Looking Beyond Label Noise: Shifted Label Distribution Matters in Distantly Supervised Relation Extraction

Qinyuan Ye*
University of Southern California
qinyuany@usc.edu

Liyuan Liu*
University of Illinois, Urbana-Champaign
ll2@illinois.edu

Maosen Zhang
Purdue University
maosenzhang.milo@gmail.com

Xiang Ren
University of Southern California
xiangren@usc.edu

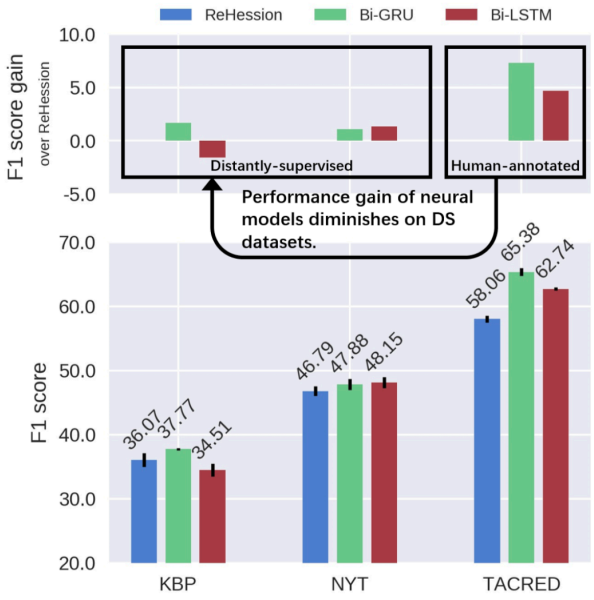
EMNLP2019 深度好文

这才是应该发表的好文章啊！

这篇是DS的，然后是sentence-level的。用的dataset包括：KBP/NYT/TACRED。其中前两个属于DS的，后面的属于manual annotated的。用的model包括feature-based和neural-based的。使用softmax得到relation 分类分布。

$$p(y = r_i|\mathbf{h}) = \frac{\exp(\mathbf{r}_i^T\mathbf{h} + b_i)}{\sum_{r_j}\exp(\mathbf{r}_j^T\mathbf{h} + b_j)},$$

paper探究的核心就是：**为什么Neural方法相对feature-based方法，在DS的提升没有人工数据集中那么高?**



于是paper就来探究这个的原因。首先试验了两种启发式方法，

- **Max Threshold** introduces an additional hyper-parameter T_m, and adjusts the prediction as (Ren et al., 2017):

$$predict(\mathbf{h}) = \begin{cases} r^*, & p(y = r^*|\mathbf{h}) > T_m \\ \text{NONE}, & \text{Otherwise} \end{cases}.$$

- **Entropy Threshold** introduces an additional hyper-parameter T_e. It first calculates the entropy of prediction:

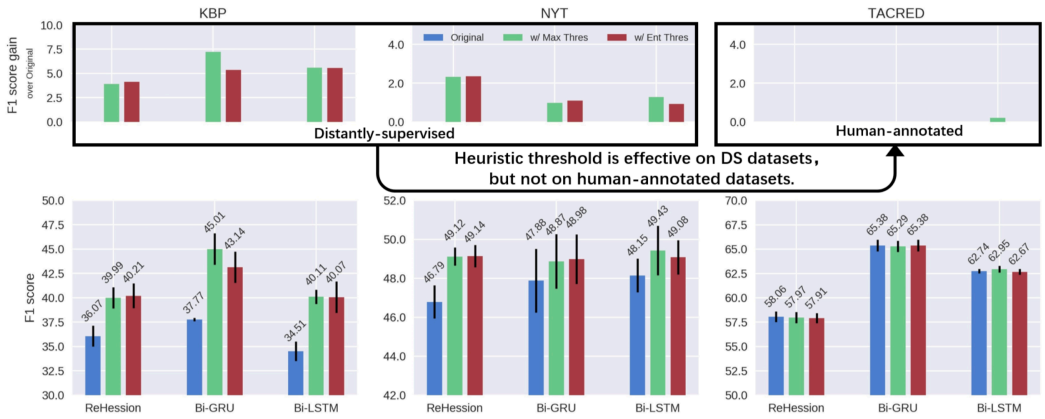
$$e(\mathbf{h}) = -\sum_{r_k} p(y = r_k|\mathbf{h}) \log p(y = r_k|\mathbf{h}),$$

then it adjusts prediction as (Liu et al., 2017):

$$predict(\mathbf{h}) = \begin{cases} r^*, & e(\mathbf{h}) < T_e \\ \text{NONE}, & \text{Otherwise} \end{cases}.$$

