中文拼写检纠错

张海鹏

2018-11-30

Where to get data(0/3)?

变异文本	正确文本	错误类型	拼音	变异词索引	分词结果
修改**侯**的立法法全文公布。	修改后的立法法全文公布。	1	hou	1	['修改', '后', '的', '立法法', '全文', '公布', '。']
修改后的立法法全文**公办**。	修改后的立法法全文公布。	2	gongb	5	['修改', '后', '的', '立法法', '全文', '公布', '。']
修改后的立法法**请问**公布。	修改后的立法法全文公布。	3	qw	4	['修改', '后', '的', '立法法', '全文', '公布', '。']

Where to get data(1/3)?

错误文本	正确文本0	正确文本1	正确文本2
妈妈在银行工作，**他**今年自己买了**一个公寓房间**	妈妈在银行工作，她今年自己买了一个公寓房间	妈妈在银行工作。她今年自己买了一间公寓。	妈妈在银行工作，她今年自己买了 一间公寓

Where to get data(2/3)?

<DOC>

<TEXT id="12004051095232014 30_2_2x2">

别只能想自己，想你周围的人。还有你，如果你是一个家庭的爸爸，你多想自己的孩子；如果你是青少年你多想自己的未来；那你可以禁烟了。

</TEXT>

<CORRECTION>

别只想自己，要想想你周围的人。还有，如果你是一个家庭的爸爸，你要多想想自己的孩子；如果你是青少年你要多想想自己的未来；那你就可以戒烟了。

</CORRECTION>

<ERROR start_off="3" end_off="3" type="R"></ERROR>

<ERROR start_off="8" end_off="8" type="M"></ERROR>

<ERROR start_off="58" end_off="58" type="M"></ERROR>

<ERROR start_off="60" end_off="60" type="S"></ERROR>

</DOC>

What’re the metrics(0/2)?

- 检错率：输入字和正确字不同时，预测字和输入字不同的频率。
- 纠错率：输入字和正确字不同时，预测字和正确字相同的频率。

What’re the metrics(0/2)?

$$P = \frac{\sum_{i=1}^{n} |e_i \cap g_i|}{\sum_{i=1}^{n} |e_i|} \qquad R = \frac{\sum_{i=1}^{n} |e_i \cap g_i|}{\sum_{i=1}^{n} |g_i|} \qquad F_1 = 2 \times \frac{P \times R}{P + R}$$

