ISSN 1000-1239/CN 11-1777/TP
50(9): 1825-1835, 2013

基于弱监督学习的海量网络数据关系抽取

陈立玮　冯岩松　赵东岩
(北京大学计算机科学技术研究所　北京　100871)
(clwclw88@pku.edu.cn)

Extracting Relations from the Web via Weakly Supervised Learning

Chen Liwei, Feng Yansong, and Zhao Dongyan
(*Institute of Computer Science and Technology*, *Peking University*, *Beijing* 100871)

Abstract In the time of big data, information extraction at a large scale has been an important topic discussed in natural language processing and information retrieval. Specifically, weak supervision, as a novel framework that need not any human involvement and can be easily adapted to new domains, is receiving increasing attentions. The current study of weak supervision is intended primarily for English, with conventional features such as segments of words based lexical features and dependency based syntactic features. However, this type of lexical features often suffer from the data sparsity problem, while syntactic features strongly rely on the availability of syntactic analysis tools. This paper proposes to make use of *n*-gram features which can relieve to some extent the data sparsity problem brought by lexical features. It is also observed that the *n*-gram features are important for multilingual relation extraction, especially, they can make up for the syntactic features in those languages where syntactic analysis tools are not reliable. In order to deal with the quality issue of training data used in weakly supervised learning models, a bootstrapping approach, co-training, is introduced into the framework to improve this extraction paradigm. We study the strategies used to combine the outputs from different training views. The experimental results on both English and Chinese datasets show that the proposed approach can effectively improve the performance of weak supervision in both languages, and has the potential to work well in a multilingual scenario with more languages.

Key words relation extraction; weakly supervised learning; maximum entropy; co-training; knowledge base construction

摘　要 在大数据时代,对于海量网络数据的信息抽取与应用已成为自然语言处理和信息检索技术发展的重要主题.其中,基于弱监督的关系抽取方法,因为具有不需要过多人工参与、适应性强的特点,受到了广泛的关注.目前针对它的研究主要集中在英语资源上,主要使用传统的词法和句法特征.然而,词法特征有严重的稀疏性问题,句法特征则对一些语言分析工具的性能有较强的依赖性.提出利用 *n*-gram 特征来缓解传统词法特征稀疏性的问题.特别地,这种特征还可以弥补传统句法特征在其他语言上不可靠的情况,对于关系抽取的跨语言应用有重要作用.在此基础上,针对弱监督学习中标注数据不完全可靠的情况,提出基于 bootstrapping 思想的协同训练方法来对弱监督关系抽取模型进行强化,并且对预测关系时的协同策略进行了详细分析.在大规模的中文和英文数据上进行实验的结果显示,把传统特征

收稿日期:2013-03-30;**修回日期**:2013-06-17
基金项目:国家"八六三"高技术研究发展计划基金项目(2012AA011101);国家自然科学基金项目(61272344,61202233)
通信作者:冯岩松(fengyansong@pku.edu.cn)

与 n-gram 特征相结合并进行协同训练,在中文和英文数据集上均可以提升弱监督关系抽取的效果,可以适应多语言的关系抽取需求.

关键词 关系抽取;弱监督学习;最大熵模型;协同训练;知识库构建

中图法分类号 TP18

近些年来,人工智能技术的发展非常迅速,并且已经被广泛地应用在人类的生活中.知识对于人工智能技术来说是一个非常重要的因素.构建一个大规模、高质量、多语言的结构化知识库对于人工智能技术的发展有着重大的促进作用.传统方法主要依靠手工构建知识库,这种方式构建起来的知识库的规模较小,领域也局限在构建知识库的专家所熟悉的范围内,难以进行扩展.因此,自动构建知识库已经是大势所趋.

互联网无疑是知识的一个重要载体,随着 Web2.0 的出现,互联网上的数据呈爆炸式增长,人们已经进入了大数据时代.但是互联网主要承载非结构化的数据,我们需要把非结构化数据转化为知识库中的结构化数据.因此,多语言的关系抽取是自动构建大规模知识库的一项核心技术,对人工智能的发展有着重要的意义.

关系抽取的主要任务是利用包含一对命名实体的自然语言文本来确定这两者之间的关系.近些年来,关于关系抽取的研究吸引了自然语言处理和信息检索领域很多研究者的关注.最初的关系抽取任务主要出现在美国国家标准与技术研究院(National Institute of Standards and Technology, NIST)组织的 MUC[1] 和 ACE① 评测中.但是,这种关系抽取主要基于规则或者有监督学习,依赖于人类专家所标注的训练数据或制定的规则,所能够抽取的关系也受限于此,因此并不能满足从海量的网络数据中抽取大量的、多种语言的关系的需求.

在网络环境中,开放域的关系抽取(open information extraction)受到了广泛的关注.TextRunner[2], ReVerb[3] 和 NELL[4] 就是几项有代表性的工作.它们不需要训练数据,只是使用基于语法、句法的模板以及启发式规则来从海量的网络信息中快速抽取出大量的句子片段,称之为"关系".然而,这种方法抽取出的"关系"难以规则化,因此很难被用来构建知识库,距离自动化构建大规模知识库还有很远的距离.

弱监督学习(weakly supervised learning)[5] 是一种基于噪声训练数据的半监督的关系抽取框架.它通过把知识库中的关系与可靠的文本集进行匹配来构建训练集,而后用这个训练集训练一个分类器来预测关系.弱监督学习整合了有监督学习和开放域关系抽取的优点:它可以自动从海量文本数据中抽取出规则化的关系,并且可以较容易地被复用到其他领域中.

以往的弱监督关系抽取的研究多数都基于英文,所使用的词法特征和句法特征都依赖于当前语言的资源和工具,比如词法特征使用的是实体间的词序列,句法特征使用的是依存关系路径.语言中的修饰语可能导致词法特征过于稀疏,句法分析工具的错误可能造成句法特征的不可靠,因此传统特征的鲁棒性比较差,难以满足多语言关系抽取的需求.在本文中,我们提出了 n-gram 特征来对传统特征进行补充.它可以从过长的词法特征中挖掘出隐含的有用信息,从而缓解词法特征稀疏性所带来的问题.更重要的是,它不依赖任何句法分析工具,因而可以弥补多语言关系抽取中分析工具不可靠带来的问题.因此,n-gram 特征相比传统特征具有更好的鲁棒性,适用于多语言的关系抽取.

