第一章：introduction

第二章：相关技术

主流方法：

深度学习

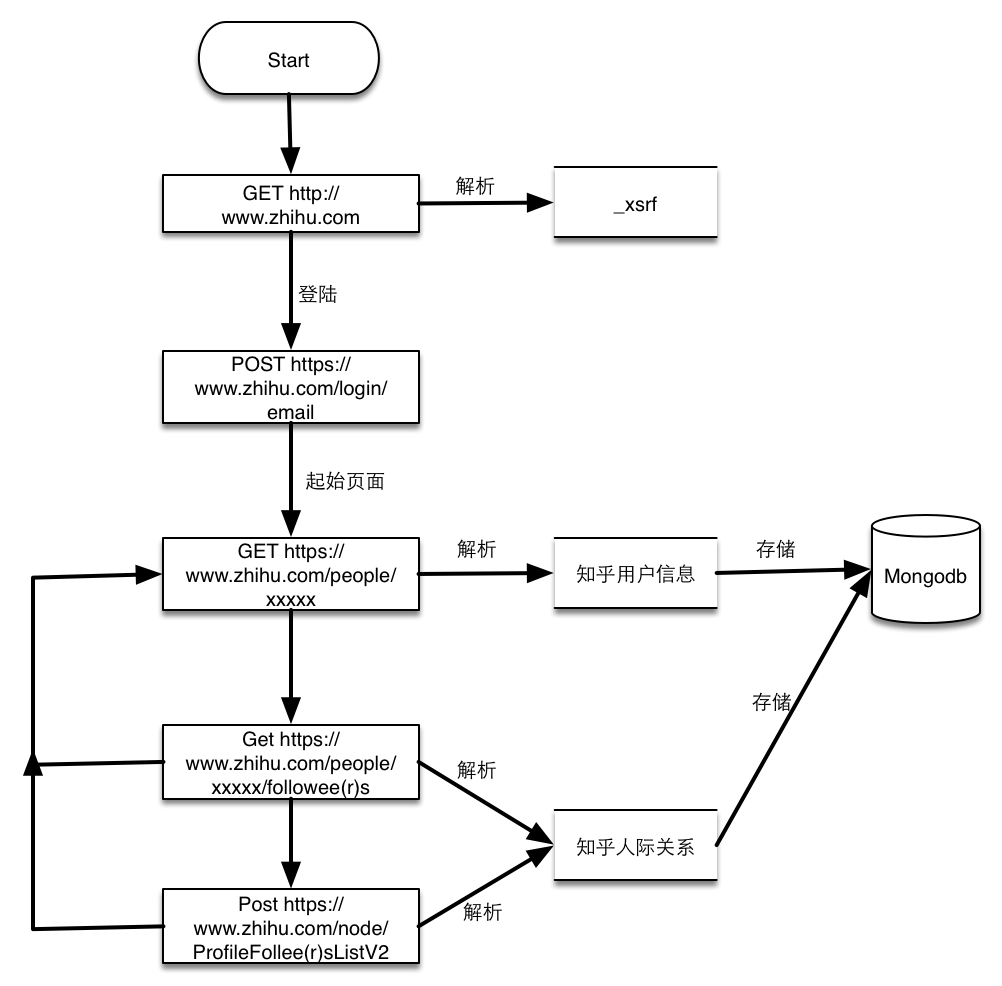
监督Semeval 2010 task 8

远程监督 NYT 10 基于注意力

对抗训练

第三章：新闻事件爬取

本章介绍爬虫技术，爬取网站以及爬取得到的新闻数据如何存储。



第四章：对抗训练

实体关系抽取技术：

监督

半监督：利用无标签数据，训练有标签样本，获得分类更准确的分类器。

远程监督

embed：word embed + pos embed + ner embed +位置embed

encoder：获取句子信息

PCNN BILSTM BERT

第五章 基于对抗训练的实体关系抽取

对比了PCNN、PCNN-ATT、RNN、RNN-ATT、BERT、

事件实体关系抽取：

远程监督缺点：噪声大，且关系种类复杂，不适用于新闻事件实体关系抽取。

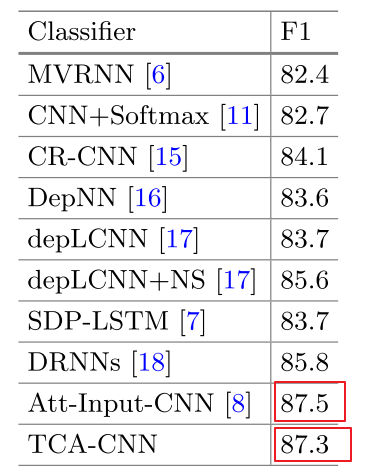
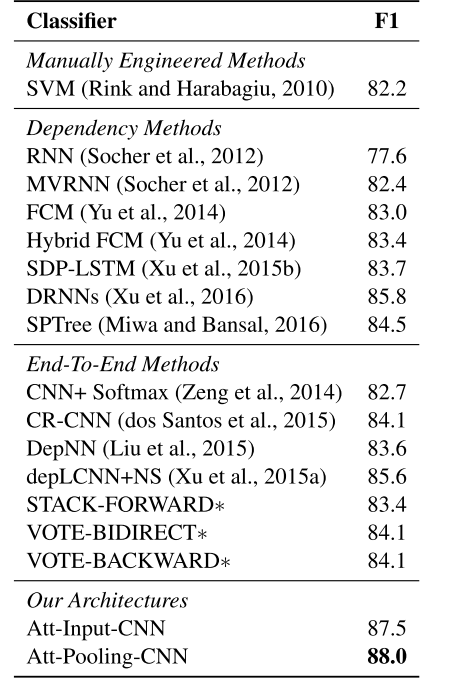
针对特定关系扩充数据的生成对抗方法更好。

