文章编号：1003-0077(2013)01-0015-06

基于双语依存关系映射的中英文词表构建研究

徐 华，刘丹丹，钱龙华，周国栋

（苏州大学 自然语言处理实验室，苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：基于上下文的双语词表构建方法是比较流行的基于可比较双语语料库的双语词表构建方法。特别地，依存上下文模型从句子的依存树上抽取词语的上下文特征，由于依存关系更能体现词语之间的共现关系，因而这种方法提高了构建双语词表的性能。该文在此基础上，进一步提出了依存关系映射模型，即通过同时匹配依存树中的上下文词语、依存关系类型和方向来实现双语词表的构建。在 FBIS 语料库上的实验表明，该方法在中文—英文和英文—中文两个方向上的双语词表构建上均取得了较好的性能，这说明了依存关系映射模型在双语词表构建中的有效性。

关键词：双语词表构建；依存上下文模型；依存关系映射

中图分类号：TP391　　**文献标识码**：A

Research on Bilingual Dependency Relationship Mapping for Chinese-English Lexicon Construction

XU Hua, LIU Dandan, QIAN Longhua, ZHOU Guodong

(Natural Language Processing Laboratory of Soochow University,
School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Currently context-based approach is a popular approach for constructing bilingual lexicons from comparable bilingual corpora. Specifically, the dependency context model extracts context features from a sentence's dependency tree. This model improves the performance of the bilingual lexicon construction since dependency relationships can better capture the co-occurrence relationship between words. Following this line, this paper further proposes a dependency relationship mapping model, which constructs bilingual lexicon by mapping dependency context words, dependency relationship types and directions simultaneously. The experiments on the FBIS corpus show that, our approach significantly outperforms a state-of-the-art system in bilingual lexicon construction from both Chinese-English and English-Chinese. This justifies the effectiveness of our dependency relationship mapping model on bilingual lexicon construction.

Key words: bilingual lexicon construction; dependency context model; dependency relationship mapping

1 引言

双语词表在机器翻译和跨语言信息检索等自然语言处理任务中发挥着重要作用。传统的双语词表构建方法是从大规模平行语料库中通过抽取词对齐信息得到双语词表[1]，该方法可获得较好的性能，然而获得高质量的大规模平行语料库需要大量的人力和昂贵的财力，因此对于许多语言对，并不存在这样的语料库。所以，近年来研究者都把研究重点转向了通过第三方中间语言或者非平行的可比较语料库来构建双语词表。

基于第三方中间语言构建双语词表的方法利用某一流行的语言（通常是英语）作为中间语言，通过

收稿日期：2011-06-21　**定稿日期**：2011-09-07

基金项目：国家自然科学基金资助项目(60873150,90920004,61003153)；江苏省自然科学基金资助项目(BK2010219)

作者简介：徐华(1988—)，男，硕士研究生，主要研究方向为信息抽取；刘丹丹(1987—)，女，硕士研究生，主要研究方向为信息抽取；钱龙华(1966—)，通信作者，男，副教授，硕士生导师，主要研究方向为自然语言处理。

现有的源语言—中间语言和中间语言—目标语言两个词表来构建源语言—目标语言的词表。该方法最早由 Tanaka 等[2]提出。Kaji 等[3]利用英语作为中间语言生成了日文—中文和中文—日文的词表。Shezaf 等[4]也利用英语这一中间语言通过加入非对齐签名(Non-Aligned Signatures,NAS)特征来改进西班牙语—希伯来语词表。

基于可比较语料库构建双语词表的方法基于这样一个假设:在可比较语料库中,意义相似的双语词语其上下文也应该相似[5]。Fung[6]从可比较语料库中抽取双语词语的上下文信息,利用词语的共现向量来计算它们之间的相似度。Garera 等[7]提出了依存上下文模型,即抽取词语在依存树中的前驱节点和后继节点词语作为其上下文。由于依存上下文很好地反映了词语和它的上下文词语之间的语法关系,摒弃了直接采用词汇上下文所带来的噪音,因而获得了较好的性能。Koehn 等[8]组合了诸如同源词、相似上下文、词频等特征,分析了这些特征的作用和贡献。不过,对于中英文词表构建来说,同源词等特征显然是不起作用的。