与传统有监督学习基于可靠的训练数据不同,弱监督学习的训练数据是基于一些不严密的假设生成的,因此训练集中可能含有大量噪声和错误.bootstrapping 方法可以通过多次的迭代来引入比较可靠以及覆盖面更广泛的样本,从而对分类器进行强化.在本文中,我们使用协同训练(co-training)方法[6] 来缓解弱监督关系抽取数据中噪声所带来的问题,并且具体分析了预测关系时所用的协同策略.

多数对弱监督关系抽取的研究都是针对英文数据进行的.中文是互联网上使用最多的语言之一,因而对中文实体关系抽取的研究也是非常重要的.虽然国内有许多学者对中文实体关系抽取进行了深入的研究[7-9],但是到目前为止,还没有针对中文进行弱监督关系抽取的研究工作.在本文中,我们通过对中文和英文数据进行实体关系抽取,验证了基于 n-gram 特征和协同训练的弱监督关系抽取可以适应多语言关系抽取的需求.

① http://projects.ldc.upenn.edu/ace

我们的主要贡献有:

1) 引入了鲁棒性更好的 n-gram 特征来对传统特征加以补充,并且研究了它所造成的影响.

2) 应用协同训练的方法来对弱监督学习进行强化,并且重点研究了不同协同策略对结果的影响.

3) 在中文、英文大规模数据集上进行了实验,验证了 n-gram 特征和协同训练在多语言环境下对弱监督关系抽取效果的提升.

1 相关工作

1.1 弱监督关系抽取

弱监督关系抽取框架在近几年被大量应用在关系抽取中. 图 1 所示的是弱监督学习的基本框架:

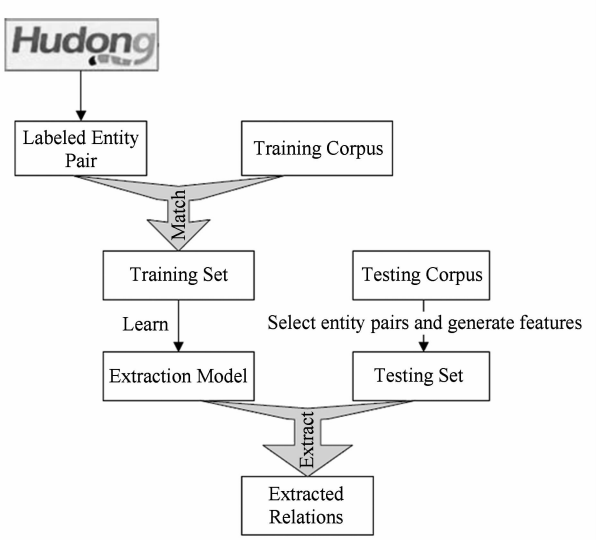

Fig. 1 Procedure of weakly supervised relation extraction.

图 1 弱监督关系抽取流程

它需要一个以三元组形式来组织的知识库(如互动百科的 infobox 构成的知识库),和来源可靠的大量文本以及一个有监督的分类器(如最大熵模型). 对于知识库中的关系,在文本中寻找同时含有它的主体和客体的句子,用来构成它的训练集. 然后,用这个训练集训练一个分类器. 对于测试文本,首先要确认需要抽取关系的实体对,而后找出同时含有这两个实体的句子作为测试数据,将测试数据输入分类器中,就可以得到最终的答案.

弱监督关系抽取最早由 Craven 和 Kumlien 提出[5],被用于从学术文献的摘要中抽取蛋白质与基因之间的关系. 文献[10-11]把这种思想应用在维基百科中,他们利用维基百科中含有结构化信息(即 infobox)的页面来生成关系抽取的训练数据. 但是,他们的实验仅仅在维基百科中不含有 infobox 的页面中进行,没有扩展到更广的领域中. Mintz 等人[12]提出了一个新的弱监督关系抽取的框架,他们利用 Freebase① 和维基百科② 中的文本来构造具有噪音的训练集,整个过程不需要人工参与. 他们构造数据集的假设是:如果两个实体之间存在关系,那么含有这两个实体的句子或多或少都描述了这个关系. Yao 等人[13]对 Mintz 等人的框架进行了改进,把关系抽取和实体的种类综合考虑,利用实体的类别来过滤掉部分错误的关系预测. Riedel 等人[14]认为 Mintz 等人使用的假设过于严格,这个假设在很多情况下并不成立. 比如在知识库中,我们找到了关于"首都"关系的一个三元组〈英国,首都,伦敦〉,并在新闻报道中找到了一个包含上述实体对的句子:"英国经济和工商业研究中心预计,今年伦敦 32.4 万名金融从业人员中将有 10%的人饭碗不保",很显然这句话并没有描述英国首都是伦敦这个关系. 因而 Riedel 等人把它放松为如果两个实体之间存在关系,那么至少有一个含有这两个实体的句子描述了这个关系. Surdeanu 等人[15]把弱监督关系抽取的框架应用在了 TAC-KBP 的属性填充(slot filling)评测中,结果表明弱监督关系抽取在这项任务中有较好的应用前景. 前面的工作都基于一对实体之间只能有一个关系的假设,然而这个假设并不符合事实. Hoffmann 等人[16]提出了一个联合概率模型来解决一个实体对之间有多个关系的问题,例如微软和比尔·盖茨之间就可能存在两个关系:〈微软,创建人,比尔·盖茨〉和〈微软,首席执行官,比尔·盖茨〉. Surdeanu 等人[17]进而提出一个新的多实例多标签学习模型,可以把关系之间的依赖因素考虑在内,从而提高关系抽取的效果.

1.2 *n*-gram 特征

n-gram 是从文本中截取出来的 n 个连续词的序列,可以捕捉到相邻词语的序列关系,在一定程度上体现了局部范围的语法习惯,并且具有较好的跨语言能力. 它被广泛应用在自然语言处理的任务中,尤其是构建语言模型. Marino 等人设计了一个基于 n-gram 的机器翻译模型[18],取得了当时最好的效果. Furnkranz 的研究显示 bigram 和 trigram 在文本分类

① www.freebase.com

② www.wikipedia.org

中是最有效的 n-gram 特征[19].