Orginal sentence:　随着通迅技术的发达我们的生活也是越来越放便。

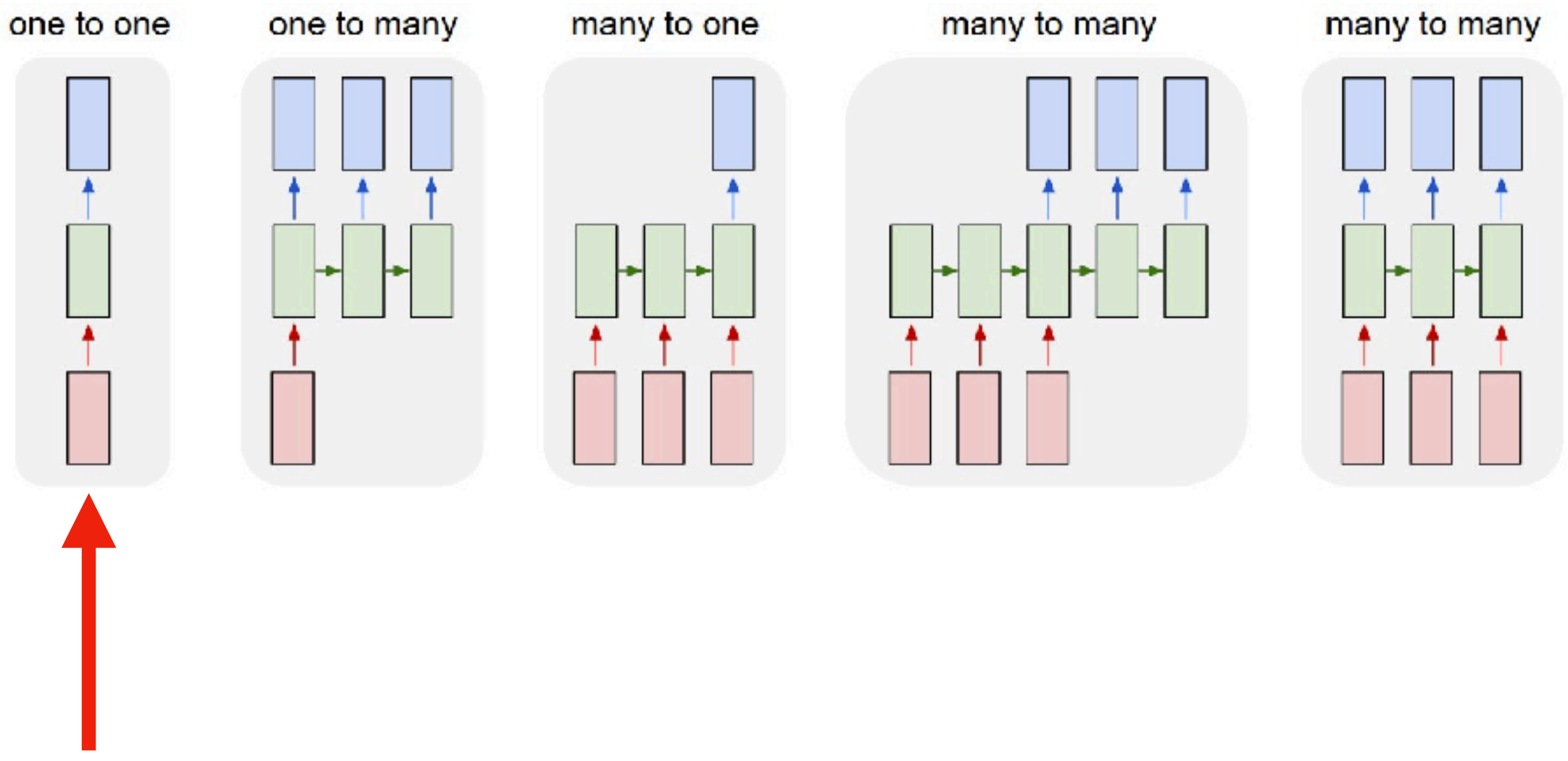
g={通迅 → 通讯, 也 → ∈, 放便 → 方便}