事件实体关系旨在找出事件中实体对的关系，现存的实体关系抽取数据集需要大量的人工标注，费时费力，导致数据集数据量和关系种类太少。基于远程监督的实体关系抽取是一种有效的解决方法来获取大量关系数据，但是现有的方法都基于精心设计的规则和远程数据库如Freebase，这些方法包含大量的噪声，且关系种类过于复杂。为了解决这一问题，我们构建了semeval task8关系抽取数据集的拓展，针对该数据集的19种关系类型构建关系数据句子候选集，利用生成对抗网络来进行候选集合中句子的选择，从而过滤掉那些噪声数据。基于我们生成的拓展数据集，我们提出了基于bert表示的生成对抗关系抽取模型，实验表明在这个拓展的数据集上我们的模型明显优于其他现有的模型，并且这种生成对抗训练方法能够有效的提升现有模型的效果。

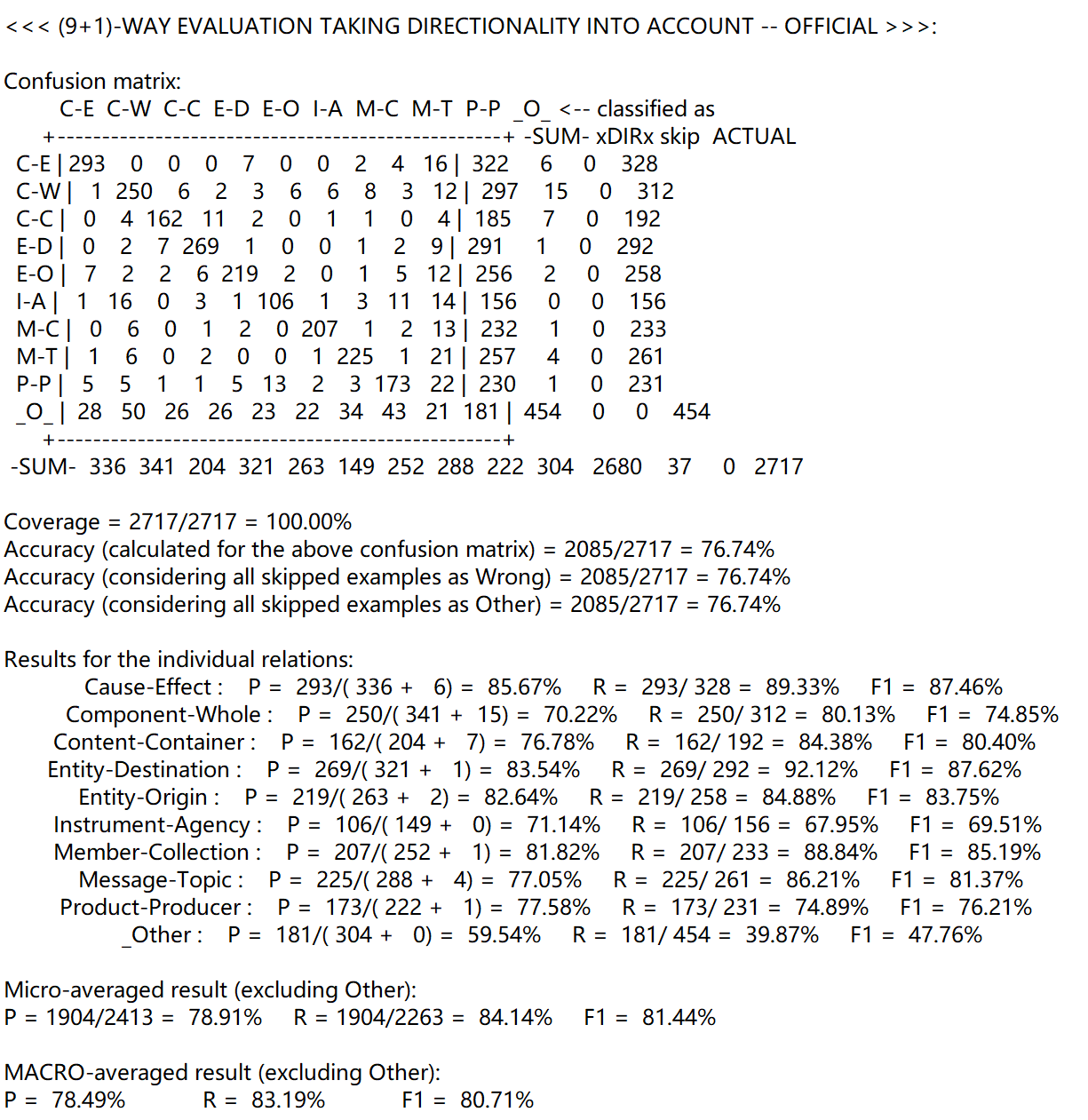
数据集：ace2005，semeval-2010 task8。

可以利用已有关系三元组构建知识库，基于注意力机制的KG选择模型，作为特征引入到关系分类中去。灵感来源于[Hongming Zhang]

