文章编号：2095-2163(2021)09-0042-06 中图分类号：TP391 文献标志码：A

融入双语词向量的韩汉名词短语对齐方法研究

刘晨阳，赵天锐
（信息工程大学洛阳校区，河南 洛阳 471000）

摘 要：针对传统短语对齐方法依赖外部资源，且较少涉及平行句对内在特征的问题，提出了融入双语词向量的韩汉名词短语对齐方法。利用平行语料，分别训练单语词向量再进行跨语言映射得到双语词向量，并构建了基于短语构成规律的短语抽取和融入双语词向量、短语长度和词性相似度的短语对齐模型。实验结果证明，融入韩汉双语词向量，能更有效地提取短语特征从而实现短语对齐。

关键词：韩语-汉语；名词短语对齐；双语词向量；平行语料库

Research on Korean Chinese noun phrase alignment method integrating bilingual word vector

LIU Chenyang, ZHAO Tianrui
(Luoyang Campus of Information Engineering University, Luoyang Henan 471000, China)

【Abstract】 Aiming at the problem that traditional phrase alignment methods rely on external resources and rarely involve the internal characteristics of parallel sentences, a Korean-Chinese noun phrase alignment method integrating bilingual word vector is proposed. After training monolingual word vectors with parallel corpus, bilingual word vectors are obtained by cross language mapping, and a phrase alignment model based on phrase composition rule is constructed, which integrates bilingual word vectors, phrase length and part of speech similarity. Experiments on Korean-Chinese parallel corpus in the field of politics and diplomacy show that phrase features can be extracted more effectively and phrase alignment can be realized by integrating Korean-Chinese bilingual word vector.

【Key words】 Korean-Chinese; noun phrase alignment; bilingual word embedding; parallel corpus

0 引 言

随着国际互联网的迅速发展，信息资源愈发呈现大规模、多语言的特征。在自然语言处理领域，以双语（或多语）平行语料库为基础的应用日益增多。如，机器翻译、词典编撰、语义消歧、跨语言信息检索等。其中，平行语料库对应单位的抽取对齐，是实现这些应用的关键技术之一。对应单位是对应源文本和目的文本中可识别的对应文本块或片段，是意义对应完整并具有清晰边界的任何片段或序列[1]。其中短语便是客观存在于平行句对之中的一种对应单位，主要表现形式为互译的多词组合。本文针对韩汉平行句对中的对齐名词短语进行抽取，构建了基于短语构成规律的短语抽取与融入双语词向量、短语长度和词性相似度的短语对齐模型，并在政治外交领域的韩汉平行语料上进行相关实验测评。其成果能广泛应用于翻译研究、语言教学、术语词典编纂和政治外交话语研究等领域，其采用的方法也可为相关研究提供参考和思路。

1 研究现状

双语短语对齐研究的基础是双语词对齐[2-3]，其原理是词语相似度的计算。词组由词构成，词对齐的部分技术方法也可迁移至短语对齐上，其关键点在于如何将词的相似度转换为短语的相似度。关于短语对齐现有研究的主流方法是先进行单语短语抽取，再进行对齐。对齐的方法有基于词典的、基于统计或二者结合的方法。

文献[4]提出了基于规则和基于统计相结合的方法，对中英文句对分类，进行句法分析后提取短语，再利用最大熵排序模型，从候选对齐句对中选取最佳结果；文献[5]基于中英平行专利语料库，使用短语对齐和组块分析技术，并借助专利语料的领域主题信息，实现了中英专利术语的高效自动抽取；文献[6]基于俄汉政治外交平行语料库，按照俄汉短语词性构成模式，使用规则获得短语，并构建了短语

作者简介：刘晨阳（1996-），男，硕士研究生，主要研究方向：计算语言学、自然语言处理；赵天锐（1997-），男，硕士研究生，主要研究方向：语言学（韩国语）、计算语言学。

收稿日期：2021-07-24

长度、词典、机器翻译三维评估模型，实现了俄汉短语单位的自动对齐。文献[7]先采用基于统计与词典融合的词对齐方法获得了韩国语-汉语的词对齐文件，再跟据韩国语名词短语结构特点抽取短语，获取词对齐文件中每个韩国语词语对应的汉语位置，最终根据卡方过滤得出匹配的名词短语对。

