基于 word2vec 的大中华区词对齐库的构建*

王明文，徐雄飞，徐凡*，李茂西

（江西师范大学 计算机信息工程学院，南昌 330022）

摘要：本文针对大陆、香港和台湾地区（简称大中华区）存在同一种语义但采用不同词语进行表达的语言现象进行分析。首先，我们抓取了维基百科以及简繁体新闻网站上的 3,200,000 万组大中华区平行句对，手工标注了一致性程度达到 95%以上的 10,000 组大中华区平行词对齐语料库。同时，我们提出了一个基于 word2vec 的两阶段的大中华区词对齐模型，该模型采用 word2vec 获取大中华区词语的向量表示形式，并融合了有效的余弦相似度计算方法以及后处理技术。实验结果表明我们提出的大中华区词对齐模型在以上两种不同文体的词对齐语料库上的 F1 值显著优于现有的 GIZA++和基于 HMM 的基准模型。此外，我们在维基百科上利用该词对齐模型进一步生成了 90,029 组准确率达 82.66%的大中华区词语三元组。

关键词：大中华区；词对齐；最长公共子序列；word2vec

中图分类号：TP391　　**文献标识码**：A

Building Word Alignment Corpus for the Greater China Region

Wang Mingwen，Xu Xiongfei，Xu Fan*，Li Maoxi

(School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

Abstract: We deal with the linguistic phenomenon that different expressions refer to the same semantic meaning among the Mainland China, Hong Kong and Taiwan, a.k.a., the greater China region(GRC). Firstly, we automatically crawl 3.2 million GCR parallel sentences from the wikipedia and the news website with simplified and traditional encoding, and then manually annotate 10,000 GCR parallel word alignment corpora with the annotation agreement above 95%. Meanwhile, we present a 2-phase GCR word alignment model based on word2vec to represent the GCR words' vector, which incorporate the effective word similarity calculation method along with post-processing technology. Expreiment results on the proposed 2 different word alignment corpus demenstrate the effectiveness of our GCR model which significantly outperforms the current GIZA++ and HMM-based models. Furthermore, we generate 90,029 triples from wikipedia with accuracy over 82.66%.

Keywords: the greater China region;word alignment; the longest common subsequence;word2vec

1 引言

词对齐的基本任务是从平行语料中找到相互对应、互为翻译的词。它在基于短语的统计机器翻译模型中具有重要的作用。目前，现有的词对齐库的源语言和目标语言通常是在两种不同的语言（双语）下进行的，例如，中英文[1]词对齐语料库，日英词对齐语料库[2-3]等等。然而，针对单语（一种语言）的词对齐语料库却非常鲜见。众所周知，香港、台湾与大陆在词语表达上面存在着一定的差异。在文字上，香港和台湾使用繁体中文，大陆则使用简体中文。在语言上，香港用广东话多一些，而台湾用福建的闽南话比较多一些。由于历史种种原因，港台词语和大陆词语产生了一定的差异，例如：香港和台湾中“熊猫”称为“猫熊”、“信息”称为“资讯”、“打印机”称为“列表机”。这些词语虽然在形式上不同，但在语义上具有同一

*基金项目：国家自然科学基金（61462045，61402208，61462044）；国家语委“十二五”规划（YB125-99）；江西省自然科学基金（20132BAB201030，20151BAB207027，20151BAB207025）

性。就口语而言，大陆、香港和台湾的表达在语音方面相差确实比较大，但就书面语而言，大陆、香港和台湾的表达方式在语法层面上没有太多区别，可以看成是同一个地区的方言。如何从海量的大中华区文本中准确地抽取出这些表达形式不同但语义相同的平行词对齐库具有重要的理论意义和实践价值。理论方面：可以利用计算机自动构建海量互联网文本中词语对齐可计算模型，从而实现计算机对词语对齐的批量生成、处理、分析和加工；实践方面：利用计算机自动生成的词语对一方面可以丰富语言学家对大陆和港台词语差异上面的认识，另一方面分析大陆与港台互联网文本在词语表达上的差异，可以为企业、政府等机构更好地理解双方的文本内容，从而提供重要的决策依据。