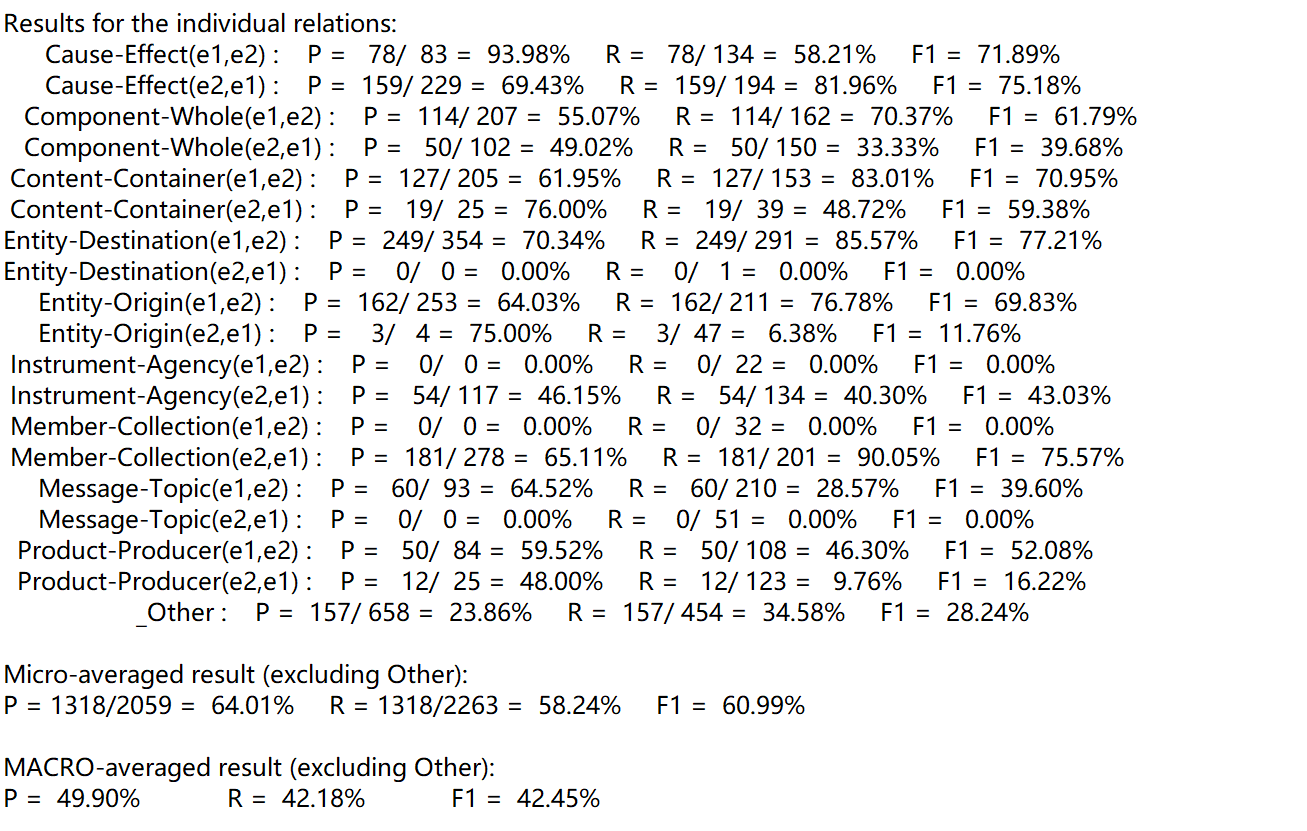
目前F1值：

PCNN结果：



把word\_embedding换成bert



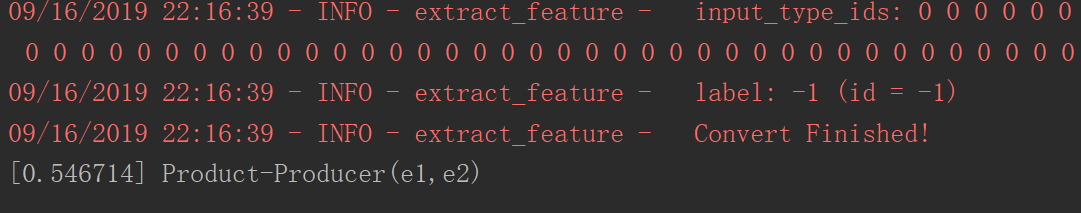
结果分析：样本多的关系P、R、F1准确率较高，样本少的明显偏低，因为样本太少没法学习到这些样本的特征。因此通过GAN网络根据样本分布生成一些数据，扩充这些样本少的数据，

1,提出了GAN用于数据增强。2,改变了一些模型结构用bert做embedding，更加充分学习token-lever语义信息。

study case：

给出一个句子和两个实体，预测实体关系：

sen = 'a child is told a lie for several years by their parents before he realizes that a santa claus does not exist'



关系分类器：



第六章：总结

参考文献：

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7469028>

Event Relations Extraction Based on Event Co-occurrence Network. Tao Liao1 , Peipei Sun1 , Zongtian Liu2.

<http://delivery.acm.org/10.1145/2340000/2339704/p1104-ritter.pdf?ip=59.64.129.128&id=2339704&acc=ACTIVE%20SERVICE&key=BF85BBA5741FDC6E%2E66A15327C2E204FC%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1567655502_d4f504bba0e1ff5ddbe298c12763cf0d>

Open Domain Event Extraction from Twitter. Alan Ritter，Mausam，Oren Etzioni.

工作总结：开放领域的事件抽取。

先标注了1000条推特，19484个token。按照TimeBank的形式。把事件抽取当作序列标注任务。用CRF模型来做。特征利用了上下文等。event terms字典从wordnet获得。

在公开数据集TimeBank上说明模型效果。

事件类型分类：即使大的数据集包含的事件类型种类也可能很少，所以用无监督的方式。

每一个event trigger都是一个mixture of types。

每一个event type都是实体分布

设置100个type，观察每个type下的event trigger，有实际意义的就起名字，没意义的标注为other。

<https://arxiv.org/pdf/1907.03663.pdf>

Knowledge-aware Pronoun Coreference Resolution.Hongming Zhang♣∗ , Yan Song♠, Yangqiu Song♣, and Dong Yu♠

工作总结：

结合KG做指代

一个代词和若干个目标名词，用BILSTM做词向量。

如何找有效的知识库知识，首先名词作为三元组的头，对应到知识库中找相关三元组，用tail的平均词向量作为一条该名词的知识，可以得到若干条词向量知识。

这样得到的知识太多，而且没用的也很多。因此通过KG ATTENTION模型来选择有效知识。计算每条知识权重，然后w\*k累加得到加权和，作为最终知识。

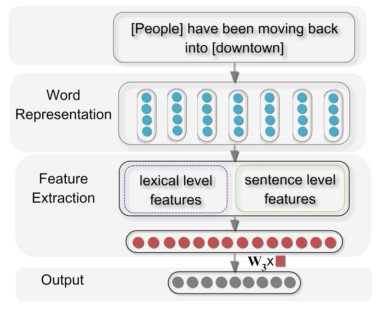
<https://arxiv.org/pdf/1601.00770.pdf>

Makoto Miwa, Mohit Bansal.End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures.

