Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

文本校对

巩锟 吴泽瀚 杨得山 万韵伟 费泽涛 沈宇辉

2021年11月15日

文本校对

- 定义
  - 利用自然语言处理，对文本中出现的书写、文法错误进行检测并校正。

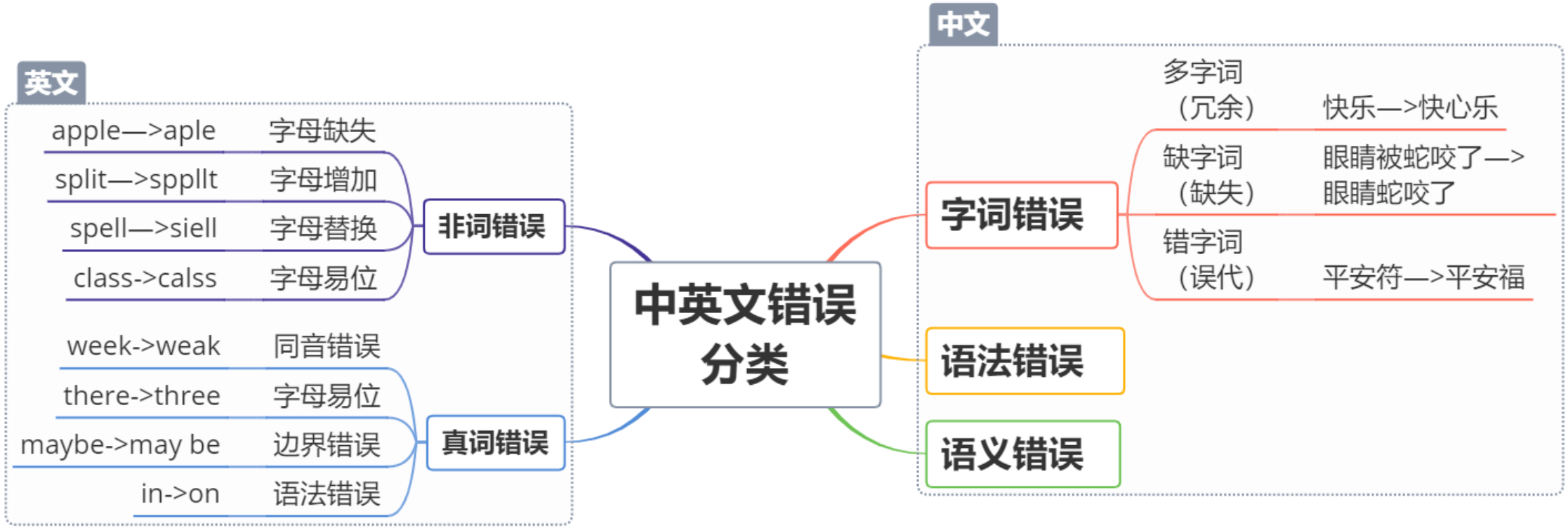
- 主要内容
  - 字词错误纠正
  - 语法错误纠正
  - 文本较对的基本应用与拓展
  - Demo展示

字词错误纠正

中英文错误分类

- 中英文文本形成原因差异：

1）输入方式不同； 2）文本结构不同； 3）词的构成规则不同； 4）字符集规模不同



字词错误纠正（传统方法）

- 中英文字词错误纠正传统方法

方法	优点	缺点
查字典法	检错和纠错一体化，效率高； 对英文非词错误检测非常有效	查准率低； 较对效果差； 很难检测中文文本中的字词错误
词形距离法	节省存储空间； 能反映一定的常见拼写错误统计规律	依赖于大词典
n-gram	包含了前n-1个词所能提供的全部信息	词与词之间不存在语义关联； 参数空间可能非常大； 数据稀疏
最小编辑距离	具有极高抗噪性； 很好描述序列间的差异性；	存在出现语义错误的可能性
基于规则的文本校对方法	准确率高	文本错误规则局限； 召回率提高难度较大

字词错误纠正（SEC）

传统方法

基于概率的文本较对方法(n-gram);

基于规则的文本较对方法

基于机器学习的方法

前馈神经网络语言模型（例：结合词向量和神经网络）；

机器学习分类方法；（结合SVM,RF,DT）

循环神经网络语言模型；（例：CSC-BiLSTM-CRF）；

MLM

BERT；

XLNET；

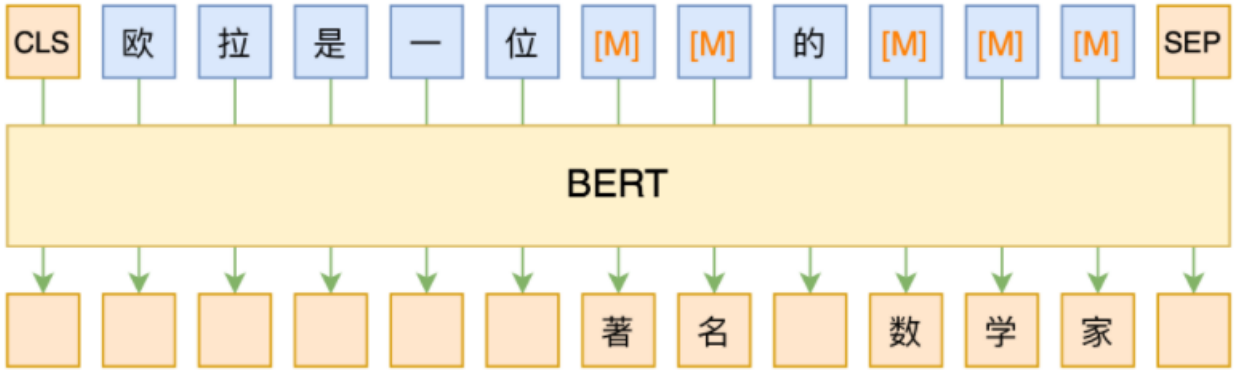
ELECTRA

字词错误纠正（CSC）

- MLM(掩蔽/掩码语言模型)
  - **完形填空**任务
  - 屏蔽给定句子中**特定百分比**的单词
  - 基于前后预测被屏蔽的单词

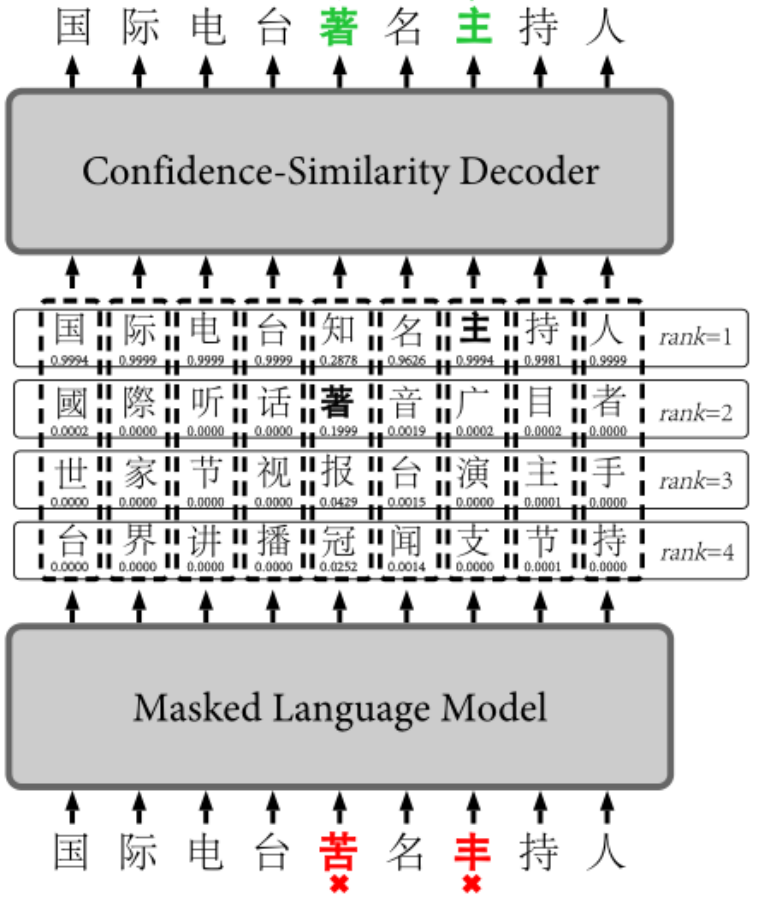
- 数据集
  - SIGHAN
  - OCR数据集
  - News Title等

- 评价指标
  - 准确率accuracy
  - 精度 precision
  - 召回率 recall
  - F1



字词错误纠正（CSC）

- FASpell
  - 解决问题
    - 中文拼写纠错语料不足
    - 混淆集进行纠错不灵活
  - 模型架构
    - 使用MLM作为**去噪自动编码器**生成候选词
    - 使用置信相似度**解码器**过滤候选词
  - 消融实验
    - 目的：取代原先单一概率阈值进行候选字的选择
      - **Confidence**(置信度):预测字的概率值
      - **Similarity**(相似度)：字音和字形
  - 效果
    - 过滤速度快,模型结构更简单
    - 校正精度更高（OCR数据集检测精度达78.5%，校正精度73.4%）



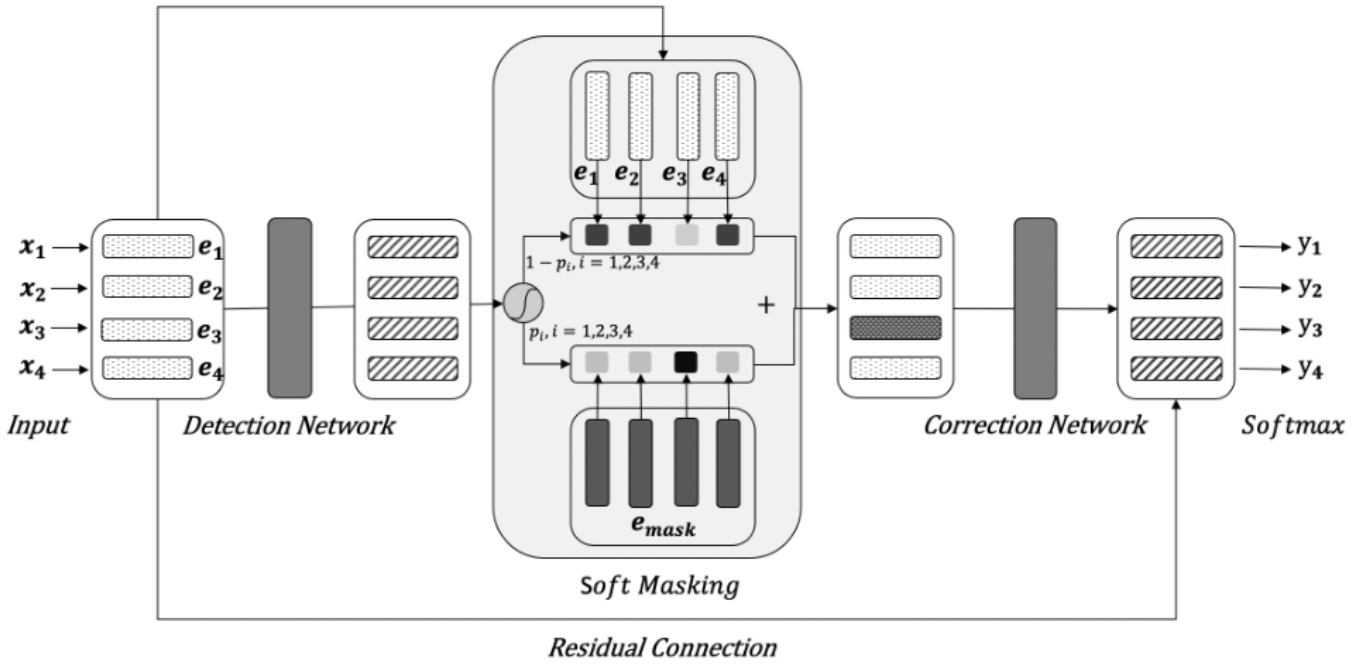
[1] FASPell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm(EMNLP 2019)

字词错误纠正（CSC）

- Soft-Masked BERT
  - 解决问题：模型的学习错误检测能力差
  - **检测**网络：双向GRU
    - 输入：嵌入序列
    - 输出：字符序列(嵌入)的错误率
    - soft-masked embedding
      - 计算输入嵌入和[MASK]嵌入的误差概率加权和

$$e_i^{'} = p_i \cdot e_{mask} + (1 - p_i) \cdot e_i$$

  - **纠错**网络：BERT
    - 输入：掩码嵌入序列
    - 输出：可以被纠正为候选字符的概率



[2] Spelling Error Correction with Soft-Masked BERT (ACL 2020)

字词错误纠正（CSC）

- Soft-Masked BERT
  - 数据集
    - SIGHAN数据集合和News Title数据集
  - 实验效果

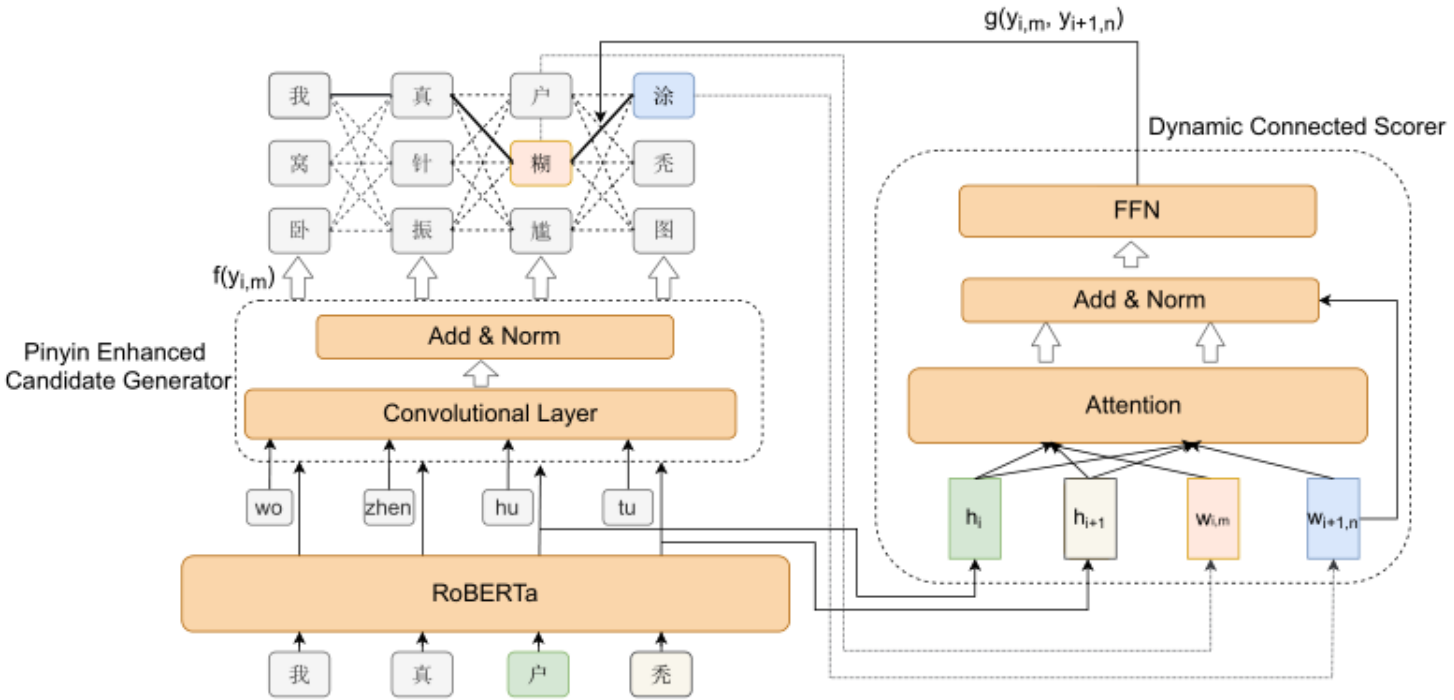
Test Set	Method	Detection				Correction			
		Acc.	Prec.	Rec.	F1.	Acc.	Prec.	Rec.	F1.
SIGHAN	NTOU (2015)	42.2	42.2	41.8	42.0	39.0	38.1	35.2	36.6
	NCTU-NTUT (2015)	60.1	71.7	33.6	45.7	56.4	66.3	26.1	37.5
	HanSpeller++ (2015)	70.1	**80.3**	53.3	64.0	69.2	**79.7**	51.5	62.5
	Hybird (2018b)	-	56.6	69.4	62.3	-	-	-	57.1
	FASPell (2019)	74.2	67.6	60.0	63.5	73.7	66.6	59.1	62.6
	Confusionset (2019)	-	66.8	73.1	69.8	-	71.5	59.5	64.9
	BERT-Pretrain	6.8	3.6	7.0	4.7	5.2	2.0	3.8	2.6
	BERT-Finetune	80.0	73.0	70.8	71.9	76.6	65.9	64.0	64.9
	Soft-Masked BERT	**80.9**	73.7	**73.2**	**73.5**	**77.4**	66.7	**66.2**	**66.4**
News Title	BERT-Pretrain	7.1	1.3	3.6	1.9	0.6	0.6	1.6	0.8
	BERT-Finetune	80.0	65.0	61.5	63.2	76.8	55.3	52.3	53.8
	Soft-Masked BERT	**80.8**	**65.5**	**64.0**	**64.8**	**77.6**	**55.8**	**54.5**	**55.2**

[2] Spelling Error Correction with Soft-Masked BERT (ACL 2020)