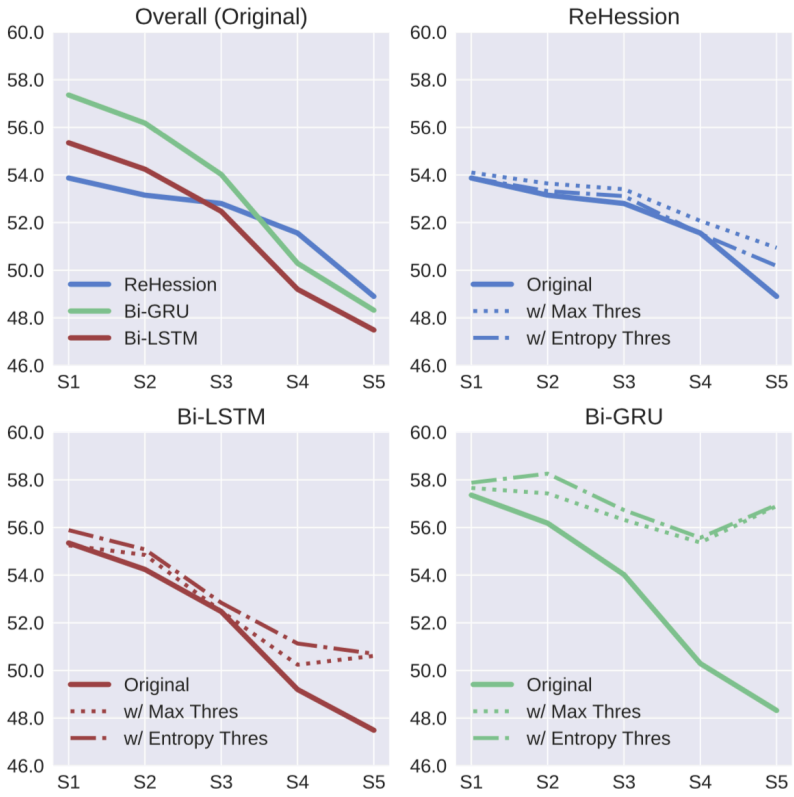
其实就是相当于对model的prediction结果直接做后处理，去掉model把握不太高的prediction。

至于里面Tm和Te这两个超参数，paper说是在test搞20%出来作为额外dev，然后在剩下的80%测。
这么做是必须的，这是连接train和test的bridge。
然后呢发现**在DS上对Neural方法有很大提升，而在人工数据集上没有**。



说明这种拒绝策略对Neural方法来说，在DS数据集上更有效，能够解决DS数据集带来的一些噪音。

作者受此启发，推测可能是因为**shifted label distribution**问题：**训练集和测试集的标签分布不一致！**（注意不是标签样本数不平衡！）
为了验证这一猜想，作者做了一个实验。改变**训练集**的标签分布，把原来的分布叫做S1，然后搞了一个随机分布的叫做S5，然后中间线性插值成S2～S4，并保证他们的数量差不多一致。
之后实验看训练效果发现



可见，这个shifted label distribution problem确实会显著影响model的performance。

之后，paper针对这个进行理论推导，得到一个理论上而不是启发式的手段。

首先把训练集叫做Dd，测试集叫做Dm。

推导基于的假设是：

$$p(\mathbf{h}|r_i, \mathcal{D}_m) = p(\mathbf{h}|r_i, \mathcal{D}_d) = p(\mathbf{h}|r_i).$$