1.3 协同训练

协同训练是 Blum 和 Mitchell 提出的一种半监督学习的框架[6]. 它将待分类数据的特征集合分为两个互相独立的视图,每个视图都可以单独训练一个分类器,两者相互学习,相互强化. 它被广泛应用在自然语言处理和信息检索领域中,如句法分析[20-22]、词义消歧[23]、图像检索[24-25]等. 但是,两个视图条件独立的要求是非常强的,在实际中很难达到. 因此,也有一些研究工作着力于放松这个限制. Abney 的工作显示了完全独立的条件可以被放松为有较弱的联系[26],而 Balcan 等人的研究对这个条件进行了进一步的放松[27]. 王娇等人提出了一个基于随机子空间的方法来解决视图不满足条件独立假设和不同的物体两个视图完全一致这两个问题[28]. Wang 和 Zhou 在此基础上又对协同训练进行了深入的研究,给出了协同训练能够生效的充要条件[29]. 对于只有一个视图的问题,一些研究工作也表明了如果使用几个差别很大的分类器,以协同训练的形式进行训练,同样可以得到理想的结果[30-31].

2 基于弱监督学习的关系抽取

基于弱监督学习的关系抽取框架最大的优点在于不需要人工参与,只要找到某领域的知识库,以及来源可靠的文本集合,就可以进行自动的关系抽取,并且可以被复用在任何可靠的文档集中. 本节主要介绍弱监督关系抽取的两个重要因素:特征和学习模型.

2.1 弱监督关系抽取的特征

弱监督关系抽取使用的特征包括传统的词法特征、句法特征以及我们新引入的 n-gram 特征. 表 1 中给出了三元组〈奥巴马,毕业院校,哥伦比亚大学〉与句子"奥巴马,美国总统,1983 年毕业于哥伦比亚大学."匹配后所抽取出的特征. 下面分别加以介绍.

1) 词法特征(lexical features, lex). 对于一个实体对(e_1,e_2),词法特征主体部分由它们之间的词序列和词性序列构成. 表 1 中展示了几个词法特征的例子. 我们可以发现,词法特征会因为句子中的修饰成分而过于具体,很难在其他的句子中再次出现. 因而这些特征往往无法为关系抽取做出贡献.

2) 句法特征(syntactic features, syn). 句法特征从句子的依存关系分析结果中获取,如图 2 所示.

我们使用句法依存关系分析器来对句子进行解析. 表 1 中展示了句法特征的例子. 句法特征依赖于依存关系的分析,因而在多语言关系抽取中,很可能因为工具不可靠导致句法特征不可靠.

3) n-gram 特征. 词法和句法特征的一个缺陷是特征表述太具体,往往与训练集中的句子密切相关,如两个实体之间的词序列,因此带来显著的数据稀疏性问题,也使得模型缺乏泛化能力. 而 n-gram 特征通常是文本中 n 个连续词组成的序列,可以捕捉到局部范围内连续词语之间的序列关系,体现语法习惯. 这样的特征,因为只包含 3 到 4 个词,因而不会像传统词法特征那样过于具体,导致特征稀疏,几乎不可能再现.

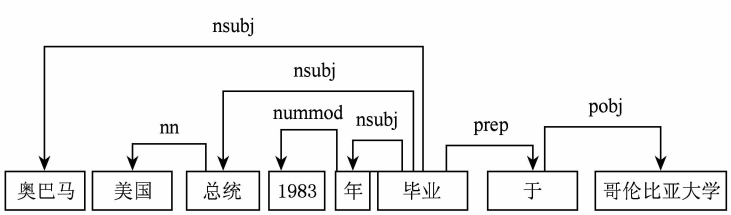

Fig. 2 The dependency parsing results of the sample sentence in section 2.1.

图 2 2.1 节例句的依存关系分析结果

另一方面,传统句法特征依赖于依存关系分析器的结果,在一些资源与工具不足的语言环境中,准确可靠的句法与依存关系分析结果并不容易获得,据此抽取的句法特征通常也存在问题. 在这种情况下,n-gram 特征则体现了很好的补充作用,几个 3~4 个词构成的局部短语结构(注意,这并不是严格意义上的短语,而且一些语法结构的片段),从一定程度上体现了这两个实体之间的句法结构,从而弥补了不可靠的句法特征所带来的问题. 从多语言资源的角度来看,n-gram 特征在不同的语言环境中都表现出较好的一致性,对于提高多语言关系抽取模型的鲁棒性具有重要意义.

除了传统的利用连续的 n 个词语构成的 n-gram (lexical n-gram, lng)以外,我们还把连续词语的词性标注组织成新的词性序列 n-gram 特征(pos n-gram, png). 这种基于词性标注的新特征比基于词

语的 n-gram 特征更加一般化,因此可以预计的是,这种特征有可能会进一步降低准确率,但是带来召回率的提升.我们还尝试把词语和它的词性标注组合起来作为一个 gram 来构造另一种新的 n-gram 特征(lexical-pos n-gram, lpng),如表 1 所示:

Table 1 Features Extracted from Sample Sentence in Section 2.1

表 1 从 2.1 节例句中抽取的特征

Feature Type	Example
Lexical	Inverse_false PER,美国总统,1983 年毕业于 ORG
	Inverse_false PER PU NR NN PU CD M VV P ORG
Syntactic	PERSON→毕业/VV←于/P←ORGANIZATION
n-gram	毕业于 ORGANIZATION
	VV P ORGANIZATION
	毕业/VV 于/P ORGANIZATION

在本文中,我们使用 tri-gram,即令 $n=3$.

2.2 弱监督学习的模型

目前在弱监督学习中广泛采用的有监督机器学习模型是逻辑斯谛回归模型(logistic regression)和最大熵模型.后者被广泛应用在自然语言处理领域.文献[32]最早将最大熵模型应用在自然语言处理中,并且通过双语消歧、词语重排序和句子切分几项任务来验证其可行性.由于最大熵模型估计的是条件概率,因而没有像生成模型那样对特征做互相独立这样的过强假设.本文使用它作为协同训练框架中的基础模型.

最大熵模型的原理是,在只掌握了关于未知分布的部分知识时,应选取符合这些知识并且熵最大(即不确定性最大)的概率分布.对于上下文 x 和标签 y,给定 x 得到 y 的条件概率可以定义为

$$p(y \mid x) = \frac{1}{Z(x)}\exp\left[\sum_{i=1}^{k}\lambda_i f_i(x,y)\right], \quad (1)$$

其中 $f_i(x,y)$是第 i 个特征的特征方程,λ_i 是特征方程 $f_i(x,y)$的权重,k 是特征总数,$Z(x)$是归一化因子,确保概率和为 1.

