基于词的分布式表示上下文向量翻译消歧方法

张春越 赵铁军
哈尔滨工业大学 哈工大语言语音教育部-微软重点实验室 哈尔滨 150001
Email:{cyzhang,tjzhao}@mtlab.hit.edu.cn

摘要：基于可比语料的双语字典抽取的任务中，种子字典的质量往往影响着最终抽取结果的准确程度。由于许多词具有一词多义的特性，导致在使用种子字典翻译上下文向量时，会引入很多的噪声。针对此问题，本文提出了一种基于词的分布式表示的无监督上下文向量翻译消歧的方法去降低种子字典的歧义性。实验表明，在汉英方向的双语字典抽取任务上，最终抽取得到的双语字典的准确率明显高于标准实现方法。
关键词：可比语料;上下文向量;种子字典;分布式表示;双语字典

Distributed Word Representation based Translation Disambiguation for Context Vector

Chunyue Zhang and Tiejun Zhao
MOE-MS Key Laboratory of Natural Language Processing and Speech
Harbin Institute of Technology, Harbin 150001,China
Email:{cyzhang,tjzhao}@mtlab.hit.edu.cn

Abstract: *In the task for extracting bilingual lexicon from comparable corpus, often the accuracy of extracted lexicon is influenced by the quality of seed lexicon. Because many words are polysemous, lots of noise will be generated when translating context vector using seed lexicon. This paper proposes a distributed word representation based disambiguation method for context vector to strengthen the seed lexicon. In the Chinese to English bilingual lexicon extraction task, the experiments show that the accuracy of extracted lexicon will be significantly improved over the standard approach method.*
Keywords: *Comparable corpus; context vector; seed lexicon; distributed word representation; bilingual lexicon;*

1 引言

双语字典（Bilingual Lexicons）在许多跨语言处理任务中都扮演着十分重要的角色，如机器翻译[Och et al.,2003]、跨语言信息检索[Ballesteros et al.,1997]等。一般地，双语字典的获取可以通过以下两种方式：其一是通过专家手工编制而成，但这样做需要花费较大的人力和物力；其二是使用平行语料（Parallel Corpus）通过统计词对齐方法获得。这种方式较第一种方式而言，不需要人工代价，但平行语料的获取同样代价较大，尤其是在某些资源缺乏的语言对上或特定领域中，平行语料是非常稀少的甚至可能根本就不存在。为此，基于可比语料（Comparable Corpus）的双语字典自动抽取技术成为解决这一问题的有效途径[Fung,1998;Rapp,1999]。可比语料主要是指在主题上相近或相关，内容上有重叠（overlapping）的双语文本。和平行语料相比，获取大量的双语可比语料是非常容易的，如维基百科[1]。因此，如何从可比语料中获取双语字典以及其他翻译知识成为近年来的一个热点问题。此类方法的核心理论依据为分布假设（Distributional Hypothesis）[Harris,1981]。这种假设可简单概括为：在两种语言中，词义相近或相同的词，它们的上下文词（context words）在词义上往往也是相近或相同的；换言之，上下文相似度较高的两个词，则有很大的可能成为互译的

[1] http://www.wikipedia.org/

词。基于此种假设来抽取双语字典的标准实现方法（Standard Approach）基本上可以归结为对源语言词和目标语言词的上下文向量（Context Vector）进行比较，相似程度越高，二者成为互译词对的可能性越大。上下文向量一般由在预先设定长度的窗口中与目标词共现（Co-occurrence）的词组成。由于源语言和目标语言的词空间不同，所以需要一个事先存在的双语字典将源语言的上下文向量翻译（转化）成目标语言空间中的向量，这个字典一般称之为种子字典（Seeds Lexicon）。

