，但需要使用同语系内的同源词信息或双语词典等领域信息资源，在面向跨语系研究时难度大大提升。此外，需要对文本进行人工标注等繁复处理，总体计算难度较大、速度较慢。

2.3 基于长度与词汇信息相结合的方法

Wu使用中国香港议会汉英语料创建特殊词表，首次将两种方法结合，有效实现句子对齐[13]。部分国外学者使用结合法构建出句子对齐工具，如Microsoft Bilingual Sentence Aligner[14]与hunalign[15]。Braune等和Trieu等则对Moore的算法[14]进行优化[16-18]。国内学者多数使用共现词特征或锚点信息与长度信息结合，双语种类包含古-白[19]、汉-英[20-21]与汉-维等[22-24]。结合法在面向不同领域、不同语言的研究中均取得较好效果[25]，是实现句子对齐的主流方法。

近年来，随着机器学习的发展，学者们尝试将机器学习的基于特征的分类思想引入句子对齐任务中。Fattah等以文本长度、特殊标点以及同源词等作为特征，训练概率神经网络、多分类支持向量机等模型实现英语和阿拉伯语的句子对齐[26-27]。刘颖等针对《史记》古文-白话文语料使用最大熵模型与BP（Back Propagation）神经网络实现短句对齐[28-29]。让子强在汉-老双语语料中以句长比例、词典匹配、词共现为特征，使用最大熵模型与支持向量机模型实现句子对齐[30]。陈相等将Fattah等的方法融合迁移学习思想，实现了生物医学领域的汉英句子对齐[31]。

纵观已有研究，在方法层面，国内外在进行句子对齐时更倾向于长度与词汇信息结合的方法，多数使用动态规划等算法进行概率最大句对的搜索，但不具有统一的标准模型。而在新兴技术的驱动下，选取对齐句对特征进行候选句对分类的句子对齐方法因为其灵活性与精准性正逐渐兴起。在语料选取层面，国外研究倾向于印欧语系内部的双语对齐，如英-法、英-德等，或是新闻常用的英语-阿拉伯语语料；而国内研究颇具多元性，包含汉-英、汉-日、汉-藏、汉-老、汉-维等跨语系句子对齐研究，但缺少对古文-英文句子自动对齐的探索。

3 研究方法与过程

本文选用融入更多信息的长度与词汇相结合的方法。鉴于动态规划等最大概率搜索算法在不同问题的阶段划分、状态识别时需要不同的方法，不具有统一的标准模型且存在维数障碍，本文在进行较为复杂的古-英句子对齐时并未使用搜索算法寻找最

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

大概率，而是引入分类思想，分类类别为“对齐句对”与“非对齐句对”两类。首先将可能对齐的双语句对列为候选对齐句对，将候选对齐句对进行类别识别，并标注该句对属于“对齐句对”（标为“S”）或“非对齐句对”（标为“O”）。然后选取特征，计算候选对齐句对的概率得分。假设每个双语句对的概率独立，根据候选句对的概率得分，概率分布最大的句对则为对齐句对，从而实现典籍双语句子对齐。

3.1　模型选择

在进行候选句对分类时基于“整体分类”与“序列标注”两种理念进行实验，参考有关研究，选取多个模型进行对比。两种实验思想的区别在于“整体分类”方法根据实验所选特征确定类别，更依赖于特征选择，无须考虑上下文的分类结果；而“序列标注”方法除特征外还依赖于上下文的分类结果。基于“整体分类”的实验思想将“对齐”与“非对齐”这两种类型的候选句对分别视为最小处理单元；而基于“序列标注”的实验思想将每一对独立候选句对均作为最小单元进行处理。

在已有使用“整体分类”理念进行句子对齐的研究中，刘颖等针对古-白句子对齐在最大熵模型（Maximum Entropy Model，MaxEnt）与 BP 神经网络模型上进行对比研究[29]；让子强使用支持向量机模型（Support Vector Machine，SVM）与最大熵模型实现汉-老句子对齐[30]；Fattah 在英-阿句子对齐时也曾选用支持向量机模型[27]，上述实验均取得较好结果。鉴于多层感知器（Multi-Layer Perception，MLP）是 BP 神经网络的基础架构，因此本文在“整体分类”方法中选择支持向量机模型（SVM）[32]、最大熵模型（MaxEnt）[33]与多层感知器（MLP）。