综上所述，短语对齐的技术多为传统方法。此类方法忽略了平行语料的内在语义特征，且依靠大量的语言学先验知识，面对低资源、小语种语言时效果欠佳。随着深度学习、神经网络的发展，词向量作为词的一种分布式表示，开始在自然语言处理领域崭露头角。词向量以原始语料作为训练集，无需外部资源便能高效地表征句法语义关系，为对应单位的相似度计算与对齐提供了新思路。

文献[8]基于英汉平行语料库，利用双向长短期记忆神经网络提取词向量，结合依存关系得到词对齐特征，并在此基础上实现了基于短语的统计机器翻译系统。文献[9]基于汉维医疗平行语料库，运用自训练的汉维双语词向量，深入词的语义一级进行双语医学术语抽取，取得了不错的效果。文献[10-11]将英语作为中间语言，通过建立对应单字的上下文向量，实现了韩法双语间的名词短语对齐，并对实验结果进行了误差分析。

由此可见，此前对短语抽取与对齐的研究中，多使用传统的方法且对词典等外部资源的依赖较多，运用神经网络语言模型且面向韩汉双语领域的研究较少。因此，将双语词向量应用于韩汉双语短语对齐相关技术，有很强的研究意义和应用价值。

2　韩语、汉语名词短语结构特点

进行短语对齐首先要进行短语的抽取，短语的构成规则与语言本身的特性息息相关。韩语属于黏着语，通过助词和词尾变化实现语法功能；汉语属于孤立语，不依赖内、外部屈折的形态变化。本文通过总结归纳韩语、汉语名词短语的结构特点，基于词性标注结果抽取相应短语。

针对韩语，采用文献[12]中基于左右边界规则获取韩国语名词短语方法总结归纳出的名词短语类型进行短语抽取；在标注工具上，使用韩国蔚山大学开发的形态素分析器 UTagger[13] 进行词性标注；UTagger 的训练基于“韩国 21 世纪世宗计划语料库”，并沿用其标注体系，支持增量训练从而不断提升分析能力。针对汉语，采用百度自然语言处理部研发的中文联合词法分析工具 LAC[14]（Lexical Analysis of Chinese）进行词性标注并沿用其标注体系。LAC 通过深度学习模型，联合学习分词、词性标注、专名识别任务以及词语重要性，整体效果 $F1$ 值超过 0.91，词性标注 $F1$ 值超过 0.94，专名识别 $F1$ 值超过 0.85。为了明晰名词短语结构从而进行短语抽取，将韩汉两种标注体系中的部分标签按规则进行统一。其规则，见表 1。

表 1　韩语、汉语词性标签对应

Tab. 1　Correspondence between Korean and Chinese tags

标签	“世宗计划”标注体系（韩语）	LAC 标注体系（汉语）
NN	NNG（普通名词）	n（普通名词）
	NNP（固有名词）	vn（名动词）
	NP（代词）	an（名形词）
	NNB（依存名词）	s（处所名词）
	NR（数词）	
VA	VA（形容词）	a（形容词）
VV	VV（动词）	v（动词）

为了进一步挖掘政治外交领域名词短语结构特点，从中国外文局、中国翻译研究院主持建设的“中国特色话语对外翻译标准化术语库”中获取了 3 000对中韩互译术语，对其进行分词与词性标注后进行相关统计，结合韩语汉语各自语法特点，归纳总结出了 12 种韩语名词短语和 10 种汉语名词短语结构，并给出了部分示例，见表 2、表 3。

表 2　韩语名词短语构成模式及部分示例

Tab. 2　Korean noun phrase formation patterns and some examples

构成模式	示例
NN+(XSN/JKG)+NN	사회/발전社会发展
JJ+ETM+NN	낡/은/차량老旧车辆
NN+(XSN/JKG)+NN+NN	혼합/소유/제/경제混合所有制经济
JJ+(ETM)+NN+NN	밝/은/발전/전망光明的发展前景
NN+(XSN/JKG)+NN+(XSN/JKG)+NN+NN	국내/경제/하락/의/압력国内经济下行压力

表 3　汉语名词构成模式及部分示例

Tab. 3　Chinese noun formation patterns and some examples

构成模式	示例
NN+NN	网络安全、贸易摩擦
JJ+NN	合理区间、良好开端
NN+NN+NN	国家领土主权
JJ+NN+NN	重要领域改革
NN+JJ+NN	领导人非正式会议

