ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2843−2852
DOI: 10.3724/SP.J.1001.2008.02843

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

基于图的半监督关系抽取*

陈锦秀[1+], 姬东鸿[2]

[1](厦门大学 智能科学与技术系,福建 厦门 361005)

[2](武汉大学计算机系,湖北 武汉 430072)

Graph-Based Semi-Supervised Relation Extraction

CHEN Jin-Xiu[1+], JI Dong-Hong[2]

[1](Department of Cognitive Science, Xiamen University, Xiamen 361005, China)

[2](Department of Computer Science, Wuhan University, Wuhan 430072, China)

+ Corresponding author: E-mail: cjx@xmu.edu.cn

Chen JX, Ji DH. Graph-Based semi-supervised relation extraction. *Journal of Software*, 2008,19(11): 2843−2852. http://www.jos.org.cn/1000-9825/19/2843.htm

Abstract: This paper investigates a graph-based semi-supervised learning algorithm, that is, label propagation algorithm for relation extraction. Labeled and unlabeled examples are represented as the nodes, and their distances as the weights of edges in the graph. The relation extraction tries to obtain a labeling function on this graph to satisfy the global consistency assumption. Experimental results on the ACE (automatic content extraction) corpus showed that this method achieves a better performance than SVM (support vector machine) when only very few labeled examples are available, and it also performs better than bootstrapping for the relation extraction task.

Key words: relation extraction; graph-based; semi-supervised learning; label propagation

摘 要: 提出利用基于图的半监督学习算法,即标注传递算法,指导计算机从非结构化的文本中自动识别出实体之间的关系.该方法首先利用图策略来建立关系抽取的模型.在这个图模型中,各个有标签和未标签的样本被表示成图上的各个节点,而样本间的距离则作为图上各边的权重.然后,关系抽取的任务就转化成在这个图上估计出一个满足全局一致性假设的标注函数.通过对 ACE(automatic content extraction)语料库的评测,结果显示,当只有少量的标签样本时,采用该标注传递的方法可以获得比基于 SVM(support vector machine)的有监督关系抽取更好的性能,同时也明显优于基于 Bootstrapping 的半监督关系抽取的方法.

关键词: 关系抽取;基于图;半监督学习;标签传递

中图法分类号: TP18 **文献标识码**: A

信息时代带来了海量的数字化文本,其中 90%的文本是非结构化和半结构化的形式.日益累积的数据使得信息的获取越来越困难.面对如此巨大的数字资源,人们的时间与精力却是有限的、不变的,然而对企业或个人有巨大价值的信息就隐藏在这些海量数据中,如何把它们提取出来呢?信息抽取(information extraction)技术在计算机自然语言处理领域正越发体现出其重要性.

* Supported by the National Natural Science Foundation of China under Grant Nos.60803078, 60773011 (国家自然科学基金)
Received 2008-02-29; Accepted 2008-08-26

实体间的关系抽取子任务是信息抽取的重要支撑技术,在许多应用领域中都扮演着重要的角色,如自动问答、生物信息技术、ontology 的创建等等.对于命名实体的识别研究,学者们已经做了大量的工作并取得了令人满意的成绩,本文主要集中于实体间关系抽取的研究,即如何指导计算机从非结构化的文本中自动识别出实体之间的关系.例如,在句子“John Smith is the chief scientist of the Hardcom Corporation.”中,传达了两个实体“John Smith”(PERSON)和“Hardcom Corporation”(ORGANIZATIONS)之间的语义关系“person-affiliation”.

关系抽取的任务是于 20 世纪 80 年代末由美国国防部的 DARPA(Defense Advanced Research Projects Agency)在 Message Understanding Conference(MUC 6)上首次被提出来的.但是,大多数 MUC 上的关系抽取研究都是基于规则的方法,可移植性差.自动内容抽取(automatic content extraction,简称 ACE. http://www.ldc.upenn.edu/Projects/ACE/)评测项目是继 MUC 会议之后又一个有关信息抽取领域的著名评测项目.该评测任务也引入了对实体关系的评测 RDC(relation detection and characterization),希望能够将实体之间的关系揭示出来.

近年来,由于语料库和自然语言工具的可利用,已有一些机器学习技术应用于关系抽取任务,如基于有监督学习的方法[1−8]、基于半监督学习的方法[9−11]和基于无监督学习的方法[12−14].其中,基于有监督学习的关系抽取方法在 ACE 的评测数据上已经取得了一定的进展,而很多抽取模型也在其中得到了应用,如 feature based,tree kernel based 等.然而,基于有监督的机器学习的关系抽取方法需要以大量有标签的训练数据为前提,这带来了大量的人力和时间上的花费.为了突破基于有监督学习方法的局限,研究者也提出了一些基于半监督学习的关系抽取方法来弥补语料库标注的缺陷.在关系抽取任务所做的努力中,主要有 3 个具有代表性的基于半监督学习的关系抽取系统,即 DIPRE(dual iterative pattern relation expansion)[10],Snowball[9]和 Zhang’s method[11].