基于此，我们抓取了维基百科以及简繁体新闻网站上的 3,200,000 万组大中华区平行句对，手工标注了一致性程度达到 95%以上的 10,000 组大中华区平行词对齐语料库。同时，我们提出了一个基于 word2vec[4]的两阶段的大中华区词对齐模型，该模型采用 word2vec 获取大中华区词语的向量表示形式，并融合了有效的余弦相似度计算方法以及后处理技术。实验结果表明我们提出的大中华区词对齐模型在以上两种不同文体的词对齐语料库上的 F1 值显著优于现有的 GIZA++和基于 HMM 的基准模型。此外，我们在维基百科上利用该词对齐模型进一步生成了 90,029 组准确率达 82.66%的大中华区词对齐三元组。

本文其余部分的结构如下：第二节介绍词对齐方面的相关工作；第三节介绍大中华区词对齐库的构建及标注过程；第四节阐述本文提出的基于word2vec的两阶段的词对齐库模型；第五节给出了该模型在词对齐语料库下的实验结果及分析；第六节给出论文的结论及后续工作安排。

2 相关工作

目前主流的词对齐模型可分为生成模型和判别模型两大类型，其中生成模型主要有Brown于1993年提出的IBM Model 1[5]模型，以及由Vogel等人于1996年提出的基于隐马尔科夫模型[6]的词对齐方法。这两个模型都是通过最大化句子的对齐概率来确定词语的对齐情况（见公式1）。

$$\arg\max P\left(f^J \mid e^I\right) = \sum_{a^J}\left(\prod_{j=1}^{J} p_d(a_j \mid a_{j-})\, p_t(f_j \mid e_{a_j})\right) \tag{1}$$

公式1中 $e^I = \{e_1,\ldots,e_I\}$ 称为源语言句子，$f^J = \{f_1,\ldots,f_J\}$ 称为目标语言句子，$a^J = \{a_1,\ldots,a_J\}$ 称为源语言句子与目标语言句子的对齐向量，例如 $a_j = i$ 表示目标语言中第 j 个词与源语言中第 i 个词对齐，$j-$ 序列表示目标语言中第 j 个词之前的一个不为空的词序号。IBM Model 1与基于隐马尔科夫的词对齐方法的根本区别在于词语的位置对齐概率计算方法不同，IBM Model 1将源语言中第 i 个词与目标语言中第 j 个词认为是等概率情况，均为 $1/J$（J 代表目标语言句子中词的总数目）,而基于隐马尔科夫模型的词对齐方法认为源语言中第 i 个词对齐的概率与目标语言中第 j 个词以及 $j-$ 词有关，其为

$p_d\left(a_j = i \mid a_{j-} = i-\right)$（$i-$代表源语言句子中第$i$个词的前一个不为空的词）。后续的词对齐模型工作主要是针对上面两个模型的改进[7-10]。

与生成模型不同，判别模型主要是计算词与词之间的翻译概率，并且将对齐概率转换到条件概率框架之下，如可以考虑将对齐转换为公式2形式。

$$\bar{a} = \arg\max \sum_{i=1}^{J} \lambda_i f_i(a^J, e^I, f^J) \tag{2}$$

公式2中$\bar{a}$表示源语言与目标语言的最终对齐序列，f_i是目标语言中第i个词的特征，λ_i是对应赋予特征的权重。同样，也存在判别模型下的词对齐库模型改进工作[11-15]。

3 大中华区词对齐库生成及标注

本节主要介绍大中华区词对齐库的生成及标注过程。

3.1 大中华区平行句对的生成及预处理

本文针对维基百科和简繁体新闻网页生成了两种不同文体的大中华区平行句对。针对维基百科，我们首先使用搜狗常用词典[1]来生成待访问的大陆简体维基百科网页链接，再根据大陆网页链接中“zh-cn”部分改为“zh-hk”、“zh-tw”来生成相对应的香港和台湾繁体网页链接；针对简繁体新闻网页，我们直接抽取了大公网[2]和台湾网[3]上的简繁体平行链接。我们采用Jsoup[4]提取了平行链接中大陆、香港和台湾对应的网页正文内容，然后对正文内容按照“。”、“！”、“？”和“；”标点符号进行分句，然后采用中科院张华平博士开发的分词工具ICTCLAS2015[5]针对句子进行分词。目前，我们共抓取了1,800,000个链接，生成了3,200,000万个大陆-香港和大陆-台湾的平行句对。同时，我们将香港和台湾的繁体句子转换成简体句子，并提取出有效平行句对（平行句对中含有不同的词情形）。