参考基于“序列标注”理念的相关研究，Grégoire 等使用循环神经网络（Recurrent Neural Network，RNN）结构进行双语对齐句对的抽取[34]。循环神经网络可以自动从上下文中学习到特征，适用于语言模型、机器翻译等情景，但在经过 Softmax 层进行最终序列标签标注时没有对标签的转移概率进行学习，存在标签偏置的问题。为解决这一问题，通常加入转移概率层，称之为 CRF，作为深层神经网络结构的输出层，数据经循环神经网络处理作为文本特征表示，并输入到 CRF 层，对标签的输出结果进行维特比解码修正。因此，本文在使用“序列标注”方法时选择以 RNN 结构为基础进行改进、性能更优的长短期记忆网络（Long Short-Term Memory，LSTM）[35]、门控循环神经网络（Gated Recurrent Unit，GRU）[36]，以及接入 CRF 层的 LSTM（LSTM-CRF）与 GRU（GRU-CRF）。以 LSTM-CRF 为例，模型结构如图 1 所示，模型输入篇章内双语句对观测序列 $(C_1E_1, C_1E_2, \cdots, C_nE_n)$，LSTM 层通过记忆单元中三个门的协作，有效提取文本深层特征信息，最终以向量化的形式输入 CRF 层。序列标注实验最终得到的序列标签表明该候选句对的类别，属于对齐句对时序列标签为 S，非对齐句对时序列标签为 O。

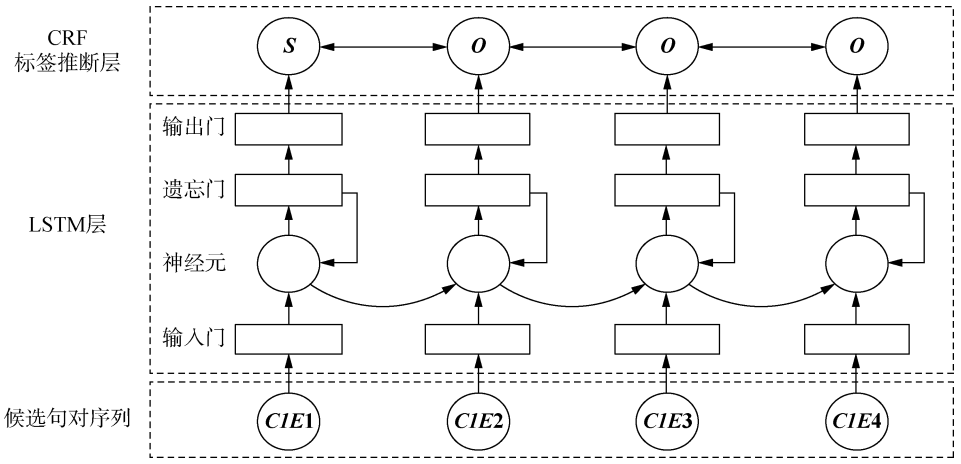


图 1　LSTM-CRF 模型示例

Fig.1　LSTM-CRF Model Sample

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3.2 典籍双语文本分析及特征选取

分析本文先秦典籍双语文本特点，参考已有双语句子对齐研究中选取的特征及实验结论，选取先秦典籍双语句子的特征。

结合近10年来双语句子自动对齐研究的特征选择可知，双语句子长度特征与对齐模式特征受语种限制较小，广泛适用于句子对齐研究。由于古文与英文断句习惯差异较大，且有学者认为句子的长度特征与对齐模式特征相结合即可代表句子的位置特征[29]，因此，字形、共现字、同源词、词典与位置特征均不适用于古-英句子对齐。但古-英双语句子中存在大量表示形式固定的实体，可将其视为锚点信息，同时双语句对多为1-1模式，因此，对齐句对的双语句末标点符号多数相同。

综上所述，在先秦典籍古-英句子对齐时选取以下4个特征：句子长度特征(F_L)、对齐模式特征(F_M)、标点符号特征(F_P)以及关键词互译特征(F_W)。

(1) 句子长度特征

句子长度特征是用来描述双语句长关系的特征。本文实验分别以中英文中的句号、问号、感叹号及分号作为双语句子切分标志，考虑到古文单字词居多且分词精准度欠佳，在排除标点后使用文本字节作为句长计算单位。

首先通过统计证明古-英双语语料长度关系特征的适用性。选用实验双语语料进行长度关系统计实验，得到对齐句对的双语句长与句长关系分布分别如图2和图3所示，可知双语句长具有相关性且符合正态分布，说明古-英互译句对间存在较为稳定的长度关系。

参考Gale等的长度特征计算公式[3]并进行调整，将双语语料的长度特征与对齐模式特征分而论之，视条件概率为长度特征，调整后的长度特征如公式(1)和公式(2)所示。

$$\delta_l\left(l_C, l_E\right)=\frac{\left(l_E - c\,l_C\right)}{\sqrt{l_C s^2}} \tag{1}$$

$$F_L = -100\log 2\left(1-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{|\delta_l|} e^{-x^2/2}\,dx\right) \tag{2}$$

