软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(7):1615−1625 [doi: 10.13328/j.cnki.jos.004630]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于双语协同训练的最大名词短语识别研究*

李业刚[1,2,3], 黄河燕[1,2], 史树敏[1,2], 鉴 萍[1,2], 苏 超[1,2]

[1](北京理工大学 北京市海量语言信息处理与云计算应用工程技术研究中心,北京 100081)

[2](北京理工大学 计算机学院,北京 100081)

[3](山东理工大学 计算机科学与技术学院,山东 淄博 255049)

通讯作者: 黄河燕, E-mail: hhy63@bit.edu.cn, http://cs.bit.edu.cn

摘 要: 针对传统方法对双语最大名词短语识别一致性差以及跨领域识别能力弱的缺点,提出一种基于半监督学习的双语最大名词短语识别算法.利用汉英最大名词短语的互译性和识别的互补性,把平行的汉语句子和英语句子这两个数据集看作一个数据集的两个不同的视图进行双语协同训练.在协同训练中,把双语对齐标注一致率作为标记置信度估计依据,进行增量标记数据的选择.实验结果表明:该算法显著提高了双语最大名词短语的识别能力,在跨领域测试和同领域测试中,F 值分别比目前最好的最大名词短语识别模型提高了 4.52%和 3.08%.

关键词: 最大名词短语;半监督学习;标注投射;双语协同训练;短语识别

中图法分类号: TP18

中文引用格式: 李业刚,黄河燕,史树敏,鉴萍,苏超.基于双语协同训练的最大名词短语识别研究.软件学报,2015,26(7):1615−1625. http://www.jos.org.cn/1000-9825/4630.htm

英文引用格式: Li YG, Huang HY, Shi SM, Jian P, Su C. Title recognition of maximal-length noun phrase based on bilingual co-training. Ruan Jian Xue Bao/Journal of Software, 2015,26(7):1615−1625 (in Chinese). http://www.jos.org.cn/1000-9825/4630.htm

Title Recognition of Maximal-Length Noun Phrase Based on Bilingual Co-Training

LI Ye-Gang[1,2,3], HUANG He-Yan[1,2], SHI Shu-Min[1,2], JIAN Ping[1,2], SU Chao[1,2]

[1](Beijing Engineering Applications Research Center of High Volume Language Information Processing and Cloud Computing, Beijing Institute of Technology, Beijing 100081, China)

[2](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

[3](College of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China)

Abstract: This article focuses on the problem of weak cross-domain ability on bilingual maximal-length noun phrase recognition. A bilingual noun phrase recognition algorithm based on semi-supervised learning is proposed. The approach can make full use of both the English features and the Chinese features in a unified framework, and it regards the two language corpus as different view of one dataset. Instances with the highest confidence score are selected and merged, and then added to the labeled data set to train the classifier. Experimental results on test sets show the effectiveness of the proposed approach which outperforms 4.52% over the baseline in cross-domain, and 3.08% over the baseline in similar domain.

Key words: maximal-length noun phrase; semi-supervised learning; label projection; bilingual co-training; phrase identification

最大名词短语(maximal-length noun phrase,简称 MNP)是不被其他任何名词短语所包含的名词短语,在句子中有稳定的外部修饰结构,是一个完整的句法单元和语义单元[1].MNP 的分析和研究,是机器翻译、信息检索和信息抽取等任务中的一个有重要价值的难题.从 20 世纪 90 年代开始,国内外众多的研究者开始对 MNP 的自动

* 基金项目: 国家重点基础研究发展计划(973)(2013CB329300); 国家自然科学基金(61132009, 61201352, 61202244)
收稿时间: 2014-02-23; 修改时间: 2014-04-14; 定稿时间: 2014-05-21

识别进行研究,其中,在英语 MNP 自动识别研究中,文献[2]提出的 NPtool 是一种基于规则方法的 MNP 自动获取工具;文献[3]则提出在浅层句法分析的基础上,利用有限状态分析识别句子中的 MNP.汉语 MNP 的自动识别研究中,文献[1]提出了组合内部结构的 MNP 识别算法,识别的正确率达到 85.4%;文献[4]提出了一种基于前后向"分歧点"的概率融合策略,利用支持向量机(support vector machine,简称 SVM),融合了前后向 MNP 标注的结果,这也是迄今为止公开的 MNP 自动识别研究中,基于单个分类器的最好的结果,其 F 值比基于 SVM 的基线模型提高了 1.05 个百分点.

综上我们发现:现有的 MNP 的研究都是针对单一语言的,仅采用单一语言的各种特征来提高 MNP 的识别性能,不同语言之间 MNP 识别的一致性比较差,这很大程度上影响了 MNP 等价对的进一步提取.实际上,不同语言的 MNP 识别利用了各自语言的不同特征,具有互补性,若取长补短,充分利用,可以提高 MNP 的识别性能和识别的一致性.另外,现有研究成果中,自动识别 MNP 效果比较好的都是基于有监督的学习方法.有监督的学习方法虽然在诸多自然语言处理任务中都有良好的表现,但其存在两个明显的不足之处:其一,需要大量的已标注数据来保证学习的准确性;其二,当已有的标注数据与待判定的数据不属于同一个领域时,有监督学习算法的性能会明显下降.改进这两个不足的方法之一就是采用半监督的学习方法,半监督学习方法在词性标注[5]、语义角色标注[6]和情感分类[7]等自然语言处理任务中都有初步的应用.

我们的研究立足于利用不同语言之间 MNP 识别的互补性,取长补短,引入了典型的半监督学习算法——协同训练,提出了双语协同训练算法,算法的框架如图 1 所示.以汉英双语 MNP 识别为例,把平行的汉语句子和英语句子这两个数据集看作一个数据集的两个不同的视图,融合汉英双语特征进行协同训练.从公开发表的论文来看,与本文最相近的研究是文献[7]的利用双语协同训练进行情感词分类,但是文献[7]在双语标注转换时采用的是词典翻译的方式,相比之下,MNP 往往是由多个词组成的,粒度更大,简单的词典翻译和词对齐等方法无能为力.因此,我们在标注投射过程中使用一个对数线性模型,修正标记示例的投射错误,减少另外一个分类器的噪音引入.

