术语关系自动抽取方法研究

孙 霞 王小凤 董乐红 吴 江
(西北大学信息技术与科技学院计算机系 西安 710127)

摘 要 将术语关系抽取转化为分类问题,给出了基于机器学习的术语关系自动抽取流程。针对现有产生式和判定学习算法的缺点,提出了混合分类算法 HC。该算法使得一部分特征值通过训练数据估计而来,另一部分特征值通过判定函数训练得到。实验结果表明,该算法优于原来的产生式学习算法和判断学习算法,在人工标注的小训练集上获得了较好的分类效果。

关键词 机器学习,术语关系抽取,混合学习算法
中图法分类号 TP391 **文献标识码** A

Study on Term Relation Extraction from Domain Text

SUN Xia WANG Xiao-feng DONG Le-hong WU Jiang
(Department of Computer Science and Technology, Northwest University, Xi'an 710127, China)

Abstract A term relation extraction approach was proposed. It was cast as a classification task. The hybrid classification algorithm combining the advantages of both naïve bayes and perceptron was also presented. In this algorithm, a subset of the features was estimated from training data, and another subset of the features was trained by discriminative function. The experimental results showed that the proposed hybrid algorithm almost always outperforms the naïve bayes algorithms and perceptron algorithms when the training set is small.

Keywords Machine learning, Term relation extraction, Classification algorithm

术语关系抽取是指从一定规模的语料中抽取出能反映某一领域文本特征的两两词语间的语义关系(如同义关系、上下位关系等)。如"传输层是在OSI七层协议之中的第四层,有时又称为传送层",该句中,计算机网络术语"传输层"和"传送层"之间存在同义关系。术语语义关系集中体现和负载了一个学科领域的核心知识,对了解和把握一个学科领域的发展现状、未来趋向等具有重要的理论和现实意义[1]。另一方面,术语关系可以广泛应用到机器翻译、本体构建等领域,为面向领域的智能系统提供知识服务。然而,单纯靠语言学专家从大规模领域文本中手工抽取术语关系费时费力,很难形成规模。因此,开发一种自动化的方法来辅助术语关系抽取显得尤为必要和迫切。

1 相关工作

目前,国内外学者就关系抽取方法做了相关研究,总体上可以分为3种:人工获取方法、模板匹配方法和机器学习方法。人工获取方法需要大量专家参与,费时费力,效率低[2,3]。模板匹配的方法,对模板表示的依赖性非常强,并且要求模板是无歧义的,抽取效果明显受到模板规模和质量的制约[4,5]。机器学习是目前主流的关系自动抽取方法,该方法将关系抽取转化为分类问题,通过构造候选关系,利用机器学习得到的分类器来标注这些候选关系属于哪一类预定义关系[6,7]。如Girju和Badulescu[8]借助WordNet训练了一个决策树分类器,获得了83%的准确率和72%的召回率。Fleischman和Hovy[9]研究了上下位关系的获取。Jeffrey[10]等人在生物医学领域做了些尝试,将出现在基因和蛋白质的邻居词作为特征,训练分类器,获取基因(gene)和蛋白质(protein)之间的关系。

上述研究中,有的学者采用了产生式分类算法,有的采用了判定分类算法。而无论是产生式分类算法还是判定分类算法,都有局限性。产生式分类算法是估计输入 x 和输出 c 的一个联合概率分布 $P(x,c)$,因此需要已知训练数据的分布形式。而实际情况往往是无法得知训练数据的真正分布函数,所以产生式分类算法被认为精度不高。判定分类算法由于不需要假设训练数据的分布,也不需要对产生数据的隐含过程做假设,只需要寻找一个判定函数,使得该函数能最大限度地分开数据。但是,判定分类算法需要足够的训练数据,只有在大量训练数据的情况下,判定分类算法才优于产生式分类算法。而人工标注数据是一件很费时费力的工作,尤其是针对汉语数据的标注工作,可利用的标注数据是有限的。

基于上述考虑,本文结合产生式和判定两种学习算法的特点,提出了一种混合分类算法(Hybrid Classification algorithm, HC)。算法的一部分特征值是通过训练数据估计而来,另一部分特征值是通过判定函数训练得到。实验表明,在少量训练数据的情况下,本文提出的混合学习算法获得了较好的抽取结果。本文第2节介绍术语关系自动抽取流程;第