3.2 大中华区词对齐库的人工标注

我们首先随机选取了10,000组维基百科和简繁体的新闻网站的平行句对，然后请两位高年级的研究生对上面抽取出来的平行句对进行词对齐标注。其中维基百科语料中大陆-香港、大陆-台湾分别选取4,000个句对；简繁体新闻语料中大陆-香港、大陆-台湾分别选取1000个句对。我们采用github上开源的标注工具[6]进行词对齐标注，当标注两个句子之间的对齐情况时，分别将句子中词与词的对齐序号标注出来。

一般而言，Kappa值可以有效地衡量语料的标注质量，然而Kappa值是针对有明确类别的标注任务，词对齐的标注任务中并没有明确表示某个词要对齐到哪些类别，所以本任务无

[1] http://pinyin.sogou.com/dict/detail/index/2441
[2] http://www.takungpao.com/
[3] http://www.taiwan.cn/
[4] http://jsoup.org/
[5] http://ictclas.nlpir.org/
[6] https://github.com/desilinguist/wordalignui

法计算出语料的Kappa值，相反，我们使用标注一致性来计算两位标注者的标注结果。计算方法如公式(3)所示。

$$Agreement = \frac{\{\#word\ alignment\}}{\{\#word\}_{CN}} \tag{3}$$

公式3中分子表示的是两个标注者同时标注为对齐的词语对数目，分母表示大陆句子中的总词语数量。表1是标注一致性结果，在两种不同文体的语料上的两位标注者的一致性程度均高于95%，其结果表明我们构建的大中华区词对齐库语料具有很高的标注质量。

表 1 大中华区词对齐语料标注一致性表

词对齐语料	Agreement (%)
维基百科语料-大陆香港	95.32
维基百科语料-大陆台湾	95.97
新闻语料-大陆香港	96.83
新闻语料-大陆台湾	97.51

4 基于 word2vec 两阶段词对齐库模型

word2vec[7]是 Google在 2013 年开源的一款将词语表征为实数值向量的高效工具，其利用深度学习的思想，通过训练，把文本内容的处理简化为K维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。word2vec已经被广泛的应用于很多NLP（Natural Language Processing）相关的工作中，如情感分析、词性分析和聚类等。

鉴于此，我们采用word2vec将大中华区句子中的每个词语生成一个200维的向量，然后通过向量与向量之间的余弦距离来计算词语与词语之间的相似度（见公式4）。

$$Sim(e_i, f_j) = \cos\theta = \frac{\vec{e_i} \cdot \vec{f_j}}{\|e_i\| \cdot \|f_j\|} \tag{4}$$

其中，公式 4 中 e_i 是源语言句子中第 i 个词，f_j 是目标语言句子中第 j 词，第 i 个词与第 j 个词之间的相似度用 $Sim(e_i, f_j)$ 表示。$\vec{e_i}$，$\vec{f_j}$ 分别对应 e_i，f_j 的 200 维向量表示。

我们提出的两阶段的大中华区词对齐模型主要利用了word2vec来计算源语言句子中的词与目标语言句子中词的相似性。阶段1的核心思想是每次都选取源语言句子和目标语言句子中相似度最大的词作为最终对齐结果，并分别判断其左、右两边是否存在1-1或1-n的对齐情况（如表2所示），如都存在则将其作为最终的对齐结果，并且对齐结束。否则对不存在的部分重复上面的方法，直到整个句子对齐结束为止。阶段2的核心思想是对阶段1的结果进行修正，主要处理空对齐以及m-n对齐（如表2所示）情况。图1为大中华区语料词对齐算法。