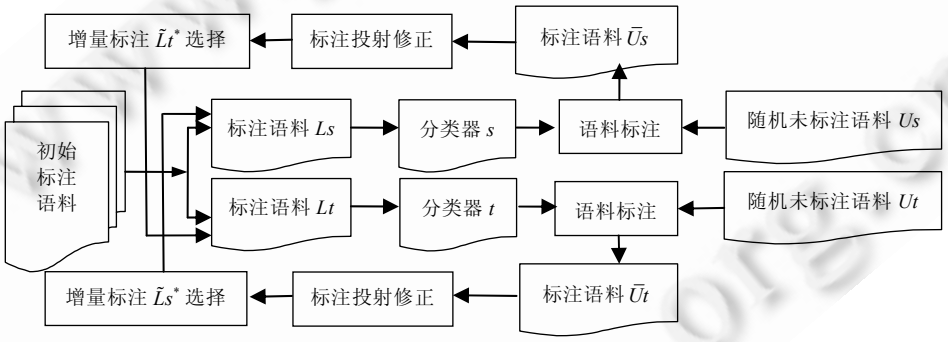


Fig.1 Framework for proposed algorithm

图 1 协同训练算法框图

1 双语 MNP

双语 MNP 要具备可互译性和双语的一致性,这对机器翻译有着重要的意义,也是我们进行 MNP 双语协同训练的基础.双语 MNP 涉及到两种不同的语言,与普通的单语 MNP 相比,有额外的约束条件:其一,双语 MNP 是句法独立的,不存在一端名词短语连续,另一端不连续的现象;第二,双语 MNP 在双语句子中句法功能相同,即,一端是名词短语,另一端也应具有名词的功能.我们对 CTB 英汉双语对照树库(English Chinese Translation Treebank V1.0)的 4 175 个句子进行了相应的统计.分析发现,汉英 MNP 具有较好的稳定性,98%以上都符合上述约束.Koehn[8]在英语-德语、汉语-英语和葡萄牙语-英语等语言对之间也做过类似的统计,虽然统计语料来源不同,但统计结果相似,这也说明等价双语 MNP 的研究适用于更大范围的语言对.限于篇幅,我们仅以汉英 MNP 对

为例进行更深入的研究.

双语 MNP 在单语中可能被其他名词短语包含,但是它不能被可互译的其他名词短语包含.也就是说,它的最大是双语一致可互译下的最大.区别于单语的 MNP,双语 MNP 可形式化定义为:

定义 1. 对于句对 $SP=\langle S,T\rangle$,S 表示汉语文本序列,$S:ws_1ws_2\ldots ws_{ns}$,T 表示英语文本序列,$T:wt_1wt_2\ldots wt_{nt}$,其中,ns 和 nt 分别表示句子 S 和 T 的长度.若存在 $\langle Mc,Me\rangle$,$Mc\subset S$,$Me\subset T$,并且满足下列条件,则称 $\langle Mc,Me\rangle$ 为双语 MNP:

$$\{\langle Mc,Me\rangle|Mc=ws_0,ws_1,\ldots,ws_m,Me=wt_0,wt_1,\ldots,wt_n;Mc\leftrightarrow Me;m<ns,n<nt\}.$$

(1) 非空性:$Mc\neq null$,$Me\neq null$;

(2) 互译性:$Mc\leftrightarrow Me$,Me 和 Mc 具有翻译上的转换充分性;

(3) 代表性:Mc 和 Me 的语义核心由一个或多个名词组成,该语义核心的成分特征决定了整个 MNP 短语结构的特征;

(4) 最大性:不存在另外一个 $\langle\overline{Mc},\overline{Me}\rangle$,$\overline{Mc}\subset S,\overline{Me}\subset T$, 且满足 $Mc\subseteq\overline{Mc},Me\subseteq\overline{Me}$.

例如:汉英句对“确保了浦东开发的有序**进行**”和“ensuring the orderly **advancement** of Pudong's development”的双语 MNP 识别结果为“确保/O 了/O 浦东/BS 开发/IS 的/IS 有序/IS **进行/IH**”,“ensuring/O the/BS orderly/IS **advancement/IH** of/IS Pudong/IS 's/IS development/IS”.汉语 MNP“浦东开发的有序进行”和英语 MNP“the orderly advancement of Pudong's development”可以互译,具转换充分性.汉语和英语 MNP 的语义核心词“进行”和“advancement”都是名词,可以代表对应 MNP 的特征.

2 双语协同训练算法

协同训练算法(co-training)[9]是一种典型的半监督学习方法,最初的协同训练是在一个数据集的两个充分冗余的视图上利用有标记示例分别训练分类器,在协同训练过程中,每个分类器从未标记示例中挑选出对示例赋予正确标记的置信度高的示例进行标记,并把标记后的示例加入另一个分类器的有标记训练集,训练过程不断迭代,达到某个停止条件为止.该算法的不足在于“充分冗余视图”这一限定条件在实际问题中往往难以满足.文献[10]的研究则证明:“充分冗余”这个条件即便是不完全满足,协同训练算法也可以在一定程度上提升分类器的性能.文献[11]提出了一种在同一个属性集上训练两个不同的分类器的协同训练算法,该算法不再要求“充分冗余的视图”,但不足之处在于:算法在少量标记数据上进行 10 倍交叉验证,经常难以稳定地估计置信度;同时,算法的泛化能力也有所降低.文献[12]提出了在同一个属性集上使用同类型分类器的 tri-training 算法,该算法首先对有标记示例集进行可重复取样,生成 3 个有标记训练集,在每个训练集上分别训练产生分类器;然后,用其中两个分类器对同一个未标记示例进行预测,如果预测结果相同,则认为该示例标记置信度较高,加入第 3 个分类器的有标记训练集中.tri-training 算法的优点是不再显式地估计置信度,而是通过多个分类器的预测结果的一致性来隐式地估计标记置信度,提高了算法的泛化能力.tri-training 算法的不足之处在于:如果初始分类器比较弱,隐式估计往往不够准确,容易错误标记未标记示例,从而给第 3 个分类器引入噪声数据.