本文在依存上下文模型的基础上,提出了双语依存关系映射模型,即通过同时匹配依存关系类型和上下文词语来改进中英文词表抽取的性能。本文的后续组织结构如下:第 2 节回顾了中英文双语词表构建的相关工作;第 3 节详细阐述了本文的方法—中英文双语依存关系映射模型;第 4 节为实验结果与分析;最后是本文总结和工作展望。

2 中英文词表构建相关工作

由于中英文语言之间的差异性较大,目前中英文词表构建系统相对较少。Fung[6]从可比较语料库中抽取双语词语的上下文信息,利用在线词典与词语共现向量来计算相似度,并分析了多义词、中文分词与英文形态信息等中英文差异性特征对词表的影响,在中英文词表抽取上达到了 30% 的准确率。张永臣等[9]在 Web 上采集中英文语料库,采用空间向量模型抽取金融领域的双语词表,并分析了种子词表的选择对双语词表性能的影响。Haghighi 等[10]采用匹配典型相关分析(Matching Canonical Correlation Analysis, MCCA)模型构建了包括英文—中文在内的多种语言对的双语词表。

Fung[11]提出了上下文异质性(Context Heterogeneity)的概念,所谓上下文异质性就是指词语前后上下文中出现词语的个数信息,它反映了该词语在语料库中的分布特征。与之类似,Yu 等[12]利用依存异质性(Dependency Heterogeneity),即词语在某些依存关系类型中中心词或依赖词的差异性,来抽取双语词表。这种方法不需要种子词表来构建双语词表,主要利用词语在语料库中的统计信息来辨别词语,不过该方法的经验性太强且缺乏相关语言学方面的理论支撑。

3 基于双语依存关系映射的中英文词表抽取

从 Garera 等[7]和 Yu 等[12]的工作中可以看出,依存信息可以有效地提高双语词表构建的性能。本节首先利用依存上下文模型构建一个中英文双语词表抽取的基准系统,然后详细介绍了本文的双语依存关系映射模型。

3.1 基准系统

Garera 等[7]的依存上下文模型通过抽取词语在依存树中一定窗口内的上下文词语来构建特征向量。实验表明,当窗口大小为 ±2 时其性能最佳。按照 Garera 等[7]的方法,我们实现了本文的基准系统,具体方法是:

- 上下文抽取。首先抽取词语在依存树中的父节点(−1)、子节点(+1)、祖父节点(−2)和孙子节点(+2)上的相关词语,保留位于种子词表中的词语;
- 特征向量构造。利用词包模型生成上下文向量,并利用点互信息(Pointwise Mutual Information,PMI)来衡量向量中某一个词语的权重。点互信息定义如下:

$$\mathrm{PMI}(w,c)=\log_2\frac{\frac{N(w,c)}{N}}{\frac{N(w)}{N}\times\frac{N(c)}{N}} \quad (1)$$

其中,$N(w,c)$代表词语 w 与其上下文词语 c 的共现频率,$N(w)$和 $N(c)$分别指词语 w 和 c 的频率,N 指语料库的总词数。由于 PMI 值的大小存在倾向于词频较少词语的缺陷,因此我们在 PMI 公式后乘上了折扣因子(Discounting Factor)[13]作为某一特征的权值。

$$\frac{N(w,c)}{N(w,c)+1}\times\frac{\min(N(w),N(c))}{\min(N(w),N(c))+1} \quad (2)$$

- 相似度计算:利用余弦相似度(Cosine Simi-

larity)来计算双语词汇向量之间的相似度，并从目标语言中选择一个相似度值最高的词汇作为源语言词语的等价词汇。

$$Sim_{\mathrm{DW}}(S,T)=\frac{\sum_i \mathrm{PMI}_{S,i}\times \mathrm{PMI}_{T,i}}{\sqrt{\sum_i \mathrm{PMI}_{S,i}}\times\sqrt{\sum_i \mathrm{PMI}_{T,i}}} \tag{3}$$