<https://aclweb.org/anthology/C14-1220>

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou and Jun Zhao.Relation Classification via Convolutional Deep Neural Network

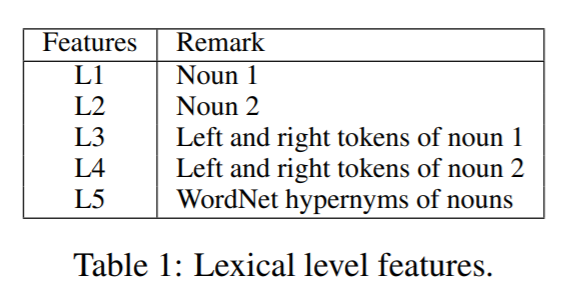
模型结构：



word embedding：不用word2vec，而是(Turian 2010ACL Word representations: a simple and general method for semi-supervised learning.)提供的word embedding.

两种features：词法级别和句子级别。

词法级别特征为5个部分，该特征直接concat到embedding后



句子级别特征：用CNN提取句子级别特征。

lexical feature：[batch,6] 6个id为e1，e1左右两侧单词id，e2,e2左右两侧单词。

embed->[batch,6,word\_dim] -> [batch,6\*word\_dim] 需要得到e1和e2的pos

sentence feature：由sentence id和posleft\posright组成，每个单词有其对应词id和与e1\e2的pos。

sentence id转变为对单词word embed -> [batch,seqlen,word\_dim]

posleft\posrigth也做embed pos ->[batch,seqlen,pos\_dim]

三者cat成sentence feature 到cnn层做卷积得到sentence embed

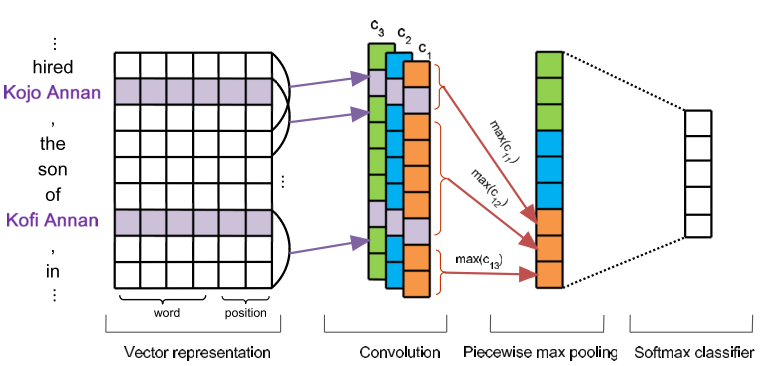
最终lexical和sentence embed cat输入到linear层变换分类，计算交叉熵。

结果：

样本少的和OTHER类别F1值较低，影响了整体F1。

Zeng (2015). Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. EMNLP

改进PCNN：



Pooling层不直接用Maxpooling，而是以实体为边界，将句子分成3个部分，分别做Maxpooling。

弱监督：

Zeng (2015). Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. EMNLP

总结：远程监督 数据集用了NYT和Freebase

方法PCNN+MIL Multi-instance learning

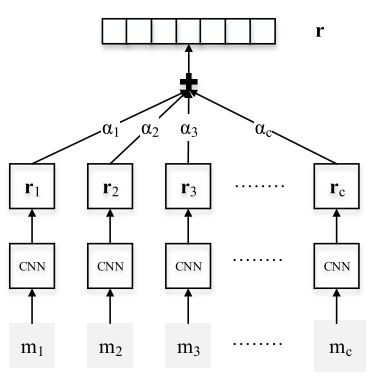
将所有包含实体对的句子当作一个bag，这个bag有一个relation标签。对每一个bag算是这个分类概率最高的句子。

<http://nlp.csai.tsinghua.edu.cn/~lyk/publications/acl2016_nre.pdf>

Yankai Lin1 , Shiqi Shen1 , Zhiyuan Liu1,2∗ , Huanbo Luan1 , Maosong Sun1.Neural Relation Extraction with Selective Attention over Instances.

在PCNN+MIL的基础上引入了att机制，

模型结构：



embedding layer 主要包含word embedding和pos embedding

输入句子、位置等信息，得到词向量、位置向量

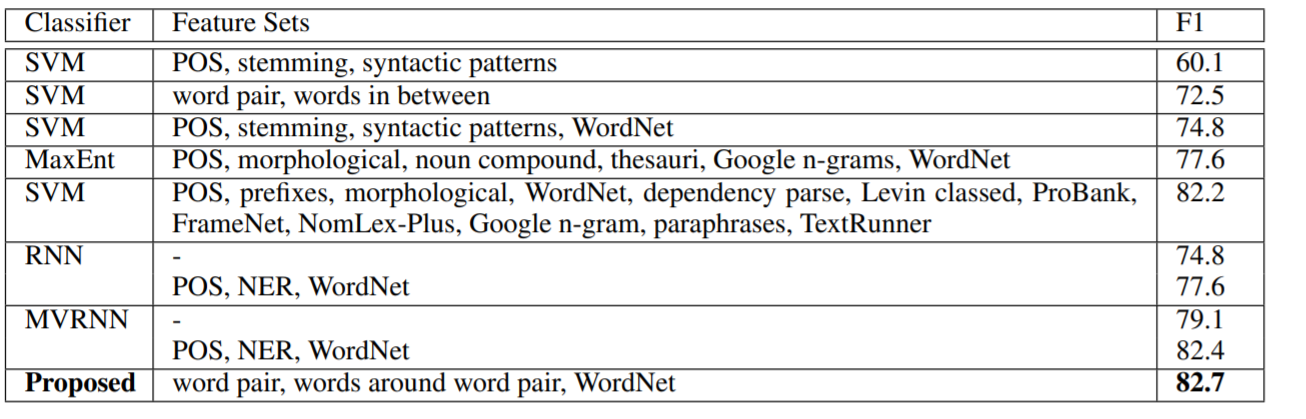
编码器 PCNN、CNN、RNN等 对嵌入的向量进行编码得到深层语义信息。

选择器 Attention 层 等

分类器 SoftMax 进行分类得到最终关系

[Lin et al., 2016] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural Relation Extraction with Selective Attention over Instances. In Proceedings of ACL.[[pdf]](http://thunlp.org/~lyk/publications/acl2016\_nre.pdf)

[Zeng et al., 2014] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In Proceedings of COLING.