DIPRE 系统是一个基于 Bootstrapping 的系统.该系统采用模式匹配方法作为分类器来挖掘模式集合和关系候选之间的重复性.该系统主要用于从互联网上抽取〈*author*,*book*〉关系.Snowball 是另外一个采用 Bootstrapping 技术从非结构化文本中抽取〈*organization*,*location*〉关系的系统.该系统和 DIPRE 系统有很多共同之处,包括建立基于 Bootstrapping 的框架和模式匹配分类技术的使用.Zhang 的方法主要是针对 ACE 的关系抽取任务提出的基于 SVM 的 Bootstrapping 模型.然而,在该系统的实现中并没有真正检测两个实体是否有关系存在,而仅仅是在假定两个实体已知存在某种关系的情况下,对各实体对进行关系类型的分类.此外,在采用特征向量表示候选关系的知识时,也没有考虑到各特征在句子中所处的位置对于实体之间这种具有上下文相关特性的对象具有一定的影响.

国外学者对信息抽取的研究起步较早,而在国内的自然语言处理领域的研究中,实体的关系抽取课题起步较晚,大部分研究主要集中在对网页信息的抽取和对命名实体的识别上,对自由文本实体关系的分析研究也是近两年才开始尝试的[15−19].这些方法大部分也都是采用了有监督学习或基于 Bootstrapping 的半监督学习的抽取模型来抽取实体之间的关系.

上述这些基于半监督学习的关系抽取方法主要都是采用 Yarowsky 在 1995 年提出的 Bootstrapping 算法作为抽取模型.Bootstrapping 算法通过循环的分类方式来分类未标签的样本,在每一次循环中,都把上一次循环分类结果中可信度较高的样本加入有标签的样本集合中,然后根据扩充了的种子集合重新学习分类模型.因此,Bootstrapping 算法是基于局部一致性假设,即未标签的样本只使用由有标签的样本数据进行训练得到的模型来分类.这样的方法忽略了对未标签的样本之间的关联性和相似性的考虑,也没有从全局一致性的观点来执行分类算法,这将导致当训练数据有限时,在推导类别边界时不能充分利用无标签数据的信息去挖掘隐藏在数据中的类结构特征.

为此,本文提出利用基于图的半监督学习算法,即标注传递算法,来自动识别出实体之间的关系,实现当训练数据不足时利用很容易获取的大量未标记样本来改善学习性能,达到关系类型标注的全局一致性.接下来,我们将首先介绍如何选择合适的特征来描述关系候选的知识;其次,建立一个基于图的关系抽取模型;然后,利用标注传递的半监督学习算法在该图模型上进行自动的关系类型标注,并给出在 ACE 语料库上的实验结果和分析;最后与相关工作进行了比较和分析总结.

1 关系候选的知识表示和度量

基于机器学习关系抽取任务的主要思路是根据给定的上下文信息,采用机器学习技术对每个关系候选进行分类,并赋予其一个相应的关系类型标注.也就是说,关系抽取任务可以被映射成一个标准的分类问题:

$$R\rightarrow(C_{pre},e_1,C_{mid},e_2,C_{post}).$$

其中,e_1 和 e_2 代表两个实体表述,而 C_{pre},C_{mid} 和 C_{post} 则分别是两个实体对之前、之间和之后的上下文信息.

因此,在使用机器学习的算法进行实体关系的抽取时,我们首先必须选取适当的特征描述来表达句子中出现的实体对的知识,从而构造出关系候选.目前我们只考虑一个句子中的两个实体之间的关系,而不考虑跨越句子的实体之间的关系.首先在句子中枚举出所有可能的实体对组合,然后从两个实体对 e_1 和 e_2 及其 3 个上下文 C_{pre},C_{mid} 和 C_{post} 中选出关于词法、句法、语法及语义等方面的特征集合.以 ACE 的英文新闻语料库为评测目标,我们从实体对及其语法分析树中选择以下有代表性的特征:

- 词(word):两个实体 e_1 和 e_2 及 3 个上下文窗口 C_{pre},C_{mid} 和 C_{post} 中的所有词.
- 两个实体所属的类别(type):PERSON,ORGANIZATION,FACILITY,LOCATION 和 GPE.
- 词性特征(part-of-speech):两实体 e_1 和 e_2 及 3 个上下文窗口 C_{pre},C_{mid} 和 C_{post} 中所有词的词性.
- Chunking 特征:
  - Chunking 标注信息(0,I-XP,B-XP):实体 e_1 和 e_2 及 3 个上下文窗口 C_{pre},C_{mid} 和 C_{post} 中的所有词.0 表示该词不在任何短语块中;I-XP 表示该词在某一个块 XP 中;B-XP 表示该词在块 XP 的开头.这里,块 XP 可以是任何短语块,如 NP 块.
  - 语法功能:两个实体 e_1 和 e_2 及 3 个上下文窗口 C_{pre},C_{mid} 和 C_{post} 中的所有词.每一个短语块的最后一个词是该短语的主词,该主词的语法功能就是整个短语块的语法功能.如 NP-SBJ 表示一个 NP 短语块作为一个句子的主语.短语块中非主词的其他词则以 NOFUNC 作为其语法功能表述.
  - IOB 链:两个实体中的主词的 IOB 链.所谓 IOB 链指的是语法树中从根节点到叶子节点的所有成员的句法范畴.

在描述上述每一个特征时,我们也同时给出它们在句子中的相对位置信息.比如,对于含有位置信息的词(word),其特征包括:

1) WE_1(WE_2): All words in e_1(e_2).

