Dual Supervision Framework for Relation Extraction with Distant Supervision and Human Annotation

2023年6月12日 8:56

泛读
==========

摘要

因为关系提取在知识图谱构建和问答系统中有重要的应用，所以它被广泛的研究。为了更好地利用人类标记的准确率以及远程监督的廉价性，作者他们提出了一个双重的监督框架，它能够有效的利用这两类数据。当然作者他们不是简单的结合这两种数据，因为远程监督数据有一种标记误差，所以他们提出了**HA-Net**和**DS-Net**各自来预测标注。
此外，作者还提出了一种额外的损失——disagreement penalty，能够让HA-Net去学习远程的监督标注。
以及利用了额外的网络μ-Net和σ-Net，通过考虑上下文信息来自适应地评估标记偏差。

文章的结构

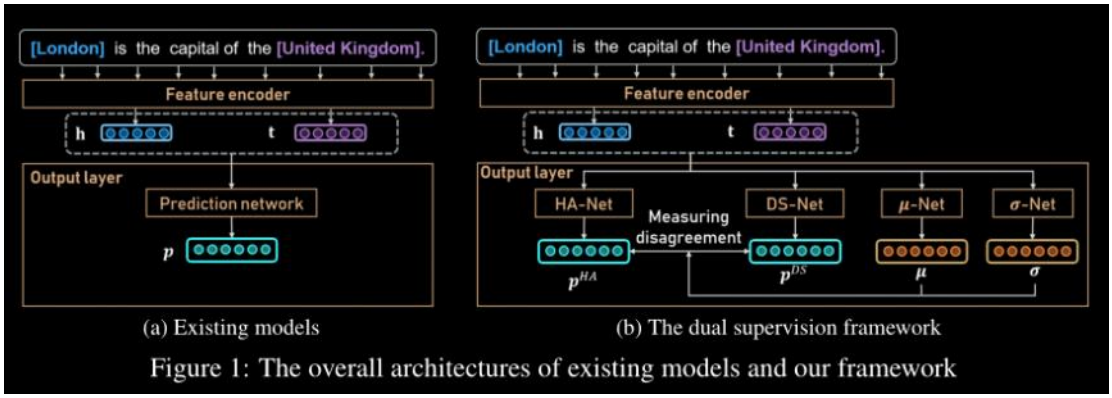
- Introduction
- Preliminaries
  Problem Statement
  Existing Works of Relation Extraction
- Dual Supervision Framework
  An Overview of the Dual Supervision Framework
  Separate Prediction Networks
  Disagreement Penalty
  Parameter Networks
  Loss Function
  Analysis of the Disagreement Penalty
  Extension to Document-level Relation Extraction
- Experiments
  Experimental Settings
  Comparison with Existing Methods
  Ablation Study 消融实验
  Quality Comparison 质量比较
  Topic-aware RE
- Related Works
- Conclusion

结论

在分析远程监督中的标签偏见基础上，作者提出了利用人类注释和远程监督的双重监督框架，他们对于关系提取模型中的输出层设计了一种新的结构，这个结构由4个子网络组成。这样的结构对于远程监督的噪声标签是鲁棒的，因为标签是分别由HA-Net和DS-Net分别预测的。
另外，作者提出了一种额外的损失项——disagreement penalty，它能够让HA-Net学习到远程监督标签。u-Net网络和σ-Net通过考虑上下文信息自适应地评估标记偏差。
此外，作者理论上分析了disagreement penalty的效果，作者的实验表明他们的dual supervison 框架比起是明显优于现有的方法。

图表及其内容 没搞明白

这个是已有模型与作者提出模型的框架图



(a) Existing models (b) The dual supervision framework

Figure 1: The overall architectures of existing models and our framework

这个表应该是不同数据的不同分布情况

Distribution	p-value
Log-normal	0.008
Weibull	0.001
Chi-square	4.6×10^{-10}
Exponential	3.6×10^{-13}
Normal	1.2×10^{-15}

