文章编号：1003-0077(2019)07-0031-09

基于词对关联网络的句子对齐研究

丁 颖，李军辉，周国栋

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：句子对齐能够为跨语言的自然语言处理任务提供高质量的对齐句子对。受对齐句子对通常包含大量对齐的单词对这种直觉的启发，该文通过探索神经网络框架下词对间的语义相互作用来解决句子对齐问题。特别地，该文提出的词对关联网络通过融合三种相似性度量方法从不同角度来捕获词对之间的语义关系，并进一步融合它们之间的语义关系来确定两个句子是否对齐。在单调和非单调文本上的实验结果表明，该文提出的方法显著提高了句子对齐的性能。

关键词：句子对齐；词对关联网络；神经网络

中图分类号：TP391 **文献标识码**：A

Word-Pair Relevance Network for Sentence Alignment

DING Ying, LI Junhui, ZHOU Guodong

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Sentence alignment provides high quality parallel sentence pairs for cross-language natural language processing tasks. Inspired by the intuition that aligned sentence pairs consists of a large number of aligned word pairs, this paper proposes the sentence alignment method by the semantic interaction between word pairs in neural network framework. In particular, this paper proposes word-pair relevance network, which first captures the semantic interaction between word pairs from different perspectives, then incorporates the semantic interaction to predict whether a sentence pair is aligned or not. Experimental results on monotonic and non-monotonic bitexts show that the proposed approach significantly improves the performance of sentence alignment.

Keywords: sentence alignment; word-pair relevance network; neural network

0 概述

句子对齐旨在从给定的双语文本中提取语义相同的句子对，为构建双语对齐语料库提供技术支持。就自然语言处理领域中的一些应用，如机器翻译[1-3]、跨语言信息检索[4-6]、多语言词汇表征[7]以及构建双语词典[8]等，都需要大规模的平行语料支持。大多数传统的句子对齐方法主要依赖于人工制定的浅层信息特征，并且受语言特性影响。例如，基于句子长度信息[9]的对齐方法适用于同语系的语言对，在印欧语言对上对齐性能较好，但在不同语系语言上对齐性能急剧下降。基于双语词汇[10-11]的方法能更充分地利用双语句对中的词汇信息，从而提高句子的对齐性能。近年来，由于神经网络强大的自主提取特征的能力，基于神经网络的方法也开始应用于句子对齐任务中[12-13]。

然而，现有的基于神经网络的句子对齐方法是基于句子级别建模，首先将每个输入句子映射成为固定长度的向量表示，然后根据这些向量表示来判断句子是否对齐[12]。该方法简单易实现，也获得了一定的对齐性能，表明了使用神经网络进行句子对齐任务是可行的，但该方法仅将源端句子和目标端句子映射成向量表示，不可避免地会丢失很多重要信息，特别是单词级别的对齐信息。

为了克服上述困难，受相互对齐的句子对包含大量相互对齐的单词对这一直觉的启发，本文通过建模词对关联网络的方法来直接捕获细粒度的单词

收稿日期：2018-09-21 **定稿日期**：2018-10-23

基金项目：国家自然科学基金(61401295,61502149)

级信息，然后使用该信息判断句子是否对齐，而不是使用句子级信息。

实验中使用中/英机器翻译数据集和OpenSubtitles2018数据集进行性能评估。实验结果表明，本文提出的基于词对关联网络的句子对齐方法能够较好地提高单调文本和非单调文本的句子对齐性能[14]。

1 相关工作

传统的句子对齐方法主要是基于统计的方法，如Gale和Church[9]提出了基于句子长度统计的方法，Moore[15]采用基于句子长度和自动派生的字典结合起来的方法，Braune等[16]利用Moore对齐模型找到最小最优可能的句子对齐关系，并提出两步聚类的方法判断句子对齐。Ma[17]则提出利用外部词典来计算句对相似性，并为频率较低的单词翻译对赋予较高的权重的方法进行句子对齐。随后，Li等[18]在Ma的基础上，将输入文本切分成更小的文本片段进行句子对齐。

虽然基于半监督或无监督方法的句子对齐已经有了大量研究①，但由于神经网络具有强大的自动学习特征表示能力，基于神经网络的监督方法的研究开始流行起来。Gregoire和Langlais[12]提出通过使用深度学习方法而不是传统的特征工程方法来提取平行句子。该方法借助双向循环神经网络(Bi-RNN)将句子编码成固定大小的向量表示，然后将该向量表示输入到全连接层来计算句子对相互对齐的概率。本文中对比方法aveRNN与上述方法相似。Grover和Mitra[13]首先获得单词对间的相似性分数矩阵，然后将动态池化操作应用于相似性分数矩阵上，最后通过卷积神经网络(CNN)进行分类。