2) WHE_1(WHE_2): Head word of e_1(e_2).

3) WMNULL: No words in C_{mid}.

4) WMFL: The only word in C_{mid}.

5) WMF,WML: The first word, the last word in C_{mid} when at least two words in C_{mid}.

6) WM_2,WM_3, ...: The second word, the third word, ... in C_{mid} when at least three words in C_{mid}.

7) WEL_1,WEL_2, ...: The first word, the second word, ... before e_1.

8) WER_1,WER_2, ...: The first word, the second word, ... after e_2.

上述所有的词法和句法特征结合它们在上下文中的位置信息生成某一个关系候选实例的特征集合.在此过程中,我们将那些只出现过 1 次的特征值删除.

2 基于图的关系抽取模型的建立

通常,基于图的学习方法都是建立在这样的假设基础上,即具有相同特征的两个节点倾向于属于同一个类别.而关系抽取任务的假设前提则是:如果两个关系实例相似度很高,即特征集合相似且语法结构相似,则它们将倾向于属于同一种关系类型.可以看出,关系抽取任务的假设前提与基于图的学习方法的假设是吻合的.因此,我们可以利用图来建立关系抽取模型,然后利用少部分有标签的数据辅助大量未标签的数据进行非监督的学习.

假设 $X=\{x_i\}_{i=1}^{n}$ 是所有实体对候选关系实例的集合,其中 n 是所有实体对候选关系实例的数目.假设 $C=\{r_j\}_{j=1}^{R}$ 是所有关系类别标号的集合,其中 r_j 代表某一关系类别,而 R 则是所有关系类型的数目.于是,我们可以建立有标签的数据样本和无标签的数据样本,具体如下:

- 有标签数据$(x_1,y_1)\ldots(x_l,y_l)$:X 中的前 l 个样本 $x_i(i\le l)$被标注上标签 $y_i(y_i\in C)$,即 $Y_L=\{y_i\}_{i=1}^{l}\in C$.
- 未标签数据 $(x_{l+1},y_{l+1})\ldots(x_{l+u},y_{l+u})$:X 中剩下的 u 个样本$(l+1\le u\le n)$是无标签样本,即 $Y_U=\{y_i\}_{i=l+1}^{l+u}$.

我们的目标就是根据 X 和 Y_L 预测出未标签数据的关系类别标注 Y_U.

为了在学习过程中有效地结合上述的未标签样本和有标签样本数据的信息,模型中定义了一个图 $G=(V,E)$.在这个图中,节点集合 V 代表了数据集中各个未标签样本和有标签样本,而任意两个节点 x_i 和 x_j 相连的边 E 则用以下公式来计算其权重:

$$W_{ij}=\exp\left(-\frac{s_{ij}^2}{\alpha^2}\right).$$

公式中的 s_{ij} 代表样本节点 x_i 和 x_j 之间的相似性.α是一个平衡因子,可以设置为有标签样本在不同类别中的平均相似度.

于是,基于图的半监督学习方法就可以看作是在这个图上估计一个标注函数 f.该标注函数必须同时满足两个限制条件:

1) 标注过程必须围绕有标签节点的标注信息.

2) 标注必须平滑地在整个图上进行.

这两个限制条件可以分别通过定义一个损失函数和一个规范式来描述.

利用图结构来平滑标注与传统的基于 Bootstrapping 的半监督学习算法有着显著的不同,它充分结合了未标签样本和有标签样本数据的信息.为了实现全局一致性的假设,关系抽取问题最终被形式化为关系类别的标签信息根据与相邻节点的相似度在图上进行全局传递的过程.

3 基于图模型的关系的自动发现

建立了基于图策略的关系抽取模型之后,图上的节点包含了少量有标签的样本数据,也包含了大量未标签的样本数据,如何选择一种算法来将有标签的样本数据中的标签信息在整个图上进行传递则是至关重要的.有很多研究人员已经提出了这类基于图的半监督机器学习算法来挖掘数据中潜在的簇结构信息[20−25],从而实现标签信息传递全局一致性的假设.

在基于图的半监督学习关系抽取任务中,我们将研究如何利用标签传递(label propagation)的算法来实现实体之间关系的发现.标签传递的算法把标签信息从任意一个节点通过加权的各边循环地传递到附近的其他节点,最终达到全局稳定的状态,从而推导出未标签节点的标注信息的目标.节点之间边的权重越大,标签信息越容易在节点间传递.因而,样本节点越相似,它们拥有同样的标签的可能性就越大(全局一致性的假设).

为此,我们首先定义两个矩阵:

- $n\times n$ 的概率转换矩阵 T:

$$T_{ij}=P(j\to i)=\frac{w_{ij}}{\sum_{k=1}^{n}w_{kj}}.$$

公式中,T_{ij} 表示从节点 x_j 跳到节点 x_i 的概率.

- $(l+u)\times R$ 的标签矩阵 Y,其中,y_{ij} 表示节点 y_i 拥有标签 x_j 的概率.

标签传递 LP 算法的具体步骤如下:

步骤 1. 初始化.

- 设置循环索引 $t=0$.
- 假设 Y^0 为图中各节点的初始标签,其中,如果 y_i 具有标签 r_j,则 $Y_{ij}^0=1$;否则,$Y_{ij}^0=0$.

