LSTM：像RNN、LSTM、BILSTM这些模型，它们在序列建模上很强大，它们能够capture长远的上下文信息，此外还具备神经网络拟合非线性的能力，这些都是CRF无法超越的地方，对于t时刻来说，输出层y_t受到隐层h_t（包含上下文信息）和输入层x_t（当前的输入）的影响，但是y_t和其他时刻的$y_{t'}$是相互独立的，感觉像是一种point wise，对当前t时刻来说，我们希望找到一个概率最大的y_t，但其他时刻的$y_{t'}$对当前y_t没有影响，如果y_t之间存在较强的依赖关系的话（例如，形容词后面一般接名词，存在一定的约束），LSTM无法对这些约束进行建模，LSTM模型的性能将受到限制。ACL2016的一篇best paper《Harnessing Deep Neural Networks with Logic Rules》中提出了一种将规则融入到神网中的模型，我们可以将一些人工的先验知识传输给这个网络，这是知识通过逻辑谓词地形式进行表示，基于Teacher-Student网络进行训练，可以较好地解决此类约束问题。

CRF：它不像LSTM等模型，能够考虑长远的上下文信息，它更多考虑的是整个句子的局部特征的线性加权组合（通过特征模版去扫描整个句子）。关键的一点是，CRF的模型为p(y | x, w)，注意这里y和x都是序列，它有点像list wise，优化的是一个序列y = (y1, y2, …, yn)，而不是某个时刻的y_t，即找到一个概率最高的序列y = (y1, y2, …, yn)使得p(y1, y2, …, yn| x, w)最高，它计算的是一种联合概率，优化的是整个序列（最终目标），而不是将每个时刻的最优拼接起来，在这一点上CRF要优于LSTM。

HMM：CRF不管是在实践还是理论上都要优于HMM，HMM模型的参数主要是“初始的状态分布”，“状态到状态的概率转移矩阵”，“状态到观测的概率转移矩阵”，这些信息在CRF中都可以有，例如：在特征模版中考虑$h(y_1), f(y_{i-1}, y_i), g(y_i, x_i)$等特征。

CRF与LSTM：从数据规模来说，在数据规模较小时，CRF的试验效果要略优于BILSTM，当数据规模较大时，BILSTM的效果可能会超过CRF。从场景来说，如果需要识别的任务不需要太依赖长久的信息，此时RNN等模型只会增加额外的复杂度，此时可以考虑类似科大讯飞FSMN（一种基于窗口考虑上下文信息的“前馈”网络）。

CNN + BILSTM + CRF：这是目前学术界比较流行的做法，BILSTM + CRF是为了结合以上两个模型的优点，CNN主要是处理英文的情况，英文单词是由更细粒度的字母组成，这些字母潜藏着一些特征（例如：前缀后缀特征），通过CNN的卷积操作提取这些特征，在中文中可能并不适用（中文单字无法分解，除非是基于分词后），这里简单举一个例子，例如词性标注场景，单词football与basketball被标为名词的概率较高， 这里后缀ball就是类似这种特征。