[Zeng et al.,2015] Daojian Zeng,Kang Liu,Yubo Chen,and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of EMNLP.

SemEval-2010 Task 8: Multi-Way Classificationof Semantic Relations Between Pairs of Nominals. Iris Hendrickx∗, Su Nam Kim†, Zornitsa Kozareva‡, Preslav Nakov§,Diarmuid ́O S ́eaghdha¶, Sebastian Pad ́o‖, Marco Pennacchiotti∗∗,Lorenza Romano††, Stan Szpakowicz

(15) (PDF) SemEval-2010 task 8. Available from: https://www.researchgate.net/publication/271452073\_SemEval-2010\_task\_8 [accessed Oct 16 2019].

冯冲, 康丽琪, 石戈, 黄河燕. 融合对抗学习的因果关系抽取. 自动化学报, 2018, 44(5): 811−818.

利用GAN生成因果关系数据，

<https://arxiv.org/pdf/1904.05255.pdf>

Peng Shi and Jimmy Lin.Simple BERT Models for Relation Extraction and Semantic Role Labeling.

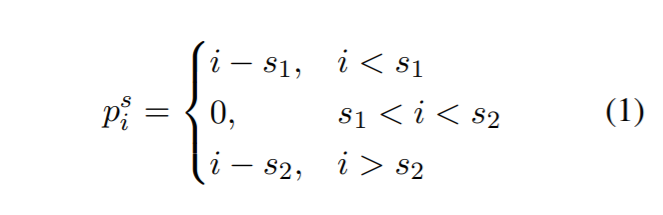
数据：TACRED

sen格式：[cls] sen [sep] e1 [sep] e2 [sep]

其中句子里的实体用S-PER O-LOC代替。表明关系里的主体/客体，类型。

pos embed：相对位置嵌入，对于SEN里的每一个token，计算其与subject和object的位置。

计算公式如下：



pos embed维度为20，concat到句子embedding后得到[batch,seqnlen,768+20]

然后输入给BILSTM。最终输入给MLP分类。

eg：句子[CLS] [S-PER] was born in [O-LOC] [SEP] Obama [SEP] Honolulu [SEP]

语义角色标注：conll2005,09,12

谓语检测，谓语分类(动词如go)，argument是识别，argument分类。

当作序列标注任务，输入句子 和 谓语标志符0/1，输出分类结果，对谓语分类

元素提取：输入(sen,v)句子和谓语对，

标注形式BIO。

eg：

[CLS] Barack Obama went to Paris [SEP] went [SEP]

<https://nlp.stanford.edu/pubs/zhang2017tacred.pdf>

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, Christopher D. Manning. Position-aware Attention and Supervised Data Improve Slot Filling.

数据：TACRED

方法：

现在CNN、RNN等模型问题：没有位置信息。

实验细节，依概率将token置为unk，概率设为0.06，entity mask将entity token变为sub/obj-<NER TAG>加入了POS和NER TAG。

数据实例：

id、relation、token、实体位置、实体类型、pos、ner、deprel、head标注

处理过程：

词向量使用glove 300d

模型：

embed：wordembed+pos embed + ner embed

LSTM：BiLSTM，隐藏层维度为

Linear：从hidden映射到num\_class

模型数据流：

输入words [batch,seqlen] wordid，填充的为0

masks[batch,seqlen] 真实数据为0，填充的为1

pos[batch,seqlen] 斯坦福pos tag，填充为0

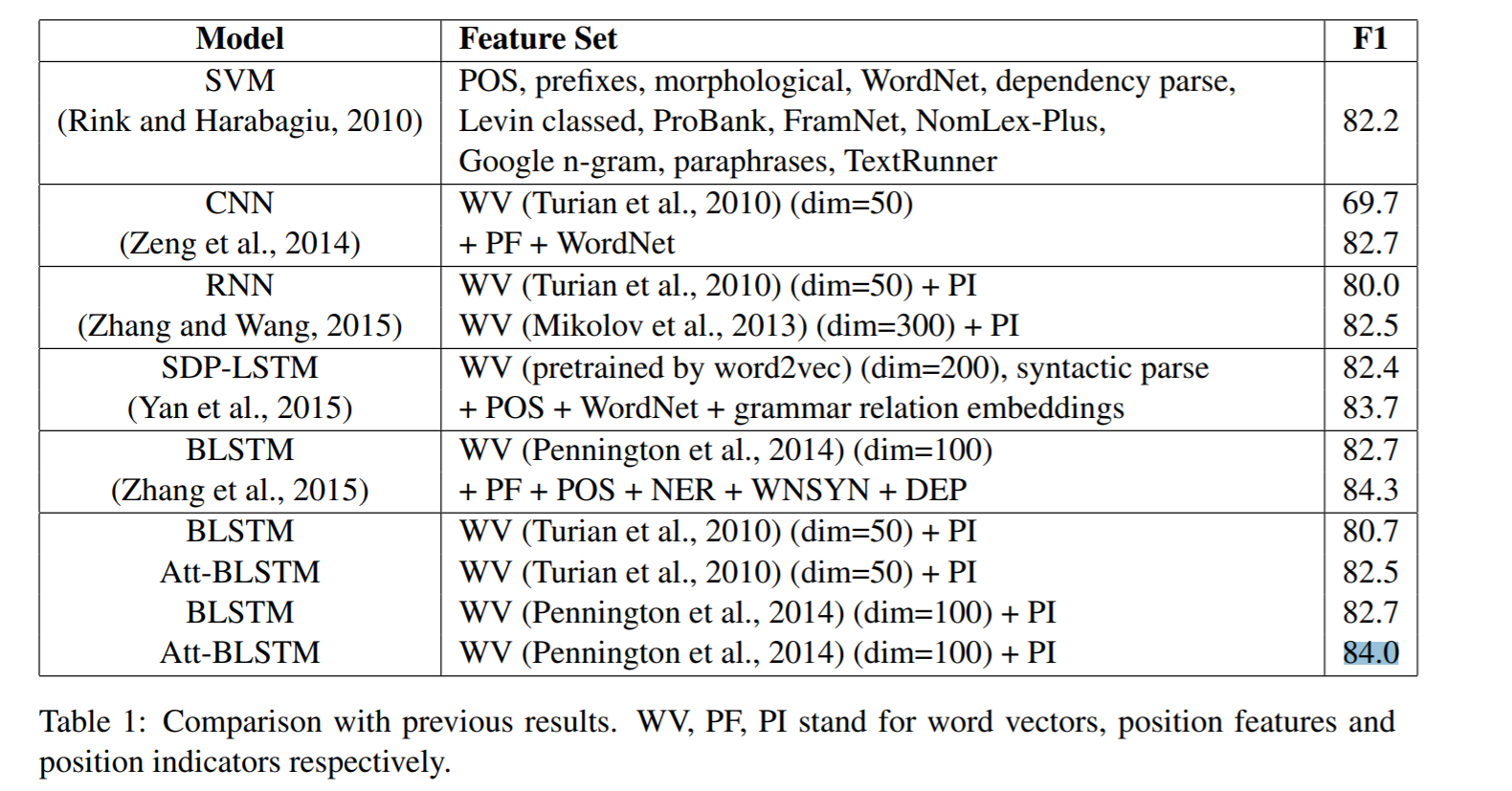
subj\_pos[batch,seqlen] 位置信息，计算公式如上篇论文，填充为0

