

基于 CRF 的汉语语块分析和事件描述小句识别

王昕, 王金勇, 刘春阳, 王奇, 付春元

黑龙江大学计算机科学技术学院

哈尔滨市学府路 74 号 (150080)

[hxw@hlju.edu.cn]

CRF-based Chinese Chunking and Event Recognition

Wang Xi, Wang Jinyong, Liu Chunyang, Wang Qi, and Fu Chunyuan

School of Computer Science and Technology, Heilongjiang University

No.74 Xuefu Road, Harbin 150080 China

(hxw@hlju.edu.cn)

摘要

我们参加了 CIPS-ParsEval-2009 测试中的三个子任务即汉语基本块分析、汉语功能块分析和事件描述小句识别的封闭测试。为方便, 我们将它们都看作是序列标注问题, 并在条件随机域模型框架下采用不同的特征组合分别加以解决。实验结果表明: 条件随机域模型在汉语基本块和功能块两个任务中取得了较好的效果。

关键词: 汉语基本块; 汉语功能块; 事件描述小句识别; 条件随机场模型

Abstract: We formalize the three sub-tasks in CIPS-ParsEval-2009, namely Chinese base chunk parsing, Chinese functional chunk parsing and Chinese event descriptive clause recognition, as sequence labeling problems, and incorporate different sets of features under the framework of Conditional Random Fields (CRFs) to resolve these problems, respectively. The CRF-based approach shows promising performance in both Chinese base chunk parsing and functional chunk parsing.

Keywords: Chinese chunking, Chinese event-clause recognition, conditional random fields

1 引言

作为自然语言处理(natural language processing, NLP)领域的一个核心问题, 高性能的句法分析在信息抽取、信息检索、机器翻译和文本挖掘等多个 NLP 应用领域发挥了重要的作用。2009 年中文信息学会句法评测(CIPS-ParEval-2009)包含五个子任务, 分别是汉语词性标注处理、汉语基本块分析、汉语功能块分析、事件描述小句识别、句法结构树分析。我们只参加其中的汉语基本块分析、汉语功能块分析、事件描述小句识别等三个子任务。为方便, 我们将这些任务都看作序列标注问题, 并在条件随机域(Conditional Random Fields, CRFs)模型框架融合不同的特征组合并分别构建了三个序列标注系统来完成上述三个字任务。本文将简要介绍关于这些任务的形式表述、特征选择以及相应的测试结果。

2 条件随机域

作为一个在给定输入节点值时计算输出节点值的条件概率的无向图模型(Lafferty 等, 2001), CRFs 既不像隐马尔科夫模型依赖于独立假设, 又在最大熵的基础上解决了标记偏置问题。因此, CRFs 目前广泛应用于各种序列标注问题中(Sha 和 Pereira, 2003; Ratinov 和 Roth, 2009)。

对于输入序列 $\vec{x}=(x_1, x_2, \dots, x_n)$ 和输出标记序列 $\vec{y}=(y_1, y_2, \dots, y_n)$ ，可以定义一个线性的 CRF 模型，即：

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \mu_k g_k(y_i, x)\right)$$

3 任务的形式化和语块的表示

为了利用现有的工具快速构建相应的分析系统，我们将三个子任务统一形式化为序列标注问题。不同的语块表示方案对语块分析性能影响很大(Ratinov 和 Roth, 2009)。为了在序列中表示基本组块和功能组块，我们采用 BILOU 方案，因为 Ratinov 和 Roth(2009)证明采用 BILOU 的短语组块分析性能一般优于广泛采用的 BIO 方案。相反，在事件小句(event-clause)表示中我们则采用 BIO 方案，因为在训练数据上进行的初步试验表明 BIO 的性能优于 BILOU。如表 1 所示，BILOU 方案定义了五个标记来表示词语构成组块时可能采用的不同模式。

表 1 BILOU 组块标记集

| 标记 | 定义 |
|----|-----------------|
| B | 多词组块的首词 |
| I | 多词组块的中间词 |
| L | 多词组块中的尾词 |
| U | 单词组块 |
| O | 组块之外的词(不属于任何组块) |

利用 BILOU 和 BIO 标记，CIPS-ParseEval-2009 测试方案中的组块标注和事件表述小句标注数据均可等价表示成词语和相应的组块标记序列。图 1 和图 2 分别给出了基本组块的 BILOU 表示和事件描述小句的 BIO 表示。注意在 CIPS-ParseEval-2009 定义的事件描述小句对应的序列表示中，每个词的标记只包含模式信息，并不包含相应的事件类型信息；但在基本组块和功能组块序列表示中，每个词的组块标记则两种信息都应包含。

4 特征的选择

考虑到训练的效率，同时参照针对通过对训练语料的封闭实验结果，我们选择了词、词性及其位置作为三个子任务的基本特征，如表 2 所示。

在表 2 的特征表示中，脚标代表当前位置的偏移量，如 w_0 为当前位置词， p_0 为当前词性， w_{-1} 为前一个位置的词。

值得注意的是表 1 给出的是三个任务缺省的通用模版，根据任务的不同我们还添加了其它的附加特征，具体说明如下：

- **基本组块分析：**根据 CIPS-ParseEval-2009 测试方案，汉语基本块主要描述句子中直接相邻的、以名词、动词、形容词等实词为中心聚合形成具有特定语义内容的词语序列，其中一般不包括各种功能词，包括连词、叹词、语气词、助词、标点符号等。为了提高基本组块分析的性能，我们先将词语的成分标记作为 unigram 特征加入模版中，继续预测词语关系标记并将结果 unigram 加入的模版中；最后以词语、词性、词语的关系标记和词语的成分标记作为特征来预测基本块成分标记-关系标记组合。
- **功能组块分析：**根据 CIPS-ParseEval-2009 测试方案，汉语功能块主要描述句子中反映不同事件内容的基本信息单元。他们一般占据了句子中的主语、谓语、宾语、状语、定语、中心语等功能位置，通过组合形成不同的事件句式完成对真实世界的不同事件内容的再现描述。为了提高功能组块分析的性能，我们将基本特征模版并结合组块成员词的位置标注信息构造 CRFs 模型。

