第 38 卷 第 4 期
Vol.38 No.4

计 算 机 工 程
Computer Engineering

2012 年 2 月
February 2012

• 软件技术与数据库 •　　文章编号：1000—3428(2012)04—0037—03　　文献标识码：A　　中图分类号：N945

基于最大熵的句内时间关系识别

王风娥，谭红叶，钱揖丽

(山西大学计算机与信息技术学院，太原 030006)

摘　要：分别对句内事件-时间对关系以及事件对之间的时间关系识别进行研究。分析影响时间关系识别的语言特征，如时间关系对之间的依存关系序列、间隔词数、信号词及其位置等，并使用基于最大熵的方法进行识别。实验结果表明，运用该方法获得的事件-时间对关系识别准确率为 87.83%，事件对之间的时间关系识别准确率为 80.79%。

关键词：时间关系；句内时间关系；最大熵；依存分析；自然语言处理

Recognition of Temporal Relation in One Sentence Based on Maximum Entropy

WANG Feng-e, TAN Hong-ye, QIAN Yi-li

(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

【Abstract】This article studies the recognition of the temporal relation between an event and a time expression and the temporal relation between two events where one event syntactically dominates the other event in one sentence. It analyzes some effective linguistic features such as dependency parsing information, relative positions, signal words, position of signal words and so on. A method based on maximum entropy model is proposed. In addition, how linguistic features could affect temporal relation recognition is analyzed. Experimental results show the accuracies of the two tasks are respectively 87.83% and 80.79%.

【Key words】temporal relation; temporal relation in one sentence; maximum entropy; dependency analysis; natural language processing

DOI: 10.3969/j.issn.1000-3428.2012.04.012

1　概述

时间关系是文本中事件之间、时间之间、事件和时间之间的时序关系。近年来，文本中时间关系的识别已经成为自然语言处理的重要任务之一。相关研究人员把时间关系识别列为一项重要评测任务，分别为 TempEval-1 和 TempEval-2。主要识别方案：TICTAC 系统[1]利用句子级句法树和自底向上地句法分析方法生成时间关系，实验表明识别句内事件-时间对和事件对的准确率分别为 64%、66%；CU-TMP 系统[2]用句法和语义特征表示事件-时间对并训练了 SVM 模型，识别句内事件-时间对准确率达到了 61%；NCUS-INDI 系统[3]采用隐马尔可夫逻辑和词汇关系相结合的机器学习技术，结果显示识别句内事件-时间对、事件对之间时间关系的准确率分别为 63%、66%。

在中文文本中，已有一些学者从不同角度对时间关系展开了研究。文献[4]采用基于转换的错误驱动学习方法识别事件和时间对的时间关系，结果显示加入转换规则集比只使用最小距离原则的错误率降低 9.74%；文献[5]分析影响多子句中事件之间时间关系的语言特征，采用基于机器学习和 NBC、PDT 异构的 CB 计算模型，识别事件对时间关系的准确率为 78%；文献[6]抽取蕴含于语法语义层面的时间关系，并通过挖掘隐含的时间关系，制定一系列与文本无关的时间关系推理规则。

本文把时间关系识别任务看成成对分类任务，抽取影响时间关系的语言特征，采用最大熵分类器对句内时间关系进行分类。

2　时间关系问题分析

文献[2]提出 6 项评测任务，包括时间和事件表达式的识别、4 项时间关系识别任务，即句内事件-时间对的关系，事件对之间的关系，事件和文档创建时间之间的关系以及相邻句中主要事件之间的关系，并提出 6 种时间关系类别：before，after，overlap，before-or-overlap，overlap-or-after，vague。因为在语料库中最大的不一致存在于 before and overlap 和 after and overlap 之间，所以用 before-or-overlap 和 after-or-overlap 来表示标注有歧义的类型，剩余的不一致则使用 vague 表示。

本文主要针对 TempEval-2 中的 2 个评测任务：句内事件-时间对时间关系和事件对时间关系的识别进行研究。该任务可形式化描述为：设有事件-时间对$<e_i,t_j>\in E\times T$，事件对$<e_i,e_j>\in E\times E$，时间关系类别 $r_k\in R$，其中，$T=\{t_1,t_2,\cdots,t_m\}$表示预先识别的时间表达式集合，$E=\{e_1,e_2,\cdots,e_n\}$表示预先识别的事件表达式集合，$R=\{r_1,r_2,\cdots,r_l\}$表示预先定义的时间关系集合。时间关系识别就是为每一对$<e_i,t_j>$或$<e_i,e_j>$分配一个时间关系类别标记 r_k[7]。

丰富多样的汉语表述给时间关系识别增加了难度，主要难点有句法结构相同，但由于词语含义不同，时间关系也不同，同一句话内包含多个事件和时间表达时，时间关系难以确定，一些句子成分缺省等情况。