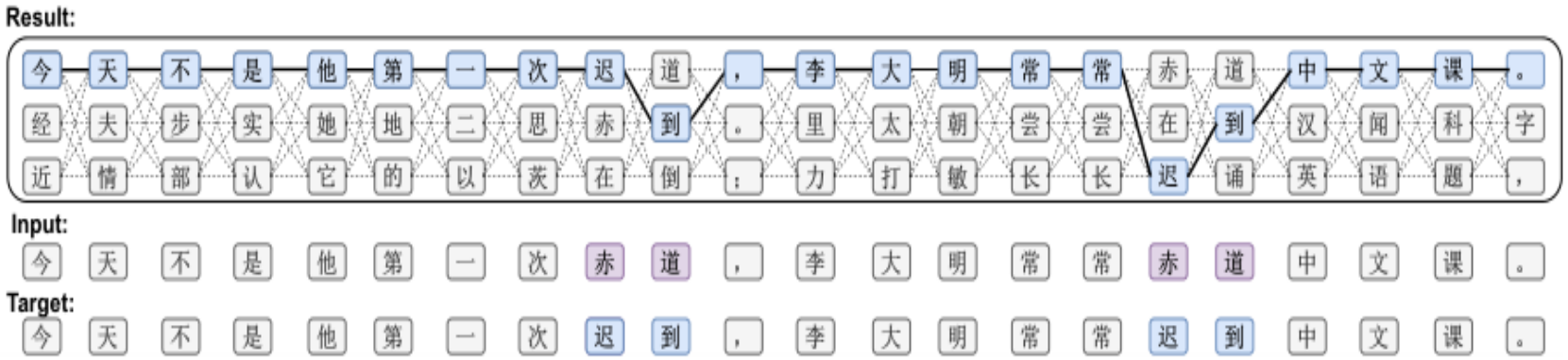
CSC前沿算法

DCN(Dynamic Connected Network)

- 算法原理
  - 解决不连贯
  - 生成候选字符
  - DCScore评分
  - 得分最高路径



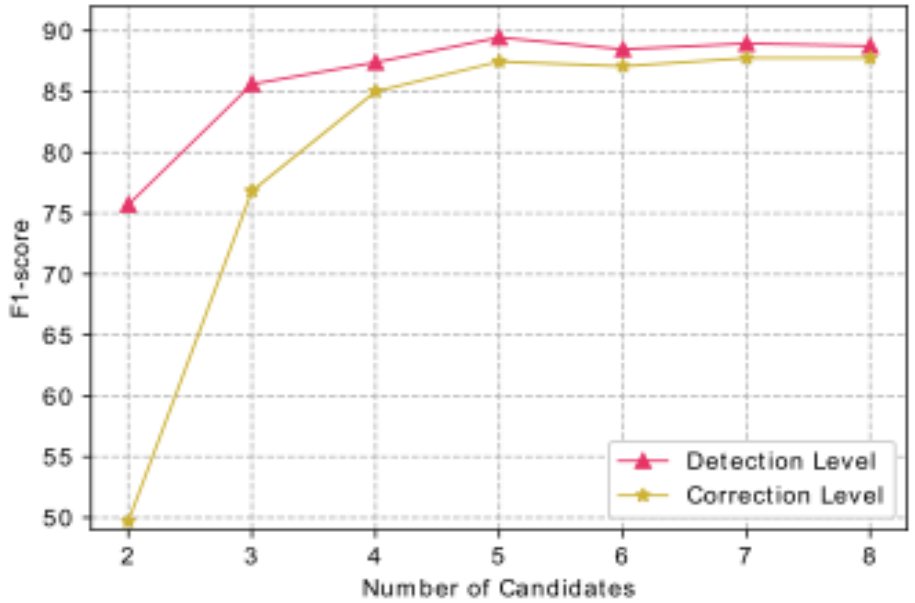
Wrong: 我忘记告诉你了，我真户秃。
Correct: 我忘记告诉你了，我真糊涂。 Translation: I forgot to tell you. I'm so confused.



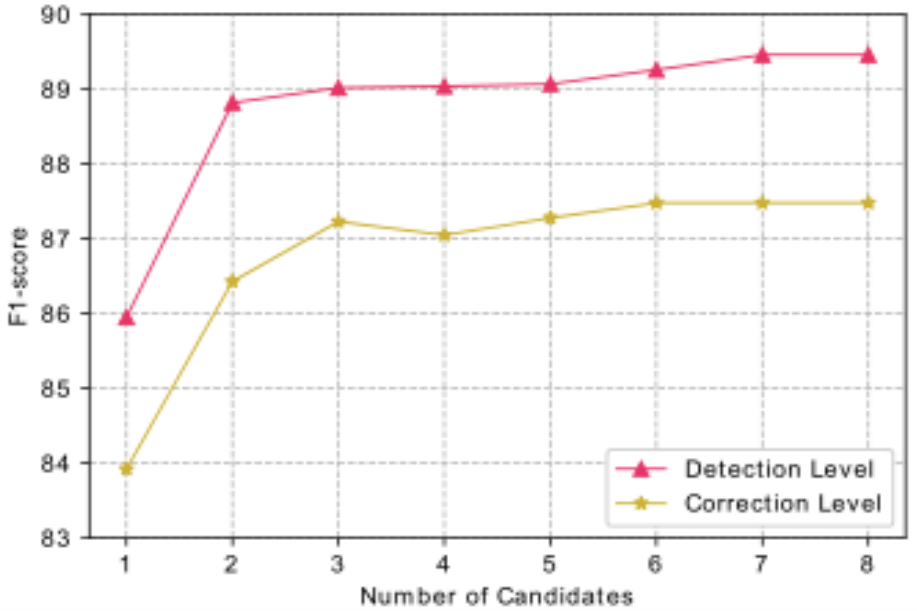
DCN(Dynamic Connected Network)

- 候选字符的有关影响
  - 候选字符生成方法
  - 候选字符数量

Sampling Method	D-F	C-F
Top-k of vocabulary	**89.7**	**88.7**
Multinomial distribution sampling	88.1	87.6
Random sampling	12.2	7.3
Top-k of confusion set	35.8	34.7



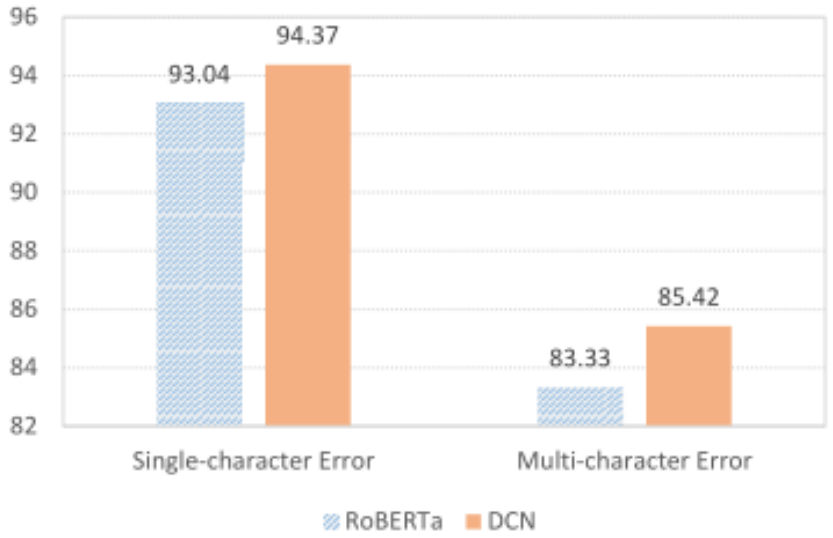
(a) Training stage.



(b) Predicting stage.

DCN(Dynamic Connected Network)

- 实验结果对比
  - FASPell
  - BERT
  - SpellGCN
  - DCN

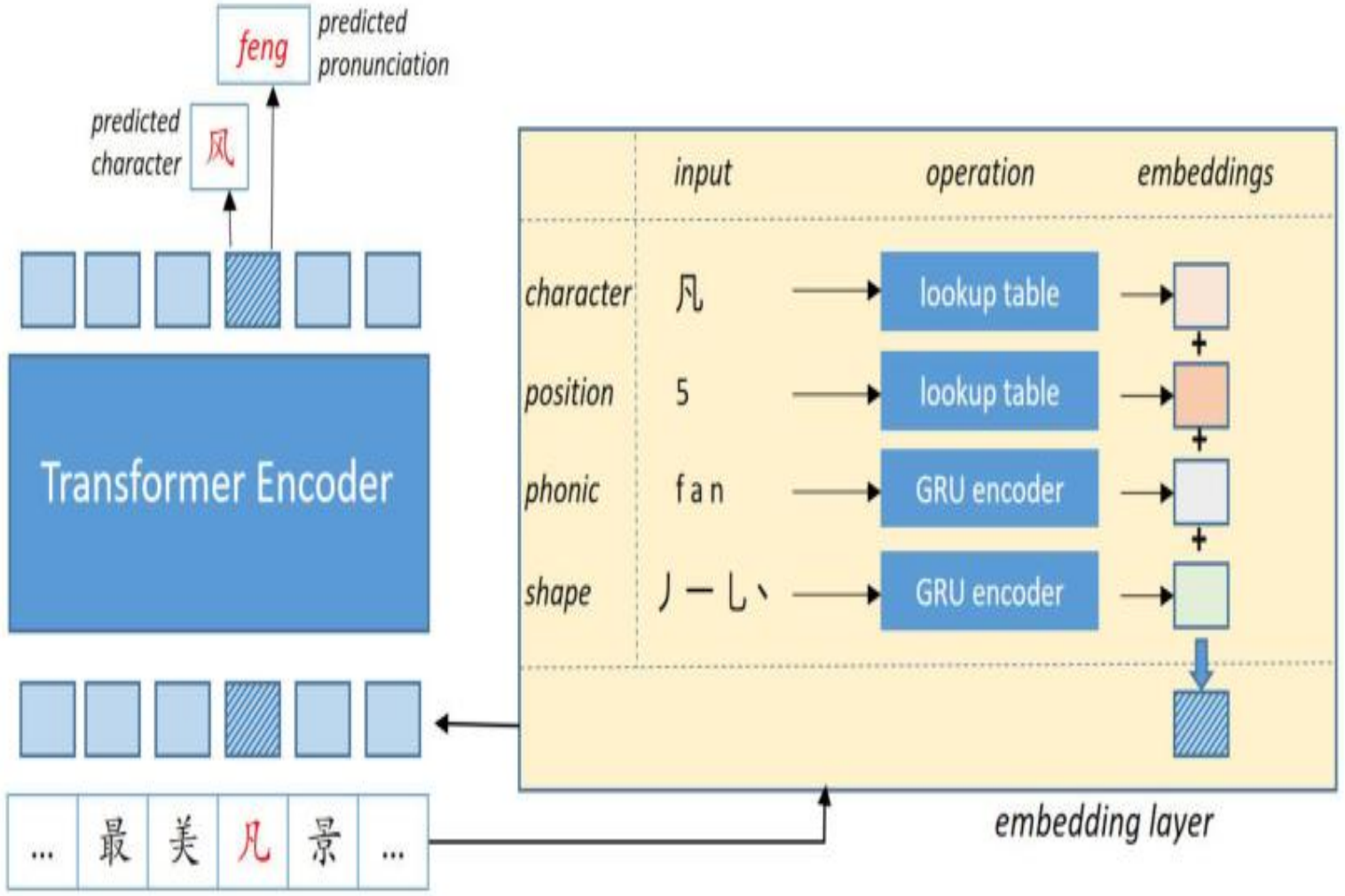


Dataset	Model	Detection-level			Correction-level		
		D-P	D-R	D-F	C-P	C-R	C-F
CSC13	FASPell (Hong et al., 2019)	76.2	63.2	69.1	73.1	60.5	66.2
	BERT (Cheng et al., 2020)	79.0	72.8	75.8	77.7	71.6	74.6
	SpellGCN (Cheng et al., 2020)	80.1	74.4	77.2	78.3	72.7	75.4
	SpellGCN*	85.2	77.7	81.2	83.4	76.1	79.6
	RoBERTa (Ours)	85.4	77.7	81.3	83.9	76.4	79.9
	RoBERTa-DCN (Ours)	86.2	78.4	82.1	84.6	76.9	80.5
	RoBERTa-Pretrain-DCN (Ours)	**86.8**	**79.6**	**83.0**	**84.7**	**77.7**	**81.0**
CSC14	FASPell (Hong et al., 2019)	61.0	53.5	57.0	59.4	52.0	55.4
	BERT (Cheng et al., 2020)	65.6	68.1	66.8	63.1	65.5	64.3
	SpellGCN (Cheng et al., 2020)	65.1	69.5	67.2	63.1	67.2	65.3
	RoBERTa (Ours)	64.2	68.4	66.2	62.7	66.7	64.6
	RoBERTa-DCN (Ours)	**67.6**	68.6	68.0	64.9	65.9	65.4
	RoBERTa-Pretrain-DCN (Ours)	67.4	**70.4**	**68.9**	**65.8**	**68.7**	**67.2**
CSC15	FASPell (Hong et al., 2019)	67.6	60.0	63.5	66.6	59.1	62.6
	Soft-Masked BERT (Zhang et al., 2020)	73.7	73.2	73.5	66.7	66.2	66.4
	BERT (Cheng et al., 2020)	73.7	78.2	75.9	70.9	75.2	73.0
	SpellGCN (Cheng et al., 2020)	74.8	80.7	77.7	72.1	77.7	*75.9*(74.8)
	RoBERTa (Ours)	74.7	77.3	76.0	72.1	74.5	73.3
	RoBERTa-DCN (Ours)	76.6	79.8	78.2	74.2	77.3	75.7
	RoBERTa-Pretrain-DCN (Ours)	**77.1**	**80.9**	**79.0**	**74.5**	**78.2**	**76.3**

PLOME(Pre-training with Misspelled Knowledge)

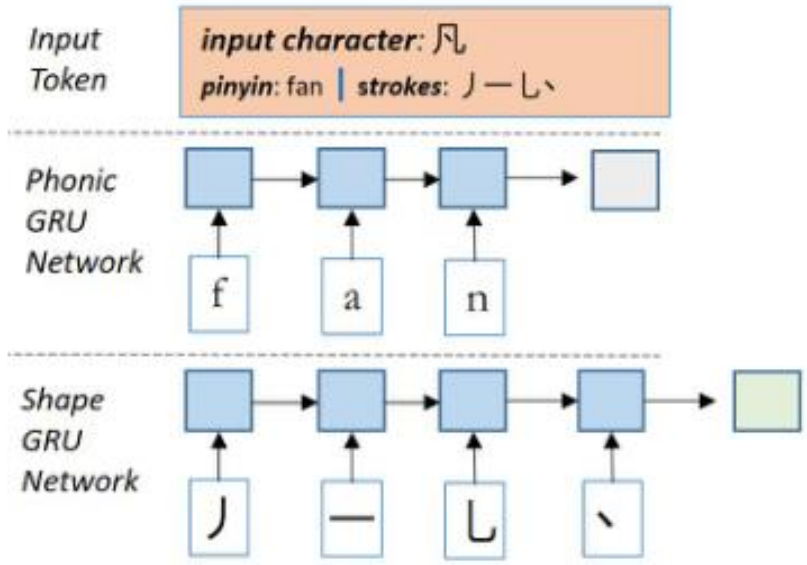
- 算法模型原理
  - 基于混淆集的掩蔽
  - 语音和笔画的嵌入
  - 对字符和语音水平建模

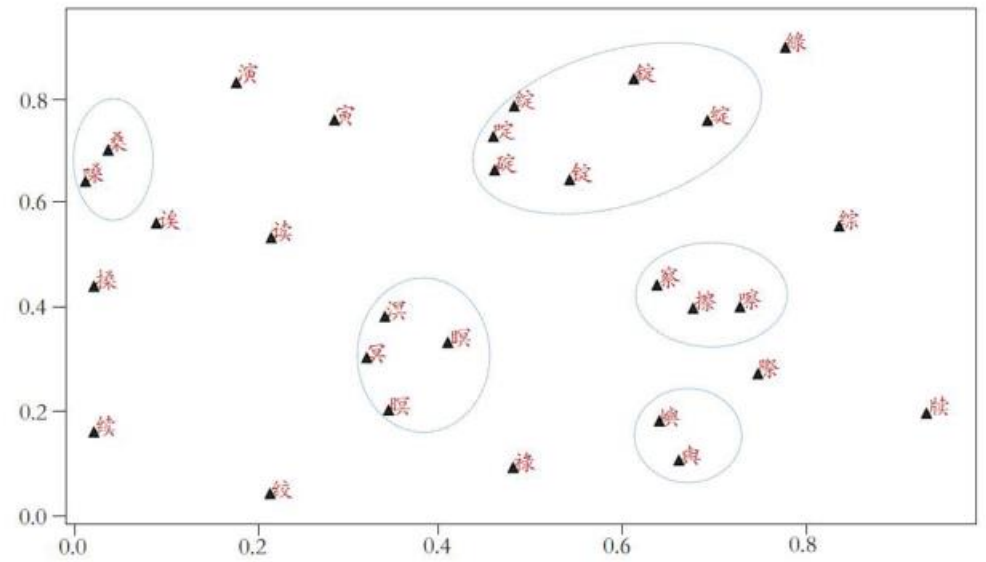
	Sentence
Original Sentence	他想明天去(qu)南京探望奶奶。
BERT Masking	他想明天[MASK]南京看奶奶。
Phonic Masking	他想明天曲(qu)南京看奶奶。
Shape Masking	他想明天丢(diu)南京看奶奶。
Random Masking	他想明天浩(hao)南京看奶奶。
Unchanging	他想明天去(qu)南京看奶奶。

