最大熵模型的事件分类

于江德[1]，李学钰[1]，樊孝忠[2]，庞文博[2]

(1. 安阳师范学院计算机与信息工程学院 河南 安阳 455002; 2. 北京理工大学计算机科学技术学院 北京 海淀区 100081)

【摘要】提出了一种基于最大熵模型的事件分类方法，该方法能够综合事件表述语句中的触发词信息及各类上下文特征对事件进行分类。对其中的两个关键问题：参数估计、特征模板与特征选择进行了详细论述，采用IIS算法学习模型参数，使用增量选择方法选择特征。应用该方法对人民日报语料中的职务变动、会见、恐怖袭击、法庭宣判、自然灾害五类事件进行了分类实验，结果表明，该方法的分类效果明显优于传统的分类方法。

关 键 词 事件信息抽取; 事件分类; 事件表述语句; 最大熵模型; 触发词

中图分类号 TP391 **文献标识码** A doi:10.3969/j.issn.1001-0548.2010.04.030

Event Classification Based on Maximum Entropy Model

YU Jiang-de[1], LI Xue-yu[1], FAN Xiao-zhong[2], and PANG Wen-bo[2]

(1. School of Computer and Information Engineering, Anyang Normal University Anyang Henan 455002;
2. School of Computer Science and Technology, Beijing Institute of Technology Haidian Beijing 100081)

Abstract An approach based on maximum entropy model is proposed for event classification. This approach can classify the events by merging the features about trigger and context in event mention sentences. The key of the method is parameter estimation and feature selection, which are discussed in detail. IIS algorithm is employed for parameter estimation and incremental method is used for feature selection. Experiments are performed on management succession, meeting, terror attack, judicial adjudicate, and natural disaster in the People Daily corpus. The results show that the method can achieve much better performance than the traditional approach.

Key words event information extraction; event classification; event mention sentence; maximum entropy model; trigger

最近几年，信息抽取(information extraction)研究受到了越来越多的关注。事件信息抽取(简称事件抽取，event extraction)是从自然语言形式的文本中自动地抽取用户感兴趣的事件以及卷入其中的特定类型的实体，并将这些信息转换为结构化数据并存储到预定义模板的过程。例如，从新闻报道中抽取恐怖事件发生的时间、地点、恐怖分子、受害者、袭击目标、袭击方式等详细情况。

事件抽取可分为两步：(1) 是对特定事件的探测和事件的分类，主要探测特定事件的候选表述语句并确定事件的类别；(2) 是从事件表述语句中抽取事件要素并填充到预定义的事件模板中。事件探测和事件分类是事件抽取的基础，事件探测旨在发现感兴趣事件的表述语句，这些语句是进一步进行事件信息抽取的数据源；而事件分类用于确定事件表述语句所叙述的事件类别，事件类别正确与否对事件模板的选择以及究竟要抽取哪些事件要素填充模板至关重要。传统的事件探测和事件分类主要依据触发词(trigger)确定，触发词是能够很好地概述事件中心意义的词。例如，职务变动事件中的“任命”、“辞去”等词语。基于触发词的事件探测和分类是将含有特定触发词的语句作为候选事件语句并依据触发词对事件进行分类。例如，文本中的语句“集团董事长辞去集团科技董事长职务”包含触发词“辞去”，就认为该语句是个离职类事件的表述语句。对大量事件表述语句进行研究发现：仅仅依据触发词就判定一个语句是某类候选事件语句容易出错。一是有些包含触发词的语句并未表述相关事件；二是一些词语在多个事件类别中充当触发词。对于该两种情况，简单地依据触发词确定事件类别是不可取的，而触发词的上下文中包含对事件类别确定有重大参考价值的各类特征，例如，触发词前后的一些特定类型的命名实体、一些用于表述某类特定事件的固定句式、短语结构和词语等。基于以上的分析，本

收稿日期：2008 – 11 – 17; 修回日期：2009 – 12 – 13
基金项目：教育部博士点基金(20050007023)
作者简介：于江德(1971 –)，男，博士，副教授，主要从事自然语言处理、信息抽取、文本数据挖掘等方面的研究.

