面向中学数学知识图谱构建的LLA-CRF数学术语抽取方法

华鑫，钟茂生*，刘淦林，吴东琦
（江西师范大学计算机与信息工程学院，江西 南昌 330022）

报告人 ： 华鑫

- 介绍
- LLA-CRF模型
- 实验
- 结束语

1 介绍

■ 任务定义

术语： 专业领域中概念的语言指称

术语抽取： 从专业领域文本中识别或抽取领域术语词

■ 应用

机器翻译、信息检索、文本分类……

■ 我们的应用目标

中学数学知识图谱的构建

在线智慧教育学习资源个性化推荐

提供知识库基础

1 介绍

■ 抽取方法

- 规则与统计混合

  C-value、互信息、词频分布变化等

  大量无关高频词、单字低频术语不敏感

- 机器学习方法

  SVM、CRF等

  适合的术语特征的选择将是一大难点

- 神经网络模型

  BiLSTM-CRF、BiGRU-CRF等

1 介绍

■ 中学数学术语特点

- 单字术语

  如和、差、商、项、面等

  易歧义 例：一方面(非术语)、线动成面(术语)

- 嵌套术语

  如直角三角形、等腰直角三角形等

  嵌套复杂；应保持完整，不可拆分

1 介绍

■ 中学数学术语抽取

单字、嵌套等复杂多样的特性，启发于NER任务的解决方式，将中学数学术语的抽取转化为序列标注任务。

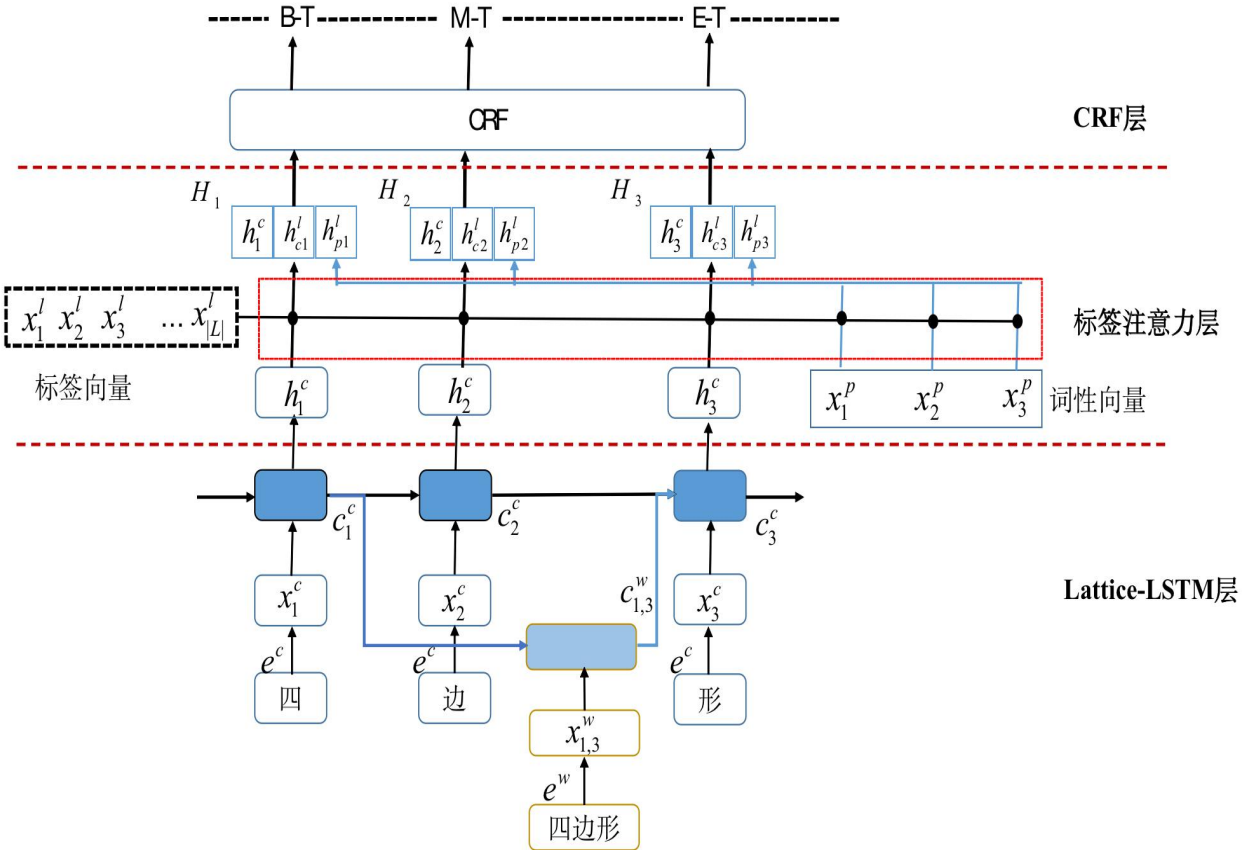
将Lattice-LSTM模型作为基线模型，额外考虑词性特征，引入对标签的注意力机制，提出面向中学数学领域的术语抽取模型Lattice-Label Attention-CRF(LLA-CRF)