在计算词性相似度时，由于中韩两种语言构词法存在差异，所以要对标注词项较多的韩语标注进行调整。根据世宗标注体系，“ETM”、“XSN”分别

是冠形词词尾、名词派生接尾词的标记,即“ETM”代表该位置是“-ㄴ/-은/-는”,词尾并无具体含义,且在中文缺少对应成分,可以对其进行省略。而语料中的名词短语内部也常常出现“XSN”,如 일인당 가처분소(NN/XSN/NN),而中文分词结果不会将词缀和原词拆分。为使词性相似度的计算更加准确,将接尾词和词干拼接,并将词性中的“XSN”省略,如일인당/가처분소득(NN/NN)。

3 融入双语词向量的韩汉名词短语对齐方法

3.1 韩汉双语词向量

词向量(Word Embedding),又称词嵌入,是一种词的分布式表示。通过将词映射至低维空间上,来表征词的句法和语义关系。文献[16]于 2013 年提出了由 NNLM[15](神经网络语言模型)改进而来的 Word2Vec 算法。其中包含了连续词袋模型(Continuous bag-of-words, CBOW)和跳字模型(Skip-Gram)。CBOW 模型的原理是根据上下文预测当前词;Skip-Gram 模型则是根据中心词预测周围的词,并使用梯度下降算法不断调整中心词的词向量。Skip-Gram 的训练特点使其在规模较小的数据集上有更好的表现。因此,选取 Skip-Gram 模型用以训练词向量。

目前,词向量的训练多针对单一语言,即单语词向量,用以表示该语言中词汇之间的句法语义关系。跨语言词向量(Cross-lingual word embedding)[17]是单语词向量的一种自然扩展,面向双语时也称为双语词向量(Bilingual word embedding)。其认为在不同语言中具有相似概念的词,在向量空间中的词向量十分接近[18]。文献[19-20]发现两种语言的单语词向量在向量空间中存在近似同态性,因此可以对多(双)语的单语词向量映射到一个共享的低维空间,在不同语言间进行知识转移,从而在多语言环境下对词义进行准确捕捉。如图 1 所示,韩汉相关词语在进行降维并映射至同一向量空间后,互译的双语词语呈现出相似的分布。因此,使用韩汉平行语料训练单语词向量,能够获取互译词语间的内在语义特征用于短语的对齐。

本文采用文献[19]提出的跨语言映射方法,该方法通过无监督初始化与自学习的方式,无需借助种子词典即可将单语种语料通过线性变换映射到共享空间中,实现该方法的主要步骤如下:

3.1.1 完全无监督初始化

设:$\boldsymbol{X}$、$\boldsymbol{Z}$ 分别为韩汉单语词向量矩阵,$\boldsymbol{M}_x = XX^T$、$\boldsymbol{M}_z = ZZ^{\mathrm{T}}$ 分别为韩汉相似度矩阵。通过对 $\boldsymbol{M}_x$、$\boldsymbol{M}_z$ 每行的值进行排序,通过最邻近匹配找到互译词,从而生成初始词典 D。

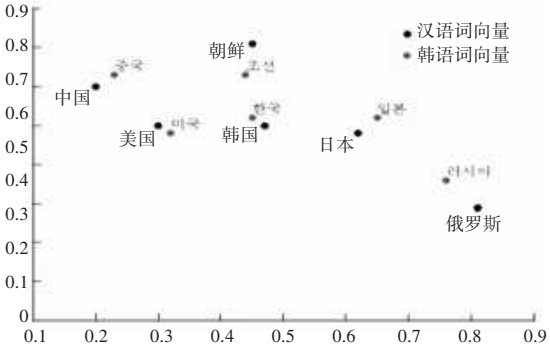

图 1 双语词向量降维、映射至同一向量空间

Fig. 1 Bilingual word vector dimensionality reduction and mapping to the same vector space

3.1.2 鲁棒自学习

首先通过计算最佳正交映射以最大化当前词典 D 的相似性,如式(1)所示。

$$\underset{\boldsymbol{W}_X,\boldsymbol{W}_Z}{\operatorname{argmax}} \sum_i \sum_j D_{ij}((X_{i*}\ \boldsymbol{W}_X)\cdot(Z_{j*}\ \boldsymbol{W}_Z)) \quad (1)$$