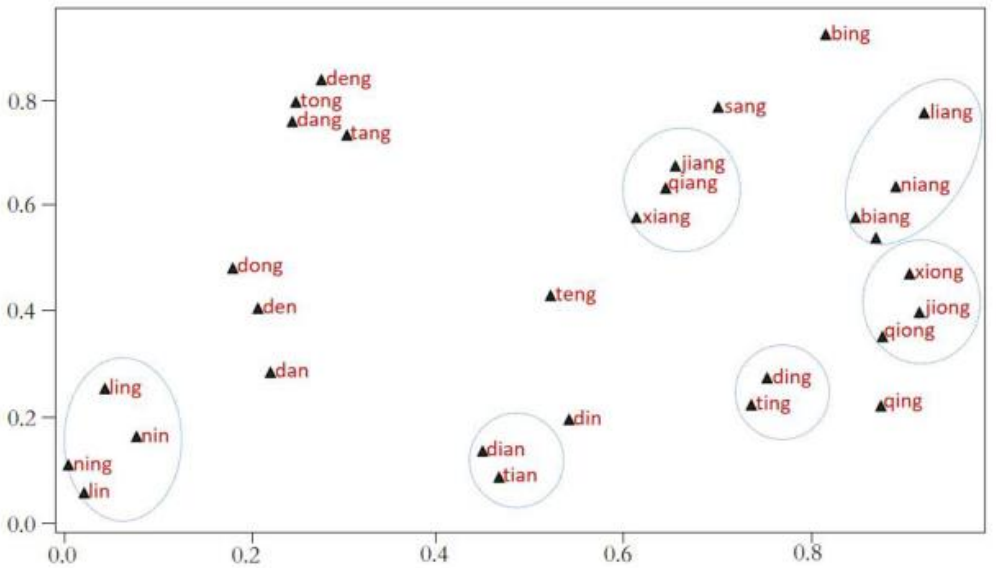


PLOME(Pre-training with Misspelled Knowledge)

- 语音嵌入
- 笔画嵌入
- 可视化结果



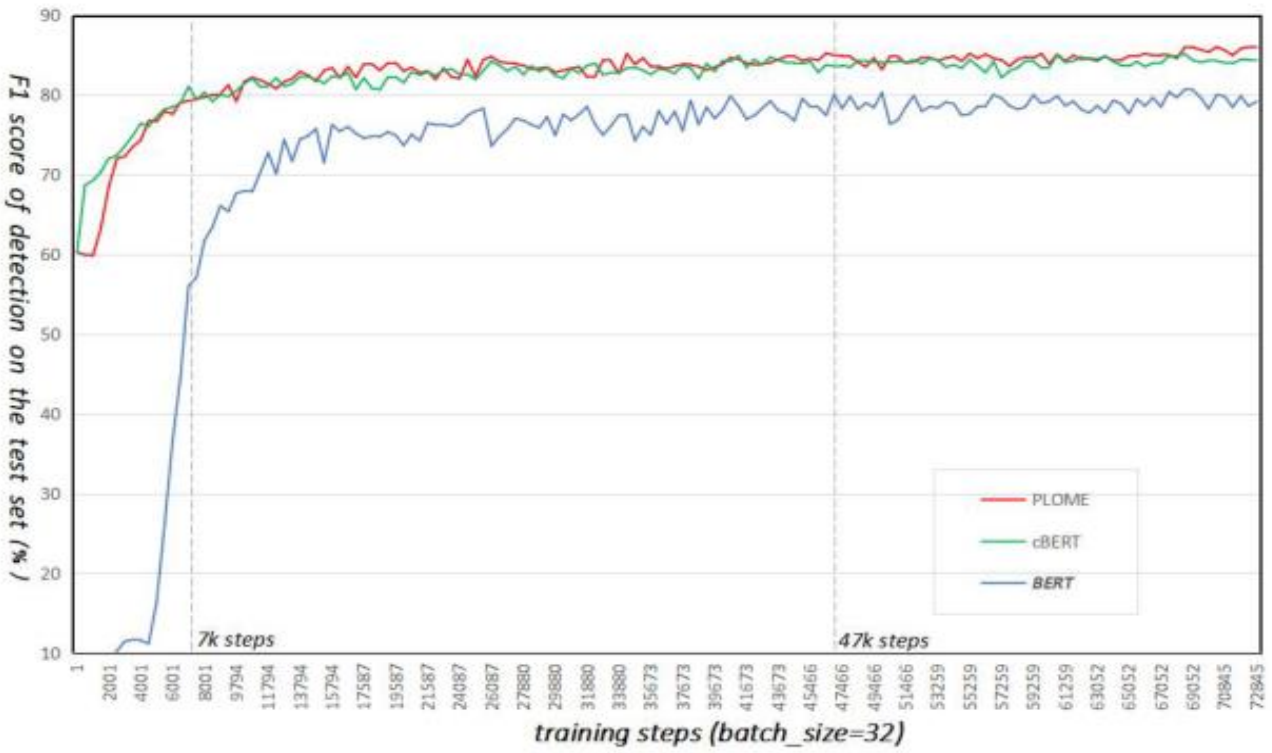




PLOME(Pre-training with Misspelled Knowledge)

- 实验结果
  - 字符和句子级别
  - 模型收敛速度

Method	Character-level on Whole Set						Sentence-level via Official Tool								
	Detection-level			Correction-level				Detection-level				Correction-level			
	P	R	F	P	R	F	FPR	A	P	R	F	A	P	R	F
SpellGCN	77.7	85.6	81.4	96.9	82.9	89.4	13.2	83.7	85.9	80.6	83.1	82.2	85.4	77.6	81.3
BERT-Finetune	76.2	83.1	79.5	96.5	80.3	87.6	14.7	81.7	85.2	76.0	80.3	80.3	84.7	73.5	78.7
cBERT-Finetune	83.0	**87.8**	85.3	96.0	83.9	89.5	**10.6**	84.5	**88.1**	79.6	83.6	82.9	87.6	76.3	81.5
PLOME-Finetune	**85.2**	86.8	**86.0**	**97.2**	**85.0**	**90.7**	10.9	**85.0**	87.9	**80.9**	**84.3**	**83.7**	**87.6**	**78.3**	**82.7**



前沿算法模型对比

- 前沿算法模型
  - FASPell
  - BERT
  - SpellGCN
  - DCN
  - PLOME

- Without Pretraining

Model	D-P	D-R	D-F	C-P	C-R	C-F
FASPell	67.6	60.0	63.5	66.6	59.1	62.6
BERT	73.7	78.2	75.9	70.9	75.2	73.0
RoBERTa	74.7	77.3	76.0	72.1	74.5	73.3
SpellGCN	74.8	**80.7**	77.7	72.1	**77.7**	74.8 (75.9)
DCN	**76.6**	79.8	**78.2**	**74.2**	77.3	**75.7**

- With Pretraining

Model	D-P	D-R	D-F	C-P	C-R	C-F
BERT_CRS + GAD	75.6	80.4	77.9	73.2	77.8	75.4
DCN-pretrain	77.1	80.9	79.0	74.5	78.2	76.3
REALISE	77.3	81.3	79.3	**75.9**	**79.9**	**77.8**
PLOME	**77.4**	**81.5**	**79.4**	75.3	79.3	77.2

Reference

[1] FASPell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm. EMNLP 2019
[2] Spelling Error Correction with Soft-Masked BERT. ACL 2020
[3] Dynamic Connected Networks for Chinese Spelling Check. Findings of ACL 2021
[4] PLOME: Pre-training with Misspelled Knowledge for Chinese Spelling Correction. ACL 2021

[5] Confusionset-guided Pointer Networks for Chinese Spelling Check. ACL 2019
[6] SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check. ACL 2020
[7] Chunk-based Chinese Spelling Check with Global Optimization. Findings of EMNLP 2020
[8] An Alignment-Agnostic Model for Chinese Text Error Correction. EMNLP 2021 short
[9] DCSpell: A Detector-Corrector Framework for Chinese Spelling Error Correction. SIGIR 2021
[10] Correcting Chinese Spelling Errors with Phonetic Pre-training. Findings of ACL 2021
[11] Global Attention Decoder for Chinese Spelling Error Correction. Findings of ACL 2021
[12] Read, Listen, and See: Leveraging Multimodal Information Helps Chinese Spell Checking. Findings of ACL 2021
[13] Exploration and Exploitation: Two Ways to Improve Chinese Spelling Correction Models. ACL 2021
[14] PHMOSpell: Phonological and Morphological Knowledge Guided Chinese Spelling Check. ACL 2021
[15] Guo Z , Chen X , Peng J , et al. Chinese Spelling Errors Detection Based on CSLM[C]// 2015 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). ACM, 2015
[16] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018

语法错误纠正

背景简介

近年来随着自媒体的热潮，人人都是信息的生产者，互联网上语法错误的内容暴增。由于用户在文本输入法，语音输入法使用上的随意性，后续又缺少审核，极易产生语法错误内容。

错误类型	错误示例
字词缺失	自然语处理->自然语言处理
字词冗余	自然语言处处理->自然语言处理
字词乱序	首个开发的空间站->开发的首个空间站
搭配不当	生活水平改善->生活水平提高
结构混乱	靠的是...取得的->靠的是...
知识错误	中国的首都是南京->中国的首都是北京
表意不明	让班长本月15日前去汇报->前/去，前去
逻辑错误	防止这类事故不再发生->防止这类事故再发生

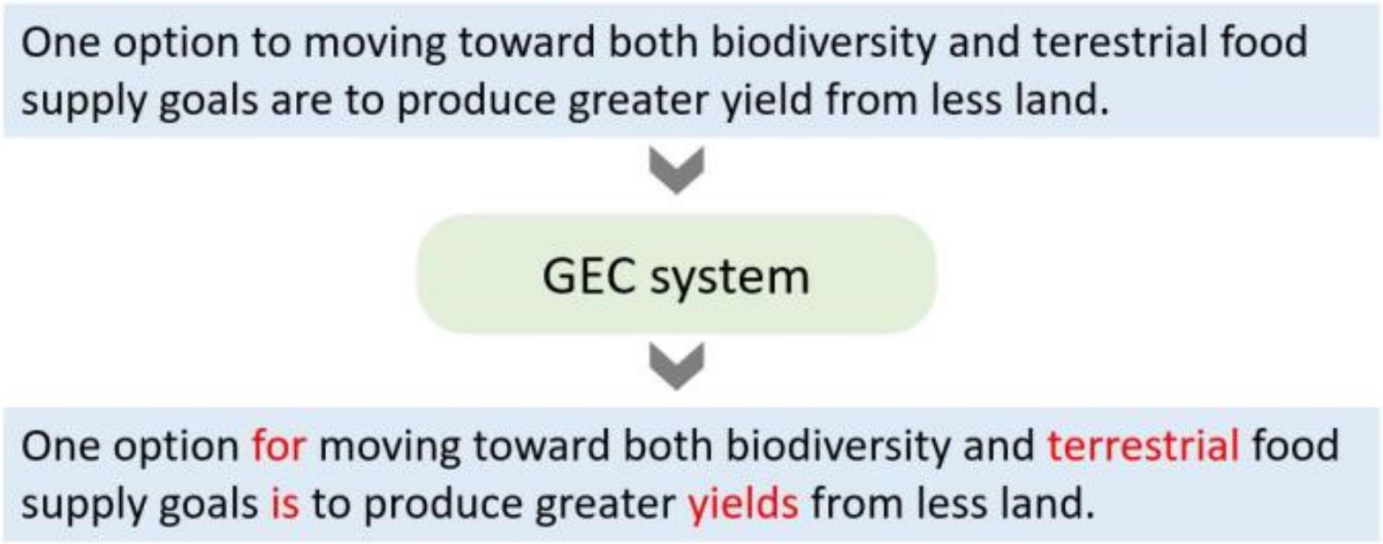
背景简介

在搜索场景中，搜索引擎会对用户输入的 query 纠错后再精准返回搜索结果。

在语音交互场景中，语音系统会将用户的语音转换成正确的文本后再进行后续的意图识别与交互。

基本概念

- 语法纠错（Grammatical Error Correction, GEC）
  - 自然语言处理领域中的一个重要任务
  - GEC任务是指自动纠正文本中存在的语法错误



语法纠错任务的示例

基本概念

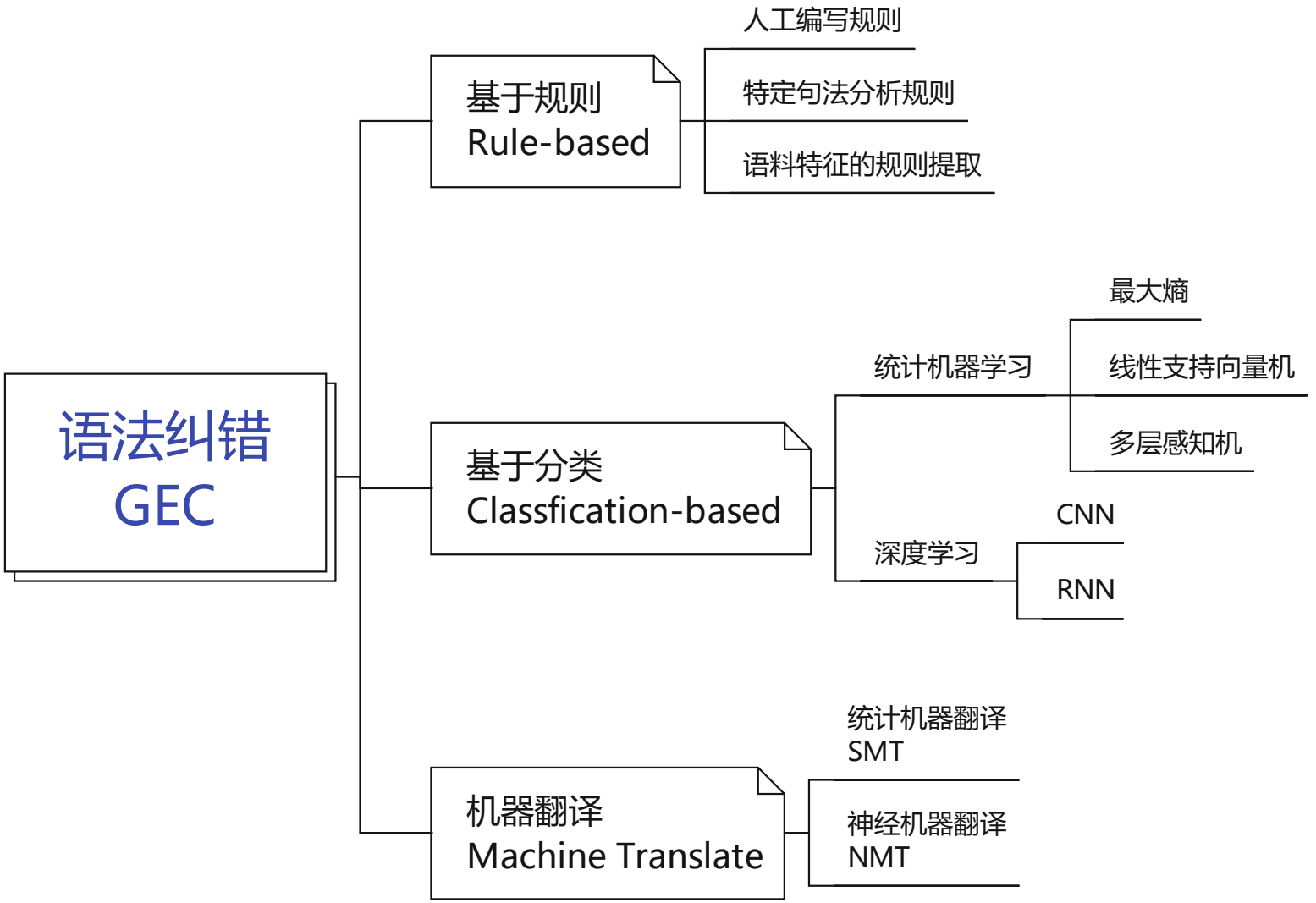
在目前的 GEC 标注体系，语法错误类型多，学习目标很难统一。

同一词语在不同语境中用法不一，为语法错误的判断带来干扰。

在训练语料中，稀疏的长距离信息的学习是机器学习的难点。

在 GEC 领域的含噪训练数据中，如何克服这些噪音以获得学习的稳定性也是难点之一。

基本方法


语法纠错
GEC
基于规则
Rule-based
人工编写规则
特定句法分析规则
语料特征的规则提取
基于分类
Classfication-based
统计机器学习
最大熵
线性支持向量机
多层感知机
深度学习
CNN
RNN
机器翻译
Machine Translate
统计机器翻译
SMT
神经机器翻译
NMT

评价指标

- BLEU
  - 根据精确率(Precision)衡量文本校对的质量
  - 分数取值范围是 0～1，分数越接近1，说明翻译的质量越高。

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} W_n \times \log P_n\right)$$

$$BP = \begin{cases} 1 & lc > \mathrm{lr} \\ \exp(1 - lr/lc) & lc \leq lr \end{cases}$$