Zhang Y, Yang J. Chinese NER using lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics, 2018:1554-1564.

Cui L, Zhang Y. Hierarchically-Refined Label Attention Network for Sequence Labeling[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China: Association for Computational Linguistics,2019:4106-4119.

2 LLA-CRF模型

- Lattice层
- 标签注意力层
- CRF层



2 LLA-CRF模型

■ Lattice层

- 字级层

$$f_j^c = \sigma(W_f[h_{j-1}^c, x_j^c] + b_f)$$

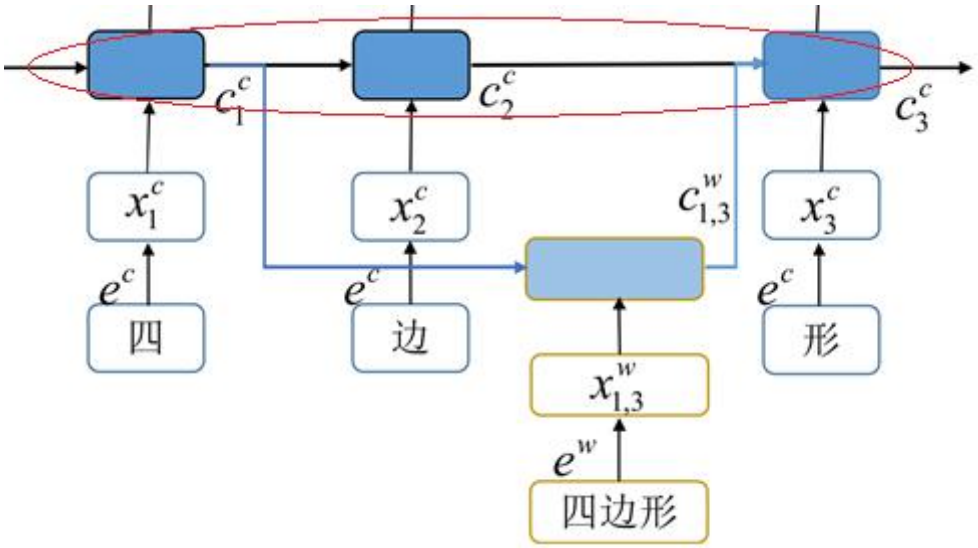
$$i_j^c = \sigma(W_i[h_{j-1}^c, x_j^c] + b_i)$$