Table 1: The result of K-S test

这个是对数据集的一个统计 我知道的有NYT数据集（新闻方面的数据集），这个应该是对不同数据集中的数据进行实验的划分 训练集与测试集吧

Data	Number of instances				# of rel. types
	Train-HA	Train-DS	Dev	Test	
KBP	378	132,369	14,103	1,488	7
NYT	756	323,126	34,871	3,021	25
DocRED	38,269	1,508,320	12,332	12,842	96

Table 2: Statistics of datasets

句子级别RE数据库（KBP和NYT）

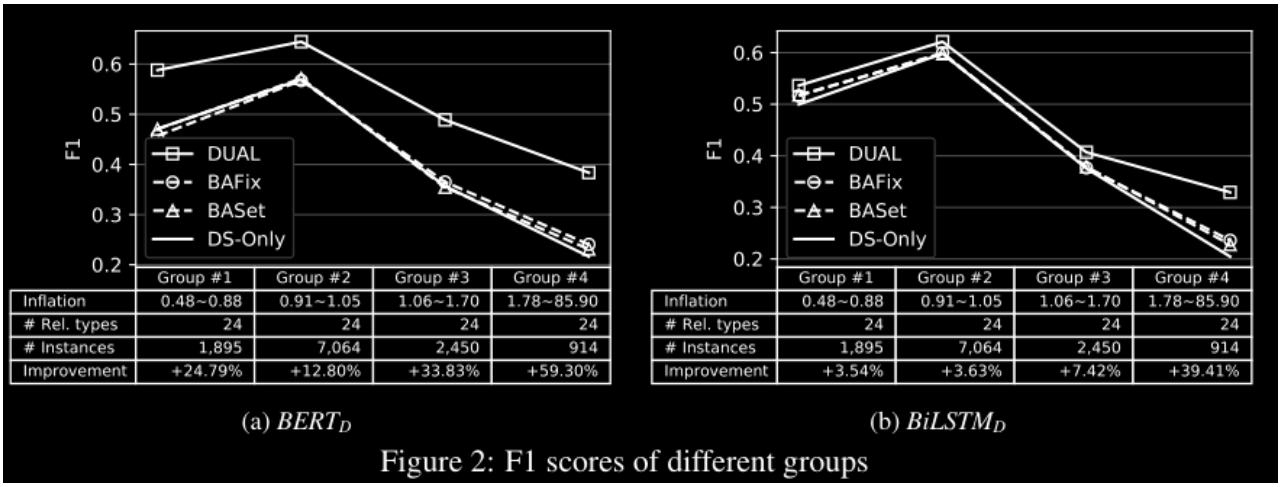
Dataset	KBP						NYT					
RE models	$BiGRU_S$	$PaLSTM_S$	$BiLSTM_S$	$PCNN_S$	CNN_S	$BERT_S$	$BiGRU_S$	$PaLSTM_S$	$BiLSTM_S$	$PCNN_S$	CNN_S	$BERT_S$
HA-Only	0.1984	0.1153	0.1787	0.3410	0.2586	0.1631	0.0884	0.1259	0.1504	0.4463	0.3978	0.1953
DS-Only	0.3909	0.3521	0.3519	0.2705	0.2810	0.3610	0.4532	0.4429	0.4297	0.4177	0.4463	0.4625
BASet	0.3972	0.4055	0.4053	0.2410	0.2400	0.3858	0.4966	0.4555	0.4561	0.3584	0.4358	0.5081
BAFix	0.4241	0.4027	0.3581	0.2931	0.2473	0.3383	0.4613	0.4507	**0.4707**	0.4023	0.4532	0.5145
MaxThres	0.4264	0.3630	0.4053	0.2815	0.2645	0.3751	0.4531	0.4462	0.4350	0.4258	0.4655	0.4952
EntThres	0.4470	0.4018	**0.4248**	0.2925	0.2826	0.3539	0.4553	0.4472	0.4210	0.4154	0.4427	0.4940
DUAL	**0.4749**	**0.4420**	0.4207	**0.3872**	**0.2969**	**0.4013**	**0.5455**	**0.5210**	0.4524	**0.4986**	**0.4744**	**0.5300**

