北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.049

面向中文专利 SAO 结构抽取的文本特征比较研究

饶齐　王裴岩　张桂平†

沈阳航空航天大学知识工程研究中心，沈阳　110136; † 通信作者, E-mail: zgp@ge-soft.com

摘要　针对中文专利文本中 SAO 结构实体关系抽取问题，使用支持向量机的机器学习方法进行关系抽取实验，分别对基本词法信息、实体间距离信息、最短路径闭包树句法信息以及词向量信息等特征的有效性进行验证分析。实验结果表明，基本的词法信息能够明显提高关系抽取性能，而句法信息没有显著提高关系抽取效果。此外，也验证了词向量在 SAO 结构关系抽取中的可行性。

关键词　SAO 结构；关系抽取；特征有效性；词向量

中图分类号　TP391

Text Feature Analysis on SAO Structure Extraction from Chinese Patent Literatures

RAO Qi, WANG Peiyan, ZHANG Guiping†

Knowledge Engineering Research Center, Shenyang Aerospace University, Shenyang 110136; † Corresponding author, E-mail: zgp@ge-soft.com

Abstract　To resolve the problem of SAO-based relation extraction from Chinese patent literatures, a series of experiments were implemented by using Support Vector Machines. It focused on the analysis of the validity of basic lexical information, syntactic information such as the shortest path enclosed trees, and distance features used in related works. The results show that simple lexical features can contribute to a good performance, while syntactic features cannot bring a remarkable improvement. Moreover, the feasibility of a new representation of words, word embeddings, is validated on SAO-based relation extraction.

Key words　SAO structure; relation extraction; effectiveness of features; word distributed representation

专利文献作为主要的技术承载体，对于科学技术创新有很高的参考价值。近年来，自然语言处理的相关技术也广泛应用于专利文献分析处理领域中。基于 SAO 结构的功能函数表示法[1]的专利定性分析技术在专利文献分析中应用较广，可以从发明的用途、原理、材料、结构和方法等方面解析专利的内容。SAO 结构抽取作为专利定性分析方法的基础，其抽取质量对于后续分析应用(如专利相似度计算[2]、专利侵权分析[3]等)有直接影响。

SAO 结构(Subject-Action-Object)，源自发明问题解决理论[4](theory of inventive problem solving, TIPS)，是表示问题解决方法的基本功能函数单元。其中主体 S 和客体 O 表示系统中的部件实体，通常由名词或名词性短语构成，行为 A 表示实体之间的操作或关系，一般由句子中的动词充当。

例句　所述数据源被设计为用于输出要发送的数据比特流。

以上例句中，“所述数据源”是主体 S，“要发送的数据比特流”是客体 O，“输出”表示主体与客体之间功能关系的 Action，抽取出的 SAO 结构用三元组的形式表示为“(所述数据源，输出，要发送的数据比特流)”。面向专利的 SAO 结构抽取，是从专利文本中抽取出(Subject, Action, Object)实体关系三元组，其中 Subject 对应于实体 1，Object 对应于实

国家“十二五”科技支撑计划项目(2012BAH14F00)资助
收稿日期: 2014–07–27; 修回日期: 2014–10–23; 网络出版时间: 2014-11-28 15:20

体 2, Action 对应于关系词。

在 SAO 结构三元组的抽取中, 实体 S 和实体 O 的抽取属于实体识别任务, 一般可以利用实体识别工具识别出, 技术相对成熟, 识别准确率高。对于表示实体 S 与实体 O 之间相互关系的指示词 A 的抽取, 则是主要难点, 与传统的关系抽取有所区别。传统的关系抽取任务中, 关系类别是预先指定的, 如雇佣关系、整体–部分关系等。SAO 结构中的关系指示词 Action 是从其所在句子中抽取, 与开放式关系抽取[5]任务很相似。与传统的关系抽取任务需要事先指定关系类型不同, 开放式关系抽取则不限定要抽取的关系的类型, 通常从实体对所在的句子中抽取关系值, 很大程度上扩展了抽取的关系的种类和数量[6]。开放式关系抽取的相关研究也是 SAO 结构关系抽取研究的重要参考。