$$o_j^c = \sigma(W_o[h_{j-1}^c, x_j^c] + b_o)$$

$$\tilde{c}_j^c = \tanh(W_{\tilde{c}}[h_{j-1}^c, x_j^c] + b_{\tilde{c}})$$

$$c_j^c = f_j^c \odot c_{j-1}^c + i_j^c \odot \tilde{c}_j^c$$

$$h_j^c = o_j^c \odot \tanh(c_j^c)$$



既能学习到字信息也能获取词信息，从而能够获取更多的上下文序列信息

2 LLA-CRF模型

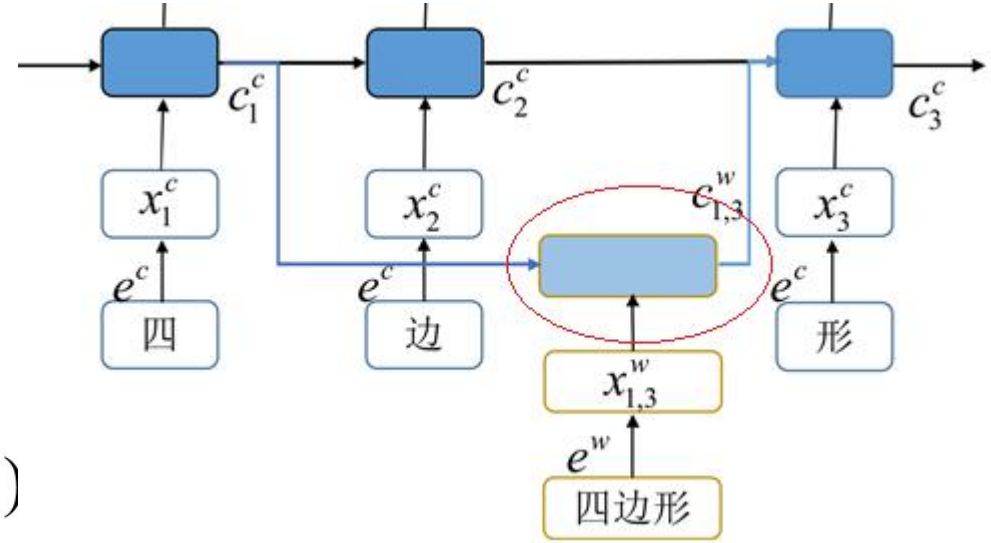
- Lattice层
  - 词级层

$$f_{b,e}^{w} = \sigma(W_f[h_b^c, x_{b,e}^w] + b_f)$$

$$i_{b,e}^{w} = \sigma(W_i[h_b^c, x_{b,e}^w] + b_i)$$

$$\tilde{c}_{b,e}^{w} = \tanh(W_{\tilde{c}}[h_b^c, x_{b,e}^w] + b_{\tilde{c}})$$

$$c_{b,e}^{w} = f_{b,e}^{w} \odot c_b^c + i_{b,e}^{w} \odot \tilde{c}_{b,e}^{w}$$



2 LLA-CRF模型

■ Lattice层

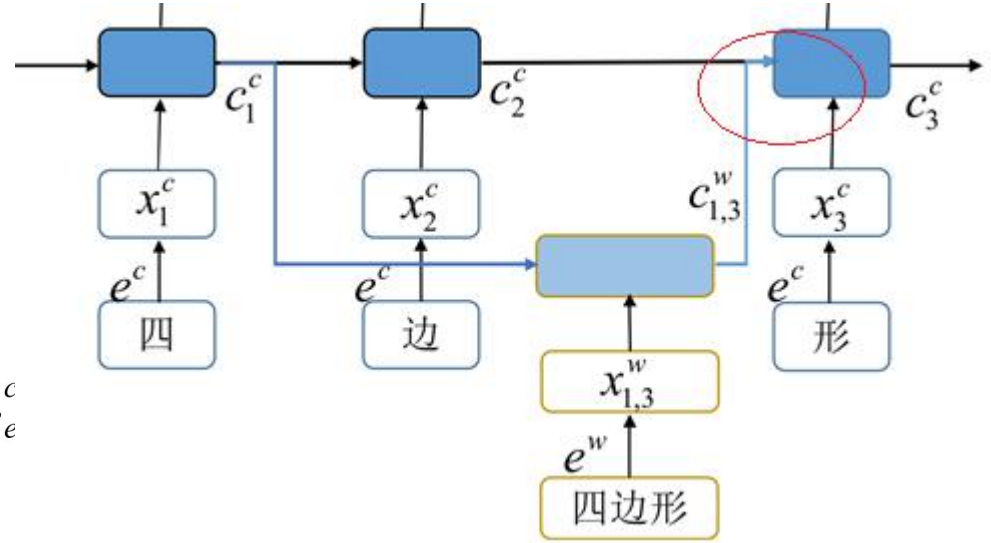
- 字、词信息融合

$$i_{b,e}^{c} = \sigma(W_l[x_e^c, c_{b,e}^w] + b_l)$$

$$c_e^c = \sum_{b\in\{b'|w_{b',e}^d \in D\}} \alpha_{b,e}^c \odot c_{b,e}^w + \alpha_e^c \odot \tilde{c}_e^c$$