到稿日期:2009-03-05 返修日期:2009-05-25 本文受陕西省教育厅项目(09JK768,09JK774,09JK738)资助。
孙 霞(1977—),女,博士,讲师,主要研究领域为信息获取技术,E-mail:raindy@nwu.edu.cn;王小凤(1979—),博士,讲师,主要研究领域为中文信息处理;董乐红(1971—),女,博士,讲师,主要研究领域为文本分类;吴 江(1962—),男,博士,教授,主要研究领域为知识工程。

3节重点介绍本文提出的混合分类算法HC;最后给出实验结果和分析。

2 术语关系自动抽取流程

2.1 基本思路

术语关系抽取基本思想是先构造候选关系集,然后使用混合分类学习算法构造分类器,对候选关系样例分类标注。符合预定义关系类别的候选将被抽取出来自动标注为该类关系。

术语关系抽取具体处理流程如图1所示。

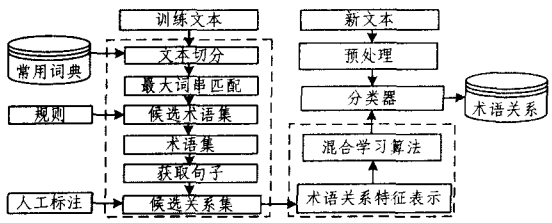

图1 术语关系自动抽取流程

主要功能模块说明如下。

1)文本切分。利用常用词典对面向领域的文本进行分词处理,采用确定性有限状态自动机识别语料中出现的时间词和数词,得到切分文档集。

2)最大词串匹配。由于常用词典不包含术语,分词时会把术语拆分成连续的若干个切分单元,例如术语"反转芯片针脚栅格阵列"经过分词处理后变为:"反转/芯片/针/脚/栅/格/阵列"。采用最大词串匹配算法(具体实现过程请见作者前期研究工作——文献[11]),尽可能合并被拆分的术语,得到候选术语集。最后利用启发式规则对候选术语集进行优化,过滤掉错误组合模式,生成术语集。

3)候选关系集。依据标点符号,抽取含有两个或两个以上术语的句子,然后枚举句子中所有可能的术语词,两两组对形成候选关系术语词对。

4)术语关系特征表示。在获得候选术语关系集后,需要将其表示成计算机可以识别的格式,以便训练分类模型,得到分类器。本文选用3类特征组成的特征向量表示每一个候选词对。这3类特征分别是词对顺序(Order)、词形特征(Term appearance)和上下文特征(Context information)(详见2.2节)。

5)混合学习算法。采用本文提出的基于产生式和判断学习相结合的混合学习算法对所有候选关系实例进行分类判别,挑选出包含预定义关系的关系实例并对其类型进行标注,获得术语关系集。

2.2 术语关系特征表示

1) 词序特征(Order)

词序是指两个具有语义关系的词在句子中出现的先后顺序,是一个二元值。0表示第一个考察词(以下简称词1)在第二个考察词(以下简称词2)的左边。1表示词1在词2的右边。

2) 词形特征(Term appearance)

词形特征主要描述术语构成的一些属性,包括:1)是否全为中文(CHIN);2)是否全为英文(ENG);3)是否为英文大小写、中英文数字混合(MIX)。这3个属性都是二元值。特征的值为1时,表示具有该属性,0表示不具有该属性。

3) 上下文特征(Context information)

本文考虑以句子为单位的上下文文本,采用词和符号进行标引。根据以往经验,词标引是非常有效的标引方式之一。除了词标引外,一些标点符号极大地暗示了句子中的两个词可能具有某种语义关系。如出现在同义解释的符号"()"、"——"、":"等前后的两个词很可能是同义关系。很多关系获取方法都没有考虑占有重要地位的符号特征。

根据词和符号在句子中所处的位置,分为左信息(Left)、中间信息(Middle)和右信息(Right)。左信息是指同时包含待考察的两个术语的句子中出现在第一个术语词左边的词和符号。类似地,中间信息定义为同时包含待考察的两个术语词的句子中两个术语词之间的词和符号。右信息是指同时包含待考察的两个术语词的句子中出现在第二个术语词右边的词和符号。出现在左信息、中间信息和右信息中的词和符号又进一步区分为相邻词汇、相邻符号以及非相邻词汇和非相邻符号,如表1所列。

