doi:10.3969/j.issn.1673-9833.2016.06.08

基于本体特征的汽车领域命名实体识别

张永平，朱艳辉，朱道杰，王天吉，李 飞

（湖南工业大学 计算机学院，湖南 株洲 412007）

摘 要：针对汽车领域命名实体识别中汽车属性名识别的准确率和召回率较低的问题，提出了一种基于本体特征的汽车领域命名实体识别方法。通过扩展现有叙词表，基于叙词表构建汽车领域本体，提取语料中的本体特征，利用CRFs模型对汽车领域命名实体进行识别。实验结果表明，本体特征能够有效地识别出汽车属性实体，准确率、召回率和*F*值分别为75.60%, 66.12%和70.54%。

关键词：命名实体识别；汽车领域；CRFs；本体；叙词表

中图分类号：TP391.4 **文献标志码**：A **文章编号**：1673-9833(2016)06-0039-05

An Ontology-Based Named Entity Recognition in Automotive Industry

ZHANG Yongping, ZHU Yanhui, ZHU Daojie, WANG Tianji, LI Fei

（School of Computer, Hunan University of Technology, Zhuzhou Hunan 412007, China）

Abstract: In view of a low accuracy rate and recall rate of named entity recognition in the automotive industry, a new method of named entity recognition based on ontology has thus been proposed. By extending the existing thesauri, and constructing an automobile domain ontology, the ontology features are to be extracted from the corpus, and a named entity recognition based on a CRFs model can be achieved. The experimental results show that the ontology features can effectively identify the vehicle attribute entities, with its accuracy rate as high as 75.60%, a recall rate as high as 66.12% and a *F*-value as high as 70.54% respectively.

Keywords: named entity recognition; automotive field; CRFs; ontology; thesaurus

0 引言

随着计算机的快速普及，互联网的迅猛发展，各式各样的信息呈爆炸式增长，如何从海量的数据中精准地抽取用户所需信息已成为研究者关注的课题。信息抽取的主要目的是将非结构化的自然语言文本转化成半结构化或者结构化数据，以便人们准确快速地获取信息。命名实体识别[1]作为信息抽取的子任务，已经成为研究的热点。其研究方法分别有基于规则[2]、基于统计[3]以及基于规则和统计[4]相结合的方法，研究领域从通用领域扩展到专业领域。在专业领域中，由于语料缺乏和属性名难以识别的特点，使其成为命名实体识别中的难点。

本文针对汽车领域命名实体进行识别，选择COAE会议[5]提供的汽车类语料，通过对汽车语料的深入分析，发现汽车属性具有以下特点：1）数量多，汽车的结构、零部件、内饰和动力总成等名称都是汽车的属性；2）口语化，比如句子“这车皮薄”

收稿日期：2016-10-13

基金项目：国家自然科学基金资助项目（61170102），国家社会科学基金资助项目（12BYY045），湖南省教育厅基金资助重点项目（15A049）

作者简介：张永平（1989-），男，贵州习水人，湖南工业大学硕士生，主要研究方向为自然语言处理，E-mail：780235260@qq.com

通信作者：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，硕士生导师，主要从事自然语言处理方面的研究，E-mail：swayhzhu@163.com

中的实体“皮薄”，句子“20寸大脚太霸气”中的实体“大脚”等，这使得汽车命名实体中属性名的抽取难度较大。针对这些问题，本文通过基于叙词表[6-7]的方法构建汽车领域本体[8]，并以本体为特征，采用条件随机场（conditional random fields，CRFs）模型[9]对汽车领域命名实体进行识别，有效提高识别的准确率。