本文提出的词对关联网络方法在一定程度上与神经机器翻译(NMT)中使用的注意力机制类似，两者都是针对源端和目标端双语表示，设计网络计算源端单词与目标端单词之间的对应关系，但两者存在着显著不同。一方面，NMT中的注意力机制使用当前时刻t的目标端状态，分别与源端每个单词的表示计算其对齐概率。该对齐概率也被认为是目标端第t个单词与源端各个单词的对齐概率。但在词对关联网络中，本文采用三种相似度度量方法，直接使用目标端单词的表示与源端单词的表示进行相似度计算。另一方面，在NMT的注意力机制中，模型使用softmax函数限制第t个目标单词与源端每个单词的对齐概率之和为1。而在词对关联网络中，并没有类似限制。同时，本文采用最大池化操作来获取词对相似度中最具信息量的部分。

此外，与本文对词对建模的目的不同的是，其他自然语言处理任务中，如语义文本相似度研究中也广泛研究了词对信息。例如，He和Lin[19]提出了建模词对间相互作用并提出相似性焦点机制来识别重要的对应关系。Wang等[20]在“匹配聚集”框架下提出了双边多视匹配(BiMPM)模型，用于更一般的句子匹配任务。Seo等[21]提出双向注意流来匹配查询和答案对。与上述研究不同的是，本文通过计算单词对的跨语言相似性来进行句子对齐研究。

2 本文方法

本节将描述基于词对关联网络的句子对齐方法，该方法将句子对齐任务看作二分类任务，通过建模词对间相似关系判断句子是否对齐。

2.1 问题描述

单调文本遵循单调性假设，即相互对齐的两个句子在两种语言文本中以相似的顺序出现，一般不出现交叉对齐的情况[22-23]。图1(a)中显示了没有交叉对齐句对的单调对齐。相比之下，非单调文本中相互对齐的句子对通常以不同的顺序出现在文本中，存在任意交叉句对的情况，图1(b)中显示了具有任意交叉对齐句对的非单调对齐。由于非单调文本中1-多/多-1的判断在实际操作时非常复杂，因此，本文假设源端的每个句子只与目标端的一个或零个句子对齐，即1-0/0-1和1-1对齐。

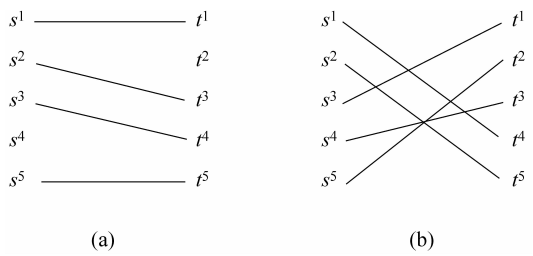

图1 文本对齐类型

句子对齐任务以源文本$X=\{x^1,x^2,\cdots,x^M\}$和目标文本$Y=\{y^1,y^2,\cdots,y^N\}$作为输入，其中M为源文本句子的个数，N为目标文本句子的个数。本

① www.statmt.org/survey/Topic/SentenceAlignment

文将句子对齐任务看作是分类任务，首先通过模型获得句子对齐的概率矩阵 $\boldsymbol{F}\in\boldsymbol{R}^{M\times N}$，其中，$F_{ij}$ 表示源文本 X 中第 i 个句子 x^i 和目标文本 Y 中第 j 个句子 y^j 的对齐概率 L 表示句对 X^i 和 y^j 是否对齐的标签，$L=1$ 表示对齐，$L=0$ 表示不对齐，定义如式(1)所示。

$$F_{ij}=p(l=1\mid x^i,y^j) \tag{1}$$

其中，$p(l=1|x^i,y^j)$ 是句子对 x^i 和 y^j 经过模型计算后输出的对齐概率值(2.2.4 节)。由于在双语文本中寻找对齐句对时，存在源文本中一个句子对应目标文本中多个句子的情况，因此需要根据矩阵 $\boldsymbol{F}$ 的值，进一步获取源文本 X 和目标文本 Y 之间的句子对齐矩阵 $\boldsymbol{A}\in\{0,1\}^{M\times N}$，其中 $A_{ij}=1$ 表示句子对 x^i 和 y^j 相互对齐，反之，$A_{ij}=0$ 表示句子对不对齐。对于单调文本来说，对齐矩阵 $\boldsymbol{A}$ 可以通过动态规划算法[17]获得，并且动态规划算法也适用于识别 1-多/多-1 对齐。对于非单调文本而言，本文使用启发式搜索算法[24]寻找局部最优来获得对齐矩阵 $\boldsymbol{A}$，包括以下两个步骤：

① 在概率矩阵 $\boldsymbol{F}$ 中选择最大非零值 $F_{ij}\geqslant 0.5$，设置 $A_{ij}=1$，表示句子对 x^i 和 y^j 相互对齐，并将 F_{i^*,j^*} $(1\leqslant i^*\leqslant M, 1\leqslant j^*\leqslant N)$ 设置为 0。

② 重复上述步骤直到 F 中所有数值均小于 0.5。

最终，根据对齐矩阵 $\boldsymbol{A}$ 得到最终 1-0/0-1 和 1-1 的句对。

2.2　基于词对关联网络的句子对齐方法

以句子对(“来自 空中 的 战争 威胁”，“war threats from the sky”)为例，图 2 展示了本文提出的基于词对关联网络的句子对齐模型，包括：