显然，种子字典是整个标准实现方法的核心元素。但是，种子字典常会出现一词多义（Polysemous）的现象，如英文单词“bat”在翻译成中文的时候会有“蝙蝠”、“球棒”和“批处理”三个翻译候选，中文单词“所有”在翻译成英文时会有“all”和“ownership”的不同意思。在将源语言上下文向量翻译成目标语言时，标准实现方法一般是将所有的翻译候选不加区分地使用。因此，这种歧义性将带来很多噪声，并影响最终获得的双语字典的准确率。

[Bouamor et al.,2013]考虑了上下文向量的歧义性问题。作者认为，假定当前为计算机主题的可比语料，如果在“bat”的上下文向量中，有“program”一词，且“program”的翻译为“程序”。则在“bat”的三个翻译候选“蝙蝠”、“球棒”和“批处理”中，“程序”一词可以用来帮助选择“bat”在当前语境中的正确翻译“批处理”。为此，作者使用基于 Word-Net 的方法对种子字典进行了翻译消歧，通过 Word-Net 上的一系列语义关联度量（Semantic Association Measures），在多个翻译候选中选择出最相关的一个或多个作为正确的翻译，去掉其余的候选项，显著提高了最终的双语翻译字典的性能。

基于 Word-Net 的方法是一种很有效的消歧方法。但是，这种方法仍然存在着一定的不足。基于 Word-Net 的消歧方法严重地依赖于 Word-Net 这种类型的语义资源库。而实际情况则是在多数语言上，这种语义资源是覆盖率比较低的甚至是不存在的，使得这种方法在其他语言上并不具有较强的扩展性。

基于此，本文提出了一种基于词的分布式表示（Distributed Word Representation）的上下文向量消歧方法。由于词的分布式表示可以用单语语料无监督地自动学习得到，这就解决了缺少 Word-Net 这类语义资源时[Bouamor et al.,2013]方法的不足。通过使用词的分布式表示，本文将词与词间的向量距离作为度量词与词关联程度的依据，进而对种子字典中的歧义词进行消歧。通过在汉英方向的双语字典抽取的实验表明，本文提出的方法显著地超过了标准实现方法，并具有更广泛的可扩展性。

在本文的剩余内容中，第 2 节简单地介绍了基于可比语料的双语字典抽取的标准实现方法；第 3 节介绍词的分布式表示的概念和本文用来获得词表示的模型；第 4 节介绍本文提出的词翻译消歧方法；本文的实验和结果在第 5 节介绍；第 6 节介绍了相关的工作；最后介绍了本文的结论和展望。

2 双语字典抽取方法的标准实现

根据分布假设，基于可比语料抽取双语字典的任务一般可以分为以下四个步骤，并将这种方式称之为标准实现方法（Standard Approach, SA）：

1. 首先建立源语言词*S*的上下文向量*VS*。*VS*向量由在指定的上下文环境中，和*S*共现的词*C*组成。其中，上下文环境的定义方式可以为不同长度的窗口，窗口可以限定在同一句子、段落甚至篇章。在这一步中，会根据停止词列表（Stop Word List）过滤停止词。此外，不同的*C*和*S*的关联程度（Association Measures）是不一样的。具体的关联程度值将在这一步被确定。关联程度的计算方法可以有TF-IDF，Log-Likelihood Ratio（LL），

Log-Odds-Ratio（LO），Pointwise Mutual Information（PMI）等，详见[Laroche et al.,2010]。一般地，完成这一步后，*S*的上下文向量*VS*会形如

$$VS = [\ldots\ldots, bat: 127.5, program: 98.2, \ldots\ldots]$$

2. 使用种子字典*SL*，将*VS*中的每一个源语言词翻译成目标语言词。如果，*VS*中的某一个词*C*在*SL*中有多个翻译，那么标准实现方法会将所有的翻译都考虑进来，并使其等分*C*在*VS*中的关联值。据此，步骤1中产生的*VS*则会变成