1 相关研究

在中文实体关系抽取研究中, 多将关系抽取任务转化为分类问题, 采用机器学习的方法来解决。那么, 机器学习算法与特征的选择是决定抽取效果的关键问题。在机器学习算法方面, 典型的有最大熵模型(MaxEnt)[7]和支持向量机(SVM)[8]。在特征方面, 重点则在于如何选择各种有效的词法、语法、语义等特征, 并有效地集成, 从而产生描述实体对象关系的各种局部和简单的全局特征[9]。在传统的实体关系抽取方面, 黄鑫等[10]在 ACE2005 中文语料上进行关系抽取实验, 对基本的词法、句法和语法特征进行组合, 使用 SVM 进行关系探测和关系分类, F 值分别达到 72.77%和 61.03%。黄晨等[11]在 ACE RDC2005 中文语料库上以最短路径包含树为关系实例的结构化表现形式, 使用卷积树核进行无指导关系抽取, F 值最高达 60.1%。在开放式关系抽取中, 在英文方面代表性的有 REVERB[12]和 OLLIE[13]等, 其中 OLLIE 在抽取中加入句法分析信息, 在以动词为核心的关系抽取中召回率达 71%。在中文方面开放式关系抽取方面, 研究相对较少, 赵奇猛等[14]在组块层次标注基础上, 应用马尔科夫逻辑网络进行中文专利领域文本的开放式实体关系抽取, F 值达 77.92%。

在中文专利实体关系抽取中, 很难再通过发现新的特征来提高关系抽取性能, 而通常使用的构成特征向量的各类特征并非全部有效。本文主要从特征分析的角度入手, 对基本的词法信息、句法信息和距离信息等特征在基于 SVM 的中文专利 SAO 结构关系抽取方法中的有效性进行系统地比较和分析验证。此外, 本文还对词向量在关系抽取中的应用进行尝试。

2 中文专利 SAO 结构抽取

2.1 SVM

SVM 是一种二类分类模型, 学习的目的是基于结构风险最小化原则, 在特征空间中, 利用间隔最大化, 求解能将样本数据正确分类的最优分类超平面。对于给定的训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$, 其中, $x_i\in R^d$ 和 $y_i\in\{-1,+1\}$ 是样本 x_i 对应的分类标记, N 是训练样本个数。最优分类超平面可以通过求解式(1)中的凸二次规划问题得到:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i \quad \text{s.t.}$$

$$\sum_{i=1}^{N}\alpha_i y_i=0,\ 0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,N \quad (1)$$

其中, 变量 α_i 为每个样本点对应的拉格朗日乘子, $K(x_i,x_j)$ 为核函数, 参数 C 为惩罚参数。分类决策函数如下所示:

$$f(x)=\text{sign}\left(\sum_{x_i\in SV}\alpha_i^* y_i K(x_i,x)+b^*\right), \quad (2)$$

其中, α_i^* 和 b^* 由式(1)中最优化问题的解得到, SV 为支持向量集合。

2.2 基于 SVM 的专利 SAO 结构抽取方法

本文将中文专利 SAO 结构关系抽取任务转化为二分类问题, 使用支持向量机(SVM)来进行 SAO 结构关系三元组的抽取, 包括 3 个步骤: 训练语料准备、分类器训练和关系实例预测。

第一步: 训练语料准备。SVM 是一种有指导的机器学习方法, 对此需要一定规模的关系实例标注语料。首先, 我们使用分词工具对中文专利句子进行分词和词性标注, 在分词的基础上由人工进行实体标注, 以保证实体的准确率, 得到实体集。对实体集中的实体进行两两组合得到实体对。通过对语料的均匀随机采样发现, 候选关系词位于实体对外侧的实例中被标记为正例的仅占 1%, 故本文考察的候选关系词在句子中的位置位于两个实体之间, 这样大幅度减小了标注的规模。将得到的实体对与实体对之间的候选关系词组合, 得到关系三元

组实例。接着, 由人工对得到的关系实例进行标注, 正例表示候选关系词能正确反映两实体之间的关系, 负例表示候选关系词不能正确反映两实体间关系或两个实体不存在相互关系。