$f_i(x,y)$一般通过上下文 x 和相应的标签 y 来定义,当 x 与 y 满足一定的条件时值为 1,否则值为 0.比如,在关系抽取中,可以定义如下的 $f_i(x,y)$:

$$f_i(x,y)\begin{cases}1, \text{ if } y = BirthPlace \text{ and}\\ \quad feature\_contains\_PERSON\text{-}WAS\text{-}\\ \quad BORN\text{-}IN\text{-}LOCATION(x) = \text{true};\\ 0, \text{ otherwise.}\end{cases} \quad (2)$$

其中,$feature\_contains\_PERSON\text{-}WAS\text{-}BORN\text{-}IN\text{-}LOCATION(x)=\text{true}$ 表示上下文 x 中含有特征 $PERSON\text{-}WAS\text{-}BORN\text{-}IN\text{-}LOCATION$.

最大熵模型使用最大似然估计来进行参数的估计.这是一项很有难度的工作.目前有两种常见的方法:通用迭代算法(generalized iterative scaling, GIS)[33]和改进迭代算法(improved iterative scaling, IIS)[34].

3 协同训练

3.1 弱监督关系抽取与 bootstrapping

在弱监督关系抽取中,训练数据是根据假设自动生成的,而这样的假设并不总是成立.因此,与传统的有监督学习拥有准确的训练数据不同,弱监督学习的训练数据中有大量的噪声以及错误标注,正确的标注可能并不多.显然在这样的训练数据上训练出来的分类模型的可靠性是值得怀疑的.而 bootstrapping 的提升方法可以通过多次迭代,每次加入一部分高置信度数据的方式来扩充训练集中的正确标注数据,从而对关系抽取模型进行强化.

特别是我们注意到,在训练数据中出现过的特征,只有一小部分在测试数据中重现.比如,在我们随机生成的一个测试集中,只有 16%的特征在对应的训练集中出现过.因此,很多测试数据可能因为训练集中相应特征过少甚至不存在,导致类别被预测错误.而在面对开放领域的文本集时候,这种情况就更加常见.因此我们也希望 bootstrapping 方法可以通过不断的迭代,从大量的无标注数据中引入新的训练数据,从而引入新的特征,进而缓解这样的问题.

在本文中,我们尝试利用协同训练来促进大规模关系抽取.需要在此强调的是,我们需要面对的是多标签多分类任务,协同训练本身的特点也许可以帮助我们实现预测多个标签.我们将在设计协同策略的时候考虑这个问题.

3.2 协同训练的框架

协同训练是一种基于 bootstrapping 思想的半监督学习框架,它被广泛应用在数据挖掘、自然语言处理等领域.协同训练假设数据的特征集合可以被分解为两个不同的视图.每个视图都可以单独预测数据的标签,并且两个视图相对标签条件独立.两个视图将被用来训练两个分类器,它们将会各自预测无标注数据的类别,并且把自己预测结果中置信度

最高的数据加入对方的训练集中.然后,两个分类器将被重新训练,这个过程不断迭代,直至收敛或者达到停止条件.协同训练的过程如下:

输入:有标注的训练数据集 L,被分为两个不同的视图 L_1,L_2,无标注的数据集 U,被分为两个不同的视图 U_1,U_2;

输出:训练后的学习模型 M.

Step1. 从无标注数据集 U 中抽取出一个大小为 p 的数据池 U',分为两个视图 U'_1,U'_2.

Step2. 循环 t 次:

Step2.1. 使用 L_1 训练一个学习模型 M_1,用它来对 U'_1 进行预测.

Step2.2. 使用 L_2 训练一个学习模型 M_2,用它来对 U'_2 进行预测.

Step2.3. 把 L_1 预测的置信度最高的 n 个数据加入到 L_2 的训练集中.

Step2.4. 把 L_2 预测的置信度最高的 n 个数据加入到 L_1 的训练集中.

Step2.5. 从 U' 中删除掉已加入训练集的数据,并且从 U 中抽取出相应数量的数据来填充 U'.

我们可以看到,协同训练中有 3 个重要的参数:数据池的大小 p,每次迭代加入到标注数据集中的数据数量 n 和迭代的次数 t.需要强调的是,这里的 n 是对每个视图而言,因此每次迭代最多加入 $2n$ 个数据.在实验中,我们将会观察这 3 个参数对于协同训练结果的影响.

另外需要提到的是,在我们的关系抽取任务中共有 3 种不同的特征,虽然不能严格区分它们的独立性,但是这 3 种特征通常可以被认为是天然分开的不同侧面,这对于协同训练是有利的.

3.3 协同训练视图切分

我们的特征空间包含 3 种特征:词法特征、句法特征和 n-gram 特征.根据特征的特点,我们把特征空间切分为词法特征+句法特征、词法特征+ n-gram 特征.很显然,这两个视图是不独立的.但是根据文献[26]的结果,在两个视图之间的联系比较弱的情况下,协同训练仍然可以提高关系抽取的效果.

3.4 协同策略

对于测试数据,我们首先使用训练好的两个模型分别来预测它的标签.需要说明的是,我们把负标注 NA(表示两个实体间没有关系)也作为最大熵模型中的一个类别.为了得到最终答案,我们要同时考虑两个分类器的结果.在这里我们提出了两个策略,并用实验来说明哪个更有效.

第 1 个策略利用最大熵模型输出的置信度的加权和来判断最终的标签.这个策略可以表示为

$$l' = \arg\max_{l}(\lambda c_1(x,l) + (1-\lambda)c_2(x,l)), \quad (3)$$

其中,x 为测试数据,$c_1(x,l)$ 是第 1 个分类器对 x 的关系为 l 给出的置信度,$c_2(x,l)$ 是第 2 个分类器为对 x 的关系为 l 给出的置信度,λ 是对分类器加权的权值.即,我们选取使得两个分类器加权置信度和最大的那个类别.在实验中,我们取 $\lambda=0.5$,因为我们并没有确切的先验信息证明哪个分类器更可信.当我们对某分类器有更多了解时,可以在此基础上人工修正 λ,或者在留存数据集上调节这个参数.我们将在将来的工作中,对此进行更加详细的探讨.