$lc =$ 机器译文的长度

$lr =$ 最短的参考翻译句子的长度

评价指标

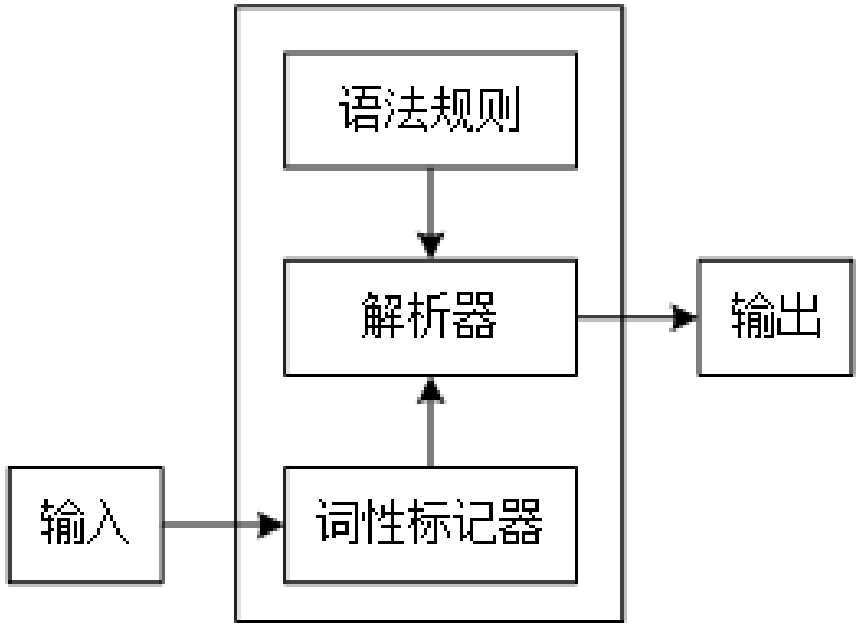
- GLEU
  - BLEU的变体，用于使用n-gram与一组参考句重叠来评估语法错误更正， 而不是特定注释错误的精确率

- F0.5 (M2)
  - 定义召回率与精确率的比重

$$F_\beta = (1+\beta^2) \cdot \frac{preciosn \cdot recall}{(\beta^2 \cdot precision) + recall}$$

基于规则的方法

- 基于规则的GEC
  - 基于人工编制的规则，并结合了解析器和语言特征
  - 根据特定的语法特征，结合规则提取算法，抽取计算机能够识别的规则

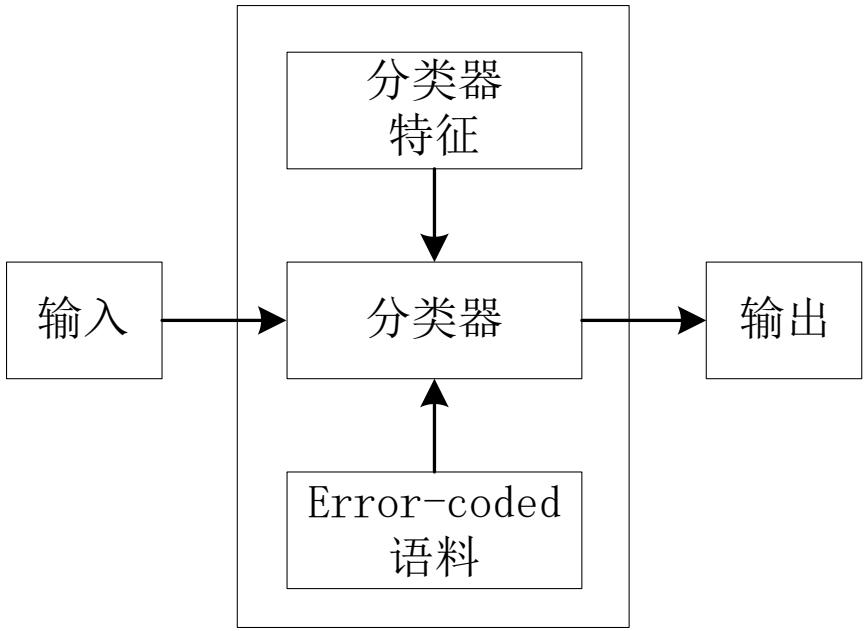


基于规则的方法

- 优势
  - 易于实现
- 不足
  - 语言规则设计的复杂性高
  - 不同语言规则之间可能存在冲突

基于分类的方法

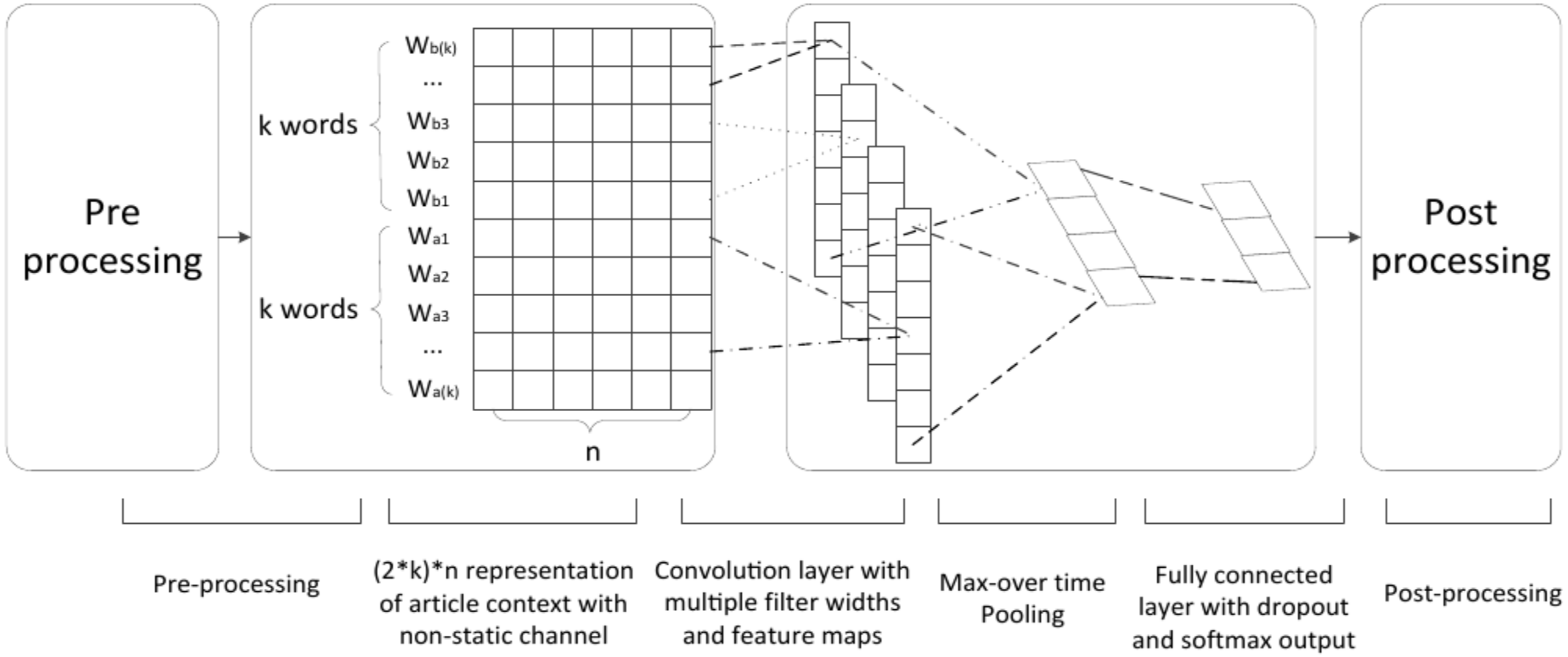
- 基于分类的GEC
  - 考虑上下文给出的语言特征，以预测正确的目标词
  - 分类器在大量无错误文本上进行训练



基于分类的方法

- 特征提取方法
  - 语言学特征工程
    - 语言学知识输入
  - 深度学习
    - 机器自主学习语言特征

基于分类的方法

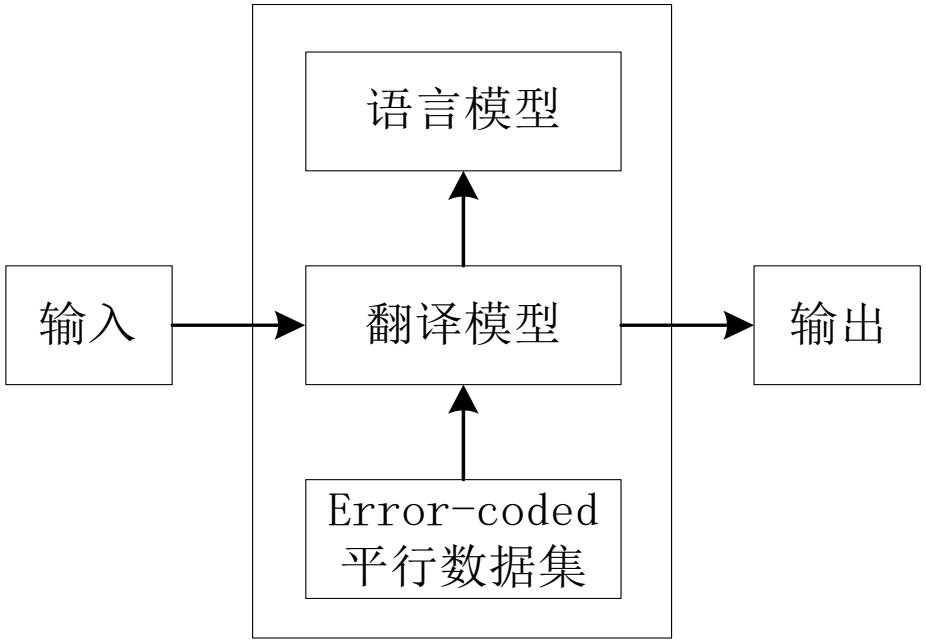


分类器+CNN

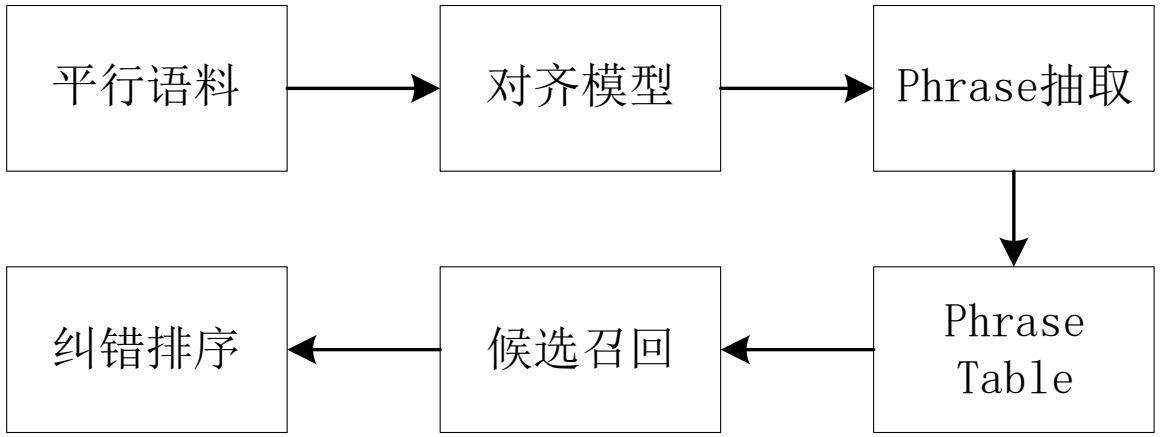
[1] Convolutional Neural Networks for Correcting English Article Errors（NLPCC 2015）

基于机器翻译的方法

- 基于统计机器翻译（SMT）的GEC
  - 将纠错看成翻译过程：错误句子 -> **正确句子**
  - 语言模型：大规模无监督数据，学习语言知识
  - 翻译模型：平行语料学习用户错误行为



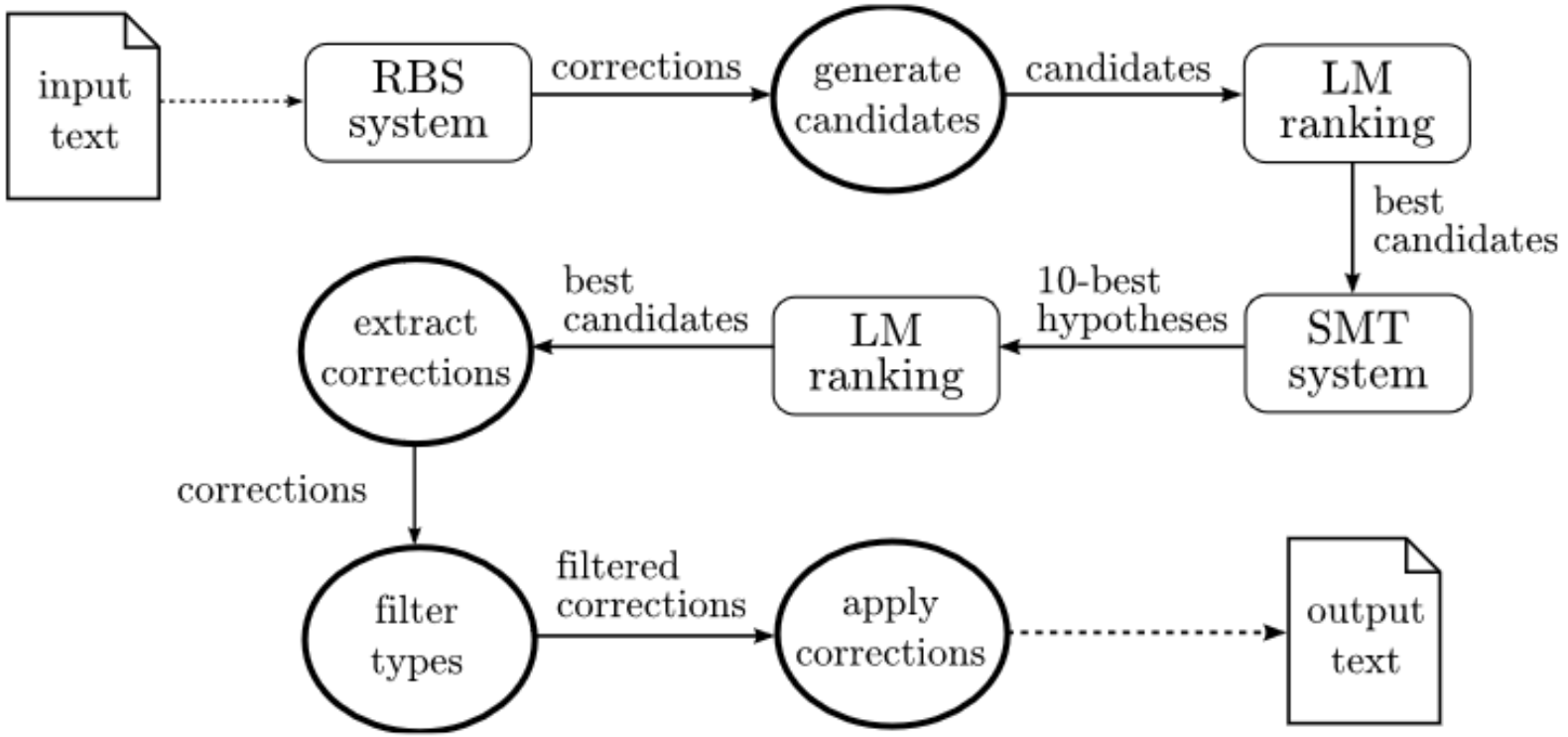
基于统计机器翻译的方法



统计机器翻译流程图

基于SMT的方法

- 结合SMT和基于规则的技术
  - 有些错误通过基于规则的技术可以更好地解决（例如，使用'a'或'an'），有些错误通过机器学习可以更好地解决（例如，定语错误）

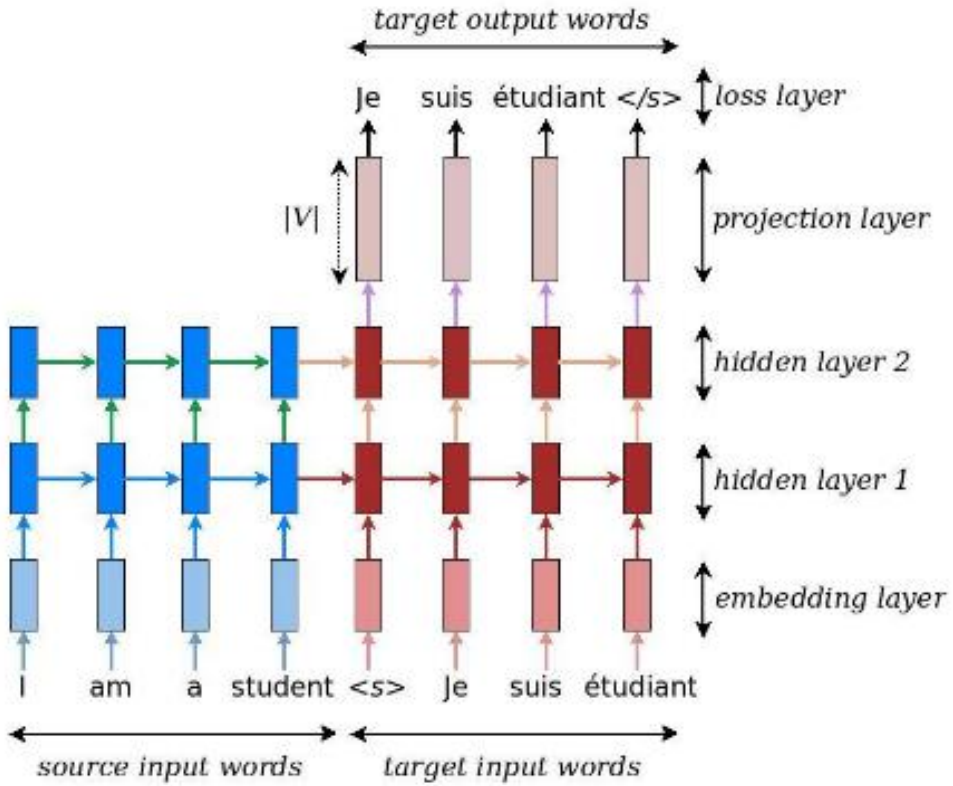


[2] Grammatical error correction using hybrid systems and type filtering（ACL 2014）