- 假设 Y_L^0 和 Y_U^0 分别是 Y^0 的前 l 行和剩余的 u 行.其中, Y_L^0 与有标签样本数据中的标签信息一致,而 Y_U^0 的初始化则可以取任意值.

步骤 2. 根据公式 $Y^{t+1}=\bar{T}Y^t$, 把标签信息从任意节点传递到相邻节点,其中 $\bar{T}$ 是转换矩阵 T 的行规范化形式,即 $\bar{T}_{ij}=T_{ij}/\sum_k T_{ik}$, 使其可以继续作为分类概率的解释.

步骤 3. 重填回有标签样本的原标签信息,即将 Y_L^{t+1} 的值替换为 Y_L^0 的值.

步骤 4. 重复步骤 2 以后的操作,直到 Y 收敛为止.

步骤 5. 标注未标签节点 $x_h(l+1\le h\le n)$为标号 $y_h,y_h=\text{argmax}_j Y_{hj}$.

上述算法的步骤 3 确保了有标签的样本不会随着标号的全局传递而改变,因为在每一轮标签传递的循环中,初始标签信息 Y_L 都会在步骤 3 被强制填回,参与下一轮的标签传递.事实上,我们也是只关心未标签样本的标注 Y_U.接下来,我们将证明该算法在步骤 4 一定会收敛到唯一的一个解.

收敛性的证明:将转换矩阵 $\bar{T}$ 分裂成分别对应于有标签样本和无标签样本的子矩阵,即

$$\bar{T}=\begin{pmatrix}\bar{T}_{LL} & \bar{T}_{LU}\\ \bar{T}_{UL} & \bar{T}_{UU}\end{pmatrix}.$$

令 $Y=\begin{pmatrix}Y_L\\ Y_U\end{pmatrix}$, 则由 $Y^{t+1}=\bar{T}Y^t$ 可得:

$$\begin{pmatrix}Y_L\\ Y_U\end{pmatrix}^{t+1}=\begin{pmatrix}\bar{T}_{LL} & \bar{T}_{LU}\\ \bar{T}_{UL} & \bar{T}_{UU}\end{pmatrix}\begin{pmatrix}Y_L\\ Y_U\end{pmatrix}^{t}.$$

于是,上述的标签传递算法等价于

$$Y_U^{t+1}=\bar{T}_{UL}Y_L^t+\bar{T}_{UU}Y_U^t.$$

由该等式及 $Y_L^t=Y_L^{t-1}=Y_L^0$ 又可以循环递推出下面的等式成立:

$$Y_U=\left(\sum_{i=1}^{n}(\bar{T}_{UU})^{(i-1)}\right)\bar{T}_{UL}Y_L^0+(\bar{T}_{UU})^{(n)}Y_U^0.$$

其中, Y_U^0 是 Y_U 的的初始值.需要证明的是 $(\bar{T}_{UU})^{(n)}Y_U^0\to 0$. 由转换矩阵 $\bar{T}$ 的定义可知, $\bar{T}$ 中的所有成员都大于 0, 此外,由于 $\bar{T}$ 是行规范式,并且 $\bar{T}_{UU}$ 是 $\bar{T}$ 的子矩阵,因而可得:

$$\exists\gamma<1,\ \sum_{j=1}^{u}(\bar{T}_{UU})_{ij}\le\gamma,\ \forall i=1,\ldots,u.$$

利用这个条件,通过矩阵相乘递推公式可得:

$$\sum_j(\bar{T}_{UU})_{ij}^{(n)}=\sum_j\sum_k(\bar{T}_{UU})_{ik}^{(n-1)}(\bar{T}_{UU})_{kj}=\sum_k(\bar{T}_{UU})_{ik}^{(n-1)}\sum_j(\bar{T}_{UU})_{kj}\le\sum_k(\bar{T}_{UU})_{ik}^{(n-1)}\gamma\le\gamma^n.$$

因此, $(\bar{T}_{UU})^n$ 按行累加和收敛到 0,意味着 $(\bar{T}_{UU})^nY_U^0\to 0$.这也证明了 Y_U 的初始值 Y_U^0 不重要,不会影响到对 Y_U 的近似值 $\hat{Y}_U$ 的估计.并且,利用无穷等比数列各项和的公式显而易见地可以得到:

$$\hat{Y}_U=\lim_{n\to\infty}Y_U^n=\lim_{n\to\infty}\left(\sum_{i=1}^{n}(\bar{T}_{UU})^{(i-1)}\right)\bar{T}_{UL}Y_L^0=\frac{1}{I-\bar{T}_{UU}}\bar{T}_{UL}Y_L^0$$

是一个唯一确定的固定值(其中,I 是一个 $u\times u$ 的单位矩阵),这也是标签传递迭代算法的唯一解.因此,算法的收敛性得证. □

4 实验和结果分析

4.1 关系候选相似度的计算

为了衡量两个关系候选的相似度,我们采用两种方法来计算,即 Cosine 相似性度量方法和 Jensen-Shannon(JS) divergence.Cosine 相似性方法通过计算两个特征向量的角度来测量语义距离.JS divergence 作为距离度量方法也应用于文本聚类中,得到了比 Cosine 相似性方法更好的性能.如果把特征向量看作是对特征的概