第二步: 分类器训练。SVM 的学习目的是, 在特征空间中寻找到一个能够将训练样本正确划分为两类的最优超平面, 并以此作为依据, 对测试样本进行预测。由第一步得到标注实例后, 提取标注实例的特征, 将其映射到 n 维的特征向量空间中, 得到对应的特征向量, 以此作为输入, 使用 SVM 工具训练得到一个分类器。

第三步: 关系实例预测。对测试集中的关系实例进行特征抽取, 将其映射到与训练样本相同的特征空间中, 然后将得到的特征向量作为训练得到的分类器的输入, 分类器依据训练得到的最优分类超平面对其进行判别, 得到预测结果。

3 特征选取

在实体关系抽取的相关研究中, 经常使用到的特征包括: 词法特征, 包括实例中出现的词语及其上下文的词和词性信息; 句法特征, 如全局的或局部的短语句法信息和依存句法信息; 距离特征, 反映实体之间的位置关系。下面对本文所分析的词法、句法、距离等特征进行详细说明。

3.1 实体词语及其上下文特征

词法信息是实例的最基本的特征, 也是最简单准确的特征。除实例本身的特征外, 通常还会使用实例的上下文特征。图 1 是上下文窗口大小为 2 时, 实例中实体、候选关系词及其上下文特征的示意图。

图 1 中“W_0 W_1 … W_{15} …”表示一个分词并进行词性标注后的中文句子, “W_i”表示句子中的一个词, 其中“$W_2W_3W_4$”和“$W_{12}W_{13}$”表示标注出的实体, “W_8”表示候选关系词。对于图 1 所示特征的说明见表 1。

表 1 中, C1 和 C6 表示实体对外侧的词特征; 实体对之间的词特征有两种, 一种形式为以上下文窗口形式表示的 C2, C3, C4, C5, 另一种则是表示实体 1 与候选关系词之间所有词的 CBL, 以及表示实体 2 与候选关系词之间所有词的 CBR。

实例的基础词法特征组合称为 BaseF, 即 BaseF=(Entity1+Rel+Entity2)。下文中具体词的向量表示, 如不做特别说明, 指基于词袋模型的 One-hot 表示法。

3.2 距离特征

距离特征在关系抽取中被广泛使用。在一个句子中, 如果两个实体之间的距离越近, 那么这两个实体之间存在交互关系的可能性也越大。本文对距离特征的有效性也进行了实验分析。距离特征包括: Dis(Entity1, Entity2), Enity1 与 Entity2 之间的词

表 1 特征说明

Table 1 Explanations of context features

特征符号	特征说明
Entity1	实体 1
Entity2	实体 2
Rel	候选关系指示词
C1	实体 1 左边的 2 个词
C2	实体 1 右边的 2 个词
C3	候选关系指示词左边的 2 个词
C4	候选关系指示词右边的 2 个词
C5	实体 2 左边的 2 个词
C6	实体 2 右边的两个词
CBL	实体 1 与候选关系词之间的短语块
CBR	实体 1 与候选关系词之间的短语块

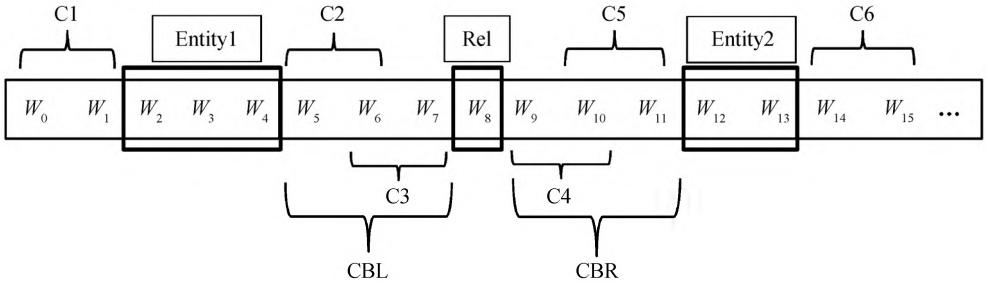

图 1 实体及其上下文特征示意图

Fig. 1 An example of entities and context features

的个数，如图 1 中，“实体 1”与“实体 2”的距离为 7; Dis(Entity1, Rel)为 Entity1 与候选关系词 Rel 之间的词的个数; Dis(Entity2, Rel)为 Entity2 与候选关系词 Rel 之间的词的个数。

