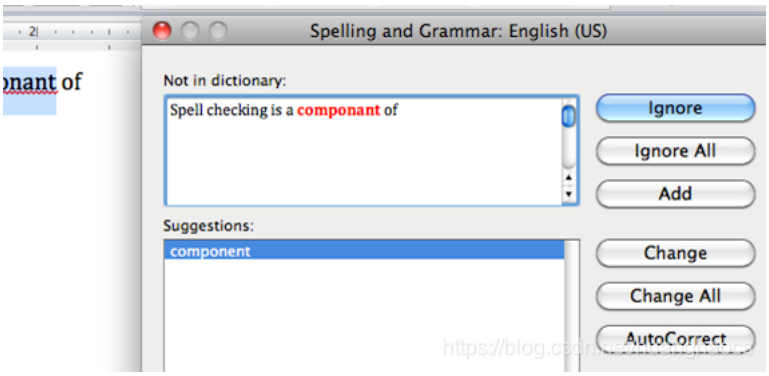
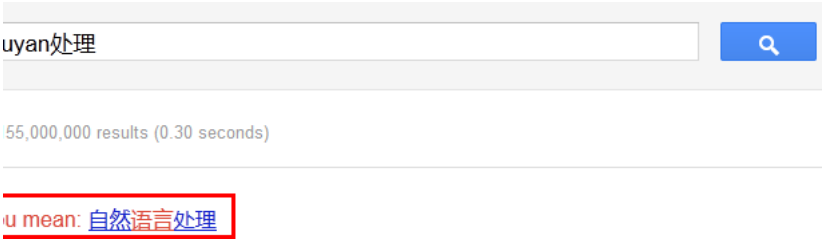
错（Spelling Correction）

章标签： NLP query rewrite query 改写 拼写纠错

弓纠错的一节，感觉讲的比较好，虽然课程很老，但是讲的知识，在目前的query改写、拼写纠错还是很

，又称拼写检查（Spelling Checker），往往被用于字处理软件、输入法和搜索引擎中，如下所示：





务：

误类型不同，分为Non-word Errors和Real-word Errors。前者指那些拼写错误后的词本身就不合法，如

者指那些拼写错误后的词仍然是合法的情况，如将“there”错误拼写为“three”（形近），将“peace”错误拼

误拼写为“too”（同音）。

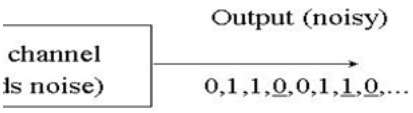
错，如把“hte”自动校正为“the”，或者给出一个最可能的拼写建议，甚至一个拼写

被词典所包含的word均被当作spelling error，识别准确率依赖词典的规模和质量。
典中与error最近似的word，常见的方法有Shortest weighted edit distance和Highest noisy channel

rd都作为spelling error candidate。
和拼写等角度，查找与word最近似的words集合作为拼写建议，常见的方法有Highest noisy channel