其中，l_C和l_E分别表示双语文本句长；c表示双语语料文本句长总和比值；s^2表示$\frac{l_E - cl_C}{\sqrt{l_C}}$的方差。

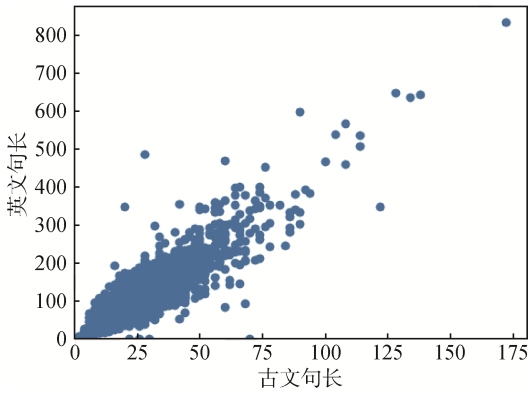


图2 对齐句对双语长度分布

Fig.2 Bilingual Length Distribution of Aligned Sentence Pairs

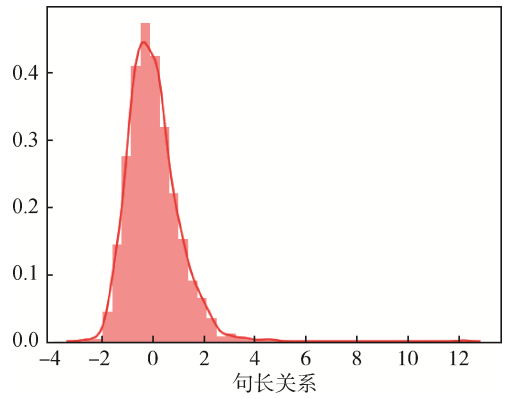


图3 对齐句对双语句长关系分布

Fig.3 Relational Distribution of the Length of Two Sentences of Aligned Sentence Pairs

(2) 对齐模式特征

对齐模式特征是用来描述双语对齐句对所含句子数量模式的特征。分别根据问号、句号、感叹号以及分号对双语文本句子进行划分并计数，本实验对齐模式分为1-1、1-2、2-1、2-2、1-0/0-1以及others模式。其中，1-0或0-1模式表明双语间未实现完整的句子对应。

在计算对齐模式特征时主要考虑句子对齐模式的概率。参考前人计算公式[3]进行对齐模式先验概率统计，1-1模式占比最高，概率为0.713；其次是1-2模式，概率为0.158。将对齐模式进行整合后进行对

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

齐模式特征的计算，如公式（3）所示。

$$F_M = -100\log(\mathrm{Prob}(M(l_C, l_E))) \quad (3)$$

（3）标点符号特征

标点符号特征是描述双语对齐句对末尾标点异同的特征。因为古文语料中的标点符号由后人根据语义、语境等语言因素添加，所以标点并未纳入句长计算，但标点却是表达句间关系的重要符号，对于判断句子对齐起重要作用。因中英文间符号差异较大，考虑共现符号精准度较差。在对句子对齐模式概率进行统计时发现，1-1 模式的对齐句对占比较高，可知多数双语对齐句对的句末标点相同，因此，将候选句对句末标点提取为特征。参考前人的符号特征计算方法[19]，当古文句末标点与英文句末点相同时特征值为 1，不同则为 0。

（4）关键词互译特征

关键词互译特征是描述双语对齐句对中存在关键词互译情况的特征。鉴于古-英不存在共现字/词与同源词且通用词典资源缺失，本文基于 Wu 创建特殊词表引入词汇信息的方法[13]，构建基于典籍双语语料的关键词词表。

古文语料中存在大量的人名、朝代等实体，对应的英文翻译多为拼音音译或是固定搭配，如古文中“颜回”音译为“YanHui”，“齊桓公”使用固定搭配“duke Huan of Qi”来表示。据此构建关键词互译词表。基于此，进行特征提取，作为实验的关键词互译特征。

在计算词典的匹配信息时，考虑到词汇对的重要程度不同，首先计算双语关键词词表中互译词对 (c_m, e_m) 在全部对齐的双语句子对中出现的总频次，然后计算在每个候选句对中出现词对的频次 $s_i tf(c_m, e_m)$，如公式（4）所示。

$$F_W = s_i tf(c_m, e_m) / \sum_{i=1, m=1} s_i tf(c_m, e_m) \quad (4)$$

3.3　语料选取及处理

（1）数据源

笔者对现有典籍古-英互译双语资源进行调研后，选取“中国哲学书电子化计划”（https://ctext.org/zhs）数据库作为实验数据源，该网站中部分先秦原典实现了古文与英文的对照翻译并以段落对齐的方式存储。实验语料选用主题相近的儒家经典著作《论语》与《礼记》，英文部分均为理雅各所译。首先对语料进行清洗，其中古文部分语料在同一位置出现两个标点，英文部分语料翻译缺失，对于此类情况统一用符号标明并进行校正与译文补充。分别以中文、英文中的句号、问号、感叹号及分号作为双语句子切分标志，对清洗后的段落对齐双语语料进行人工句子对齐，其中《论语》双语对齐句对 1 555 对，《礼记》双语对齐句对 4 386 对，两本典籍对齐句对共 5 941 对。

（2）候选句对生成

候选句对包含对齐句对与非对齐句对，生成候选句对旨在获取对齐可能性较大的双语平行句对，会影响总体实验的召回率。若将原始语料中所有双语句子进行多对多组合，可使实验召回率达到 100%，但因生成的候选句对数量过多且语义对应较差而不具有可行性。