其中,$\boldsymbol{W}_x$、$\boldsymbol{W}_z$ 为线性变换矩阵;$\boldsymbol{W}_{i*}$、$\boldsymbol{W}_{j*}$ 分别表示第 i、j 个单词各自的词向量;$\boldsymbol{D}_{ij}$ 为初始词典编码而成的稀疏矩阵,当 $\boldsymbol{D}_{ij}=1$ 时表示韩语中第 i 个单词与汉语中第 j 个单词互译。

在映射嵌入的相似性矩阵 $\boldsymbol{X}\boldsymbol{W}_Z\boldsymbol{W}_z^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}$ 中使用汉语到韩语的最邻近检索,为每个汉语单词分配了韩语中最接近的单词,将映射的汉语嵌入和韩语嵌入之间的点积用作相似度度量。

即 $j=\operatorname{argmax} k(X_{i*}\ \boldsymbol{W}_X)\ (Z_k,\boldsymbol{W}_Z)$ 时,$D_{ij}=1$ 否则 $D_{ij}=0$。

3.1.3 对称重加权

对两种语言对称地应用重加权,可以使映射方向中立,从而获得更好的效果。给定 $\boldsymbol{X}$ 的奇异值分解 $USV^{\mathrm{T}}=X^{\mathrm{T}}DZ$,使 $\boldsymbol{W}_X=US^{1/2}$、$\boldsymbol{W}_z=VS^{1/2}$,即获得两种语言的映射矩阵。

韩汉单语词向量进行映射嵌入的训练过程如图 2 所示。

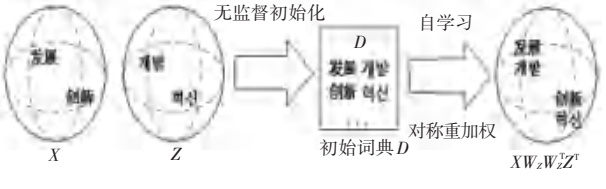

图 2 韩汉双语词向量训练过程

Fig. 2 Vector training process of Korean Chinese bilingual words

3.2　韩汉双语短语长度、词性相似度

基于长度的方法最初应用在句对齐领域，最初由文献[21]提出。其依据是源语言与译文文本长度具有关联性，并多以字节、字符或词数作为长度计量单位。之后的研究者又将句子所含的词性等元素加入，用以计算句子长度。如文献[22]中将句子所含的动词、名词、形容词等词语作为句长计量单位，在英汉句对齐任务上取得了良好的效果。同样互译的短语在长度和词性构成上也具有一定的关联性。

本文以构成短语的字符作为短语长度计量单位，以构成短语词的词性匹配数量，用以计算短语相似度，对先期获得的 3 000 对互译短语随机打乱顺序，进行定量统计，见表 4。

表 4　对齐与非对齐短语相关特征

Fig. 4　Features related to aligned and non-aligned phrases

	对齐短语	非对齐短语
短语字数比标准差	0.33	1.17
词性相似度标准差	0.15	0.58

由此可以看出，两种特征在一定程度上对于短语是否对齐有一定的区分度。但由于短语的自身特性，当抽取出的候选短语过多时，短语长度相似度和词性相似度就难以对其进行区分，此时就要从深层语义出发获取短语的内在特征。