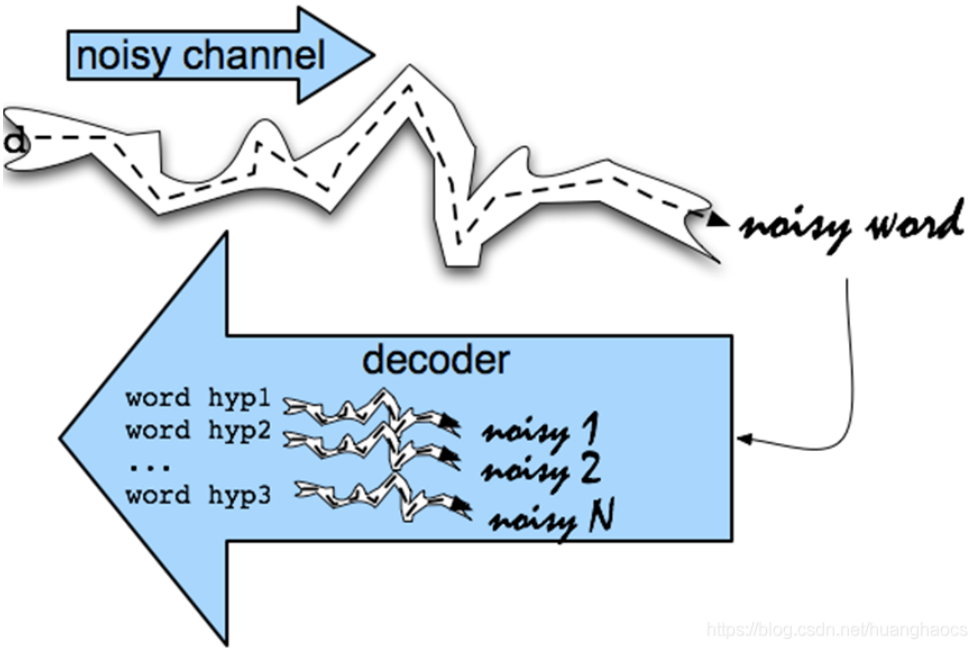
odel的拼写纠错

模型，或称信源信道模型，这是一个普适性的模型，被用于语音识别、拼写纠错、机器翻译、中文分
立用领域。其形式很简单，如下图所示：



号恢复输入信号，形式化定义为：

$$\frac{P(O|I)P(I)}{P(O)} = \arg\max_I P(O|I)P(I)$$



作original word通过noisy channel转换得到，现在已知noisy word（用x表示）如何求得最大可能的
如下：

$$\cdot \mid x)$$

$$\frac{c \mid w)P(w)}{P(x)}$$ 贝叶斯公式

$$c \mid w)P(w)$$ 对于特定的x，P(x)不变

率，二者可以基于训练语料库建立语言模型和转移矩

error correction的例子加以解释：
词典匹配容易确定为“Non-word spelling error”；然后通过计算最小编辑距离获取最相似的candidate
这里的最小编辑距离涉及四种操作：

f two adjacent letters

andidate orrection	Correct Letter	Error Letter	Type
ctress	t	–	deletion
ress	–	a	insertion
aress	ca	ac	transposition
ccess	c	r	substitution
cross	o	e	substitution
cres	–	s	insertion
cres	–	s	insertion

为1，几乎所有的拼写错误编辑距离小于等于2，基于此，可以减少大量不必要的计算。

建议候选集（candidate w），此时，我们希望选择概率最大的w作为最终的拼写建议，基于噪声信道模型
|w)。

可以很容易建立语言模型，即可得到P(w)，如下表所示，计算Unigram Prior Probability（word总数：

cy of word	P(word)
9, 321	. 0000230573
220	. 0000005442
686	. 0000016969
37, 038	. 0000916207
120, 844	. 0002989314
12, 874	. 0000318463

ed word x=x1 x2 x3 … xm, correct word w=w1 w2 w3 … wn>pair计算del、ins、sub和trans四种转移矩

$$
\begin{aligned}
&\text{count}(\text{xy typed as x})\\
&\text{count}(\text{x typed as xy})\\
&\text{count}(\text{x typed as y})\\
:\quad &\text{count}(\text{xy typed as yx})
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\text{del}[w_{i-1},w_i]}{\text{count}[w_{i-1}w_i]}, && \text{if deletion}\\
&\frac{\text{ins}[w_{i-1},x_i]}{\text{count}[w_{i-1}]}, && \text{if insertion}\\
&\frac{\text{sub}[x_i,w_i]}{\text{count}[w_i]}, && \text{if substitution}\\
&\frac{\text{trans}[w_i,w_{i+1}]}{\text{count}[w_iw_{i+1}]}, && \text{if transposition}
\end{aligned}
$$

orrect etter	Error Letter	x|w	P(x|word)
	–	c|ct	.000117
	a	a|#	.00000144
a	ac	ac|ca	.00000164
	r	r|c	.000000209
	e	e|o	.0000093
	s	es|e	.0000321
	s	ss|s	.0000342

Error Letter	x|w	P(x|word)	P(word)	10^9 *P(x|w)P(w)
–	c|ct	.000117	.0000231	2.7
a	a|#	.00000144	.000000544	.00078
ac	ac|ca	.00000164	.00000170	.0028
r	r|c	.000000209	.0000916	.019
e	e|o	.0000093	.000299	2.8
s	es|e	.0000321	.0000318	1.0
s	ss|s	.0000342	.0000318	1.0

上更大。
am，也可以推广到bigram，甚至更高阶，以较好的融入上下文信息。
ess whose combination of sass and glamour…”，计算bigram为：
/hose|actress) = .0010
/hose|across) = .000006