3.3 句法特征

句子的句法分析树可以很好地反映句子中词语间的长距离依存关系，准确的句法分析信息有助于实体关系抽取。本文使用 Stanford Parser 3.2.0 (http://nlp.stanford.edu/software/lex-parser.shtml) 对句子进行句法分析。在进行句法解析之前，用“实体_1”和“实体_2”等代号对句子中的具体的实体进行替换，简化句子结构。例如句子“在 进气管 401 下方 设有 第二 管 402 ， 该 管 大致 呈 水平 状 。”中下划线所示的短语表示已标识出的实体。替换之后句子为“在 实体_1 下方 设有 实体_2 ，实体_3 大致 呈 实体_4 。”。对应的完全句法分析树如图 2 所示。

完全句法分析树虽然包含实例句子的完整句法结构信息，但也引入较多的噪声，会对关系抽取产生负面作用。本文参照文献[15]的策略，使用节点之间的最短路径闭包树 SPT (the shortest path enclosed tree)。SPT 指连接句法分析树中两个节点的最短路径所包围的节点构成的一颗子树。在前面例句中，“实体_1”与“实体_2”之间的最短路径为(“实体_1-NN-NP-PP-VP-VP-NP-NN-实体_2”)，其对应的 SPT 如图 3(a)所示。

本文选择的句法树特征包括：SPT(Entity1, Entity2)，如“实体_1”与“实体_2”之间的最短路径闭包树，如图 3(a)所示; SPT(Entity1, Rel)，如“实体_1”与候选关系词“设有”之间的最短路径闭包树，如图 3(b)所示; SPT(Rel, Entity2)，如候选关系词“设有”与“实体_2”之间的最短路径闭包树，如图 3(c)所