本文提出的双语协同训练算法在现有丰富的汉英平行语料基础上,利用双语 MNP 的互译性和识别的互补性,把平行的汉语句子和英语句子这两个数据集看作一个数据集的两个不同的视图进行双语协同训练.不同于普通的协同训练任务,双语协同训练过程中还有一个标记投射的问题.由于语言的差异性和词对齐技术的差强人意,单纯依靠词对齐进行词投射得到对应 MNP 的方法性能比较差.因此,我们在投射过程中使用一个对数线性模型修正投射标记,降低标记示例的投射错误,减少另外一个分类器的噪音引入,从而提高协同训练的质量.另外,在利用分类器对未见示例进行预测时,我们引入双语对齐标注一致率作为标记置信度估计的衡量指标,摆脱了对小样本标记数据的依赖,隐式估计标记置信度,从而提高了算法的泛化能力,具有领域适应性.算法描述如算法 1 所示.

算法 1. 双语 MNP 协同训练算法.

1. 已知条件:

1.1. 汉英句子级别对齐的已标注语料集合 Ls,Lt;

1.2. 汉英句子级别对齐的未标注语料集合 Us,Ut.

2. 初始化分类器:

2.1. 在 Ls 上训练分类器 $Classifier(s)$;

2.2. 在 Lt 上训练分类器 $Classifier(t)$.

3. 循环 m 次:

3.1. 从 Us 和 Ut 中抽取 n 个对齐的句子,分别利用分类器 $Classifier(s)$和 $Classifier(t)$进行标注,形成 $\bar{U}s$ 和 $\bar{U}t$, 计算 $conformity\_ration(\bar{U}s,\bar{U}t)$,$\max \leftarrow conformity\_ration(\bar{U}s,\bar{U}t)$, 初始化标注语料增量集合 $\tilde{L}t^* \leftarrow null, \tilde{L}s^* \leftarrow null$;

3.2. 循环 n 次:

3.2.1. 随机地从 $(\bar{U}s,\bar{U}t)$ 中抽取 k 个句对形成 $(\hat{L}s,\hat{L}t)$, 依据词对齐原则从 $\hat{L}s$ 到 $\hat{L}t$ 进行标注投射,投射结果融合 $\hat{L}t$ 修正后形成 $\tilde{L}t$;

3.2.2. 在 $Lt \cup \tilde{L}t$ 上重新训练分类器, $classifier(t) \leftarrow classfier(Lt \cup \tilde{L}t)$;

3.3.3. 利用分类器 $Classifier(t)$对 $\bar{U}t$ 进行标注,重新计算 $conformity\_ration(\bar{U}s,\bar{U}t)$, 如果 $conformity\_ration(\bar{U}s,\bar{U}t) > \max$, 则 $\max \leftarrow conformity\_ration(\bar{U}s,\bar{U}t), \tilde{L}t^* \leftarrow \tilde{L}t$;

3.3. $Lt \leftarrow Lt \cup \tilde{L}t^*$, 在 Lt 上重新训练分类器 $classifier(t) \leftarrow classifier(Lt)$;

3.4. 循环 n 次:

3.4.1. 随机地从 $(\bar{U}s,\bar{U}t)$ 中抽取 k 个句对形成 $(\hat{L}s,\hat{L}t)$, 依据词对齐原则从 $\hat{L}t$ 到 $\hat{L}s$ 进行标注投射,投射结果融合 $\hat{L}s$ 修正后形成 $\tilde{L}s$;

3.4.2. 在 $Ls \cup \tilde{L}s$ 上重新训练分类器 $classifier(s) \leftarrow classfier(Ls \cup \tilde{L}s)$;

3.4.3. 利用分类器 $Classifier(s)$对 $\bar{U}s$ 进行标注,重新计算 $conformity\_ration(\bar{U}s,\bar{U}t)$, 如果 $conformity\_ration(\bar{U}s,\bar{U}t) > \max$, 则 $\max \leftarrow conformity\_ration(\bar{U}s,\bar{U}t), \tilde{L}s^* \leftarrow \tilde{L}s$;

3.4.4. $Ls \leftarrow Ls \cup \tilde{L}s^*$, 在 Ls 上重新训练分类器 $classifier(s) \leftarrow classifier(Ls)$.

3 双语对齐标注一致率

在协同训练过程中,一旦某个增量标注出错,这个错误将被将继续学习和加强,导致算法性能下降.这就需要采取有效的措施减少噪声数据引入.双语 MNP 具备互译性,也就是说,正确识别的汉英 MNP 应该具有标注的一致性.所以我们在词对齐的基础上,以对齐词的 BIO 标注一致率作为衡量指标,选择对齐标注一致率最高的部分(top z)作为增量标记数据.对齐标注一致率的计算如公式(1)所示.

$$conformity\_ratio = \frac{1}{n}\sum_{U}\frac{1}{K}\sum_{k=1}^{K} conformity(ws_i, wt_j)_k \tag{1}$$

其中, $conformity(ws_i, wt_j)_k = \begin{cases} 1, & T(ws_i) = T(wt_j) \\ 0, & T(ws_i) \neq T(wt_j) \end{cases}$, $(ws_i, wt_j)_k$ 表示平行句对的第 $k(1 \leqslant k \leqslant K)$个词对;$T(ws_i)$,$T(wt_j)$分别表示 MNP 汉英两端的 BIO 标记;$U$ 表示未标注语料集;n 表示 U 中的句子数.由于汉语和英语在语序上有较大的差异,在对齐标注一致率计算时我们忽略标记"*B*"和"*I*"的差异,认为它们是相同的标记.以 Lt 的增量标注选择为例,双语对齐标注一致率的获取过程如框图 2 所示.