文提出了一种基于最大熵模型的事件分类方法，将候选事件语句中的触发词及其上下文信息融合到最大熵模型中进行事件类别判定，运用该方法对《人民日报》语料中的职务变动、会见、恐怖袭击、法庭宣判、自然灾害等5类事件进行分类实验，结果表明，该方法效果较优。

1 最大熵模型及相关研究

1.1 最大熵理论

最大熵理论反映了自然界的一条基本原则：事物是约束和自由的统一体，事物在约束下总是争取最大自由度，即最大熵。因此，在已知条件下，熵最大的事物，最可能接近它的真实状态。具体来说，对于一个事件，往往只了解它的部分情况，对于其他情况则一无所知，对该事件建立模型时，对于已知的部分要尽量地拟合，使模型符合已知的情况；对于未知的情况，则保持均匀分布，即使该事件的熵最大。

1.2 最大熵模型的一般形式

对于分类问题，给定一些训练样本(x,y)，其中x表示上下文，y表示问题的类别，可根据已知的样本构建一个能够对实际问题进行准确描述的统计模型$p(y|x)$，用于预测未知事件。该模型的概率分布与训练语料中的经验概率分布应该相符。最大熵原理表明，x、y的正确分布应该是，在满足已知条件(约束)的情况下，使熵的分布最大，所构建的模型就是最大熵模型，其一般形式为：

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

式中 $Z(x)=\sum_{y}\exp\left[\sum_{i=1}^{k}\lambda_i f_i(x,y)\right]$ 为归一化因子，保证对所有可能的上下文x有 $\sum_{y}p(y|x)=1$；$f_i(x,y)$是特征函数；k为特征函数的数目；参数λ_i指示特征f_i对于模型的重要程度。研究者引入特征函数描述已知的约束条件，特征函数一般情况下是一个二值函数$f(x,y)\to\{0,1\}$，形式为：

$$f(x,y)=\begin{cases}1 & \text{如果}(x,y)\text{满足某种约束}\\ 0 & \text{否则}\end{cases} \quad (3)$$

1.3 相关研究

综上所述，最大熵模型是一种性能良好且适应性、灵活性极好的统计模型，它可以从数据中提取各种相关或不相关的特征进行综合处理，并且具有坚实的数学基础。自从文献[1]首次将它应用于自然语言处理以来，最大熵模型被广泛地应用于自然语言处理中，包括文本切分[2]、词性标注[3]、组块分析[4]、歧义消解[5]、文本分类[6-7]等。文献[6]使用词频作为特征函数的值进行文本分类的研究，并且对基于最大熵模型和简单Bayes模型的分类方法进行了比较。文献[7]使用分词和N-Gram两种中文文本特征生成方法研究文本分类，并比较了最大熵模型和Bayes、KNN、SVM等3种常用的文本分类方法。文献[8]首先从训练语料中提取一些触发词，然后采用《同义词词林(扩展版)》扩展这些触发词构建触发词表，并探讨了基于触发词表和触发词表加机器学习方法识别事件类别。本文依据最大熵模型对触发词发现的事件表述语句进行类别判定，即根据事件表述语句中的特征将，事件划分到预先定义好的类别中，而正文文本分类依据正文文本的特征将文本划分到预先定义好的类别中。事件分类和文本分类相比有如下特点：(1) 分类的文本短，大部分都是一个完整的句子，最长的也不过2～3个完整的语句；(2) 因为语句是事件表述语句，所以语句中包含的信息量大；(3) 要分类的事件表述语句已经进行了分词、词性标注及命名实体识别。本文针对这些特点所提出的基于最大熵的事件分类方法与一般的基于最大熵的文本分类方法有所不同，主要表现在以下几个方面：(1) 采用命名实体和分词相结合的特征生成方法；(2) 对触发词进行词频统计，并将统计结果也作为一类特征；(3) 融合触发词的特征和触发词上下文中的命名实体、短语等各种特征进行事件分类。

2 基于最大熵的中文文本事件分类

使用最大熵模型进行事件分类首先要建立模型，其中的两个关键步骤是参数估计和特征选择。参数估计是从训练数据集学习每一个特征的权重参数，即求解向量$\lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$的过程。而特征选择是筛选出对最大熵模型有表征意义的特征，包括特征模板构建和依据特征模板进行有效特征选择。

2.1 参数估计

传统的参数估计方法是在GIS(generalized iterative scaling algorithm)算法[9]的基础上提出的IIS(improved iterative scaling algorithm)算法[10]。本文采用IIS算法求解模型参数。