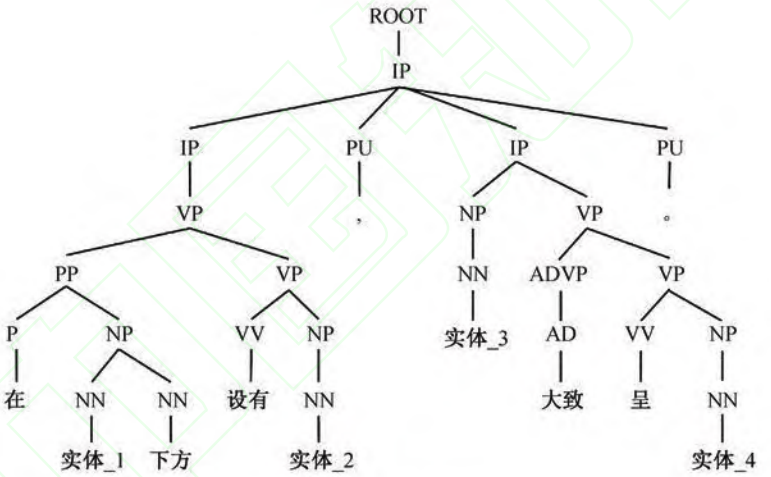

图 2 完全句法分析树

Fig. 2 Complete syntactic parsing tree

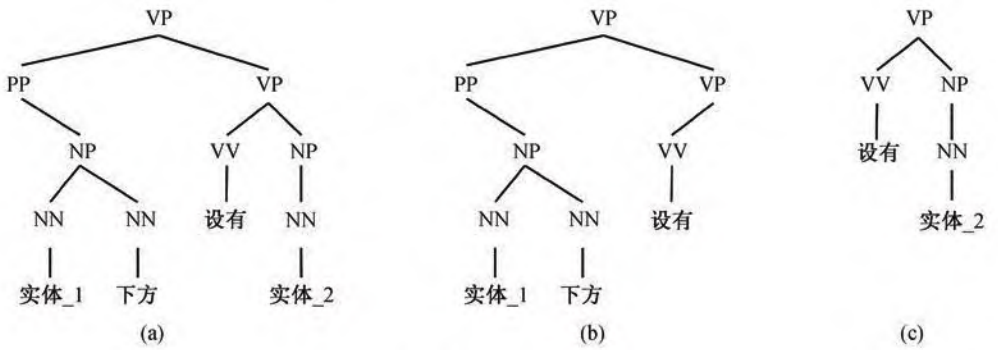

图 3 最小路径包含树

Fig. 3 Shortest path enclosed trees

示。

3.4 词向量

对于词特征的表示, 传统的表示方法是一种基于词袋模型的 One-hot 表示法, 通常是一个稀疏向量, 向量维数为词表的大小, 对应的词表索引位置的值为 1。这种表示法有一个明显的缺点, 即语义鸿沟问题, 词表中的任意两个词都是孤立的, 即使是语义上很相近的两个词, 也无法直接发现它们之间的联系。

与这种稀疏表示法不同, 在深度学习[16]中提供了一种词语的分布表示法 (distributed representation), 通常称之为词向量[17]。通过建立神经网络语言模型, 学习得到词的低维稠密向量表示形式。Turian 等[18]将词向量作为特征应用到命名实体识别和短语识别任务中, 识别效果得到进一步提升。在中文方面, 来斯惟等[19]在中文分词任务中引入词向量特征进行了有益的尝试。

本文对 92 万篇中文专利摘要文本使用 ICTCLAS (http://ictclas.nlpir.org)进行分词处理, 然后使用开源工具 word2vec (https://code.google.com/p/word2vec/)训练得到词向量模型, 得到的词表大小为 452 636 词。对于词表中出现的每个词, 由一个 200 维的稠密向量表示。在本文的实验中, 将比较基于词袋模型的词表示方式与词向量表示方式的关系抽取效果, 以验证词向量在关系抽取中的可行性。

4 实验与分析

4.1 实验设置

本文使用的实验语料为来源于生物、化学、计算机和机械领域的中文专利文本的摘要部分, 包括 2591 个句子, 语料的平均句长为 31.10 个词。由人工对语料进行实体和关系词标注, 得到 5595 个标注实例的标注集, 其中包含 2472 个正例和 3123 个负例。为提高实验的可靠性, 本文将标注集随机分成大致均等的 10 份, 每次选择其中 9 份作为训练集, 余下的 1 份作为测试集。实验中均采用 10 折交叉验证。实验使用的支持向量机工具包为 SVM-LIGHT-TK (http://disi.unitn.it/moschitti/Tree-Kernel.htm), 参数值设置为工具包中的默认值。

4.2 评价指标

对于抽取性能的评价, 本文使用精确率(A)、准确率(P)、召回率(R)和 F 值(F-measure)作为性能评价指标, 计算方法如下:

A=(正确识别的样本数/测试集总样本数)×100% ,

P=(正确识别出的正例样本数/预测结果中被标记为正例的样本数)×100%,

R=(正确识别出的正例样本数/测试集中的正例样本数)×100%, $F=2PR/(P+R)$。

此外, 本文对每组对比实验的结果进行显著性水平 $\alpha=0.05$ 的成对 T 检验, 以验证实验结果差异的统计显著性。

4.3 实验结果与分析

4.3.1 上下文特征的有效性验证实验

按照图 1 所示的词法特征表示, 以实例的词法特征作为基础特征组合, 即 BaseF, 上下文窗口大小设置为 2。在 BaseF 基础上加入实例的上下文特征。对上下文特征有效性验证的实验数据如表 2 所示。

从表 2 可以看出, 在实例的基础特征组合 BaseF 上加入实例的上下文特征后, F 值均有大幅度提升, 说明实例的上下文特征对于关系抽取性能的提高是很有效的。实例中的实体和关系词一般由领域术语构成, 而实例的上下文多由通用的介词、代词、方位词等组成。由于语料规模的限制, BaseF 是比较稀疏的, 引入上下文特征提供了更多的信息,