$$VS' = [\ldots\ldots, 蝙蝠: 42.5, 球棒: 42.5, 批处理: 42.5, 程序: 98.2 \ldots\ldots]$$

从这一例子中可以看出，在步骤2，由于种子字典的歧义性，那么将在*VS'*中引入噪声。因此，这一步是本文关注的问题所在。

3. 基于与步骤1同样的方式，目标语言的每一个词*T*的上下文向量*VT*也被创建。

在步骤 2 和步骤 3 后，源语言词的上下文向量 *VS'*就和 *VT* 在同一个特征空间中了，因此，可以根据相似度量方式，将 *VS'*与目标语言中每一个候选词 *T* 的上下文向量 *VT* 进行相似度的计算，并进行排序。最常用的相似度量方式为余弦（Cosine）距离。最终，排名第一的或者前几个 *VT* 所对应的目标语言词 *T* 将被视为翻译候选输出。

3 词的分布式表示及其模型

3.1 词的分布式表示

近年来，基于深度神经网络（Deep Neural Network, DNN）的各种模型得到了研究界的广泛关注，并且在很多自然语言处理的任务中收获了非常好的效果[Socher et al.,2011;Mikolov 2011]。之所以 DNN 能够在自然语言处理问题获得性能上提升，除了 DNN 本身的深层结构之外，另一个重要的原因在于这类工作大多会学习到词的一种分布式表示（Distributed Representation）[2]。过去，大部分的自然语言处理的模型都是建立在所谓的原子（atom）词表示基础上的。这种表示为：假设单语词典 *V* 共有|V|个不同的词，那么 *V* 中的每一个词 *w* 可用一个|V|维的向量 *VW* 表示，*VW*=[...,0,0,0,1,0,0,0,...]。这种表示方式习惯上称之为 One-Hot 形式。One-Hot 形式最大的不足在于它没法表征两个语义相近或相关的两个词。如“school”和“university”这样两个词在词义上很相近，但使用 One-Hot 形式的表示则无法得到两者相近的关系。而词的分布式表示（Distributed Word Representation）则把离散的词转变为一种连续的、低维度的、实值的向量形式。直觉上，词的分布式表示的每一个维度代表一个词的某种隐含特征，并能够在某种程度上代表这个词的语义或者语法属性[Bengio et al.,2006]。词的分布式表示通常使用大量的单语文本自动学习得到，然后针对具体的任务做微量调整。当在大量的语料上训练得到词的分布式表示之后，这些表示能够捕获很多重要的语义信息。得到词的分布式表示的方式有很多，如语言模型[Bengio et al.,2006]，Auto-encoder[Socher et al.,2011]。通过此类模型，词义相近或相关的词，会有着非常相似的分布式表示。

3.2 Skip-gram 模型

在[Mikolov et al., 2013]中，作者使用了 Skip-gram 和 Continuous Bag-of-Words（CBOW）模型来学习词的分布式表示。这两种模型都使用一个简单的神经网络体系结构，Skip-gram 与 CBOW 的区别在于前者用当前词预测上下文词，后者用上下文词预测当前词。两个模型的结构如图 1 所示。前者对低频词的效果更好，而后者的训练速度更快。由于其模型很简单，

[2] 注意与前文提到的分布假设的不同

使得其可以在大量的语料上进行训练，并且有相关的开源工具 word2vec[3]，因此本文使用 Skip-gram 模型获得词的分布式表示。

Skip-gram 模型训练的目标是想学习到一种词的分布式表示，这种表示能够较好的预测同一个句子中当前词的上下文词。这个模型的时间复杂度非常低，可以高效地在大规模单语数据上进行训练。在实践中，当单语数据比较小的时候，Skip-gram 能给出更好的分布式表示。

形式上，给定一个训练词序列$w_1, w_2, \ldots, w_T$，Skip-gram 模型的训练目标是最大化下述目标函数：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0} logP(w_{t+j}|w_t) \quad \ldots\ldots\ldots\ldots (1)$$

