Regular expressions

+(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]

word normalization:Lemmatization：将所有词转化成词根(Morphological Parsing)

Sentence Segmentation:tokenize first,done by rules

Word Representation

Co-occurrence matrices(sparse vector)

行：该词的向量表示,列：共现词; **PMI矩阵**(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

Dense vector(word2vec)

skip-grams(c预测o):$p(w_{t+j}|w_t)$=点积softmax; **c嵌入**：v_w，o嵌入u_w; **优化**：$p(cooccur|c,o) = sigmoid(u_o \cdot v_c)$，负采样调整似然; **CBOW**:bag of o预测c

Evaluation:Intrinsic(more fast) & Extrinsic

Document Representation

Co-occurence matrices(行:词表,列:篇章)

tf$= \log_{10}(count(t,d)+1)$(对于每个文档中的每个词)**文档频数**df **idf**=$\log_{10} \frac{N}{df_i}$ （衡量词的出现和文档的相关性） $w_{t,d} = tf_{t,d} * idf_t$

Dense-vector：SVD(LSA)，neural methods

Text Classification Rule-based:char-level or word-level regular expressions

Text Classification ML

Naive Bayes:词袋模型

learning:MLE,闭式解; **Smoothing**：计算$P(w_i|c_j)$时,分子+1,分母+V; **ignoreUKW**

logistic-Regression:文档的特征向量二/多分类:正则化最小交叉熵，梯度学习

Mixture of Gaussian(MoG)

unsupervised learning:EM:**E**:$P(y_i = k|x_i, \theta^t)$;**M**: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

Intrinsic:Perplexity:$l = -\frac{1}{M}\sum_{i=1}^{m}\log_2 p(x_i)$

N-gram

Smoothing:每个N-grams频数+lambda，重算概率

Method2:Backoff and Interpolation(回退|插值)

RNN($O(n)$) 无稀疏性问题、不储存n-grams

 # Training:对$J(\theta)$SGD; **Problems**：Exploding(clipping)

LSTM(forget/input/output gate)&GRU(update/reset gate)

Multi-layer RNN
Bidirectional RNN(正反序独立再拼接,整序列建模)

Attention:每时间步都为$O(n)$
Causal Masking(-inf),**Scaled**($\sqrt{d_k}$),**Multi-head**:m种低维空间,m个A(Q,K,V)=$Softmax(\frac{QK^T}{\sqrt{d_k}})V$,多头线性拼接

Transformer:基于先前所有词的Attn预测下一词的概率
PE:文本无关的2d正弦编码or可学习向量; Relative Embeddings:解决泛化等问题(偏移对应编码,加在k,v; 注意力递减; RoPE矩阵Θ_{n-m}; NoPE); layer normalization(避免shifting，Attn FFN)
复杂度T^2d_k,Sparse/Linear attn

ELMo:2层正反向LSTM,CNN初始嵌入,残差连接,共享嵌入&softmax,每层每时间步hidden加权求和; end-task与finetune
BERT:BPE,Transformer,MLM; finetuning,prompting,prompt tuning(soft),NSP&CLS表句子信息
GPT:单向,attn mask; finetuning(最后token连接下游),prompting/incontext learning/chain-of-thought prompting
BART:噪声文本预训练,解码器解除噪声
T5:有/无监督数据转为text2text训练
GLM:decoder-only,autoregressive blank infilling

Pretraining:Scaling law(参数,数据,迭代,loss与x轴); Emergence能力
微调 Inst-Finetuning,PEFT{Prompt tuning,prefix tuning(KV加参数),Adaptor,LoRA(3个LoRA)}
RLHF:对齐偏好(RM,比较二分类,正则化); DPO(E[R]闭式解); (for chain of thought)
Parallel:Jacobi Decoding(并行自回归); Speculative Decoding(drafting,LLM并行自回归)(增加命中率:树形drafting)
KV Cache:Head:MQA GQA MLA(矩阵压缩矩阵还原);Layer:LCKV YOCO CLA;Token:pruning merging
RAG:IG信息检索(Retriver),检索到的信息作为LLM query信息的一部分(Generator)