1 基于叙词表的汽车本体构建

本研究选用的叙词表是《汽车工程叙词表》，但由于这个叙词表发行时间较早，没有进行更新修订，有很多新概念及属性都没有，所以首先要对叙词表进行升级优化，优化算法如下。

Step 1 从《汽车工程叙词表》中取出“汽车结构”概念为新叙词表，其中包括“汽车结构”和“汽车零部件”的概念、定义以及等级关系。

Step 2 从汽车百科网上获取关于汽车的所有名词以及释义。

Step 3 把Step1和Step2得到的内容组合去重。

Step 4 按照老叙词表中概念的等级关系框架，逐个把Step 3的概念和释义添加到新叙词表中。

基于叙词表构建汽车领域本体，基本思路是：1）根据叙词表确定核心概念集；2）确定概念间关系；3）添加汽车领域概念属性；4）为本体添加实例。具体算法如下。

Step 1 选择叙词表中“汽车种类”和“汽车结构”下的名称为父亲概念，然后添加叙词表中相应的子概念，得到核心概念集。

Step 2 确定概念间关系：确定了核心概念集后，利用中间展开[10]的方法，在叙词表中逐步抽取概念间的关系。

Step 3 添加概念属性，把叙词表中对概念的释义当作属性。

Step 4 添加实例，实例是对概念的举例，可以从“太平洋汽车”网中“分类”板块获得，添加到本体相应的概念中。

2 汽车命名实体识别

条件随机场模型是给定一组输入随机变量，求另外一组具有隐马尔可夫性质的输出随机变量的条件概率分布的无向图。在自然语言处理任务中，很多地方都用到了条件随机场，例如新词识别、中文分词、依存关系等。基于条件随机场的主要实现工具有CRF++，FlexCRF等，本文使用的模型训练和测试工具为CRF++。

本文提出的基于领域本体的汽车命名实体识别方法的基本流程图如图1所示。

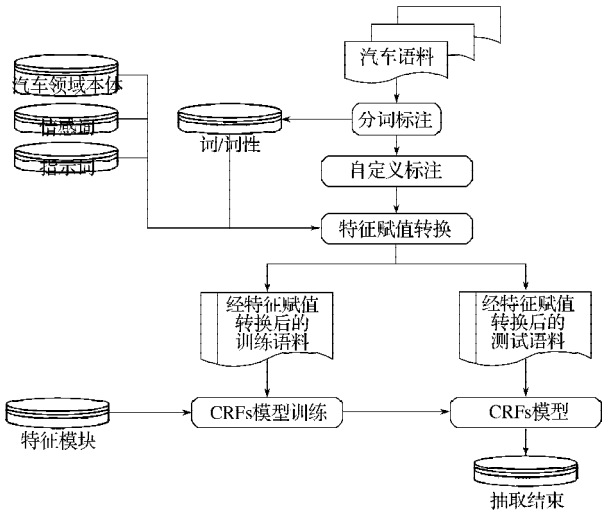

图1 汽车领域命名实体识别流程图

Fig. 1 Named entity recognition process in the automotive field

2.1 特征抽取

汽车领域本体特征表示的是词汇单元所具备的领域及其语义特征，反映领域属性共识。汽车命名实体识别最大的难度是汽车属性名的识别，课题组利用汽车本体可以对汽车属性名和其它实体之间的关系进行描述，从而建立起了属性名和其它实体之间的关系。通过这种“关系”，可以对汽车属性进行有效的识别。为了进行对比，除了本体特征外，本文还提取了词和词性、指示词、情感倾向这3个特征，并组成特征模板。

1）词和词性特征

词特征为实验语料经过分词后的词汇单元本身，即将分词后的结果作为一类特征，可以表示词在句子中的位置；词性特征表示词在句子中的词性标注，利用NLPIR[11]分词工具可以获得这2个特征。词和词性特征模板如表1所示。

表1 词和词性特征模板

Table 1 A feature template for words and its part of speech

编号	特征	描述
1	Word(0)	当前词
2	Word(-1)	前一个词
3	Word(-2)	前两个词
4	Word(1)	后一个词
5	Word(2)	后两个词
6	Pos(0)	当前词的词性
7	Pos(-1)	前一个词的词性
8	Pos(1)	后一个词的词性
9	Pos(-2)	前两个词的词性
10	Pos(2)	后两个词的词性

2）指示词特征

指示词是指在命名实体周围具有指示性的词语，

如：“新款马自达阿特兹在性能方面很可靠”中的“性能”、“进口的 A4 和国产的 A4L 有啥区别？”中的“进口”和“国产”就是指示词。本文提出的指示词抽取算法如下。