Table 3: Sentence-level RE datasets (KBP and NYT)

文档级别的RE数据集（DocRED）

	Dev					Test				
RE models	$BERT_D$	$BiLSTM_D$	CA_D	$LSTM_D$	CNN_D	$BERT_D$	$BiLSTM_D$	CA_D	$LSTM_D$	CNN_D
HA-Only	0.5513	0.4992	0.4986	0.4817	0.4788	0.5478	0.4982	0.4992	0.4815	0.4681
DS-Only	0.4683	0.4951	0.4890	0.4877	0.4166	0.4587	0.4809	0.4772	0.4713	0.4160
BASet	0.4807	0.5123	0.5024	0.5012	0.4349	0.4716	0.4949	0.4905	0.4905	0.4320
BAFix	0.4802	0.5136	0.5070	0.5166	0.4365	0.4730	0.5061	0.4989	0.4977	0.4354
DUAL	**0.5880**	**0.5510**	**0.5372**	**0.5392**	**0.4967**	**0.5774**	**0.5379**	**0.5306**	**0.5277**	**0.4909**

不同组的F1分数



	Group #1	Group #2	Group #3	Group #4
Inflation	0.48~0.88	0.91~1.05	1.06~1.70	1.78~85.90
# Rel. types	24	24	24	24
# Instances	1,895	7,064	2,450	914
Improvement	+24.79%	+12.80%	+33.83%	+59.30%

(a) $BERT_D$

	Group #1	Group #2	Group #3	Group #4
Inflation	0.48~0.88	0.91~1.05	1.06~1.70	1.78~85.90
# Rel. types	24	24	24	24
# Instances	1,895	7,064	2,450	914
Improvement	+3.54%	+3.63%	+7.42%	+39.41%

(b) $BiLSTM_D$

Figure 2: F1 scores of different groups

更改HA数据的大小

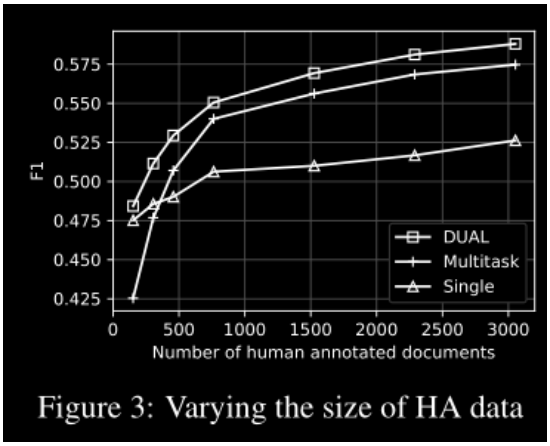


Figure 3: Varying the size of HA data

通过作者的模型框架从文档中进行关系提取的例子

Document	Title: Kungliga Hovkapellet [1] Kungliga Hovkapellet is a Swedish orchestra, originally part of the Royal Court in **[Sweden]**'s capital **[Stockholm]**. [2] Its existence ...	Title: Loopline Bridge [1] The Loopline Bridge (or the Liffey Viaduct) is a railway bridge spanning the River Liffey and several streets in **[Dublin]**, **[Ireland]**. [2] It joins ...
Relations	True label: ⟨**Sweden**, `capital`, **Stockholm**⟩ *DUAL*: ⟨**Sweden**, `capital`, **Stockholm**⟩ *BAFix*: ⟨**Sweden**, `capital`, **Stockholm**⟩ *DS-Only*: ⟨**Sweden**, `capital`, **Stockholm**⟩	True label:NA *DUAL*:NA *BAFix*: ⟨**Ireland**, `capital`, **Dublin**⟩ *DS-Only*:⟨**Ireland**, `capital`, **Dublin**⟩

