Word Normalization

Regilar expressions

- 普通字符
- +(匹配前方表达式一次及以上) *(匹配0次及以上) ?(匹配0或1次)
- \w:任何字符(大小写&数字)***[A-Z]:大小写***.:"."***$:"$"***\d:[0-9]

Subword tokenization

- BPE
  - 起始为所有字母&终止符
  - 每次合并最高频的相邻子词,语料库合并&词表加入该子词
  - segment算法:按照词表中子词合并顺序依次合并语料库

Word Normalization

- Case folding:全部转化为小写
- Lemmatization:将所有词转化成词根
  - Morphological Parsing
- Stemming:暴力切断

Sentence Segmentation

tokenize first,done by rules

Text Representation

Word Representation

- Co-occurrence matrices(sparse vector)
  - 行:该词的向量表示；列:共现词
  - PMI矩阵:防止it a 等词的影响(PMI越大,相关性越大)$\log_2 \frac{xy}{x \cdot y}$
  - PPMI矩阵:PMI最小值为0

    - 防止bias(极低频词)情况:smoothing
        - co-occurence matrix频数全部加上k
- Dense vector
    - Word2Vec(skip-grams:根据中心词预测上下文词):
        - skip-grams:给定中心词,预测上下文词概率($p(w_{t+j}|w_t)$)(softmax,得分为词向量点积)
        - 中心词嵌入:v_w,上下文词嵌入u_w
        - 优化:$p(co-occur|c,o)=sigmoid(u_o\dot v_c)$,负采样调整似然函数
    - Word2Vec(CBOW:给定上下文词预测中心词)(词袋模型)
- Evaluation
    - Intrinsic(more fast) & Extrinsic

Document Representation

- Co-occurence matrices
    - 行:词表；列:篇章；
    - TF-IDF加权:避免高频词减弱区分度
        - 词频重算tf(\log_{10}(count(t,d)+1))(对于每个文档中的每个词)
        - 文档频率df:一个词在多少篇文档中出现(对于每个词)
        - idf:$\log_{10}\frac{N}{df_i}$(衡量词的出现和文档的相关性)
        - $w_{t,d}=tf_{t,d}*idf_t$:即:重算后的词频与词&文档相关性相乘的结果作为词&文档矩阵。有效压低了过高频词的影响(tf取对数、idf削弱词的权重)
- Dense-vector:SVD(LSA),neural methods

Text Classification

Rule-based methods

- char-level or word-level regular expressions
  ;

Machine learning

- Generative-Classifiers:建模每个类别的特征作为先验分布,结合似然得到后验分布
    - Naive Bayes
        - 假设:每个class中的词独立生成,与位置也无关(词袋模型)
        - learning:MLE,存在闭式解

        - Smoothing:计算$P(w_i|c_j)$时,分子+1,分母+V
        - 对于test data中出现,但词表中没有的词:直接ignore
- Discriminatice-Classifiers:直接判别条件似然
    - logistic-Regression(对文档的特征向量二/多分类)
        - 正则化最小交叉熵,梯度学习

Evaluation(衡量分类器分类结果)

Text Clustering

Mixture of Gaussian(MoG)

- model
- unsupervised learning:EM
    - E:已知参数,更新数据:$P(y_i = k|x_i, \theta^t)$:在参数θ^t下,以x_i为条件,标签$y_i = k$的概率,数值取决于高斯分布表达式(i会有多个取值,相应地代表i号标签对应的高斯分布不同)
    - M:已知数据,更新参数:求 μ, Σ, π,存在闭式解
- purity=$\frac{1}{N}\sum_m max_d(m,d)$,inverse purity=$\frac{1}{N}\sum_d max_m(m,d)$,m为聚类,d为golden种类

Language modeling

Overall

- Goal:预测句子出现的概率
- unknown words(训练集中没有但任务中遇到):
- Evaluation:
    - Extrinsic(下游任务)
    - Intrinsic:Perplexity
        - $l = -\frac{1}{M}\sum_{i=1}^{m}\log_2 p(x_i)$(概率为一句中平均词概率)
        - 限定条件:词表一致
        - 特殊取值:1 V inf