3.3　融入双语词向量的韩汉名词短语对齐模型

融入双语词向量的韩汉名词短语对齐模型如图 3 所示。主要由短语抽取、短语对齐、相似度排序评估 3 部分组成。

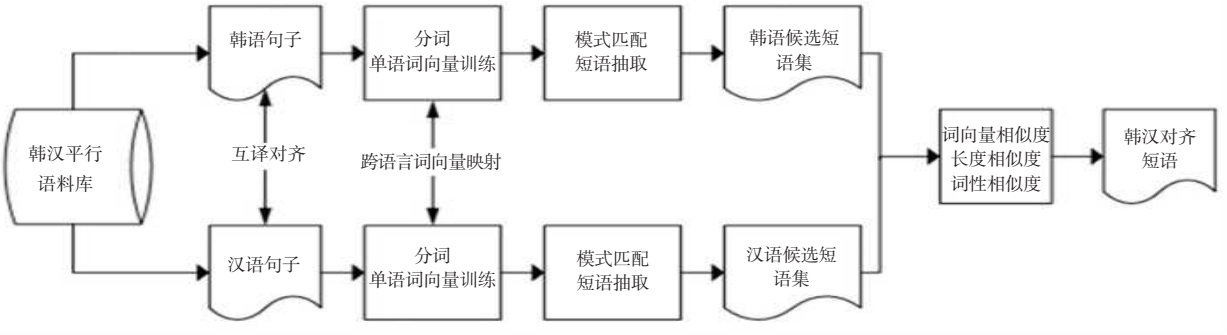

图 3　融入双语词向量的韩汉短语对齐模型

Fig. 3　Korean Chinese phrase alignment model with bilingual word vector

(1)短语抽取：对双语平行语料进行分词和词性标注。分词结果用于训练单语词向量并进行跨语言映射，词性标注结果基于韩汉短语构成规律进行短语抽取，形成短语集。

(2)短语对齐：将韩汉名词短语的词向量相似度、短语长度相似度与短语词性相似度进行加权求和，形成短语相似度。

(3)对候选韩汉名词进行相似度排序评估，根据匹配结果得到韩汉名词短语对齐集。

定义汉语短语 P_{zh}，由 m 个词组成。每个词为 $X_{i(i=1,2,\cdots,m)}$，则有 $P_{zh}=(x_1,x_2,\cdots,x_m)$；韩语短语 P_{kr} 由 n 个词组成，每个词为 $Y_{j(j=1,2,\cdots,n)}$，则有 $P_{kr}=(y_1,y_2,\cdots,y_n)$。定义短语词向量相似度 S_E、短语长度相似度 S_L 和短语词性相似度 S_P，如式(2)~(4)所示：

$$S_E=\sum_{\substack{1\leqslant i\leqslant m\\1\leqslant j\leqslant n}}\cos\ (W_i,W_j) \tag{2}$$

式中，W_i、W_j 分别为对应词的词向量权重。

$$S_L=\frac{\min(L_{zh},L_{kr})}{\max(L_{zh},L_{kr})} \tag{3}$$

式中，L_{zh}、L_{kr} 分别为汉语、韩语短语字长度。

$$S_P=\frac{N}{\max(m,n)} \tag{4}$$

式中，N 为韩汉对应短语中词性相同词的个数。

最终得到韩汉短语相似度，如式(5)所示。

$$\text{Similarity}\ <P_{zh},P_{kr}\geqslant W_1*S_E+W_2*S_L+W_3*S_P \tag{5}$$

其中，W_1、W_2、W_3 分别为 S_E、S_L 和 S_P 的权重，默认权重值为 1/3。

4　实验与分析

4.1　语料介绍与数据预处理

本文以中国政府工作报告(中韩对照版)、当代中国与世界研究院、中国翻译研究院和中国外文局联合编译的《中国关键词》(中韩对照版)，以及通过网络爬虫获取的政治外交领域的双语文章作为原始语料。在此基础上，使用自动对齐于人工校对的方

式进行句对齐,最终得到韩汉双语平行句对11 672对。

对于汉语句子,使用 LAC 工具进行分词、去停用词并进行词性标注;对于韩语句子,使用 UTagger 工具进行分词、去停用词并进行词性标注。之后采用 Word2Vec 中的 Skip-Gram 模型,分别训练处理过的韩汉句子集合。训练参数分别为: *Size*(词向量维度) = 100, *Window*(窗口大小) = 3, *Iter*(迭代次数) = 10, 其它均为默认参数,分别得到韩语和汉语单语词向量,并使用 Vecmap2 工具将其映射至同一向量空间,得到韩汉双语词向量。

4.2 实验设计与测评指标

对于每组平行句对,基于规则抽取出短语后形成短语集。对于短语集中的每个短语,计算与对应短语集中每个短语的相似度后,选取相似度最大的作为对齐短语。此外,设定了两种对齐情况:完全对齐(对齐结果与正确结果完全一致)与未对齐(对齐结果与正确结果完全不一致)。见表 5。