Table 5: Examples of documents and extracted relations

看内容应该是评估指标 F1和AUC

	F1	AUC
HA-Only	0.6569	0.6456
DS-Only	0.6624	0.6978
DUAL	**0.6930**	**0.7125**

Table 6: Topic-aware RE

泛读总结

1、论文要解决什么问题

我觉得它主要是想要结合人工标注的准确性和远程监督的廉价性，从而对文档级或者句子级进行关系抽取，对于远程监督的噪声问题，作者也通过其它的方式进行了解决，使得最后的label 具有鲁棒性。

2、论文采用了什么方法?

采用了远程监督和人工标注结合的方法，它对框架的输出层进行了改进，提出了四个子网，其中的有些网络能够对远程监督的标记偏差进行改善。

3、论文达到什么效果?

论文提出的框架模型明显优于现有的方法。实验部分可能还需要一些评估指标，但是这里我没有发现。

精读
==========

Introduction

第一段讲了RE模型的应用——知识库建设和问答系统等，RE的目标就是发现文章中实体之间的关系，更加具体的说，就是从文章中提取出三元组——<eh,r,et>
eh是头实体 et是尾实体 r是实体间关系

第二段讲了RE模型需要大量的标签训练数据，它们的形式是文章-三元组对的形式存在的，但是因为人工标注的标签是耗时以及耗钱的，所以Mints等人就提出了远程监督这种方式——通过额外的知识库来产生大量的标签数据，但是因为远程监督的特点，它很容易产生错误的标签问题，例如数据库中存在<U K, capital, London>这样的三元组，那么通过远程监督可能会对这样的句子’ London is the largest city of the UK' 生成上述标签。

第三段主要讲了作者将这两种标记方法进行结合，人工标记human-annotated叫做HA数据，远程监督distantly supervised称为DS数据。
其中有个下划线部分没有搞懂，总之作者认为DS数据中的关系类型均值概率与HA数据中的一样就认为这个关系类型是无偏的。然后通过对文档级别的RE 数据集DocRED检查发现关系类型的膨胀范围从0.48到了85.9。这表明对于某些关系类型，远程监督往往会产生大量的虚假标签。

第四段主要讲了Ye等人之前提出了使用HA data来改善DS data标注偏差问题，它也的确提升了准确率，但是它认为偏差是静态的，但是偏差其实是随着上下文的信息变动的，所以大部分通过远程监督得到的关系标注都是假阳性的（false positive），因此最后作者提到需要考虑上下文信息以及考虑标签偏差来更加准确地提取关系。

第五段其实是主要内容的一个概括，比较重要。
首先，作者认为HA数据和DS数据是非常不同的，所以他们将使用这两种数据的关系提取模型看作是一个多任务学习的问题——所以有HA-Net和DS-Net两个部分。
然后为了能够让HA-Net从DS 数据中进行学习，他们提出了一个附加的损失项——disagreement penalty。
接着使用了u-Net和σ-Net，通过考虑上下文信息来自适应地估计对数正态分布。
最后，作者在文档级别的数据集和句子级别的数据集中验证了它们的有效性，实验表明，它们的模型明显优于其它方法。

Preliminaries

1. Problem Statement

由于一个句子通常描述它们之间的单一关系，因此句子级关系提取通常被视为一个multi-class classification（多分类）问题，一对实体之间可能有多个关系，那么这些关系可能在一个文档中表达，因此文档级的关系提取通常被定义为multi-label classification（多标签分类）问题

定义（sentence-level 关系提取）

原文
For a pair of the head and tail entities eh and et, a
relation type set R and a sentence s annotated with entity mentions, we determine the relation r ∈ R between eh and et in the sentence. Note that R includes a special relation type NA which indicates that
there does not exist any relation between eh and et.