由于本文基于段落对齐开展，语义对应限定在段落范围内，因此仅在段落范围内进行双语句子组合即可。通过观察语料进一步将候选句对在段落内的对齐模式进行重新限定。对生成的候选句对进行筛选后，共生成《论语》16 636 对候选句对、《礼记》20 092 对候选句对，作为实验的候选句对数据集。

（3）数据平衡处理

在机器学习或深度学习中的基础假设中，往往假设实验中不同类别的样本数据处于均衡状态，但类别样本数据差距较大对实验存在影响[37]。

本实验从数据层面处理不平衡样本问题。过采样方法以 SMOTE（Synthetic Minority Oversampling Technique）算法为代表，通过合成新的少数类样本，从而对小样本数据进行添加；欠采样方法则通过 NearMiss 等算法添加启发式规则选择样本，从而减少数量较大的样本，实现数据平衡。若单纯使用 SMOTE 算法，当边界样本与其他样本进行过采样产生差值时会产生噪声数据，因此在进行过采样之后需要使用下采样的方法对样本进行清洗，此种方法被称为上采样与下采样相结合的方法，选用上采样与下采样结合的方法 SMOTE-Edited Nearest Neighbors 进行数据平衡与归一化处理。

4　实验及结果讨论

在序列标注实验中选用准确率 P 值、召回率 R

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

值以及二者的调和平均数F值作为模型句子对齐效果的测评标准，在整体识别实验评价指标中增加一致性评价指标*Kappa*值，基于混淆矩阵来衡量整体分类精度。其中，A表示模型正确识别的对齐句对的数量；B表示将非对齐句对识别成对齐句对的数量；C表示未识别出的对齐句对的数量；P_o表示每一类正确分类的样本数量之和除以总的样本数，即总体分类精度；P_e表示每一类的预测数量与实际数量乘积的和再除以样本总数的平方。具体如公式(5)-公式(8)所示。

$$P = \frac{A}{A + B} \times 100\% \tag{5}$$

$$R = \frac{A}{A + C} \times 100\% \tag{6}$$

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \tag{7}$$

$$Kappa(k) = \frac{P_o - P_e}{1 - P_e} \tag{8}$$

4.1 整体分类实验

在整体分类实验中，选取支持向量机(SVM)、最大熵模型(MaxEnt)以及多层感知机(MLP)进行实验。

(1) 模型参数

监督学习中模型的参数对学习效果影响较大，在对数值型特征与类别型特征进行归一化后，对模型中的参数进行选择，旨在使模型性能最佳。

在SVM中，超参数包含核函数Kernel的种类和C值。基于实验语料进行分析，数据特征数量较小且样本数量正常，拟选用径向基函数(Radial Basis Function，RBF)进行实验。设置对照实验进行验证，控制其余超参数为默认值，使用线性核函数进行实验时*Kappa*值为71.3%，而使用RBF核函数进行实验的*Kappa*值为80.6%，与之前假设相符，因此选用RBF高斯核函数。使用Grid Search算法进行超参数优化实验，在经过特征归一化处理后，通过5折交叉验证，选取C=10，gamma=1作为最终SVM的实验参数。

最大熵模型与多层感知机需要调整的参数较少。在最大熵模型中选取改进迭代算法IIS(Improved Iterative Scaling)并设置最大迭代次数为1 000；在多层感知机中，激活函数选用神经网络使用较多的ReLU函数，设置4层隐藏层并控制神经元个数。

(2) 实验结果分析

4种经过超参数优化的模型的实验结果对比如表1所示。

表1 整体分类模型句子对齐实验结果

Table 1 Experimental Results of the Overall Classification Model

模型	*P*	*R*	*F*	*Kappa*
SVM	90.19%	90.02%	90.63%	80.02%
MaxEnt	98.57%	76.01%	85.83%	73.59%
MLP	88.18%	89.25%	88.71%	76.19%

由表1可知，在整体分类实验中，基于支持向量机模型的句对分类效果最好，F值达90.63%，明显优于最大熵模型与多层感知机。多层感知机作为浅层人工网络，需要具有更多隐藏单元的测试数据控制网络复杂性，相较之下更适用于大样本数据。最大熵模型对齐句对识别准确率较高，但召回率较低，原因如下：最大熵统计模型获得的是所有满足约束条件的模型中信息熵极大的模型，预测风险最小，因此其具有较高的准确率；观察结果发现未被识别的多为对齐模式中的“others”模式句对，这是因为在生成候选句对时以1-1与1-2模式为主，以致其他类型对齐句对缺失，因此最大熵模型在进行学习时其他模式句对的预测风险增大导致判别概率降低。