$$\alpha_{b,e}^c = \frac{\exp(i_{b,e}^c)}{\exp(i_e^c) + \sum_{b'\in\{b''|w_{b'',e}^d \in D\}} \exp(i_{b',e}^c)}$$

$$\alpha_e^c = \frac{\exp(i_e^c)}{\exp(i_e^c) + \sum_{b'\in\{b''|w_{b'',e}^d \in D\}} \exp(i_{b',e}^c)}$$



2 LLA-CRF模型

■ 标签注意力层

- 标签嵌入

给定预测标签序列 $L = \{l_1, l_2, ..., l_{|L|}\}$标签映射成标签向量:

$$x_k^l = e^l(l_k)$$

标签嵌入为随机生成，在模型训练过程中自动调优。

2 LLA-CRF模型

■ 标签注意力层

- 词性嵌入

词性为一个词语的词性，因模型为字单位标签预测，将词语的词性分割为单个字的词性。如“三角形”词性为“n”,对应分割为单字对应的“n-l”，“n-m”，“n-r”,最终词性序列为 $s^p = \{c_1^p, c_2^p, ..., c_n^p\}$

$$x_j^p = e^p(c_j^p)$$

词性嵌入由Word2Vec模型生成。

2 LLA-CRF模型

■ 标签注意力层

- 标签注意力机制

标准缩放点积注意力机制公式：

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_h}})V$$

多头注意力机制获取更多的潜在信息：

$$Multi\ (Q,K,V)\ =\ concat\ \left(head_1,\ \ldots\ ,\ head_{\mathrm{k}}\right)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

最终的注意力机制计算公式为：

$$H^l = Multi(Q,K,V)+Q$$

2 LLA-CRF模型

■ 标签注意力层

- 标签注意力机制

$$H^l = Multi(Q,K,V) + Q$$

$K = V = x^l, x^l \in \mathbb{R}^{|L| \times d_h}$ 为标签集合表示。

$Q = H^c$ 时，可得隐藏层输出与标签的注意力机制输出 H_c^l

$Q = H^p$ 时，可得输入语句词性与标签的注意力机制输出 H_p^l

最终标签注意力层的输出为：

$$H = [H^c; H_c^l; H_p^l]$$

2 LLA-CRF模型

■ CRF层

最后，顶层CRF层接收标签注意力层的输出，标签序列概率为：

$$P(y \mid s) = \frac{\exp(\sum_i (W_{CRF}^{l_i} H_i + b_{CRF}^{(l_{i-1}, l_i)})}{\sum_{y'} \exp(\sum_i (W_{CRF}^{l'_i} H_i + b_{CRF}^{(l'_{i-1}, l'_i)})}$$

3 实验

■ 数据集

- 从中学教材、考纲、教案等资源人工收集了10934条句子
- 进行句子去重、数据集划分和自动标注，再人工校正
- 三个数据集内包含的术语词不重叠以测试模型的真实泛化性能