基金项目：国家自然科学基金资助项目(61100138, 61005053)；山西省高校科技开发基金资助项目(20091001)；山西省自然科学基金资助项目(2011011016-2)

作者简介：王风娥(1981－)，女，硕士研究生，主研方向：中文信息处理；谭红叶、钱揖丽，副教授、博士

收稿日期：2011-08-17　**E-mail：**hytan_2006@126.com

3 句内时间关系识别

本节中具体介绍影响事件-时间对之间时间关系和事件对之间时间关系的语言特征及基于最大熵方法的时间关系识别过程。

3.1 语言特征

文中所选取的语言特征包括 3 类：与事件，与时间，与时间关系相关的特征。

3.1.1 与事件、时间相关的语言特征

与事件、时间相关的语言特征使用了训练语料中的事件和时间的标注属性及取值，并用粗糙集正域约简方法进行约简，保留了对中文时间关系识别有益的属性，其中保留的事件属性有 Aspect、Class、Pos，时间属性有 Type，具体取值及意义可以查看 2010 年 TempEval-2 事件和时间的标注大纲。

3.1.2 与时间关系相关的语言特征

假设待分析的时间关系是事件 e_i 和时间 t_j 之间或事件 e_i 和事件 e_j 之间的时间关系。时间关系相关的特征主要有：

(1)Is-dependency：事件-时间对或事件对之间存在的依存关系。具体取值及意义如下：

1)-Link：以 t_j(或 e_j)为触发词，通过依存关系查找 e_i 的关系链。

2)Link：以 e_i 为触发词，通过依存关系查找 t_j(或 e_j)的关系链。

(2)Signal-position：信号词与<e_i,t_j>词对或<e_i,e_j>词对的相对位置。把能使文本中 2 个实体(时间和事件，时间和时间或事件和事件)之间关系清晰的文本元素称为信号词。在中文文本中，通常用一些时间方位词来转换时间关系。具体取值及其意义为：

1)Signal-left、Signal-mid、Signal-right：分别表示信号词与词对在同一子句中，且出现在词对左侧、中间或右侧。

2)Signal-e_i-left、Signal-e_i-right：分别表示信号词仅与 e_i 在同一子句中，且出现在 e_i 左侧或右侧。

3)Signal-e_j/t_j-left、Signal-e_j/t_j-right：分别表示信号词仅与 t_j(或 e_j)在同一子句中，且出现在 t_j(或 e_j)左侧或右侧。

(3)Pair-distance：e_i 与 t_j 或 e_j 中间的间隔词数。具体取值及其意义为：

1)0：e_i 和 t_j(或 e_j)不在同一子句。

2)-i、i：e_i 和 t_j(或 e_j)在同一子句，且 t_j(或 e_j)是 e_i 左侧或右侧第 i 个词。

3.2 基于最大熵方法的时间关系识别

最大熵原理是指，预测一个随机事件的概率分布时，预测应该满足全部已知的条件，而对未知的情况不做任何主观假设。在这种情况下，概率分布最均匀，预测的风险最小[8]。

本文构建的最大熵分类器完全基于最大熵原理，不额外增加任何在训练数据中没有观察到的信息。本文中时间关系的具体识别思想为：首先将中文文本进行词性标注和依存句法分析，获得信号词词典和依存关系，然后提取语言特征，把得到的文本按 4：1 的比例分成训练文本和测试文本，用训练文本训练最大熵分类器，输入测试文本进行分类，最后输出时间关系。时间关系识别的具体过程如图 1 所示。