00021*.0010 = 210 x10-10
00021*.000006 = 1 x10-10

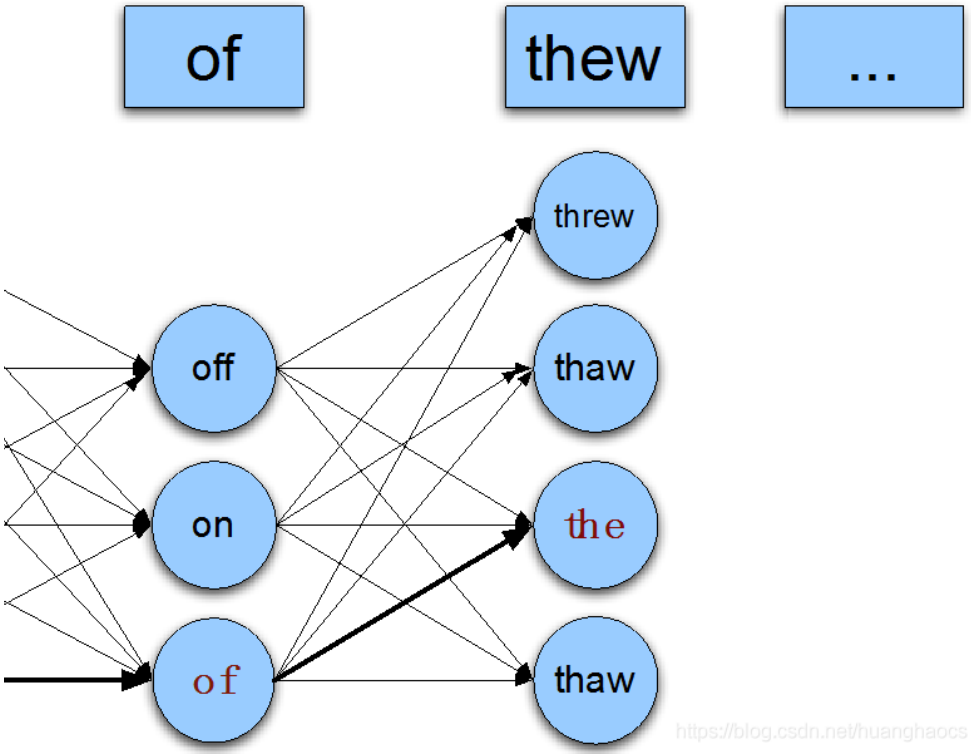
的拼写错误都属于Real-word类型，与Non-word类型相比，纠错难度更大，因为句子中的每个word都被
分两步：

rd in sentence
ndidate set
itself
-letter edits that are English words
at are homophones
candidates
annel model
cific classifier