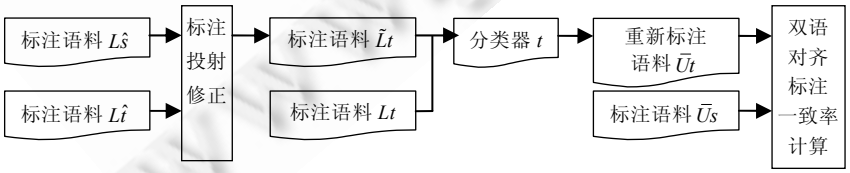


Fig.2 Framework of conformity-ration acquisition

图 2 双语对齐标志一致率的获取过程框图

文献[11]采用的是在少量标记数据上进行 10 倍交叉验证来进行增量标记数据的选择.这种显式的估计方法的不足之处在于:增量标记数据的选择依赖于少量的标记数据,这会降低算法的泛化能力.为了增强算法的领域适应性,我们采用了类似文献[12]的隐式估计方法,在每次迭代更新的未标注的数据集上进行验证,把在未标注集上对齐标注一致率最高的增量标记数据作为最优增量标记数据.

4 标记投射修正模型

双语协同训练不同于普通的协同训练,两个分类器分别是在不同的语言上训练.因此,一个分类器标记后的示例不能直接加入另一个分类器的有标记训练集,还需要在两个不同语言句子之间进行标注投射.由于汉英之间的语言差异较大,MNP 粒度又比较大,仅依靠词对齐进行源语言到目标语言的标注投射获得目标语言 MNP,结果会不尽如人意.为了提高标注投射的可靠性,我们融合目标语言 MNP 的单语特征和双语 MNP 的对齐特征,对投射结果进行修正.我们首先对从源语言到目标语言的 MNP 投射区域进行扩展,使之容纳更多的目标语言 MNP 假设,每个 MNP 投射假设与源语言 MNP 组成一个双语 MNP 假设;然后,我们构造一个线性对数模型,融合目标语言名词短语的句法置信度和双语 MNP 的对齐置信度,对所有的双语 MNP 假设综合打分;最后,通过一个贪心搜索得到句对最优的双语 MNP 假设集合.源语言在目标语言端的最优投射结果就是与源语言名词短语组成最优双语 MNP 假设的那个目标语言 MNP.

4.1 投射MNP扩展

源语言 MNP 表示为 Mc_{c1}^{c2},通过词对齐方式,得到目标语言端连续的且包含投射中心词的中心词块作为最小候选区域 Me_{a1}^{a2},把包含所有投射词的投射区域 $\overline{Me}_{b1}^{b2}$ 的两端分别向外扩展 4 个词(到达句首或者句尾可能不到 4 个词)作为最大候选区域.

在目标语言端建立一个滑动窗,从最小候选区域出发,不断向句子任意一侧扩充词,直至达到最大候选区域边界为止,从而扩展产生一系列的目标语言端候选 MNP 假设.每个目标语言端 MNP 假设与 Mc_{c1}^{c2} 组合,形成一个双语 MNP 假设,表示为 $H_k=(Mc,M\tilde{e})$.

4.2 MNP单语句法置信度

周强[1]的实验证明边界分布信息特征在 MNP 自动识别中的有效性.为了确保目标语言端 MNP 投射满足名词短语的句法特征,我们也选用了左右边界分布概率作为目标语言 MNP 的句法置信度.边界分布概率包含了左边界二元词性共现频率和右边界二元词性共现频率.

- 左边界二元词性共现频率定义如公式(2)所示.

$$P(Mxl\mid M\tilde{x}_a^b,S)=\max\left(\frac{c(t_i,t_{i+1},lw)}{c(lw)},\frac{c(t_{i-1},t_i,lw)}{c(lw)}\right) \tag{2}$$

- 右边界二元词性共现频率的定义如公式(3)所示.

$$P(Mxr\mid M\tilde{x}_a^b,S)=\max\left(\frac{c(t_i,t_{i+1},rw)}{c(rw)},\frac{c(t_{i-1},t_i,rw)}{c(rw)}\right) \tag{3}$$

其中,公式中的 t_i,t_{i-1},t_{i+1} 分别表示边界词 w_i 的词性、边界词 w_i 的前一个词 w_{i-1} 的词性和边界词 w_i 的后一个词 w_{i+1} 的词性,$c(*,*,*)$表示语料库中 MNP 边界词 w_i 的二元词性组合出现的次数,而 $c(rw_i)$和 $c(lw_i)$分别表示左、右边界在语料中出现的次数.数据平滑处理使用了 Katz back-off[13],计算方法如公式(4)所示.

$$P_{smooth}(t_i\mid t_{i-n+1}^{i-1})=\begin{cases}P_{smooth}(t_i\mid t_{i-n+1}^{i-1}), & \text{if } C(t_{i-n+1}^{i-1})>0\\ \gamma(t_{i-n+1}^{i-1})P_{smooth}(t_i\mid t_{i-n+2}^{i-1}), & \text{if } C(t_{i-n+1}^{i-1})=0\end{cases} \tag{4}$$

融合左、右边界信息,投射 MNP 的单语句法置信度的计算如公式(5)所示.

$$P(Mx\mid M\tilde{x},S)=P(Mxl\mid M\tilde{x}_a^b,S)P(Mxr\mid M\tilde{x}_a^b,S) \tag{5}$$