N-gram:以前n-1次预测下一词的概率

- idea:链式法则的简化替换(条件简化)
- Estimating
  - w_i条件概率计算(和也算token)
    - Problem
  - Method1:Smoothing
    - 每个N-grams频数+lambda,重算概率
  - Method2:Backoff and Interpolation
    - Backoff:N规模缩小回退
    - Interpolation:加权插值

RNN($O(n)$):以先前所有词的编码向量预测下一词的概率

- Background:Fixed window NN
  - one hot $\rightarrow$ word embeddings $\rightarrow$ hidden layer vec $\rightarrow$ output distribution(after softmax)
- RNN idea:每个时间步读取一个输入,与之前时间步的综合信息vec(hidden states)产生当前时间步的hidden states,以此预测当前时间步的输出词概率(无稀疏性问题、无需储存n-grams)
- Training:每个时间步有损失函数$J^{(t)}(\theta)$,SGD更新参数
  - Problems:Vanishing(距离过远导致不更新:无法判断梯度消失还是达到最优点,破坏长句依赖,很难解决) or Exploding(可通过Gradient clipping解决) gradient
- LSTM(forget/input/output gate):改善梯度问题
- GRU(update/reset gate):改善梯度问题
- Multi-layer RNN
- Bidirectional RNN
  - Tips:正序RNN和倒序RNN为独立关系,无直接传递关系,二者hidden states直接拼接成总hidden states
  - 不可用于LM:因为建模的是整个序列,而非从左到右逐个预测

Attention

- idea:A(q,K,V)/A(Q,K,V)(self-attn:q k v都由同一句子的hidden states提供)
  - Causal Masking:LM中不能提前看到未来的token:相应位置的注意力点积得分置为-inf
  - Scaled Dot-Product Attn(注意力得分缩放sqrt(dk)倍)
- Multi-head Attn:
  - idea:QKV线性映射至m种低维空间$\rightarrow$形成m-head Attn,每个head遵循原先的A(Q,K,V)计算方法($Softmax(\frac{QK^T}{\sqrt(d_k)})V$),最终多头线性拼接回到高维

- Range of Attn:每时间步复杂度为$O(n)$(RNN为常数)

Transformer:基于先前所有词的Attn预测下一词的概率

- Embedding
  - Method1:与文本无关的固定编码(2d维度的正弦编码)
  - Method2:可学习向量
  - Absolute Embeddings:根据绝对位置编码:不同位置同义词可能判定为不同&训练句过长可能难以泛化
  - Relative Embeddings:解决泛化等问题
    - Method1:同一相对位置偏移量对应同一编码向量,加在key/value向量
    - Method2:注意力随距离变长而递减
    - Method3:RoPE:qk点积结果中间插入矩阵Θ_{n-m}
    - No embedding NoPE:Transformer LM中只提供Masking(其他情境表现不好)
- FNN:修复Mul-head Attn的纯线性问题(对于Attn输出向量,先升维激发(Non-linear activation:ReLU)再降维)
- Tricks:
  - Residual connections
  - layer normalization
- 复杂度(假设输入序列长度为T)
  - QK^T计算:
  - Softmax计算:T^2
  - 右乘V计算:T^2d_k
  - 总复杂度:T^2d_k
  - 简化方案:Sparse attn/Linear attn

Seq to Seq

- idea:RNN
  - encoder & decoder 共享hidden states
- idea:Attn
  - q为上步输出、k v为每步hidden states
- idea:Transformer
  - Encoder:同LM Transformer完全一样(无Mask):将所有输入tokens编码成一个信息向量
  - Decoder:第一层为普通masked-mul-head attn,第二层增加mul-head cross attn(q来自第一层attn输出,kv来自encoder最顶层输出)

- Learning
  - MLE:似然函数(链式法则)
  - 优化:optimized as a single system,backpropagation operates end to end
  - 训练方法:teacher forcing
    - Problem:exposure bias:never seen own errors
    - Sol:Scheduled sampling
- Decoding
  - Greedy:每次给出最有可能的token
  - Beam search
  - Stopping criterion
  - Problem:Diversity→每步进行概率取样,只取top-k或top-p概率词(nucleus sampling)
  - Problem:repetition
  - Non-autoregressive