率分布,那么,JS divergence 就可以看成是对两个概率分布的距离进行计算.JS divergence 根据 KL-divergence 来定义,具体如下:

$$JS(q,r)=\frac{1}{2}[D_{KL}(q\|\overline{p})+D_{KL}(r\|\overline{p})].$$

其中, $\overline{p}=\frac{1}{2}(q+r)$, 且

$$D_{KL}(q\|\overline{p})=\sum_y q(y)\left(\log\frac{q(y)}{\overline{p}(y)}\right),\ D_{KL}(r\|\overline{p})=\sum_y r(y)\left(\log\frac{r(y)}{\overline{p}(y)}\right).$$

4.2 实验设置

本文采用 ACE2003 语料库来评测我们的基于标签传递(LP)的关系抽取算法,实现 ACE 评测中对于关系子类型的识别和分类任务(relation subtype detection and characterization).该语料库包含了来自 broadcast,newswire 和 newspaper 三个来源的 519 个文件.我们只处理句子中的显式关系,并假设所有的实体都已经事先在 ACE 的实体识别的子任务 EDT 中识别出来.表 1 列出了训练和测试语料库中各关系候选分布于不同的关系子类型 Subtype 中的频度.我们通过从训练数据集中随机取样,建立存在关系的有标签样本数据集 L_1.并按照同样的取样百分比从那些无关的实体对集合中随机取样,建立不存在关系的样本集合 L_2,并将其归为一类,记为 NONE 类,则 $L=L_1+L_2$ 就共同构成了算法中的有标签样本数据集,并利用建立起的 NONE 类来帮助我们检验给定的实体对是否存在关系,即关系检测任务(relation detection).如果实体对在标注过程中被标为 NONE 类,则该实体对不存在关系;反之,若实体对被标为其他 24 类关系子类型(subtype),则表示存在关系.此外,我们把训练数据集中取样剩下的样本连同 ACE 的测试数据集都归为算法中的无标签样本数据集,通过 LP 算法的运行对它进行标注关系类型.而测试数据集的标注结果用于最后的评估计算.评估的性能标准是 Precision,Recall 和 *F*-measure.

Table 1 Frequency of relation subtypes in the ACE training and devtest corpus

表 1 ACE 训练和测试语料库中各关系候选分布于各关系子类型的频度

Type	Subtype	Training	Devertest
ROLE	General-Staff	550	149
	Management	677	122
	Citizen-Of	127	24
	Founder	11	5
	Owner	146	15
	Affiliate-Partner	111	15
	Member	460	145
	Client	67	13
	Other	15	7
PART	Part-Of	490	103
	Subsidiary	85	19
	Other	2	1
AT	Located	975	192
	Based-In	187	64
	Residence	154	54
SOC	Other-Professional	195	25
	Other-Personal	60	10
	Parent	68	24
	Spouse	21	4
	Associate	49	7
	Other-Relative	23	10
	Sibling	7	4
	GrandParent	6	1
NEAR	Relative-Location	88	32

4.3 实验结果:LP vs. SVM

SVM 是有监督学习的关系抽取任务中常用的分类方法.在这个实验中,我们采用的是 LIBSVM(LIBSVM:a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm)的工具和线性的核函数公式.

为了比较 SVM 和 LP 算法,我们用不同的取样大小来建立有标签的样本集,包括:1%×N_{train},10%×N_{train},25%×N_{train},50%×N_{train},75%×N_{train},100%×N_{train}(N_{train} 是 ACE 中训练集的样本数目).如果任何一个关系子类型 Subtype 不在我们取样后的有标签样本集中出现,则必须重新取样.对每一个取样值都执行 20 次,最后再取 20 次结果的平均值作为结果.表 2 是关系检测任务的结果比较,我们将每一次根据不同取样大小得到的有标签数据集同时也作为 SVM 算法的训练集.其中,SVM 和 LP 算法分别在不同大小的有标签样本集上测试.LP 算法分别用 Cosine 和 JS 两种相似性方法进行计算.表中 P,R,F 分别表示 Precision,Recall 和 *F*-measure,下同.从表 2 中我们可以看出,LP_{cosine} 和 LP_{JS} 都获得了比 SVM 更高的 Recall 值.特别是,当有标签的样本集较小时(取样百分比≤25%),LP 的性能更是大为改善.当取样百分比从 50%增加到 100%时,LP_{cosine} 仍然获得了和 SVM 差不多的 *F*-measure 值,而 LP_{JS} 的 *F*-measure 值则比 SVM 算法稍高些.