自我理解
对于一对头尾实体eh和et，一个关系型集合R和一个句子s（这个句子被标注提及了上述的eh和et实体），那么我们要决定这个句子中eh和et实体的关系是什么。
需要注意：R关系集当中可以有NA，它表示eh和et之间不存在任何的关系。

定义（Document-level 关系提取）

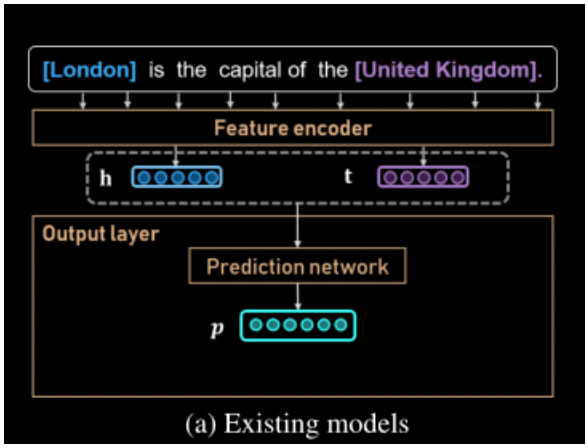
原文

F or a pair of the head and tail entities eh and et,
a relation type set R and a document d annotated with entity mentions, we find the set of all relations
R∗ ⊂ R between eh and et appearing in document d. Note that R does not include NA in this case since
it can be represented by an empty set of R∗.

自我理解

对于一对实体eh和et，一个关系数据集R和一篇文档（这篇文档被标注有上述提到的实体）我们需要发现一个R*数据集（它属于R）这个R*数据集表示在文档d中实体eh和et所有的关系。
需要注意：R关系集中不能有NA，因为它可以用R*的空集表示

2. RE现有工作

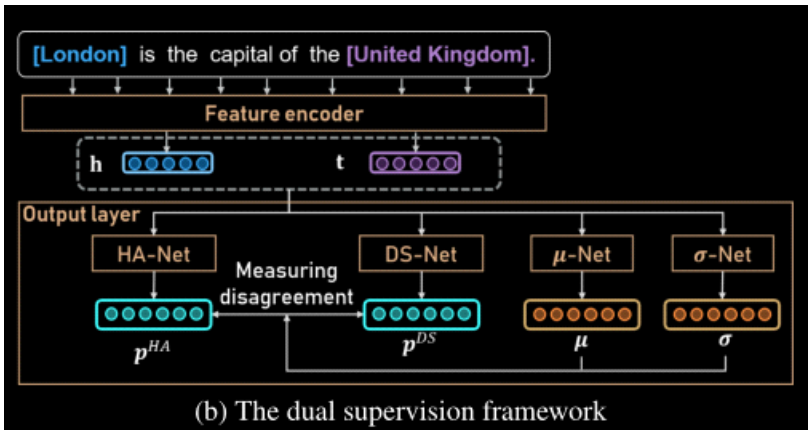


RE model组成：一个特征编码器（a feature encoder） + 预测网络（a prediction network）

feature encoder：将一个text转换成头实体和尾实体这样的隐式表示（以前使用Bi-LSTM、BERT、CNN来编码text，以及位置嵌入等等）

prediction network：输出实体间关系的概率分布。
sentence-level RE使用softmax 分类器作为预测网络，使用分类交叉熵（categorical cross entropy）作为损失函数。
document-level RE使用sigmoid分类器和二元交叉熵（binary cross entropy）

双重监督框架 Dual Supervision Framework



(b) The dual supervision framework

1. 双重监督框架的概述

该框架允许人类标记标签与远程监督标签不同，使用了HA-Net和DS-Net分别在HA数据集和DS数据集中预测标签。同时使用HA-Net从测试数据中提取关系，预测网络的分开也防止了远程监督带来的准确率下降。

为了让HA-Net能够从远程监督标签中学习，引入了一个额外的损失项叫做disagreement penalty，它通过使用对数正态分布的极大似然估计
对HA网络和DS网络输出的不一致进行了建模，此外，通过考虑上下文信息，框架还使用了u网络和σ网络来达到自适应的预估对数正态分布的参数。