其中，c 为训练中设定的窗口的大小，里层的求和是计算给定当前词w_t，正确地预测词w_{t+j}的对数概率，外层的求和则是遍历训练数据中的所有词。

在 Skip-gram 模型中，每个词 w 被关联到两个参数向量u_w和v_w。相应地，它们代表了词 w 的输入向量和输出向量。最终，给定词w_i，正确预测词w_j的概率定义如下：

$$P(w_i|w_j) = \frac{exp(u_{w_j}^T v_{w_j})}{\sum_{w=1}^{W} exp(u_{w_j}^T v_{w_j})} \ldots\ldots\ldots\ldots (2)$$

其中，W 是词典中的总词数。

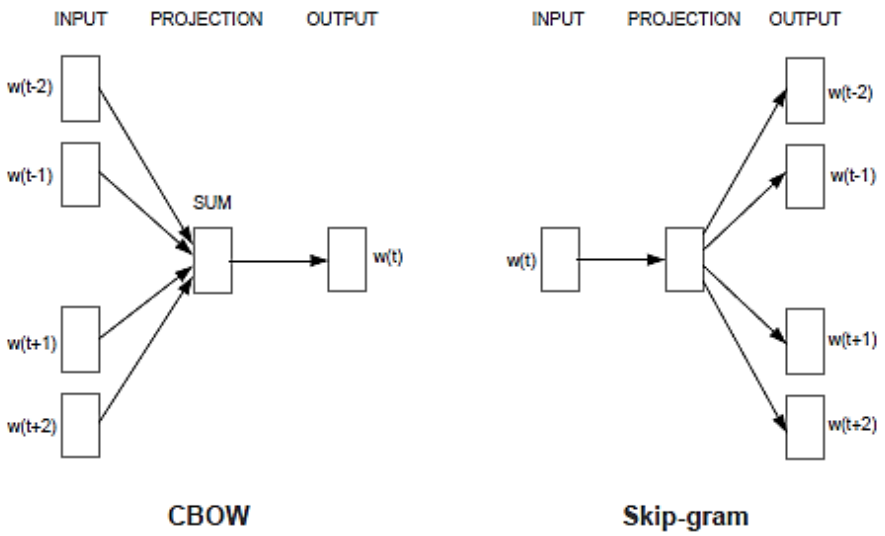

图 1. CBOW 和 Skip-gram 结构图

4 基于词的分布式表示消歧方法

词的分布式表示能够捕获词的语义信息，具体表现为语义越相关或相近的词在词的分布式表示空间上的距离越近[Mikolov et al.,2013]。基于这种属性，可以使用词的分布式表示作为词翻译消歧的依据。假设“bat”有不同的翻译项“蝙蝠”、“球棒”和“批处理”，同时在

[3] http://code.google.com/p/word2vec/

上下文向量中，有另一个词“program”，其翻译为“程序”。在词的分布式表示空间中，“批处理”和“程序”的距离会更近。因此，这个信息可以作为词翻译消歧的依据。

根据引言中的介绍，将源语言的上下文向量翻译成目标语言的上下文向量时，需要对多义词进行消歧。[Bouamor et al.,2013]提出根据无歧义词（*monosemic*）来作为参考词，这里的无歧义词指的是在种子字典中只有唯一的翻译候选的词。根据词的分布式表示的消歧能力可以认为，在当前上下文环境中，应该选择多义词的多个翻译候选中和这些无歧义词在分布式表示空间中距离较接近的词作为正确的翻译候选。基于此，本文提出了以下两种消歧方法。

4.1 单独消歧方法（Single Disambiguation Method）

设 w 为一个源语言词，其上下文向量为

$$[m_1:A_{m_1}, m_2:A_{m_2}, \dots, m_k:A_{m_k}; p_1:A_{p_1}, p_2:A_{p_2}, \dots, p_n:A_{p_n}],$$