vn，为每个wi生成candidate set，如下：
w'"1 ,…}
w'"2 ,…}

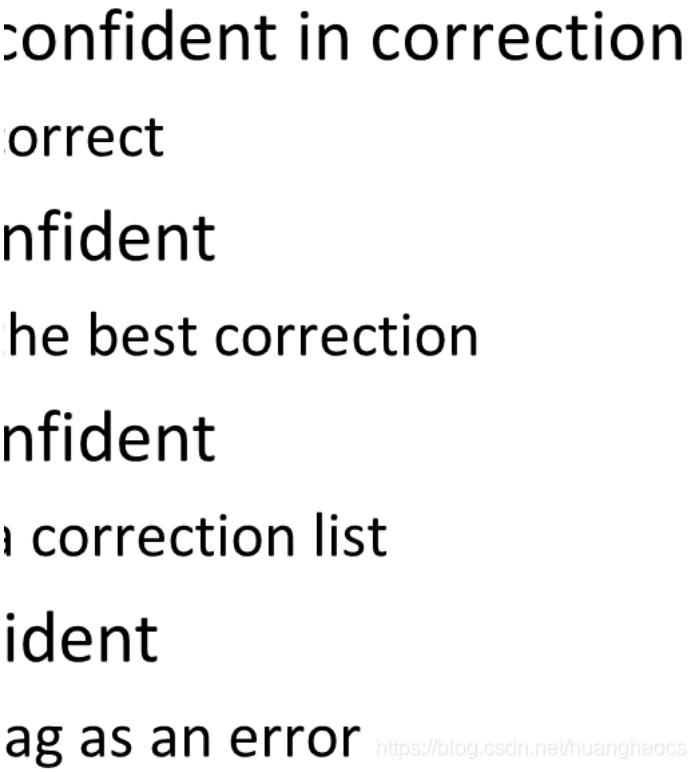
w'"n ,…}

自动纠错后的句子，与中文分词、音字转换等应用相同，可以表示成词网格形式，转化为HMM的解码过



子中最多有一个word存在splling error（事实上，所出现的情况也的确如此）。

口下HCI（Human Computer Interface）准则：

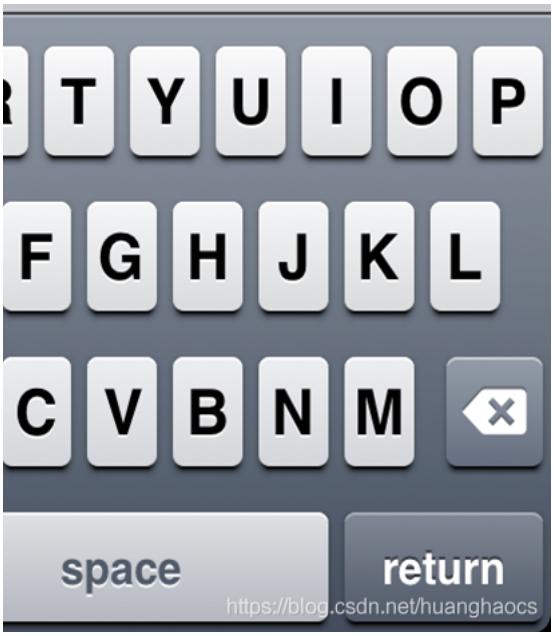


sitivity），需要对语言模型进行特别的处理，如：

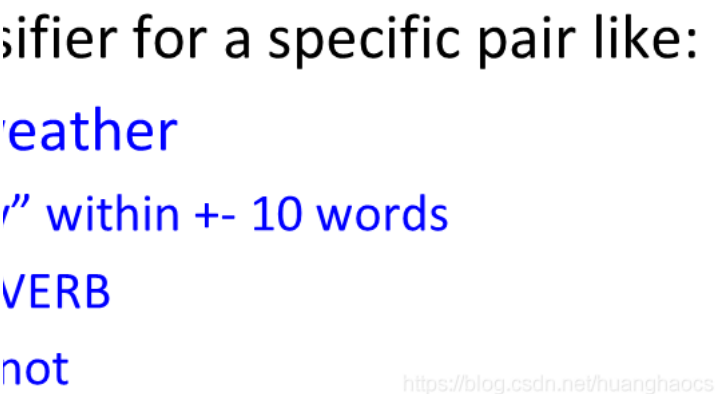
$x \mid w)P(w)^{\lambda}$ Learn λ from a development test set

能同音导致，所以，有些系统将“error model”转化为“l

引入spelling error pair，据此可以对转移矩阵进行加权。



分类问题，通过构建训练语料库，抽取features，训练分类模型，预测新实例等一系列过程解决，如



haimizhao的白板 2073
y-Document(摘要)看成源语言和目标语言，用翻译模型计算二者的短语与短语之间的对齐关系，扩展Query的同时起…

拼写纠错（Spelling Correction）” overstack的专栏 3168
在Coursera启动了在线自然语言处理课程，由NLP领域大牛Dan Jurafsky 和 Chirs Manning教授授课： https://class…

抢沙发
高权重 评论

weixin_33835103的博客 851
题还是在于用户query和商品描述之间存在GAP，特别是中长尾query。把问题分成以下几种类型： 多种描述：划痕…

mingo220的博客 450

错技术和架构_chengyong... 12-25
成的query纠错相对更加精准,每个词存在上下文关系约束,整个query的意图更加明确。通过对query分词,查找每个词…

(NMT)的语法纠错算法_怎... 1-4
rrection,下同)是NLP的一个子领域,在搜索query纠错、语音纠错、舆情文本纠错等诸多直接与普通大众交互的场景得…

错技术和架构
术和架构 1 背景 如今，搜索引擎是人们的获取信息最重要(

点赞Mark关注该博主, 随时了解TA的最新博文

点赞 评论 分享 收藏 举报 关注 一键三连