算法的输入是源语言和目标语言的句对，输出是最终的词对齐情况。

阶段 1：初始词对齐

初始词对齐对应图 1 中 4-14 行，第 4 行对应采用最长公共子序列算法(The Longest Common Subsequence LCS)[16]，将源语言句子 e^I 与目标语言句子 f^J 公共子序列部分进行对齐，存入 $a(e^I, f^J)$ 中；第 5 行是去除掉 e^I 和 f^J 中的公共子序列并分别存入 e^{remove}, f^{remove} 中。接下来我们对剩余的 e^{remove}, f^{remove} 部分进行对齐：第 8 行计算 e^{remove}, f^{remove} 中每两个词的余弦相似度（如公式 4）；第 10 行选取出源语言句子中第 i 个词对应目标语言句子第 j 个词的最大相似度作为 i 与 j 的对齐概率 $p(i \mid j)$；第 12 行找出整个句子中最大的 $p(i \mid j)$ 作为对齐情况；第 13-14 行是在 12 行对齐基础上找该对齐左边和右边是否存在 1-1 或 1-n 的对齐情况（如表 2 所示）。如果都存在，则阶段 1 结束，否则对不存在的那部分重复图 1 中第 12-14 行部分的操作。

算法：基于word2vec的大中华区语料词对齐算法

1. *Input*: e^I, f^J
2. *Output*: $a(e^I, f^J)$
3. *BEGIN*
4. $a(e^I, f^J) \leftarrow LCS(e^I, f^J)$
5. $e^{remove}, f^{remove} \leftarrow$ *remove LCS from* e^I *and* f^J
6. *For* i *from* 1 *to size*(e^{remove})
7. *For* j *from* 1 *to size*(f^{remove})
8. $Sim(e_i^{remove}, f_j^{remove}) \leftarrow COS_{Word2vector}(e_i^{remove}, f_j^{remove})$
9. *End For*
10. $p(i \mid j) \leftarrow MAX_{Sim(e_i^{remove}, f_j^{remove})}$
11. *End For*
12. $a(e^I, f^J) \leftarrow MAX_{p(i|j)}$
13. $a(e^I, f^J) \leftarrow Alignment(e_i^{Left}, f_j^{Left})$
14. $a(e^I, f^J) \leftarrow Alignment(e_i^{Right}, f_j^{Right})$
15. $Process(a(e^I, f^J))$
16. *End*

图 1　基于 word2vec 的大中华区语料词对齐算法

从图1的分析可以看出，如算法所示的词对齐模型只处理了1-1和1-n的两种情况，然而无法处理m-n的对齐情况。同时我们每次都是选取最大的词语来作为最终的对齐情况，如果选取的最大对齐情况有误，那么该错误将会影响到整个句子后续的对齐情况，并且造成无法挽回的错误。为了解决这种情形，我们在图1所示的词对齐结果上给定一个校验与修正模型。

表2　1-1、1-n、m-n三种对齐情况实例表

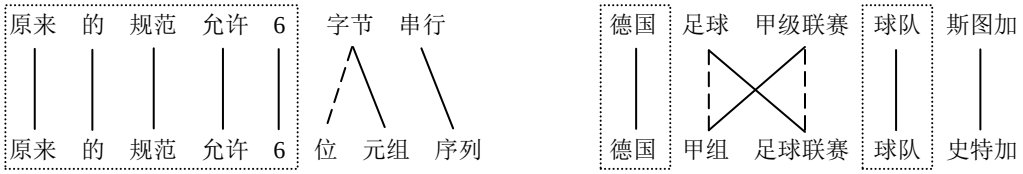
对齐情况	1-1 对齐	1-n 对齐	m-n 对齐

[7] http://word2vec.googlecode.com/svn/trunk/

阶段 2：词对齐后处理

词对齐后处理对应图1算法的第15行，是对阶段1的对齐结果 $a(e^I, f^J)$ 进行修正。通过对阶段1具体分析，我们可以发现如果阶段1生成的词对齐出现交叉或者对齐为空的情况，则说明该句子的对齐情况需要修正。具体错误实例如图2所示：

