Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

匮乏资源命名实体识别

吴杭颐　硕士研究生

2021年10月31日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1. 了解命名实体识别的研究历史及现状
  - 2. 了解跨语言NER、跨领域NER、多任务学习的基本概念
  - 3. 了解匮乏资源命名实体识别的方法
  - 4.了解命名实体识别的实际应用

背景简介

背景简介

- 命名实体识别（ Named entity recognition，简称NER ）
  - 提取文本中具有特定意义的实体并对这些实体进行分类，包括人名、地名、机构名、日期、时间、货币、百分比等
  - 不仅需要找出实体的位置，还需要对实体进行分类

General domain(news)

- Person name
- Organization name
- Location name
- Numeric expression

Biomedical domain

- Gene
- Protein
- Disease
- Chemical
- Cell

不同领域的实体

小丽 在 北京理工大学 的 信息科学实验楼 听了一场学术报告

PER ORG LOC

Protein DNA RNA Chemical Disease

Glucocorticoid receptors in peripheral blood lymphocytes of patients with bronchial asthma. Quantitation of glucocorticoid receptors (GCR) and the study of their affinity for glucocorticosteroids (GCS) were made in peripheral blood lymphocytes of bronchial asthma (BA) patients in consideration of GCR treatment and serum levels of endogenous cortisol.

NER的任务：找+分类

背景简介

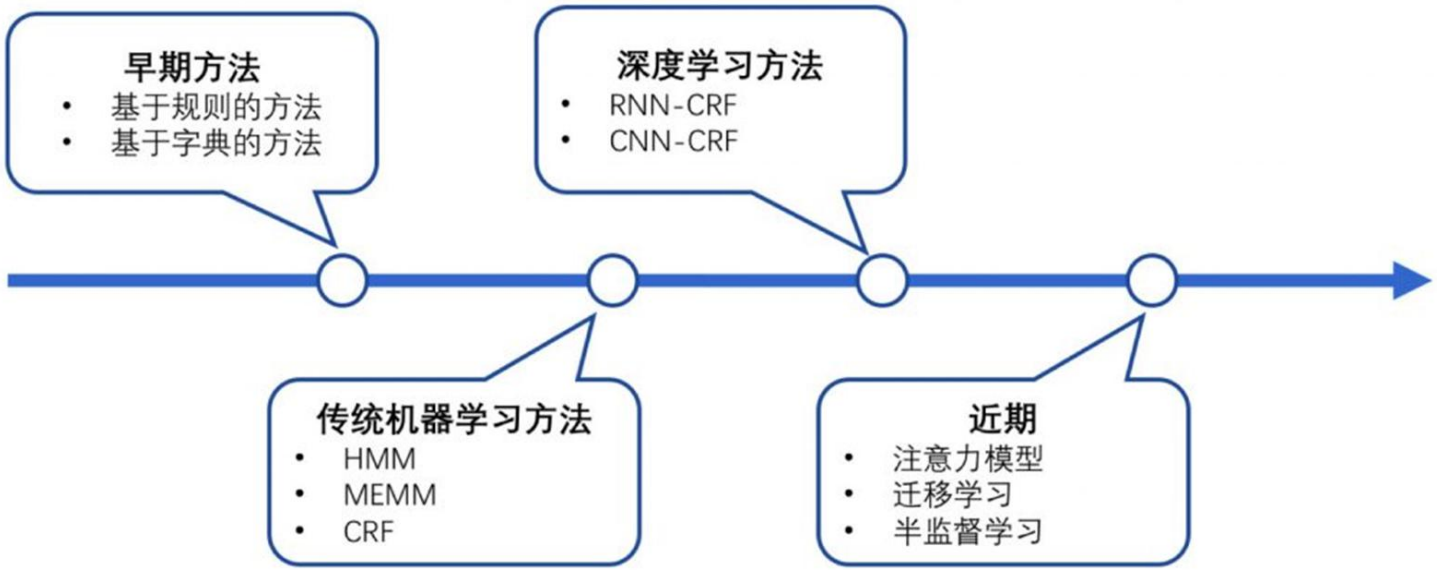
- 命名实体识别难点：
  - 歧义多
  - 边界界定困难
    - 武汉市长江大桥
  - 领域命名实体识别局限性
    - 目前NER只是在有限的领域和实体类型中取得了较好的成绩，但这些技术无法很好地迁移到其他特定领域中
      - 一方面，由于不同领域的数据往往具有领域独特特征
      - 另一方面，由于领域资源匮乏造成标注数据集缺失

领域资源匮乏涵盖了不同资源条件的一系列情况，包括濒危语言的相关工作，也包括英文等高资源语言在某些专业领域和任务上

根据不同的任务有不同的标准和指导准则

背景简介

- NER发展趋势



- 缺乏标注数据使得无法适应资源匮乏领域，而收集大量标注数据是昂贵、困难、甚至不可能

↓

如何解决资源匮乏领域的命名实体识别难题?

↓

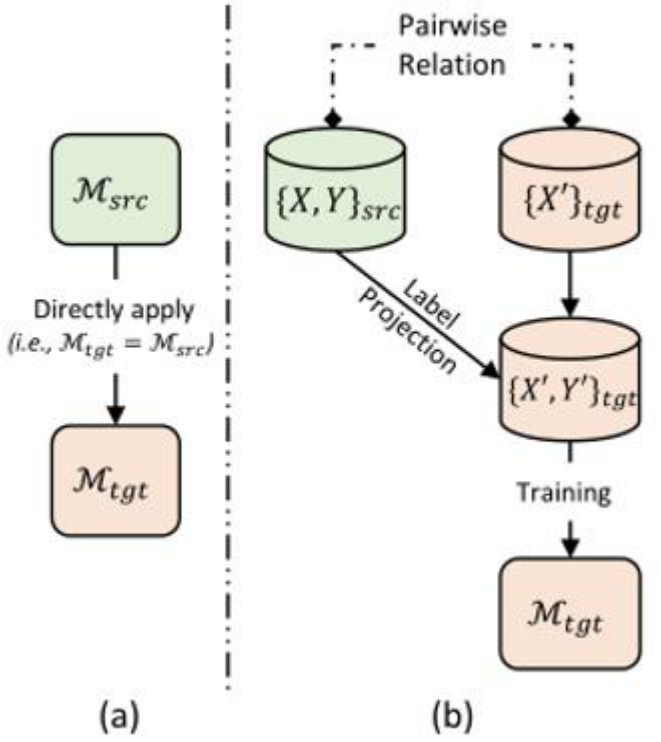
迁移学习、多任务学习等

基本概念

基本概念

- 对资源匮乏的语种来说，获取大量带标注数据的代价太大
- 跨语言NER（源语言->目标语言）方法：
  - 1. 基于模型的迁移
    - 侧重于在源语言中的标注数据上训练一个共享的NER模型
    - 缺乏词汇化特征，而这些特征对NER任务具有预测能力
    - 不利用可能包含有用信息的目标语言中的未标记数据
  - 2. 基于标注数据的投影
    - 对齐平行语料
    - 侧重于使用源语言中的标注数据生成目标语言中的伪标注数据，以训练NER模型
    - 不适用于无法访问所需标注数据的情况