表 2 上下文特征的有效性验证

Table 2 Validity analysis of context features

序号	特征集合	A/%	P/%	R/%	F/%
#1	BaseF	72.00±2.69	73.16±3.62	57.76±4.57	64.50±4.03
#2	BaseF+C1+C6	77.07±2.20	75.14±2.89	72.25±4.34	73.56±2.28
#3	BaseF+CBL+CBR	79.20±1.92	75.90±2.81	77.65±2.70	76.73±2.09
#4	BaseF+CBL+CBR+C1+C6	82.77±1.69	81.84±2.87	78.62±3.68	80.11±1.89
#5	BaseF+C2+C3+C4+C5	79.74±2.37	75.91±2.98	79.46±3.63	77.60±2.68
#6	BaseF+C2+C3+C4+C5+C1+C6	83.08±2.22	82.25±3.87	79.03±3.97	80.50±2.38

得到的特征向量能够更准确地表示标注实例。这与我们对未知词的理解是类似的。对于未知词，我们可以通过比较与其上下文相同或相似的词来猜测其含义。

表 2 中#3 与#5, #4 与#6 对比，比较两种不同形式的实体对间特征对关系抽取性能的影响，发现与使用 CBL 和 CBR 特征时相比，使用(C2, C3, C4, C5)类型特征的#5 和#6 的召回率均有提升。通过对实验语料的统计发现，实体 1 与候选关系词之间平均有 1.45 个词，实体 2 和候选关系词之间平均有 0.97 个词。当上下文窗口设置为 2 时，部分实例的 C2, C3 特征与 CBL 特征是相同的，即这两种特征形式引入的上下文信息基本相同，特征区分度不高，故在正确率和 *F* 值上并没有产生显著变化。

4.3.2 距离特征的有效性验证实验

在将距离信息加入到特征向量中时，我们并不是直接使用绝对距离值，而是对其进行了一定程度的弱化，即对距离值乘上一个合适的弱化系数 k。通过对 k 值的试验，发现当 k 取 0.2 时能得到较好的实验效果。表 3 给出对距离特征有效性的验证实验，使用的基准特征组合为 Baseline=(BaseF+C1+C2+C3+C4+C5+C6)，距离特征组合 Dis=(Dis(Entity1, Entity2)+Dis(Entity1, Rel)+Dis(Entity2, Rel))。

从表 3 可见，加入距离特征后，*A*, *P*, *R* 和 *F* 平均值较未加入距离特征前分别提升了 0.79, 1.15, 0.5 和 0.83，其中准确率和 *F* 值的增加是具有统计显著性的。对实验结果分析发现，准确率的提高一方面是由于对正例中的短距离实例(实体间距离≤5)正确识别数量的增加，另一方面是被识别为正例的实例总数的减少，即识别出更多正确的负例，其中负例中的长距离实例(实体间距离＞5)被正确识别的数量增多。距离特征的加入有利于对正例中的短距离实例以及负例中的长距离实例的正确识别。对实验语料中的实体间距离信息进行统计，得到正例的实体间距离平均值为 2.29 个词，负例的实体间平均距离为 4.31 个词，表明距离特征是有区分度的，在 SAO 结构关系抽取中距离特征是有效的。

4.3.3 句法特征的有效性验证实验

表 4 给出句法树特征与特征向量组合的对比实验数据。“Vec”表示仅使用特征向量，构成向量的特征组合为“(BaseF+C1+C2+C3+C4+C5+C6)”。

表 3 距离特征的有效性验证

Table 3 Validity analysis of distance features

序号	特征集合	*A*/%	*P*/%	*R*/%	*F*/%
#1	Baseline	83.08±2.22	82.25±3.87	79.03±3.97	80.50±2.38
#2	Baseline+Dis	83.87±1.90	83.40±2.92	79.53±3.87	81.33±2.03

表 4 句法特征的有效性验证

Table 4 Validity analysis of syntactic features

序号	特征集合	*A*/%	*P*/%	*R*/%	*F*/%
#1	Vec	83.08±2.22	82.25±3.87	79.03±3.97	80.50±2.38
#2	SPT	83.02±2.47	84.14±3.39	76.07±4.31	79.82±2.94
#3	SPT+Vec	84.21±2.43	85.62±3.22	77.41±4.50	81.22±2.94