输入：执法/vN 部门/n 是/vC 反/v 腐败/a 斗争/vN 、/wD 搞好/v 廉政/vN 建设/vN 的 /uJDE 重点/n 部门/n 之一/rN 。/wE

基本组块分析结果：[np-ZX 执法/vN 部门 /n] [vp-SG 是/vC] [np-ZX 反/v 腐败/a 斗争 /vN] 、/wD [vp-SG 搞好/v] [np-ZX 廉政 /vN 建设/vN] 的/uJDE [np-ZX 重点/n 部门 /n] [np-SG 之一/rN] 。/wE

基本组块的 BILOU 表示：执法/vn/B-np-ZX 部门/n/L-np-ZX 是/vC/U-vp-SG 反/v/B-np-ZX 腐败/a/I-np-ZX 斗争/vN/L- np-ZX 、 /O-wD 搞好/v/U- vp-SG 廉政/vN/B-np-ZX 建设/vN/L-np-ZX 的/O-uJDE 重点/n/B-np-ZX 部门/n/L-np-ZX 之一/rN/U- np-SG 。 /O-wE

图 1 基本组块的 BILOU 表示

输入：执法/vN 部门/n 是/vC 反/v 腐败/a 斗争/vN 、/wD 搞好/v 廉政/vN 建设/vN 的 /uJDE 重点/n 部门/n 之一/rN 。/wE

事件描述小句识别结果：[执法/vN 部门/n 是/vC 反/v 腐败/a 斗争/vN 、/wD 搞好/v 廉政/vN 建设/vN 的/uJDE 重点/n 部门/n 之一/rN] 。/wE

事件描述小句的 BIO 表示：执法/vN/B 部门/n/I 是/vC/I 反/v/I 腐败/a/I 斗争/vN/I 、 /wD/I 搞好/v/I 廉政/vN/I 建设/vN/I 的 /uJDE/I 重点/n/I 部门/n/I 之一/rN /I 。 /wE/O

图 2 事件描述小句的 BIO 表示

表 2 基本特征模版

| |
|--|
| 1.词 unigram 特征: $w_0, w_1, w_{-1}, w_2, w_{-2}$ |
| 2.词 bigram 特征: $w_{-1}w_0, w_0w_1$ |
| 3.词性 unigram 特征: $p_0, p_1, p_{-1}, p_2, p_{-2}$ |
| 4.词性 bigram 特征: $p_{-1}p_0, p_0p_1$ |
| 5.词性 trigram 特征: $p_{-2}p_{-1}p_0, p_{-1}p_0p_1$ |

- **事件描述小句识别：**事件描述小句识别是从完整的句子中识别出各个事件描述单元，它们一般具有以下特征：(1)以逗号、分号、句号、问号等点号分隔而形成的词语序列；(2)

内部包含完整的主、状、谓、宾等描述结构，考虑到各种省略情况，其中至少应包含一个谓语块。通过这阶段处理，可以把汉语中复杂的流水复句等描述形式转化为一个个基本的事件描述单元，包括各个小句或句首状语成分，便于后续分析处理。为了提高功能组块分析的性能，我们将基本特征模版结合构成事件描述小句的所有成员词的位置标注信息(即 BIO 标记信息)来构造 CRFs 模型。

5 测试结果

根据上面对任务的形式化描述和特征选择，我们利用CRF++¹工具包分别构建三个系统分别完成基本块分析、功能块分析和事件描述小句识别三个子任务。在三个系统的训练均只采用CIPS-ParseEval-2009所提供的训练语料，没有使用任何其它数据。评价指标使用传统的准确率(Precision)、召回率(Recall)和F1-measure。测试结果如表3所示。

表3 测试结果

| 任务 | | F1 | Rank |
|----------------|------------------------|-------|------|
| Task2: 基本块分析 | boundary+type | 87.93 | 2 |
| | boundary+type+relation | 86.52 | |
| Task3:功能块分析 | | 85.90 | 1 |
| Task4:事件描述小句识别 | | 69.08 | 5 |

6 结论

我们把汉语基本组块分析、功能组块分析和事件描述小句识别三个任务转化为序列标注问题，并采用 CRF++工具包分别构建了处理这三个问题的系统。CIPS-ParseEval-2009 测试结果表明，CRFs 在基本组块和功能组块测试任务中取得了比较满意的结果，但在事件描述小句识别测试结果还不尽如人意。在将来的工作我们将挖掘更多的特征，比如事件相关的实词等特征来改进事件描述小句识别性能。

致谢

本次任务是在韩习武和付国宏二位老师的指导下完成的，并得到国家自然科学基金(编号：60773069、60873169)和哈尔滨市科技创新人才研究专项资金项目(留学回国人员)(编号：2009RFLXG007)资助。

参考文献

- F. Sha and F. Pereira. 2003. *Shallow parsing with conditional random fields*. Proc. of HLT/NAACL 2003.
- J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Proc. of ICML, pp.282-289.
- L. Ratinov, and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. Proc. of CoNLL'09, pp.147-155.

¹ <http://crfpp.sourceforge.net/>