其中$m_i(1 \le \mathrm{i} \le \mathrm{k})$为无歧义词，$A_{m_i}$为$m_i$与$w$的关联度，其翻译为$C_{m_i}$，其对应的词表示向量为$VC_{mi}$；$p_i(1 \le i \le n)$为有歧义词，$A_{p_i}$为$p_i$与$w$的关联度，其在种子字典中的翻译为$C_{p_{ij}}(1 \le j \le L)$, 表明其有 L 个翻译，每一个翻译对应的词表示向量为$VC_{p_{ij}}$；$d(w_1, w_2)$为w_1和w_2对应的向量之间的距离函数，距离函数可以使用 *cosine* 距离，欧几里德距离等等，本文使用的是欧几里德距离。对每一个$C_{p_{ij}}$，可以根据下述公式计算出其和无歧义词m_i的翻译C_{m_i}的语义相关（相似）程度。

$$Score\left(C_{p_{ij}}\right) = \frac{1}{\frac{1}{K}\sum_{i=1}^{K} d\left(VC_{m_i}, VC_{p_{ij}}\right)} \dots\dots\dots\dots (3)$$

根据词的分布式表达的语义作用假设，可以从公式（3）中看到，若C_{m_i}和$C_{p_{ij}}$越相关，则$d(VC_{m_i}, VC_{p_{ij}})$越小，*Score* 则越大；反之$d(VC_{m_i}, VC_{p_{ij}})$越大，*Score* 则越小。从而，对$C_{p_{i1}} \dots\dots C_{p_{iL}}$，我们得到了词义消歧的依据。$Score$越大，表明这个翻译越相关。因此可以根据排序信息选择前面的一个或几个翻译作为消歧之后的结果。

4.2 加权消歧方法（Weighted Disambiguation Method）

事实上，词的分布式表示的准确率并没有达到完美的程度，因此，仅仅选择歧义度最低的一个或几个作为消歧结果是有一定风险的。因此，可以使用这些歧义度对关联程度进行重新分配。歧义度越低，在重新分配的关联度中占有的比重应越大，反之则越低。为此，按如下方式定义每一个$C_{p_{ij}}$的关联度，其中A_{p_i}为m_i在源语言向量中的关联程度（见第 2 节）：

$$WeightedScore\left(C_{p_{ij}}\right) = A_{p_i} \times \frac{Score(C_{p_{ij}})}{\sum_j Score(C_{p_{ij}})} \dots\dots\dots\dots (4)$$

5 实验

5.1 实验数据和配置

实验的翻译方向为中文到英文。可比语料数据来自于 CWMT2008[4]发布的科技领域的双语数据。由于该数据是平行双语语料，因此为了获得可比双语语料，我们沿用[Daumé III et al.,2011]的做法：中文语料使用的是平行语料的中文部分的后 50%，英文语料采用的是平行语料中的英文语料的前 50%。注意这样产生的语料同维基百科上的数据相比，可比程度更差。

[4] http://nlpr-web.ia.ac.cn/cwmt-2008/

为了得到两个语言的分布式表示，使用 CWMT2008 评测发放的新闻领域的平行双语数据作为训练语料。中文数据使用 stanford-chinese-segmentor[Chang P.C et al.,2008]进行分词，英文数据进行了 tokenize，lowcase 和 lemmatization 操作，中、英文都根据各自的停止词列表去除了停止词。可比语料库的规模和词数如表 1 所示。

表 1. 语料规模

语料	Token 数量
中文	11.5M
英文	12.9M

其他需要配置的参数包括共现词的窗口尺寸、词与词的关联性度量方法和向量的相似性度量方法。[Laroche et al. 2010]详细地讨论了各种参数对最终获取的双语字典质量的影响。参照其在文章的结论，本文窗口尺寸定义为 3（左右各三个词），词与词的关联性度量方法选择了 Discounted Log-Odds Ratio[Laroche et al.,2010]。为了减少低频词对性能的影响，我们在构造上下文向量时需保证目标词和共现词的共现次数不小于 5，源语言词向量的非 0 维数目不小于 5。这样就得到了源语言和目标语言的上下文向量 *VS* 和 *VT*。