- 双向循环神经网络(Bi-directional recurrent neural network (Bi-RNN) layer)，用于对输入的句子进行上下文建模，并作为后续网络层的基础。
- 词对关联网络层(Word-pair relevance network layer)，用于从多个角度捕获词对间语义关系，计算相似性分数；
- 池化层(Max Pooling)，用于获取相似性分数矩阵中最具信息量的部分，并将其重塑为一个向量；
- 多层感知器层(Multi-layer perceptron, MLP)，用于句子分类，其中 1 表示句子对齐，0 表示句子不对齐。

为方便起见，将本文提出的基于词对关联网络的模型简称为 WPRN 模型。以下以句对 (x,y) 为例，分别描述各个网络层的相关细节。其中，源端句子表示为 $x=(x_1,x_2,\cdots,x_m)$，目标端句子表示为 $y=(y_1,y_2,\cdots,y_n)$，m 表示源端句子的单词个数，n 表示目标端句子的单词个数，d_h 表示源端和目标端隐藏状态的大小。

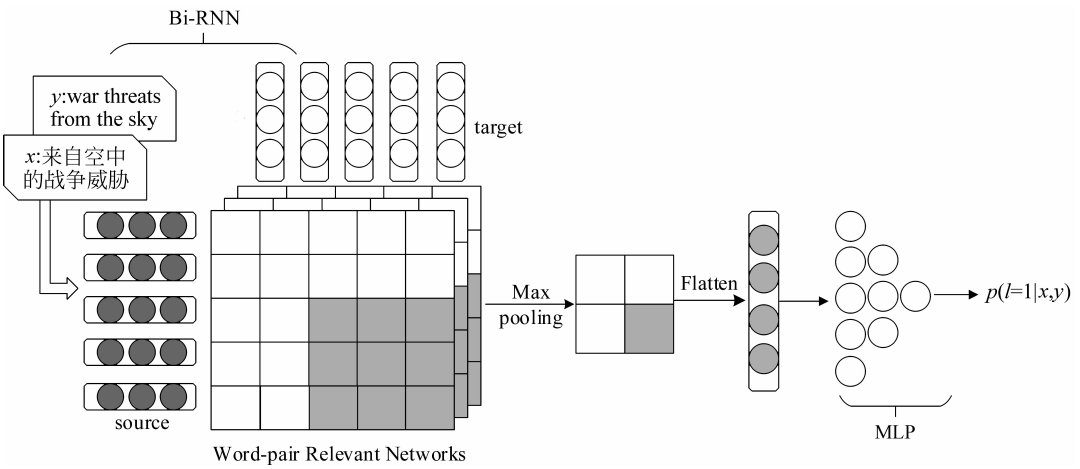

图 2　基于词对关联网络的句子对齐模型结构

2.2.1　双向循环神经网络

本文通过双向循环神经网络(Bi-RNN)对源端和目标端句子进行编码。为了简洁起见，下文仅描述源端句子的编码过程：从左到右读取输入的句子序列 $x=(x_1,x_2,\cdots,x_m)$，并输出其隐藏层状态的前向序列 $(\overrightarrow{h}_1,\overrightarrow{h}_2,\cdots,\overrightarrow{h}_m)$；从右到左读取输入的句子序列并获得其隐藏层状态的后向序列 $(\overleftarrow{h}_1,\overleftarrow{h}_2,\cdots,\overleftarrow{h}_m)$。那么，每个单词 x_i 的隐藏状态 h_i 则通过拼接向量 $\overrightarrow{h}_i$ 和 $\overleftarrow{h}_i$ 而获得。类似地，可以获得目标端每个单词的隐藏状态。

前向RNN网络考虑了单词前面文本的信息，后向RNN网络考虑了单词后面的文本信息，使得经过双向循环神经网络编码的每个单词的隐藏状态均考虑到了整个句子信息。特别地，本文使用相同的Bi-RNN对源端和目标端句子编码。

2.2.2 词对关联网络

现有的基于特征的句子对齐系统表明词对特征是判断句子是否对齐的关键性因素。如图2中例句所示，该句子存在包括(来自，from)、(空中，sky)、(战争，war)、(威胁，threats)等四组对齐的词对，这为判断此句对是对齐的提供了有力证据。受对齐的句子对通常包含大量对齐的单词对这种直觉的启发，本节提出的词对关联网络即通过建模词对间相似度计算的方法来获得句对间的对齐词对信息。

对于给定的源端句子 $x=(x_1,x_2,\cdots,x_m)$ 和目标端句子 $y=(y_1,y_2,\cdots,y_n)$，获得其隐藏状态 $h_x=(h_{x_1},h_{x_2},\cdots,h_{x_m})$ 和 $h_y=(h_{y_1},h_{y_2},\cdots,h_{y_n})$ 之后，词对关联网络从不同角度来计算每个词对 (h_{x_i},h_{y_j}) 的相似度分数。具体来说，该网络使用以下三种方法来计算词对间相似度分数。

