文章编号：2095-2163(2021)01-0028-04 中图分类号：TP391 文献标志码：A

融入多特征的汉韩双语自动句对齐方法

刘晨阳[1]，唐慧丰[2]
（1 信息工程大学 洛阳校区，河南 洛阳 471003；2 信息工程大学，郑州 450001）

摘 要： 为解决汉韩双语平行语料库资源匮乏以及传统句对齐算法面向跨语系语言准确率较低的问题，提出了融合特征的汉韩双语句对齐方法。首先将 Bi-LSTM 融入孪生神经网络构建句对齐模型，用以分别提取汉语和韩语句子的特征并进行对齐。之后基于语料的特点提取句对齐特征融入输入层。通过与传统 Bi-LSTM 和不同特征组合的孪生 Bi-LSTM 的对比实验证明，融入特征的孪生 Bi-LSTM 方法在句对齐任务中具有更优越的性能。

关键词： 汉语-韩语；自动句对齐；孪生 Bi-LSTM；多特征；平行语料库

Automatic sentence alignment method for Chinese-Korean bilinguals with multi-features

LIU Chenyang[1], TANG Huifeng[2]
(1 Luoyang Campus, Information Engineering University of PLA Strategic Support Forces, Luoyang Henan 471003, China;
2 Information Engineering University of PLA Strategic Support Forces, Zhengzhou 450001, China)

【Abstract】 In order to solve the problem of the lack of resources in the Chinese-Korean bilingual parallel corpus and the low accuracy of traditional sentence alignment algorithms for cross-lingual languages, a method of Chinese-Korean bilingual sentence alignment with multiple-features is proposed. Firstly, Bi-LSTM is integrated into Siamese Neural Network to construct sentence alignment model, which is used to extract the features of Chinese and Korean sentences and align them. After that, sentence alignment features are extracted based on the features of corpus and integrated into the input layer. Compared with the traditional Bi-LSTM and the Siamese Bi-LSTM with different feature combinations, it is proved that the Siamese Bi-LSTM method with features has better performance in sentence alignment task.

【Key words】 the Chinese-Korean; automatic sentence alignment; Siamese Bi-LSTM; multi-features; parallel corpus

0 引 言

双语句对齐是指将语料中的双语互译句对进行匹配，该技术可以将篇章、段落级别对齐的语料进一步细化为句对齐语料，从而构建高质量的双语平行语料库。目前，汉韩双语平行语料库资源较为匮乏，构建方式也多以人工为主，因此通过自动句对齐技术高效地构建汉韩双语平行语料库，对以机器翻译为代表的汉韩自然语言处理任务有着重大意义。

1 相关研究

研究可知，双语句对齐的方法主要有基于长度的方法、基于词汇的方法和基于长度与词汇相结合的方法。对此拟展开研究论述如下。

1.1 基于长度的句对齐方法

该方法最初由 Gale&Church 提出，其依据是源语言与译文文本长度具有关联性，据此区分对齐与非对齐句对[1]。传统的对齐算法多以字节、字符或词数作为长度计量单位，此后的研究者利用其他元素计算句子长度，如张霞等人[2]将句子所含的动词、名词、形容词等词语作为句长计量单位，在英汉句对齐上取得了良好的效果。

基于长度的方法忽略了词汇形态、词义等信息，因此在印欧语系语言上对齐效果较好，却不适合分属汉藏语系和阿尔泰语系的汉语和韩语。

1.2 基于词汇的句对齐方法

该方法利用双语词典和词汇信息进行句对齐，其依据是句中的词汇信息在一定程度上可以明确该句子的主题。Kay 等人[3]基于对英德语料库的研究提出将包含互译词汇最多的句对作为对齐句对。Ma[4]开发了基于词典的句子对齐工具 Champollin。

基于词汇的方法虽然提升了句对齐的精准度，但对双语资源的要求较高，在面对跨语系、低资源的语言对齐时难度大大提升。

作者简介：刘晨阳（1996-），男，硕士研究生，主要研究方向：自然语言处理、计算语言学；唐慧丰（1973-），男，博士（后），教授，主要研究方向：智能信息处理、机器学习。
通讯作者：唐慧丰 Email：lcy_96108@163.com

收稿日期：2020-10-29

1.3　基于长度和词汇相结合的句对齐方法

该方法结合两者的优点，在降低计算复杂度的同时提高了鲁棒性。Wu[5]在基于长度的基础上，利用香港议会汉英语料创建特殊词表，将长度与词汇相结合有效实现句对齐。

近年来，随着机器学习与深度神经网络的发展，基于特征的分类思想被引入句对齐任务。让子强[6]以句长比例、词共现等为特征，使用最大熵与支持向量机模型实现汉语、老挝语句对齐；贾善崇等人[7]利用 Bi-LSTM 模型，使用汉语老挝语词向量和长度特征实现汉老双语句对齐；梁继文等人[8]将对齐模式、关键词互译等特征融入神经网络，完成了先秦典籍汉英句子的自动对齐。