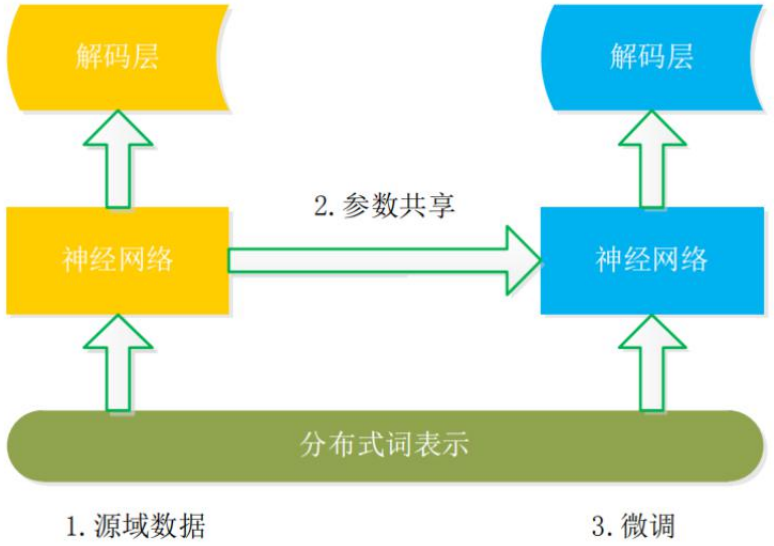
cat<--->猫
dog<--->狗

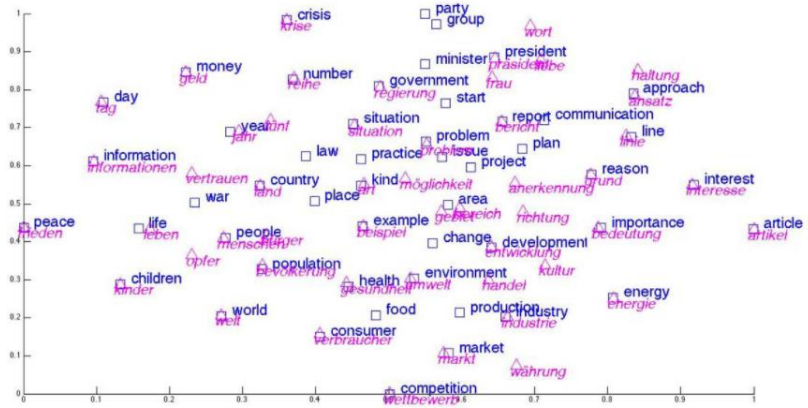


(a)模型迁移方法　(b) 标注投影方法

基本概念

- 跨领域NER（源领域->目标领域）方法：
  - 1. 基于参数转移
    - 将目标域、源领域的NER模型一同进行训练，共享参数
    - 利用目标领域的标记数据进行fine-tune
  - 2. 基于特征转移
    - 建立源领域、目标领域的实体标签空间之间的联系
    - 将标签embedding作为特征，输入NER模型
    - 通过特征提取器，将含有标签知识的表示作为特征加入跨领域的NER模型中

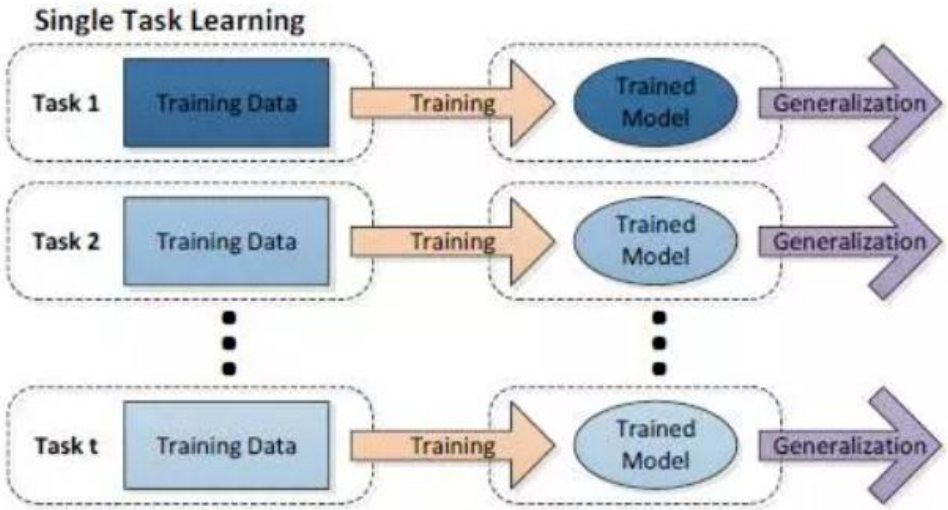


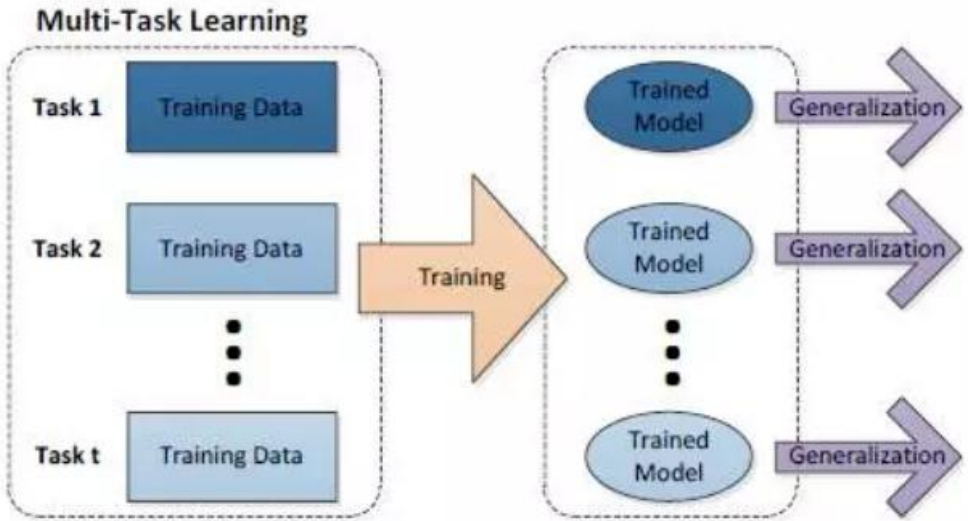


基本概念

- 多任务学习
  - 把多个相关的任务放在一起学习
    - 考虑了问题之间所富含的丰富关联信息
    - 在学习中共享学到的知识，能更好地泛化
    - 比如一个正在笑的人会张开嘴，有效地发现和利用这个相关脸部属性将帮助更准确地检测嘴角