Pre-trained LMs

- Encoder-only PLMs(ELMo BERT)
  - overview:随上下文文本进行词嵌入
  - ELMo
    - ideas
      - 一个2层正向LSTM,一个2层反向LSTM,二者分开
      - CNN计算初始词嵌入
      - 残差连接(输入层和第二层)
      - 2个LSTM共享词嵌入&Softmax层参数(底层输入&顶层处理,其他不共享)
    - 输入词最终表示:每层的该时间步hidden state加权求和
    - Task
      - 句子输入ELMo,生成每个词的嵌入,用于end-task
      - 对于end-task model:直接使用(不再更新ELMo参数),或保持较小学习率进行finetune
  - BERT
    - idea:使用MLM进行上下文兼顾(普通LM只能正向或反向),LSTM换位Transformer,使用BPE子词分割
    - Utilizing in downstream tasks
      - Finetuning
      - Prompting:不改变参数,仅输入提示文本作为task
      - Prompt tuning:调整一部分参数而非全局参数(Tunable soft prompt)

        - NSP,CLS表示句子信息
- Decoder-only PLMS(GPT)
    - overview:单向(Transformer with attn mask)
    - Utilizing
        - Finetuning:最后一个token包含完整输入→最后一个token连接下游任务进行微调
        - Prompting/in-context learning/chain-of-thought prompting
- Encoder-Decoder PLMs(BART T5 GLM)
    - BART:噪声文本预训练,解码器解除噪声
    - T5:text2text,预训练数据有/无监督
    - GLM:decoder-only,autoregressive blank infilling

LLMs

LLM training

- Pretraining
    - Scaling law
        - 参数、训练数据、训练迭代次数与表现正相关(loss为y轴,与x轴对数成线性关系)
        - Abillity of Emergence(涌现能力):对于大模型而言会出现任务性能显著提高
- Instruction Finetuning
- Parameter-Efficient Finetuning(PEFT):调整一部分模型参数:Prompt tuning,prefix tuning(KV加参数),Adaptor,LoRA(3个LoRA)
- Reinforcement Learning with Human Feedback(RLHF)
    - idea:已进行一部分指令微调的模型:人工为相应进行评分,不模仿人类响应,而是同人类偏好保持一致
    - Problem1:成本
    - Problem2:量化不准确
    - 正则化优化
    - Direct Preference Optimization(DPO) (不再是强化学习,而是一种监督学习:找到了E[R]闭式解)
    - for chain of thought
- Parallel Decoding
    - Jacobi Decoding(算法(并行autoregressive decoding on all tokens,迭代若干次)、steps数量上限)
    - Speculative Decoding(算法:快速drafting,LLM并行自回归验证,反复迭代)(增加命中率:树形结构drafting)

- KV Cache
  - Head:MQA GQA MLA(矩阵压缩，再矩阵还原)
  - Layer:LCKV YOCO CLA
  - Token:pruning merging
- Retrival-Augmented Generation(RAG)
  - idea:数据库存储信息 IG信息检索(Retriver),检索到的信息作为LLM query信息的一部分(Generator)

Seq Labeling

Overview

BIES,BIESO

HMM

- Idea(数学表达式:转移*发射)
- Decoding/Inference
  - 期望结果
  - 算法:Viterbi(状态定义,初始化,转移方程,终止状态)
    - 复杂度分析
  - Marginal Inference
    - 算法:Forward-Algorithm(Viterbi的max改成sum)(状态定义、初始化、转移方程、终止状态,复杂度分析)
- Learning
  - Supervised
    - 最大似然估计(似然函数、闭式解、稀疏性处理)
  - Unsupervised(apply:POS)
    - EM:Baum-Welch Algorithm(maximize p(sentense))
      - E step(给定参数算分布)
        - expected count(求解label&transition&emission、Forward-Backward算法)
      - M step(给定分布迭代参数)
        - 似然函数、闭式解(normalizing E-counts)
    - 除EM外算法:梯度下降
      - 似然函数为p(sentenses),forward计算,再backprop
      - 与Forward-Backward非常相似

