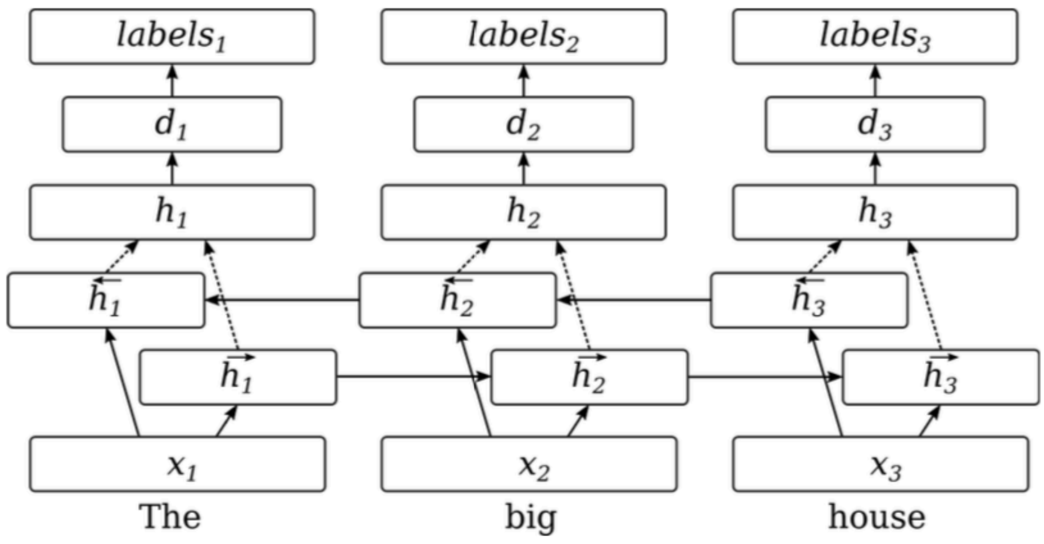
源码：https://github.com/marekrei/sequence-labeler

模型第一部分：bilstm+crf部分



之前bilstm输出的隐层状态输给CRF，这里多了个d隐层

We include an extra narrow hidden layer on top of the LSTM, which proved to be a useful modification based on development experiments. An additional hidden layer allows the model to detect higher-level feature combinations, while constraining it to be small forces it to focus on more generalisable patterns:

$$d_t = tanh(W_d h_t) \tag{2}$$

where W_d is a weight matrix between the layers, and the size of d_t is intentionally kept small.

作者的解释是这样的效果更好，可以捕捉到“更高层”特征且压缩维度，个人猜测是bilstm的输出仍比较稀疏，之前也有实验表明压缩维度有一定效果。

模型第二部分：embedding部分

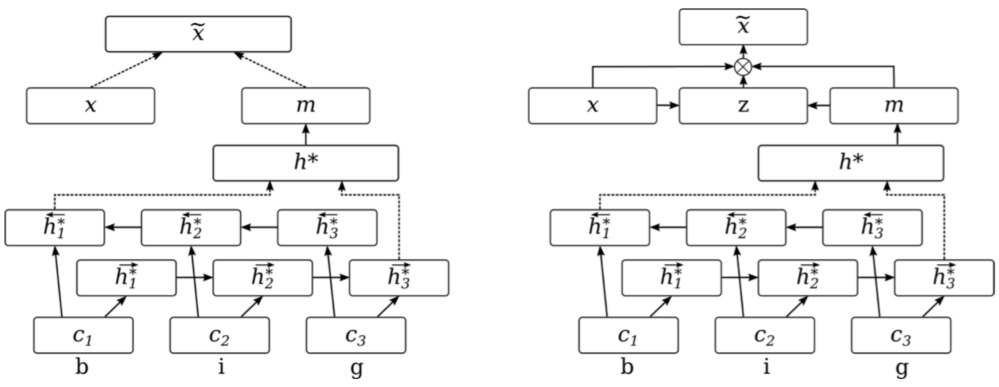


Figure 2: Left: concatenation-based character architecture. Right: attention-based character architecture. The dotted lines indicate vector concatenation.

$$h^* = [\overrightarrow{h_R^*}; \overleftarrow{h_1^*}] \qquad m = tanh(W_m h^*)$$

这里h1和hr进行concat 然后映射到m

attention部分：

$$z = \sigma(W_z^{(3)} tanh(W_z^{(1)} x + W_z^{(2)} m)) \qquad \widetilde{x} = z \cdot x + (1 - z) \cdot m$$

使用的是无交互的attention加权

字符和词向量训练时的交互：

$$\widetilde{E} = E + \sum_{t=1}^{T} g_t(1 - cos(m^{(t)}, x_t)) \qquad g_t = \begin{cases} 0, & \text{if } w_t = OOV \\ 1, & \text{otherwise} \end{cases}$$

E为原来的交叉熵，后面这一项是对于out-of-vocabulary的词来说，通过cos值使得字符向量和词向量更加接近。作者在这里解释对于训练语料未出现的词，词向量的结果还是值得字符向量去接近的，反之则效果不佳

作者的其他解释：
1.优点在于处理OOV词时可以平衡词向量和字符向量的权重，也可以提取部分前后缀特征（只提取前后缀用CNN也可以）
2.参数量少了，相对于concat，attention部分的z维度更小

实验结果：

	CoNLL00		CoNLL03		PTB-POS		FCEPUBLIC	
	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Word-based	91.48	91.23	86.89	79.86	96.29	96.42	46.58	41.24
Char concat	92.57	92.35	89.81	83.37	97.20	97.22	46.44	41.27
Char attention	**92.92**	**92.67**	**89.91**	**84.09**	**97.22**	**97.27**	**47.17**	**41.88**

	BC2GM		CHEMDNER		JNLPBA		GENIA-POS	
	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Word-based	84.07	84.21	78.63	79.74	75.46	70.75	97.55	97.39
Char concat	87.54	87.75	82.80	83.56	76.82	72.24	98.59	98.49
Char attention	**87.98**	**87.99**	**83.75**	**84.53**	**77.38**	**72.70**	**98.67**	**98.60**

Table 2: Comparison of word-based and character-based sequence labeling architectures on 8 datasets. The evaluation measure used for each dataset is specified in Section 6.

16年还可以的效果，另外提到了常用的conll2003数据集有其特殊性，90%的state of art是由于从原来IOB的输出标签扩展到IOBES，使用了考虑词序的嵌入方式