绿色 红色
蔬菜 肉





算法原理

算法原理

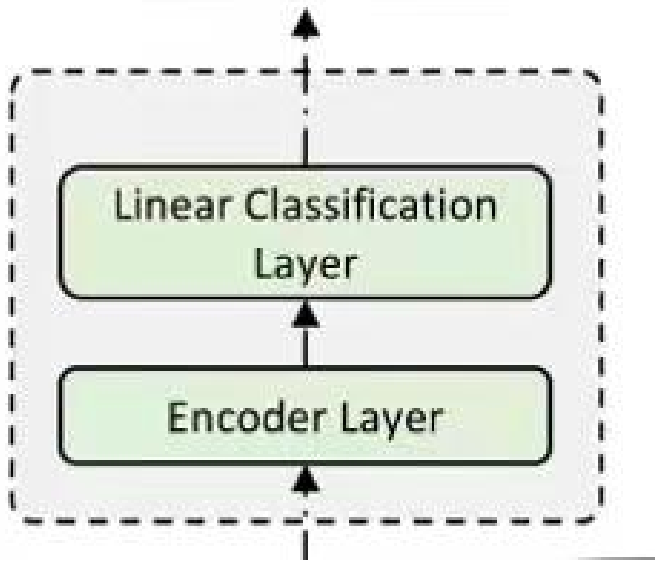
T	为了更好地解决标注数据很少或没有标注数据语言的命名实体识别(NER)问题
I	源语言中的NER模型
P	1、利用多语言BERT作为基础模型来产生独立于语言的特性； 2、源语言的先前训练的NER模型被用作教师模型，以预测目标语言的伪标签； 3、最后，使用伪标签数据为目标语言训练学生NER模型
O	训练好的学生NER模型

P	先前基于标注投影和基于模型转移的方法存在局限性
C	存在先前训练好的源语言NER模型被用作教师模型
D	关注跨语言NER的极端情况，目标语言中没有可用的标注数据
L	ACL 2020

算法原理

- 模型结构：
  - 编码器层

$$x = \{x_i\}_{i=1}^{L}$$
$$h = \{h_i\}_{i=1}^{L}$$
$$h = f_\theta(\mathrm{x})$$



  - 线性分类器层

$$p(x_i, \Theta) = softmax(Wh_i + b)$$

对应token x_i实体标签的概率分布

  - 输出

$$y = \{y_i\}_{i=1}^{L}$$

算法原理

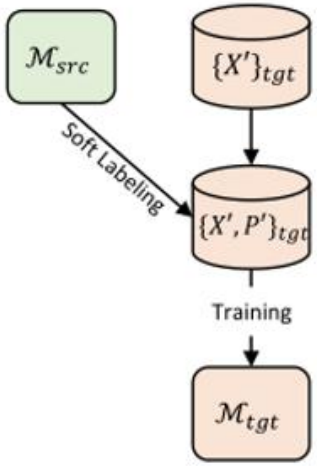
- Teacher-Student Learning：
  - Training
    - 训练学生模型模仿教师模型输出在目标语言中未标记数据D_{tgt}上实体标签的概率分布
    - $\mathcal{L}(x_i', \Theta_s) =$

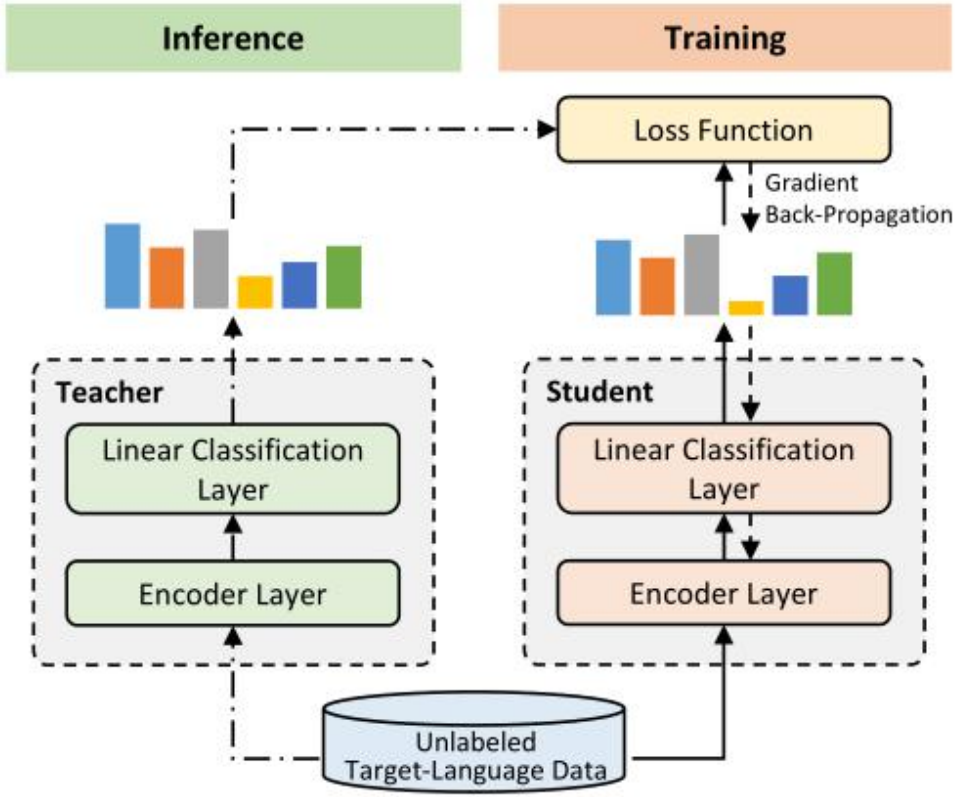
      $\frac{1}{L}\sum_{i=1}^{L} MSE(\hat{p}(x_i', \Theta_s), \tilde{p}(x_i', \Theta_T))$

