1 Related Work

1.1 Chinese Spelling Error Correction

该文中提出了一种spelling error correction的方式，可以对于文章进行字符级别的查错与纠正。该文章的目标是将输入语句$X = (x_1, x_2, ..., x_n)$进行纠错之后输出为$Y = (y_1, y_2, ...y_n)$，其中x_i为语句中的第i个字符，y_i为x_i对应的修改后的字符。该系统主要由Detection Network和Correction Network两个部分构成，其中的Detection Network对于本项工作有一些帮助。

Detection Network以sequence of embedding $E = (e_1, e_2, ..., e_n)$作为输入，其中e_i是x_i的embedding，更具体的说，是word embedding，position embedding 和 segment embedding的和。该模型输出$G = (g_1, g_2, ..., gn)$，其中g_i表示第i个字符是否正确，0表示字符正确，1表示错误。按照上述的定义，detection network采用了Bi-GRU，定义每个字符错误的概率p_i

$$p_i = P_d(g_i = 1|X) = \sigma(W_d h_i^d + b_d) \tag{1.1}$$

h_i^d是Bi-GRU的hidden state，定义为：

$$\overrightarrow{h}_i^d = GRU(\overrightarrow{h}_{i-1}^d, e_i) \tag{1.2}$$

$$\overleftarrow{h}_i^d = GRU(\overleftarrow{h}_{i+1}^d, e_i) \tag{1.3}$$

$$h_i^d = \left[\overrightarrow{h}_i^d, \overleftarrow{h}_i^d\right] \tag{1.4}$$

将embedding 和 soft embedding的加权和作为soft-mask embedding：

$$e_i^{'} = pi * e_{mask} + (1 - p_i) * e_i \tag{1.5}$$

该模型利用已经标注好的数据进行训练，即(X_i, Y_i)这样的数据对。对于我们的工作，不规范语句的标注而言，这样的训练对比较难以获得，因为规范的定义比较抽象，我们对于是否符合规范的语句不好进行定义。但是这篇文章的工作提供了一个规范化的一个方面，也是比较重要的一个方面。

可以尝试对定义规范化的一些规则，例如人称问题、seplling问题等。对于逐条规则进行check，然后对于不符合规则的进行highlight。