Table 2 Performance of relation detection on relation subtypes

表 2 在关系子类型上检测是否存在关系的实验结果

Percentage (%)	SVM			LP_{cosine}			LP_{JS}		
	P	R	F	P	R	F	P	R	F
1	35.9	32.6	34.4	58.3	56.1	57.1	58.5	58.7	58.5
10	51.3	41.5	45.9	64.5	57.5	60.7	64.6	62.0	63.2
25	67.1	52.9	59.1	68.7	59.0	63.4	68.9	63.7	66.1
50	74.0	57.8	64.9	69.9	61.8	65.6	70.1	64.1	66.9
75	77.6	59.4	67.2	71.8	63.4	67.3	72.4	64.8	68.3
100	79.8	62.9	70.3	73.9	66.9	70.2	74.2	68.2	71.1

表 3 是关系检测和分类任务的结果,性能评估描述了各主要关系子类型的 Precision,Recall 和 *F*-measure 的平均值.其中,SVM 和 LP 算法分别在不同大小的有标签样本集上测试.从表 3 中我们可以看出,几乎对于所有取样大小不同的有标签数据集,由于 Recall 值的提高,LP_{cosine} 和 LP_{JS} 都获得了比 SVM 更高的 *F*-measure 值.当有标签数据集相对较小时(取样百分比≤50%),LP 算法与 SVM 算法之间的结果差距会更大.当取样百分比从 75%增加到 100%时,LP 仍然获得了和 SVM 差不多的性能.此外,基于 JS 的 LP 算法结果也总是略好于基于 Cosine 的方法.

Table 3 Performance of relation detection and classification on relation subtypes

表 3 在关系子类型上进行关系检测和识别的实验结果

Percentage (%)	SVM			LP_{cosine}			LP_{JS}		
	P	R	F	P	R	F	P	R	F
1	31.6	26.1	28.6	39.6	37.5	38.5	40.1	38.0	39.0
10	39.1	32.7	35.6	45.9	39.6	42.5	46.2	41.6	43.7
25	49.8	35.0	41.1	51.0	44.5	47.3	52.3	46.0	48.9
50	52.5	41.3	46.2	54.1	48.6	51.2	54.9	50.8	52.7
75	58.7	46.7	52.0	56.0	52.0	53.9	56.1	52.6	54.3
100	60.8	48.9	54.0	56.2	52.3	54.1	56.3	52.9	54.6

4.4 实验结果:LP vs. Bootstrapping

为了与 Zhang[11]提出的基于 SVM 的 Bootstrapping 方法进行比较,我们采用了该方法中所采用的相同的特征集合,同时从训练数据集中随机地选取了 100 个样本作为初始的有标签的样本数据.表 4 列出了 Zhang 的方法和我们的 LP 方法对于每一个关系类型的分类结果.从表中我们可以看出,LP 算法在 ROLE,PART,AT 和 NEAR 四个关系类型上都获得了比 Zhang 的方法更好的性能,而在关系类型 SOC 上也取得了差不多的结果.

4.5 实验总结

从以上的实验结果可以看出,当有标签的样本数据不足时,LP 关系抽取算法取得了比基于 SVM 或基于 Bootstrapping 的方法更好的性能.原因如下:本文提出的基于图的半监督学习方法可以充分利用图的结构来平滑未标签样本的标签信息.因而,未标签样本的标签信息不仅由附近的有标签样本决定,而且还由附近的未标签样本决定.而对于有监督的学习方法,如 SVM,太少的有标签样本不足以揭示每一个关系类别的分类边界,而它又无法借助于未标签样本的信息来发现各类结构信息,所以无法相对准确地学习出分类超平面.从而导致运行

的效果不佳.所以,当有标签的样本数据太少时,我们提出的基于图的关系抽取方法在关系检测和分类上都能获得更好的结果.

Table 4 Comparison of the performance of the bootstrapped SVM method by Zhang[11] and LP method with 100 seed labeled examples for relation type classification task

表 4 Zhang[11]的基于 SVM 的 Bootstrapping 方法和 LP 算法对于关系类型分类任务的运行结果比较(作为种子的有标签样本的数目为 100)

Relation type	Bootstrapping			LP_{JS}		
	P	R	F	P	R	F
ROLE	78.5	69.7	73.8	81.0	74.7	77.7
PART	65.6	34.1	44.9	70.1	41.6	52.2
AT	61.0	84.8	70.9	74.2	79.1	76.6
SOC	47.0	57.4	51.7	45.0	59.1	51.0
NEAR	–	–	–	13.7	12.5	13.0

目前,针对 ACE 的 RDC 任务的工作都以基于有监督学习的方法为主.表 5 对 3 个典型的系统在 ACE 上的关系检测和分类的运行结果作了比较.从表 5 中可以看出,Zhou 等人[8]对于关系子类型的检测和分类取得了最好的结果,为 63.1%/49.5%/55.5%.与 Zhou 等人的方法相比,我们在 ACE 上报告的结果相对要差些,可能的原因是我们使用了一个比 Zhou 等人的方法中相对较小的特征集.在本文中,我们集中考察如何对关系抽取任务建立图模型,以及在这个图模型上进行关系的自动发现.在将来的工作中,我们也会进一步挖掘能够最贴切地表示实体对的知识来指导关系抽取的任务,从而提高抽取的准确度.