      $x' \in D_{tgt}$的师生学习损失
    - $\mathcal{L}(\Theta_s) = \sum_{x' \in D_{tgt}} \mathcal{L}(x_i', \Theta_s)$

      整个训练的损失函数

      最小化$\mathcal{L}(\Theta_s)$导出学生模型



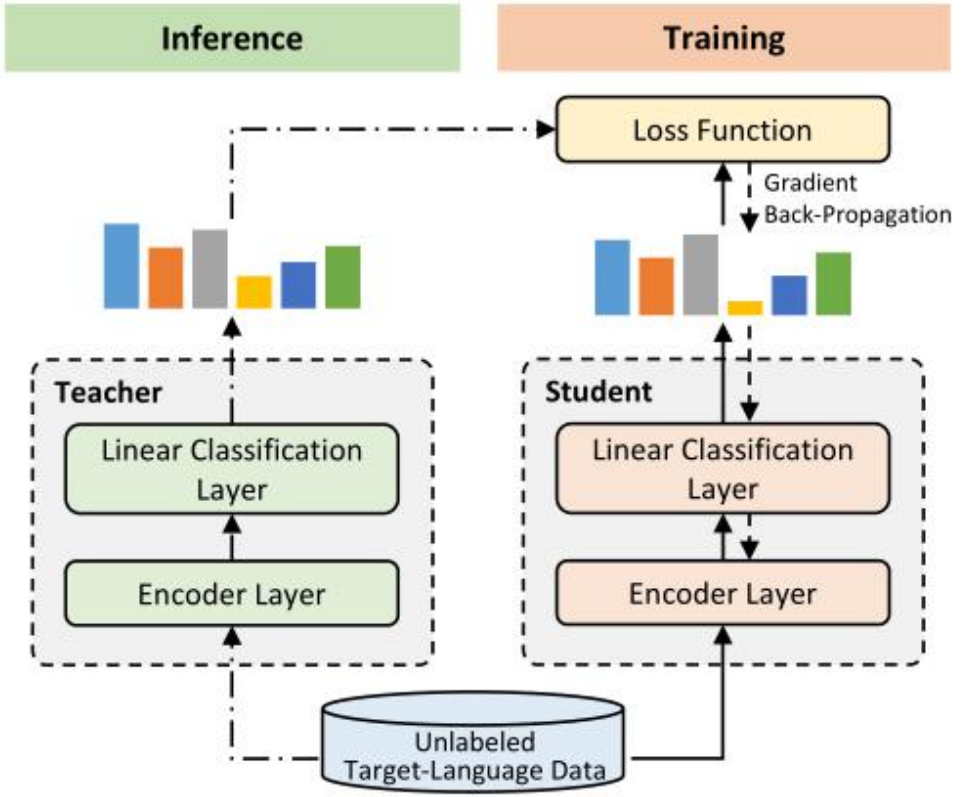


跨语言NER的师生学习方法框架

算法原理

- Teacher-Student Learning
  - Inference
    - 只利用通过学习到的学生模型去预测句子x的每个token实体标签概率分布
    - 用预测到的最高概率的实体标签作为最后的标签y_i

$$y_i = \arg^{max}_{c} \hat{p}(x_i', \Theta_s)_c$$



跨语言NER的师生学习方法框架

实验分析

- 实验设置：
  - 对3种目标语言（即西班牙语、荷兰语和德语）进行广泛的实验
  - 2个NER基准数据集
    - CoNLL-2002 (Spanish and Dutch)
    - CoNLL-2003 (English and German)
      - 都被标注为4种实体类型：PER，LOC，ORG，MISC

实验分析

- 实验结果：

	es	nl	de
Täckström et al. (2012)	59.30	58.40	40.40
Tsai et al. (2016)	60.55	61.56	48.12
Ni et al. (2017)	65.10	65.40	58.50
Mayhew et al. (2017)	65.95	66.50	59.11
Xie et al. (2018)	72.37	71.25	57.76
Wu and Dredze (2019)†	74.50	79.50	71.10
Moon et al. (2019)†	75.67	80.38	71.42
Wu et al. (2020)	76.75	80.44	73.16
Ours	**76.94**	**80.89**	**73.22**

与直接模型迁移相比，师生学习方法利用了未标记的目标语言数据的好处

实验分析

- 实验结果（进行消融实验）
  - 消融实验
    - 直接标注投影 HL
      - HL传递的知识信息少，泛化能力弱
    - 直接模型转换 MT
      - 没有未标记的目标语言数据，退化为直接应用源语言模型

	es	nl	de
Single-source:			
Ours	**76.94**	**80.89**	**73.22**
HL	76.60 (-0.34)	80.43 (-0.46)	72.98 (-0.24)
MT	75.60 (-1.34)	79.99 (-0.90)	71.76 (-1.46)

所有跨语言NER设置中一致性能下降

实验分析

- 实验结果
  - 解决了以前基于标注投影和基于模型迁移的方法的局限性
  - 借助于未标记的目标语言数据，提高了模型预测准确性
  - 引入师生学习，使用伪标签可以传递更多的信息