4.2 序列标注实验

在序列标注实验中，选取LSTM、GRU、LSTM-CRF以及GRU-CRF共4种模型。

(1) 模型参数

实验使用Python程序语言，均采用TensorFlow框架搭建神经网络进行序列识别，在搭载4GB显存“NVIDIA”QuadroK1200型号GPU的Linux操作系统中进行。激活函数为ReLU，其中神经单元参数可在LSTM与GRU两种神经单元中进行选取，并选择是否接入CRF层作为输出层。实验中每层隐藏神经元为200，最大迭代次数为100，选取Adam作为模型优化器，机器能承受的最大训练样本量Batch为64，Dropout为0.5。

(2) 实验结果分析

基于4种模型进行5折交叉验证的实验结果如表2所示。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表2　序列标注模型实验结果

Table 2　Experimental Results of Sequence Labeling Model

模型	*P*	*R*	*F*
LSTM	84.07%	93.64%	88.60%
LSTM-CRF	97.35%	88.42%	92.67%
GRU	81.68%	78.65%	80.61%
GRU-CRF	90.35%	89.60%	89.97%

（1）总体上，在序列标注实验中，接入CRF模型作为输出层的深层网络与原始网络相比，效果有明显提升；

（2）LSTM-CRF深层神经网络的对齐句对识别效果最好；

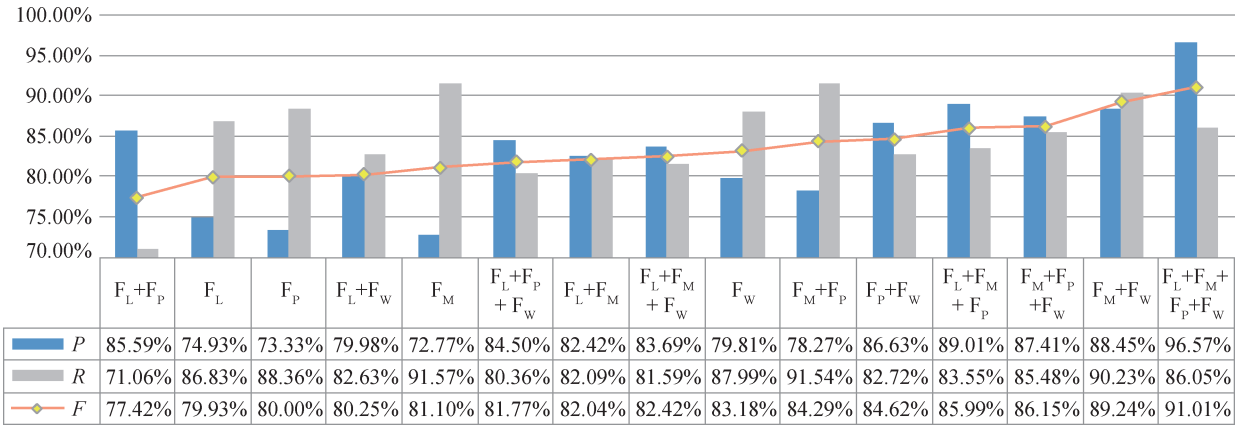
（3）未接入CRF的LSTM神经网络准确率较低，而未接入CRF的GRU神经网络的召回率较低，两种模型虽然结构相似，但因记忆单元不同在相同的任务中表现差异较大。

将基于整体分类与序列标注思想的模型句子对齐实验结果进行对比，LSTM-CRF深层神经网络的句子对齐效果最好；基于序列标注思想的句子对齐效果略优于整体对齐的机器学习与浅层人工神经网络模型。

综合来看，使用5 941对古-英双语句对，效果最好的模型*F*值达92.67%，实验总体效果较好。与前人古文相关类似研究[29]相比，本文在准确率方面略占优势。

4.3　句子对齐模型的特征选择

为更全面地进行LSTM-CRF句子对齐模型的整体性能分析，使用15组特征组合进行5折交叉验证实验，并将不同特征组合按照*F*值升序展示实验结果，如图4所示。



	F_L+F_P	F_L	F_P	F_L+F_W	F_M	$F_L+F_P+F_W$	F_L+F_M	$F_L+F_M+F_W$	F_W	F_M+F_P	F_P+F_W	$F_L+F_M+F_P$	$F_M+F_P+F_W$	F_M+F_W	$F_L+F_M+F_P+F_W$
P	85.59%	74.93%	73.33%	79.98%	72.77%	84.50%	82.42%	83.69%	79.81%	78.27%	86.63%	89.01%	87.41%	88.45%	96.57%
R	71.06%	86.83%	88.36%	82.63%	91.57%	80.36%	82.09%	81.59%	87.99%	91.54%	82.72%	83.55%	85.48%	90.23%	86.05%
F	77.42%	79.93%	80.00%	80.25%	81.10%	81.77%	82.04%	82.42%	83.18%	84.29%	84.62%	85.99%	86.15%	89.24%	91.01%