Table 5 Comparison of the performance of previous methods on ACE RDC task

表 5 典型的系统在 ACE 上的运行结果比较

Method		Relation detection			Relation detection and classification					
					On types			On subtypes		
		P	R	F	P	R	F	P	R	F
Culotta, *et al.*[2]	Tree kernel based	81.2	51.8	63.2	67.1	35.0	45.8	–	–	–
Kambhatla[3]	Feature based, maximum entropy	–	–	–	–	–	–	63.5	45.2	52.8
Zhou, *et al.*[8]	Feature based, SVM	84.8	66.7	74.7	77.2	60.7	68.0	63.1	49.5	55.5

5 相关技术比较分析

近年来,基于图的分类算法由于对全局一致性假设的关注,引起了越来越多学者的研究兴趣[20−25],其中包括了谱分析、随机游走、图切割等各种各样的算法策略.无论采用何种策略,这些基于图的半监督分类方法都假设了在整个图上的标注过程是平滑的,即满足全局一致性的假设,同时一般也都具有非参数、有判别力和直推式学习的特性.它们的主要区别在于如何实现一致性的假设,也就是如何定义和选择损失函数和规范式,即对全局一致性假设的两个限制条件的不同阐述.

本文所采用的 LP 算法是一个简单的循环标签传递过程,把图中一个节点的标签通过节点间的相似度传递给所有的节点,同时在循环传递标签的每一次迭代中始终保持有标签节点的标签信息固定.该算法利用了由大量无标签数据所定义的簇结构,并且假设该结构和分类的目标是相关联的.值得一提的是,该 LP 算法基于图的标签传递思想与 Zhou 等人[23]的工作初衷很类似.Zhou 等人所提出的基于图的分类算法关键点也是把每一个节点的标签信息循环传递到它的相邻节点,直到达到一个全局稳定的状态.两种算法最显著的区别之一就是如何使用有标签节点的标签信息.Zhou 等人的方法在每一次循环传递标签的迭代中,每一个节点都接收了两方面的信息:其一是来自相邻节点的信息,其二是自身的初始标签信息.然后,再通过参数α来指定两者信息的相关度.通过这样的方式可以将标签信息扩散开来,但是,那些有标签节点的初始标签信息的作用在扩散过程中因为参数α的限制而被淡化.此外,如何自动指定合适的参数α的值也是一个难点.Zhou 等人在实验中指定了参数α的值为 0.99,也即指明了来自相邻节点的信息的相关度为 0.99,而自身的初始标签信息的相关度为 0.01,可见,有标签节点的初始标签信息在整个算法的标签扩散中并没有得到最充分的利用.而在我们的 LP 算法中,有标签节点的标

签信息在每一次标签的全图传递后都将被填回,因此,这些标签信息就像是固定资源一样被循环推开,直到整个系统达到一个稳定的状态.有标签节点的初始标签信息在整个算法的标签传递中得到了最充分的利用.所以,当拥有相当数目的有标签节点时,LP 算法有可能针对微弱或模糊边界数据,获得比 Zhou 等人的方法更为精确的分类结果.此外,两种算法的区别还在于对 Laplacian 算子的定义上,Zhou 等人的方法在规范式中使用了正规 Laplacian 算子,即 $D^{-1/2}WD^{-1/2}$,其中 D 是对角矩阵 ($d_i = \sum_j w_{ij}$),而我们的 LP 算法中使用的是图的组合 Laplacian 算子.在后续工作中,我们将通过实验进一步验证我们的上述分析结果,也将针对不同的 Laplacian 算子的定义对 LP 算法的影响继续展开研究,确定出最适合关系抽取任务的 Laplacian 算子.

6 结束语

随着信息技术的飞速发展,收集大量未标记的样本已经相当容易,如何利用大量的未标记样本改善学习性能已成为当前机器学习研究中备受关注的问题.本文的特色就是利用未标记样本进行半监督学习关系的自动抽取,解决有标签样本数据不足的问题,从而减少标记样本时人力、物力的耗费.其主要创新点有:

(1) 提出用图策略来建立关系抽取的模型,当有标签样本不足时,依靠无标签样本的信息能够尽可能准确地挖掘出数据样本内部所潜在的结构信息来协助候选关系的分类.

(2) 充分利用建立的关系候选的图模型,提出利用半监督学习的标签传递算法在该模型上进行标签的全局传递,克服已有的基于 SVM 和 Bootstrapping 算法的关系抽取模型中所呈现的因为有标签样本不足时局部一致性假设所带来的弊端,实现全局一致性的目标,提高分类的准确度.

致谢 在此,我们向对本文的工作给予支持和建议的新加坡国立大学和新加坡信息通信研究局的同行表示感谢,并对各位论文评审专家给出的宝贵意见表示感谢.

References:

[1] Bunescu R, Mooney RJ. A shortest path dependency kernel for relation extraction. In: Proc. of Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP). Morristown: Association for Computational Linguistics, 2005. 724−731.

[2] Culotta A, Soresen J. Dependency tree kernels for relation extraction. In: Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004). Morristown: Association for Computational Linguistics, 2004. 423−430.

[3] Kambhatla N. Combining lexical, syntactic and semantic features with maximum entropy models for extracting relations. In: Proc. of the ACL Interactive Poster and Demonstration Sessions. Morristown: Association for Computational Linguistics, 2004. 178−181.

[4] Miller S, Fox H, Ramshaw L, Weischedel R. A novel use of statistical parsing to extract information from text. In: Proc. of the 6th Applied Natural Language Processing Conf. Morristown: Association for Computational Linguistics, 2000. 226−233.

[5] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. In: Haji J, Matsumoto Y, eds. Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). Morristown: Association for Computational Linguistics, 2002. 71−78.