2　韩-汉双语句对齐模型

2.1　词向量层

词向量（Word Embedding）能够将词转化成一种分布式表示，即将每个词映射到低维度的连续向量空间中，词的分布式表示易于神经网络处理，同时可以展现词间的深层语义关系。

对分词后的汉语和韩语句子使用 Word2Vec 算法的 CBOW（Continuous Bag-of-Words）模型计算词向量，得到汉语句子的词向量 $\boldsymbol{V}_{cn}$ 和韩语句子的词向量 $\boldsymbol{V}_{kr}$ 作为神经网络的输入。

2.2　孪生神经网络

孪生神经网络（Siamese Neural Network，SNN）是基于 2 个人工神经网络建立的耦合架构。狭义的孪生神经网络要求 2 个神经网络结构相同且共享权重；广义的孪生神经网络，又称“伪孪生神经网络（Pseudo-Siamese Network）”则可以由 2 种任意的神经网络组合而成，比如卷积神经网络（CNN）、循环神经网络（RNN）。孪生神经网络以 2 个样本为输入，输出其嵌入高维空间的表征，因此在比较样本相似度上取得了较好效果。

2.3　Bi-LSTM

长短期记忆网络（Long Short-Term Memory Network，LSTM）[9]最早由 Hochreiter 和 Schmidhuber 提出，是循环神经网络（RNN）的一个变体，通过输入、输出和遗忘三种门控制上下文信息的选择，有效解决了传统 RNN 的梯度爆炸或消失问题。LSTM 网络的循环单元结构如图 1 所示。输入门 $\boldsymbol{i}_t$ 控制当前时刻的候选状态 $\tilde{c}_t$ 有多少信息需要保存，其计算方式如下：

$$\boldsymbol{i}_t = \boldsymbol{\sigma}(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i), \tag{1}$$

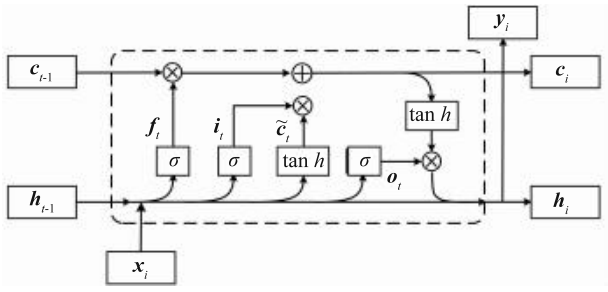

图 1　LSTM 单元

Fig. 1　LSTM unit

遗忘门 $\boldsymbol{f}_t$ 控制上一个时刻的内部状态 c_{t-1} 需要遗忘多少信息，其计算方式如下：

$$\boldsymbol{f}_t = \boldsymbol{\sigma}(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f), \tag{2}$$

控制门 $\boldsymbol{o}_t$ 控制当前时刻的内部状态 $\boldsymbol{c}_t$ 有多少状态输出给外部状态 $\boldsymbol{h}_t$，其计算方式如下：

$$\boldsymbol{o}_t = \boldsymbol{\sigma}(\boldsymbol{W}_0 \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o). \tag{3}$$

LSTM 网络循环单元的计算过程为：首先使用上一时刻的外部状态 $\boldsymbol{h}_{t-1}$ 和当前时刻的输入 $\boldsymbol{x}_t$ 计算出 3 个门，之后利用遗忘门 $\boldsymbol{f}_t$ 和输入门 $\boldsymbol{i}_t$ 更新记忆单元 $\boldsymbol{c}_t$，最终结合输出门 $\boldsymbol{o}_t$ 将内部状态的信息传递给外部状态 $\boldsymbol{h}_t$。

双向长短期记忆网络（Bi-directional Long Short-Term Memory Network，Bi-LSTM）将前向 LSTM 与后向 LSTM 结合在一起，能够从前后两个方向对序列进行训练，从而更好地获取上下文信息，在长序列处理任务中效果更好。

2.4　孪生 Bi-LSTM

基于长度、词汇等传统的句对齐方法在计算句子相似度时难以挖掘句子的深层特征，而传统的神经网络方法将句对齐视为分类问题[10]，忽略了双语语句的内在特征。因此结合孪生神经网络处理双输入样本的特点和 Bi-LSTM 善于提取长序列特征的优势，将其作为孪生神经网络的网络结构，得到孪生 Bi-LSTM 进行双语句对齐任务。最终的句对齐模型如图 2 所示。