将word embed\pos\ner cat后，输入到BiLSTM获得隐层向量ht和句子表示out

out,(ht,ct) = nn.BiLSTM()

<https://www.aclweb.org/anthology/P16-2034.pdf>

Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification



google bert RE

Livio Baldini Soares Nicholas FitzGerald Jeffrey Ling∗ Tom Kwiatkowski.Matching the Blanks: Distributional Similarity for Relation Learning.

<https://www.aclweb.org/anthology/D17-1187.pdf>

Adversarial Training for Relation Extraction. Yi Wu, David Bamman, Stuart Russell

<http://nlp.csai.tsinghua.edu.cn/~lzy/publications/coling2018_amnre.pdf>

刘知远 对抗训练RE

Adversarial Multi-lingual Neural Relation Extraction Xiaozhi Wang1∗ , Xu Han1∗ , Yankai Lin1 , Zhiyuan Liu1† , Maosong Sun1,2.

对于一个实体对，包含该实体对的句子有Sn个，每个句子的关系为r，

DIScriminator：用于识别句子属于哪个language

用两层MLP分类。

encoder的目的：encode出dis无法分类的句子。

如果两个实体具有某种关系R，则所有instances包含这两个实体的句子都被标注为R。这样就构成了一个候选set。通过对抗训练，从这个set中选择出有效的句子，并且能够提升ENCODER的性能。将数据集划分为reliable和unreliable两个set，设计分类器和生成器。

分类器目的：识别一个instance是否标注正确，

生成器：选择最具有迷惑性的instance来扰乱分类器的分类效果。

分类器训练数据由两部分构成：reliable data作为正例，生成器选择的例子作为负例

训练过程中，生成器选择噪声数据来提升分类器性能，分类器影响生成器选择更具有迷惑性的例子。当这种对抗训练达到平衡时，分类器提高了对噪声的抵抗力，能够更加有效的分类实体关系。生成器也能够选择有效的例子给分类器。

实验证明这种方法能够有效获得大量特定关系数据，且能够有效减小噪声带来的影响，这个模型比现有其他模型效果要好。刘知远组提出了对抗训练用于关系抽取，他们的方法是获得多语言一致性信息。