Step 1　建立一个空的指示词库。

Step 2　依次读取已经经过分词处理的词汇。

Step 3　若当前词是命名实体，则转到 Step 4，否则转到 Step 2。

Step 4　以命名实体为中心，选择窗口大小 N，即取当前词的前 N 个词和后 N 个词共同组成集合 *boundary*。

Step 5　把集合 *boundary* 中的词和指示词库中的词一一对比，若有相同的词，则该词的词频加 1；若无相同的词，则把该词加入到指示词库，并将词频设为 1。

Step 6　判断当前词是否为语料的最后一个词，是则转到 Step 7，否则转到 Step 2。

Step 7　设置一个阈值，将指示词库中的词频小于阈值的词移出指示词库。

指示词特征模板如表 2 所示。

表 2　指示词特征模板

Table 2　A feature template for demonstratives

编号	特　征	描　述
1	Indicate(0)	当前词是否为指示词
2	Indicate(−1)	前一个词是否为指示词
3	Indicate(1)	后一个词是否为指示词
4	Indicate(−2)	前两个词是否为指示词
5	Indicate(2)	后两个是否为指示词

3）情感倾向特征

文本的情感倾向是指文本中的用户所表达的态度，通过对情感倾向的分析可以看出评论者对事物态度是积极还是消极，其中评论者和评论对象很有可能是命名实体。本文采用文献[12]的方法抽取情感特征，情感特征模板如表 3 所示。

表 3　情感特征模板

Table 3　A feature template for affective words

编号	特　征	描　述
1	Sentiment(0)	当前词是否为情感词
2	Sentiment(−1)	前一个词是否为情感词
3	Sentiment(1)	后一个词是否为情感词
4	Sentiment(−2)	前两个词是否为情感词
5	Sentiment(2)	后两个词是否为情感词

4）本体特征

领域本体中的类别有概念、属性和实例，本体特征是指分词后的词语是否属于本体类别中的种类，提取过程是将分词后的词汇在构建的本体系统中进行等级关系的判定，返回词汇所属的类别。本体特征模板如表 4 所示。

表 4　本体特征模板

Table 4　Ontology template

编号	特征	描　述
1	Noumenon(0)	当前词在本体中存在
2	Noumenon(1)	后一个词在本体中存在
3	Noumenon(2)	后两个词在本体中存在
4	Noumenon(−1)	前一个词在本体中存在
5	Noumenon(−2)	前两个词在本体中存在

2.2　特征转换

本次实验所使用的工具是 CRF++0.54[13]，使用时须把具体特征转变为标注符，称为特征标记取值，结合 2.1 节所介绍特征，为各个特征制定一个转换标注，具体如表 5 所示。

表 5　特征标记取值

Table 5　Characteristic marks

特征	特征标记	描　述
词和词性特征	W	词本身
	POS	词性标注
指示词特征	Ind-Y	当前词是指示词
	Ind-N	当前词不是指示词
情感特征	Sen-P	当前词为正面评价的词
	Sen-N	当前词为负面评价词
	Sen-O	当前词不是评价词
本体特征	Nou-C	当前词在本体中以类存在
	Nou-P	当前词在本体中以属性存在
	Nou-I	当前词在本体中以实例存在
	Nou-O	当前词没有在本体中出现

2.3　结果标注集

在利用 CRFs 进行训练和测试时，要指定一个标注集，本文采用的标注集如表 6 所示。

表 6　结果标注集

Table 6　Result annotation set

输出标签	含　义
Bra-B	品牌实体开端部分
Bra-M	品牌实体中间部分
Bra-E	品牌实体结尾部分
Bra-S	单独构成品牌实体
Bra-F1	品牌实体被上文割裂或单独割裂的前面部分
Bra-F2	品牌实体被下文割裂或单独割裂的其他部分
Ser-B	系列实体开端部分
Ser-M	系列实体中间部分
Ser-E	系列实体结尾部分
Ser-S	单独构成系列实体
Ser-F1	系列实体被上文割裂或单独割裂的前面部分
Ser-F2	系列实体被下文割裂或单独割裂的其他部分
Att-B	属性实体开端部分
Att-M	属性实体中间部分
Att-E	属性实体结尾部分
Att-S	单独构成属性实体
Att-F1	属性实体被上文割裂或单独割裂的前面部分
Att-F2	属性实体被下文割裂或单独割裂的其他部分
N	非命名实体