算法原理

算法原理

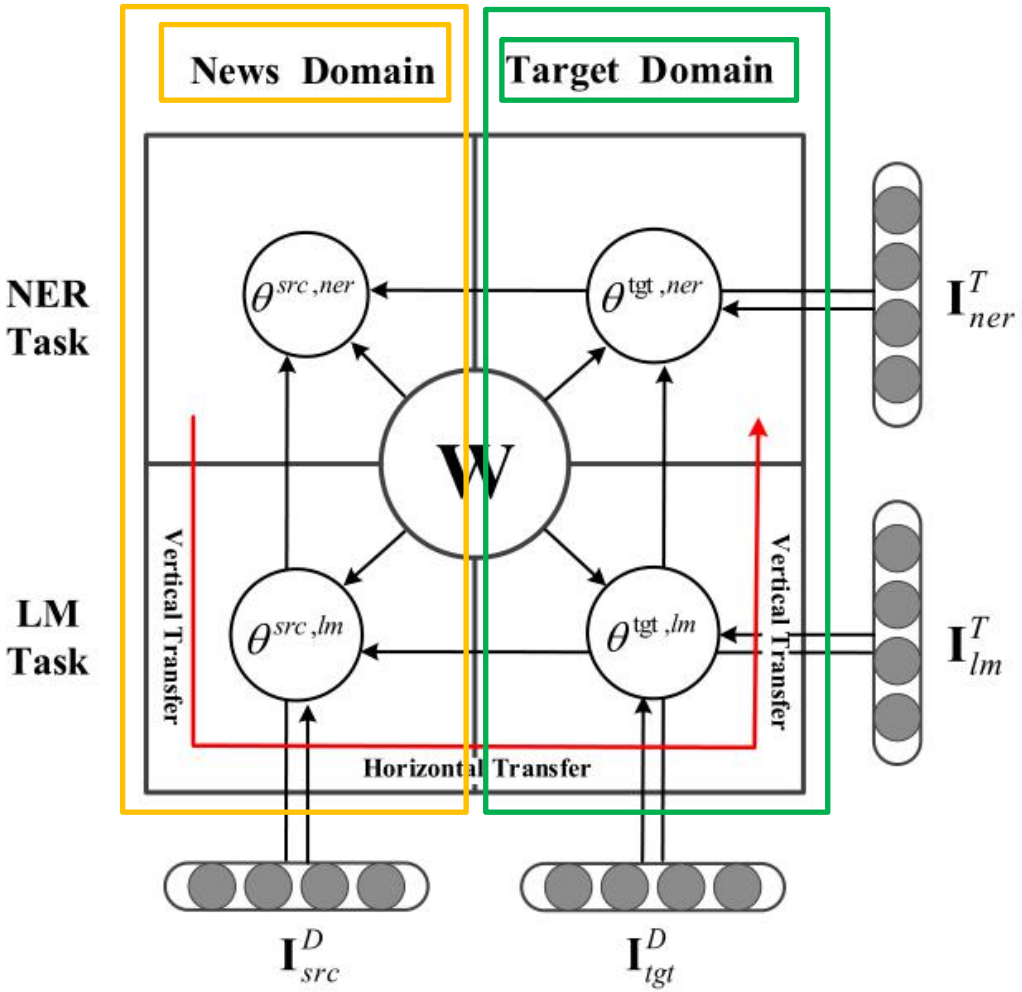
T	研究跨领域和跨任务环境下可迁移知识的模块化
I	源领域和目标领域NER数据、未标注数据
P	1、句子通过共享的嵌入层得到一个单词级的表示 2、通过参数生成网络计算一系列任务和领域参数
O	目标模型

P	多任务学习应用于跨领域NER的实用性
C	存在LM训练过程能够捕获名词实体以及上下文
D	由于文本类型和实体名称的不同，跨域NER增加了建模的难度
L	CCF A类

算法原理

- 思想
  - 同一领域（源领域）的跨任务迁移
  - 同一任务（LM任务）的跨领域知识迁移
  - 同一领域（目标领域）的跨任务迁移

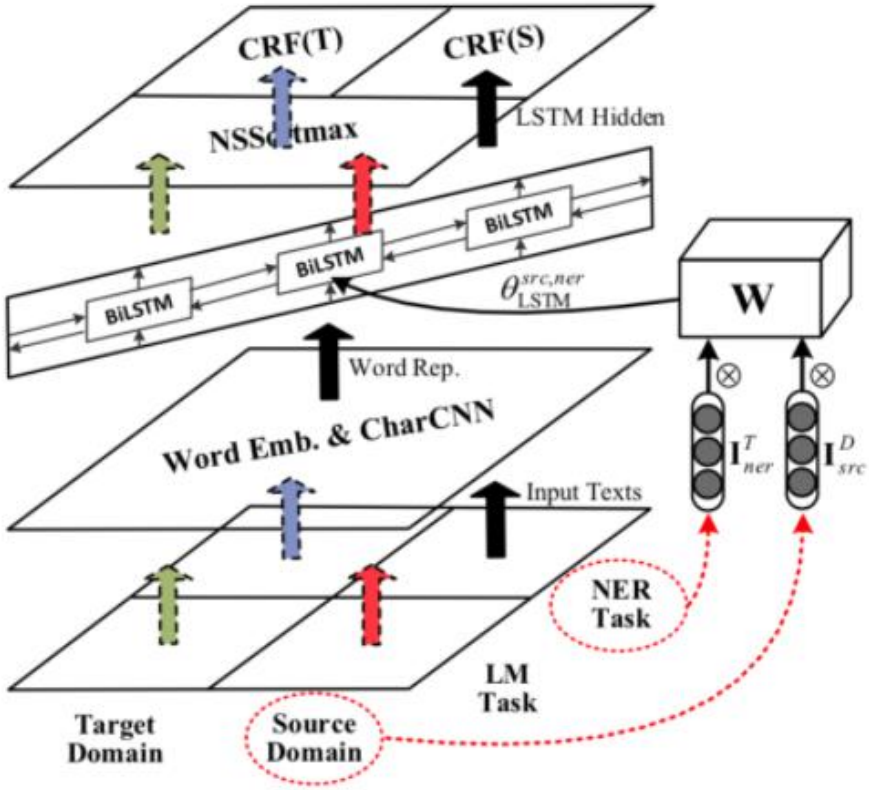
- 模型架构
  - 不同的task对应不同domain，形成四个模块
  - NER-Source Domain学到的内容能够很好地帮助NER-Target Domain进行抽取



模型图

算法原理

- 1. 输入
  - 数据：
    - 源域NER训练集：$S_{ner} = \{(x_i, y_i)\}_{i=1}^{m}$
    - 目标域NER训练集：$T_{ner} = \{(x_i, y_i)\}_{i=1}^{n}$
    - 源域原始文本集：$S_{lm} = \{x_i)\}_{i=1}^{p}$
    - 目标域原始文本集：$T_{lm} = \{x_i)\}_{i=1}^{q}$
- 2. 共享embedding层
  - 对于每个词，将其word-embedding和经过CNN处理过的char-embedding拼接起来，得到每个词的词表示
    - $v_i = [e^w(x_i) \oplus CNN(e^c(x_i))]$