框架：encoder、对抗训练、应用

每个instance都标注有实体对<e1,e2>和关系R，如果没有关系则标为NA

3.2encoder

PCNN

BILSTM

BERT

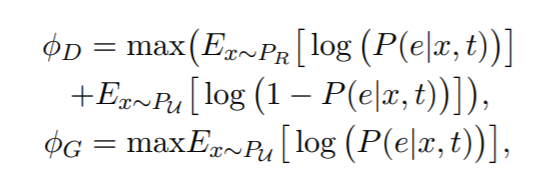
3.3对抗训练

包含一个dis和一个generator，

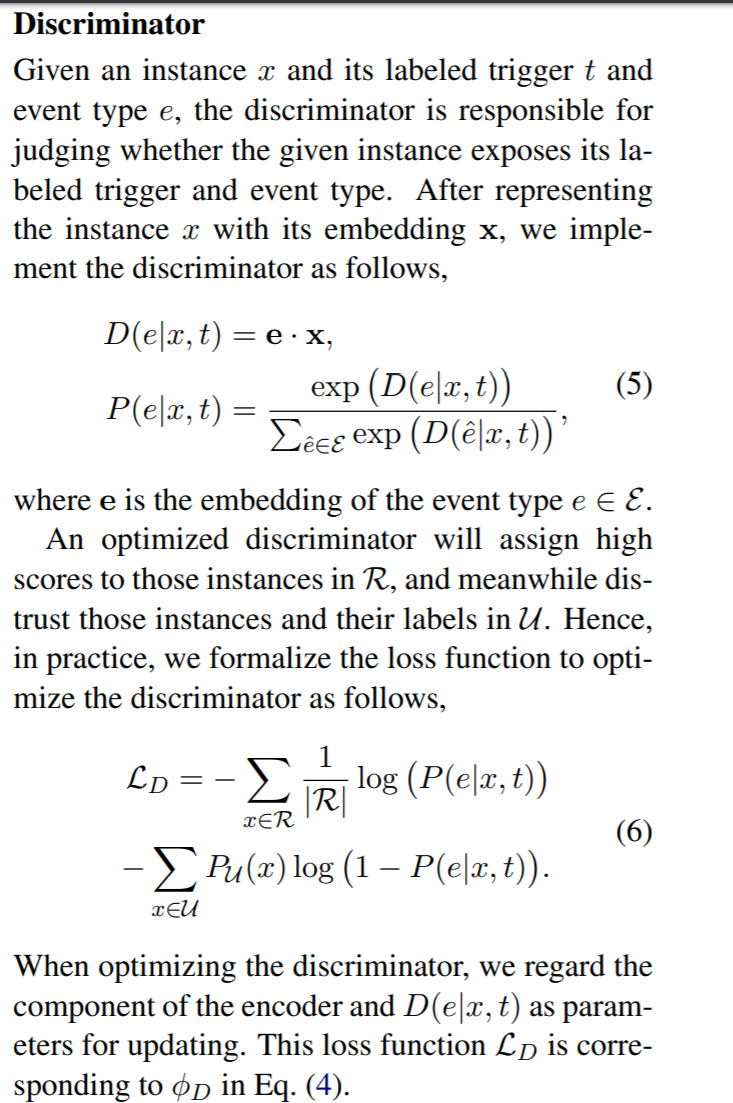
生成器从unreliable instances中选择一个输入到分类器中

集合U中的实例，被标注错误的概率是一定的。所以分类器的目标函数：是最大化两个条件概率P(r|x, e1,e2), x ∈ R and 1 − P(r|x, e1,e2), x ∈U。

生成器选择P(e|x, t), x ∈U，整个训练过程如下所示：

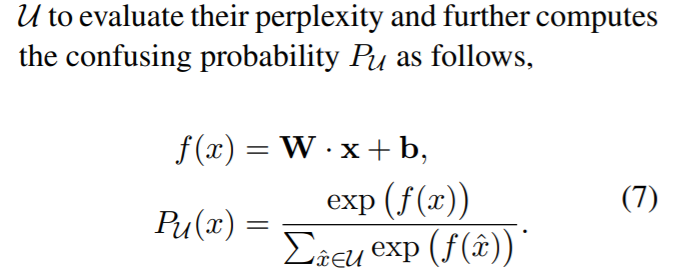


经过足够的训练后，生成器和分类器达到平衡，

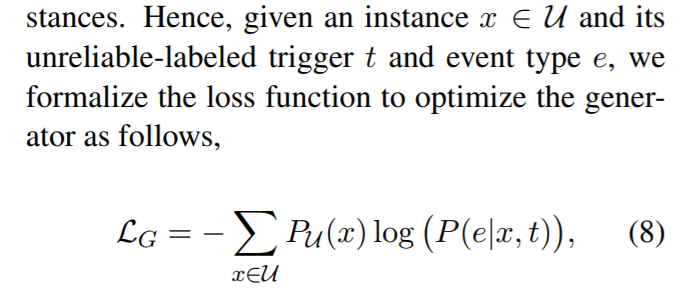


生成器:

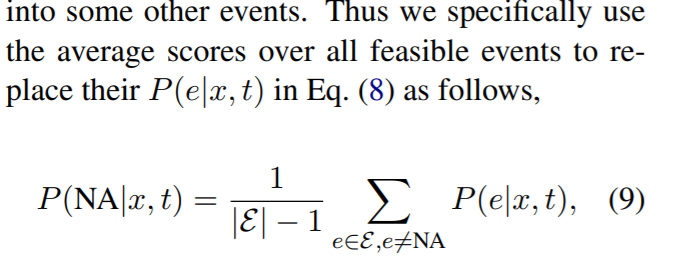
计算每个Instance in U的迷惑率，



X是embedding



有些instance in U 标记为NA，也可能给错预测成其他relation。



为关系类型数。

3.4 extend datasets

本节介绍自动标注方法和如何分成R和U两个set用于对抗训练。

自动标注方法：假设有一对实体在训练样本中被标记为关系r，则所有包含这两个实体的句子都被标记为r。这种方法虽然简单，但是可以获得更大范围的大量数据用于对抗训练。且可以针对某对实体或某种关系进行特定的数据扩充，以解决样本不均衡的问题。

半监督方法：

先用小范围已经标注的样本预训练encoder和分类器，使得他们具有分类关系的能力。然后再基于扩充策略扩大数据集，用预训练的encoder和分类器对所有实例进行标注，构成U集合。有标注的集合为R。这样就可以进行对抗训练了。训练一定epochs后，被生成器选择且被分类器正确分类的instance就从U划分到R。从而扩充标注数据集。

远程监督：

用auto-labeled的数据pretrain encoder和分类器。对所有实例进行打分，高于阈值的为R，低的为U。

实验：

半监督

远程监督

扩大训练集，训练后，用原始测试集测试。

We extend the ACE-2005 training set with the new dataset, and then test the models trained on the extended training set on the original test set. Our models trained on the original training set are named DMCNN and DMBERT, and our bootstrapped models trained on the extended dataset are named DMCNN+Boot and DMBERT+Boot.

证明了这种方法能够提升P、R、F1。

如何证明扩充的训练集是准确的？

用现有的模型（或者manual evaluation）来预测这些数据，看结果是否符合我们的select策略。

Case Study：

在SemEval数据中，XXX和XXX具有关系R。我们的策略发现的数据里有

XXX…XXX，语法结构与SEMEVAL数据不同，但是这两个entity具有这种关系。

<https://arxiv.org/pdf/1511.06390.pdf>

UNSUPERVISED AND SEMI-SUPERVISED LEARNING WITH CATEGORICAL GENERATIVE ADVERSARIAL NETWORKS。Jost Tobias Springenberg University of Freiburg 79110 Freiburg, Germany.

GAN用于多分类

<https://arxiv.org/pdf/1605.07725.pdf>

GAN用于文本分类

事件抽取相关：

<https://arxiv.org/pdf/1909.03546.pdf>事件抽取 EMNLP 2019

David Wadden† Ulme Wennberg† Yi Luan‡ Hannaneh Hajishirzi†∗. Entity, Relation, and Event Extraction with Contextualized Span Representations.

<https://arxiv.org/pdf/1909.05360.pdf> EMNLP 2019

Rujun Han,1,2 Qiang Ning,3 Nanyun Peng1,2.Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction

<http://nlp.csai.tsinghua.edu.cn/~lzy/publications/naacl2019_adv4ned.pdf> acl2019

Xiaozhi Wang1∗ , Xu Han1∗ , Zhiyuan Liu1† , Maosong Sun1 , Peng Li2.Adversarial Training for Weakly Supervised Event Detection.