e={通迅 → 通讯, 放便 → 方便}

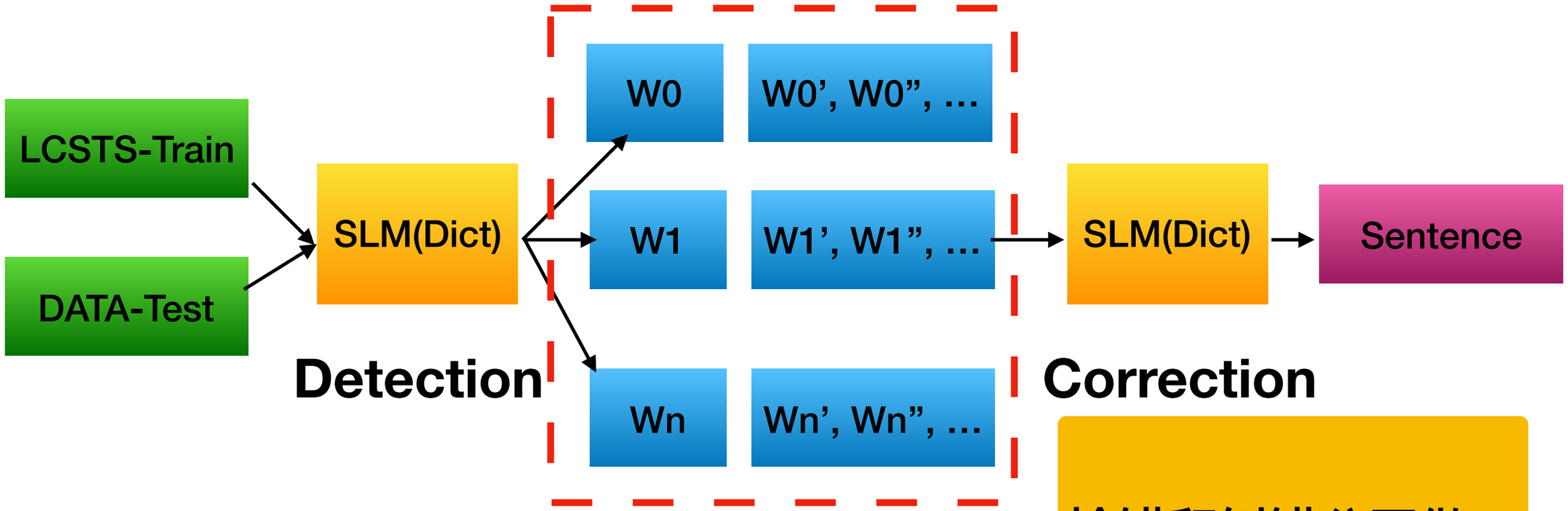
P=2/3, R=2/2=1, F1 =2RP/(R+P)=4/5

MaxMatch

How to model?