其中 S 和 T 分别指源语言和目标语言词语的上下文向量，$\mathrm{PMI}_{S,i}$ 和 $\mathrm{PMI}_{T,i}$ 分别指第 i 个在种子词表中能匹配的源语言和目标语言的词语互信息值，Sim_{DW} 为双语词语依存上下文的相似度。

该模型利用了双语词语与种子词表中词语的共现程度来衡量相似度，由于采用词包模型，且只考虑了依存上下文中的词汇信息，忽略了其他关键信息，如依存关系类型等，因而其性能不够理想。

3.2　双语依存关系映射模型

中英文双语依存关系类型存在着一定的对应关系，Lin[14] 提出了一种基于依存路径转换的机器翻译模型，根据依存路径创建转换规则，把源语言的依存路径转换为目标语言的依存树片段。基于他的工作，我们发现在中英文双语语料库中词汇之间的依存信息可以很好地进行匹配。图 1 举例说明了中英文之间的依存关系类型的映射关系。

从图 1 可以看出，显然在两个平行句子中，对应词语及其依存关系大都可以很好地匹配。通过对双语词汇的上下文进行观察，我们发现，对于一个双语等价翻译对，与其共现的上下文词语和依存关系类型也能够进行匹配。如表 1 所示，“业绩”和其等价翻译词“achievement”的上下文中，它们的依存关系类型和上下文词语就可以很好地匹配。不过，由于中英文语言之间的差异性和标记集的不同，并不是所有的依存关系类型可直接匹配，有些依存关系可能对应另外一种语言的多种依存关系。例如，中文依存关系 nn，可以匹配英文依存关系中的 amod、nn 和 prep_of。需要说明的是，虽然一种语言的依存关系可能映射到另一种语言的多种依存关系，但在实际匹配时，由于在一个句子中一对词语之间的依存关系是唯一的，因此只能选择一种依存关系进行匹配。

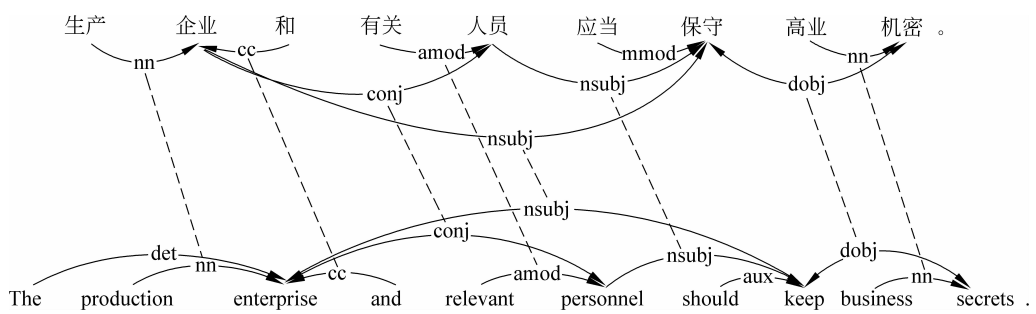

图 1　中英文依存关系类型映射关系

表 1　“业绩”和“achievement”的依存上下文中依存关系类型和上下文词语的匹配

业绩	Achievement
中文上下文	英文上下文
dobj_创造	dobj_create
conj_经验	conj_experience
nn_经营	nn_operation
amod_伟大	amod_great
nn_管理	nn_management

通过分析中英文两种语言各自依存关系的特点，我们得到了中文—英文和英文—中文的依存类型的映射关系，如表 2 和表 3 所示。根据这些依存类型的映射关系，我们抽取了带有依存关系类型的上下文词汇作为上下文特征，并且在特征匹配时两者都必须匹配。需要注意的是，依存关系直接发生在一对词语之间，因此，此时的窗口大小为±1。与基准系统类似，我们仍然采用点互信息来衡量带依存关系的上下文向量的权重，并计算其余弦相似度。此时，双语之间的相似度同时考虑基准系统中的依存上下文特征和依存关系映射特征，其计算公式如式(4)：

$$Sim_T(S,T)=\alpha\times Sim_{\mathrm{DW}}(S_1,T_1)+(1-\alpha)\times Sim_{\mathrm{DRM}}(S_2,T_2) \tag{4}$$