图4　特征组合句子对齐综合性能比较

Fig.4　Performance of Feature Combination Sentence Alignment

（1）整体上来看，当使用4种特征时*F*值最大为91.01%，其次是对齐模式特征与关键词互译特征进行组合（F_M+F_W）时，*F*值为89.24%；

（2）对齐模式特征F_M对应的召回率最高，为91.57%，但准确率最低；4种特征全组合对应的准确率最高，为96.57%。

上述结果证实了选取句长、对齐模式、标点符号以及关键词互译信息作为特征集合的合理性。值得注意的是，因为特征间存在相互作用，所以并非添加任意特征都会提升模型效果。为更加明确单一特征与组合特征对模型效果的影响，以及加入某一特征后模型性能的变化，进行两种特征组合与单一特征的*F*值对比，结果如图5所示。

（1）4种独立特征在句子对齐模型中的有效性由高到低为：F_W>F_M>F_P>F_L。其中，F_W在句子对齐中作用最大，这也证实了在句子对齐中词汇信息的重要性，但不排除关键词表指向较强等影响因素存在；而相较之下仅使用F_L时，*F*值偏低，这一结论符合Gale等提出的在跨语系特殊语料中长度特征对平行句对区分能力较差的观点[3]。

（2）当为单一特征添加其余特征后，除F_M外，其余融入句长特征的组合特征实验效果与原始单特征

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

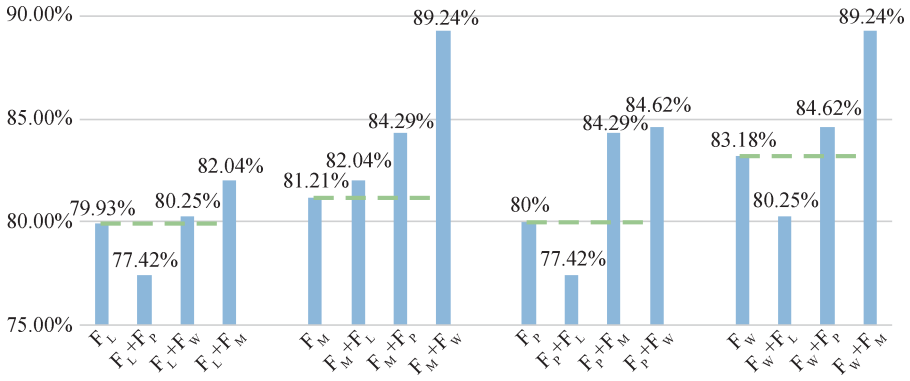


图5　特征组合与单一特征实验结果对比

Fig.5　Experimental Results Between Feature Combination and Single Feature

相比均有所下降，验证了部分前人研究中将句长与对齐模式的乘积视为一个单独特征的合理性。

（3）F_M与其他特征相结合有助于改进句子对齐效果。

（4）同时考虑F_M与F_W的模型性能最好，F值为89.24%；而同时考虑F_L与F_P时模型性能最差，F值仅为77.42%。

5　结　语

本文探索了先秦典籍古-英双语句子的自动对齐，将句子对齐视为双语候选句对中“对齐句对”与“非对齐句对”分类问题，结合实验文本特点提取较为有效的特征，基于“整体分类”与“序列标注”两种理念实现句子对齐并围绕特征选取与组合展开讨论。实验结果表明，基于“序列标注”思想的LSTM-CRF神经网络在处理对齐句对识别任务时具有一定优势，且同时使用4种特征进行组合的模型性能最优。句子级别对齐的先秦典籍双语语料为后续典籍双语平行语料库构建提供了数据支持，可衍生出有效的双语词典，并在一定程度上有助于跨语言文本检索与挖掘、语义提取以及领域本体的探索。

本文可在以下方面进行改进：首先在语料方面，增加类别丰富的原始语料，探索不同规模与类别的语料对句子对齐效果稳定性的影响，以及不同的候选句对生成方式对实验结果的影响；增加“不完全对齐”类别的句对，用以提升召回率。此外，可引入深度学习的方法进行命名实体识别，自动提取双语实体，实现关键词互译词典的自动构建。

参考文献：

[1] Guo M, Shen Q L, Yang Y F, et al. Effective Parallel Corpus Mining Using Bilingual Sentence Embeddings[OL]. arXiv Preprint, arXiv:1807.11906.

[2] Brown P F, Lai J C, Mercer R L. Aligning Sentences in Parallel Corpora[C]//Proceedings of the 29th Annual Meeting on Association for Computational Linguistics. 1991:169-176.

[3] Gale W A, Church K W. A Program for Aligning Sentences in Bilingual Corpora[J]. Computational Linguistics, 1993, 19(1): 75-102.

[4] 张霞, 昝红英, 张恩展. 汉英句子对齐长度计算方法的研究[J]. 计算机工程与设计, 2009, 30(18): 4356-4358.(Zhang Xia, Zan Hongying, Zhang Enzhan. Study on Length Computation Method of Chinese-English Sentence Alignment[J]. Computer Engineering and Design, 2009, 30(18): 4356-4358.)

[5] Chuang T C, Yeh K C. Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria[J]. International Journal of Computational Linguistics & Chinese Language Processing, 2005, 10(1): 95-122.