Statistical Language Model

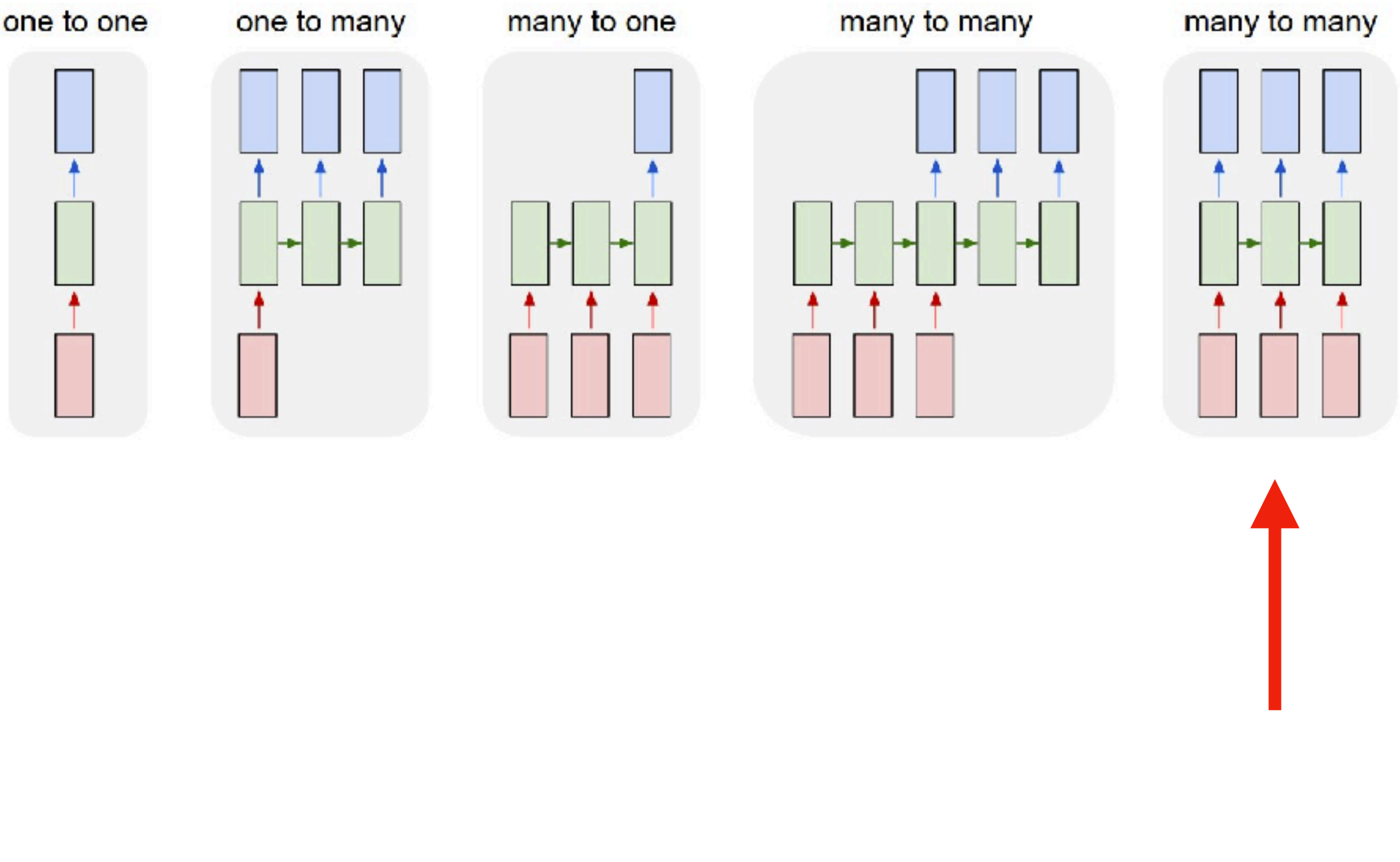


检错和纠错分开做!

模型(SLM)	检错率
包含错字的序列->序列得分(n-gram)	0.30

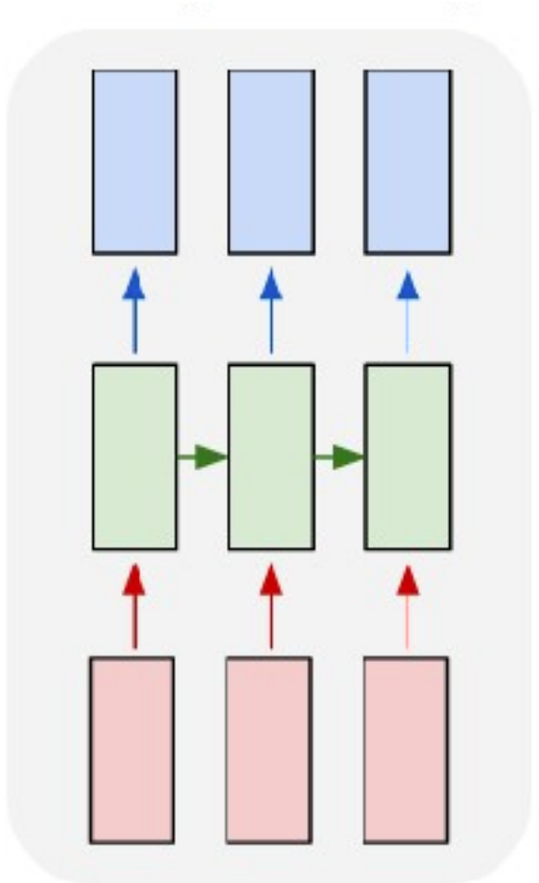
https://github.com/kpu/kenlm

How to model?