表 5 短语"中国经济"匹配配对示例

Tab. 5 Example of matching the phrase China´s economy

候选韩语短语	词向量相似度	长度相似度	词性相似度	加权	是否对齐
중국경제 中国经济	**0.790**	**1.0**	**0.50**	**0.763**	**是**
고품질성장 高品质增长	0.270	0.80	0.50	0.523	否
문화적교류 文化交流	0.194	0.80	0.33	0.441	否
匹配结果		中国经济—중국경제			

为有效评测融入双语词向量的短语自动对齐方法的性能,从平行语料中随机抽取出 2 000 对句对,采用专家人工审校方式进行短语对齐,将结果作为标准测试语料。

本文设计了 3 组对比实验:第一组实验,通过对比融入单语与双语词向量后的对齐效果,用以验证双语词向量的有效性;第二组实验,将训练词向量时的迭代次数和特征权重作为自变量进行实验,用以探究最佳的权重参数设置;第三组实验,通过对比训练词向量不同迭代次数后的对齐效果,探究迭代次数对结果的影响。

本文采用准确率 P、召回率 R 和 $F1$ 值指标作为衡量模型对齐短语的性能指标。其具体表达如式(6)~(8)所示。

$$Recall = \frac{|\ TP\ |}{|\ TP + FN\ |} \tag{6}$$

$$Precision = \frac{|\ TP\ |}{|\ TP + FP\ |} \tag{7}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100 \tag{8}$$

其中, TP 为短语对齐结果与测试集完全匹配的数量; TP 为测试集中未与短语对齐结果匹配的数量; FN 为短语对齐结果中未与测试集匹配的数量。

4.3 实验结果与分析

第一组实验结果见表 6。

表 6 融入词向量对比实验研究

Tab. 6 Comparative experimental study on integrated word vector %

方法	P	R	$F1$
无词向量	39.15	51.85	44.61
单语词向量	40.95	54.25	46.67
双语词向量	**47.86**	**63.40**	**64.55**

从中可以看出:融入未经映射的单语词向量相比于未融入词向量略有提升。准确率 P、召回率 R 和 $F1$ 值分别提升了 1.80%、2.40% 和 2.06%;而融入双语词向量后,相比于单语词向量有较大提升,准确率 P、召回率 R 和 $F1$ 值分别提升了 6.93%、9.15% 和 7.88%。由此可知双语词向量对短语对齐的提升作用比较明显。

第二组实验结果见表 7。

表 7 权重组合对比实验结果

Tab. 7 Weight combination comparison experimental results %

权重组合	P	R	$F1$
(0.33 0.33 0.33)	47.86	63.40	54.55
(0.50 0.25 0.25)	46.71	61.87	53.23
(0.25 0.50 0.25)	47.04	62.31	53.61
(0.25 0.25 0.5)	47.70	63.18	54.36
(0.5 0.5 0.0)	46.46	61.55	52.95
(0.5 0.0 0.5)	43.26	57.30	49.30

从结果看出:经过多组权重对比实验,词向量相似度、长度相似度与词性相似度的权重均对结果有一定影响。词向量特征具有较强的正向作用,长度特征和词性特征具有一定的正向作用。在三者权重相当时,模型整体性能最好。

第三组实验结果见表 8。

表 8　词向量训练迭代次数对比实验结果

Tab. 8　Comparison of word vector training iteration times and experimental results　%

迭代次数	*P*	*R*	*F*1
10	47.86	63.40	54.55
15	55.10	72.98	62.80
20	64.14	84.97	73.10
25	64.72	85.73	73.76
30	51.48	68.19	58.70
35	49.42	65.47	56.32

可以看出：词向量训练时的迭代次数会对模型性能产生较大影响。随着迭代次数的增加，各项指标呈现先上升后下降的趋势。迭代次数为 25 时效果最好，相比于默认的 10 次迭代，准确率 *P*、召回率 *R* 和 *F*1 值分别提升了 16.86%、22.33% 和 19.21%。说明适当增加训练迭代次数，对模型的性能有很大提升。

5　结束语

本文提出了融入双语词向量的韩汉名词短语对齐方法，并构建了基于短语构成规律的短语抽取和融入双语词向量、短语长度和词性相似度的短语对齐模型。在政治外交领域的韩汉平行语料上进行实验分析，得到以下结论：