公式介绍

$$L_{h,t} = I_{HA} \cdot L_{h,t}^{HA} + (1 - I_{HA}) \cdot L_{h,t}^{DS} + \lambda \cdot L_{h,t}^{DS\text{-}HA}$$

通过对每种类型的数据使用单独的预测网络并引入不一致惩罚，HA-Net从远程监督标签中学习，同时减少对噪声DS数据的过拟合

2. 独立预测网络

HA网络用于从训练数据中预测人工注释的标签，并从测试数据中预测关系
两个网络之间不会共享参数，输出都是R维的一个向量

公式

3. Disagreement Penalty

远程监督标签有偏差，偏差的大小因关系而异，此外，偏差可能会根据许多其他特征而变化，例如头部和尾部实体的类型以及文本的内容。

Distribution of inflations

- [ ] 这里之后可能需要自己计算一下inflations以及计算它的概率

inflations的定义：

Recall that the inflation of a relation type
is the ratio of the average frequencies of the relation type per text in
DS data and HA data, respectively.

个人理解inflations： 一个关系类别的inflation是DS数据和HA数据中每个文本的关系类型的平均频率的比率

为了研究inflations的分布情况
1）作者计算了DocRED数据中的96种关系类型的inflations
2）使用了ks test对DocRED data中的inflations进行检验
因为inflations的范围是[0,+∞)
所以假设有这些分布：对数正态分布、weibull分布、卡方分布、指数分布、正态分布
3）经过检验，得到inflations最有可能属于对数正态分布log-normal
4）基于观察结果，对两个预测网络的输出之间的disagreement penalty进行了建模

Modeling the disagreement penalty

在极大似然估计的基础上发展了disagreement penalty
Xr作为随机变量，表示pDSr与pHAr的比值
由于inflations也是DS数据和HA数据中标签数量的比率，因此Xr表示关系类型r的条件inflation，该flation以具有头部和尾部实体的文本作为条件。
（因为前面假设了inflation是对数正态分布，所以这里Xr也假设是对数正态分布，就可以得到以下公式：

$$f(x) = \frac{1}{x\sigma_r\sqrt{2\pi}} exp\left(-\frac{(\log x - \mu_r)^2}{2\sigma_r^2}\right)$$

而disagreemnet penalty LDS-HA h,t被定义为条件inflation pDSr/pHAr的负对数似然，所以经过推导有下面公式：

$$-\log f\left(p_r^{DS}/p_r^{HA}\right) = \frac{1}{2}\left(\frac{\log p_r^{DS} - \log p_r^{HA} - \mu_r}{\sigma_r}\right)^2 + \log p_r^{DS} - \log p_r^{HA} + \log \sigma_r + \frac{\log 2\pi}{2}.$$

同时，u和σ不能设置为一个定值

4. Parameter Networks

u-Net和σ-Net输出向量 u = [u1,u2, … ,uR] 和 σ = [σ1,σ2, … , σR]，它们同时作为条件inflation的参数表示，

u-Net和σ-Net的结构与预测网络的结构相同，除了输出激活函数之外。
使用tanh函数 hyperbolic tangent function 作为u-Net的激活函数 （因为u既可以是正的，又可以是负的）
使用softplus函数作为σ-Net的激活函数 （总是正的positive）

$$\boldsymbol{\mu} = tanh(\mathbf{h}^\top \mathbf{W}^\mu \mathbf{t} + \mathbf{b}^\mu),\ \boldsymbol{\sigma} = softplus(\mathbf{h}^\top \mathbf{W}^\sigma \mathbf{t} + \mathbf{b}^\sigma) + \varepsilon$$