第 2 个策略是投票策略.对于一个测试数据,如果两个模型预测的类别相同,则该类别为最终结果,最终的置信度设为较高的一方;如果两个模型预测的类别不同,则分为两种情况:如果有一个模型预测的类别为 NA,那么取另一个模型的结果和置信度作为最终答案;如果两个模型预测的类别都不是 NA,则我们认为两个模型的答案都是正确的.我们知道,两个实体之间的关系可能有多个,而最大熵模型只会预测出一个.我们制定的这个策略可以在一定程度上缓解这个问题:两个分类器从各自不同的角度来预测的结果,有可能都是对的.

实际上,我们所设计的协同策略可以被看成是一种重排序.对于从某个视角获取的预测结果,使用另一个视角训练出的模型进行重新加权和排序,进而得到一个最终的结果.McClosky 等人使用类似的策略在句法分析任务上取得了一些提升[35].

4 实验设置

4.1 知识库

弱监督关系抽取需要结构化知识库来构建训练集.知识的表达形式为三元组,即〈主体,关系,客体〉的结构.

我们选用了 DBpedia[36] 作为英文知识库.DBpedia 主要通过抽取维基百科页面中的结构化信息(如 Info box)来构建.它包含超过 3.46 亿个实体.英文版的 DBpedia 共包含 3.85 亿个来自不同领域的三元组.

对于中文,利用互动百科① 来构造知识库.互动百科是目前最大的中文在线百科全书之一,拥有超

① www.baike.com

过 500 万个页面.与维基百科类似,互动百科也是由志愿者们共同创建的.很多页面中包含一个结构化的信息盒(Info Box),含有关于当前页面所描述实体的大量结构化知识.我们从这些页面中抽取出这些结构化知识,作为中文的知识库.

4.2 文档集

对于英文,我们选用《纽约时报》2005—2007 年的新闻语料集作为文档集.与文献[14]一致,我们使用 2005 年和 2006 年两年的新闻数据作为训练集,2007 年的新闻数据作为测试集.对于中文,我们使用 4 份全国性经济类报纸(《经济观察报》、《经济参考报》、《第一财经日报》、《21 世纪经济报道》)在 2009 年的新闻数据作为文档集.我们使用 4 份报纸前 7 个月的新闻数据作为训练集,后 5 个月的数据作为测试集.

4.3 预处理

对于英文数据,我们使用斯坦福大学的自然语言处理工具[37-38]来进行命名实体识别、依存关系分析等预处理操作.对于中文数据,我们使用耶宝分词系统来进行分词和命名实体识别,Mate Parser[39]来进行依存关系分析.

4.4 基准算法和评测标准

我们一共使用了 4 个基准算法.前两个基准算法是分别在两个视图(词法特征+句法特征、词法特征+n-gram 特征)上训练的最大熵模型,分别用 lex+syn 和 lex+3lpng 表示.对于 3 种 n-gram 特征的选择将在 5.2 节详细讨论.

第 3 个基准算法是在整个特征空间上训练的最大熵模型,用 lex+syn+3lpng 表示.

MultiR 算法[16]是我们所使用的第 4 个基准算法.MultiR 由 Hoffmann 等人提出,它可以为一个实体对预测多个关系,是一个多实例多标签(multi instance multi label)的分类模型,是目前弱监督关系抽取中效果最优的算法之一.

对于算法的评测标准,我们与以前的相关工作一样[12-14,16-17],使用 P-R 曲线来比较算法的效果.他们计算 P-R 曲线的方法是:把所有实体对按照其关系预测的置信度从大到小排序,忽略掉被预测为 NA 的实体对,而后按置信度从大到小遍历该序列,在每一个实体对处计算所有排在它前面的实体对预测的精确率和召回率.精确率为预测正确的实体对数目除以预测实体对的总数,召回率为预测正确的实体对数目除以有关系实体对的总数.每个点都计算完后,就可以构成一条 P-R 曲线.通常,P-R 曲线分布越接近坐标系的右上方,延伸的越远,表明模型的效果越好.

5 实验结果与讨论

5.1 关系抽取的结果分析

图 3 和图 4 展示了在两种语言上,协同训练与基准算法的 P-R 曲线.

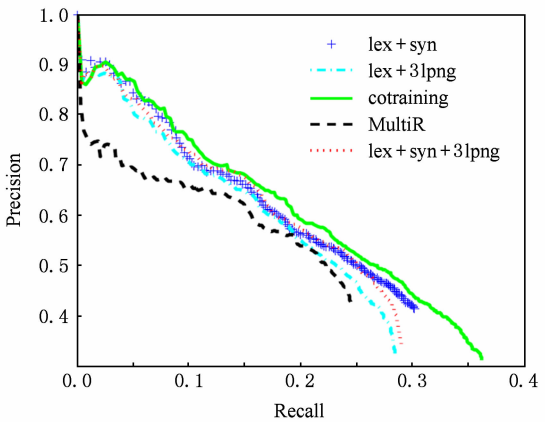

Fig. 3 Relation extraction performance on the English dataset.

图 3 英文数据集上关系抽取算法的效果

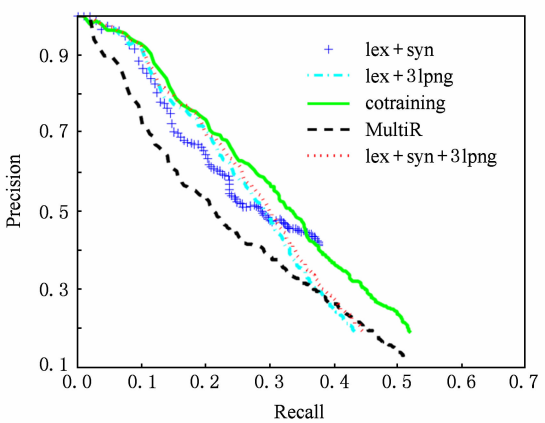

Fig. 4 Relation extraction performance on the Chinese dataset.