4.3 MNP双语对齐置信度

最大熵[14]模型能够融合不同类型的特征,对于双语 MNP 的对齐置信度 $P(A|Mc,M\tilde{e},CS,ES)$, 我们构造特征函数 $f_m(a_k,Mc_a^b,M\tilde{e}_c^d,CS,ES),m=1,2,...,M$, 利用最大熵模型进行建模,如公式(6)所示.对于每一个特征函数 f_m,对应的模型参数为 λ_m,m=1,2,…,M.

$$P(a_k|Mc_a^b,M\tilde{e}_c^d,CS,ES)=\frac{\exp\left(\sum_{m=1}^{M}\lambda_m f_m(a_k,Mc_a^b,M\tilde{e}_c^d,CS,ES)\right)}{\sum_{A}\exp\left(\sum_{m=1}^{M}\lambda_m f_m(a_k,Mc_a^b,M\tilde{e}_c^d,CS,ES)\right)} \tag{6}$$

公式中,CS 和 ES 分别表示中文句子和英文句子.我们采用 3 个特征对双语 MNP 对齐置信度进行建模,分别为双语 MNP 词性组合共现特征、双语 MNP 互译特征以及双语 MNP 长度关联特征.

- 双语 MNP 词性组合共现特征.

词性组合共现特征是指双语 MNP 中对应的汉英词性序列在整个语料库中的共现频率.具体计算如公式(7)所示.

$$f_m(a_k,Mc_a^b,M\tilde{e}_c^d,CS,ES)=f_m(a_k,t\_Mc_a^b,t\_M\tilde{e}_c^d,CS,ES)=\frac{c(t\_Mc_a^b,t\_M\tilde{e}_c^d)}{\sum c(t\_Mc_a^b,*)}+\frac{c(t\_Mc_a^b,t\_M\tilde{e}_c^d)}{\sum c(*,t\_M\tilde{e}_c^d)} \tag{7}$$

其中, $c(t\_Mc_a^b,t\_M\tilde{e}_c^d)$ 表示双语 MNP 词性组合在语料中共现的次数,*表示语料中任意词性的组合.

- 双语 MNP 互译特征.

Brown 等人[15]使用公式(8)计算源语言文本串 $F=f_1,f_2,\ldots,f_m$,翻译成目标语言文本串 $E=e_1,e_2,\ldots,e_n$,的翻译概率.

$$P(F|E)=\frac{1}{(n+1)^m}\prod_{j=1}^{m}\sum_{i=1}^{n}t(f_j|e_i) \tag{8}$$

对于候选双语 MNP,我们把源语言 MNP 与目标语言端投射 MNP 之间的相互翻译概率分别用 $P(Mc_a^b|M\tilde{e}_c^d)$ 和 $P(M\tilde{e}_c^d|Mc_a^b)$ 来表示,则双语 MNP 互译特征如公式(9)所示.

$$f_m(a_k,Mc_a^b,M\tilde{e}_c^d,CS,ES)=\log(P(Mc_a^b|M\tilde{e}_c^d))+\log(P(M\tilde{e}_c^d|Mc_a^b)) \tag{9}$$

- 双语 MNP 长度关联特征.

对于最优的双语 MNP (Mc_a^b,Me_c^d) 而言, Mc_a^b,Me_c^d 的长度差异近似满足标准正态分布[16],由此,我们定义长度关联特征如公式(10)所示.

$$f_m(a_k,Mc_a^b,M\tilde{e}_c^d,CS,ES)\approx f_m(a_k,|Mc_a^b|,|M\tilde{e}_c^d|)=\frac{|Mc_a^b|-\delta|M\tilde{e}_c^d|}{\sqrt{(|Mc_a^b|+1)^{\sigma^2}}} \tag{10}$$

其中, $\delta=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{c(Me_i)}{c(Mc_i)}\right),\sigma^2=\frac{1}{n}\sum_{j=1}^{n}\left(\frac{c(Me_j)}{c(Mc_j)}-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{c(Me_i)}{c(Mc_i)}\right)^2\right)$, $c(*)$表示*包含的字符数,英语为字母数,汉语为汉字数.

4.4 最优双语MNP假设搜索

我们把扩展双语 MNP 假设集合 $H_k=(Mc,M\tilde{e})$ 中每个假设 $h_i(Mc,M\tilde{e})$ 的分值表示为公式(11)的形式.

$$score(h_i)=\log(P(a_i|Mc,M\tilde{e},CS,ES))+\log(P(Me|M\tilde{e},S)) \tag{11}$$

利用公式(11)对句对中的所有扩展双语 MNP 假设进行打分,通过以下的贪心搜索过程选出句对的最优双语 MNP 假设集合,从而得到最优的目标语言 MNP 投射.

(1) 初始化该最优双语 MNP 假设集合为空;

(2) 根据公式(11)计算句对中所有双语 MNP 假设的 $score(h_i)$,并按降序排列;

(3) 依次选取一个和当前最优双语 MNP 假设集合中的双语 MNP 没有边界冲突的扩展双语 MNP 假设

h_i放入最优双语 MNP 假设集合;

(4) 重复步骤(3),直到找不到满足条件的扩展双语 MNP 假设为止.

4.5 不同特征选择比较

我们进一步考察选择不同的特征时标记投射修正模型的性能差异.Baseline 直接采用了词对齐信息获得投射 MNP.我们在 Baseline 的投射结果的基础上加入不同的特征进行修正,其中,Feature1 表示第 4.3 节所述的双语对齐特征,Feature2 表示第 4.2 节所述的单语句法特征,Feature1+Feature2 表示第 4.4 节所述的双语对齐特征和单语句法特征组合.我们采用正确率来评价 MNP 投射性能,正确率(accuracy)=投射正确的 MNP 数/总 MNP 数.从华建公司提供的汉英平行语料中随机抽取 500 句对,人工标注后作为测试集.实验结果见表 1.