3 实验

■ 数据集

因为以句子包含的术语词不重叠的前提下划分数据集，在语料术语词稠密的情况下，在划分过程中会过滤掉大部分句子。

表1. 数据集信息

	训练集	验证集	测试集
句子数	3.5k	0.22k	0.25k
字数	12.5k	8.9k	9.5k
术语数	13909	231	336

3 实验

■ 实验结果与分析

本文的基线模型为同样基于字的BiLSTM-CRF、BiGRU-CRF和Lattice-CRF等序列标注模型。

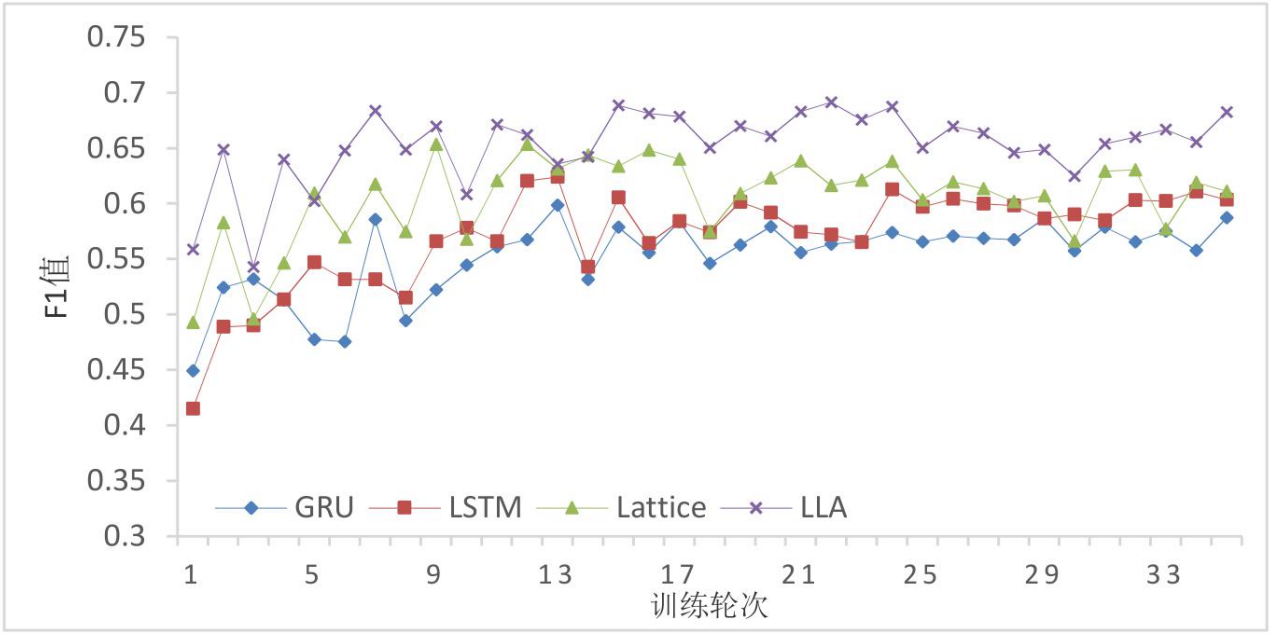


图2 在验证集上35个训练轮次的F1值结果

3 实验

■ 实验结果与分析

在测试集上的最终实验结果

表2. 实验结果

模型	P	R	F1
BiGRU-CRF	59.64	69.05	64.00
BiLSTM-CRF	59.57	70.24	64.57
Lattice-CRF	65.06	68.15	66.57
LLA-CRF	**69.30**	**70.54**	**69.91**

3 实验

■ 消融实验

表 3. 消融实验结果

注意力机制	P	R	F1
隐层特征	66.95	**71.13**	68.98
词性特征	67.83	69.64	68.72
两者	**69.30**	70.54	**69.91**

3 实验

- 抽取实例分析
  - 歧义术语区分抽取
  - 语料语境情景不全面

表 4. 部分易歧义数学术语抽取实例

句子片段	抽取结果	术语
当学生独自面对问题的时候	非术语	面
在试卷讲评方面	非术语	
线构成面	术语	
提高教学质量	非术语	高
高一年级	非术语	
高为10cm	术语	
图形中的点	术语	点
学生易错点	非术语	
易错题的特点	非术语	
任意多边形的外角的和	非术语	和
两个数之和	术语	
有两角和它们的夹边	非术语	

4 结束语

■ 总结

- 本文将中学数学术语抽取转化为序列标注任务来完成，额外考虑词性特征，引入对标签的注意力机制，提出了中学数学领域的术语抽取模型LLA-CRF。

- 本文模型与对比序列标注模型在中学数学术语抽取任务上，有着更高的精确率、召回率和F1值。

- 相比于传统的统计方法，能够在句级上区分抽取术语。

江西師範大學

谢谢！

huax1102@163.com