表1 上下文特征

No.	类型	特征
		左信息(Left)
1	Float	与词1左相邻词汇,记为 LNW_1
2	Float	与词1左相邻符号,记为 LNS_1
3	Float	出现在词1左边词汇,记为 LW_1
4	Float	出现在词1左边符号,记为 LS_1
		中间信息(Middle)
5	Float	与词1右相邻词汇,记为 MNW_1
6	Float	与词1右相邻符号,记为 MNS_1
7	Float	出现在词1和词2中间词汇,记为 MW
8	Float	出现在词1和词2中间符号,记为 MS
9	Float	与词2左相邻词汇,记为 MNW_2
10	Float	与词2左相邻符号,记为 MNP_2
		右信息(Right)
11	Float	与词2右相邻词汇,记为 RNW_2
12	Float	与词2右相邻符号,记为 RNS_2
13	Float	出现在词2右边词汇,记为 RW_2
14	Float	出现在词2右边符号,记为 RS_2

图2给出了上下文信息示意图。

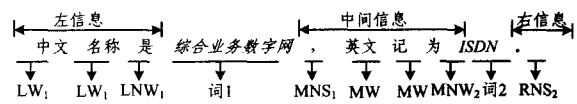

图2 上下文信息示意图

本文借鉴向量空间模型的文本表示方法,构造术语关系的特征向量。向量空间模型由于不考虑特征在文本中的位置信息,因此被认为是一个词袋(A Bag of Words)模型。本文提出的上下文特征,考虑了两个待观察术语之间存在的长距离依赖特性,将上下文特征进一步分为左信息、中间信息和右信息。也正是由于把上下文特征划分为左信息、中间信息和右信息,因此为向量空间模型增加了特征在文本中的位置信息,提高了特征表示的能力。

3 混合分类学习算法

首先考虑简单但表现良好的产生式学习算法——朴素贝叶斯分类算法(Naïve Bayes,简称NB)。朴素贝叶斯分类算法分类原理是:从给定类别的训练样例中估计类的概率 $P(c_j)$ 和类的条件概率 $P(t_i|c_j)$,然后计算新样例属于类集合 C 中每一个类 c_i 的概率,比较概率值,概率值最大的类为新样例的类别。这里只考虑两类的情况,即 $C=\{c_1,c_2\}$。

对于新样例 $\vec{x}=(t_1,t_2,\cdots,t_n)$,为了判断所属类别,分别估计该样例属于 c_1 类和 c_2 类的概率值。

$$A=P(c_1|t_1,t_2,\cdots,t_n)=P(t_1,t_2,\cdots,t_n|c_1)P(c_1) \quad (1)$$

假定给定目标值时特征之间相互条件独立，则有

$$A=P(t_1,t_2,\cdots,t_n|c_1)P(c_1)=P(c_1)\prod_i P(t_i|c_1) \quad (2)$$

同理

$$B=P(c_2|t_1,t_2,\cdots,t_n)=P(t_1,t_2,\cdots,t_n|c_2)P(c_2)$$
$$=P(c_2)\prod_i P(t_i|c_2) \quad (3)$$

当 $A\geqslant B$ 时，新样例 $\vec{x}$ 属于 c_1 类，反之属于 c_2 类。由于选择 3 类特征：词序、词形和上下文特征，其中上下文特征又被分为 14 个子区域（如表 1 所列），因此最终将特征划分为 16 个区域。t^1 表示词序特征，t^2 表示词形特征，$t^3,\cdots,t^{16}$ 分别表示上下文特征。于是，式(2)和式(3)改写为

$$A=P(c_1|t_1,t_2,\cdots,t_n)$$
$$=P(c_1)\times\prod_{i=1}^{n_1}P(t_i^1|c_1)\times\prod_{i=1}^{n_2}P(t_i^2|c_1)\times\cdots\times\prod_{i=1}^{n_{16}}P(t_i^{16}|c_1) \quad (4)$$

$$B=P(c_2|t_1,t_2,\cdots,t_n)$$
$$=P(c_2)\times\prod_{i=1}^{n_1}P(t_i^1|c_2)\times\prod_{i=1}^{n_2}P(t_i^2|c_2)\times\cdots\times\prod_{i=1}^{n_{16}}P(t_i^{16}|c_2) \quad (5)$$