其中，图2中用虚线框住的部分为最长公共子序列部分，词与词连线为实线的部分是阶段1对齐结果，词与词之间用虚线连接的部分为阶段2修正的对齐结果。如图2(a)所示，首先采用word2vec最先将“字节”与“元组”进行对齐，然后再将“串形”与“序列”进行对齐，这样导致目标语言中的“位”字没有对齐，在修正模型中我们从左到右扫描，发现这种对齐为空的情况，就将其合并入下一组对齐中（即“字节”对齐“位、元组”）。图2(b)所示阶段1对齐结果为“足球”对齐“足球联赛”；“甲级联赛”对齐“甲组”，阶段2可以识别这两个对齐存在交叉部分，我们可以将这种对齐看成m-n的对齐情况。因此，通过阶段2修正后，我们可以处理m-n的情况，如图2(b)中“足球、甲级联赛”对齐“甲组、足球联赛”的情况。



(a) (b)

图 2　词对齐模型对齐情况实例图

5 实验

为了验证本文标注的大中华区词对齐语料的可行性，同时为了验证本文提出的基于word2vec的两阶段的大中华区词对齐模型的有效性，我们给出了如下实验。

5.1 实验设置

数据集：我们标注的两种类型的词对齐语料，其一是维基百科上的大陆-香港以及大陆-台湾词对齐语料、其二是简繁体大陆-香港以及大陆-台湾词对齐语料。

基准模型：我们选择了目前几个主流的词对齐模型作为基准模型，分别是2003年由Och and Ney提出的GIZA++模型[17]，以及Berkeley aligner工具包中的两种方法：由2006年由Liang提出的基于HMM模型的词对齐方法[18]和2007年由DeNero和Klein提出的基于句法的HMM（SYN-HMM）模型的词对齐方法[19]。

5.2 实验结果及分析

本节分别设计了大中华区词对齐和三元组的识别实验，实验结果及分析如下两节所示。

5.2.1 大中华区词对齐识别性能

我们统计对齐结果的时候将最长公共子序列部分去除，这样可以更直观的分析各个系统对大中华区平行语料的对齐情况。

本文采用P、R和F1值来评价系统性能。表3是词对齐结果(表格中的*代表成对的显著性检验p<0.01)。结果显示本文的模型在维基百科词对齐语料和简繁体新闻词对齐语料的对齐结果都要优于GIZA++、基于HMM的方法和SYN_HMM的对齐结果。我们的模型阶段2的F1值优于基于GIZA++ (grow-diag-final)的方法1.5-4.8%。实验结果表明我们的模型要明显优于其它基准系统，原因在于word2vec能够很好的将词语表示成全局的向量形式，进而可以采用余弦相似度来衡量大中华区词语间的相似程度，同时我们提出的词对齐算法的第1阶段每次都选取整个句子中概率最大的对齐情况，并结合1-1和1-n对齐规则，并且算法的阶段2修正了阶段1不能处理的空对齐与m-n对齐情况。

表3　各个系统词对齐P,R,F1值表

词对齐模型	维基百科词对齐语料			新闻词对齐语料		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
			大陆与香港			
GIZA++ (→)	72.27	73.82	73.04	71.37	73.16	72.25
GIZA++ (←)	80.13	81.84	80.98	81.11	82.13	81.62
GIZA++ (grow-diag-final)	86.44	87.32	86.63	86.31	90.69	88.45
HMM	80.21	81.47	80.84	80.63	82.52	81.57
SYN_HMM	78.45	79.36	78.91	78.83	80.35	79.58
OUR_STAGE1	83.46	86.27	84.84	85.52	88.46	86.97
OUR_STAGE1+STAGE2	**89.04***	**88.10***	**88.56***	**90.54***	**90.38***	**90.46***
			大陆与台湾			
GIZA++ (→)	72.45	73.72	73.08	70.29	71.27	70.78
GIZA++ (←)	80.86	81.93	81.39	81.16	82.65	81.89
GIZA++ (grow-diag-final)	85.98	87.62	86.79	86.31	89.48	87.86
HMM	80.55	81.89	81.22	80.23	80.99	80.61
SYM_HMM	79.20	80.35	79.77	78.13	78.87	78.50
OUR_STAGE1	83.54	85.85	84.68	87.64	89.18	88.40
OUR_STAGE1+STAGE2	**88.88***	**87.72***	**88.29***	**93.00***	**92.32***	**92.66***