基于NMT的语法纠错

基于NMT的方法

- NMT, Neural Machine Translation
  - 神经机器翻译(NMT) 的目标是建立一个单一的神经网络，可以共同调整以最大化翻译性能。
  - Encoder-Decoder



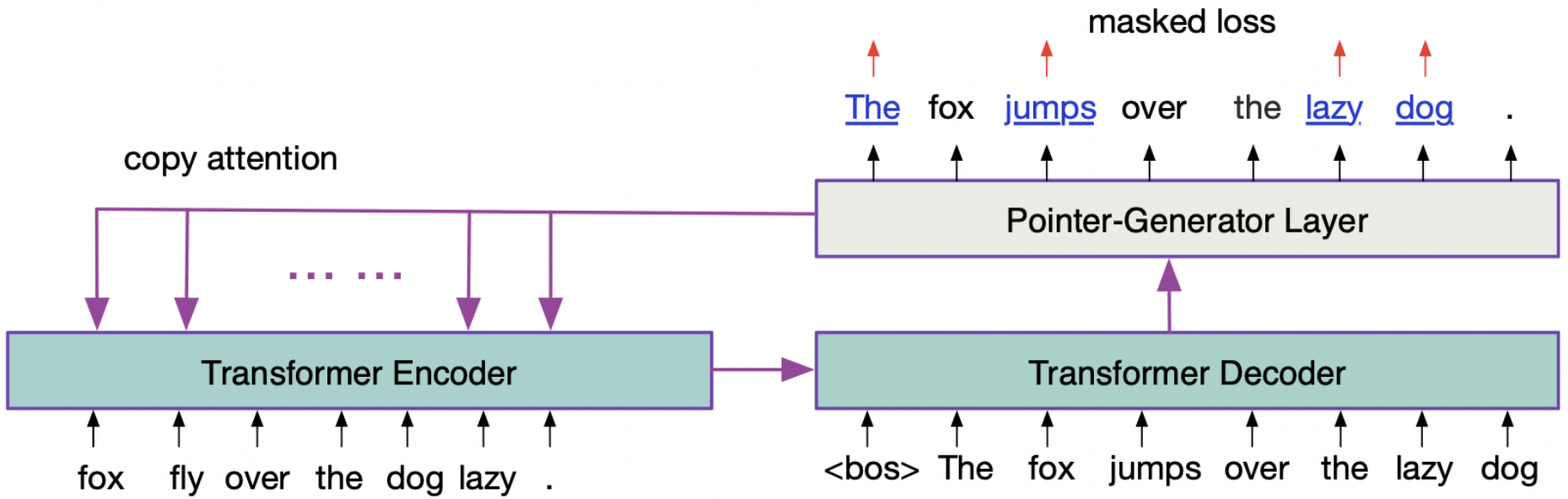
基于NMT的方法

- NMT-GEC
  - 基于 NMT 的 GEC 将有语法问题的“坏”句子翻译成“好”句子。
  - 神经编码器-解码器模型直接从训练数据中学习从源到目标的映射

Type I	今天我感到飞长高兴!	I feel fly long happy today!
Type II	今天我到非常常高兴!	I am always happy when I come to Fei today!
Type III	今天我非常感到高兴!	I very feel happy today!
Correct	今天我感到非常高兴!	I feel very happy today!

语料自动扩充方法

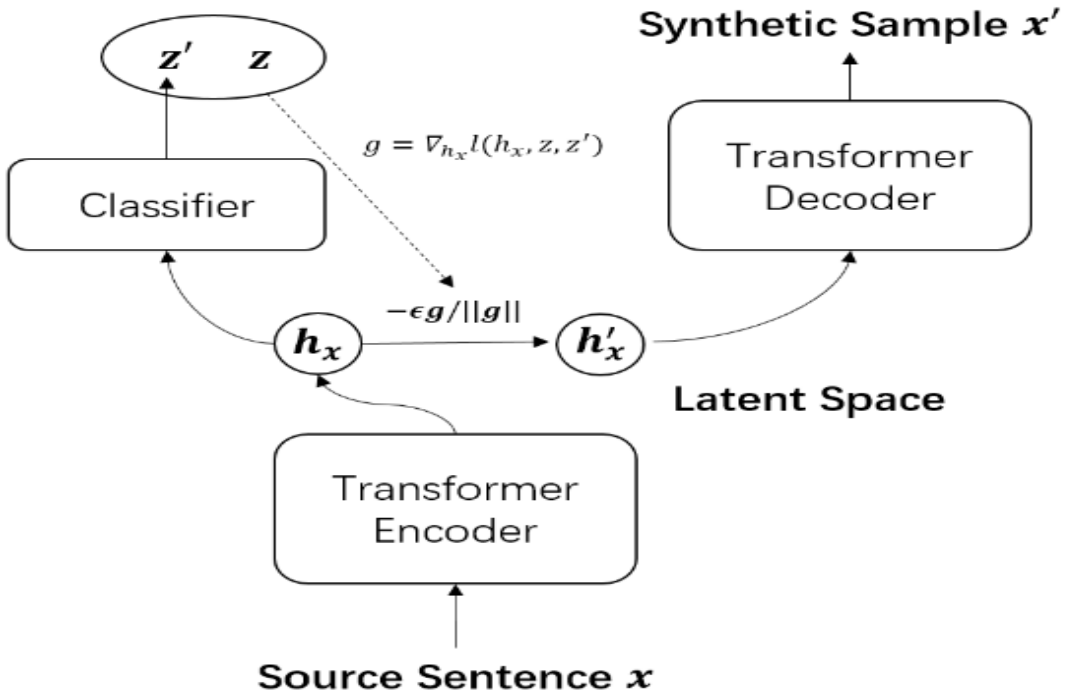
- 基于规则的方法
  - 根据错误类别制定策略构建数据，包括随机删词，随机加词，随机乱序等
  - 借鉴去噪自编码的方法，对文本进行随机处理加噪声，然后使用seq2seq对其复原



[3] Denoising based Sequence-to-Sequence Pre-training for Text Generation（EMNLP 2019）

语料自动扩充方法

- 基于文本生成的方法
  - 生成指定错误类型的数据
  - 伪数据生成



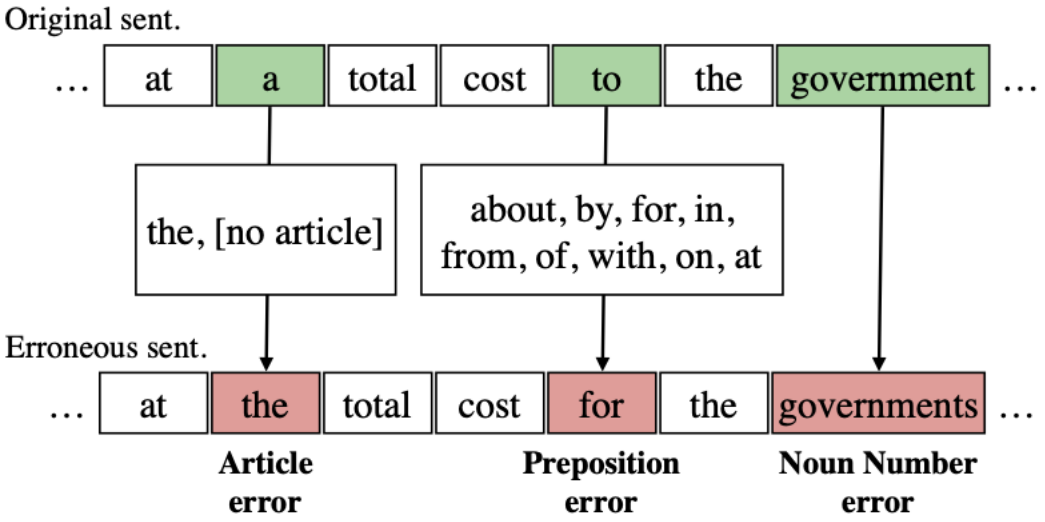
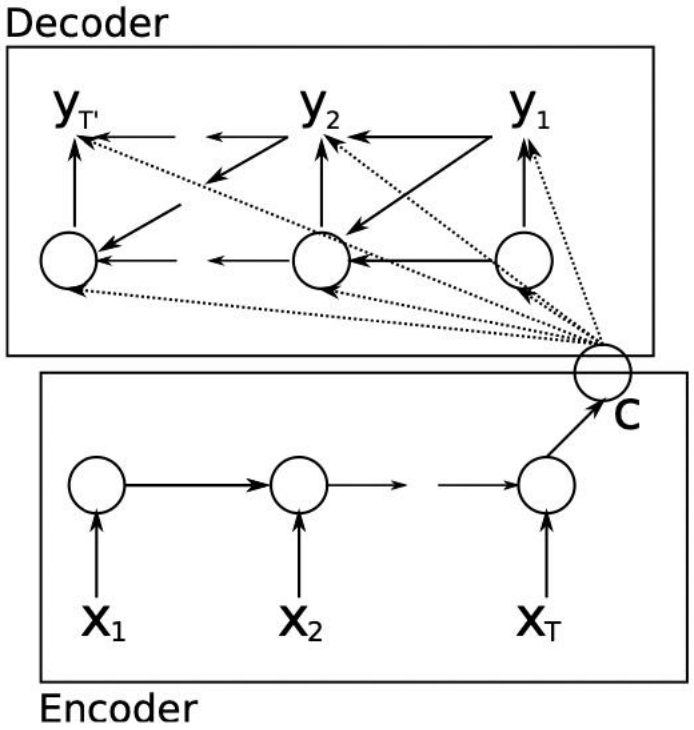


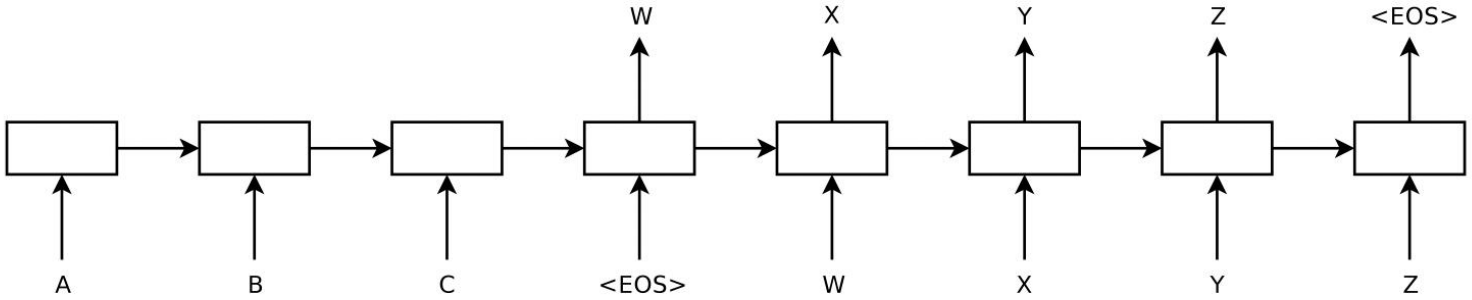
Figure 1: Example of pseudo error generation.

[4] Improving Grammatical Error Correction with Data Augmentation by Editing Latent Representation（COLING 2020）
[5] Grammatical Error Correction Using Pseudo Learner Corpus Considering Error Tendency of Learners（ACL 2021）

seq2seq

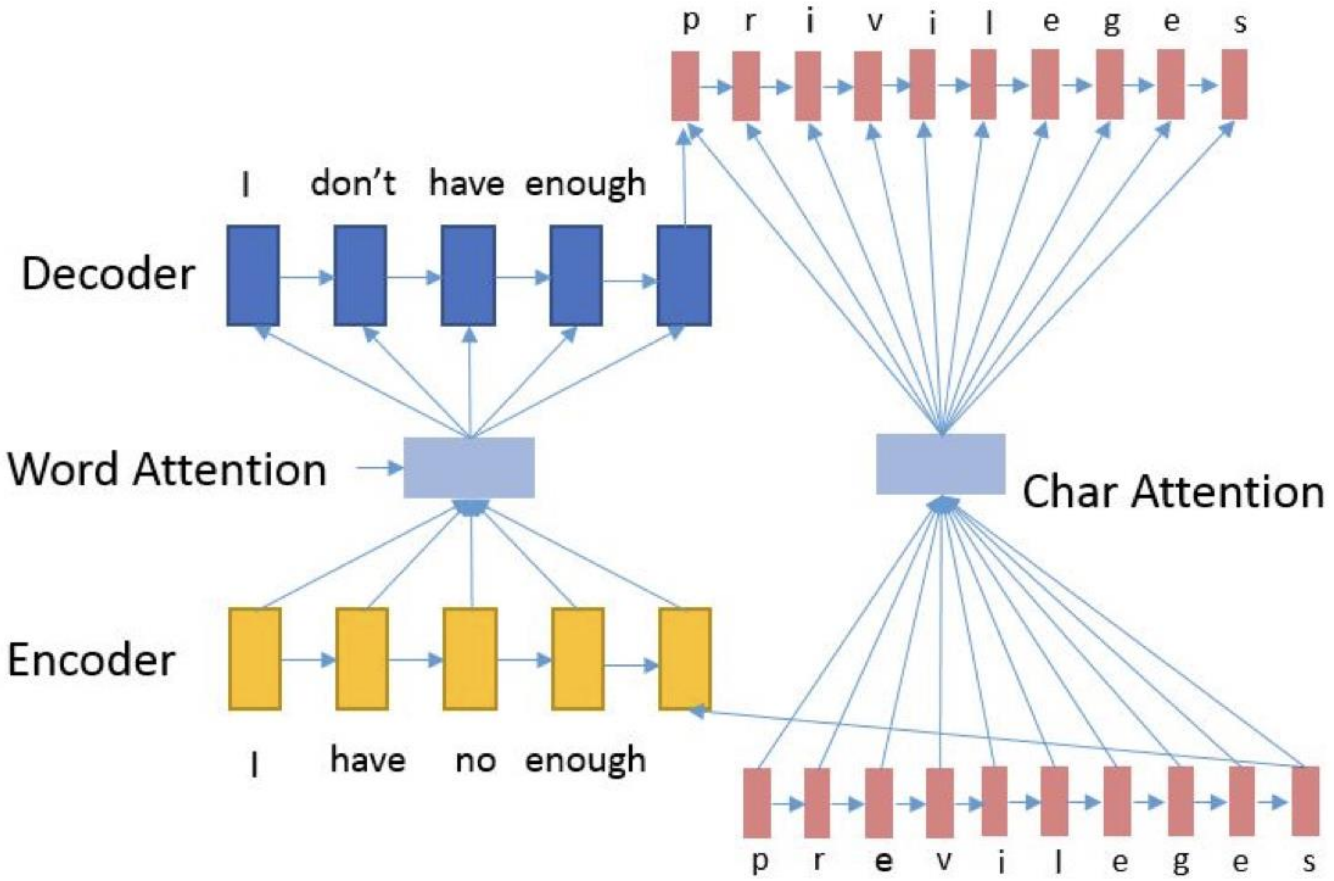
- Sequence-to-Sequence
  - 输入一个序列，用一个 RNN （Encoder）编码成一个向量u，再用一个 RNN （Decoder）解码成一个序列输出，且输出序列的长度是可变的。





神经混合模型

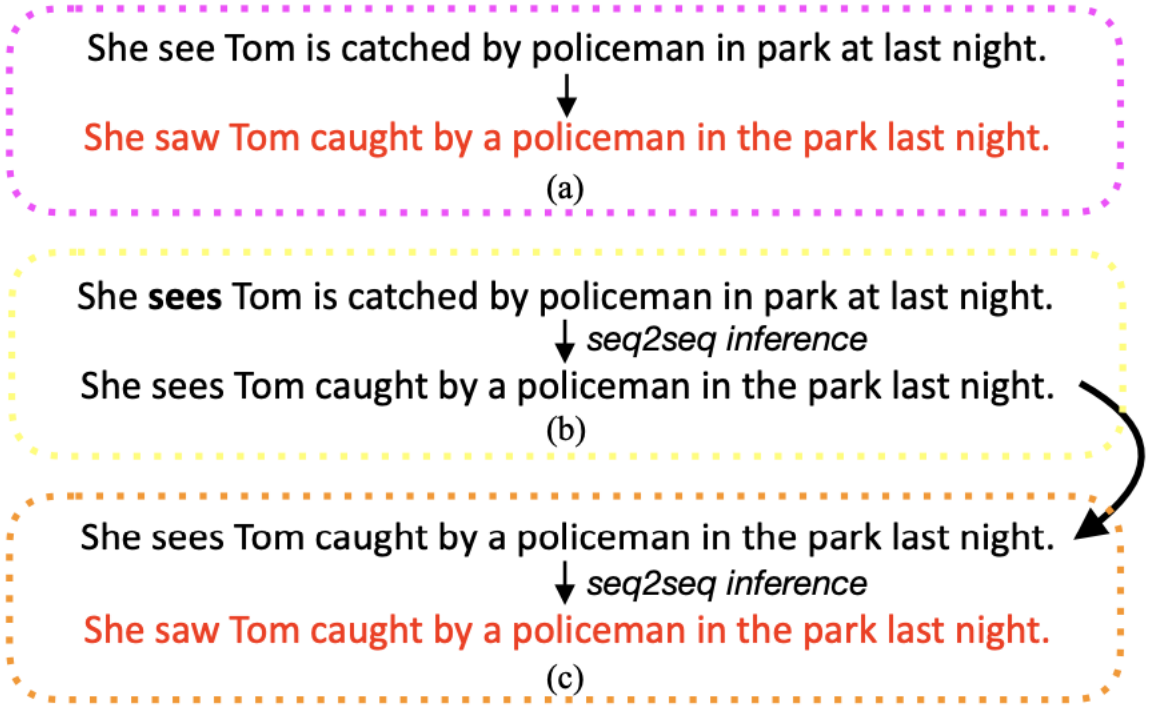
- 嵌套注意力的神经混合模型
  - 以词级的seq2seq模型作为骨干
  - 字符级编解码器
  - 注意力组件