5. Loss function

句子级别：使用categorical cross entropy loss 分类交叉熵损失作为预测损失
由于这里只需要判断标签是否是属于HA数据还是DS数据，所以这里可以看作是一个二分类问题。

$$L_{h,t} = I_{HA} \cdot L_{h,t}^{HA} + (1 - I_{HA}) \cdot L_{h,t}^{DS} + \lambda \cdot L_{h,t}^{DS\text{-}HA}$$
$$= - I_{HA} \cdot \log p_r^{HA} - (1 - I_{HA}) \log p_r^{DS} + \lambda \left[\frac{1}{2} \left(\frac{\ell_r - \mu_r}{\sigma_r} \right)^2 + \ell_r + \log \sigma_r \right]$$

IHA：如果标签来自HA数据就为1，否则为0
其它内容在本子上待整理

6. 对Disagreement Penalty进行分析

通过比较损失函数中Wha的梯度得到disagreement penalty的影响
对损失函数中的pHAr进行求导
如果是人工标注的标签 求导后的梯度为
推导公式在本子上

$$\nabla L_{h,t} = \nabla L_{h,t}^{HA} + \mathbf{0} + \lambda \nabla L_{h,t}^{DS\text{-}HA} = - \left(1 + \lambda(1+\phi_r)\right) \frac{1}{p_r^{HA}} \nabla p_r^{HA}.$$

如果是远程监督标注的标签 求导后的梯度为

$$\nabla L_{h,t} = \mathbf{0} + \mathbf{0} + \lambda \nabla L_{h,t}^{DS\text{-}HA} = -\lambda \left(1 + \phi_r\right) \frac{1}{p_r^{HA}} \nabla p_r^{HA}.$$

从公式上来看，上述两个梯度的方向是下降的，因为都为-，它表示人工标注标签和远程监督标签对于HA-Net由相同的影响，只不过具体的下降程度不一样（系数不同）所以HA-Net不仅能够从人工标注标签进行学习，也能通过这个disagreement penalty进行学习。

L(u，σ)：符合对数正态分布

7. 文档级RE的扩展

使用二元交叉熵损失函数

$$L_{h,t} = -I_{HA} \left(\sum_{r \in R_{h,t}} \log p_r^{HA} + \sum_{r \in R \setminus R_{h,t}} \log \left(1 - p_r^{HA}\right) \right)$$
$$-(1 - I_{HA}) \left(\sum_{r \in R_{h,t}} \log p_r^{DS} + \sum_{r \in R \setminus R_{h,t}} \log \left(1 - p_r^{DS}\right) \right) + \lambda \sum_{r \in R_{h,t}} \left[\frac{1}{2} \left(\frac{\ell_r - \mu_r}{\sigma_r} \right)^2 + \ell_r + \log \sigma_r \right].$$

在测试时如果pHAr大于在开发数据集上调整的阈值，则模型输出三元组<eh,r,et>

实验

对句子级（Ye等人，2019）和文档级RE (Yao et al., 2019; Wang et al., 2019)进行了性能研究.
环境：pyTorch、 V100 GPU

1、实验设置

数据集：
sent：KBP + NYT
doc：DocRED

比较的方法 methods：
DUAL与BASet和BAFix
对于sent-level，DUAL与MaxThres和EntThres

RE models：
sent-level RE：BiGRUS、PaLSTMS、 BiLSTMS、CNNS、PCNNS、BERTS
doc-level RE：BERTD、 CNND、 LSTMD、BiLSTMD、 CAD 、

2、与已存在的方法进行比较
Sentence-level RE：

实验那一块当中有好多的方法不知道，以及看了文章之后其实有很多的比较是怎么比较的不太懂，所以这里先跳过，直接看Related works这一块，还有这个作者的论文写的很好，我觉得，会有公式提示，以及相关研究的发展。