图 4 中文数据集上关系抽取算法的效果

我们可以看到,在两种语言的 P-R 曲线上,协同训练模型与在两个视图上单独训练或者在全部特征上训练的最大熵模型相比,效果都有提升或者与之相当.而协同训练模型在两种语言上的效果均高于 MultiR,并且提升非常明显.这说明了 MultiR 算法在开放数据集和多语言的环境下并没有较好的适应性,也说明了协同训练确实可以提升弱监督关系抽取的效果,并且在多语言的环境里一样可以有较好的表现.我们发现在英文数据集中,协同训练所带来

的提升比在中文数据集中要小. 我们猜测这是因为英文数据的句法特征比较可靠,因而导致 n-gram 特征不会对其有明显的补充作用. 我们还注意到,在中文上,词法特征+n-gram 特征的结果比词法特征+句法特征具有一定的优势,而在英文上则处于微弱劣势. 通常情况下,中文句法依存关系的分析结果比英文要差,导致在中文环境下句法特征可靠性下降;而此时,n-gram 特征起到了很好的补充作用. 由此可见,n-gram 特征对于多语言关系抽取是非常重要的.

5.2 对 n-gram 的选择

我们共设计了 3 种不同的 n-gram 特征,基于词序列的 lexical n-gram(lng)、基于词性标注序列的 pos n-gram(png)和两者结合的 lexical-pos n-gram (lpng). 在英文和中文数据集上,我们分别测试了这 3 种 n-gram 特征的效果. 这里所使用的分类算法为最大熵模型,所得的 P-R 曲线如图 5 和图 6 所示:

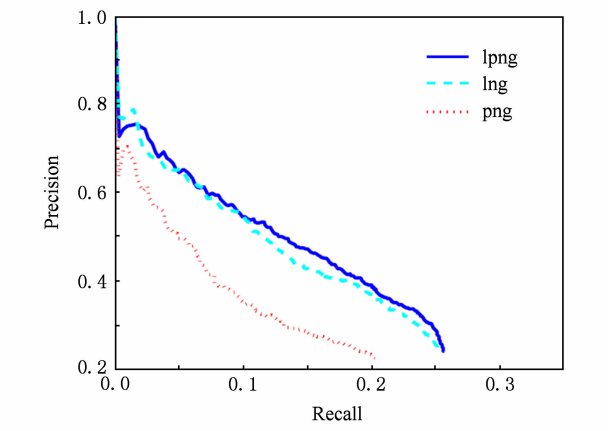

Fig. 5 Comparison of three kinds of n-gram features on the English dataset.

图 5 英文数据集上使用 3 种不同 n-gram 的效果比较

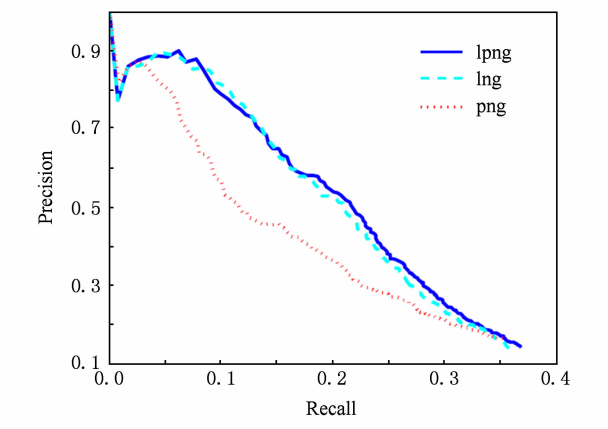

Fig. 6 Comparison of three kinds of n-gram features on the Chinese dataset.

图 6 中文数据集上使用 3 种不同 n-gram 的效果比较

从图 5 和图 6 中我们可以看到,在中文和英文中,词与词性标注结合产生的 n-gram 效果都是最理想的. 因此,在对协同训练的效果进行实验时,我们使用的即为这种 n-gram. 我们还发现,词性的 n-gram 并没有带来召回率的提升,这可能是因为这种 n-gram 导致大量实体对之间被预测为没有关系(NA),因而召回率不但没有提升,甚至还可能有所下降.

5.3 协同预测的策略

我们为协同训练的预测设计了两种不同的策略,置信度求和的策略(sum)和投票策略(vote). 我们在两种语言的数据集上都进行了实验,以测试哪个策略更好. 需要重点说明的是,我们的投票策略在两个模型输出的答案不一致时,将会认为两个都是正确的,我们认为这样可以借助两个不同视图下的分类器,一定程度上解决为一个实体对预测多个关系的问题. 为了验证这个想法是否正确,我们设计了一个基准策略(VoteSingle),在投票策略的基础上加了以下变化:如果两个模型预测结果不一致并且都不是 NA(没有关系),只选取置信度高的一方作为最终答案. 这 3 个策略的结果如图 7 和图 8 所示.

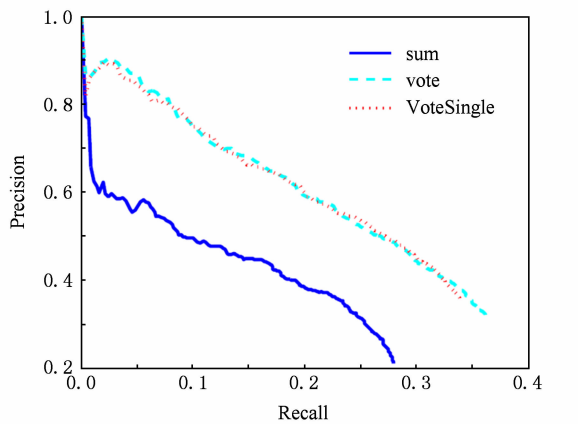

Fig. 7 Comparison of three different co-training strategies on the English dataset.

图 7 英文数据集上 3 种不同协同策略的比较

在图 7 和图 8 中我们可以看到,投票策略的效果远好于置信度求和的策略. 我们认为出现这样的现象的原因是置信度求和的策略很容易受到噪声的影响. 比如,虽然某个错误关系在两个模型中置信度都不是最高的,但它的置信度的和却是最高的,这种策略就会把这样的关系挑出来. 我们还发现,在中文上,我们原始的投票策略的比起只输出一个结果的投票策略具有一定优势;在英文上,虽然两者多数时间结果接近,但原始策略的曲线却延伸得更长一些,

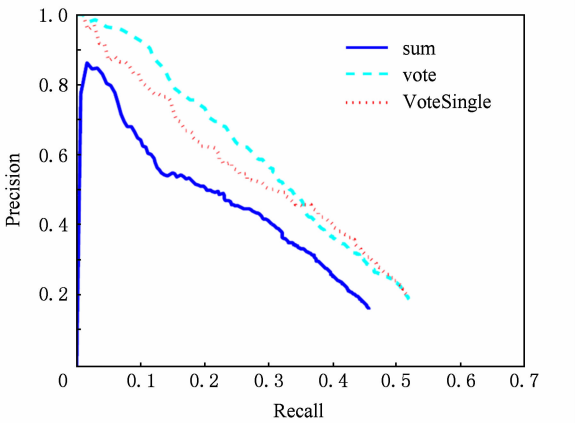

Fig. 8 Comparison of three different co-training strategies on the Chinese dataset.