（1）双语词向量无需借助外部资源（如双语词典、术语库等）就能够高效地表示平行句对中对应单位的深层语义特征，从而提升对应单位对齐的准确率。

（2）语言学知识对于短语抽取与对齐和类似自然语言处理任务仍起着重要作用。部分情况下，短语长度和词性相似度仍能进行有效短语对齐，对于对齐结果有正向提升。

由于时间及水平所限，本文尚存在许多不足。一是韩汉名词短语的种类有待进一步扩充。基于短语结构使用词性抽取的方法需要依靠语言学知识制定大量规则，且只能覆盖部分类别的短语，后续将尝试使用统计的方法进行短语抽取，扩充短语的种类。二是语料的规模有待进一步增加。词向量的训练基于大规模语料，而目前高质量的平行语料仍属稀缺资源，因此如何自动高效地获取句对齐平行语料仍是研究的方向。三是面向韩汉自然语言处理领域的语言学知识有待进一步归纳。本文短语对齐的相关指标仍不能令人满意，其主要原因是韩汉双语间的语言差异导致短语抽取、特征提取效果不佳。因此进一步挖掘深层的句法语义知识有助于自然语言处理领域相关任务的实现。

参考文献

[1] 李文中. 平行语料库设计及对应单位识别[J]. 当代外语研究，2010(9)：26-31，65.

[2] 牛翊童. 基于汉越双语平行语料库的词对齐方法研究[D]. 昆明：昆明理工大学，2017.

[3] 杨飞扬，赵亚慧，崔荣一，等. 基于平行语料和翻译概率的多语种词对齐方法[J]. 中文信息学报，2019，33(12)：37-44.

[4] 王思宽. 基于规则和基于统计相结合的中英双语平行句对短语对齐方法[D]. 北京：北京邮电大学，2010.

[5] 孙茂松，李莉，刘知远. 面向中英平行专利的双语术语自动抽取[J]. 清华大学学报(自然科学版)，2014，54(10)：1339-1343.

[6] 原伟. 基于俄汉政治外交平行语料库的短语对应单位抽取研究[J]. 解放军外国语学院学报，2020，43(5)：38-45.

[7] 凌天斌，毕玉德. 基于统计和词典方法相结合的韩汉双语语料库名词短语对齐[J]. 中文信息学报，2018，32(8)：27-31.

[8] 周嘉剑. 基于英汉平行语料库的双语词对齐系统[D]. 重庆：重庆邮电大学，2019.

[9] 于清，常乐，徐健，等. 基于汉维医疗平行语料的双语术语抽取研究[J]. 内蒙古大学学报(自然科学版)，2018，49(5)：528-533.

[10] SEO H W，KIM J H. Analyzing Errors in Bilingual Multi-word Lexicons Automatically Constructed through a Pivot Language[J]. Journal of the Korean Society of Marine Engineering，2015，39(2)：172-178.

[11] SEO H W，KWON H S，CHEON M A，et al. Bilingual multi-word lexicon construction via a pivot language[J]. Contemporary Engineering Sciences，2014，7(23)：1225-1233.

[12] 安帅飞，毕玉德. 韩国语名词短语结构特征分析及自动提取[J]. 中文信息学报，2013，27(5)：205-210.

[13] [illegible]

[14] Jiao Z，Sun S，Sun K. Chinese lexical analysis with deep bi-gru-crf network[J]. arXiv preprint arXiv：1807.01882，2018.

[15] BENGIO Y，RÉJEAN DUCHARME，VINCENT P，et al. A Neural Probabilistic Language Model.[J]. Journal of Machine Learning Research，2003，3：1137-1155.

[16] Mikolov T，Chen K，Corrado G，et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv：1301.3781，2013.

[17] RUDER S，VULI I，SGAARD A. A Survey Of Cross-lingual Word Embedding Models[J]. Journal of Artificial Intelligence Research，2017，65：569-631.

[18] 彭晓娅，周栋. 跨语言词向量研究综述[J]. 中文信息学报，2020，34(2)：1-15，26.

[19] ARTETXE M，LABAKA G，AGIRRE E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018：64-73.

[20] ARTETXE M，LABAKA G，AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings[J]. arXiv preprint arXiv：1805.06297，2018.

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

[22] 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.