其中，$t^j=t_1,\cdots,t_{n_j}$；n_1 和 n_2 分别表示词序特征和词形特征的个数；$n_3,n_4,\cdots,n_{16}$ 分别表示上下文特征中 14 个子特征的个数。将式(4)和式(5)两边取对数后得

$$\log A=\log P(c_1)+\sum_{i=1}^{n_1}\log P(t_i^1|c_1)+\sum_{i=1}^{n_2}\log P(t_i^2|c_1)+\cdots+\sum_{i=1}^{n_3}\log P(t_i^{16}|c_1) \quad (6)$$

$$\log B=\log P(c_2)+\sum_{i=1}^{n_1}\log P(t_i^1|c_2)+\sum_{i=1}^{n_2}\log P(t_i^2|c_2)+\cdots+\sum_{i=1}^{n_3}\log P(t_i^{16}|c_2) \quad (7)$$

式(6)和式(7)中，给 $t^1,t^2,\cdots,t^{16}$ 分别赋予不同的权重，则有

$$\log A'=\log P(c_1)+\frac{k_1}{n_1}\sum_{i=1}^{n_1}\log P(t_i^1|c_1)+\frac{k_2}{n_2}\sum_{i=1}^{n_2}\log P(t_i^2|c_1)+\cdots+\frac{k_{16}}{n_{16}}\sum_{i=1}^{n_3}\log P(t_i^{16}|c_1) \quad (8)$$

$$\log B'=\log P(c_2)+\frac{k_1}{n_1}\sum_{i=1}^{n_1}\log P(t_i^1|c_2)+\frac{k_2}{n_2}\sum_{i=1}^{n_2}\log P(t_i^2|c_2)+\cdots+\frac{k_{16}}{n_{16}}\sum_{i=1}^{n_3}\log P(t_i^{16}|c_2) \quad (9)$$

当 $\log A'-\log B'\geqslant 0$ 时，新样例 $\vec{x}$ 属于 c_1 类，反之属于 c_2 类。由式(8)和式(9)得

$$\log A'-\log B'=k_0+k_1\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\log P(t_i^1|c_1)-\frac{1}{n_1}\sum_{i=1}^{n_1}\log P(t_i^1|c_2)\right]+\cdots+k_{16}\left[\frac{1}{n_{16}}\sum_{i=1}^{n_{16}}\log P(t_i^{16}|c_1)-\frac{1}{n_{16}}\sum_{i=1}^{n_{16}}\log P(t_i^{16}|c_2)\right] \quad (10)$$

其中，$k_0=\log P(c_1)-\log P(c_2)$。式(10)可以看成判定分类算法感知器(Perceptron)的判别函数，即 $f(x)=\log A'-\log B'$。当 $f(x)\geqslant 0$ 时，新样例 $\vec{x}$ 属于 c_1 类。反之，新样例 $\vec{x}$ 属于 c_2 类。至此，得到了一个混合分类算法。

4 实验

实验语料来自 Web 上选取的计算机网络领域文档。将获得的约 20M 文本格式的科技语料按 4∶1 的比例随机划分为两部分，一部分用于训练，另一部分用于测试。经预处理，得到 63,498 个计算机网络方面的 3 种候选术语关系（同义关系、上下位关系、整体-部分关系），其中 4,886 个正例、58,612 个负例。在测试文本上进行相同的预处理过程，得到 12,292 个计算机网络方面的 3 种候选术语关系，其中 1,062 个正例、11,230 个负例。

4.1 实验步骤

Step1 样例特征化

从术语关系候选集中抽取 3 类特征：词序特征、词形特征、上下文特征。为候选集合中的每一个词对构造一个带类别标记的特征向量 $\vec{F}_i'$，其格式为

$\vec{F}_i'::=\langle$Feature:Weight,Class$\rangle$

Feature::=⟨Order, Term appearance, Context information⟩

Class::={Syn|~Syn}

Class 的含义就是“是同义关系词对”“不是同义关系词对”，分别记为“Syn”和“~Syn”。

Step2 估计类概率和类的条件概率

从训练文本中估计类概率和类的条件概率。类概率的估计值计算公式为

$$P(c_1=\text{Syn})=\frac{\text{正例所在句子个数}}{\text{所有样例所在句子总数}} \quad (11)$$

$$P(c_2=\sim\text{Syn})=\frac{\text{负例所在句子个数}}{\text{所有样例所在句子总数}} \quad (12)$$

类的条件概率估计值计算公式为

$$P(t_i^j|c_1)=\frac{m_{1i}+1}{m_1+s_i} \quad (13)$$

$$P(t_i^j|c_2)=\frac{m_{2i}+1}{m_2+s_i} \quad (14)$$

其中，t_i 为 Step1 获得的特征向量中某一个特征项，$j=\{1,2,\cdots,16\}$。为了避免某个特征在句子中出现的次数太少，采用 m-估计方法。m_{1i} 为特征 t_i 出现在正例所在句子中的次数。m_{2i} 为特征 t_i 出现在负例所在句子中的次数，m_1 是所有正例总数，m_2 是所有负例总数，s_i 为特征 t_i 可能取值的个数。

