中国科学院软件研究所学术年会’2020
暨计算机科学国家重点实验室开放周

学术论文

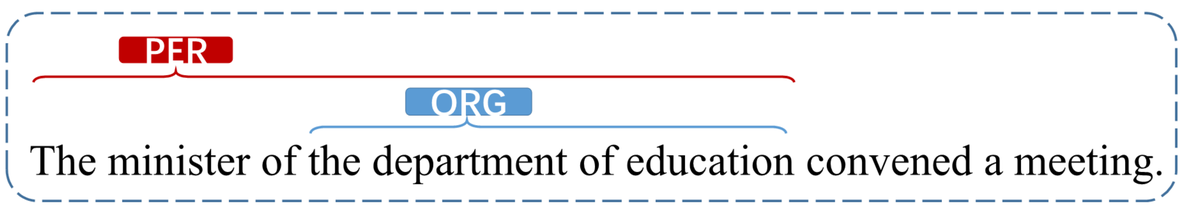
用于嵌套命名实体识别的串到块锚点区域网络

林鸿宇，陆垚杰，韩先培，孙乐
TEL: 13581158099 E-mail: xianpei@iscas.ac.cn

In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (CCF-A)*

简介

- 传统的基于序列标注的模型无法解决带有嵌套结构的命名实体识别问题

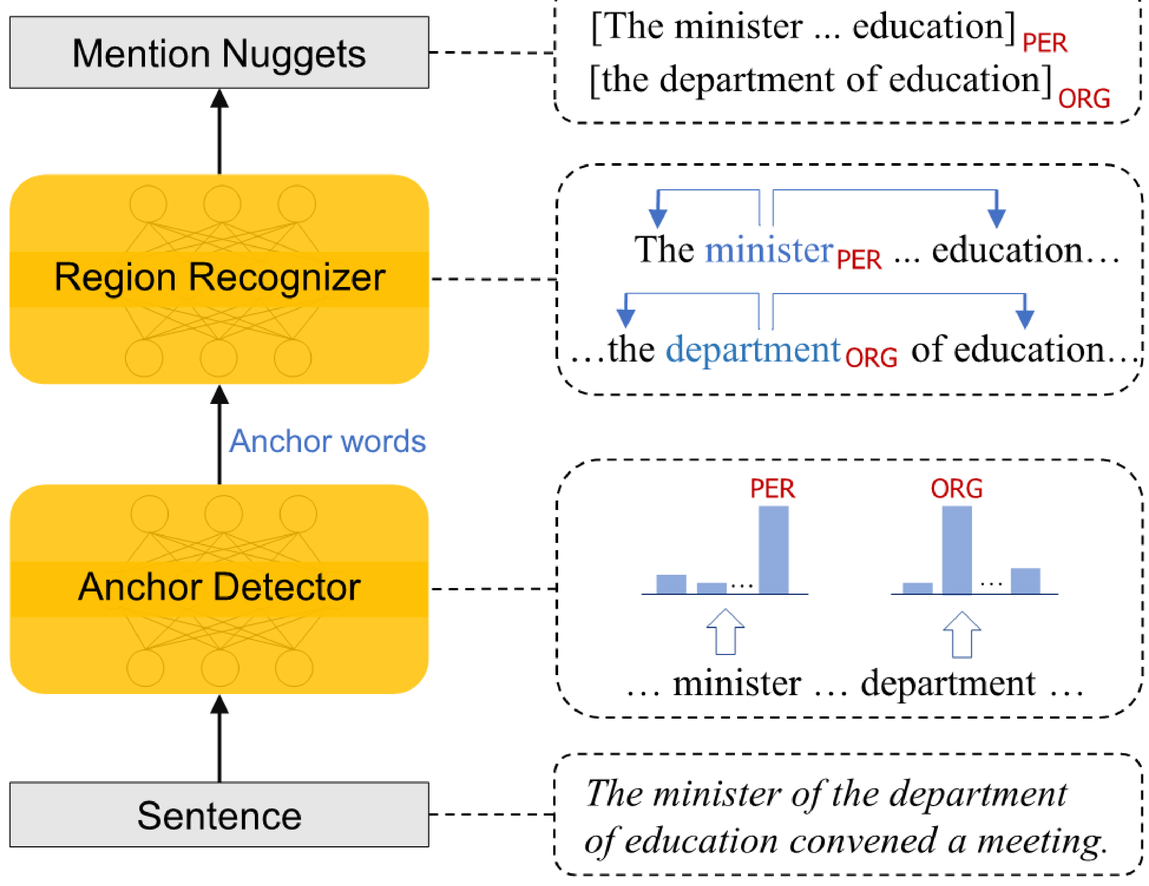


- 先前的解决方案：
  - 1) **基于区块模型的方法**：高时间复杂度
  - 2) **基于模式的方法**：模式带有歧义，歧义越低的模式时间复杂度越高

本文核心思想

- 每一个实体提及有至少一个的锚点词，该锚点词显著地暗示了实体提及的类别
- 即使嵌套实体提及存在，不同的实体提及对应有不懂的锚点词
- 根据上述思想，将嵌套命名实体提及过程分为以下两步：
  - 首先使用字级别分类器**检测锚点词**
  - 使用指针网络**识别锚点词对应的提及边界**
- 使用**包损失函数**在没有锚点词标注的情况下联合训练上述模型

锚点区域网络



包损失函数

- 核心思想：
  - 所有同属于一个最能层实体提及的词语构成一个包
  - 包中与实体类型关联程度越大的词语越可能是锚点词

$$\mathcal{L}(x_i;\theta) = \omega_i \cdot [-\log P(c_i|x_i) + L^R(x_i;\theta)] + (1-\omega_i)\cdot[-\log P(NIL|x_i)]$$

$$\omega_i = [\frac{P(c_i|x_i)}{\max_{x_t \in B_i} P(c_i|x_t)}]^{\alpha}$$

实验结果

Model	ACE2005			GENIA			KBP2017		
	P	R	F1	P	R	F1	P	R	F1