2.2 特征模板及特征选择

最大熵模型的一个主要优点是能够在同一个模型中集成不同的特征。所以，建立最大熵模型的另

一个关键步骤是针对特定的任务为模型选择合适的特征集，用简单的特征集表示复杂的语言现象，包括特征模板构建和有效特征选择。

2.2.1 特征模板

特征模板的主要功能是定义上下文中某些特定位置的语言成分或信息对某类事件的出现概率是否有影响。由于本文是根据一个候选事件表述语句确定该语句表述的事件类型，因此就由该语句中出现的语言成分及这些语言成分出现的位置确定特征集合，即要考虑该语句中的触发词以及触发词前后的词、命名实体所具有的特征，图1是可能的特征空间的图示。根据确定事件类别时影响因素的类别差异，具体定义特征空间为：(1) 触发词信息，即触发词及其词性、词频等的信息，其中触发词的词频信息用触发词的触发频和出现频两个值表征(触发频是训练集中包含该触发词且归属为特定类型事件的语句占所有包含该词的语句的比率，反映该触发词出现的语句是特定类型事件表述语句的概率；出现频是训练语料中所有该事件类型的语句中包含该触发词的比率，反映特定类型事件表述语句中该触发词出现的概率)；(2) 触发词上下文中的命名实体的信息，包括命名实体的类别、相对于触发词的位置等；(3) 句子中其他词或词组的词性标注、位置等的信息；(4) 句子中的否定词的信息，包括是否出现否定词、以及否定词是否改变事件表述语句的意义等；(5) 句子中的时态信息，该特征空间是针对事件属性中的时态设定的；(6) 事件表述语句的整体或局部的简单句法结构的信息。

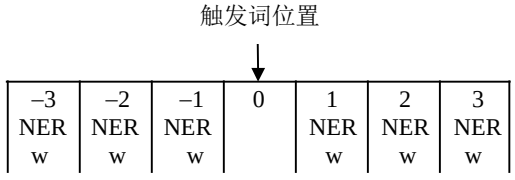

图1 可能的特征空间

根据上面给出的特征空间，本文定义了最大熵模型中的特征模板，如表1所示。由于该表中每个模板只考虑一种因素，故称之为原子模板。原子模板也可以看作是对于当前上下文的一个特征函数。当特征函数取特定值时，则该模板被实例化，得到具体的特征。当模板的取值确定后，就可以产生一个特征，该特征称为原子特征。

由于语言现象十分复杂，仅仅用原子特征不足以表示上下文中的所有特征。通过对表1中各种原子模板进行组合，构成一些复合特征模板表示更复杂的上下文环境，如表2所示。原子特征模板和各种复合特征模板共同构成模型的特征模板集。同样，对于复合特征模板，也是首先通过对各个原子模板实例化，对模板函数取值，从而产生一个特征，称为复合特征。复合特征表示为二值特征函数的形式，与原子特征相似，只是取值时需要满足的约束条件更复杂。

表1 部分原子特征模板列表

原子特征模板	特征模板含义描述
Trigger	触发词
TriggerPOSTag	触发词词性标注
TriggerFreq	触发词触发频
TriggerTermFreq	触发词出现频
NERType－1	触发词前后位置上的命名实体类型
NERType－2	
NERType－3	
NERType＋1	
NERType＋2	
NERType＋3	
NP	名词短语
PP	介词短语
PrivativeAttribute	否定词属性
Word－1	触发词前后位置上的词
Word－2	
Word－3	
Word＋1	
Word＋2	
Word＋3	
WordPOSTag－1	触发词前后词的词性标注
WordPOSTag－2	
WordPOSTag－3	
WordPOSTag＋1	
WordPOSTag＋2	
WordPOSTag＋3	
TenseAttribute	时态属性

表2 部分复合特征模板列表

复合特征模板	特征模板含义描述
NERType–1&NERType=Per	触发词前一个位置为命名实体，其类别为人(Person)
NERType+1&NERType=Per	触发词后一个位置为命名实体，其类别为人(Person)
Trigger–1 & NP	触发词前一个位置是名词短语(NP)
Trigger+1 & PP	触发词后一个位置是介词短语(PP)

2.2.2 特征选择

定义特征模板后，要对训练语料中实际出现的特征进行特征选择，因为并不是所有的特征对模型都有贡献，太多的特征会增加模型训练的时间。常用的特征选择方法有基于频次的特征选择方法(count cutoff feature selection，CCFS)[6,11]和增量特征选择方法(incremental feature selection，IFS)[1]。CCFS方法是给定一个阈值K，模型只考虑在训练集中出现的次数大于K的特征。虽然该方法实施简单，但不能