Step3 训练感知器参数

由式(11)－式(14)可分别获得式(10)中 $k_0=\log P(c_1)-\log P(c_2)$ 值和 $\frac{1}{n_j}\sum_{i=1}^{n_j}\log P(t_i^j|c_1)-\frac{1}{n_j}\sum_{i=1}^{n_j}\log P(t_i^j|c_2)$ 值，$j=\{1,2,\cdots,16\}$。然后，用 k_0 值和 $\frac{1}{n_j}\sum_{i=1}^{n_j}\log P(t_i^j|c_1)-\frac{1}{n_j}\sum_{i=1}^{n_j}\log P(t_i^j|c_2)$ 值重新改写每一个候选词对的特征向量 $\vec{F}_i'$，即为

$\vec{F}_i'::=\langle$Feature:Weight,Class$\rangle$

Feature::=$\langle t^1,t^2,t^3,\cdots,t^{16}\rangle$

$$\text{Weight}::=\langle\frac{1}{n_1}\left(\sum_{i=1}^{n_1}\log P(t_i^1|c_1)-\sum_{i=1}^{n_1}\log P(t_i^1|c_2)\right),\frac{1}{n_2}\left(\sum_{i=1}^{n_2}\log P(t_i^2|c_1)-\sum_{i=1}^{n_2}\log P(t_i^2|c_2)\right),\cdots,\frac{1}{n_{16}}\left(\sum_{i=1}^{n_{16}}\log P(t_i^{16}|c_1)-\sum_{i=1}^{n_{16}}\log P(t_i^{16}|c_2)\right)\rangle$$

Class::={Syn|~Syn}

最后，用改写后的带类标记的特征向量训练感知器分类器。训练一个感知器分类器意味着选择 $k_1,\cdots,k_n$ 的值。为了得到可接受的常数向量 $\vec{k}=k_1,\cdots,k_n$，本文采用 delta 法则。关于 delta 法则请见相关文献。

Step4 分类新样例

对于新文本，为了获得该文本中含有的术语关系，首先获取该文本中的所有候选术语关系词对和词对所在的句子；然后将这些候选术语表示成没有类标记的特征向量；最后将特征向量输入 Step3 得到的混合分类器中，即可判断哪些候选

（下转第 215 页）

World Wide Web Conf. (WWW 2003). Budapest: ACM Press, 2003:123-134

[3] Xiong L, Liu L. PeerTrust: supporting reputation-based trust for p2p electronic communities[J]. IEEE Trans. Knowledge and Data Engineering, 2004, 16(7): 843-857

[4] Zhou R, Hwang K. PowerTrust: A robust and scalable reputation system for trusted P2P computing[J]. IEEE Trans. Parallel and Distributed Systems, 2007, 18(4): 460-473

[5] Song S S, Hwang K, Zhou R F, et al. Trusted p2p transactions with fuzzy reputation aggregation[J]. IEEE Internet Computing, 2005, 9(6): 24-34

[6] Gambetta D. Can we trust trust [M]. Reprinted in electronic edition from Department of Sociology, University of Oxford: Basil Blackwell, 1988

[7] Jøsang A, Gray L, Kinateder M. Simplification and analysis of transitive trust networks[J]. Web Intelligence and Agent Systems Journal, 2006, 4(2): 139-161

[8] Wang Yonghong, Singh M P. Trust representation and aggregation in a distributed agent system[C]// Proceedings of the 21st National Conference on Artificial Intelligence (AAAI). Boston, MA, USA, 2006

[9] Dong-Huynh T, Jennings N R, Shadbolt N R. Fire: An integrated trust and reputation model for openmulti-agent systems[J]. Journal of Autonomous Agents and Multi-Agent Systems, 2006, 13(2): 119-154

[10] Zhou R, Hwang K, Cai M. Gossip Trust for fast reputation aggregation in peer-to-peer networks[J]. IEEE TKDE, 2008, 20(9): 1282-1295