LSTM-CRF [12]	70.3	55.7	62.2	75.2	64.6	69.5	71.5	53.3	61.1
Multi-CRF	69.7	61.3	65.2	73.1	64.9	68.8	69.7	60.8	64.9
FOFE(c=6) [21]	76.5	66.3	71.0	75.4	67.8	71.4	81.8	62.0	70.6
FOFE(c=n) [21]	76.9	62.0	68.7	74.0	65.5	69.5	79.1	62.5	69.8
Transition [23]	74.5	71.5	73.0	78.0	70.2	73.9	74.7	67.0	70.1
Cascaded-CRF [32]	74.2	70.3	72.2	78.5	71.3	74.7	-	-	-
LH [1]	70.6	70.4	70.5	79.8	68.2	73.6	-	-	-
SH(c=6) [25]	75.9	70.0	72.8	76.8	71.8	74.2	73.3	65.8	69.4
SH(c=n) [25]	76.8	72.3	74.5	77.0	73.3	**75.1**	79.2	66.5	72.3
KBP2017 Best [36]	-	-	-	-	-	-	72.6	73.0	72.8
Anchor-Region Networks (c=6)	75.2	72.5	73.9	75.2	73.3	74.2	76.2	71.5	73.8
Anchor-Region Networks (c=n)	76.2	73.6	**74.9**	75.8	73.9	74.8	77.7	71.8	**74.6**

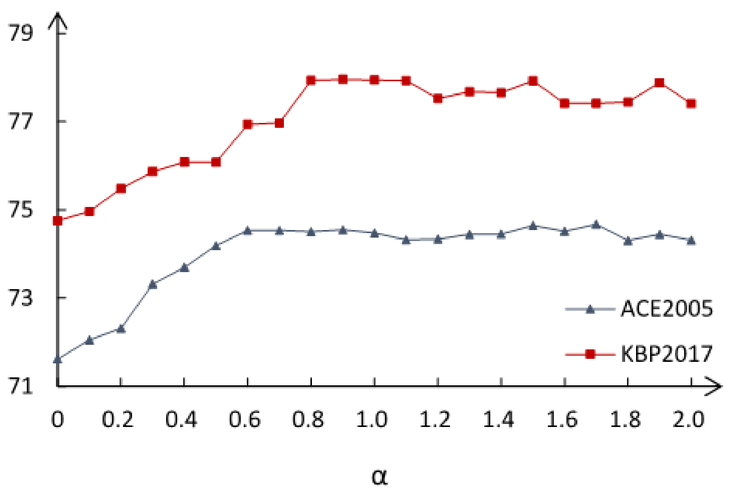


Fig. 4. The F1-score w.r.t. different α in Bag Loss on development sets. When $\alpha = 0$, the model ablates Bag Loss and will treat all words in the same innermost mention as anchor words during training.