3 实验结果及分析

3.1 实验语料

本次实验选取了从COAE2008至COAE2015所有的汽车类语料，从中筛选出22 303句，其中14 000句为训练语料，剩下的句子为测试语料。

3.2 实验工具介绍

本实验采用CRF++外部开发包来完成CRFs模型的训练和测试。CRF++是目前综合性能最佳的条件随机场开源工具，其对训练语料的格式要求是：训练语料的列为特征，并且至少有两列。使用CRF++工具包还需要定义一个特征模板文件，也就是特征的组合方式，本文的特征组合方式总共有6种，在下一节详细介绍。训练过程中只要把训练语料和特征模板作为输入，利用CRF++工具训练，输出就是训练好的模型，这个模型可以用来做测试。

3.3 实验结果与分析

本实验采用CRF++外部开发包来完成CRFs模型的训练和测试，实验结果如表7所示。

表7 实验对比结果

Table 7 Contrast of experimental results %

特征组合	准确率 P	召回率 R	F 值
①词 + 词性特征	66.11	46.46	54.57
②词 + 词性 + 指示词特征	68.03	48.98	56.95
③词 + 词性 + 情感倾向特征	71.30	50.82	59.34
④词 + 词性 + 领域本体特征	72.78	52.26	60.84
⑤词 + 词性 + 指示词 + 情感倾向特征	72.51	51.00	59.88
⑥词 + 词性 + 指示词 + 情感倾向 + 领域本体特征	75.60	66.12	70.54

在表7中，特征组合②、③、④的对比实验表明，本体特征要优于情感倾向特征和指示词特征，其中准确率比②高4.75%；特征组合④、⑤的对比实验表明，虽然⑤在特征数量上比④更多，但识别效果却并不比其好，说明本体特征要优于同时拥有指示词和情感倾向特征的模板，并且在模型训练时效上特征组合④也优于⑤；特征组合⑤、⑥的对比实验表明，在⑤的基础上加入本体特征，准确率、召回率和F值分别高出3.09%, 15.12%和10.66%，特别在召回率上大幅领先；以上几组对比数据表明，本体特征能够对汽车命名实体进行有效的识别。

4 结语

本文提出了基于本体特征的汽车领域命名实体识别方法，首先通过叙词表构建了汽车领域本体，并选择本体作为特征，基于CRFs模型进行汽车命名实体识别。通过与指示词特征、情感特征进行对比，实验表明，基于本体特征的识别效果最好，特别是在属性名的识别上。虽然本文研究取得了一定成果，但不足之处是本体的构建方法偏于简单，导致命名实体识别效果整体偏低，因此构建一个质量较优的本体是将来要进一步研究的工作。

参考文献：

[1] 张晓艳，王 挺，陈火旺. 命名实体识别研究[J]. 计算机科学，2005，32(4)：44-48.
ZHANG Xiaoyan，WANG Ting，CHEN Huowang. Research on Named Entity Recognition[J]. Computer Science，2005，32(4)：44-48.

[2] 周 昆. 基于规则的命名实体识别研究[D]. 合肥：合肥工业大学，2010.
ZHOU Kun. Research on Named Entity Recognition Basd on Rules[D]. Hefei：HeFei University of Technology，2010.

[3] 俞鸿魁，张华平，刘 群，等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报，2006，27(2)：87-94.
YU Hongkui，ZHANG Huaping，LIU Qun，et al. Chinese Named Entity Identification Using Cascaded Hidden Markov Model[J]. Journal of Communications，2006，27(2)：87-94.

[4] 向晓雯，史晓东，曾华琳. 一个统计与规则相结合的中文命名实体识别系统[J]. 计算机应用，2005，25(10)：2404-2406.
XIANG Xiaowen，SHI Xiaodong，ZENG Hualin. Chinese Named Entity Recognition System Using Statistics-Based and Rules-Based Method[J]. Computer Application，2005，25(10)：2404-2406.