表 5 上下文窗口大小对实验结果影响

Table 5 Impact of context window size on experiment results

序号	窗口大小	*A*/%	*P*/%	*R*/%	*F*/%
#1	win=1	82.67±2.42	81.94±3.25	78.21±4.25	79.94±2.65
#2	win=2	83.08±2.22	82.25±3.87	79.03±3.97	80.50±2.38
#3	win=3	83.46±2.16	82.63±3.55	79.56±3.95	80.96±2.24
#4	win=4	83.75±2.15	83.15±3.01	79.55±4.26	81.22±2.43
#5	win=5	83.37±2.17	82.57±2.93	79.32±4.16	80.82±2.29

“SPT”表示仅使用 SPT 句法树信息, SPT 特征组合为 SPT=(SPT(Entity1, Rel)+SPT(Entity2, Rel)+SPT(Entity1, Entity2))。“SPT+Vec”表示同时使用两者。

#1 和#2 对比, 与仅使用特征向量相比, 仅使用 SPT 时能得到更高的准确率, 但召回率要低。将特征向量与 SPT 句法树组合后, 准确率要比仅使用特征向量的表示方法要高 3.3%, 但召回率下降了 1.6%。通过对实验数据的显著性检验, 我们发现 *F* 值无显著变化。#3 中, 当特征向量与 SPT 句法树组合时, 能得到最高的准确率。但与#1 相比 *F* 值并无统计差异性。SPT 句法信息的引入并没有显著提高系统的性能。相比基本的上下文等词法信息, 句法信息对于远距离的关系实例的识别有较好表现, 而本文实验标注的关系实例中远距离的关系实例(实体间距离＞5)占比很小, 约为 15%, 短距离的标注实例(实体间距离≤5)约占 85%。对于短距离的关系实例, 基本的词法信息已能起到很好的识别作用, SPT 句法信息在此并没有表现出优势。进一步分析发现, SPT 句法信息的引入带来更多的句法结构限制, 使得识别为正例的实例数量减少, 更倾向于识别出更多的负例, 这也是相比于仅使用特征向量时, 精确率上升而召回率下降的原因。

4.3.4 上下文窗口大小的影响

使用上下文特征, 上下文窗口的不同也会对实验结果产生不同程度的影响。窗口选择过小, 则引入的上下文信息不够, 若窗口设置过大, 又会引入一些噪声。表 5 列出上下文窗口大小由 1 逐渐增加至 5 的实验结果。使用的特征组合为“BaseF+C1+C2+C3+C4+C5+C6”。

从表 5 可见, 当上下文窗口为 4 时能够得到最高的精确率, 准确率和 *F* 值。我们通过对窗口大小相差为 1 的各组实验数据做差异显著性检验, 发现相邻两组实验结果差异无统计显著性。对窗口大小相差为 2 的实验数据(#1 和#3, #2 和#4, #3 和#5)进行 *T* 检验发现, 其中#1 和#3, #2 和#4 的数据差异是有统计显著性的。由此可知上下文窗口的大小对关系抽取的性能有影响。适当加大上文窗口, 引入更多有效的上下文信息, 对于关系抽取是有益的。

4.3.5 词特征表示方式的比较

为尝试词向量信息在 SAO 结构关系抽取中的可行性, 本文进行使用词向量的表示方法(WordVec)与基于词袋模型(BoW)的 One-hot 表示形式在 win=1 和 win=2 下的对比实验, 结果如表 6 所示, 其中, 特征组合选择的是“BaseF+C1+C2+C3+C4+C5+C6”。BoW 中的每个词由 One-hot 形式的稀疏向量表示。WordVec 中的每个词的向量表示是从 3.4 节中所述词向量模型中获得的, 为 200 维的稠密向量形式。

从表 6 可见, 在 win=1 和 win=2 时, WordVec 的准确率都略高于 BoW, 而召回率略低于 BoW, *F* 值持平。通过对实验结果的显著性检验, 我们发现, 只有在 win=2 时, WordVec 较 BoW 的召回率的提高是显著的, 其他性能指标的差异并无统计显著性。BoW 与 WordVec 两种词特征表示方式的实验效果相当, 表明在中文专利 SAO 结构关系抽取中, 使用词向量特征表示词特征是可行的。