[6] A Nested Attention Neural Hybrid Model for Grammatical Error Correction （ACL 2017）

流畅度提升

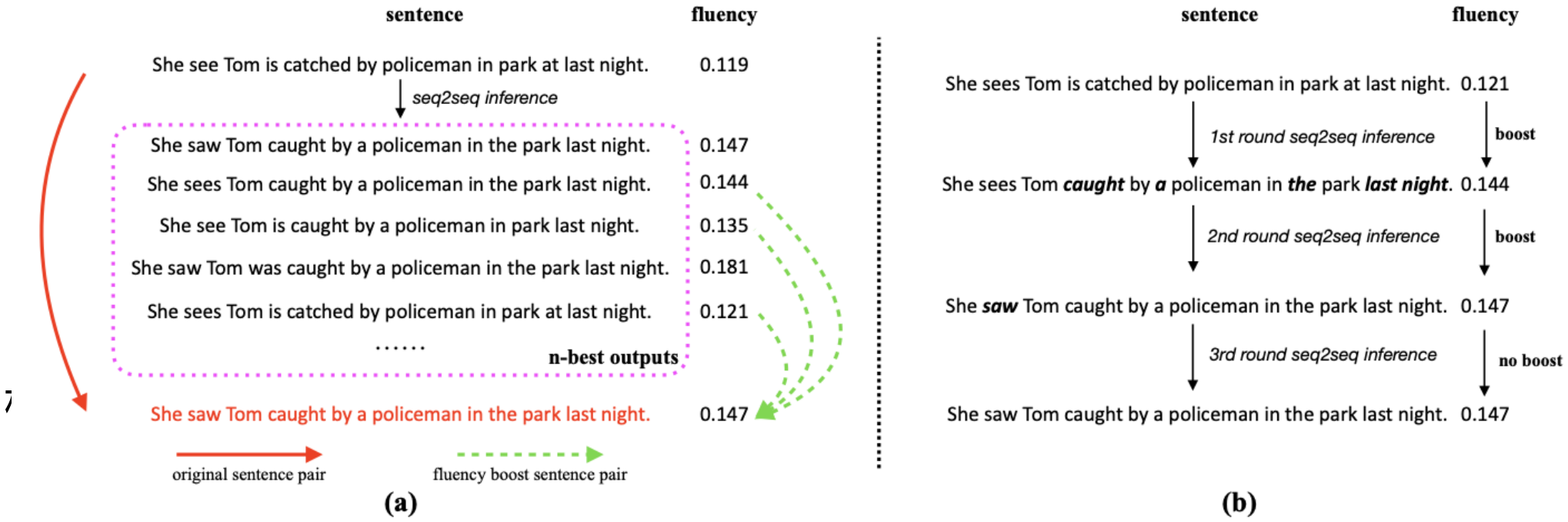
- 基于流畅度提升的语法纠错
  - 传统的seq2seq单次纠错在面对句子有多项错误时，容易纠错不全，甚至越纠越错



[7] Reaching human-level performance in automatic grammatical error correction（ACL 2018）

流畅度提升

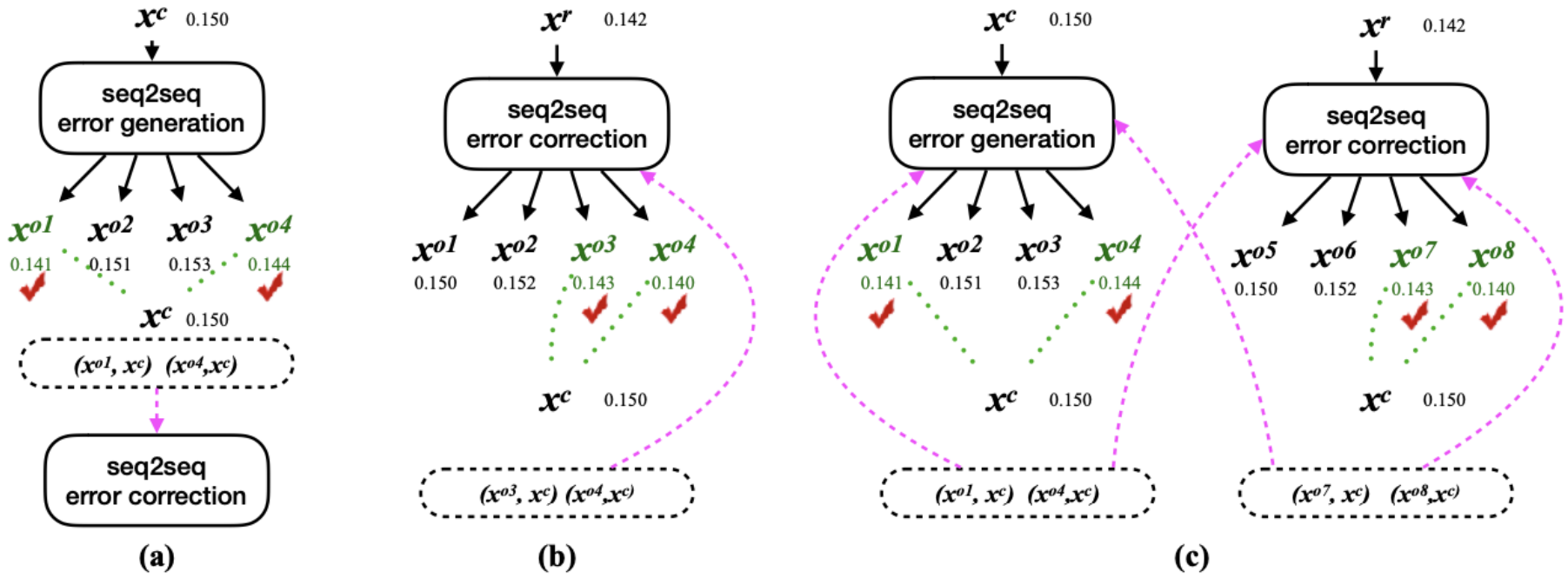
- 流畅度定义
  - 流畅度 f(x)



[7] Reaching human-level performance in automatic grammatical error correction（ACL 2018）

流畅度提升

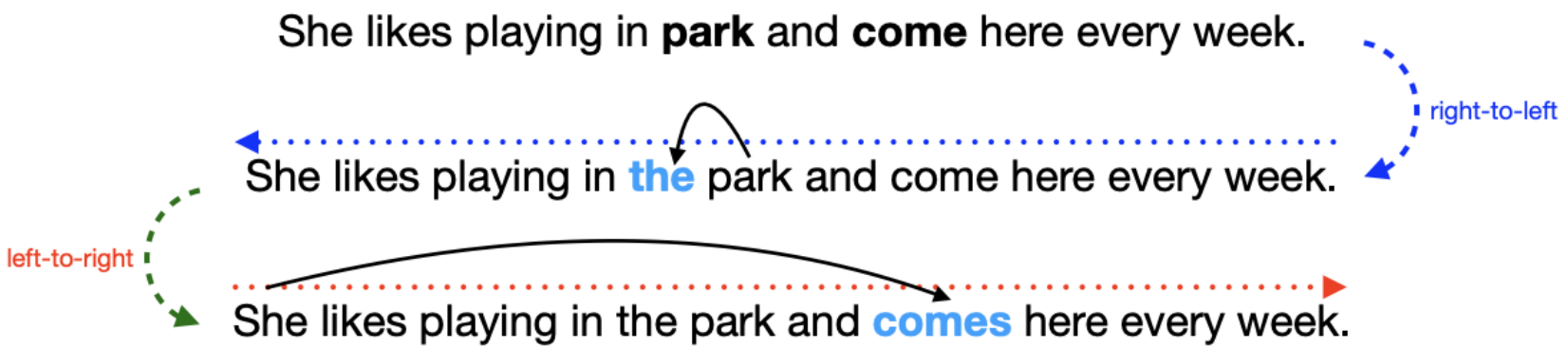
- 提升策略
  - 反向提升，自提升，双向提升



[7] Reaching human-level performance in automatic grammatical error correction（ACL 2018）

流畅度提升

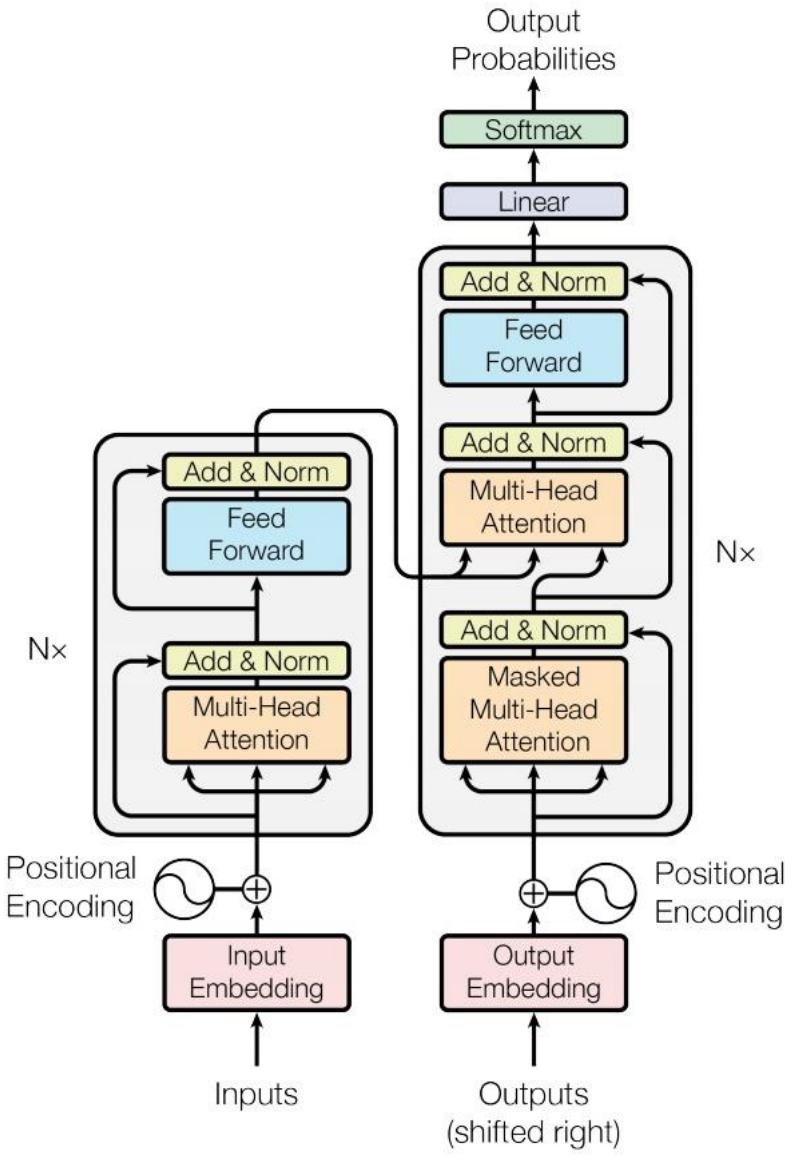
- 往返纠错
  - 某些类型的错误（例如，冠词错误）由从右到左的 seq2seq 模型会更容易纠错，而某些错误（例如主谓一致）由从左到右的 seq2seq 模型更容易纠错。



[7] Reaching human-level performance in automatic grammatical error correction（ACL 2018）

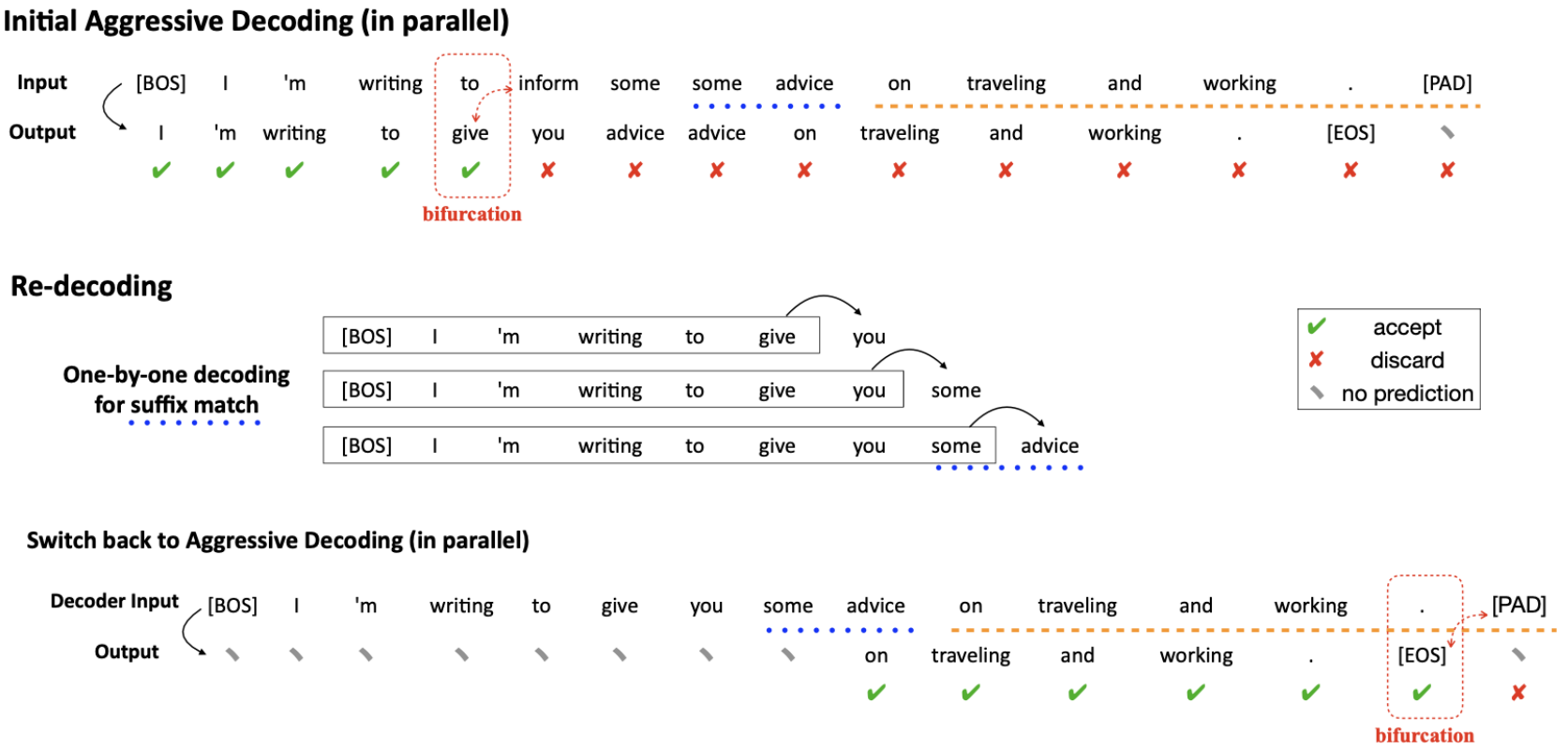
Transformer

- 特点
  - Multi-Head Attention
  - Positional Encoding
- 优点
  - 每层计算复杂度低
  - self-attention
    - 并行计算
    - 解决长时依赖



SAD-即时GEC

- 基于浅层积极解码（SAD）的即时语法纠错
  - 并行解码
  - 浅层解码器



[8] Instantaneous Grammatical Error Correction with Shallow Aggressive Decoding（ACL 2021）

SAD-即时GEC

- 性能对比

Model	Synthetic Data	Total Latency (s)	Speedup	CoNLL-13		
				P	R	$F_{0.5}$
Transformer-big (beam=5)	No	440	1.0×	**53.84**	18.00	**38.50**
Transformer-big (greedy)	No	328	1.3×	52.75	**18.34**	38.36
Transformer-big (aggressive)	No	**54**	**8.1×**	52.75	**18.34**	38.36
Transformer-big (beam=5)	Yes	437	1.0×	**57.06**	23.62	44.47
Transformer-big (greedy)	Yes	320	1.4×	56.45	**24.70**	**44.91**
Transformer-big (aggressive)	Yes	**60**	**7.3×**	56.45	**24.70**	**44.91**

Model	NLPCC-18			
	P	R	$F_{0.5}$	Speedup
Transformer-big (beam=5)	36.0	17.2	**29.6**	1.0×
Levenshtein Transformer⋆	24.9	15.0	22.0	3.1×
LaserTagger⋆	25.6	10.5	19.9	**38.0×**
Span Correction⋆	**37.3**	14.5	28.4	2.7×
Our approach (9+3)	33.0	**20.5**	29.4	**12.0×**

[8] Instantaneous Grammatical Error Correction with Shallow Aggressive Decoding（ACL 2021）

前沿分析

- 低资源场景下的GEC
  - 语料库中的并行数据不足，会损害基于MT的GEC系统性能
  - 数据生成模型不可控，难解释
- 不同系统的组合
  - 探索组合策略以更好地整合不同 GEC 系统的优势，这些 GEC 系统可能专门针对不同的错误类型、主题和句子熟练程度
- 更好的评价指标
  - 虽然现有的评估指标捕获了语法纠正和流畅性，但没有人衡量意义的保留程度。理想的指标应该解释系统输出的语法、流畅度和意义保留程度。