Table 1 Comparison of MNP projection with different feature

表 1 在不同特征时的 MNP 投射性能比较

Feature (English-Chinese)	Accuracy (%)	Feature (Chinese-English)	Accuracy (%)
Baseline	72.91	Baseline	72.84
Baseline+Feature1	86.76	Baseline+Feature1	86.32
Baseline+Feature2	80.47	Baseline+Feature2	81.02
Baseline+Feature1+Feature2	90.83	Baseline+Feature1+Feature2	90.35

从实验结果来看,标记投射修正模型采用双语对齐特征和单语句法特征组合时,MNP 的投射正确率最高.汉英和英汉方向都比基线方法提高了近 20 个百分点.

5 实验结果及分析

5.1 实验设置

实验中使用了华建公司提供的 243 540 句对多领域的汉英平行语料、宾州树库 V5.0《新华日报》语料和东北大学 NiuTrans 开源统计机器翻译系统的部分训练语料(10 000 句对的汉英平行树库)[17].宾州树库语料不是句子级对齐的,我们将其中 325 个中英文源文件进行了句子对齐处理,得到了 3 849 个对齐的句子,其中,后 1 000 句对(2 850~3 849)作为同领域测试集,其余的作为协同训练的已标注训练集.以东北大学 NiuTrans 训练语料作为最大熵模型参数训练语料,并从中选取 1 000 句对作为跨领域测试集.我们从华建的语料库中随机选取不同领域的词数大于 15 的 10 000 句对作为实验中协同训练的无标注语料,用 GIZA++[18]获得了汉英、英汉词对齐.从 NiuTrans 树库和宾州树库分别抽取 MNP 并标注,为了保障抽取到的 MNP 语料的公开性和算法的通用性,我们采用了一种简单通用的 MNP 抽取方法:直接抽取树库的最顶层名词短语节点作为双语 MNP,如果顶层节点不互译,则依次向下抽取,直到抽取到互译的 MNP 对为止,并对抽取到的汉英 MNP 进行了人工校正.我们在 IOB2[19]标注体系的基础上添加了两个标记符号:H 和 S,用来区分 MNP 中心词和非中心词.这样,共有 5 类标记用于 MNP 及其中心词的识别:BH,BS,IH,IS 和 O.

5.2 评价指标

本文使用 F 值作为 MNP 自动识别的评价指标.单语端的 F 值与传统 MNP 识别中 F 值的定义相同;双语 MNP 的 F 值定义为:假设模型标注出的 MNP 数目为 $C1$,其中,正确标注的 MNP(双语标注都正确)数目为 $C2$,测试集中 MNP 的数目为 $C3$,那么准确率定义为 $P=C2/C1$,召回率为 $R=C2/C3$,F 值为 $F=2PR/(P+R)$.

5.3 Baseline实验

Baseline 使用了文献[4]中的实验设置,这是截止到目前为止,在单个分类器下 MNP 自动识别性能最好的.采用了基于 TinySVM 分类器的开源序列标注工具 Yamcha(http://www.chasen.org/~tAKu/software/yamcha/),特征采用了词和词性,静态特征窗口设置为 9,动态特征考虑当前位置之前的 4 个历史标记,即,使用五元历史标注,汉英两端使用了相同的实验设置.实验结果见表 2,其中,MNPc 表示双语 MNP 汉语端的识别结果,MNPe 表示英语端的识别结果,BMNP 表示双语一致识别的结果.similar_domain 表示在同领域测试集上的测试,cross_domain

表示在跨领域测试集的测试.从 Baseline 实验结果可以看出,双语 MNP 一致识别的 F 值要远远低于 MNPc 和 MNPe.这是因为 MNPc 和 MNPe 的识别错误在对齐过程中混合扩大,很大程度上影响了双语 MNP 的识别效果.另外,双语 MNP 在跨领域测试集上的识别 F 值与在同领域测试集上相比降低了接近 10 个百分点,这可以归结于有监督的学习方法对训练语料的依赖性.

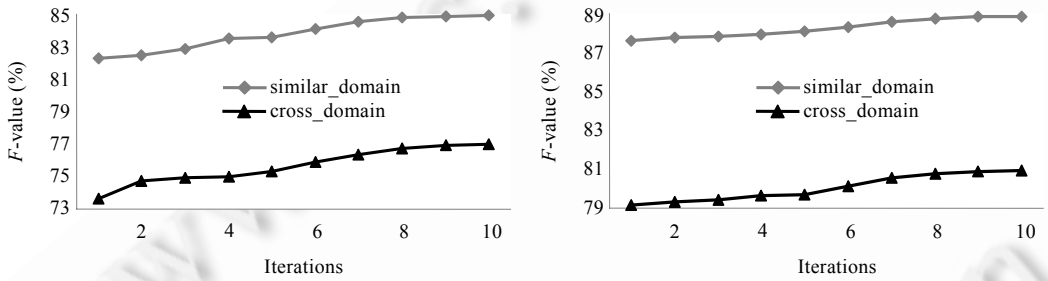
Table 2 Baseline identifying results (%)

表 2 Baseline 识别结果(%)

Domain	MNPc	MNPe	BMNP
similar_domain	82.31	87.56	76.67
cross_domain	73.60	79.03	68.82

5.4 双语co-training算法实验

缓冲区大小 n 和标注数据增量 z 这两个参数的选择直接影响到 co-training 算法的性能.文献[20]研究表明:算法每轮迭代时,取前 10%的无标记数据辅助分类器的训练能取得较好的效果.因此,我们将缓存区的大小 n 设置为 1 000(可以容纳 1 000 个句子),标记集增量 z 设为 100,即,每次循环时选取使对齐标注一致率最高的 100 个句子作为另一端语言的标注数据增量集.图 3 显示了双语 co-training 算法单语 MNP 识别的实验结果,图 4 显示了双语 co-training 算法双语 MNP 一致识别的实验结果.其中,横坐标表示协同训练的轮次,纵坐标表示识别结果的 F 值.