我们使用了一个本研究室内部编纂的汉英方向通用领域的翻译字典作为种子字典。其具体情况参见表 2。

表 2. 种子字典信息表

总条数	45869
中文词数	30566
平均词条	1.5
最多翻译数	22

我们从种子字典中选取中文端在 *VS* 中出现，英文端在 *VT* 中出现的词，共计 600 个词，每个词只有唯一的翻译，其中 300 个词做 validation set，剩余 300 个词做测试集。
在使用标准实现方法和本文提出的方法后，可以得到 *VS'*和 *VT*，使用 cosine 距离来计算向量 *VS'*和 *VT* 的相似性程度。

本文选择 Precision 和 MRR（mean rank reciprocal）两个评价指标来衡量本文方法的有效性。

Precision 的定义为

$$P_N = \frac{\sum_t m(t)}{T}$$

其中，T 为测试样例个数，若 t 的正确翻译出现在前 N 个翻译候选中，则 $m(t) = 1$。P_N 值越大越好。

MRR 的定义为

$$MRR = \frac{1}{T} \times \sum_{i=1}^{T} \frac{1}{rank_i}$$

其中，T 为测试样例的数量，$rank_i$ 表示在第 i 个测试样例的翻译候选中，正确候选所在的排序，如果不在，则为 0。

5.2 实验结果和分析

本文提出的单独消歧方法可以给出每一个翻译的相关程度，根据这个值，可以对所有的

翻译进行排序，因此是用最相似的一个翻译还是使用多个翻译自然是一个值得关注的问题。为此，我们考虑了不同数量的翻译候选，记为 WN-Ti，i 表示翻译候选的数量。

此外，基于分布式表示的方法，词表示的维度是一个重要的参数，因此，本文也考虑了不同的维度值，本文使用了 25，50，100，200 这几个维度值。

产生词表示的工具 word2vector 也有一些参数需要配置，这里设置 min-count 为 1，contextsize 为 5，模型为 skip 模型，其余为默认配置。

实验结果如表 3 所示：

表 3. Top1 上 Validation set 性能

SA	0.096				
	WN-T1	WN-T3	WN-T5	WN-T7	Weighted
D25	0.056	0.093	0.1	0.096	0.103
D50	0.056	0.096	**0.106**	0.096	**0.116**
D100	0.056	0.096	0.1	0.096	0.103
D200	0.053	0.08	0.1	0.096	0.106

表 3 中的每一项表示P_1指标的性能。SA 表示标准实现方法的性能，Weighted 表示加权消歧方法的性能，DK 表示基于词表示的维度为 K。从表中看出，在使用单独消歧方法时，仅使用 Top1 时性能是下降的，在 Top5 时才有提升，当到 Top7 时，已经退化到标准实现方法了。而加权消歧方法则可以显著地提升基线系统的性能。根据在 validation set 上的性能，最终选择词表示的维度为 50，WN-Ti 设置$i = 5$。最终在测试集上的性能见表 4。

表 4. 测试集性能

	TOP1	TOP5	TOP20
SA	0.09/0.09	0.186/0.130	0.33/0.140
WN-T5	0.106/0.106	0.2/0.141	0.346/0.154
Weighted	**0.116/0.116**	**0.213/0.149**	**0.366/0.161**

表 4 中的每一项形如 A/B，其中 A 为P_N值，B 为 MRR 的值。从表 4 可以看出，基于加权消歧方法显著地提升了标准实现方法的性能，并在输出多个翻译候选时同时提高了 Precision 和 MRR。

6 相关工作

由于平行语料获取的代价非常高昂，近来，越来越多的研究者开始探索如何利用可比语料来获取双语翻译字典。代表性研究工作有[Rapp,1995;Fung,1998;Koehn et al.,2002;Haghighi et al., 2008;Daume III et al.,2011]。