总结：现有的ED主要基于特征工程，token-level的，structured特征。神经网络将文本信息编码成向量，基于这些向量进行ED。基于人工标注的数据，作有监督训练。

如果一个词，作为已知事件的触发词，那么所有包含这个词的句子都表示这个事件。

想法：如果一个句子包含这个词且包含这个词相关的两个实体，再认定该句子表示这个事件。

框架：encoder：将一个句子{n1,n2…t,..n}包含n个词和触发词candidate 编码成向量形式。

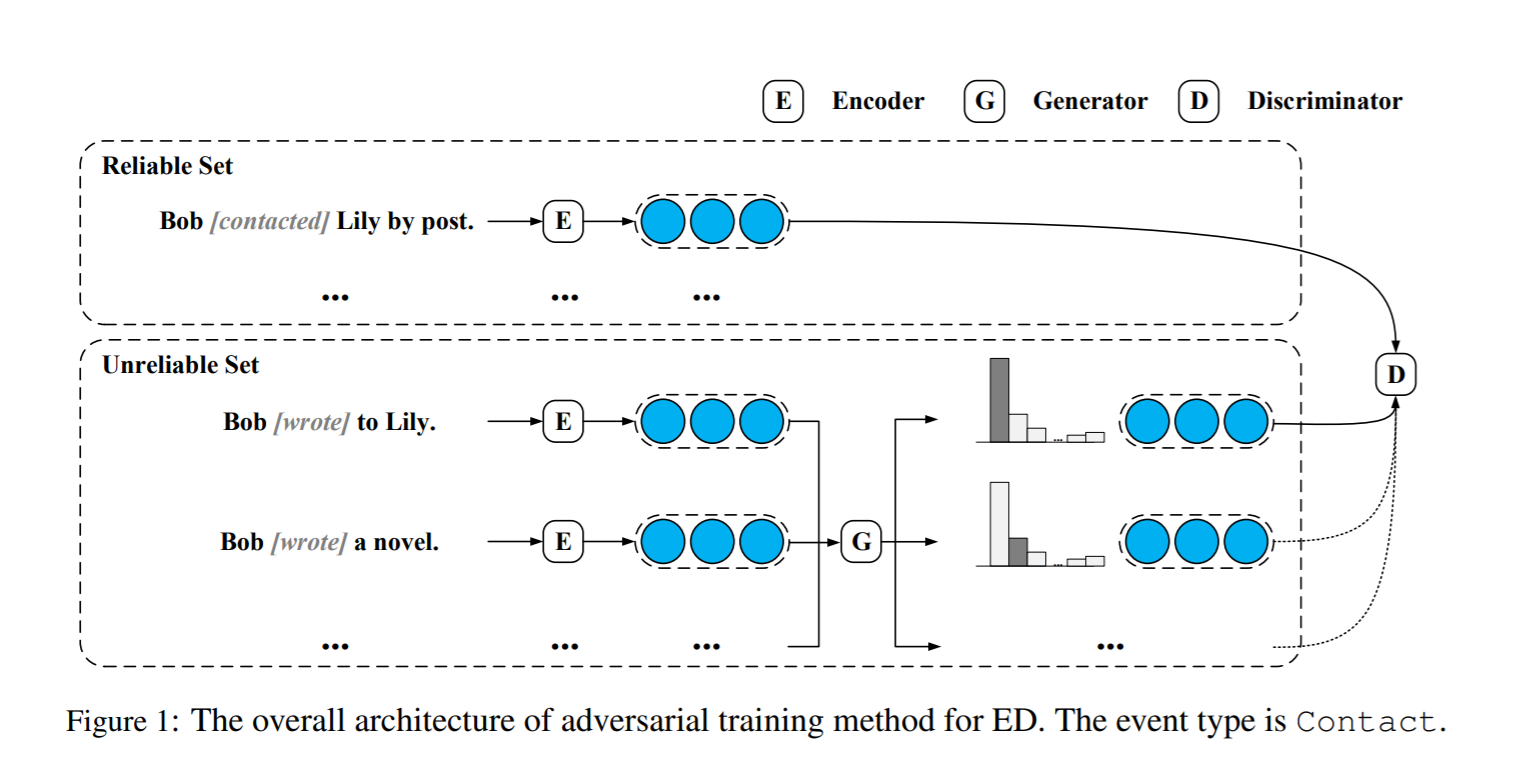
training strategy：将句子分为可信R 和 不可信U的集合。

label：为一个trigger t和event type e。没有trigger则标记NA。

生成器和判别器：输入句子vector，输出类型22

生成器：从不可信集合U里选择句子用于输入到判别器里，输入句子vector，输出0或者1.表示是否输入到判别器里。

想法：生成器引入注意力机制，进行选择，提升降噪性。



对抗训练过程：

U集和R集里的每一个instance都会标有label，实体对e1,e2和关系r。

属于R集的实例，都有确定的关系r，和实体对。

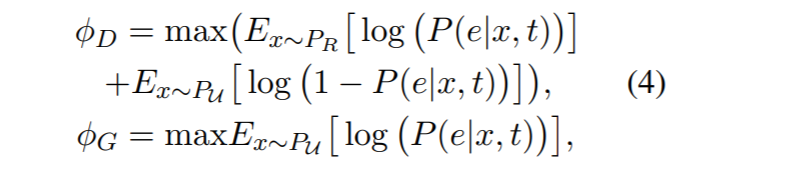
属于U集的实例，其关系r不可信，即有一个确定的概率，是其被标记错误的概率。

因此分类器：

最大化：P(r|x,e1,e2) x∈R和1-P(r|x,e1,e2) x∈U。即对真实数据预测其关系概率为1，对不可靠数据预测其关系为0。

生成器从U集里选择数据来迷惑分类器，依据概率P(r|x,e1,e2),x∈U：

所以整个对抗训练过程如下公式所示：



PR是R集数据分布，PU是U集数据分布，

分类器：给定实例i，和实体对e1,e2，关系r，预测其关系，看是否符合r。

训练好的分类器应当对R集合的数据

相关工作之：