(a) MNPc (b) MNPe

Fig.3 Learning curves of bilingual co-training

图 3 双语 co-training 算法的 MNP 识别性能曲线

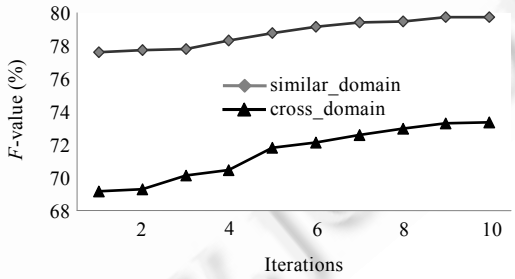


Fig.4 BMNP learning curves for bilingual co-training

图 4 双语 co-training 算法的 BMNP 识别性能曲线

由图 3 可以很明显地观察到:MNPc 和 MNPe 的识别性能都有较大的提高;随着协同训练轮次的增加,F 值上升的趋势比较明显;当协同训练轮次达到 10 次左右时,算法趋于收敛.

由图 4 可以很明显地观察到:双语 MNP 识别的性能也有较大的提高;随着协同训练轮次的增加,F 值上升的趋势比较明显.

10 次迭代后,实验结果见表 3.从英语单语 MNP 识别来看,与基线模型相比,同领域测试的 F 值提升了 1.3 个百分点;跨领域的更高一点,提高了 1.9 个百分点.汉语单语 MNP 识别情况好于英语,同领域测试的 F 值提升了 2.6 个百分点,跨领域达到了 3.39 个百分点.这表明,双语 co-training 算法提高了双语 MNP 的单语言端 MNP 的识别性能.这也证明,把平行的汉英句集看作同一个数据集的两个不同视图进行协同训练的算法是有效的.

Table 3 Bilingual co-training identifying results (%)

表 3 双语 co-training 算法的 MNP 识别结果(%)

Domain	MNPc	MNPe	BMNP
similar_domain	84.91 (+2.6)	88.86 (+1.3)	79.75 (+3.08)
cross_domain	76.99 (+3.39)	80.93 (+1.9)	73.34 (+4.52)

在双语 MNP 的单语言端,从 MNPc 和 MNPe 实验对比可以看到:算法在汉语端的识别性能提升幅度要远高于英语端,其中,同领域的测试中高出 1.3 个百分点,跨领域的测试中高出 1.49 个百分点.分析其中的原因不难发现:由于汉语端分类器的性能要比英语端低很多(大约 5%~6%),因此在协同训练的过程中,汉语端提供给英语端的带标数据噪音比例要高于英语端提供给汉语端的,更多噪声数据的引入,造成英语端识别性能提高幅度要低于汉语端.

在不同领域测试集的测试比较中,跨领域测试的提升幅度要高于同领域测试(英语端增幅为 0.6 个百分点,汉语端增幅为 0.79 个百分点).这是因为大量跨领域无标注数据的引入以及隐式的置信度估计,提升了协同训练算法的泛化能力.

在跨领域测试中,BMNP 的 F 值提高了 4.52 个百分点,虽然算法在同领域测试集中的表现略低于跨领域测试,但也有 3.08 个百分点的提升.这可以归结为:随着单语 MNP 识别性能的提高,MNPc 和 MNPe 识别错误减少,从而在对齐过程中,识别错误混合放大的情况在一定程度上得到了缓解.我们从实验数据中还可以观察到,BMNP 提升的幅度要高于 MNPc 和 MNPe.这表明:把平行的汉英句子集看作同一个数据集的两个不同视图进行协同训练的方法提升了 MNPc 和 MNPe 识别的一致性,这也进一步论证了 MNPc 和 MNPe 识别的互补性,这对于进一步提取等价双语 MNP 对有着重要的意义.

5.5 增量标注选择策略比较

从第 5.4 节的实验结果来看,第 2 节所述的 co-training 算法(Method0)取得了较好的效果.但是 Method0 需要多次循环执行,计算量较大.因此,我们构建了另外一种增量标注选择方法(Method1),直接选取 k 个标注一致率最高的平行句对来扩充带标数据集合.另外,为了验证隐式估计方法的有效性,我们构建了显式的增量标注选择方法(Method2),在少量标记数据上(我们手工标注了 500 句对的新闻语料)进行 10 倍交叉验证来进行增量标记数据的选择.Method1 和 Method2 在分别进行 10 次迭代协同训练后的实验结果见表 4.

Table 4 The results of Method1 and Method2 (%)

表 4 Method1 和 Method2 的 MNP 识别结果(%)

Method	Domain	MNPc	MNPe	BMNP
Method1	similar_domain	83.53 (+1.22)	88.38 (+0.82)	78.25 (+1.58)
	cross_domain	75.08 (+1.48)	80.02 (+0.99)	71.13 (+2.31)
Method2	similar_domain	84.95 (+2.64)	88.97 (+1.41)	79.95 (+3.28)
	cross_domain	73.68 (+0.18)	79.14 (+0.11)	69.11 (+0.29)

虽然 Method1 的效率高于 Method0,平均训练次数降到大约原来的 1/10,但是从实验结果来看,Method1 的性能远远低于 Method0,整体下降接近 50%.这是因为把标注一致率最高的平行句对作为增量标注训练分类器,与在数据集上标注效果最好的分类器不是等价的.其中,跨领域性能下降幅度更大,归结于标注一致性高的句对往往与已标注语料的领域相似性比较高,直接用来扩充带标数据集合,会导致算法的泛化能力的下降.Method2 虽然在同领域测试中略优于 Method0,比如,BMNP 识别 F 值提高了 0.2%,但在跨领域测试中性能则大幅度下降,其中,BMNP 识别 F 值降低了 4.23%.这也验证了 Method0 具有更好的领域适应性.