模型架构

算法原理

- 3. 参数生成网络
  - 某特定领域特定任务的Bi-LSTM参数：

$$\theta_{LSTM}^{d,t} = W \otimes I_d^D \otimes I_t^T$$

    - 领域和任务之间的相关度通过两个向量进行计算：
      - 1）特定task-embedding：

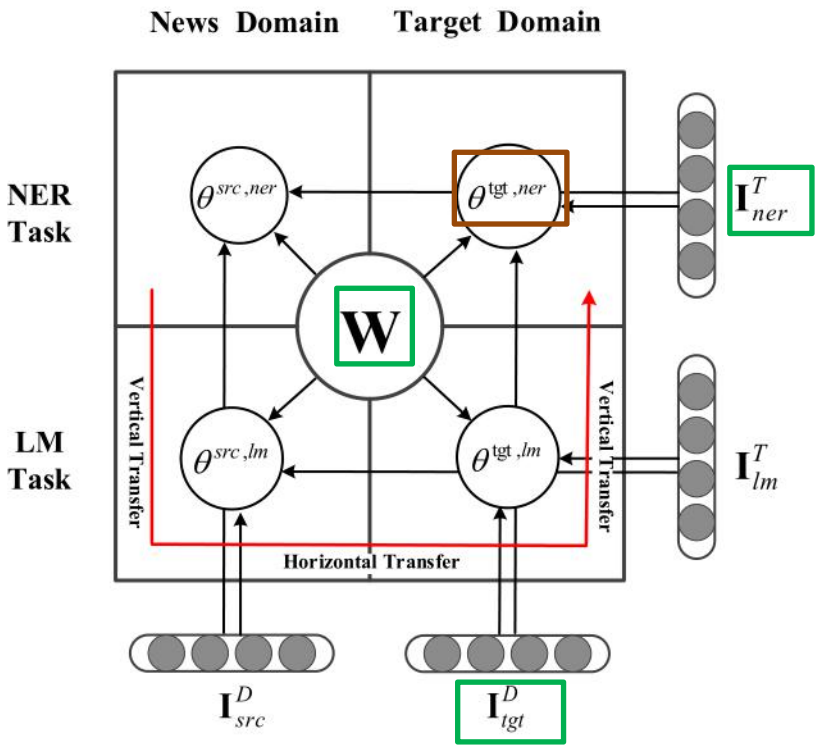
$$I_t^T (t \in \{ner, lm\})$$

      - 2）特定domain-embedding：

$$I_d^D \ (d \in \{src, tgt\})$$

    - 3）针对特定任务和领域知识的可变参数组W
- 训练目标

$$\mathcal{L} = \sum_{d \epsilon \{scr, tgt\}} \lambda^d (\mathcal{L}_{ner}^d + \lambda^t \mathcal{L}_{lm}^d) + \frac{\lambda}{2} ||\Theta||^2$$



模型图

算法原理

- 多任务学习算法：
  - 第 4–5 行、第 7–8 行、第 11–12 行、第 15–16 行分别代表之前提到的四种任务
  - 每种任务都是同样的步骤：
    - 首先通过生成参数网络生成对应的 LSTM 网络参数
    - 继而计算梯度并得到输出
    - 最后更新参数

Algorithm 1 Multi-task learning

Input
Para
- Para
- Outp
Outp

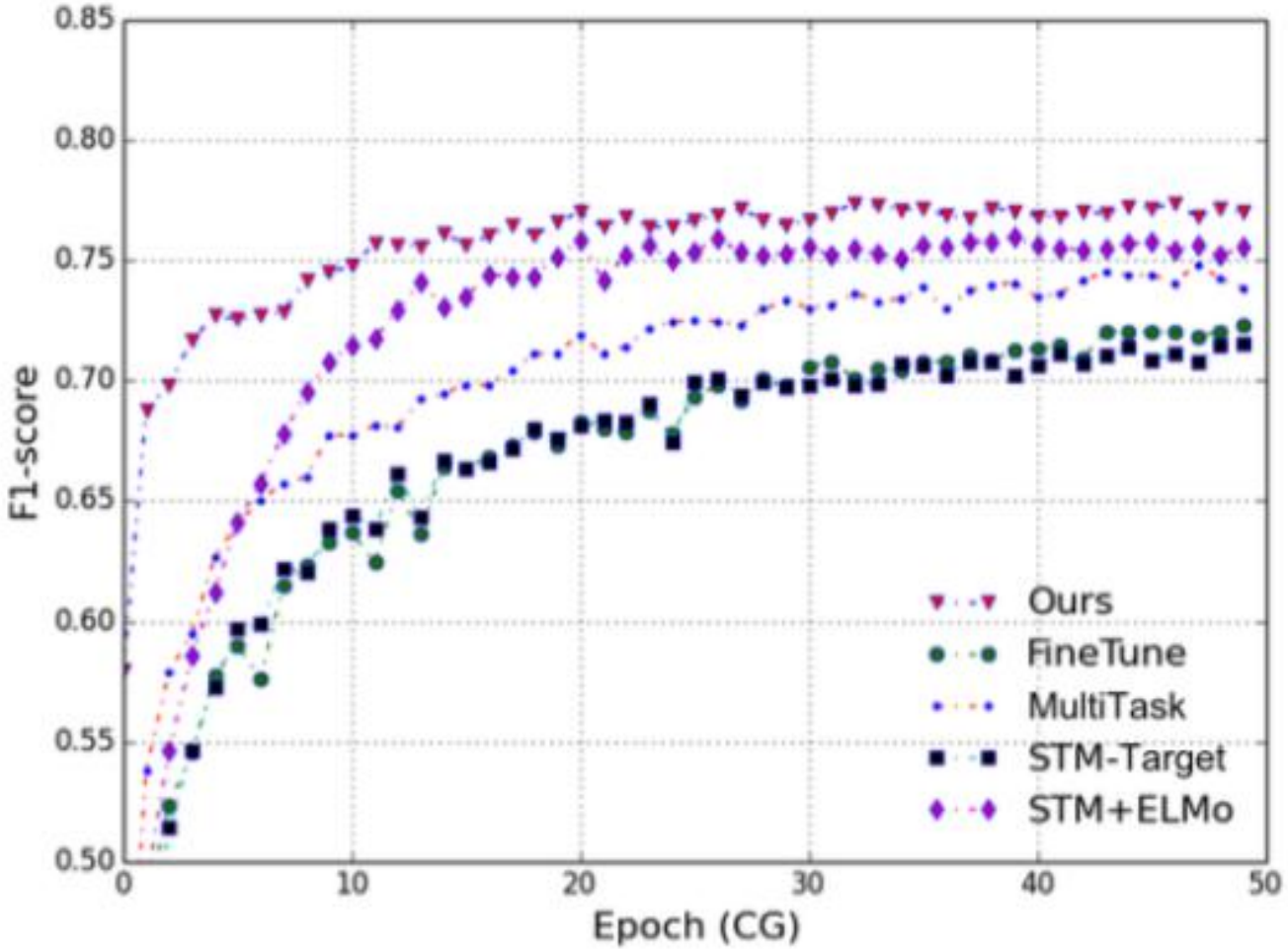
1: **while** training steps not end **do**
2: split training data into minibatches:
$B^{ner_s}, B^{ner_t*}, B^{lm_s}, B^{lm_t}$
3: # source-domain NER
4: $\theta^{src,ner}_{\text{LSTM}} \leftarrow f(\mathbf{W}, \mathbf{I}^D_{src}, \mathbf{I}^T_{ner})$
5: $\Delta\mathbf{W}, \Delta\mathbf{I}^D_{src}, \Delta\mathbf{I}^T_{ner}, \Delta\theta_{crf_s} \leftarrow train(B^{ner_s})$
6: # source-domain LM
7: $\theta^{src,lm}_{\text{LSTM}} \leftarrow f(\mathbf{W}, \mathbf{I}^D_{src}, \mathbf{I}^T_{lm})$
8: $\Delta\mathbf{W}, \Delta\mathbf{I}^D_{src}, \Delta\mathbf{I}^T_{lm}, \Delta\theta_{nss} \leftarrow train(B^{lm_s})$
9: **if** do supervised learning **then**
10: # target-domain NER
11: $\theta^{tgt,ner}_{\text{LSTM}} \leftarrow f(\mathbf{W}, \mathbf{I}^D_{tgt}, \mathbf{I}^T_{ner})$
12: $\Delta\mathbf{W}, \Delta\mathbf{I}^D_{tgt}, \Delta\mathbf{I}^T_{ner}, \Delta\theta_{crf_t} \leftarrow train(B^{ner_t})$
13: **end if**
14: # target-domain LM
15: $\theta^{tgt,lm}_{\text{LSTM}} \leftarrow f(\mathbf{W}, \mathbf{I}^D_{tgt}, \mathbf{I}^T_{lm})$
16: $\Delta\mathbf{W}, \Delta\mathbf{I}^D_{tgt}, \Delta\mathbf{I}^T_{lm}, \Delta\theta_{nss} \leftarrow train(B^{lm_t})$
17: Update $\mathbf{W}, \{\mathbf{I}^D\}, \{\mathbf{I}^T\}, \theta_{crf_s}, \theta_{crf_t}{}^*, \theta_{nss}$
18: **end while**