[6] Simard M, Foster G F, Isabelle P. Abstract Using Cognates to Align Sentences in Bilingual Corpora[C]//Proceedings of the 4th International Congress on Theoretical & Methodological Issues in Machine Translation. 1992: 67-81.

[7] Church K W. Char_align: A Program for Aligning Parallel Texts at the Character Level[C]//Proceedings of the 31st Annual Meeting on Association for Computational Linguistics. 1993: 1-8.

[8] Melamed I D. Bitext Maps and Alignment via Pattern Recognition[J]. Computational Linguistics, 1999, 25(1): 107-130.

[9] Kay M, Röscheisen M. Text-translation Alignment[J]. Computational Linguistics, 1993, 19(1): 121-142.

[10] Ma X Y. Champollion: A Robust Parallel Text Sentence Aligner [C]//Proceedings of LREC-2006. 2006: 489-492.

[11] 李秀英. 基于历史典籍双语平行语料库的术语对齐研究[D]. 大

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

连:大连理工大学, 2010.(Li Xiuying. Term Translation Pair Alignment Based on a Bilingual Parallel Corpus of Chinese Historical Classics [D]. Dalian: Dalian University of Technology, 2010.)

[12] 李闻. 汉语古现句子对齐研究[C]//第十一届全国机器翻译研讨会(CWMT 2015), 中国, 合肥. 2015: 90-96.(Li Wen. Research on Alignment of Ancient Chinese Sentences to Modern Ones[C]// Proceedings of China Workshop on Machine Translation, Hefei, China. 2015: 90-96.)

[13] Wu D K. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria[C]//Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. 1994: 80-87.

[14] Moore R C. Fast and Accurate Sentence Alignment of Bilingual Corpora[C]//Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users. 2002: 135-144.

[15] Varga D, Halácsy P, Kornai A, et al. Parallel Corpora for Medium Density Languages[J]. Amsterdam Studies in the Theory and History of Linguistic Science Series 4, 2007. DOI：10.1075/cilt.292.32var.

[16] Braune F, Fraser A. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora[C]// Proceedings of the 23rd International Conference on Computational Linguistics: Posters. 2010: 81-89.

[17] Trieu H L, Nguyen P T, Nguyen K A. Improving Moore's Sentence Alignment Method Using Bilingual Word Clustering [C]//Proceedings of the 5th International Conference KSE. 2014: 149-160.

[18] Trieu H L, Nguyen P T, Nguyen L M. A New Feature to Improve Moore's Sentence Alignment Method[J]. VNU Journal of Science: Computer Science and Communication Engineering, 2015, 31(1): 32-44.

[19] 郭锐, 宋继华, 廖敏. 基于自动句对齐的相似古文句子检索[J]. 中文信息学报, 2008, 22(2): 87-91, 105.(Guo Rui, Song Jihua, Liao Min. Ancient Sentence Search Based on Sentence Auto Alignment in Parallel Corpus of Ancient and Modern Chinese[J]. Journal of Chinese Information Processing, 2008, 22(2): 87-91,105.)

[20] 钱丽萍, 赵铁军, 杨沐昀, 等. 基于译文的英汉双语句子自动对齐[J]. 计算机工程与应用, 2000, 36(12): 123-125.(Qian Liping, Zhao Tiejun, Yang Moyun, et al. Translation-based Automatic Alignment of English and Chinese Parallel Corpora[J]. Computer Engineering and Applications, 2000, 36(12): 123-125.)

[21] 张艳, 柏冈秀纪. 基于长度的扩展方法的汉英句子对齐[J]. 中文信息学报, 2005, 19(5): 33-38, 60.(Zhang Yan, Kashioka Hideki . Aligning Sentences in Chinese-English Corpora with Extended Length-based Approach[J]. Journal of Chinese Information Processing, 2005, 19(5): 33-38, 60.)

[22] 塞麦提·麦麦提敏, 侯敏, 吐尔根·依布拉音. 基于锚点句对的汉维句子对齐方法[J]. 计算机工程, 2015, 41(4): 166-170.(Saimaiti Maimaitimin, Hou Min, Tuergen Yibulayin. Chinese-Uyghur Sentence Alignment Method Based on Anchor Sentence Pairs [J]. Computer Engineering, 2015, 41(4): 166-170.)

[23] 田生伟, 吐尔根·依布拉音, 禹龙, 等. 多策略汉维句子对齐[J]. 计算机科学, 2010, 37(4): 215-218, 292.(Tian Shengwei, Tuergen Yibulayin, Yu Long, et al. Chinese-Uyhur Sentence Alignment Based on Hybrid Strategy[J]. Computer Science, 2010, 37(4): 215-218, 292.)

[24] 李文刚, 周杰, 杨保群. 基于词典和句长及位置的双语对齐方法的改进[J]. 现代电子技术, 2011, 34(14): 25-27.(Li Wen'gang, Zhou Jie, Yang Baoqun. Improvement of Bilingual Sentence Alignment Method Based on Sentence Length and Location Information with Bidirectional Dictionary [J]. Modern Electronics Technique, 2011, 34(14): 25-27.)