模型训练的过程如下，将分词过后的汉语句子 *Sentence_cn* 和韩语句子 *Sentence_ko*，各自向量化后输入至一组 Bi-LSTM 中；由该组 Bi-LSTM 得到各自输入序列的特征向量 $\boldsymbol{h\_cn}$ 和 $\boldsymbol{h\_ko}$，经融合后通过线性层映射至对齐结果。

2.5　汉-韩双语文本分析及特征选取

2.5.1　句子长度特征 F_L

定义句子长度特征 F_L 作为双语句长关系的特征。以去除标点后的字节数作为句长单位进行长度关系统计，得到句对长度分布如图 3 所示。相比于

未对齐句对,已对齐句对的句长具有相关性。

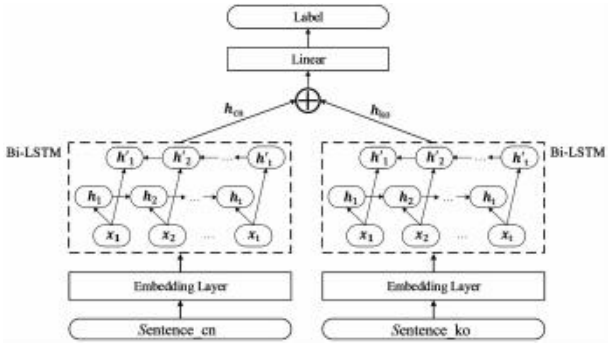

图 2 基于孪生 Bi-LSTM 的汉韩句对齐模型

Fig. 2 Chinese-Korean sentence alignment model based on the Siamese Bi-LSTM

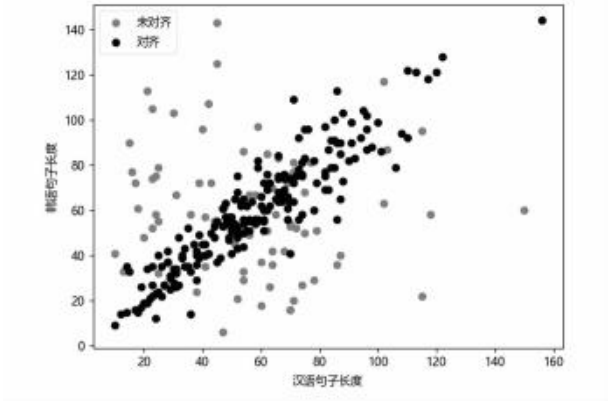

图 3 对齐、未对齐汉语、韩语句长关系

Fig. 3 The length relationship between aligned, unaligned Chinese and Korean sentences

因此定义长度特征为:

$$F_L = \frac{L_i - L_{\min}}{L_{\max} - L_{\min}}. \tag{4}$$

其中,$L_{\max}$,$L_{\min}$ 分别表示汉(韩)句子的最大长度和最小长度,L_i 表示当前句子长度。

2.5.2 字对齐特征 F_N

定义数字对齐特征 F_N 作为双语句对中数字对齐的特征。数字对齐特点见表 1。通过观察发现,语料中双语句对中阿拉伯数字的表达方式相同,因此将存在数字对齐的句对特征设为 1,不存在则设为 0。

表 1 语料中汉语、韩语的数字对齐特点

Tab. 1 Numerical alignment of Chinese and Korean in corpus

汉语	韩语
3 300 万	3300 만
7 时 18 分	7 시 18 분

在模型的输入层,将预训练的词向量与特征向量首尾向量,形成联合向量输入至神经网络。

3 实验分析

3.1 语料介绍及数据预处理

本实验所用的数据集来自人工进行句对齐的汉-韩双语新闻。原始数据集见表 2。

表 2 原始语料格式

Tab. 2 Original corpus format

汉语	韩语
2017 年按统计结算结果售出排位列出的 20 个国营企业。	2017년회계결산결과 매출순으로분류된20개공기업기준.
在最近的新闻中最关心的话题是?	최근뉴스에서가장관심을가졌던화제는?
今年 800 万观众崩溃。	올해 800만 관중이 붕괴했다.

实验选取 8 000 对句对作为总数据集,从中选取 5 000 对对齐句对作为正例数据,3 000 对完全打乱顺序作为负例数据。实验中训练集占 80%,测试集占 20%。在数据预处理时,将特殊符号及标点符号进行清洗。

对汉语句子使用 jieba 进行分词处理,对韩语句子使用 Konlpy 韩语自然语言处理库进行分词处理。

训练语料的格式为一行文本,即“汉语 韩语对齐标志”,其中对齐标志“1”表示对齐,“0”表示未对齐,语料的具体格式见表 3。

表 3 训练语料格式

Tab. 3 Training corpus format

汉语	韩语	标签
最近 在读 什么 书 呢	요즘읽고있는책은무엇인가	1
美国 高校 棒球队应该向 他们学习	미국 고교 야구팀이 배워야 한다고 했다	1
大家 好 我 是 宋仲基	이하송중기입장문전문이다	0

