Bilingual Expert

Binlingual Expert：

模型结构：

• Multi-head self-attention mechanism and transformer NN to build LM, used as feature extractor

• Extracted features combined with human-crated features, and fed into a Bi-LSTM predictive model

• Greedy ensemble selection method to decrease individual model errors and increase model diversity

训练过程： 将模型 在 平行语料上进行训练， src 输入到transformer的encoder中。 t 输入到bi-transformer 的encoder，但双向 会把right or left方向的 mask掉。当具体地预测一个token的时候，正向的，会把 token之前的

token以及src的信息join，反向的会把token之后的以及src的信息join,这样就得到要预测token的前后信息特征，就会预测出来。

抽取mis-mature的feature 包含 强制解码为当前词的概率，最大概率值，两个概率值差，以及是否是一致的

导出模型特征（z,e）

连接两个特征输入到bilstm 模型，进行训练

Destination:

构建一个双语专家模型，在没有参考译文的情况下，进行对machine translation进行评估。主要针对 句子级别 计算hter的值，以及进行单词的标注。

Model:

• Multi-head self-attention mechanism and transformer NN to build LM, used as feature extractor

• Extracted features combined with human-crated features, and fed into a Bi-LSTM predictive model

• Greedy ensemble selection method to decrease individual model errors and increase model diversity

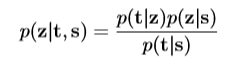
den 
emb 
press 
drück 
the 
auf 
button 
knopf 
Joint Attention 
Encoder 
Self Attention 
machine translation 
Categorial distribution for the 3"' token 
prediction from rnodel prior 
machine translation output 
mis•matching gap 
mis•matching 
features 
Bi-LSTM Quality Estimator 
model 
derived 
press 
the 
button 
Token Reconstruction 
drock knopf 
88888 
driick 
Forward 
Self Attention 
kno 
drock den 
Backward 
Self Attention 
knopf 
features (t,e) 
Bilingual Expert Model 

**训练过程：**

在平行语料（s,t）

总体目标函数：

双语专家模型：需要导出模型features, z,e:深层语义 ,需要包含src以及t的信息



p(z|t,s)表示机器翻译的一个latent variable的一个分布

使用机器学习中的变分推断 方法，用q近似表示z真实的后验分布

q(z|t,s)表示的是平行语料的latent variable的一个分布

C:\Users\Xida\AppData\Local\Temp\msohtmlclip1\02\clip_image008.png

C:\Users\Xida\AppData\Local\Temp\msohtmlclip1\02\clip_image009.png

p(tli), q(zlt, s) 

**VAE 方法** ==>变分自编码

p(z|s) 标准的高斯分布

变分推理--->VAE 是连在一起的

Let the prior over tlw latent variables be the centered isotropic multivariate Gaussian pg(z) 
N (z; 0, I). Note that in this case, the prior lacks parameters. We let p9(xlz) be a multivariate 
Gaussian (in case of real-valued data) or Bernoulli (in case of binal)' data) whose distribution pa- 
rameters are computed from z with a MLP (a fully-connected neural network with a single hidden 
layer, see O). Note the true posterior pg(zlx) is in this case intractable. While there is 
much freedom in thenorm qé(zlx), we'll assume the true (but intractable) posterior takes on a ap- 
proximate Gaussian form with an approximately diagonal covariance. In this case, we can let the 
variational approximate posterior be a multivariate Gaussian with a diagonal covariance struc 
log = logy (z; ILO), 

Mlp:

C.2 Gaussian MLP as encoder or decoder 
In this case let encoder or decoder be a multivariate Gaussian with a diagonal covariance structure: 
logp(xlz) = log N (x; 'L, ff21) 
log + = Wöh+ b5 
h = tanh(W3z + b3) 
(12) 
where {W 3, W , Wö, b3, b, , bö} ale the weights and biases of tlw MLP and part of when used 
as decoder. Note that when this network is used as an encoder q.(zlx), then z and x are swamrd, 
and the weights and biases are variational parameters 4. 

**模型导出的特征** ： z 包含source 的信息 以及 target 的context

C:\Users\Xida\AppData\Local\Temp\msohtmlclip1\02\clip_image013.png

深层语义：

· 前 一 个 词 的 词 向 量 一 1 
后 一 个 词 的 词 向 量 etk+l 

**mis-matching 特征：**

Categorial distribution for 3rd token 
im prediction from model prior 
mk machine translation by error 
mis-matching gap 
ein einen einu 
Bilingual Expert Model 
press the button 
source 
drück auf knopf 
machine translation 

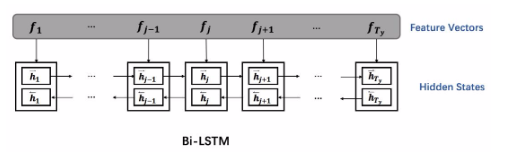
吕 标 端 强 制 解 码 为 当 前 词 的 概 率 信 息 
概 率 最 高 词 谌 的 概 率 信 息 
· 强 制 解 码 为 当 前 词 与 解 码 为 概 率 最 高 词 的 概 率 信 愿 差 异 一 
当 前 词 与 预 测 词 是 否 一 致 m 唁 “ 

抽取特征：

(lk,mk' k tk 
k, mk 

3: Extract features fk = Concat(z , k, for 
QE training data (s, m). 

**双向LSTM:**



level score HTER prediction can be formulated as a regres- 
sion problem (9), and the word error prediction is a sequence 
labeling problem (10), 
h1:T, = 
(8) 
2 
arg min llh — sigmoid (wT[hT, Il 
(9) 
2 
arg min EXENT(Yk, W[hk, FKI) 
(10) 

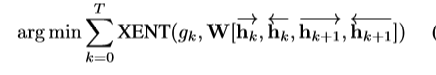
xent为交叉熵

yk为ok or bad

一起训练

Algorithm 1 Translation Quality Estimation with Bi- 
Transformer and Bi-LSTM 
Require: QE training data (s, m, t, h, QE inference data 
2: 
3: 
4: 
5: 
(s, m), and parallel corpus (s, 
Combine the parallel corpus with 10 copies Of QE training 
parallel corpus C = (so, U 10 X (Srn, 
Pre-train bilingual expert model via the bidirectional trans- 
former on the combined corpus C. 
Extract features fk = Concat(a h, etk_l , , warn) for 
QE training data (s, m). 
Train Bi-ISTM model via objectives (9)(10). 
return Predict h, y for QE inference data 

Gap prediction:



m. Note that we need to define a token for gap 
prediction, meaning that nothing needs to be inserted. There- 
fore, it also results in a side product — automatic post-editing. 

The ﬁrst expectation term in(3) can be readily considered as a conditional auto-encoder system if we use one sample Monte Carlo integration during optimization

p(t|z)--> a conditional auto-encoder system

p(z|s)--> standard Gaussian distribution, playing as a model regularization for latent variables 如果我们实际将先验p(z|s)设为标准高斯分布，则可作为潜在变量的模型正则化