一般认为，使用更具有表达能力的上下文向量，更能获取高质量的双语翻译字典。因此，为了能改进标准实现方法的性能，研究人员提出了很多改进上下文向量的方法。在这类工作中，文献[Chiao et al.,2002]使用了辅助的语言学资源作为特殊字典（specialized dictionaries），文献[Prochasson et al., 2011]将种子字典与音译词（transliterated words）进行融合来翻译上下文向量。然而，较少的工作关注由种子字典引起的上下文向量歧义性问题。[Hazem et al.,2012]通过利用词性信息和领域相关度量信息来过滤种子字典，但性能并未有显著改进。[Gassusier et al., 2004]利用典型关联分析（Canonical Correlation Analysis, CCA）方法来解决词的歧义性问题。[Bouramor et al., 2013]中使用 Word-Net 的方法进行种子字典的翻译消歧，与之相比，本文的方法不需要使用 WordNet 这样的语义资源。

7 结论

针对可比语料的双语字典抽取任务，本文提出了一种基于词的分布式表示的无监督上下文向量翻译消歧方法。具体地，使用词的分布式表示作为对上下文向量进行消歧的依据。在汉英方向的科技领域可比语料的双语抽取任务实验上，本文提出的方法显著地超过了标准实现方法。进一步的工作可在其他语言方向上进行更为详细地比较实验。

致谢

本文作者真诚地感谢三位审稿人负责任的审稿态度、专业的审稿意见和无价的修改意见。本文研究受国家自然科学基金面上项目（资助号：61173073）和国家国际科技合作专项（资助号：2014DFA11350）的支持。

参考文献

Ballesteros L, Croft W B. Phrasal translation and query expansion techniques for cross-language information retrieval[C]//ACM SIGIR Forum. ACM, 1997, 31(SI): 84-91.

Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]//Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.

Bouamor D, Popescu A, Semmar N, et al. Building Specialized Bilingual Lexicons Using Large-Scale Background Knowledge[J].EMNLP,2013

Bouamor D, Semmar N, France C, et al. Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora[J]. ACL. Sofia, Bulgaria (Citépage 35), 2013.

Chang P C, Galley M, Manning C D. Optimizing Chinese word segmentation for machine translation performance[C]//Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008: 224-232.

Chiao Y C, Zweigenbaum P. Looking for candidate translational equivalents in specialized, comparable corpora[C]//Proceedings of the 19th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 2002: 1-5.

Daumé III H, Jagarlamudi J. Domain Adaptation for Machine Translation by Mining Unseen Words[C]//ACL (Short Papers). 2011: 407-412.

Fung P. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora[M]//Machine Translation and the Information Soup. Springer Berlin Heidelberg, 1998: 1-17.

Gaussier E, Renders J M, Matveeva I, et al. A geometric view on bilingual lexicon extraction from comparable corpora[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 526.

Haghighi A, Liang P, Berg-Kirkpatrick T, et al. Learning Bilingual Lexicons from Monolingual Corpora[C]//ACL.2008, 2008: 771-779.

Harris Z S. Distributional structure[M]. Springer Netherlands, 1981.

Hazem, Amir, and Emmanuel Morin. Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora.LREC. 2012.

Koehn P, Knight K. Learning a translation lexicon from monolingual corpora[C]//Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9. Association for Computational Linguistics, 2002: 9-16.

Laroche A, Langlais P. Revisiting context-based projection methods for term-translation spotting in comparable corpora[C]//Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics, 2010: 617-625.

Mikolov T, Kombrink S, Burget L, et al. Extensions of recurrent neural network language model[C]//Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 5528-5531.

Mikolov T, Le Q V, Sutskever I. Exploiting Similarities among Languages for Machine Translation[J]. arXiv preprint arXiv:1309.4168, 2013.

Morin E, Prochasson E. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora[C]//Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. Association for Computational Linguistics, 2011: 27-34.

Och F J, Ney H. A systematic comparison of various statistical alignment models[J]. Computational linguistics, 2003, 29(1): 19-51.

Prochasson E, Fung P. Rare word translation extraction from aligned comparable documents[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 1327-1335.

Rapp R. Identifying word translations in non-parallel texts[C]//Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1995: 320-322.

Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 151-161