图 8 中文数据集上 3 种不同协同策略的比较

即最终召回率更高一些. 因此,从整体上来说,我们的原始策略是更占优一些的. 这也说明了协同训练确实可以在一定程度上解决多标签分类的问题.

5.4 协同训练的参数

第 3.2 节提到,协同训练共有 3 个参数,在本节中,我们将观察这 3 个参数对于算法性能所带来的影响.

我们从{1 000,2 000,3 000,4 000,5 000}中选取数据池大小 p,从{20,40,60,80,100}中选取每次迭代加入到标注集中的数据数量 n,迭代次数从 1~50. 我们使用 P-R 曲线上 F-measure 的最高值来表示一种参数设置下的结果. F-measure 是精确率和召回率的调和平均数. 表 2 和表 3 列出了两个数据集上每组 p 和 n 的组合下最好的结果,F-measure 后面的括号内是达到该结果时的迭代次数.

我们发现,在多数情况下,协同训练的效果先随迭代次数上升,一般在迭代 1~2 次后达到最好,之后效果会随着迭代一直下降. 发生这样的现象原因也是比较直观的:当迭代次数较少的时候,每次加入到标注数据集中的数据比较可靠,因此更可能对抽取模型的训练起到正向的作用. 随着迭代次数增加,加入到标注集中的数据可靠性越来越差(McClosky 等人[35]在句法分析任务上也有类似的发现),对模型的训练也开始起负面的作用. 由于弱监督学习的训练数据本身质量就不高,因此多数情况下,迭代 1~2 次后再加入到标注集中的数据的可靠性就下降了,从而导致关系抽取结果的下降. 数据池的大小 p 和每次迭代扩充到标注数据数量 n 对协同训练的效果并没有呈现出规律性的影响. 在我们的实验中,通过留存估计,我们对英文数据集选取的参数是 $p=3\,000$,$n=20$,$t=1$;对中文数据集选取的参数是 $p=2\,000$,$n=20$,$t=2$.

Table 2 Co-Training Performance vs. Parameters on the English Dataset

表 2 英文数据集上协同训练效果随参数的变化

p	$n=20$	$n=40$	$n=60$	$n=80$	$n=100$
1 000	0.368 1(1)	0.367 1(3)	0.369 4(9)	0.361 2(1)	0.364 1(4)
2 000	0.366 2(1)	0.368 2(2)	0.363 3(1)	0.368 8(1)	0.365 8(1)
3 000	0.369 2(1)	0.369 7(1)	0.368 8(2)	0.368 6(1)	0.366 9(3)
4 000	0.363 4(4)	0.368 5(1)	0.368 6(2)	0.365 6(1)	0.367 3(7)
5 000	0.368 3(9)	0.369 1(1)	0.364 7(2)	0.367 5(1)	0.365 3(1)

Table 3 Co-Training Performance vs. Parameters on the Chinese Dataset

表 3 中文数据集上协同训练效果随参数的变化

p	$n=20$	$n=40$	$n=60$	$n=80$	$n=100$
1 000	0.409 4(9)	0.416 5(1)	0.415 2(1)	0.403 6(1)	0.404 3(1)
2 000	0.425 5(2)	0.417 1(5)	0.413 2(1)	0.408 9(2)	0.402 8(1)
3 000	0.412 8(8)	0.415 2(2)	0.405 8(4)	0.411 4(2)	0.416 9(2)
4 000	0.417 9(2)	0.410 0(2)	0.411 1(1)	0.406 8(1)	0.405 0(1)
5 000	0.422 7(5)	0.420 3(2)	0.410 6(2)	0.404 8(1)	0.404 5(1)

6 结束语

在本文中,我们首先针对弱监督关系抽取的特征过于具体并且依赖于句法分析工具的缺陷,引入了可以提高多语言关系抽取模型鲁棒性的 n-gram 特征,在此基础上,我们针对弱监督关系抽取的训练数据含有较多噪声和错误的问题,提出使用协同训练的方法来向训练集中引入比较可靠并且多样的标注数据,从而强化关系抽取模型,提高关系抽取的性能. 我们重点研究了协同预测的策略,并且提出了一个可以一定程度上解决同一实体对含有多个关系的问题策略. 我们在中文和英文的数据集上进行了实验,验证了 n-gram 和协同训练的结合确实可以提高弱监督关系抽取的效果,并且使得它在多语言关系抽取中具有更强的鲁棒性.

在后续的工作中,我们将继续研究协同训练的协同策略,并且把弱监督关系抽取应用在面对网络上海量数据的关系抽取任务中,将抽取出的关系构建成一个结构化的知识库.

参 考 文 献

[1] Sundheim B, Chinchor N. Survey of the message understanding conferences [C] //Proc of HLT'93. Stroudsburg, PA: ACL, 1993: 56-60

[2] Banko M, Cafarella M, Soderland S, et al. Open information extraction from the Web [C] //Proc of IJCAI 2007. New York: ACM, 2007: 2670-2676

[3] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction [C] //Proc of EMNLP 2011. Stroudsburg, PA: ACL, 2011: 1535-1545

[4] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning [C] //Proc of AAAI 2010. Palo Alto, CA: AAAI, 2010: 1306-1313

[5] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources [C] //Proc of the 7th Int Conf on Intelligent Systems for Molecular Biology. Palo Alto, CA: AAAI, 1999: 77-86

[6] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training [C] //Proc of ICML 1998. New York: ACM, 1998: 92-100

[7] Che Wanxiang, Liu Ting, Li Sheng. Automatic entity relation extraction [J]. Journal of Chinese Information Processing, 2005, 19(2): 1-6 (in Chinese)
(车万翔, 刘挺, 李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2): 1-6)