其中，Sim_{DW} 是指在基准系统的依存上下文模型中，双语词语之间的相似度，Sim_{DRM} 指在依存关系映射模型中的相似度，而 Sim_T 为总的相似度。S_1，T_1 分别表示在基准系统中的双语词语的依存上下文向量，而 S_2，T_2 则表示包含依存关系类型的依

存上下文向量，α 为复合参数。根据实验测试，当 $\alpha=0.8$ 时系统性能最好。

表2 中文—英文的依存关系映射

中文依存关系	描述	英文依存关系	描述
nsubj	名词性主语	nsubj	名词性主语
dobj	直接宾语	dobj	直接宾语
conj	连词	conj	连词
nn	名词修饰	amod	形容词修饰
		nn	名词修饰
		prep_of	介词“of”修饰
amod	形容词修饰	amod	形容词修饰
assmod	“的”修饰词	prep_of	介词“of”修饰
		poss	所有格修饰

表3 英文—中文的依存关系映射

英文依存关系	描述	中文依存关系	描述
nsubj	名词性主语	nsubj	名词性主语
dobj	直接宾语	dobj	直接宾语
conj	连词	conj	连词
nn	名词修饰	nn	名词修饰
amod	形容词修饰	amod	形容词修饰
		nn	名词修饰
prep_of	介词“of”修饰	nn	名词修饰
		assmod	“的”修饰词
poss	所有格修饰	nn	名词修饰
		assmod	“的”修饰词

4 实验与分析

本节首先介绍了本文实验所使用的语料库，然后详细说明了种子词表和测试词表的生成方法，最后分别讨论了不同依存关系类型和各种不同特征对构建中英文双语词表性能的影响。

4.1 语料库

我们以中英文“对外广播信息服务”(Foreign Broadcast Information Service，FBIS) 平行语料库作为双语词表抽取的训练和测试语料库。FBIS 是新闻领域语料库，包含约 24 万句平行句对，约 690 万中文词，890 万左右英文词。我们把 24 万句语料库分成两部分：11 万句和 13 万句，利用中文语料的第一部分和英文语料的第二部分构成非平行的可比较语料库。此方法与 Haghighi 等[10] 和 Ismail 等[15] 构建可比较语料库的方法类似，是常见的从平行语料库中提取非平行的可比较语料库的方法。

对于语料库的预处理，我们首先对语料库进行句法分析，使用 Stanford Parser[16] 获取依存关系和词性信息。由于英文中存在名词复数、动词时态语态等形态特征，我们对英文语料库进行形态处理以获取英文词语的原型形式。

4.2 种子词表和测试词表

种子词表是已知对齐的双语词表，它是构建新的双语词表的基础。在上下文模型中，利用待对齐的双语词语与种子词表中的已知词语的搭配信息来计算双语词语之间的上下文相似度，并通过选择相似度最高的词语来构建双语词表。大多数基于上下文的双语词表构建方法都使用种子词表来匹配上下文词语，例如，Rapp[5] 和 Fung[6] 均使用规模在 20k 左右的词典作为种子词表，而 Haghighi 等[10] 和 Ismail 等[15] 都使用 100～1 000 左右的小型种子词表。与 Haghighi 等[10] 和 Ismail 等[15] 类似，我们也试图在小型种子词表的基础上提高双语词表构建的性能。我们通过对齐 FBIS 语料库并去掉停用词后，获取频率最高的 1 000 个词作为我们的种子词表。

我们选取名词作为测试词表。在去除种子词表包含的名词后，选取频率最高的 500 个名词作为测试词表。在目标语言中，选取 5 000 个名词作为候选词与测试词语进行匹配，即 5 000 个词语中与测试词语相似度最大的词作为测试词语的等价翻译词。