6 结论和展望

双语 MNP 的自动识别在机器翻译、辅助机器翻译和跨语言信息检索等领域具有至关重要的意义.本文提出了一种基于半监督学习的双语 MNP 识别算法,使汉英两种语言的 MNP 识别优势互补,既可以提高单语 MNP 的识别性能,也可以提升双语 MNP 的识别一致性,并对双语协同训练算法中增量标记的选择进行了详细的介绍.实验结果表明:该算法不仅显著提高了双语 MNP 的识别能力,大量跨领域未标注数据的加入,还增强了双语 MNP 识别的领域适应性.

我们今后的研究重点将放在尝试更好的融合算法,更多加入主动学习的因素,提高增量标记的置信度,进一步增强算法对新特征的学习能力,从而达到更好的识别效果.另外,继续考察其他的语言对,提高算法的应用范围,也是我们进一步研究的内容.

References:

[1] Zhou Q, Sun MS, Huang CN. Automatic identification of Chinese maximal noun phrases. Ruan Jian Xue Bao/Journal of Software, 2000,11(2):195−201 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/20000206.htm

[2] Voutilainen A. NPtool, a detector of English noun phrases. In: Church K, ed. Proc. of the Workshop on Very Large Corpora. Columbus: Association for Computational Linguistics, 1993. 48−57.

[3] Chen KH, Chen HH. Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation. In: Proc. of the 32nd Annual Meeting on Association for Computational Linguistics. New York: Association for Computational Linguistics, 1994. 234−241. [doi: 10.3115/981732.981764]

[4] Jian P, Zong CQ. A new approach to identifying Chinese maximal-length phrases by combining bidirectional labeling. CAAI Trans. on Intelligent Systems, 2009,4(5):406−413 (in Chinese with English abstract).

[5] Sogaard A. Semi-Supervised condensed nearest neighbor for part-of-speech tagging. In: Matsumoto YJ, Mihalcea R, eds. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Portland: Association for Computational Linguistics, 2011. 48−52.

[6] Lang J, Lapata M. Unsupervised semantic role induction via split-merge clustering. In: Matsumoto YJ, Mihalcea R, eds. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Portland: Association for Computational Linguistics, 2011. 1117−1126.

[7] Wan XJ. Bilingual co-training for sentiment classification of Chinese product reviews. Computational Linguistics, 2011,37(3): 587−616. [doi: 10.1162/COLI_a_00061]

[8] Koehn P, Knight K. Feature-Rich statistical translation of noun phrases. In: Proc. of the 41st Annual Meeting on Association for Computational Linguistics, Vol. 1. Sapporo: Association for Computational Linguistics, 2003. 311−318. [doi: 10.3115/1075096.1075136]

[9] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proc. of the 11th Annual Conf. on Computational Learning Theory. Wisconsin: ACM Press, 1998. 92−100. [doi: 10.1145/279943.279962]

[10] Abney S. Bootstrapping. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002. 360−367.

[11] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data. In: Proc. of the 17th Int'l Conf. on Machine Learning. San Francisco: Int'l Machine Learning Society, 2000. 327−334. http://dblp.uni-trier.de/rec/bib/conf/icml/2000

[12] Zhou, ZH, Li M. Tri-Training: Exploiting unlabeled data using three classifiers. IEEE Trans. on Knowledge and Data Engineering, 2005,17(11):1529−1541. [doi: 10.1109/TKDE.2005.186]

[13] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. on Acoustics, Speech and Signal Processing, 1987,35(3):400−401. [doi: 10.1109/TASSP.1987.1165125]

[14] Berger AL, Pietra VJD, Pietra SAD. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):39−71.

[15] Brown PF, Pietra VJD, Pietra SAD, Mercer RL. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 1993,19(2):263−311.

[16] Church KW. Char_Align: A program for aligning parallel texts at the character level. In: Proc. of the 31st Annual Meeting on Association for Computational Linguistics. Ohio: Association for Computational Linguistics, 1993. 1−8. [doi: 10.3115/981574.981575]

[17] Xiao T, Zhu JB, Zhang H, Li Q. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation. In: Li HZ, Lin CY, Osborne M, Lee GG, Park JC, eds. Proc. of the ACL 2012 System Demonstrations. Jeju: Association for Computational Linguistics, 2012. 19−24.

[18] Och FJ, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003,29(1):19−51. [doi: 10.1162/089120103321337421]

[19] Sang EFTK, Veenstra J. Representing text chunks. In: Proc. of the 9th Conf. on European Chapter of the Association for Computational Linguistics. Bergen: Association for Computational Linguistics, 1999. 173−179.

[20] Pierce D, Cardie C. Limitations of co-training for natural language learning from large datasets. In: Lee L, Harman D, eds. Proc. of the 2001 Conf. on Empirical Methods in Natural Language Processing. Pittsburgh: Association for Computational Linguistics, 2001. 1−9.

附中文参考文献:

[1] 周强,孙茂松,黄昌宁.汉语最长名词短语的自动识别.软件学报,2000,11(2):195−201. http://www.jos.org.cn/1000-9825/20000206.htm

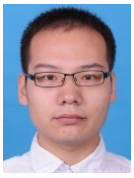
[4] 鉴萍,宗成庆.基于双向标注融合的汉语最长短语识别方法.智能系统学报,2009,4(5):406−413.

李业刚(1975－),男,山东淄博人,博士生,副教授,主要研究领域为自然语言处理,机器翻译.

黄河燕(1963－),女,博士,研究员,博士生导师,主要研究领域为自然语言处理,机器翻译.

史树敏(1978－),女,博士,讲师,CCF 会员,主要研究领域为自然语言处理,本体方法论及应用.

鉴萍(1982－),女,博士,讲师,CCF 会员,主要研究领域为自然语言处理,机器学习.



苏超(1988－),男,博士生,主要研究领域为自然语言处理,机器学习.