推导出

$$
\begin{aligned}
& p(y=r_i|\mathbf{h},\mathcal{D}_m) \\
=& \frac{p(y=r_i|\mathbf{h},\mathcal{D}_d)\frac{p(r_i|\mathcal{D}_m)}{p(r_i|\mathcal{D}_d)}}{\sum_{r_j} p(y=r_j|\mathbf{h},\mathcal{D}_d)\frac{p(r_j|\mathcal{D}_m)}{p(r_j|\mathcal{D}_d)}} \\
=& \frac{\exp(\mathbf{r}_i^T\mathbf{h}+b_i')}{\sum_j \exp(\mathbf{r}_j^T\mathbf{h}+b_j')},
\end{aligned}
$$

where

$$
b_i' = b_i + \ln p(r_i|\mathcal{D}_m) - \ln p(r_i|\mathcal{D}_d).
$$

具体过程可以去看origin paper。

这个什么意思呢？就是我们在由h做r的prediction的时候，可以把偏置项改成这个，从而来弥补shifted label distribution。

- 对于p(ri | Dd) 的先验，我们可以在训练集上统计得到
- 对于p(ri| Dm)，我们预留20%的test来估计。

然后paper还提出了两种应用方式：

- BA-Set：在train的时候不变，evaluate的时候把b变成b‘
- BA-Fix：在train的时候，bi = ln p(ri | Dd)，到了evaluate的时候换成bi' = ln p(ri | Dm)

然后最后还和attention的那种DS方法比较效果。当然最后evaluate都是sentence-level的。

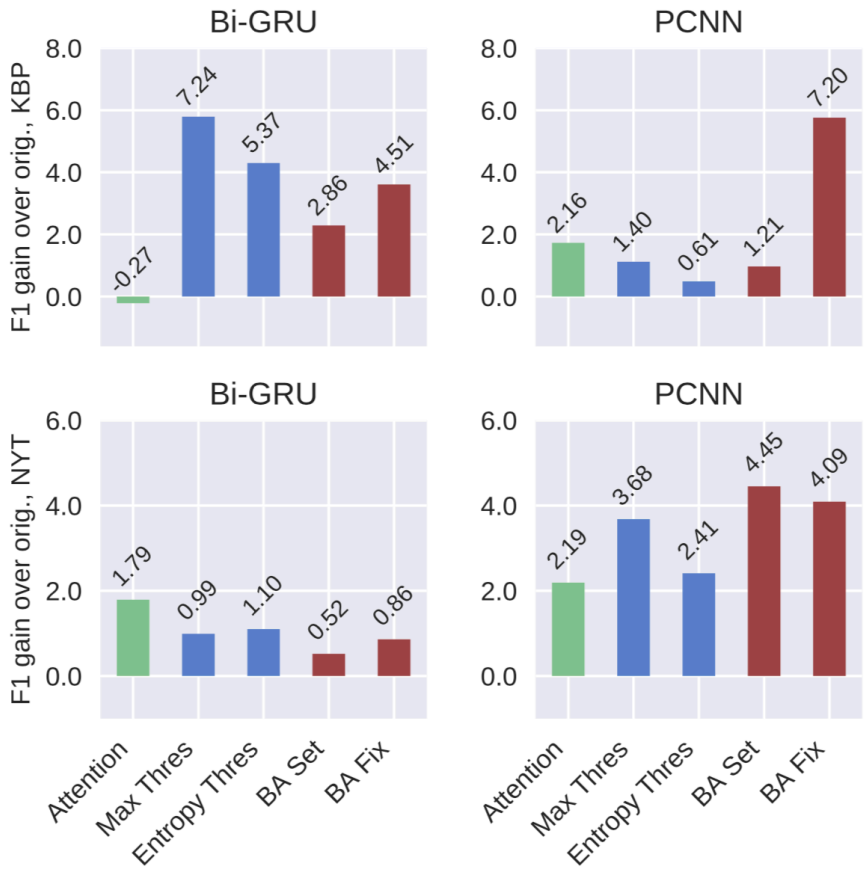


Figure 6: Comparison among selective attention, threshold heuristics and bias adaption approaches. Threshold heuristics and bias adaption approaches bring more significant improvements in some cases, indicating that shifted label distribution is a non-negligible problem.

虽然说这个需要预留test 20%来做估计，泄露了test dataset的信息，理应更高。

但是**仅仅通过计算这点信息，就可以得到如此高的提升，还是很牛逼的**!

对了，为什么说DS的会不一致。

因为在这篇paper里，DS的训练集是DS的，但是测试集是人工标的。