5 结语

本文针对中文专利文本中的 SAO 结构关系抽取问题, 使用基于 SVM 的机器学习方法, 在中文专利语料上进行 SAO 结构关系抽取实验, 对于基本的词法特征、句法结构特征和距离特征在关系抽取中的有效性进行了比较分析。实验分析表明, 简单的上下文词法特征能显著地提高抽取性能。其中, 相对于实体对外侧的上下文特征, 实体对之间的上下文特征具有更好的识别效果, 距离特征对于关系抽取也有积极作用。在关系抽取中加入 SPT 短语句法树信息, 综合性能并没有得到明显提升。最后, 本文还尝试了词向量在关系抽取中的表现,

表 6 BoW 与 WordVec 的比较

Table 6 Comparison of BoW and WordVec

窗口大小	特征表达形式	*A*/%	*P*/%	*R*/%	*F*/%
win=1	BoW	82.67±2.42	81.93±3.25	78.21±4.25	79.94±2.65
	WordVec	82.92±2.44	82.97±3.19	77.50±4.18	80.05±2.54
win=2	BoW	83.08±2.22	82.25±3.87	79.03±3.97	80.50±2.38
	WordVec	83.25±2.38	83.09±3.55	78.27±3.98	80.51±2.50

得到了词袋模型表示法相当的实验效果，表明词向量在关系抽取中是可行的。基于本文的特征分析结果，在下一步工作中，我们将尝试在特征向量中引入一些深层的语义信息，对特征进行有效的组合，以期得到更好的抽取性能。

参考文献

[1] Cascini G, Fantechi A, Spinicci E. Natural language processing of patents and technical documentation. Document analysis systems Ⅵ. Berlin: Springer, 2004: 508–520

[2] Moehrle M G, Walter L, Geritz A, et al. Patent-based inventor profiles as a basis for human resource decisions in research and development. R&D Management, 2005, 35(5): 513–524

[3] Bergmann I, Butzke D, Walter L, et al. Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips. R&D Management, 2008, 38(5): 550–562

[4] Altshuller G S. Creativity as an exact science: the theory of the solution of inventive problems. New York: Gordon and Breach Science Publishers, 1984

[5] Banko M, Cafarella M J, Soderland S, et al. Open information extraction for the web // The Twentieth International Joint Conference on Artificial Intelligence. Hyderabad, 2007, 7: 2670–2676

[6] 赵军, 刘康, 周光有, 等. 开放式文本信息抽取. 中文信息学报, 2011, 25(6): 98–110

[7] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations // Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Barcelona, 2004: Article 22

[8] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods //roceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, 2005: 419–426

[9] 奚斌, 钱龙华, 周国栋, 等. 语言学组合特征在语义关系抽取中的应用. 中文信息学报, 2008, 22(3): 44–49

[10] 黄鑫, 朱巧明, 钱龙华, 等. 基于特征组合的中文实体关系抽取. 微电子学与计算机, 2010, 27(4): 198–200

[11] 黄晨, 钱龙华, 周国栋, 等. 基于卷积树核的无指导中文实体关系抽取研究. 中文信息学报, 2010, 24(4): 11–17

[12] Etzioni O, Fader A, Christensen J, et al. Open information extraction: the second generation // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. Barcelona, 2011, 11: 3–10

[13] Schmitz M, Bart R, Soderland S, et al. Open language learning for information extraction // Processing of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, 2012: 523–534

[14] 赵奇猛, 王裴岩, 冯好国, 等. 面向中文专利的开放式实体关系抽取研究[J/OL]. 计算机工程与应用. (2013-09-12) [2014-01-28]. http://www.cnki.net/kcms/detail/11.2127.TP.20130912.1433.006.html

[15] 李丽双, 刘洋, 黄德根. 基于组合核的蛋白质交互关系抽取. 中文信息学报, 2013, 27(1): 86–92

[16] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507

[17] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems. Lake Tahoe, 2013: 3111–3119

[18] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning // Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics. Uppsala, 2010: 384–394

[19] 来斯惟, 徐立恒, 陈玉博, 等. 基于表示学习的中文分词算法探索. 中文信息学报, 2013, 27(5): 8–14