① 余弦相似度(cosine)：$\cos(h_{x_i},h_{y_j})$ 通过计算两个向量的夹角余弦值来评估它们的相似度。

② 双线性模型(bilinear model)：定义如式(2)所示。

$$b(h_{x_i},h_{y_j})=h_{x_i}^{T}Mh_{y_j} \tag{2}$$

其中，$M\in R^{d_h\times d_h}$ 是要学习的模型参数[25-26]。双线性模型能够简单有效地捕获两个向量之间的强线性相互作用。

③ 单层神经网络(single layer network (SLN))：定义如式(3)所示。

$$s(h_{x_i},h_{y_j})=u^{\mathrm{T}}f(V[h_{x_i},h_{y_j}]+b) \tag{3}$$

其中，$u\in R^k$，$V\in R^{k\times 2d_h}$，$b\in R^k$ 是要学习的模型参数，f 是非线性激活函数，k 是可以任意设置的超参[27]。单层神经网络可以用来捕获两个向量间的非线性相互作用，同时也可以看作双线性模型的补充。

词对关联网络通过余弦相似度、双线性模型和单层神经网络三种计算相似度的方法从多个角度计算词对间的相似关系，充分考虑了词对间的语义关系，从而得到一个大小为 $3\times m\times n$ 的相似性分数矩阵。

2.2.3 池化层

两个句子之间的相似关系通常是由一些强烈的语义关系决定的。因此，本文采用最大池化策略将词对关联网络得到的相似性分数矩阵，如图2所示，划分为一组非重叠子区域，并获取每个子区域的最大值。假设最大化池的大小为 $3\times k_1\times k_2$，则最终可以获得大小为 $(m/k_1)\times(n/k_2)$ 的最具信息量的相似性分数矩阵。值得注意的是，此处最大池的第一维大小设置为2.2.2节所述的计算相似性方法的个数。也就是说，最大化池的输出与计算相似性方法的数量无关。

最后，将相似性分数矩阵重塑为一个向量，作为多层感知器层的输入。

2.2.4 多层感知器层

多层感知器(MLP)层由两个隐藏层和一个输出层组成。池化层的输出经过两个全连接隐藏层获得了更加抽象的表示，并最终连接到输出层。对于句子对齐任务而言，输出为二分类的分类概率。本文通过预测概率 $p(l=1|x,y)$ 来判断两个句子是否对齐，定义如式(4)所示。

$$p(l=1\mid x,y)=\sigma(z) \tag{4}$$

其中，z 是MLP层的输出，σ 是sigmoid函数。

3 实验

本节中，将使用中/英非单调文本和单调文本来评估WPRN模型抽取平行句对的性能。

3.1 实验设置

3.1.1 数据集

训练数据来源于NIST机器翻译评测数据，由中英平行语料库LDC2003E14、LDC2004T07、LDC2005T06和LDC2005T10等构成，共包含61 968篇文章和1.25MB平行句对，含有27.9MB中文单词和34.5MB英文单词。本文以所有的平行句对作为句子对齐的正例，同时以文档为单位，对于每个源端句子，从目标端随机选择一个句子来获得负例，从而构建相同规模的负例作为训练数据。实验中，使用NIST MT 02数据集作为开发集，保存最优模型。

本文使用两种不同测试集来评估WPRN模型的性能，即NIST MT测试集和OpenSubtitles测试集①。NIST测试集是由NIST MT 03、04、05数据集(分别包含919、1 788、1 082个句子对)人工生成的、领域内的数据集。由于NIST MT数据集最

① https://www.opensubtitles.org/zh

初仅包含 1-1 对齐，为了获得 1-0 和 0-1 对齐，本文从源端随机选择 90 个句子删除，从目标端随机选择 60 个句子删除。需要注意的是，此处随机删除的句子个数可以设为任意数，仅是为了使 NIST MT 测试集中包含 1-0 和 0-1 对齐而设置，随机删除的句子数越多，表明 1-0 和 0-1 对齐的数量也越多。此时，获得 NIST MT 单调测试集。此外，将上述数据集中句子顺序打乱来获得 NIST MT 非单调测试集。表 1 显示了该测试集的统计数据。

表 1　NIST MT 测试集的句子数量及对齐关系统计

Dataset	# Src	# Trg	1-0/0-1	1-1
NIST02	788	818	138	734
NIST03	829	859	144	772
NIST04	1 698	1 728	146	1 640
NIST05	992	1 022	140	937
Total	3 519	3 609	430	3 349

OpenSubtitles(OSs)测试集是来自 OpenSubtitles2018 真实的、领域外数据集。本文从 OpenSubtitles2018 中随机选择 8 篇中英文文档作为另一个测试集，该测试集为单调文本，包含 1-0/0-1、1-1、1-2/2-1 和 1-3/3-1 对齐。表 2 显示了该测试集的统计数据。

表 2　OpenSubtitles(OSs)测试集的句子数量及对齐关系统计

Dataset	# Src	# Trg	1-0/0-1	1-1	1-2/2-1	1-3/3-1
OSs	4 225	3 806	207	2 840	576	104