保证得到一个最小的有效特征集合。设F是一个候选特征集合，其中只有一部分特征是语言模型的有效特征，有效特征子集设为S，增量特征选择就是从候选特征集F中选取有效特征子集S的过程。方法为：开始设有效特征集S为空，然后不断地向S中增加候选特征，每次向S中增加的特征由训练数据决定。每增加一些特征需要对所有的候选特征调用IIS算法对λ重新计算，还要利用训练数据对新模型的对数似然进行计算，以判断该特征是否使模型的对数似然增量最大，实现相对困难，但能够获得有效的特征集。本文采用IFS方法选择有效特征。

3　实验结果及其分析

为验证本文提出的基于最大熵模型的事件分类方法，进行了多组实验。首先根据第2节介绍的方法在3.1小节给出的训练数据集上对用于事件分类的最大熵模型进行参数估计和特征选择，构建最大熵模型。然后用构建好的最大熵模型对测试数据集中的5类事件语句进行分类实验。

3.1　实验数据集

训练数据集借助辅助工具依据触发词从《人民日报》1995年全年的生语料中抽取出来的职务变动、会见、恐怖袭击、法庭宣判、自然灾害5类事件的27 824条语句。这些语句经过简单的人工筛选后剩下18 650条。再经过分词、词性标注、命名实体识别、事件类别划分后作为训练数据集，用于构建最大熵模型。测试数据集包含依据触发词从《人民日报》1998年1月的熟语料中抽取出来的上述5类事件的候选语句，这些语句经过命名实体识别用于验证基于最大熵模型的事件分类方法。在构建实验数据集时，使用的触发词表是手工构建的，构建时参阅了《人民日报》1995年的全年语料和《同义词词林》。

3.2　实验结果及分析

本文使用国际上常用的文本分类评价指标。准确率和召回率的盈亏平衡点(precision/recall breakeven point)为类器的微平均准确率(MicroP)。

实验过程中，分别进行如下性能比较：

(1) 只使用分词、分词+命名实体相结合两种不同语句特征生成方法时的分类器性能；(2) 选取不同数量特征时的分类器性能；(3) 最大熵模型分类和只根据触发词进行的分类的性能；(4) 与Naïve Bayes和KNN分类器性能。

在事件语句的分类阶段，本文分别使用分词、分词+命名实体相结合两种不同方法生成每一个事件语句的特征。在进行测试时也基于该两种不同的方法生成特征，对使用不同特征生成方法、不同特征数目时基于最大熵模型的分类器性能做出评价。在特征数目为50～500的情况下，对基于最大熵模型的事件分类方法进行实验的实验结果如表3所示。

表3　不同特征生成方法的性能比较

特征数/个	词特征(微平均准确率)	词和命名实体(微平均准确率)
50	0.767	0.812
100	0.819	0.876
150	0.856	0.926
200	0.872	0.945
250	0.891	0.938
300	0.901	0.924
400	0.896	0.919
500	0.881	0.913

从表3可以得出如下结论：(1) 生成语句特征使用分词+命名实体的方法，要优于只使用分词的方法；(2) 随着特征数目的增加，分类准确率提高，达到一定数目后，准确率不再升高，反而有所下降；(3) 基于分词+命名实体的语句特征生成方法可用更少的特征数达到更好的分类性能。

为了对基于最大熵模型的事件语句分类方法和其他分类方法进行比较，选择了基于触发词的分类方法、Naïve Bayes和KNN，其中，Naïve Bayes使用多项式模型，KNN方法K值取60。实验结果如表4所示，基于触发词的事件分类不受特征数目的影响。

表4　不同分类方法的性能比较

特征数/个	Bayes (MicroP)	KNN (MicroP)	触发词 (MicroP)	最大熵 (MicroP)
50	0.781	0.821	0.784	0.812
100	0.836	0.869		0.876
150	0.892	0.896		0.926
200	0.905	0.925		0.945
250	0.912	0.937		0.938
300	0.893	0.934		0.924
400	0.895	0.929		0.919
500	0.873	0.914		0.913