[5] 廖祥文，许洪波，孙 乐，等. 第三届中文倾向性分析评测(COAE2011)语料的构建与分析[J]. 中文信息学报，2013，27(1)：56-63.
LIAO Xiangwen，XU Hongbo，SUN Le，et al. Construction and Analysis of the Third Chinese Opinion Analysis Evaluation (COAE2011) Corpus[J]. Journal of Chinese Information Processing，2013，27(1)：56-63.

[6] 常 春，卢文林. 叙词表编制历史、现状与发展[J]. 农业图书情报学刊，2002(5)：25-28.

CHANG chun，LU Wenlin. The History，Current Situation and Development of Compilation of the Thesaurus [J]. Journal of Library and Information Science in Aricultural，2002(5)：25-28.

[7] 杨秋芬，陈跃新. Ontology方法学综述[J]. 计算机应用研究，2002，19(4)：5-7.
YANG Qiufen，CHEN Yuexin. A Survey of Ontology Methodology[J]. Computer Application Research，2002，19(4)：5-7.

[8] BORST W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[J]. Universiteit Twente，1997，18(1)：44-57.

[9] LAFFERTY J D，Mccallum A，Pereira F C N. Conditional Random Fields：Probabilistic Models For Segmenting And Labeling Sequence Data[C]// ICML 2001 Proceedings of the Eithteenth International Conference on Machine. San Francisco：Morgan Kaufmann Publishers，2001：282-289.

[10] 唐爱民，真　溱，樊　静. 基于叙词表的领域本体构建研究[J]. 现代图书情报技术，2005(4)：1-5.
TANG Aimin，ZHEN Zhen，FAN Jing. Thesaurus-Based Approach to Build Domain Ontology[J]. New Technology of Library and Information Service，2005(4)：1-5.

[11] ZHOU L，ZHANG D. NLPIR：A Theoretical Framework for Applying Natural Language Processing to Information Retrieval[J]. Journal of the American Society for Information Science & Technology，2003，54(2)：115-123.

[12] 朱艳辉，栗春亮，徐叶强，等. 一种基于多重词典的中文文本情感特征抽取方法[J]. 湖南工业大学学报，2011，25(2)：42-46.
ZHU Yanhui，LI Chunliang，XU Yeqiang，et al. A Method of Emotional Feature Extraction in Chinese Text Based on Multiple Lexicons[J]. Journal of Hunan University of Technology，2011，25(2)：42-46.

[13] Source Forge. CRF++[EB/OL]. [2016-07-19]. https://sourceforge.net/projects/crfpp/.

（责任编辑：申　剑）

（上接第6页）

[4] 包春燕，姜谙男，唐春安，等. 单轴加卸载扰动下石灰岩声发射特性研究[J]. 岩石力学与工程学报，2011，30（增刊2）：3871-3877.
BAO Chunyan，JIANG Annan，TANG Chun'an，et al. Study of Acoustic Emission Characteristics of Limestone Under Cycle Uniaxial Loading-Unloading Perturbation[J]. Chinese Journal of Rock Mechanics and Engineering，2011，30(S2)：3871-3877.

[5] 曹树刚，刘延保，李　勇，等. 不同围压下煤岩声发射特征试验[J]. 重庆大学学报，2009，32(11)：1321-1327.
CAO Shugang，LIU Yanbao，LI Yong，et al. Experimental Study on Acoustic Emission Characteristics of Coal Rock at Different Confining Pressure[J]. Journal of Chongqing University，2009，32(11)：1321-1327.

[6] 陈景涛. 岩石变形特征和声发射特征的三轴试验研究[J]. 武汉理工大学学报，2008，30(2)：94-96.
CHEN Jingtao. Experimental Study on Triaxial Compression Deformation and Acoustic Emission Property of Rock[J]. Journal of Wuhan University of Technology，2008，30(2)：94-96.

[7] 苏承东，高保彬，南　华，等. 不同应力路径下煤样变形破坏过程声发射特征的试验研究[J]. 岩石力学与工程学报，2009，28(4)：757-766.
SU Chengdong，GAO Baobin，NAN Hua，et al. Experimental Study on Acoustic Emission Characteristics During Deformation and Failure Processes of Coal Samples Under Different Stress Paths[J]. Chinese Journal of Rock Mechanics and Engineering，2009，28(4)：757-766.

（责任编辑：邓光辉）