Note: * means none in unsupervised learning

实验分析

- 数据集：
  - 源领域的NER数据
    - 新闻领域
      - 来自 CoNLL-2003
  - 目标领域的NER数据：
    - 生物医药领域
      - BioNLP13PC(13PC)
      - BioNLP13CG(13CG)

实验分析



- STM-Target表现最差
  - 源领域数据是对NER结果有好处的；
- FineTune方法结果不是总比STM-Tartget结果好
  - 领域之间数据集的差异可能反而使模型效果变差
- STM+ELMo表现较好
  - 利用了源域和目标域的生文本信息，可以提升模型效果
- Ours同时使用了源领域的标记数据与目标领域的无标记数据，始终在图中给出了最佳F1-score表现

实验分析

- 实验结果:

Methods	Datasets	
	13PC	13CG
Crichton et al. (2017)	81.92	78.90
STM-TARGET	82.59	76.55
MULTITASK(NER+LM)	81.33	75.27
MULTITASK(NER)	83.09	77.73
FINETUNE	82.55	76.73
STM+ELMO	82.76	78.24
CO-LM	84.43	78.60
CO-NER	83.87	78.43
MIX-DATA	83.88	78.70
FINAL	**85.54**†	**79.86**†

在数据集13PC和13CG的F1-score表现

- 为了检测从CoNLL数据集迁移至13PC和13CG数据集上之后的NER任务的表现

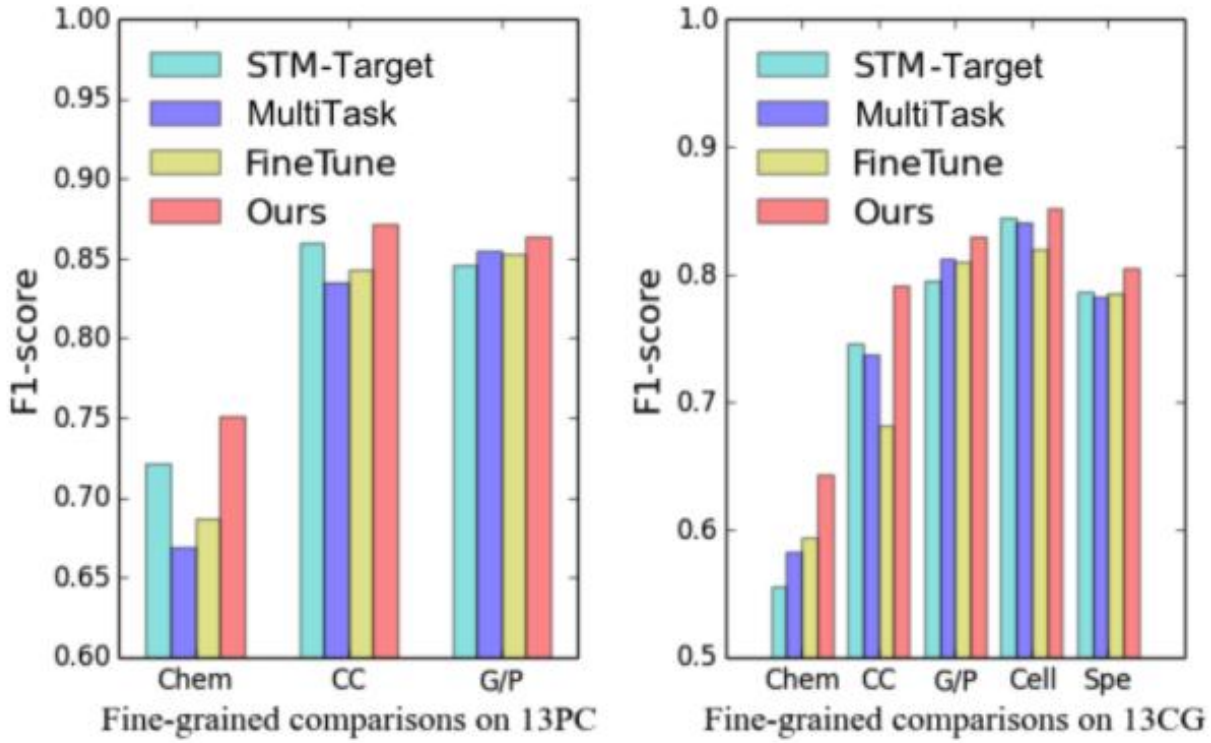
- MUILTITASK(NER+LM)表现最差
  - 使用4个任务的共享参数，以及与最终模型完全相同的数据条件
  - 在面对风格迥异以及包含相互冲突的信息时，跨领域和跨任务设置的挑战