3.2 实验设置

实验中各参数设置如下:$batch\_size = 64$,$word\_dim = 128$,$epoch = 10$,优化器选取 Adam,学习率为自适应的学习率函数,其他参数为默认参数。

模型使用准确率(*Precision*)、召回率(*Recall*)以及 *F* 值(*F-Score*)作为评价指标。计算公式如下:

准确率=(所有预测正确的样本/总的样本)*100%, (5)

召回率=(将正类预测为正类/所有的正类)*100%, (6)

F 值=准确率*召回率*2/(准确率+召回率). (7)

3.3 实验结果与分析

实验选取了 2 种模型和不同的特征组合分别进

行对比，首先选取 Bi-LSTM 和孪生 Bi-LSTM 两种模型进行对比实验，结果见表 4。

表 4　Bi-LSTM 与孪生 Bi-LSTM 对比实验结果

Tab. 4　Comparison of experimental results between Bi-LSTM and Siamese Bi-LSTM　/%

特　征	准确率	召回率	*F* 值
Bi-LSTM	56.7	80.0	66.5
孪生 Bi-LSTM	79.4	85.6	82.3

从表 4 中可以看出，孪生 Bi-LSTM 模型在准确率、召回率、*F* 值三项指标上相比传统的 Bi-LSTM 都有较大的提升，这是由于该模型分别提取了汉韩句子的特征而不是将其合成一个整体作为输入。由此可以推断孪生 Bi-LSTM 模型符合句对齐任务的内在逻辑，可以取得较好的句对齐效果。

再将原始未加入特征的词向量作为基准，选取孪生 Bi-LSTM 进行多类型特征融入的对比实验，结果见表 5。

表 5　多类型特征组合实验结果

Tab. 5　Experimental results of multi-features combination　/%

特　征	准确率	召回率	*F* 值
词向量	79.4	85.6	82.3
词向量+长度特征	80.1	94.1	86.5
词向量+数字对齐特征	78.4	88.6	83.1
词向量+长度特征+数字对齐特征	79.6	93.1	86.0

可以看出，加入长度特征后的实验结果有了较大提升，模型在分类准确率上提升了 0.7%，召回率上提升了 8.5%，从而在 *F* 值上提升了 4.2%；数字对齐特征的加入，使模型在分类准确率上降低了 1.0%，但在召回率上提升了 3.0%，从而使 *F* 值提升了 0.8%；在同时加入长度特征和数字特征后，模型在分类准确率上提升了 0.2%，召回率上提升了 7.5%，从而在 *F* 值上提升了 3.7%。

从实验结果上看，在句对齐任务中，长度特征对模型效果有较好的正向作用，而数字对齐特征对模型效果作用较小，因此融入句子长度特征的孪生 Bi-LSTM模型能够有效提升汉韩双语句对齐的效果。

4　结束语

本文面向汉韩双语句对齐任务，将 Bi-LSTM 融入孪生神经网络结构中，同时针对汉韩双语对齐句子的相关特征，提取句子长度特征、数字对齐特征融入输入层，实现了汉韩双语的自动句对齐。实验结果表明，孪生 Bi-LSTM 在句对齐任务上的表现优于传统 Bi-LSTM，且相关特征对于提升模型效果有正向作用。

本文仍存在一定不足。一方面，语料中对齐、非对齐双语句对的比例关系对结果的影响有待进一步探究，可增加原始语料的类别和“完全对齐”类别的句对；另一方面，选取的双语句对齐特征数量有待进一步扩充，可引入更多特征进行对比实验，找出最佳特征组合。

参考文献

[1] GALE W A, CHURCH K W. A program for aligning sentences in bilingual corpora[J]. Computational Linguistics, 1993, 19(1):75-102.

[2] 张霞, 咎红英, 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.

[3] KAY M, RSCHEISEN M. Text - translation alignment[J]. Computational Lingus, 1993, 19(1):121-142.

[4] MA Xiaoyi. Champollion: A robust parallel text sentence aligner[C]// Fifth International Conference on Language Resources & Evaluation. Genoa, Italy: dblp, 2006:489-492.

[5] WU Dekai. Aligning a parallel English - Chinese corpus statistically with lexical criteria[C]//Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 1994:80-87.

[6] 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.

[7] 贾善崇,周兰江,张建安,等. 融入多特征的汉-老双语对齐方法[J]. 中国水运(下半月),2020,20(3):78-80.

[8] 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(9):123-132.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[10] 和志强, 杨建, 罗长玲. 基于 BiLSTM 神经网络的特征融合短文本分类算法[J]. 智能计算机与应用, 2019, 9(2):21-27.