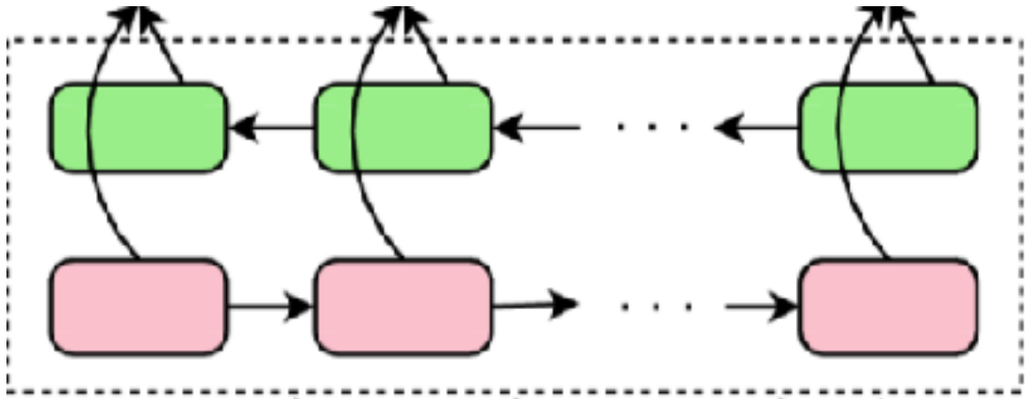


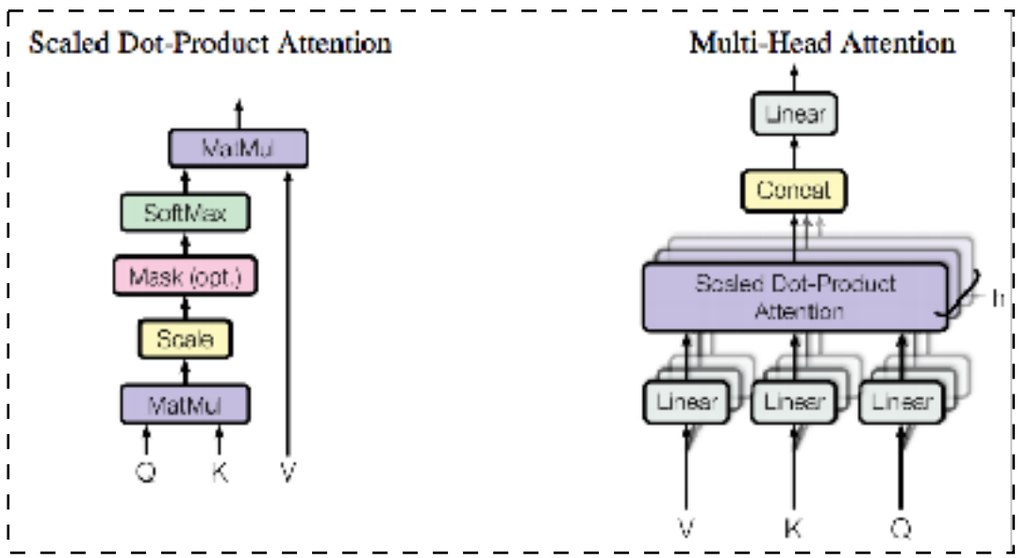
BiLSTM+X

修改后的立法法全文公布。



修改后的立法法全文公布。





修改**侯**的立法法全文公布。

xiugaihoudelifafaquanwengongbu(tone/non-tone)