3.1.2　模型设置

根据训练语料分别选择词频最高的前 30K 个单词作为源端词表和目标端词表，分别占总词汇量的 98.4%和 99.0%。对于所有不在词表中的单词，将其统一映射到特殊标记 UNK 上。为了初始化本文模型的词向量，使用由 Zou 等[28]提供的 50 维预先训练好的中英双语词向量，并在训练过程中更新词向量。

为了有效地训练神经网络模型，源端和目标端句子长度被限制在 50 之内。在 Bi-RNN 层，本文使用 GRU[29]作为 RNN 的激活函数，并设置其隐藏状态大小为 150。此时，词对关联网络层将输出大小为 3×50×50 的相似性分数矩阵。其中，设置单层神经网络(SLN)中 k 值为 2，非线性激活函数 f 为 tanh。在 Max Pooling 层，设置最大化池大小为 3×3×3，并获得维度为 289 的向量(即(50/3)×(50/3))。最后，在 MLP 层中，隐藏层大小分别设置为 2 000 和 1 000。

在训练过程中，使用 AdaDelta[30]来优化模型参数，其中 $\varepsilon=10^{-6}$、$\rho=-0.95$。对于除了词向量外的所有模型参数，使用[−0.1，0.1]中的均匀分布来随机初始化它们。此外，dropout 设置为 0.5，批处理大小设置为 80。模型基于 Theano 深度学习框架开发，在训练集上进行 10 轮迭代，使用单个 GeForce GTX 1080 GPU，需要约 24h。

3.1.3　模型训练

给定训练句对 $(X,Y)=\{x^i,y^i\mid 1\leqslant i\leqslant N\}$ 和它们的真实标签 $L=\{l^i\mid l^i\in\{0,1\}\}$，训练目标是最小化预测结果 $\hat{P}=\{\hat{p}^i\}$ 和真实标签的交叉熵，定义如式(5)所示，其中，Θ 表示模型中所有参数的简写。

$$L(X,Y,L;\Theta)=\sum_{i=1}^{N}l^i\log(\hat{p}^i(x^i,y^i;\Theta)) \quad (5)$$

3.1.4　Baseline 系统

与本文提出的 WPRN 模型进行比较的 Baseline 系统如下：

① aveRNN：使用 Bi-RNN 对源端和目标端句子进行编码，取其所有单词隐藏状态的平均值作为句子级别的向量表示。最后将两个源端和目标端句子的向量表示拼接后输入到 MLP 层进行句子对齐分类。

② attRNN：该模型类似于 aveRNN，区别在于使用结构化注意机制[31]的方法获得句子级别的表示①。

③ G&M(2017)：该模型在本文实验数据的基础上重现了 Grover 和 Mitra[13]提出的方法，使用余弦相似度计算相似性分数矩阵，并使用动态池化将其映射到固定维度上，然后使用卷积神经网络(CNN)进行分类。

④ NMT：该系统使用神经机器翻译(NMT)方法获得在给定源端句子 x^i 时，翻译为目标句子 y^j 的概率，即对源端句子 x^i 进行强制解码，获得目标译文 y^j 的分数②。本文使用基于注意力机制的 NMT 系统[2]，并使用与本文相同的实验数据集来训练该 NMT 模型。

① 本文在重现 Lin 等[31]的方法时设置 $r=4$。

② 为了避免长翻译倾向于具有低翻译概率的问题，本文定义如 Wu 等[32]的评分函数。

上述 Baseline 系统中，aveRNN 和 attRNN 是基于句子级向量表示的方法，G&M(2017)是基于词对相似度建模的方法，最后一个是利用 NMT 技术的翻译方法。

3.1.5 评估指标

为了评估模型的性能，本文采用精确度(P)、召回率(R)和 F_1 值作为实验的评估指标。同时，计算精确度、召回率和 F_1 值的平均性能(Micro-P/R/F_1)来评估所有对齐的总体性能。

3.2 实验结果

3.2.1 非单调文本句子对齐

表 3 给出了 NIST MT 非单调测试集的句子对齐性能，从中可以发现：

表 3 NIST MT 非单调测试集的句子对齐性能

	1-0/0-1			1-1			Micro		
	P	R	F_1	P	R	F_1	P	R	F_1
aveRNN	23.2	24.4	23.8	61.7	61.5	61.6	57.1	57.3	57.2
attRNN	38.0	58.1	46.0	84.1	81.2	82.6	76.3	78.8	77.4
G&M(2017)	34.5	76.7	47.6	92.4	85.1	88.6	78.7	84.2	81.3
NMT	12.2	2.6	4.2	88.9	93.5	91.1	87.0	83.1	85.0
WPRN	**66.4**	**83.0**	**73.8**	**97.0**	**95.4**	**96.2**	**92.7**	**94.0**	**93.3**
Moore	5.6	62.4	10.2	0.3	0.1	0.2	4.6	7.2	5.6
Gargantua	6.0	36.0	10.3	0.2	0.1	0.1	3.3	4.2	3.7
Champollion	6.3	39.3	10.9	1.1	0.8	0.9	4.0	5.1	4.5