注：表中 OUR_STAGE1 代表本文模型阶段 1 试验结果，OUR_STAGE1+STAGE2 代表本文模型（阶段 1 和阶段 2）试验结果，GIZA++ (→)、(←)分别表示不同对齐方向，GIZA++ (grow-diag-final)是将 GIZA++两个对齐方向进行融合。

为了更细粒度的分析实验结果，本文对语料中去除最长公共子序列部分的 1-1、1-n 和 m-n 三种对齐情况进行了分析（见表 4 所示的大陆-香港对齐情况）。结果表明在 1-1 和 1-n 对齐情况下基于 HMM 的方法、SYN_HMM 和 GIZA++方法均低于我们提出模型的阶段 2，在

这些基准模型中 GIZA++ (grow-diag-final)方法效果比较好。同时在 m-n 对齐情况下，仅有 GIZA++ (grow-diag-final)以及我们提出模型的阶段 2 能够处理，其它基准系统无法处理该部分对齐。因此可以看出我们提出的模型的结果优于其它三个基准模型的原因在于，我们的模型能够更好的处理 1-n 与 m-n 的对齐情况。此外，大陆-台湾的 1-1、1-n 和 m-n 对齐情况结果类似，限于篇幅，本文不再赘叙。

表4　各系统1-1, 1-n, m-n三种词对齐情况结果表

词对齐模型	维基百科词对齐语料				新闻词对齐语料		
	对齐情况	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
GIZA++ (→)	1-1	71.60	99.05	83.12	70.64	98.92	82.42
	1-n	63.77	12.10	20.34	87.03	14.02	24.16
	m-n	-	-	-	-	-	-
GIZA++ (←)	1-1	80.31	97.20	87.95	86.31	91.98	89.06
	1-n	84.06	69.20	75.91	82.08	68.36	74.59
	m-n	-	-	-	-	-	-
GIZA++ (grow-diag-final)	1-1	83.65	99.05	90.70	86.30	98.83	92.14
	1-n	77.96	80.84	79.37	81.19	79.75	80.47
	m-n	10.30	12.81	11.42	10.09	14.35	11.85
HMM	1-1	81.10	97.52	88.55	80.70	98.44	88.69
	1-n	75.35	44.01	55.56	80.69	48.65	60.70
	m-n	-	-	-	-	-	-
SYN_HMM	1-1	79.45	96.62	87.19	78.93	97.61	87.28
	1-n	72.26	38.42	50.17	78.68	42.98	55.59
	m-n	-	-	-	-	-	-
OUR_STAGE1	1-1	81.63	95.23	87.91	83.02	95.22	88.70
	1-n	91.84	69.44	79.08	94.22	77.91	85.29
	m-n	-	-	-	-	-	-
OUR_STAGE1+STAGE2	1-1	93.66	92.24	**92.95***	93.65	93.31	**93.48***
	1-n	82.91	82.95	**82.93***	89.22	88.96	**89.09***
	m-n	23.36	23.55	**23.46***	15.38	14.28	**14.81***

注：图中"-"表示系统无法处理 m-n 对齐情况。

5. 2. 2 大中华区三元组识别性能

为了进一步验证本文提出的词对齐算法在更大规模平行句对上的执行情况，我们自动抽取了3,200,000组维基百科上的平行句对，同时利用本文词对齐方法生成了90,029组大中华区词对齐三元组。表5列出了一些大中华区三元组样式。

表 5　大陆、香港和台湾词典实例表

大陆词语	香港词语	台湾词语
方便面	即食面	速食面
熊猫	熊猫	猫熊
粒子	体子	体子
出租车	的士	计程车

分辨率	解像度	解析度
存档	封存	封存
福布斯	福布斯	富比士
琼斯	钟斯	琼丝
乔布斯	乔布斯	贾伯斯
信息产业部	资讯工业部	资讯工业部

从表5数据中数据可以看出，大陆、香港和台湾有些词比较相似，如“熊猫”和“猫熊”。然而有些词的差异很大，如“粒子”和“体子”，而有些词是由于音译时采用的字不相同所导致的，如“乔布斯”和“贾伯斯”。