Seq to Seq
Transformer(unmask)(MaskMH+CrossKV)
Learning:链式法则似然MLE,end2end优化,teacher forcing训练,Sol为scheduled sampling
Decoding:greedy|beam(取平均对数似然); **Diversity**$\rightarrow$top-k/p sampling; NAT(预测长度k,并行无码解码)

HMM
Inference:Viterbi,$O(nY^2)$; Marginal Inference:Forward-Algorithm(改sum)
SL:eq积似然MLE,统计闭式解,稀疏性
USL(apply:POS): P(sentence)似然EM; **E**:expected count(求标签转移发射,Forward-Backward); **M**:联

合对数似然,归一闭式解; **还可GD**：forward算p(sentenses)似然再backprop，像Forward-Backward
MEMM:每步($s = s_e + s_q$softmax)相乘; label bias(weights)

CRF
Inference: **每步得分和**softmax似然取对数; **Viterbi**
SL:softmax对数似然MLE,Forward-Algorithm求Z; Forward-Backward(求EC)做GD|SSVM(可能用vitebi,考虑标签损失&偏重boundary,loss不可微)
USL:E-Decoder,forward算loss,GD优化(不可算P(sentence))

Neural
Neural CRF:神经方法计算potentials emission
Inference:**不用CRF**:逐位置独立neural softmax; **用CRF**:Viterbi
Learning:似然估计|边际损失(similar to CRF learning)

Constituency Parsing:得分取和
Span-based
每点打分:Discriminative:feature of span,或词嵌入&Biaffine
Parsing:CYK(Bottom up DP)(求和转移,s(i,j)取maxl)
SL:$\sum(i,j,l)$**softmax**似然MLE,**Z**:Inside Algorithm(s'(i,j)取suml,相乘转移); **SGD**;
Alternative:margin-based loss

Context-Free(P/SCFG,WCFG)
Parsing:Bottom-up DP:CYK(CNF,PCFG)(概率积转移)
 # SL-Gene(PCFG):概率积似然MLE,闭式解; **SL-Dis(WCFG)**:权重积softmax似然,inside algorithm(3参数,两子树*rule概率求和)算Z,SGD,margin-based loss alternative
USL:结构|参数{P(sentence)MLE,**E**算树分布,用inside-outside算ECounts,**M**算参数,ECounts归一闭式解},**或梯度下降**,P(sentence)用inside算

Transition-based
Parsing:Greedy,Beam-search
learning:训练分类器; potential flaw(只见过正确):Dynamic oracle

Graph-based Dependency
CYK:$O(n^3|G|) = O(n^5)$
Proj:Eisner:($O(n^3)$),Non-proj:MST:($O(n^3)$)
SL:**s(t)softmax**似然,**Z**用sum Eisner|Kirchff; **head**-selection,梯度优化
USL:**Gene**:EM|SGD2P(sentence),类PCFG; **Dis**:CRF-autoencoder(每词预测head),SGD

Lexical Semantics
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 synset,relations; distance为上下位最短路
WSD:seq labeling

Formal meaning representation
Semantic Graphs:word(DM PSD)part(EDS)unanchord
Parse2formal:**SCFG**:非终结符→(自然&非终结)/(形式&非终结),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|semantic graph(基于转移|图)
learning:weak supervised(运行结果)

Semantics Role Labeling
PropBank(roles较少,general),FrameNet(frame,谓词集和角色)
标注:seq labeling,graph-based,seq to seq

Information Extraction:NER(span classification),实体链接,关系提取(dependency),事件信息(seq2seq)

coherent:{Lexical Chains,Cerefenrece Chains,Discourse Markers}
连贯性关系:RST:satellite to nucleus
篇章结构:节点为EDU(seq labeling解析),边为RST(成分|依存解析)

Coreference:(1):mention(POS,成分,命名实体分析)(rules,二分类等); (2)clustering:{1.clustering:二分类器,远端困难; 2.ranking:选一个,语义特征或神经法训练,transitice closure解码,(inverse)purity}