前沿分析

- 近两年（2020～2021）ACL、EMNLP等相关顶会论文共有28篇
  - 多语言语法校对，使用大规模的多语言语言模型
  - 即时语法校对，提升在线推理效率
  - GEC预训练模型，伪数据的生成
  - 文档级的语法校对，上下文感知
  - ……

Reference

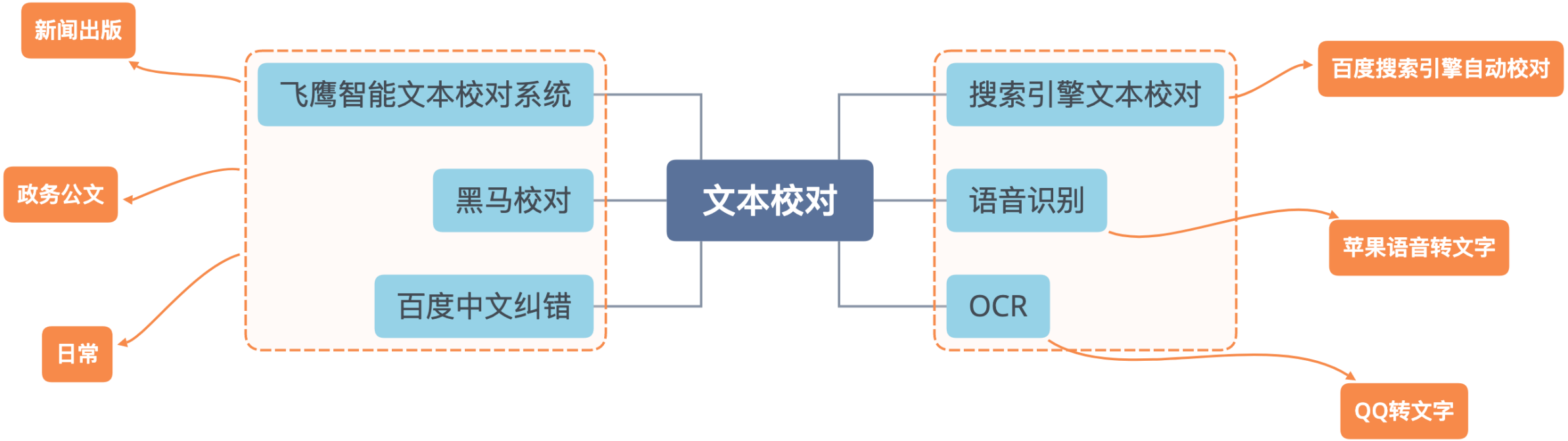
[1] Convolutional Neural Networks for Correcting English Article Errors. NLPCC 2015
[2] Grammatical error correction using hybrid systems and type filtering. ACL 2014
[3] Denoising based Sequence-to-Sequence Pre-training for Text Generation. EMNLP 2019
[4] Improving Grammatical Error Correction with Data Augmentation by Editing Latent Representation. COLING 2020
[5] Grammatical Error Correction Using Pseudo Learner Corpus Considering Error Tendency of Learners. ACL 2021
[6] A Nested Attention Neural Hybrid Model for Grammatical Error Correction. ACL 2017
[7] Reaching human-level performance in automatic grammatical error correction. ACL 2018
[8] Instantaneous Grammatical Error Correction with Shallow Aggressive Decoding. ACL 2021

[9] A Simple Recipe for Multilingual Grammatical Error Correction. ACL 2021
[10] Tail-to-Tail Non-Autoregressive Sequence Prediction for Chinese Grammatical Error Correction. ACL 2021
[11] Grammatical Error Correction as GAN-like Sequence Labeling. ACL 2021
[12] Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction. ACL 2020
[13] Improving the Efficiency of Grammatical Error Correction with Erroneous Span Detection and Correction. EMNLP 2020
[14] Improving Grammatical Error Correction Models with Purpose-Built Adversarial Examples. EMNLP 2020
[15] Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data. NAACL 2019
[16] Fluency Boost Learning and Inference for Neural Grammatical Error Correction. ACL 2018

基本概念

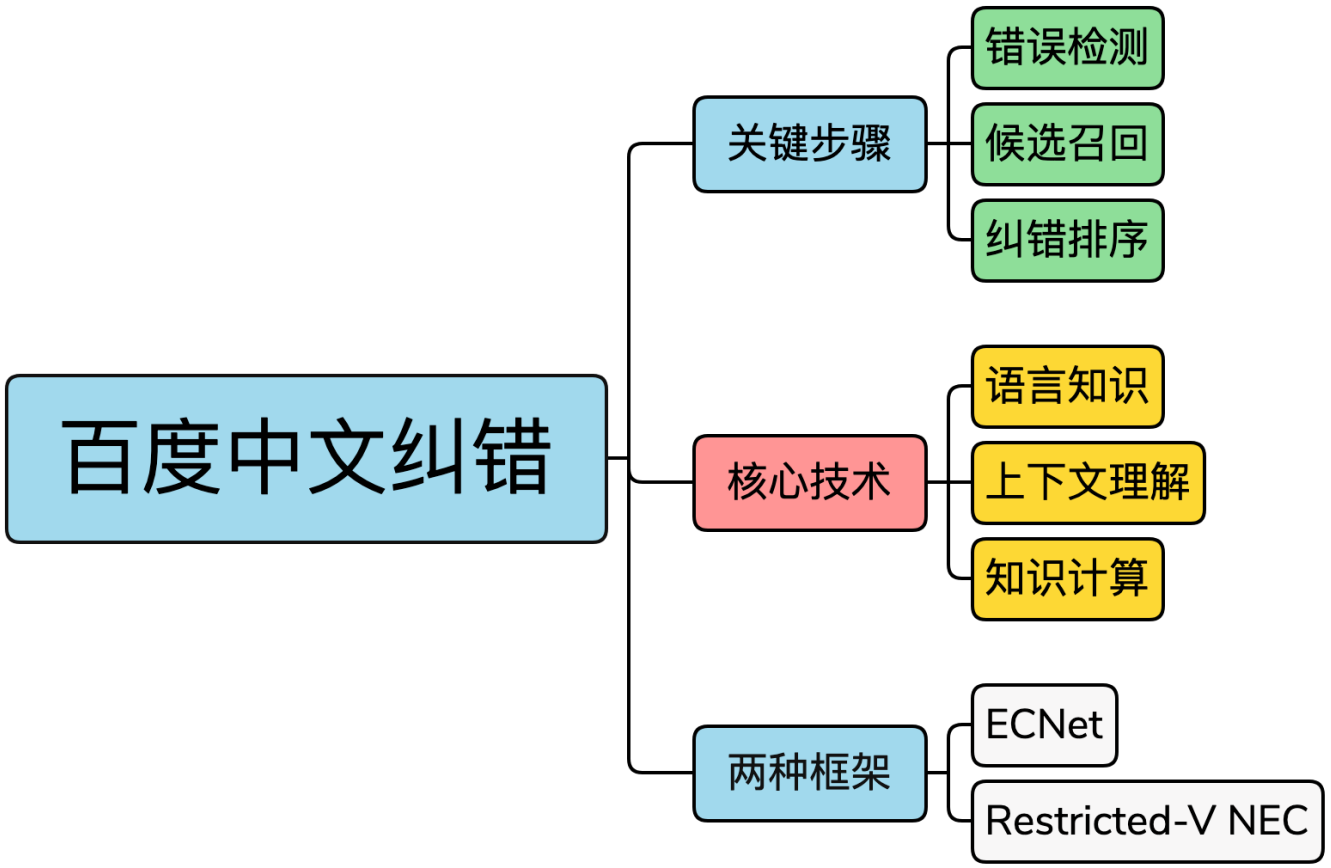
文本校对的基本应用

基本应用与拓展



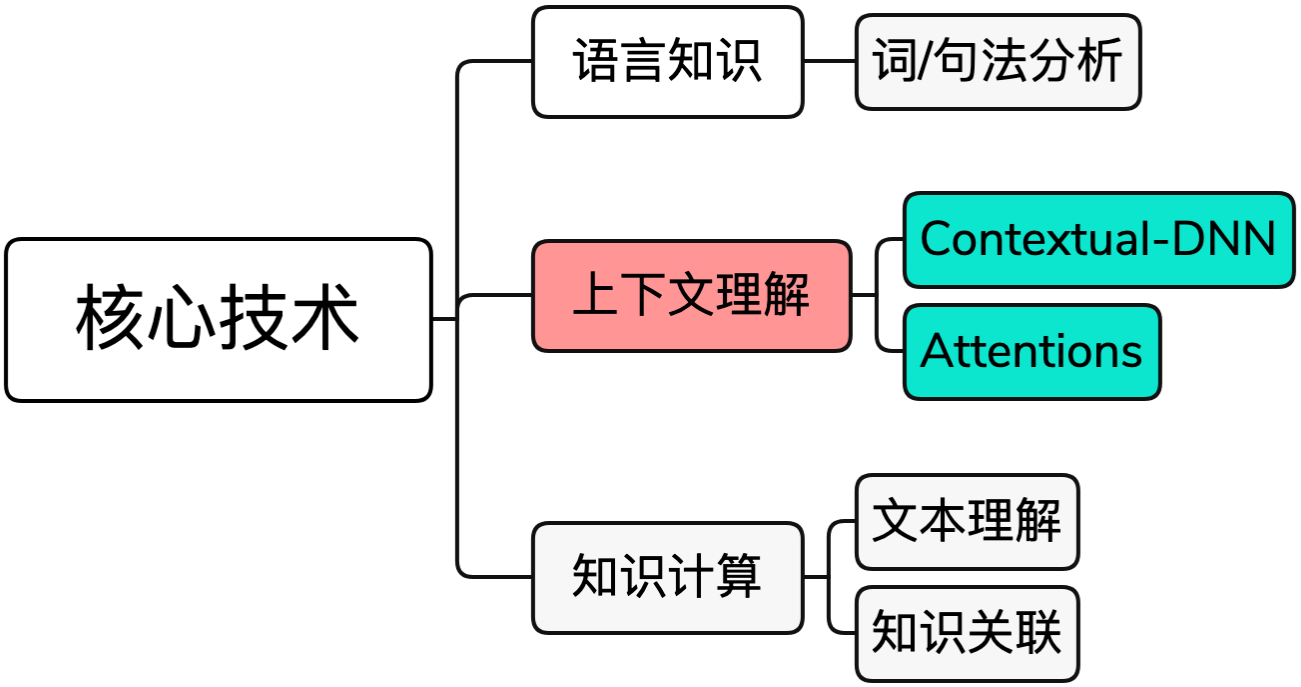
基本 **拓展**

百度中文纠错



- 错误检测是识别输入语句存在的问题，采用(Transformer/LSTM)+CRF
  - 充分利用词法/句法分析等先验知识
  - 特征设计方面，采用DNN以及hard统计特征
  - 根据字粒度和词粒度的特点，解决字对齐
- 候选召回是指结合历史错误行为，以及音形等特征召回纠错候选
- 纠错排序是将候选集中正确的结果排在第一位

百度中文纠错



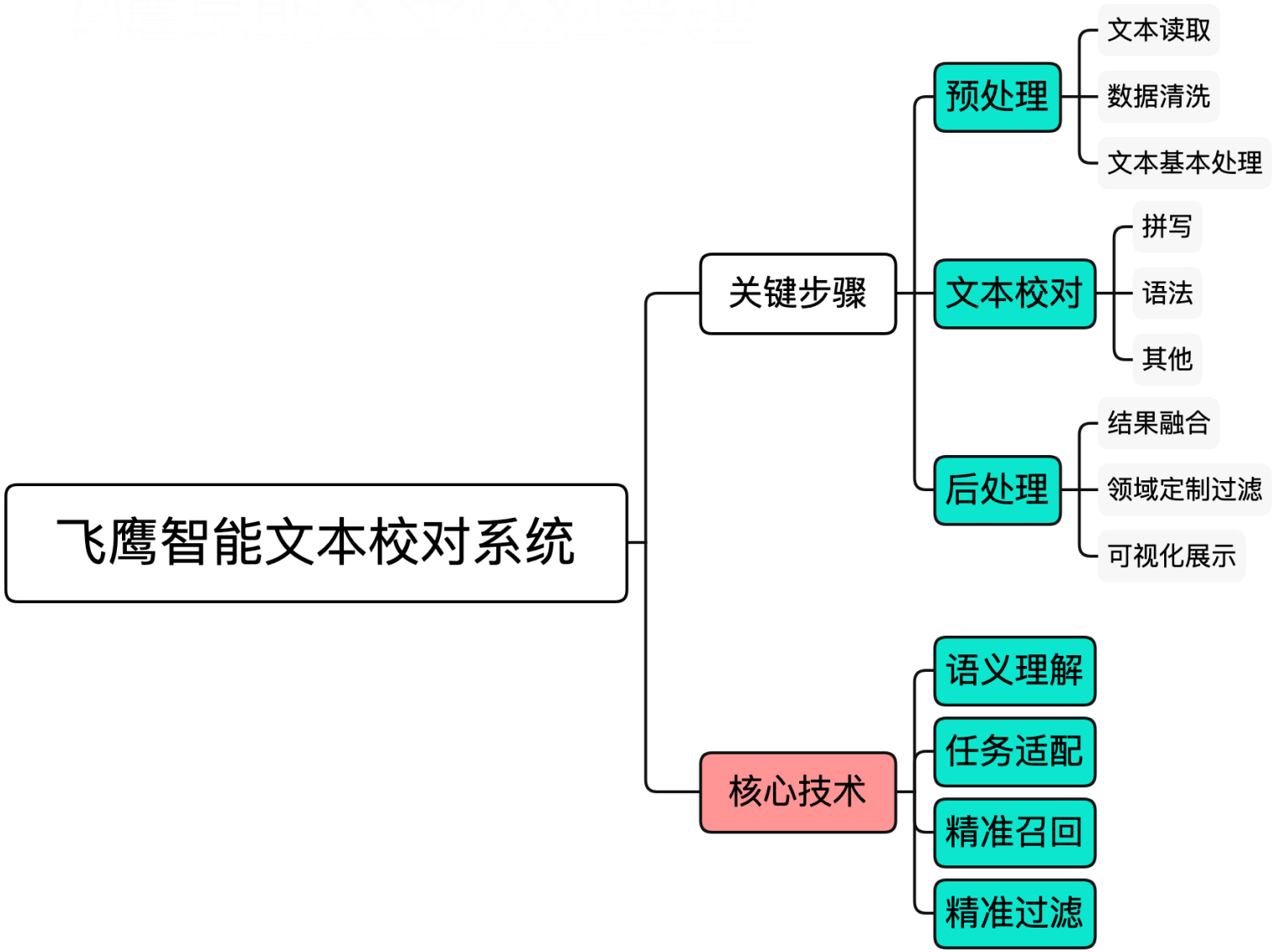
- 语言知识完成对语言规则以及结构的学习

- 上下文理解

使用Contextual-DNN模型来学习，并通过AOA attention 机制解决长依赖问题

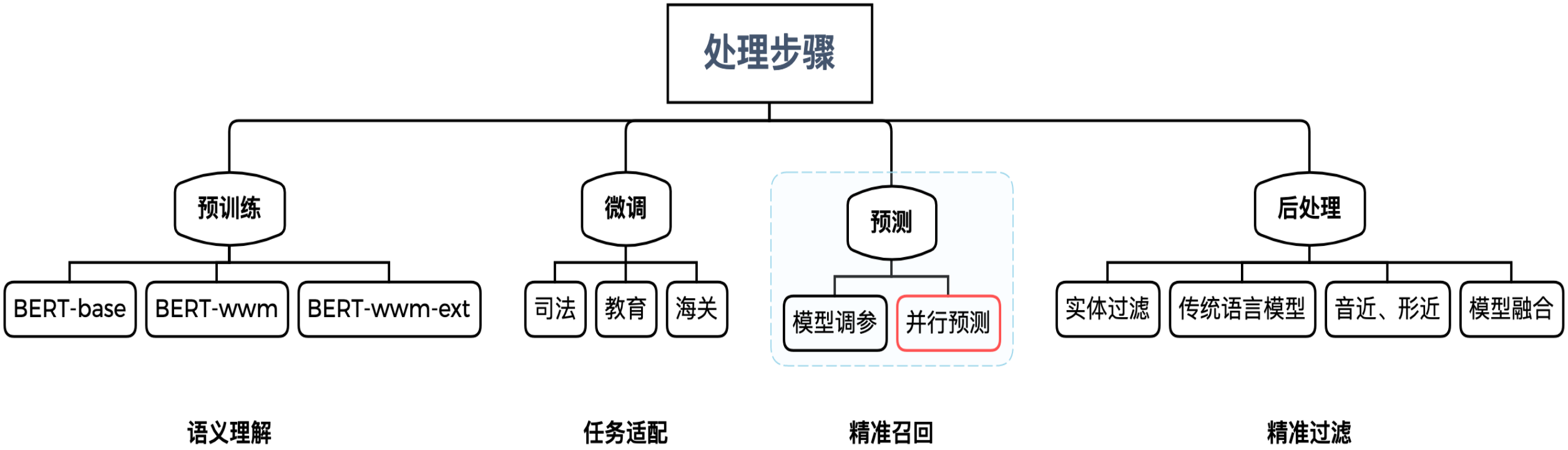
- 知识计算的重点是做好文本理解与知识关联

飞鹰智能文本校对系统



- 飞鹰智能文本校对系统由哈工大讯飞联合实验室基于深度学习自然语言处理技术研发打造。

飞鹰智能文本校对系统


处理步骤
预训练
BERT-base
BERT-wwm
BERT-wwm-ext
微调
司法
教育
海关
预测
模型调参
并行预测
后处理
实体过滤
传统语言模型
音近、形近
模型融合
语义理解
任务适配
精准召回
精准过滤