[25] Sennrich R, Volk M. MT-based Sentence Alignment for OCR-generated Parallel Texts[C]//Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA 2010). 2010.

[26] Fattah M A, Bracewell D B, Ren F J, et al. Sentence Alignment Using P-NNT and GMM[J]. Computer Speech and Language, 2007, 21(4): 594-608.

[27] Fattah M A. The Use of MSVM and HMM for Sentence Alignment[J]. Journal of Information Processing Systems, 2012, 8 (2): 301-314.

[28] 刘颖, 王楠. 古汉语与现代汉语句子对齐研究[J]. 计算机应用与软件, 2013, 30(11): 127-130.(Liu Ying, Wang Nan. Research on Classical and Modern Chinese Sentence Alignment[J]. Computer Applications and Software, 2013, 30(11): 127-130.)

[29] 刘颖, 王楠. 最大熵模型和BP神经网络的短句对齐比较[J]. 计算机工程与应用, 2015, 51(7): 112-117.(Liu Ying, Wang Nan. Comparison of Clause Alignment Based on Maximum Entropy Model and Back Propagation Neural Network Model[J]. Computer Engineering and Applications, 2015, 51(7): 112-117.)

[30] 让子强. 汉老双语句子对齐方法研究[D]. 昆明: 昆明理工大学, 2017.(Rang Ziqiang. Research on Chinese-Lao Bilingual Sentence Alignment Methods[D]. Kunming: Kunming University of Science and Technology, 2017.)

[31] 陈相, 林鸿飞, 杨志豪. 基于高斯混合模型的生物医学领域双语句子对齐[J]. 中文信息学报, 2010, 24(4): 68-73.(Chen Xiang, Lin Hongfei, Yang Zhihao. Sentence Alignment for Biomedicine Texts Based on Gaussian Mixture Model[J]. Journal of Chinese Information Processing, 2010, 24(4): 68-73.)

[32] Cortes C, Vapnik V. Support-vector Networks[J]. Machine Learning, 1995, 20(3): 273-297.

[33] Jaynes E T. On the Rationale of Maximum-entropy Methods[J].

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Proceedings of the IEEE, 1982, 70(9):939-952.
[34] Grégoire F, Langlais P. A Deep Neural Network Approach to Parallel Sentence Extraction[OL]. arXiv Preprint, arXiv: 1709.09783.
[35] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[36] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation[OL]. arXiv Preprint, arXiv: 1406.1078.
[37] Hensman P, Masko D. The Impact of Imbalanced Training Data for Convolutional Neural Networks[EB/OL].[2019-03-02]. https://www. kth. se/social/files/588617ebf2765401cfcc478c/PHensmanDMasko_dkand15.pdf.

作者贡献声明：

梁继文：研究方案设计，语料获取及处理，进行实验，撰写论文；
江川：模型构建及调参，论文修改；
王东波：确定论文选题，完善研究思路，论文最终版本修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，E-mail：1943769394@qq.com。
[1] 梁继文,王东波. all_par. zip. 先秦典籍段落对齐双语平行语料.
[2] 梁继文,王东波. sen. zip. 先秦典籍句子对齐双语平行语料.
[3] 梁继文,王东波. key_word. xlsx. 典籍双语实体关键词词表.
[4] 梁继文,王东波. data. zip. 实验数据.
[5] 梁继文,王东波. result_all. xlsx. 实验结果.

收稿日期：2019-03-11
收修改稿日期：2019-11-22

Chinese-English Sentence Alignment of Ancient Literature Based on Multi-feature Fusion

Liang Jiwen[1] Jiang Chuan[2] Wang Dongbo[2, 3]

[1](School of Information Management, Nanjing University, Nanjing 210023, China)
[2](College of Information Science & Technology, Nanjing Agricultural University, Nanjing 210095, China)
[3](Facultair Onderzoekscentrum ECOOM, KU Leuven, Leuven B-3000, Belgium)

Abstract: [**Objective**] This paper proposes a method automatically aligning Chinese sentences from Pre-Qin Literature with their English translations, aiming to construct bilingual sentence-level parallel corpus and support cross-language retrieval. [**Methods**] First, we modified classification method for parallel sentence pairs to align bilingual sentences from historical literature. Based on the characteristics of bilingual corpus, we retrieved features of bilingual sentence pairs. Finally, with "sequence labeling" and "overall classification", we identified aligned pairs from candidate sentences. [**Results**] In the sequence labeling experiment, the LSTM-CRF model yielded the best performance with its F value reaching 92.67%. In the overall classification experiment, the SVM had the best results with a F value of 90.63%. In the experiment combining all four features, the F value was 91.01%. [**Limitations**] The corpus size needs to be expanded. [**Conclusions**] The LSTM-CRF model with four features could effectively align ancient Chinese sentences with their English translations.

Keywords: Sentence Alignment Multilingual Information Processing Chinese-English Parallel Corpus Pre-Qin Literature Digital Humanities

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net