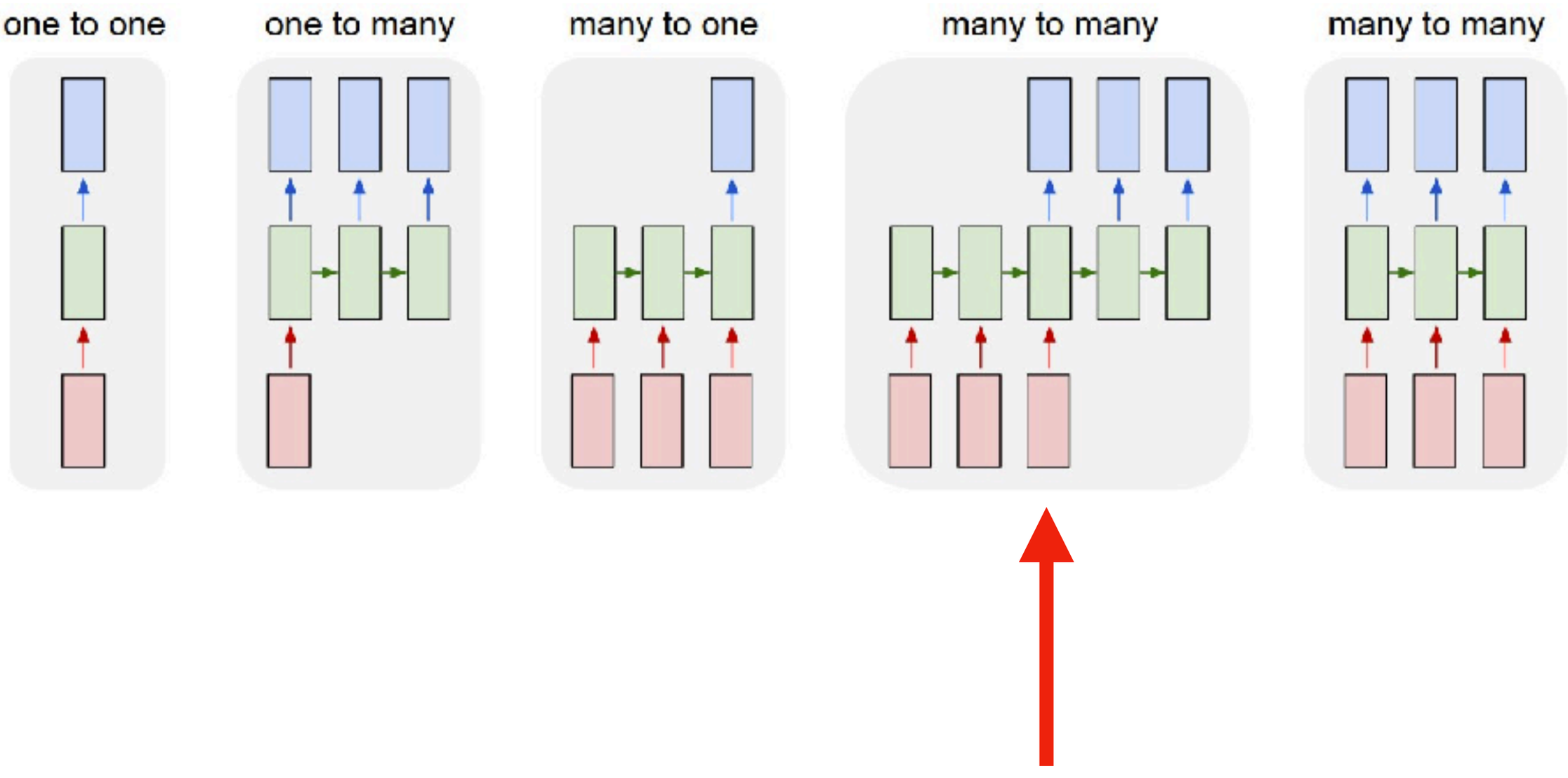
Evaluation

检错和纠错同时做！

模型(BiLSTM)	纠错率	检错率
汉字->汉字	0.12	0.18
汉字+拼音（有音调）->汉字	0.12	0.14
汉字+拼音（有音调+无音调）->汉字	0.11	0.13

模型(Multi Head Self Attention+BiLSTM)	纠错率	检错率
X->汉字	0.50	0.50

How to model?



Evaluation

模型(Transformer)	纠错率	检错率
包含错字的序列->正确序列	0.50**(+0.30)**	0.50**(+0.30)**

汽车**形式**在这条隧道上

中国人工**只能**布局基本划分

想不想**在来**一场辩论赛呢

你不**觉的**高兴吗

权利的游戏第八季什么时候播出

检错和纠错同时做！

https://github.com/tensorflow/tensor2tensor

What else?

- SLM: rethink the n-gram score of a given sentence(http://view.zsxq.com/view/5bfcdcdbed01db2204c896a9)

- Sequence Tagging: low resource task as CGED in EMNLP-IJCNLP 2019

- Seq2Seq: does copy/coverage works?

Takehome

- Rethink “95% of tasks **do not** require deep learning”, statistical language model
- Data v.s. Model
- N-gram/Sequence Tagging/Seq2Seq
- SLM/LSTM/BiLSTM/Transformer, etc.
- Find another **test set** that diff with train/dev/**test**