从表4可以得出如下结论：(1) 基于最大熵模型的事件语句分类、Naïve Bayes分类、KNN分类方法都优于基于触发词的事件分类方法。(2) 基于最大熵的事件语句分类优于Naïve Bayes分类，与KNN分类方法的性能相近。从表4还可以看出，最大熵模型较KNN分类方法在较少的特征数目下能达到最优性能，在特征数目最少的50时和特征数目多于300时，KNN分类性能要优于最大熵模型分类性能，反映了

最大熵模型对特征的增减灵敏性略高，由于KNN要计算并统计待分类事件表述语句和样本语句的距离，最大熵模型则根据已知样本构建一个能够对分类语句进行准确描述的统计模型 $p(y|x)$ 预测事件的类别，所以该差别的内在机理有待进一步研究。(3) 基于最大熵的事件语句分类效果要比基于最大熵模型的文本分类效果[7]高出一个多百分点。

4 结 论

在文本事件抽取中，发现事件表述语句并进行分类有着重要的现实意义。本文使用基于最大熵模型的方法对事件进行分类，并且就模型参数估计、特征模板的构建和特征选择、特征数目对基于最大熵模型的事件分类器的性能影响进行了实验和分析，比较了几种不同的事件分类方法。实验结果表明，使用分词和命名实体相结合的特征，利用最大熵分类器对事件表述语句进行事件分类有较好的效果，比单纯使用触发词确定事件类别及子类别的进行准确率要高很多。今后将扩大所处理的事件类型和子类型的规模，进行更大范围的实验和研究。

参 考 文 献

[1] BERGER A L, DELLA P S A, DELLA P V J. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1): 39-71.

[2] BEEFERMAN D, BERGER A, LAFFERTY J. Statistical models for text segmentation[J]. Machine Learning, 1999, 34(1-3): 177-210.

[3] 赵 岩, 王晓龙, 刘秉权, 等. 融合聚类触发对特征的最大熵词性标注模型[J]. 计算机研究与发展, 2006, 43(2): 268-274.
ZHAO Yan, WANG Xiao-long, LIU Bing-quan, et al. Fusion of clustering trigger-pair features for POS tagging based on maximum entropy model[J]. Journal of Computer Research and Development, 2006, 43(2): 268-274.

[4] 李素建, 刘 群, 杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报, 2003, 26(12): 1722-1727.
LI Su-jian, LIU Qun, YANG Zhi-feng. Chunk parsing with maximum entropy principle[J]. Chinese Journal of Computers, 2003, 26(12): 1722-1727.

[5] 张 锋, 樊孝忠. 基于最大熵模型的交集型切分歧义消解[J]. 北京理工大学学报, 2005, 25(7): 590-593.
ZHANG Feng, FAN Xiao-zhong. Resolution of overlapping ambiguity strings based on maximum entropy model[J]. Transactions of Beijing Institute of Technology, 2005, 25(7): 590-593.

[6] NIGAM K, LAFFERTY J, MCCALLUM A. Using maximum entropy for text classification[C]//Proceedings of the IJCAI99 Workshop on Information Filtering. Stockholm, Sweden: MII Press, 1999.

[7] 李荣陆, 王建会, 陈晓云, 等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展, 2005, 42(1): 94-101.
LI Rong-lu, WANG Jian-hui, CHEN Xiao-yun, et al. Using maximum entropy model for Chinese text categorization[J]. Journal of Computer Research and Development, 2005, 42(1): 94-101.

[8] 赵妍妍, 王啸吟, 秦 兵, 等. 中文事件抽取中事件类别的自动识别[C]//第三届学生计算语言学研讨会. 北京: 清华大学出版社, 2006: 240-245.
ZHAO Yan-yan, WANG Xiao-yin, QIN Bing, et al. Automatic event type extraction in Chinese event extraction[C]//Proceedings of the 3rd Students Workshop of Computational Linguistics. Beijing: Tsinghua Univesity Press, 2006: 240-245.

[9] DARROCH J N, RATCLIFF D. Generalized iterative scaling for log-linear models[J]. Annals of Mathematical Statistics, 1972, 43(5): 1470-1480.

[10] DELLA P S, DELLA P V, LAFFERTY J. Inducing features of random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(4): 380-393.

[11] RATNAPARKHI A. Maximum entropy models for natural language ambiguity resolution[D]. Ponnsy lvania: University of Pennsylvania, 1998.

编 辑 蒋 晓