信号词词典的建立。信号词词典由与时间词连用的方位名词组成。如“本世纪/nt 末/nd”，“末”被标注为方位名词，则把“末”加入到信号词词典中。

提取语言特征。由于篇幅有限这里仅列出 Is-dependency 的提取思想，下面通过例 1 说明依存关系提取的思想，例 1 的依存关系如图 2 所示。

例 1 [今年]t9 以来，国际 跨国公司 [继续]e32 踊跃 [投资]e33。

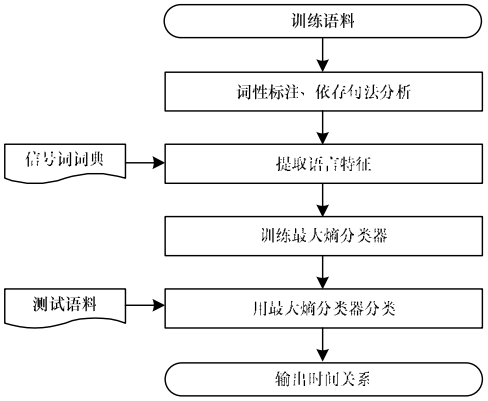

图 1 时间关系识别过程

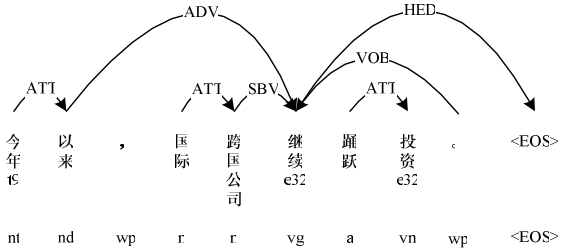

图 2 例 1 的依存关系

这里分别以时间词“今年”、事件词“继续”和“投资”为触发词，通过依存关系递归查找事件词。以“今天”为触发词，通过依存关系 ATT-ADV 和 ATT-ADV-VOB 可以找到“继续”和“投资”，即：<e_{32}, t_9>、<e_{33}, t_9>的 Is-dependency 值为-ATT-ADV 和-ATT-ADV-VOB。以“继续”为触发词，找不到“投资”，但以“投资”为触发词，通过依存关系 VOB 可以找到“继续”，则<e_{32}, e_{33}>、<e_{33}, e_{32}>的 Is-dependency 的值为-VOB 和 VOB。

4 实验结果与分析

实验数据选取的是 2010 年 TempEval-2 提供的中文训练语料，包括 1 274 对事件-时间和 1 094 对事件。实验数据准备工作使用哈尔滨工业大学信息检索实验室的依存分析工具得到词性标注和依存关系。实验过程中使用张乐博士写的最大熵工具包。

为分析本文中修改和补充的事件类别、依存关系、间隔词数和信号词位置等语言特征对时间关系的影响，设计 6 个实验，每个实验选用不同的语言特征。实验把语料平均分成 5 份，依次选择每一份为测试语料，其他 4 份为训练语料，进行 5 倍交叉验证，得到不同的分类准确率，由于篇幅有限，这里仅列出了各项测试的平均准确率，如表 1 所示，其中，#语言特征表示不包含该语言特征；*表示包含全部语言特征。

表 1 句内时间关系识别结果

序号	语言特征	<e_i,t_j>的准确率/(%)	<e_i,e_j>的准确率/(%)
1	*	87.83	80.79
2	#Class&#Is-dependency	87.44	74.67
3	#Class	87.36	74.77
4	#Is-dependency	88.46	80.14
5	#Pair-distance	87.12	80.71
6	#Signal-position	88.01	78.09

可以看出，事件类别对于句内时间关系识别，尤其是对

于事件对的时间关系识别有积极影响，分别对比 1、3 两组和 2、4 两组实验结果可以看出，加入事件类别后准确率分别提高 6.02%和 5.47%；加入词对间隔词数也会提高时间关系识别的准确率；另外，加入依存关系和信号词位置这 2 个语言特征后，事件-时间对之间时间关系识别准确率会略有下降，事件对之间时间关系识别准确率会略有提高。

分析该实验得到的错误的识别结果发现，导致错误的原因主要有以下 3 个方面：(1)由于句子中省略了部分成分使得到的依存关系并不准确；(2)同一个词在不同语言环境中被标注不同的语言特征值，事件的语义信息很多情况下，还要通过词本身含义体现；(3)本文实验采用的信号词词表中只考虑了时间方位词，没有考虑连接时间的符号以及一些与时间词连用的动词如“截止”，使时间关系发生变化的情况。

从以上实验结果可知，在中文语料中，依据所选语言特征，用最大熵分类方法识别句内时间关系是较为有效的。2010 年 TempEval-2 提供的中英文语料中句内时间关系分布比例差距较大，如事件-时间对的样例中 overlap 占到了近 90%，而 vague 为 0，英文语料中 overlap 仅不到 54%，而 vague 也占有 2%，因此，中文语料中时间关系的识别较英文语料更容易。

5　结束语

本文提出一种依据语言特征识别时间关系的最大熵分类方法，并尝试加入依存关系、间隔词数等语言特征，较有效地解决了句内时间关系识别问题。实验结果表明，事件类别、间隔词数对于时间关系识别效率的提高有一定帮助。本文工作的不足之处在于，有些依存关系分析并不准确，没有考虑事件词、时间词及信号词的含义，选取的信号词有限。下一步工作将准确分析依存关系，健全信号词表，并继续分析研究事件词、时间词及信号词含义对时间关系的影响。

参考文献

[1] Georgiana P. Temporal Relation Identification by Syntactico Semantic Analysis[C]//Proc. of the 4th Int’l Workshop on Semantic Evaluation. Prague, Czech Republic: [s. n.], 2007.

[2] Steven B. Temporal Relation Classification Using Syntactic and Semantic Features[C]//Proc. of the 4th Int’l Workshop on Semantic Evaluations. Prague, Czech Republic: [s. n.], 2007.