此外，我们利用上述算法作用于维基百科，生成了90,029个大陆-香港-台湾词语的三元组，随机先取了5,000组大中华区三元组进行人工标注，本文算法可以取得82.66%的accuracy。实验结果进一步说明了本文提出的词对齐算法的可行性。

6 总结与展望

针对大中华区利用不同词语表达同一语义的语言现象进行了分析。我们首先人工标注了10,000组规模的来自维基百科和简繁体新闻网页上的两种不同文体的大中华区词对齐语料库，同时提出了一个基于word2vec的两阶段的词对齐模型，并且在以上两种词对齐语料库上进行了词对齐实验，实验表明我们提出的模型的词对齐性能显著优于现有的基于HMM的词对齐和GIZA++的基准模型。此外，我们利用本文提出的词对齐方法进一步生成了90,029组识别准确率达82.66%的大中华区词语三元组。

接下来我们将进一步扩充词典和词对齐语料的规模，同时我们计划将澳门、新加坡和马来西亚等其它大中华区的语言进行词对齐分析。

参考文献

[1]Ayan N F, Dorr B J. Going beyond AER: An extensive analysis of word alignments and their impact on MT[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 9-16.

[2]Takezawa T, Sumita E, Sugaya F, et al. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World[C]//In Proceedings of the 3rd International Conference on Language Resources and Evaluatio. 2002: 147-152.

[3] Mihalcea R, Pedersen T. An evaluation exercise for word alignment[C]//Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3. Association for Computational Linguistics, 2003: 1-10.

[4] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[5] Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational linguistics, 1993, 19(2): 263-311.

[6] Vogel S, Ney H, Tillmann C. HMM-based word alignment in statistical translation[C]//Proceedings of the 16th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1996: 836-841.

[7] Neubig G, Watanabe T, Sumita E, et al. An unsupervised model for joint phrase alignment and extraction[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 632-641.

[8] Songyot T, Chiang D. Improving word alignment using word similarity[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1840-1845.

[9] Kondo S, Duh K, Matsumoto Y. Hidden Markov Tree Model for Word Alignment[C]//8th Workshop on Statistical Machine Translation. 2013: 503.

[10] Chang Y W, Rush A, DeNero J, et al. A Constrained Viterbi Relaxation for Bidirectional Word Alignment[J]. Annual Meeting of the Association for Computational Linguistics. 2014: 1481-1490.

[11] Tamura A, Watanabe T, Sumita E. Recurrent neural networks for word alignment model[C]//Proceedings of EMNLP. 2014: 1470-1480.

[12] Yang N, Liu S, Li M, et al. Word Alignment Modeling with Context Dependent Deep Neural Network[C]// Annual Meeting of the Association of Computational Linguistics. 2013: 166-175.

[13] Blunsom P, Cohn T. Discriminative word alignment with conditional random fields[C]//Annual Meeting of the Association of Computational Linguistics, 2006, 65-72.

[14] Blunsom P, Cohn T. Discriminative word alignment with conditional random fields[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 65-72.

[15] Taskar B, Lacoste-Julien S, Klein D. A discriminative matching approach to word alignment[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 73-80.

[16] Chvatal V, Sankoff D. Longest common subsequences of two random sequences[J]. Journal of Applied Probability, 1975: 306-315.

[17] Katoh K, Misawa K, Kuma K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform[J]. Nucleic acids research, 2002, 30(14): 3059-3066.

[18] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[19] DeNero J, Klein D. Tailoring word alignments to syntactic machine translation[C]//Annual Meeting of the Association of Computational Linguistics. 2007, 45(1): 17–24.

作者简介：

王明文（1964—），男，博士，教授，主要研究方向为信息检索、数据挖掘、机器学习。

Email: mwwang@jxnu.edu.cn

徐雄飞（1989—），男，硕士研究生，主要研究方向为自然语言处理。

Email: xuxiongfei1989@sina.com

徐凡（1979—），男，博士，讲师，主要研究领域为自然语言处理和中文信息处理。

Email:xufan@jxnu.edu.cn **(通讯作者)**