RE相关工作

2009年Mintz提出远程监督，以克服人类注释标签数量的限制
2010、2011年Riedel和Hoffmann等人使用人工特征来发现文章中的实体关系
由于RE models采用NLP的输入特征，因此由NLP 工具产生的误差被传播到了RE模型。
2014、2015、2016年，为了解决这样的误差问题，使用了神经网络例如CNN、LSTM、和BERT来编码文章 去发现关系，而不是使用 handscrafted features。
2019年 由于许多的关系事实是在多个句子中表达的，因此最近的工作研究了文档级RE，yao等人还提供了DocRED数据集并比较了改编自句子级RE模型
2015、2016、2018、2019年，远程监督中的错误标签问题已经在以前的许多工作中得到了解决。 其中，一种bag-of-sentences-level的方法出现，与已有的RE不同，它对于一些应用例如知识问答方面有一些限制。
2019年 与作者研究工作最相近的就是：Ye的论文，也是使用了人工标注标签和远程监督标签，然而，它们不使用HA数据来训练模型。。。

有实验细节和代码：！！！ 在附录当中

额外知识：

k-s检验：检测数据是否满足一定的统计分布，ks test是一个有用的非参数假设检验，主要用来检验一组样本是否来自于某个概率分布，或者比较两组样本的分布是否相同。

对数正态分布：
如果一个随机变量的对数服从正态分布，那么该随机变量服从对数正态分布。

wikidata：英文开源知识图谱
wikidata是一个大型数据库，存储了维基百科、Freebase中的海量信息，为了便于机器识别、算法调用，在存储时Wikidata将数据结构化成了固定的格式——RDF。

项： Q300918

标签： cat

描述： Unix utility that concatenates and lists files

条目： part of GNU Core Utilities

instance of standard UNIX utility

wikidata支持的是以三元组为基础的知识条目items的自由编辑，一个三元组代表一个关于该条目的陈述（statements）
每个实例对应知识图谱中的一个项item，比如上图就是一个项，对应的实例就是一个Linux命令“cat”
每个项都有标签label、描述description、别名aliases...
每个项中的具体数据被称为条目statement，一个实例可以有多个条目，表现了实例不同方面的特征，条目由属性property和数值value构成。例如上图中的cat命令包含了两个条目，其中一个条目的属性为part of，数值为GNU Core Utilities

Wikipedia：维基百科
是一个基于维基技术的多语言百科全书协作计划，用多种语言编写的网络百科全书。

DocRED： A Large-Scale Document-level Relation Extraction Dataset
大型文档级关系抽取数据集
是一个从Wikipedia和Wikidata构建的大规模人工标注的文档级RE数据集，具有以下特征：
(1)DocRED包含132375个实体和56354个关系事实，标注在5，053个维基百科文档上，使其成为最大的人工标注文档级RE数据集。
(2)由于DocRED中至少有40.7%的关系事实只能从多个句子中抽取，因此DocRED需要阅读文档中的多个句子来识别实体，并通过综合文档的所有信息来推理其关系。这使得DocRED区别于那些句子级的RE数据集
(3)还提供了大规模的远距离有监督数据来支持弱监督的RE研究。

Kungliga Hovkapellet

[1] ***Kungliga Hovkapellet*** (The ***Royal Court Orchestra***) is a *Swedish* orchestra, originally part of the *Royal Court* in ***Sweden***'s capital *Stockholm*. [2] The orchestra originally consisted of both musicians and singers. [3] It had only male members until *1727*, when *Sophia Schröder* and *Judith Fischer* were employed as vocalists; in the *1850s*, the harpist *Marie Pauline Åhman* became the first female instrumentalist. [4] From *1731*, public concerts were performed at ***Riddarhuset*** in *Stockholm*. [5] Since *1773*, when the ***Royal Swedish Opera*** was founded by *Gustav III* of ***Sweden***, the ***Kungliga Hovkapellet*** has been part of the opera's company.

Subject: ***Kungliga Hovkapellet; Royal Court Orchestra***
Object: ***Royal Swedish Opera***
Relation: part_of **Supporting Evidence:** 5

Subject: ***Riddarhuset***
Object: ***Sweden***
Relation: country **Supporting Evidence:** 1, 4

DocRED中的一个样本

DocRED: 大型文档级关系抽取数据集 - 知乎 (zhihu.com)

双曲正切函数Tanh函数

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$