[3] Eun H, Alok B. Modeling Temporal Relations with Markov Logic and Lexical Ontology[C]//Proc. of the 5th Int’l Workshop on Semantic Evaluation. Uppsala, Sweden: [s. n.], 2010.

[4] 王　昀, 苑春法. 基于转换的时间—事件关系映射[J]. 中文信息学报, 2004, 18(4): 23-30.

[5] Li Wenjie, Wong Kam-Fai, Cao Guihong, et al. Applying Machine Learning to Chinese Temporal Relation Resolution[C]//Proc. of the 42nd Annual Meetings of Association for Computational Linguistics. New York, USA: [s. n.], 2004.

[6] 林　静, 苑春法. 汉语时间关系抽取与计算[J]. 中文信息学报, 2009, 23(5): 62-67.

[7] 谭红叶, 郑家恒, 梁吉业. 时间关系识别研究进展[J]. 中文信息学报, 2011, 25(5): 44-52.

[8] 张仰森. 基于最大熵模型的汉语词义消歧与标注方法[J]. 计算机工程, 2009, 35(18): 15-18.

编辑　陈　文

(上接第 36 页)

测试条件步骤：

(1)首先连接好整个诊断系统，确保被诊断 ECU 上电，确保诊断上位机启动。在诊断上位机界面输入被诊断 ECU 的 ID，双击左侧树形控件执行相应诊断服务。

(2)双击上位机的编程模式服务请求，确保 ECU 处于编程模式。

(3)选取安全等级为 1 的安全访问服务进行 ECU 解锁。

(4)在 ECU 解锁状态下进行读取数据服务。

5.2　测试结果分析

本文所测试的诊断服务执行过程如图 5 所开发的诊断工具上位机界面所示[5]，上位机发送编程模式请求(10 02)，ECU 回复会话模式跳转正定响应进入编程模式。上位机发送请求种子服务(27 01)，ECU 响应 2 Byte 随机种子(bf b1)，上位机根据安全访问算法计算密钥(b1 bf)并发给 ECU，ECU 发送正定响应(67 02)，表示已解锁。上位机进行读数据服务，数据标识符为 f1 8c，由通信报文可见，此项服务为多帧传输，ECU 首先发送第 1 帧(10 07 62 f1 8c 11 11 11)，上位机接收到第 1 帧后发送流控帧，ECU 再继续发送后续帧。

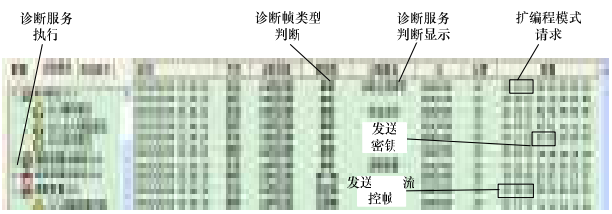

图 5　本文开发的诊断工具报文分析界面

由上所述测试结果分析可知，本文所设计整个诊断系统测试结果符合预期要求，所测试诊断服务按诊断协议执行，通信规则符合 15765-2 单帧及多帧传输的传输规则。

6　结束语

本文通过研究分析车载网络国际诊断标准 IS015765，根据车载网络诊断结构，设计了基于 ISO15765 的车载网络诊断。开发的基于低成本 USBCANII 的诊断工具可实现对被诊断 ECU 实现基于 ISO15765 的诊断通信。同时，在开发支持诊断功能的 ECU 阶段，可实现诊断服务的执行与响应测试，为支持诊断功能 ECU 的开发商提供一种低成本、方便灵活的诊断测试工具。通过实际测试车载网络中的节点验证了所开发系统的可行性与可靠性，对基于 ISO15765 的车载网络诊断的后续设计与开发具有一定参考价值。

参考文献

[1] 张　宏, 詹德凯, 林长加. 基于 CAN 总线的汽车故障诊断系统研究与设计[J]. 汽车工程, 2008, 30(10): 934-937.

[2] 马　英, 阴晓峰, 张德旺. 基于 CAN 的汽车电控系统故障诊断技术[C]//2008 年中国汽车工程学会年会论文集. 北京: 机械工业出版社, 2008.

[3] ISO. ISO 15765-2004 Road Vehicles——Diagnostics on Controller Area Networks(CAN)[S]. 2004.

[4] 蒋建春, 陈洪霞, 郑太雄. 基于 CCP 的 ECU 在线编程技术的实现[J]. 计算机工程, 2011, 37(5): 241-243.

[5] Li Renjun, Liu Chu, Luo Feng. A Design for Automotive CAN Bus Monitoring System[C]//Proc. of VPPC’08. Harbin, China: [s. n.], 2008.

编辑　索书志