百度中文纠错-语料测试

```
from aip import AipNlp
import time
app_id      =    '25158499'
api_key     =    'MqUSbUCiT7ULiGgNy7kRSolD'
secret_key  =    'iPsilsGII5QDUd9LiunmapynpseacE3u'

client = AipNlp(app_id, api_key, secret_key)

client = AipNlp(app_id, api_key, secret_key)


text_list = [
            '好象是我们错了。',
            '现在从应当由教育行政部门办成公办园或委托办成普惠性民办园。',
            '一年有四百天。',
            '国院各部委各直属机构，'
            '就难免必理不平衡。',
            '还有进口香皂、家居服、花艺样样聚全。',
            '成立城镇小区配套幼儿园幼儿园治理工作小组',
            '你说我不该不该，不该在这时候来桶泡面',
            '你说我不该不该，不该在这时候才说爱你。']
```

– 调用百度文本纠错API，进行测试

飞鹰智能文本校对系统-语料测试

结果对比

序号	错误类型	原始文本	百度中文纠错	飞鹰智能文本校对系统
1	拼写错误	好**象**是我们错了。	好像是我们错了。	好像是我们错了。
2	语法错误	现在**从**应当由教育行政部门办成公办园或委托办成普惠性民办园。	现在从应当由教育行政部门办成公办园或委托办成普惠性民办园。	现在从应当由教育行政部门办成公办园或委托办成普惠性民办园。
3	语义错误	一年有**四百天**。	一年有四百天。	一年有四百天。
4	语法错误，标点错误	**国院**各部**委**各直属机构，	国务院各部委各直属机构：	国务院各部委各直属机构：
5	拼写错误	就难免**必理**不平衡。	就难免必理不平衡。	就难免比例不平衡。
6	语法错误，拼写错误	**还有**进口香皂、家居服、花艺样样**聚全**。	还有进口香皂、家居服、花艺样样聚全。	还有进口香皂、家居服、花艺样样俱全。
7	语法错误	成立城镇小区配套**幼儿园**幼儿园治理工作小组	成立城镇小区配套幼儿园幼儿园治理工作小组	成立城镇小区配套幼儿园幼儿园治理工作小组
8	标点错误	你说我不该**不该，不该**在这时候来桶泡面	你说我不该不该，不该在这时候来桶泡面	你说我不该，不该，不该在这时候来桶泡面。
9	无	你说我不该不该，不该在这时候才说爱你。	你说我不该不该，不该在这时候才说爱你。	你说我不该不该，不该在这时候才说爱你。

应用对比

	百度中文纠错	飞鹰智能文本校对系统
多类型	√	√
场景迁移	地图检索、语音通话	司法、教育、海关
多模态	支持文本、语音等形式	×
特点	上下文理解	并行化校对

- 百度中文纠错对拼写错误的纠错准确率较高但从拼写纠错，语法纠错，标点纠错上的准确率来说，都没有飞鹰智能文本校对系统好

- 百度中文纠错支持文本、语音等形式

- 飞鹰智能文本校对系统、百度纠错的语义纠错准确率都偏低

拓展应用

搜索引擎文本校对

- 对查询词进行校正可以提高查询效率

- 常用的纠错工具有Pycorrector，可用于中文拼音、笔画输入法的错误纠正

语音识别（ASR)

00:33 - 00:36
如果有一双这样的鞋，

00:36 - 00:40
人在心理上就会产生比别人更多的优越感。

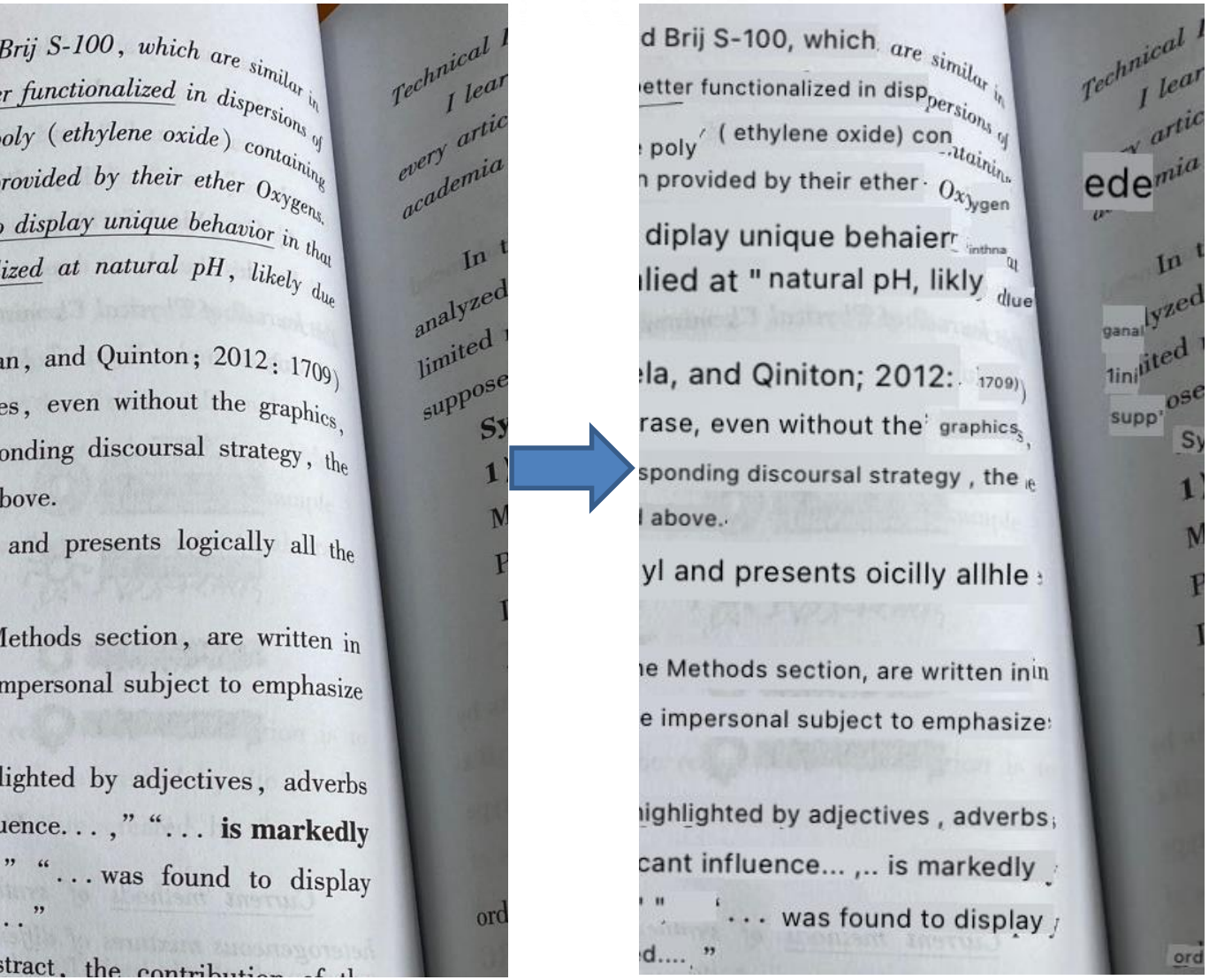
00:40 - 00:44
因为这种鞋子不是人人都买得起的，

00:46 - 00:48
这就是刷存在感。

00:51 - 00:56
在没有电话的年代里，电报是很昂贵的通讯工具！

- ASR纠错，着重于辨别“音似”，“发音相近”导致的错误
- 普通错别字纠错，着重于“拼写”导致的错误

光学字符识别（OCR）



– OCR就是将纸上的字符识别出来。

– 通过对Yolo检测后的倾斜文本进行文本校正，可以提高OCR识别的准确率

缺点及目前的解决方案

Dynamic Connected Networks for Chinese Spelling Check

Baoxin Wang[1,2], Wanxiang Che[1], Dayong Wu[2], Shijin Wang[2,3], Guoping Hu[2], Ting Liu[1]
[1]Research Center for SCIR, Harbin Institute of Technology, Harbin, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[3]iFLYTEK AI Research (Hebei), Langfang, China
{bxwang2,dywu2,sjwang3,gphu}@iflytek.com
{car,tliu}@ir.hit.edu.cn

Abstract

Chinese spelling check (CSC) is a task to detect and correct spelling errors in Chinese text. Most state-of-the-art works on the CSC task adopt a BERT-based non-autoregressive language model, which relies on the output independence assumption. The inappropriate independence assumption prevents BERT-based models from learning the dependencies among target tokens, resulting in an incoherent problem. To address the above issue, we propose a novel architecture named Dynamic Connected Networks (DCN), which generates the candidate Chinese characters via a Pinyin Enhanced Candidate Generator and then utilizes an attention-based network to model the dependencies between two adjacent Chinese characters. The experimental results show that our proposed method achieves a new state-of-the-art performance on three human-annotated datasets.

Wrong: 我忘记告诉你了，我真户秃。
Correct: 我忘记告诉你了，我真糊涂。 Translation: I forgot to tell you. I'm so confused.

Table 1: An example of Chinese spelling errors. Here, "户秃" should be corrected to "糊涂" (confused).

character will be considered as a spelling error and corrected to the most likely character. Based on the powerful generalization ability of BERT (Devlin et al., 2019), these works have achieved better performance than other models.

However, these works on the CSC task rely on the incorrect independence assumption, which may lead to an incoherent problem. Concretely, they assume that the predicted tokens are independent of each other, which generally does not hold in natural language (Yang et al., 2019; Gu and Kong, 2020). For the CSC task, one spelling error may have mul-

计算语言学协会的研究结果：ACL-IJCNLP 2021，第2437-2446页，2021年8月1-6日。
2437汉语拼写检查动态连接网络 Baoxin Wang1,2，Wanxiang Che1，Dayong Wu2，Shijin Wang2,3，郭平 Hu2，Ting Liu11中国哈尔滨工业大学SCIR研究中心哈尔滨，中国2科大讯飞研究院认知智能国家重点实验室，中国3IFLYTEK人工智能研究（河北），廊坊，中国{bxwang2，dywu2，sjwang3，gphu}@iflytek.com{car，tliu}@ir.hit.edu.cnAbstractChinese拼写检查（CSC）是一项检测和纠正中文文本拼写错误的任务。大多数关于CSCtask的最新工作采用基于BERT的非自回归语言模型，该模型依赖于outputindependence假设。不适当的独立性假设阻止基于BERT的模型学习目标标记之间的依赖关系，从而导致不可理解的问题。为了解决上述问题，我们提出了一种称为动态连接网络（dynamicconnectednetworks，DCN）的新体系结构，该结构通过一个pinyin增强的候选生成器生成候选汉字，然后利用基于注意的网络来模拟两个相邻汉字之间的依赖关系。实验结果表明，我们提出的方法在三个人体数据集上取得了新的性能。

- 文献翻译的文本校对精度仍然不够高

缺点及目前的解决方案

- 对歧义字段、生词、新词以及专用名词没有很好的辨识能力。
  建立姓氏名字用字频率表、称谓表、指界动词

- 人工校对与自动校对有较大差异
  在校对时增加自动校对系统的背景知识和上下文关联信息

拓展应用

Demo展示

算法性能对比

- 技术路线
  - 框架
    - Pycorrector
  - 参与对比的模型
    - kenlm（Pycorrector提供，作为baseline）
    - BERT（自行训练）
    - MacBERT（自行训练）
  - 数据集
    - SIGHAN 2013/2014/2015
    - SIGHAN基础上生成的语料[1]
    - NLPCC2018 中文语法错误修正任务数据集

Pycorrector

SIGHAN

[1] A Hybrid Approach to Automatic Corpus Generation for Chinese Spelling Checking (EMNLP2018)

算法性能对比

- 技术路线
  - 测试集
    - Pycorrector corpus500

	准确率	召回率	预测时间（500句）
kenlm	49.2%	28.2%	61s
BERT	68.6%	63.6%	527s
MacBERT	70.8%	65.4%	20s

```
original sentence:炮弹把船底炸了一个洞，江水立既俑了进来。 => 炮弹把船底炸了一个洞，江水立即涌了进来。, err:[('既', '即', 14, 15), ('俑', '涌', 15, 16)]
original sentence:少先队员因该为老人让坐 => 少先队员应该为老人让座, err:[('因', '应', 4, 5), ('坐', '座', 10, 11)]
original sentence:机七学习是人工智能领遇最能体现智能的一个分知 => 机器学习是人工智能领域最能体现智能的一个分支, err:[('七', '器', 1, 2), ('遇', '域', 10, 11), ('知', '支', 21, 22)]
original sentence:我很高兴受到你们结婚的邀请单，你们到底决定结婚了！ => 我很高兴收到你们结婚的邀请单，你们到底决定结婚了！, err:[('受', '收', 4, 5)]
original sentence:天气真不错，他们很开心的一边聊天一边烤肉 => 天气真不错，他们很开心的一边聊天一边烤肉, err:[]
original sentence:你这个垃鸡模型只能做错别字检测 => 称这个垃圾模型只能做错别字检测, err:[('你', '称', 0, 1), ('鸡', '圾', 4, 5)]
```

实际场景测试

- 测试结果
  - 模型虚警率相对较高

```
original sentence:全会期间，记者采访了15位与会同志。他们有的任职部委，有的主政一方，也有的来自田间地头和实验室。同志们评价，党的十九届六中全会是我们党的历史上的一座里程碑，决议是一篇马克思主义的纲领性文献，是我们党百年奋斗的皇皇巨著。 => 全会期间，记者采访了15位与会同志。他们有的任直部委，有的属政一方，也有的来自田间地头和实验室。同志们评价，党第十九届六中全会是我们党的历史上的一座里程碑，决议是一篇马克思主义的纲领性文献，是我们党百年奋斗的皇皇巨著。，err:[('职', '直', 23, 24), ('主', '属', 29, 30), ('的', '第', 55, 56)]
```

全会期间，记者采访了15位与会同志。他们有的任职（直）部委，有的主（属）政一方，也有的来自田间地头和实验室。同志们评价，党的（第）十九届六中全会是我们党的历史上的一座里程碑，决议是一篇马克思主义的纲领性文献，是我们党百年奋斗的皇皇巨著。

实际场景测试

- 测试结果
  - 无法应对缺字词和多字词问题
  - 局限于日常领域，对于其他领域的泛化性差
  - 缺少“灵活变通”
  - 难以应对新词和概念漂移带来的问题

艾条燃烧后又淡淡的白色烟雾	艾条燃烧后有淡淡的白色烟雾
更加应证了如今的国际局势愈发波谲云诡	更加印证了如今的国际局势愈发波谲云诡
稍稍一动弹就就是一身汗	稍稍一动弹就就是一身汗
国院办公厅关于开展城镇小区配套幼儿园治理工作的通知	国院办公厅关于开展城镇校区配套幼儿园治理工作的通知
一叶易色而知天下秋	一叶一色而知天下秋
“你喝茶就喝茶呀哪来这么多话，莫讲勒些花哨话，听着就很假！”	“你喝茶就喝茶呀哪来这么多话，莫讲那些花哨话，听着就很假！”
能不能帮我想一个给力的例子	能不能帮我想一个吃力的例子

“今日头条”文章规范性分析

- 分析方法
  - 使用爬虫爬取今日头条网站上财经、国际、养生三个板块各500篇文章
  - 规范性分数 $= \frac{\text{纠错总数}}{500} \times \frac{\text{抽样正确纠错个数}}{\text{抽样纠错总数}}$

“今日头条”文章规范性分析

中医“镇魂七方”，治

文知　8评论　2小时前

早上起床后有这5个表

光明网　2148评论　1月前

老年痴呆的十大危险信

上观新闻　107评论　1月前

爱出汗就是“虚”？还

海外网　54评论　1月前

睡觉流口水是睡得香？

光明网　1评论　5天前

治疗慢性咳喘的“小药

中医名家祁文强　0评论　3小时前

可可在线 21小时前

美国排出世界大学排名，排名

清华大学，北京大学，复旦大

分享　0　9

钞票也能“中国制造”？我

“印钞大国”

阿森侃文　22评论　2天前

勇担时代责任 发挥引领作

人民网　5评论　14小时前

苹果将支付3000万美元与

时的包包和设备检查

cnBeta　9评论　14小时前

“碳中和+新能源”，化工行业迎全面价值重估——化工行业2022年度投资策略

金融界　96评论　6天前

人民财评：为近30万年轻人拒绝“买买买”点赞

人民网　1332评论　2天前

如何缴纳社保才能保障晚年无忧？企业缴纳社保15年和个人缴纳社保15年，退休金会有什么差别

中国甘肃网　157评论　2月前

直男讲财经 13小时前 · 证券投资顾问 优质财经领域创作者　+关注

最新消息！今天星期日，明天中国股市就要开了，一觉醒来，消息面上很不平静，让人兴奋不已，究竟是怎么回事？在刚刚，证券市场爆出了3个利好消息，快来看看吧，或将影响你持仓的股票，给1.9亿股民简单交待一下：　1、全国首家元宇宙协...

分享　2　144

个税合理避税的12种方法！公开

“今日头条”文章规范性分析

- 分析方法
  - 使用爬虫爬取今日头条网站上财经、国际、养生三个板块各500篇文章
  - 规范性分数 $= \frac{\text{纠错总数}}{500} \times \frac{\text{抽样正确纠错个数}}{\text{抽样纠错总数}}$

	纠错总数	抽样纠错正确/纠错总数	规范性分数（错误/篇）
养生	805	13/39	0.537
财经	573	7/31	0.259
国际	534	4/24	0.178

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。