HMM to CRF

- 2 Problems for HMM
- MEMM
  - Background:generative or discriminative model的似然函数
  - 似然函数:每步似然(s函数softmax)相乘
  - 问题:label bias(用weights)

CRF

- 似然函数:exp每步似然和,softmax
- Inference/Decoding
  - 期望结果
  - 算法:Viterbi
- Supervised Learning
  - 对数似然函数
    - 归一函数Z的求解:Forward-Algorithm
  - 学习方法1:梯度下降
    - 梯度式涉及期望计数:Forward-Backward/直接通过auto-differentiation
  - 学习方法2:SSVM
    - 似然函数
    - Advantages
      - 考虑了Delta(标签预测损失函数)
      - 注重decision boundary而非整体分布
    - 似然函数可能通过Viterbi快速求解
    - 优化方法
- Unsupervised learning
  - Encoder & Decoder
  - 似然函数及计算(Forward-Algorithm)
  - 优化(梯度下降)

Neural

- RNN problems & Bidirectional RNN
  - 每个时间步接收上一步的总和信息vec&当前时间步的word vec,输出当前时间步的label
- Transformer
- Neural CRF
  - idea:神经方法计算CRF potentials:emission得分
- Inference

    - 不用CRF:neural softmax:逐位置独立预测label
    - 用CRF:Viterbi解码
- Learning
  似然估计 or 边际损失(similar to CRF learning)

Constituency Parsing

Overview

- idea
    - constituency parsing tree、scoring(score each part of the tree,取分数之和)
- Parsing
    - Generative and discriminative、各自的打分解码方法、对应的**-based 算法
    - S得分函数计算方法
        - Assume1:独立、DP、全局最优
        - Assume2:非独立、贪心、局部最优
- Learning
    - Supervised:tree bank学习
    - Unsupervised:tree bank评估
    - 评估:各点元组化、计算precision&recall&F1
    - 多句评估:Macro/Micro averange F1

Span-based

- idea
    - 二叉树、每点状态表示
    - 打分方法(特别是根节点)
        - Discriminative Parsing Method:feature of span,或词嵌入&Biaffine评分(对于i,j,l)
- Parsing
    - 期望结果
    - 算法:CYK(Bottom up DP)(求和转移,s(i,j)取maxl)
- Supervised learning(MLE)
    - 似然函数:$\sum(i,j,l)$做softmax
    - Z(x)计算:Inside Algorithm(s'(i,j)取suml,相乘转移)
    - 优化:SGD
    - Alternative:margin-based loss

Context-Free

- idea
  - terminal nonterminal S rules
  - 转移语法打分:PCFG/SCFG & WCFG
- Parsing
  - Bottom-up DP:CYK(CNF)(状态定义及跨度表示、Base case、状态转移(总概率计算)(用于Probabilistic CYK 消除歧义算法))
    - CYK(span-based) vs CYK(PCFG)(max取值点不同)
- Learning
  - Supervised:Generative Methods
    - 似然函数
    - 闭式解及MLE效果、原因
  - Supervised:Discriminative Methods(for WCFG)
    - 似然函数:$\Pi W(r|x)$做softmax
      - Z(x)求解:inside algorithm
    - 优化:SGD
    - Alternative:margin-based loss
    - Inside Algorithm:状态表示、base case、状态转移(use sum)
      - Forward Algorithm:inside的特殊情况
      - Viterbi Algorithm:CYK的特殊情况
  - Unsupervised
    - Structure search
    - Parameter Learning
      - MLE:P(sentense)(marginalized)
      - E step:计算解析树分布(expected counts inside-outside algorithm)
      - M step:根据解析数分布更新参数(闭式解:expected counts归一)
      - 直接梯度下降(inside algorithm)

Transition-based

- 基本操作
- Parsing:Greedy,Beam-search
- learning:训练分类器
  - 步骤
  - potential flaw(只见过正确):Dynamic oracle
- 操作次数:3L-1