实验分析

- 实验结果:

Methods	Datasets	
	13PC	13CG
Crichton et al. (2017)	81.92	78.90
STM-TARGET	82.59	76.55
MULTITASK(NER+LM)	81.33	75.27
MULTITASK(NER)	83.09	77.73
FINETUNE	82.55	76.73
STM+ELMO	82.76	78.24
CO-LM	84.43	78.60
CO-NER	83.87	78.43
MIX-DATA	83.88	78.70
FINAL	**85.54**†	**79.86**†

在数据集13PC和13CG的F1-score表现



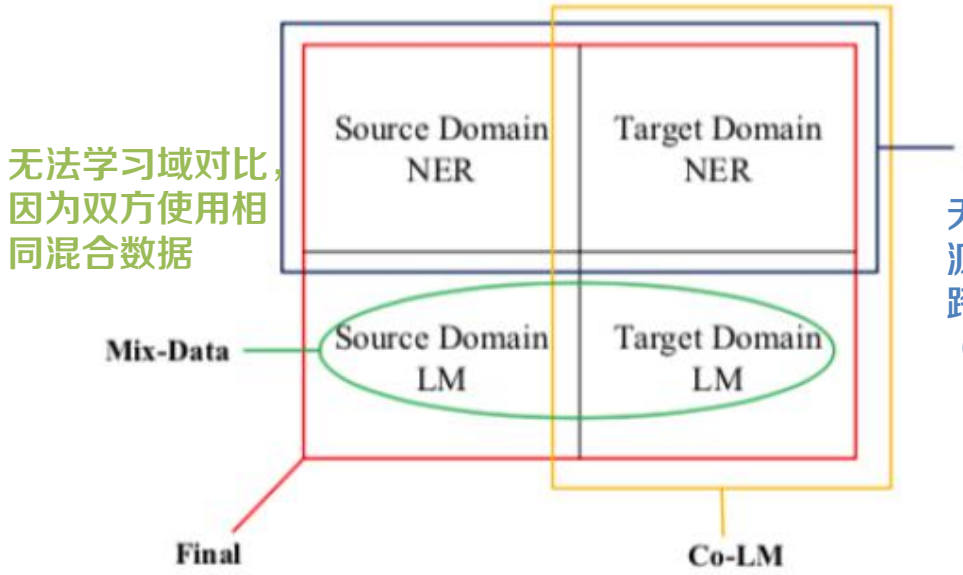
一些类型的实体识别直接使用多任务方法效果是非常差的，而本文的方法一直是最优的。可见，同样是多任务学习，参数生成网络带来的提升是巨大的。

实验分析

- 实验结果:

Methods	Datasets	
	13PC	13CG
Crichton et al. (2017)	81.92	78.90
STM-TARGET	82.59	76.55
MULTITASK(NER+LM)	81.33	75.27
MULTITASK(NER)	83.09	77.73
FINETUNE	82.55	76.73
STM+ELMO	82.76	78.24
CO-LM	84.43	78.60
CO-NER	83.87	78.43
MIX-DATA	83.88	78.70
FINAL	**85.54**†	**79.86**†

在数据集13PC和13CG的F1-score表现



无法学习域对比，因为双方使用相同混合数据

无任务，联合训练源域和目标域NER跨领域学习（新闻->医疗领域）

无源域，联合训练目标域NER和LM跨任务学习（LM->NER）

该论文模型超越了以上，说明该模型学习到了领域和任务之间的差别，帮助了NER任务进行提升

实验分析

- 优势
  - 能够有效地结合领域之间的差异知识
  - 通过一种新的参数生成网络进行跨领域语言建模
  - 在有监督的领域自适应方法中是非常有效

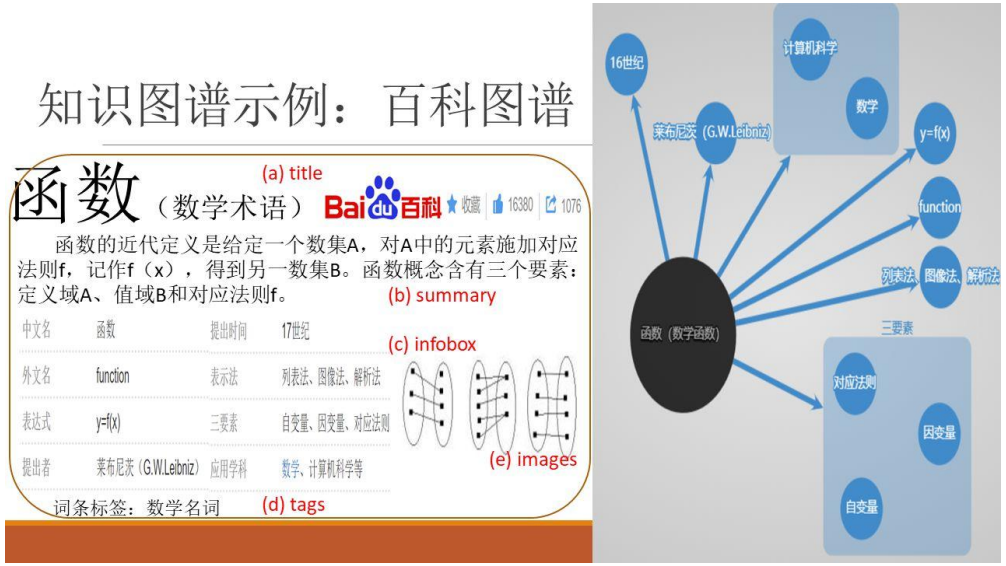
- 劣势
  - 无监督的领域自适应方法中效果需要提升

应用总结

应用总结

- 总结：
  - “多任务学习”是解决匮乏资源NER问题的一个有效方案，可以考虑能否将其很好地应用到无监督的领域自适应方法中
  - 引入teacher-student learning，使用伪标签传递更多信息也是不错的选择

- 应用：
  - 构建知识图谱
  - 机器翻译
  - 对话系统



参考文献

[1] Wu Q , Lin Z , Karlsson B F , et al. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language[J]. 2020.
[2] Jia C , Liang X , Zhang Y . Cross-Domain NER using Cross-Domain Language Modeling[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
[3] Liu Z , Winata G I , Fung P . Zero-Resource Cross-Domain Named Entity Recognition[C]// 2020.
[4] https://www.jianshu.com/p/b05243cbe7e5

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。