[11] Zhuge Hai, Chen Xue, et al. Trust- based probabilistic search with the view model of p2p networks[J]. Concurrency and Computation: Practice and Experience, 2006, 18(14): 1839-1855

[12] Menasce D A, Kanchanaplli L. Probabilistic scalable p2p resource location services[J]. ACM SIGMETRICS Performance Evaluation Review, 2002(20): 48-58

[13] Wang Ping, Qiu Jing. A search algorithm based on referral trust in unstructured P2P systems[C]// Proceedings of the International Conference on ISECS. Nanchang, China, 2009: 644-649

[14] 陈学梁，万晓榆，樊自甫. 无线接入网中 P2P 对网络业务性能的评估[J]. 重庆工学院学报：自然科学版，2006，20(5)：92-96

(上接第 191 页)

关系词对具有同义关系。

4.2 实验结果和分析

图 3 给出了混合分类算法(HC)与贝叶斯分类算法(NB)和感知器分类算法(Perceptron)随训练集大小变化的曲线图。横坐标表示训练数据的大小，以千为单位。纵坐标是 F1 值。

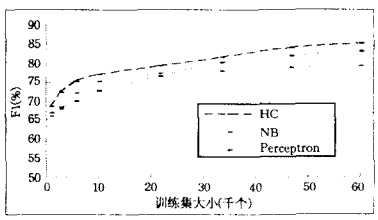

图 3 分类算法的比较结果

实验数据表明，本文提出的混合分类算法 HC 获得了较好的性能。NB 分类算法在训练数据量少的情况下，F1 值高于 Perceptron。但是随着训练数据的不断增加，Perceptron 分类算法的 F1 值高于 NB 分类算法。这是由于产生式分类算法需要已知训练数据的分布形式，然而实际情况往往无法得知训练数据的真正分布函数，因此 NB 的精度不高。但是，判定分类算法在足够的训练数据的前提下优于产生式分类算法。本文提出的混合分类算法，一部分特征值是通过训练数据估计而来，另一部分特征值是通过判定函数训练得到。因此，在训练数据较少的情况下 HC 曲线表现出和 NB 曲线一样快速上升的趋势。而在训练数据较多的情况下，NB 曲线基本保持不变，而 HC 曲线和 Perceptron 曲线仍呈上升趋势。

结束语 术语关系自动抽取逐渐成为当前热点研究方向，国内外学者做了许多有益尝试。其中，将关系抽取转化为分类问题，被认为是一种有效的抽取方法。然而，无论产生式分类算法还是判定分类算法，都有局限性。鉴此，本文提出了一种混合分类算法 HC。结合 Naïve Bayes 算法和 Perceptron 算法的优点，一部分特征值是通过训练数据估计而来，另一部分特征值是通过判定函数训练得到。实验表明，在相同大小的训练数据集下，本文提出的混合学习算法优于 Naïve Bayes 算法和 Perceptron 算法。

参 考 文 献

[1] Boguraev B, Kennedy C. Applications of term identification technology: domain description and content characterization[J]. Natural Language Engineering, 1999, 5(1): 17-44

[2] Appelt D E. Introduction to Information Extraction[J]. AI Communications, 1999, 12(3): 161-172

[3] Aone C, Ramos M, Rees S. A large-scale relation and event extraction system[C] // Proceedings of the 6th Applied Natural Language Processing Conference. New York: ACM Press, 2000: 76-83

[4] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora[C]// 14th International Conference on Computational Linguistics. Nantes, France, 1992: 539-545

[5] Yu Hong, et al. Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles[C]// Proceedings of the American Medical Informatics Association 2002 Symposium (AMIA'2002). 2002: 919-923

[6] 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

[7] Li Wenjie, et al. A Novel Feature-based Approach to Chinese Entity Relation Extraction[C]// Proceedings of ACL-08: HLT. Columbus, USA, 2008: 89-92

[8] Girju R, Badulescu A, Moldovan D. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations [C]// Edmonton, Canada. Proceedings of HLT-NAACL. Edmonton, Canada, 2003: 80-87

[9] Fleischman M, Hovy E. Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked[C]// Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan, 2003: 1-7

[10] Chang J T. Using Machine Learning to Extract Drug and Gene relationships from Text[D]. Stanford University, 2003

[11] 孙霞，郑庆华，王朝静，等. 一种基于生语料的领域词典生成方法[J]. 小型微型计算机系统，2005，26(6)：1088-1092