4.3 评价标准

我们采用准确率(Precision)和平均排名倒数(Mean Reciprocal Rank，MRR)作为评价标准[12]。准确率是双语词表构建中常用的评价标准，指的是在相似度最高的前 n 个候选词中的平均准确度。MRR 是指正确翻译词在候选词中排名倒数的平均值，衡量正确翻译词的相似度在候选词中的排名次序。本文中准确率只考虑相似度最高的一个候选词的情况，定义如下：

$$\text{Precision} = \frac{\text{count}_{\text{top1}}}{N} \tag{5}$$

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i} \tag{6}$$

其中，$count_{top1}$ 指相似度最高的一个候选词中正确的个数，$rank_i$ 是正确翻译词在候选词中的排名，N 是测试词表的个数。与准确率不同，MRR 不需要考虑 n 的大小，因而更能全面地衡量自动构建出来的双语词表的性能。

4.4　实验结果与分析

• 不同依存类型对抽取性能的影响

表 4 列出了在中文—英文和英文—中文两个方向构建词表时，不同依存关系类型对性能的影响。为了提高计算效率，我们在基准系统的基础上采取了重排序的策略，即在基准系统的结果中选取相似度最高的 50 个候选词，添加后续特征后重新计算测试词语与该 50 个候选词的相似度。参考 Stanford Parser 的依存关系类型，我们将上述依存关系映射特征分为论元关系（Argument）、连接关系（Conjunction）和修饰关系（Modifier）三大类进行排序，并采用累加的方式逐步添加到系统中，即每一种依存关系映射特征按照相应顺序逐一添加到系统中。

表 4　采用双语依存关系映射的中英文词表抽取性能

中文—英文			英文—中文		
特征	准确率/%	MRR/%	特征	准确率/%	MRR/%
基准系统	41.8	51.34	基准系统	33.8	42.23
＋nsubj-nsubj	35.2	46.11	＋nsubj-nsubj	31.6	40.92
＋dobj-dobj	39.8	50.07	＋dobj-dobj	35.8	45.13
＋conj-conj	41.6	52.66	＋conj-conj	38.8	47.50
＋nn-amod	42.6	53.15	＋nn-nn	40.0	49.56
＋nn-nn	43.4	53.86	＋amod-amod	42.2	50.54
＋nn-prep_of	44.2	54.68	＋amod-nn	41.8	50.65
＋amod-amod	44.2	54.81	＋prep_of-nn	41.8	50.96
＋ assmod-prep_of	45.0	55.36	＋prep_of-assmod	42.6	51.58
＋ assmod-poss	45.0	55.38	＋poss-nn	42.8	51.62
—	—	—	＋poss-assmod	43.0	51.89

从表 4 中可以看出，在开始添加特征时，性能有所下降，这是因为在少量特征下，上下文向量较稀疏，不足以区分词语的语义，反而会引入噪音，导致了性能的降低，但随着加入特征的增多，上下文逐渐丰富，性能也逐渐提高。最后，中文—英文的总体性能 Precision 和 MRR 分别比基准系统高出 3.2 和 4.04，而英文—中文词表的总体性能 Precision 和 MRR 分别比基准系统高出 9.2 和 9.66。这说明依存关系映射特征能显著提高中英文词表构建的性能。另外，虽然由于中文词性的歧义性，使得英文—中文的基准系统性能明显低于中文—英文基准系统的性能，但是双语依存关系映射特征能很好地弥补这一缺陷，从而大幅度地提高其词表构建的性能。

• 不同特征对性能的影响

表 5 考察了不同特征对中英文双词词表构建性能的影响，其中①为基准系统，②为仅使用依存关系映射特征，第 3 行表示依存上下文特征和依存关系映射特征的线性复合（即式(4)），第 4 行表示在第 3 行基础上再考虑位置特征，即在匹配词语和依存类型时，还要同时考虑依存方向。

表 5　不同特征对双语词表构建性能的影响

特征	中文—英文		英文—中文	
	准确率/%	MRR/%	准确率/%	MRR/%
①基准系统	41.8	51.34	33.8	42.23
②依存关系映射特征	35.4	47.79	33.2	41.67
①＋②	45.0	55.38	43.0	51.89
①＋②＋方向特征(－1/＋1)	46.0	56.18	44.0	52.36

表 5 的实验结果表明，单独使用依存关系映射特征时，无论是中文—英文还是英文—中文的双语词表构建，其性能均低于基准系统，这是由于同时匹配词语和依存关系会导致特征更加稀疏而引起的。