[6] Zhang M, Zhang J, Su J, Zhou GD. A composite kernel to extract relations between entities with both flat and structured features. In: Proc. of the 21st Int'l Conf. on Computational Linguistics and the 44th Annual Meeting of the ACL. Morristown: Association for Computational Linguistics, 2006. 825−832.

[7] Zhao SB, Grishman R. Extracting relations with integrated information using kernel methods. In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2005. 419−426.

[8] Zhou GD, Su J, Zhang J, Zhang M. Exploring various knowledge in relation extraction. In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2005. 427−434.

[9] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections. In: Proc. of the 5th ACM Int'l Conf. on Digital Libraries (ACMDL 2000). New York: ACM Press, 2000. 85−94.

[10] Brin S. Extracting patterns and relations from world wide Web. In: Atzeni P, Mendelzon AO, Mecca G, eds. Proc. of the WebDB

Workshop at the 6th Int'l Conf. on Extending Database Technology (WebDB'98). Heidelberg: Springer-Verlag, 1998. 172−183.

[11] Zhang Z. Weakly-Supervised relation classification for information extraction. In: Proc. of ACM the 13th Conf. on Information and Knowledge Management (CIKM 2004). Washington: ACM Press, 2004. 581−588.

[12] Chen JX, Ji DH, Chew LT, Niu ZY. Automatic relation extraction with model order selection and discriminative label identification. In: Dale R, Wong KF, Su J, Kwong OY, eds. Proc. of the 2nd Int'l Joint Conf. on Natural Language Processing (IJCNLP 2005). Heidelberg: Springer-Verlag, 2005. 390−401.

[13] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora. In: Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2004. 415−422.

[14] Zhang M, Su J, Wang DM, Zhou GD, Chew LT. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In: Dale R, Wong KF, Su J, Kwong OY, eds. Proc. of the 2nd Int'l Joint Conf. on Natural Language Processing (IJCNLP 2005). Heidelberg: Springer-Verlag, 2005. 378−389.

[15] Che WX, Liu T, Li S. Automatic entity relation extraction. Journal of Chinese Information Processing, 2005,19(2):1−6 (in Chinese with English abstract).

[16] Dong J, Sun L, Feng YY, Huang RH. Chinese automatic entity relation extraction. Journal of Chinese Information Processing, 2007,21(4):80−85, 91 (in Chinese with English abstract).

[17] He TT, Xu C, Li J, Zhao JX. Named entity relation extraction method based on seed self-expansion. Computer Engineering, 2006, 32(21):183−184,193 (in Chinese with English abstract).

[18] Liu KB, Li F, Liu L, Han Y. Implementation of a kernel-based Chinese relation extraction system. Journal of Computer Research and Development, 2007,44(8):1406−1411 (in Chinese with English abstract).

[19] Zhang SX, Wen J, Qin Y, Yuan CX, Zhong YX. Study about automatic entity relation extraction. Journal of Harbin Engineering University, 2006,27(B07):370−373 (in Chinese with English abstract).

[20] Belkin M, Niyogi P. Using manifold structure for partially labeled classification. In: Thrun BS, Obermayer K, eds. Advances in Neural Information Processing Systems 15. Cambridge: MIT Press, 2003. 926−936.

[21] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: Carla EB, Andrea PD, eds. Proc. of the 18th Int'l Conf. on Machine Learning. (ICML 2001). San Fransisco: Morgan Kaufmann Publishers, 2001.

[22] Blum A, Lafferty J, Rwebangira MR, Reddy R. Semi-Supervised learning using randomized mincuts. In: Carla EB, ed. Proc. of the 21st Int'l Conf. on Machine Learning. Banff: ACM Press, 2004. 934−947.

[23] Zhou DY, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. In: Thrun S, Saul LK, Scholkopf B, eds. Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2004.

[24] Zhu XJ, Ghahramani ZB. Learning from labeled and unlabeled data with label propagation. Technical Report, CMU-CALD-02-107, CMU CALD, 2002.

[25] Zhu XJ, Ghahramani ZB, Lafferty J. Semi-Supervised learning using Gaussian fields and harmonic functions. In: Fawcett T, Mishra N, eds. Proc. of the 20th Int'l Conf. on Machine Learning. AAAI Press, 2003. 912−919.

附中文参考文献:

[15] 车万翔,刘挺,李生.实体关系自动抽取.中文信息学报,2005,19(2):1−6.

[16] 董静,孙乐,冯元勇,黄瑞红.中文实体关系抽取中的特征选择研究.中文信息学报,2007,21(4):80−85,91.

[17] 何婷婷,徐超,李晶,赵君喆.基于种子自扩展的命名实体关系抽取方法.计算机工程,2006,32(21):183−184,193.

[18] 刘克彬,李芳,刘磊,韩颖.基于核函数中文关系自动抽取系统的实现.计算机研究与发展,2007,44(8):1406−1411.

[19] 张素香,文娟,秦颖,袁彩霞,钟义信.实体关系的自动抽取研究.哈尔滨工程大学学报,2006,27(B07):370−373.

陈锦秀(1978－),女,福建厦门人,博士,助教,主要研究领域为自然语言处理,信息抽取,机器学习.

姬东鸿(1967－),男,博士,教授,博士生导师,主要研究领域为自然语言处理,数据挖掘.