Dependency Parsing

Overview

- idea
  - arc:from head to dependent
  - adv、disadv
  - dependency & constituency转化
- parsing
  - 期望目标
  - evaluating:UAS LAS
  - multiple sentences evaluating:macro micro averange

Graph-based

- idea
  - Scoring:每边一个score(由两词特征决定),总分为sum(head指向depend)
- Parsing
  - 独立选所有正边(？)
  - head-selection(？)
  - MST(？)
  - CYK(？)
    - idea:转换形式类似CNF(每2词一个arc)→CYK解析
    - $O(n^3|G|) = O(n^5)$
  - Eisner(4中图示、4种合并转移、状态表示、状态转移、应用前提、时间复杂度、最终状态表达式)
  - non-projective:MST($O(n^3)$)
- second-order
- Learning
  - Supervised learning
    - 似然函数:s(t)做softmax
    - Z函数计算:
      - projective:sum Einsner algorithm:similar to inside algorithm
      - non-projective:Kirchhoff
    - 似然函数分解:head-selection
    - 优化:Gradient-based
  - Unsupervised learning

        - Generative:类似PCFG的MLE:EM SGD P(sentence)
        - Discriminative:CRF-autoencoder(解码为:每个词预测其head),SGD优化

Transition-based

- 基本操作

Semantics

Lexical Semantics

- Semantic relations
    - synonymy(同义),Antonymy(反义),Hyponymy(前者is a后者),Hypernymy(前者contains后者),Meronymy(A is part of),Holonymy(A has a)
- Wordnet
    - 每节点为一个synset,边为relations
    - semantic distance:两个synsets间通过hypernymy/hyponymy的最短路
- WSD:seq labeling

Formal meaning representation(形式化表示语义)

- First order logic(FOL)
- Semantic Graphs
    - 0:single word&relation(DM PSD)
    - 1:part of sentense&relation(EDS)
    - 2:unanchored&relation
- Parsing:(句子→形式化语义)
    - syntax-driven of neural approach:同步CFG(SCFG)
        - 非终结符→(自然词汇&非终结符序列)/(形式化语义序列),两树节点对应(自然语言构建左树,替换右树,形式化表示)
    - Neural-parsing
        - seq to seq
        - parsing to semantic graph(基于转移or基于图)
- learning
    - 监督:成本高
    - weak supervised(知道语义表示的运行结果,无需标注语义图)

Semantics Role Labeling

- PropBank(roles较少,general)
- FrameNet(较多,specific)
  - 多个frame,定义了谓词集合和角色
- 标注(输出span和role of span)
  - seq labeling:一次标一个谓词
  - graph-based:arcs表示span role
  - seq to seq

Information Extraction

- 命名实体&嵌套命名实体()识别
- 实体链接
- relation extraction(关系提取)
- 提取事件及相关信息
- NER Methods:
  - seq labeling(BIO),span classification
- relation extraction methods:
  - dependency prediction
- seq to seq

Discourse Analysis

Overview

- coherent
  - Lexical Chains
  - Cerefenrece Chains(共同指代)
  - Discourse Markers(逻辑关系标记词)

Discourse

- 连贯性关系:RST:建模coherence relations(nucleus,satellite),用图解分析(satellite指向nucleus)
- 篇章结构:RST关系形成的树结构(节点为elementary discourse unitsEDU(可用seq labeling解析出来),边为RST关系边(可用成分解析或者依存解析)

- parsing

Coreference

- Step1:找出mention
  - POS tagger(词性标注)、成分分析、命名实体分析
  - 特殊情况处理:rules,Binary classification,inference with mention clustering
- Step2:clustering
  - idea1(mention clustering):训练二分类器,每一个mention对于所有其他的mentions识别是否位于同一聚类当中⇒距离很远则预测困难
  - idea2(mention ranking):每个mention只找出匹配概率最高的mention(包括NA)
    - Training:通过语义学特征或神经方法
    - Inference:transitive closure
    - Evaluation:purity & inverse purity