另外，Garera 等[7]的实验表明，在依存上下文模型中，共现词语的依存方向对词表构建性能没有促进作用。我们在中英文两个方向的词表构建实验表明，在依存关系映射模型中，方向特征均能提高1个点的准确率，MRR 值也都有所提高。这说明在依存类型匹配的前提下，依存方向特征有助于双语词表的构建。

5 结论与展望

本文提出了基于依存关系映射模型的中英文双语词表构建方法，即在依存上下文模型的基础上增加了依存关系映射特征，它包含了依存上下文词语及其类型和方向等三个因素，因而可以更准确地反映双语等价翻译词之间的对应关系。实验表明，双语依存关系映射模型在中英文两个方向的双语词表构建上都取得了较好的效果，显著提高了双语词表抽取的性能，同时也表明了该方法对不同语言对具有潜在的适用性。

目前的双语依存关系映射是通过人工的特征工程方法来实现的，其映射特征并非是最佳特征，也较难应用到不同的语言对上。因此在下一步工作中，我们将利用机器学习的方法自动发掘语言对之间的依存映射关系，进一步提高系统的性能和领域适用性。

参考文献

[1] Dekai Wu, Xuanyin Xia. Learning an English-Chinese Lexicon from a Parallel Corpus[C]//Proceedings of the 1st Conference of the Association for Machine Translation in the Americas, Columbia, Maryland, 1994: 206-213.

[2] Kumiko Tanaka, Kyoji Umemura. Construction of a bilingual dictionary intermediated by a third language [C]//Proceedings of Conference on Computational Linguistics. 1994.

[3] Hiroyuki Kaji, Shin'ichi Tamamura, Dashtseren Erdenebat. Automatic construction of a Japanese-Chinese dictionary via English[C]//Proceedings of the 6th Edition of the Language Resources and Evaluation Conference. Marrakech, Morocco, 2008: 699-706.

[4] Daphna Shezaf, Ari Rappoport. Bilingual Lexicon Generation Using Non-Aligned Signature[C]//Proceedings of ACL 2010. Uppsala, Sweden, 2010: 98-107.

[5] Reinhard Rapp. Automatic identification of word translations from unrelated English and German corpora[C]//Proceedings of ACL, 1999: 519-526.

[6] Pascale Fung. A statistical view on bilingual lexicon extraction: from parallel corpora to nonparallel corpora [C]//Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas. 2000.

[7] Nikesh Garera, Chris Callison-Burch, David Yarowsky. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences[C]//Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL), Boulder, Colorado, June 2009: 129-137.

[8] Philipp Koehn, Kevin Knight. Learning a translation lexicon from monolingual corpora[C]//Proceedings of ACL Workshop on Unsupervised Lexical Acquisition, 2002.

[9] 张永臣，孙乐，李飞，等. 基于 Web 数据的特定领域双语词典抽取[J]. 中文信息学报，2006，20(2)：16-23.

[10] Aria Haghigi, Percy Liang, Taylor Berg-Krikpatrick, et al. Learning bilingual lexicons from monolingual corpora[C]//Proceedings of the ACL, Ohio, USA, 2008: 771-779.

[11] Pascale Fung. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus[C]//Proceedings of 3rd Annual Workshop on Very Large Corpora. Boston, Massachusetts: Jun. 1995: 173-183.

[12] Kun Yu, Junichi Tsujii. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity[C]//Proceedings of NAACL-HLT, short papers, 2009: 121-124.

[13] Dekang Lin, Patrick Pantel. Concept Discovery from Text[C]//Proceedings of Coling 2002: 42-48.

[14] Dekang Lin. A path-based transfer model for machine translation[C]//Proceedings of Coling 2004, Geneva, Switzerland, 2004: 625-630.

[15] Azniah Ismail, Suresh Manandhar. Utilizing contextually relevant terms in bilingual lexicon extraction [C]//Proceedings of Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, Boulder, Colorado, USA, 2009: 10-17.

[16] M-C de Marneffe, B MacCartney, C D Manning. Generating typed dependency parses from phrase structure parses[C]//Proceedings of LREC 2006.