首先，细粒度的词级信息是判断句子是否对齐的重要信息。WPRN 模型的整体 F_1 值分别高出 aveRNN、attRNN 和 NMT 方法 36.1、15.9、和 8.3 的 F_1 值，性能明显优于基于句子级向量表示的方法和基于机器翻译的方法，表明将整个句子信息仅仅汇入一个句子级向量是不够的，这容易造成句子语义间重要信息的丢失。同时，由于 NMT 模型训练的目标是最大化整个目标句子的翻译概率，并没有强调词对之间的翻译，因此往往容易造成词的错翻；而 WPRN 模型的目的在于获取双语句对的词对间相关性，三种相似度计算方法可以从三种不同角度来建模词对关系，可以在一定程度上缓解 NMT 模型对齐错误的现象。

其次，WPRN 模型性能高出 G&M(2017)模型 12.0 F_1 值，说明了使用不同相似度计算方法来从不同角度捕获词对间语义关系优于仅使用一种相似度计算方法。

另外，从表 3 中还可以发现 1-0/0-1 对齐比 1-1 对齐更难识别，这也与 Quan[14] 中实验结果保持了一致。1-0/0-1 对齐性能较低的一个重要原因在于本文的预测模型是判断两个句子是否对齐，并不是为了专门预测某个句子在另一端是否存在对齐句子。

最后，本文同时与三个现有的句子对齐工具进行比较。Moore① 是一种基于句子长度和自动派生的双语词典的句子对齐工具[15]；Gargantua② 是用于对称和非对称平行语料库的无监督句子对齐工具[16]；Champollion③ 是基于词典的句子对齐工具，为潜在噪声的平行文本而设计[17]。表 3 中的最后三行比较了它们在非单调文本上的对齐性能，从中可以发现这些句子对齐工具的性能受非单调性的影响非常严重，并且不适用于非单调的双语文本。

3.2.2 单调文本句子对齐

本文提出的句子对齐方法同样可以在单调文本上取得优越的性能。如 2.1 节所述，本文采用动态规划方法[17] 来获得最优对齐结果。表 4 给出了 NIST MT 单调测试集的性能。同时，为了验证 WPRN 模型的可行性，本文也在跨领域的、真实的 OpenSubtitles 测试集上进行了评估，对齐性能如表 5 所示。

① https://www.dssz.com/905003.html
② https://github.com/braunefe/Gargantua
③ http://champollion.sourceforge.net

表 4　NIST MT 单调测试集的句子对齐性能

	1-0/0-1			1-1			Micro		
	P	R	F_1	P	R	F_1	P	R	F_1
WPRN	**100.0**	**92.1**	**95.9**	**99.5**	**100.0**	**99.8**	**99.6**	**99.1**	**99.3**
Moore	53.8	89.3	67.1	98.8	94.6	96.6	90.6	94.0	92.3
Gargantua	43.5	79.8	56.3	97.0	91.9	94.4	86.4	90.5	88.4
Champollion	32.7	59.5	42.2	91.1	86.3	88.7	79.6	83.3	81.4

表 5　OpenSubtitles 测试集的句子对齐性能

	1-0/0-1			1-1			1-2/2-1			1-3/3-1			Micro		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
WPRN	**23.8**	**39.1**	**29.6**	**84.8**	**86.1**	**85.4**	**59.9**	**55.0**	**57.4**	**64.3**	**51.9**	**57.5**	**75.5**	**77.7**	**76.6**
Moore	5.0	8.1	9.5	76.2	65.6	70.5	—	—	—	—	—	—	35.1	54.5	42.7
Gargantua	8.9	27.1	13.4	67.8	75.6	71.5	56.1	30.2	39.3	61.3	18.3	28.1	57.9	64.3	61.0
Champollion	3.2	5.3	4.0	52.2	48.3	50.2	20.7	8.7	12.2	8.8	12.5	10.4	40.3	38.8	39.5

从表 4 和表 5 可以看出，使用词对关联网络模型：

首先，OpenSubtitles 测试集上的句子对齐比 NIST MT 测试集的句子对齐更具挑战性。一方面，前者是跨领域数据集，而后者是领域内数据集；另一方面，OpenSubtitles 测试集包含 1-N/N-1(N>1)的情况，句子对齐情况相对复杂。

其次，在不同类型的对齐中，1-0/0-1 对齐是最难识别的，其次是 1-2/2-1 和 3-1/1-3 对齐，而 1-1 对齐相对简单。

对比表 4 和表 5 中 WPRN 和其他三种方法的性能，可以发现虽然现有的对齐工具都取得了良好的对齐性能，但 WPRN 模型在 NIST MT 测试集上取得了高达 99.3 的 F_1 值，在 OpenSubtitles 测试集上取得了 76.6 的 F_1 值，明显优于其他三种对齐工具，表明了监督学习有利于提高句子对齐性能，并且 WPRN 模型能够较好地捕获句子之间的语义关系。