[8] Liu Kebin, Li Fang, Liu Lei, et al. Implementation of a kernel-based Chinese relation extraction system [J]. Journal of Computer Research and Development, 2007, 44(8): 1406-1411 (in Chinese)
(刘克彬, 李芳, 刘磊, 等. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展, 2007, 44(8): 1406-1411)

[9] Dong Jing, Sun Le, Feng Yuanyong, et al. Chinese automatic entity relation extraction [J]. Journal of Chinese Information Processing, 2007, 21(4): 80-85, 91 (in Chinese)
(董静, 孙乐, 冯元勇, 等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007, 21(4): 80-85, 91)

[10] Wu Fei, Hoffmann R, Weld D. Information extraction from Wikipedia: Moving down the long tail [C] //Proc of ACM SIGKDD 2008. New York: ACM, 2008: 731-739

[11] Wu Fei, Weld D. Autonomously semantifying wikipedia [C] //Proc of CIKM 2007. New York: ACM, 2007: 41-50

[12] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data [C] //Proc of ACL 2009. Stroudsburg, PA: ACL, 2009: 1003-1011

[13] Yao Limin, Riedel S, McCallum A. Collective cross-document relation extraction without labeled data [C] //Proc of EMNLP 2010. Stroudsburg, PA: ACL, 2010: 1013-1023

[14] Riedel S, Yao Limin, McCallum A. Modeling relations and their mentions without labeled text [J]. Machine Learning and Knowledge Discovery in Databases, 2010, 6323: 148-163

[15] Surdeanu M, McClosky D, Tibshirani J, et al. A simple distant supervision approach for the TAC-KBP slot filling task [C] //Proc of the TAC-KBP 2010 Workshop. 2010: 1-5

[16] Hoffmann R, Zhang Congle, Ling Xiao, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C] //Proc of ACL-HLT 2011. Stroudsburg, PA: ACL, 2011: 541-550

[17] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C] //Proc of EMNLP 2012. Stroudsburg, PA: ACL, 2012: 455-465

[18] Marino J, Banchs R, Crego J, et al. N-gram-based machine rranslation [J]. Journal of Computational Linguistics, 2006, 32(4): 527-549

[19] Furnkranz J. A study using *n*-gram features for text categorization [R]. Austrian Research Institute for Artificial Intelligence. Wien, AT: Austrian Research Institute for Artificial Intelligence, 1998

[20] Sarkar A. Applying co-training methods to statistical parsing [C] //Proc of NAACL 2001. Stroudsburg, PA: ACL, 2001: 95-102

[21] Hwa R, Osborne M, Sarkar A, et al. Corrected co-training for statistical parsers [C] //Proc of the Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, ICML 2003. New York: ACM, 2003: 95-102

[22] Steedman M, Osborne M, Sarkar A, et al. Bootstrapping statistical parsers from small data sets [C] //Proc of EACL 2003. Stroudsburg, PA: ACL, 2003: 331-338

[23] Mihalcea R. Co-training and self-training for word sense disambiguation [C] //Proc of CoNLL 2004. Stroudsburg, PA: ACL, 2004: 33-40

[24] Zhou Zhihua, Chen Kejia, Jiang Yuan. Exploiting unlabeled data in content-based image retrieval [G] //LNCS 3201: Proc of ECML2004. Berlin: Springer, 2004: 525-536

[25] Zhou Zhihua, Chen Kejia, Dai Hongbin. Enhancing relevance feedback in image retrieval using unlabeled data [J]. ACM Trans on Information Systems, 2006, 24(2): 219-244

[26] Abney S. Bootstrapping [C] //Proc of ACL 2002. Stroudsburg, PA: ACL, 2002: 360-367

[27] Balcan M, Blum A, Yang Ke. Co-training and expansion: Towards bridging theory and practice [C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2005: 89-96

[28] Wang Jiao, Luo Siwei, Zeng Xianhua. A random subspace method for co-training [J]. Acta Electronica Sinica, 2008, 36(12A): 60-65 (in Chinese)
(王娇, 罗四维, 曾宪华. 基于随机子空间的半监督协同训练算法[J]. 电子学报, 2008, 36(12A): 60-65)

[29] Wang Wei, Zhou Zhihua. A new analysis of co-training [C] //Proc of ICML 2010. New York: ACM, 2010: 1135-1142

[30] Goldman S, Zhou Yan. Enhancing supervised learning with unlabeled data [C] //Proc of ICML 2000. New York: ACM, 2000: 327-334

[31] Zhou Zhihua, Li Ming. Tri-training: Exploiting unlabeled data using three classifiers [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[32] Berger A, Pietra V, Pietra S. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996, 22(1): 39-71

[33] Darroch J, Ratcliff D. Generalized iterative scaling for log-linear models [J]. The Annals of Mathematical Statistics, 1972, 43(5): 1470-1480

[34] Pietra S, Pietra V, Lafferty J. Inducing features of random fields [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(4): 380-393

[35] McClosky D, Charniak E, Johnson M. Effective self-training for parsing [C] //Proc of HLT-NAACL 2006. Stroudsburg, PA: ACL, 2006: 152-159

[36] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia—A crystallization point for the Web of data [J]. Web Semantics, 2009, 7(3): 154-165

[37] de Marneffe M, MacCartney B, Manning C. Generating typed dependency parses from phrase structure parses [C] // Proc of LREC 2006. Paris: ELRA/ELDA, 2006: 449-454

[38] Finkel J, Grenager T, Manning C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling [C] //Proc of ACL 2005. Stroudsburg, PA: ACL, 2005: 363-370

[39] Bohnet B. Top accuracy and fast dependency parsing is not a contradiction [C] //Proc of CoLing 2010. Stroudsburg, PA: ACL, 2010: 89-97

Chen Liwei, born in 1986. PhD candidate in the Institute of Computer Science and Technology, Peking University. His current research interests include natural language processing, information extraction.

Feng Yansong, born in 1981. PhD. Lecturer in the Institute of Computer Science and Technology, Peking University. His current research interests include natural language processing, machine learning.

Zhao Dongyan, born in 1969. PhD. Professor in the Institute of Computer Science and Technology, Peking University. Senior member of China Computer Federation. His research interests are text mining, semantic Web, graph data management and digital publishing(zhaodongyan@pku.edu.cn).