3.3　实验分析

3.3.1　部分非单调文本句子对齐

由于完全非单调双语文本在实际应用中很少见，而部分非单调双语文本却不是。本文根据 Quan[14] 的分析方法来说明本文提出 WPRN 模型的实用性。本文通过随机打乱测试数据集中 0%、10%、20%、40%、60%、80%、100%的句子来构建七个版本测试集，测试 WPRN 方法的性能，如图 3 所示。理论上，文本的非单调性比例对本文方法的性能没有影响，因此，对于任意比例的非单调性文本应有相同的性能，图 3 也表明了这一点。从图 3 中还可以发现 Moore、Gargantua 和 Champollion 在近似单调文本上表现出良好的性能，但是当非单调比例增加时，它们的性能显著下降。此外，当文本几乎完全单调时，WPRN 模型的性能也优于其他方法，说明本文的方法适用范围更加广泛。

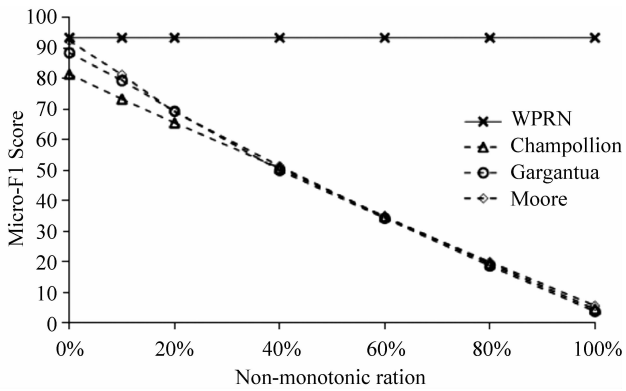

图 3　WPRN 模型在不同程度的非单调文本上的性能

3.3.2　实例分析

本文以相同的源端句子和不同的目标端句子为例，来说明 WPRM 模型相比于其他模型的优越性。如表 6 所示，其中(Source，Target1)不是相互对齐的句对，而基于句子级向量(如 aveRNN 和 attRNN)的方法将它们预测为高概率(即对齐概率 0.98 和 0.81)的对齐句对。另外，虽然 C&M

(2017)方法也预测出它们不是对齐句对,但WPRN模型给出了更接近于0的对齐概率。(Source, Target2)是相互对齐的句对,所有方法都正确地将其预测为对齐句对,但概率值之间存在相当大的差异。例如,aveRNN获得了0.91的概率值,甚至低于与Target1对齐的概率,说明基于句子级向量的方法虽然能进行句子对齐任务,但性能并不理想;C&M(2017)虽然获得了0.67的概率值,但在最终对齐策略中很容易将其判断为不对齐句对;而WPRN获得了几乎100%的概率。同时,从(Source, Target2)句对中可以发现(鲍尔, powell)、(视察, inspect)、(泰国, Thailand)、(印尼, Indonesia)、(斯里兰卡, lanka)、(灾情, disaster)、(高峰, summit)等许多一一对应的同义词对,而(Source, Target1)句对中几乎没有同义词对,WPRN分别给出了0.05和0.999 6的对齐预测概率,准确地判断两个句对之间的对齐关系。上述两个例子对比表明了词级信息是判断句子对齐的重要信息,WPRN模型能够较好地捕获词对间的语义关系,从而能够更准确地判断句对之间关系。

表6 模型预测句子对齐的对比实例

Source	鲍尔 将 率 团 视察 泰国 、印尼 与 斯里兰卡 的 灾情 ，再 参与 六日 的 高峰 会议 。
Target1	china's npc delegation led by li tieying will also go to uruguay and brazil for goodwill visits .
Prob1	aveRNN: 0.98; attRNN: 0.81; C&M(2017): 0.21; WPRN: 0.05
Target2	powell will head the delegation to inspect the disaster in thailand, indonesia and sri lanka before attend-ing the summit on the 6th .
Prob2	aveRNN: 0.91; attRNN: 0.86; C&M(2017): 0.67; WPRN: 0.9996

4 总结

本文提出了一种基于词对关联网络(WPRN)的句子对齐方法,该方法的主要特点是采用不同的相似性度量方法从不同的角度捕捉单词对的语义交互信息。在非单调和单调文本上的实验结果表明,对词对间的相似性建模是进行句子对齐任务至关重要的步骤,词对关联网络能够有效准确地捕获词对间的语义信息。在将来的工作中,将进一步改进现有的WPRN模型,同时探索研究从可比语料库中提取平行句对、提高跨领域文本的句子对齐性能等,这将更具挑战性。

参考文献

[1] Stephan Vogel, Alicia Tribble. Improving statistical machine translation for a speech-to-speech translation task[C]//Proceedings of ICSLP 2002, 2002: 1901-1904.

[2] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate[C]//Proceedings of ICLR 2015, 2015.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need[C]//Proceedings of NIPS 2017, 2017.

[4] Jian-Yun Nie, Michel Simard, Pierre Isabelle, et al. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web[C]//Proceedings of SIGIR 1999, 1999: 74-81.

[5] Wessel Kraaij, Jian-Yun Nie, Michel Simard. Embedding web-based statistical translation models in cross-language information retrieval[J]. Computational Linguistics, 2003, 29(3): 381-419.

[6] Giovanni Da San Martinom, Salvatore Romeo, Alberto Barroon-Cedeno, et al. Cross-language question re-ranking[C]//Proceedings of SIGIR 2017, 2017: 1145-1148.

[7] Karl Moritz Hermann, Phil Blunsom. Multilingual models for compositional distributed semantics[C]//Proceedings of ACL 2014, 2014: 58-68.

[8] Judith Klavans, Evelyne Tzoukermann. The BICORD System: Combining lexical information from bilingual corpora and machine readable dictionaries[J]. Computational Linguistics, 1990, 62(4): 174-179.

[9] William A Gale, Kenneth W Church. A program for aligning sentences in bilingual corpora[C]//Proceedings of ACL 1991, 1991: 177-184.

[10] Martin Kay, Martin Roscheisen. Text-Translation Alignment[J]. Computational Linguistics, 1993, 19(1): 121-142.

[11] 刘昕, 周明, 朱胜火, 等. 基于自动抽取词汇信息的双语句子对齐[J]. 计算机学报, 1998, 21(s1): 151-158.

[12] Francis Gregoire, Philippe Langlais. A deep neural network approach to parallel sentence extraction[J]. avXiv preprint arXiv: 1709.09783v1,2017.

[13] Jeenu Grover, Pabitra Mitra. Bilingual word embeddings with bucketed CNN for parallel sentence extraction[C]//Proceedings of ACL: Student Research Workshop 2017, 2017: 11-16.

[14] Xiaojun Quan, Chunyu Kit, Yan Song. Non-monotonic sentence alignment via semisupervised learning [C]//Proceedings of ACL 2013. 2013: 622-630.

[15] Robert C Moore. Fast and accurate sentence alignment of bilingual corpora[C]//Proceedings of AMTA 2012, 2012: 135-144.

[16] Fabienne Braune, Alexander Fraser. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora [C]//Proceedings of COLING 2010, 2010: 81-89.

[17] Xiaoyi Ma. Champollion: A robust parallel text sentence aligner[C]//Proceedings of International Conference on Language Resources and Evaluation, 2006: 489-492.

[18] Peng Li, Maosong Sun, Ping Xue. Fast-champollion: A fast and robust sentence alignment algorithm[C]//Proceedings of COLING 2010, 2010: 710-718.

[19] Hua He, Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement [C]//Proceedings of NAACL 2016, 2016: 937-948.

[20] Zhiguo Wang, Wael Hamza, Radu Florian. Bilateral multi-perspective matching for natural language sentences[C]//Proceedings of IJCAI 2017, 2017: 4144-4150.

[21] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, et al. Bidirectional attention flow for machine comprehension[C]//Proceedings of ICLR 2015, 2015.

[22] Philippe Langlais, Michel Simard, Jean Véronis. Methods and practical issues in evaluating alignment techniques[C]//Proceedings of COLING-ACL 1998, 1998: 711-717.

[23] Dekai Wu. Alignment[M]. Handbook of Natural Language Processing, CRC Press. 2010: 367-408.

[24] Chunyu Kit, Jonathan J Webster, King Kui Sin, et al. Clause alignment for Hong Kong legal texts: A lexical-based approach[J]. International Journal of Corpus Linguistics, 2004: 9(1): 29-52.

[25] Ilya Sutskever, Ruslan Salakhutdinov, Joshua B Tenenbaum. Modelling relational data using Bayesian clustered tensor factorization [C]//Proceedings of NIPS 2009, 2009: 1821-1828.

[26] Rodolphe Jenatton, Nicolas Le Roux, Antoine Bordes, et al. A latent factor model for highly multi-relational data[C]//Proceedings of NIPS 2012, 2012: 3167-3175.

[27] Ronan Collobert, Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning [C]//Proceedings of ICML 2008, 2008: 160-167.

[28] Will Y Zou, Richard Socher, Daniel Cer, et al. Bilingual word embeddings for phrase-based machine translation[C]//Proceedings of EMNLP 2013, 2013: 1393-1398.

[29] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of EMNLP 2014, 2014: 1724-1734.

[30] Matthew D Zeiler. ADADELTA: An Adaptive learning rate method[J]. arXiv preprint arXiv: 1212.5701, 2012.

[31] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, et al. A structured self-attentive sentence embedding[C]//Proceedings of ICLR 2017, 2017.

[32] Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. avXiv preprint arXiv: 1609.08144v2, 2016.

丁颖(1994—)，硕士研究生，主要研究领域为句子对齐、机器翻译。
E-mail: 20165227025@stu.suda.edu.cn

周国栋(1967—)，博士，教授，主要研究领域为自然语言处理。
E-mail: gdzhou@suda.edu.cn

李军辉(1983—)，博士，副教授，主要研究领域为自然语言处理、机器翻译。
E-mail: lijunhui26@gmail.com