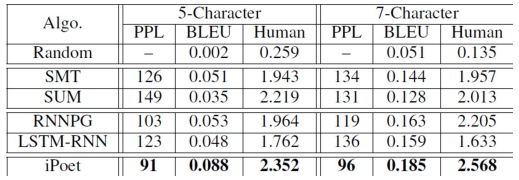
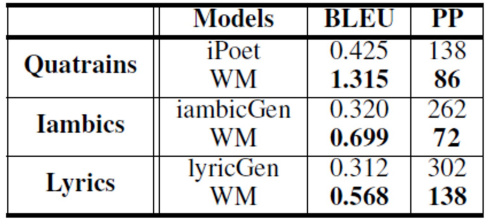
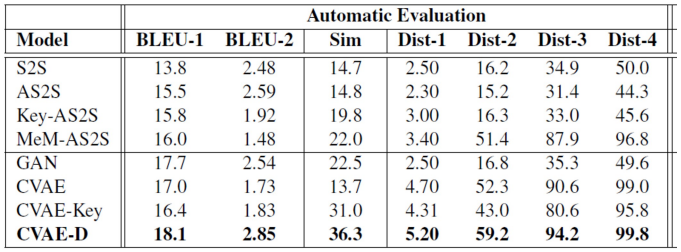
baselines

2019年3月13日 14:29

		baselines	evaluations	automatic evaluations
planning	baidu 16 COLING			
polishing (I, Poet)	Yan 16		Perplexity BLEU human	
memory	Zhang 17 ACL	None	(all human) Compliance Fluency Theme Aesthetic Scenario Consistency Innovation Consistency	
working memory	Sun 18 EMNLP	Planning Polishing Memory human	BLEU(仅在研究memory slot使用到) (human:) Fluency Meaning Coherence Relevance Aesthetics	
CVAE	Yan 18 EMNLP	S2S, AS2S, Key+AS2S(2016 Qixin), Memory (Jiyuan)	BLEU-1 BLEU-2 (unigrams and bigrams (p < 0.01)) Similarity (thematic consistenccy, cosin) Distinctness-n (差异性。不同ngram的比例 n = 1 to 4) (human:) Consistency Fluency Meaning Poeticness	

Perplexity (PPL). For most of the language generation research, language perplexity is a sanity check. Our first set of experiments involved intrinsic evaluation of the "perplexity" evaluation for the generated poems. Perplexity is actually an entropy based evaluation. In this sense, the lower perplexity for the poems generated, the better performance in purity for the generations, and the poems are likely to be good ones.

Algo.	5-Character			7-Character		
	PPL	BLEU	Human	PPL	BLEU	Human
Random	–	0.002	0.259	–	0.051	0.135
SMT	126	0.051	1.943	134	0.144	1.957
SUM	149	0.035	2.219	131	0.128	2.013
RNNPG	103	0.053	1.964	119	0.163	2.205
LSTM-RNN	123	0.048	1.762	136	0.159	1.633
iPoet	**91**	**0.088**	**2.352**	**96**	**0.185**	**2.568**

	Models	BLEU	PP
Quatrains	iPoet	0.425	138
	WM	**1.315**	**86**
Iambics	iambicGen	0.320	262
	WM	**0.699**	**72**
Lyrics	lyricGen	0.312	302
	WM	**0.568**	**138**

	Automatic Evaluation						
Model	**BLEU-1**	**BLEU-2**	**Sim**	**Dist-1**	**Dist-2**	**Dist-3**	**Dist-4**
S2S	13.8	2.48	14.7	2.50	16.2	34.9	50.0
AS2S	15.5	2.59	14.8	2.30	15.2	31.4	44.3
Key-AS2S	15.8	1.92	19.8	3.00	16.3	33.0	45.6
MeM-AS2S	16.0	1.48	22.0	3.40	51.4	87.9	96.8
GAN	17.7	2.54	22.5	2.50	16.8	35.3	49.6
CVAE	17.0	1.73	13.7	4.70	52.3	90.6	99.0
CVAE-Key	16.4	1.83	31.0	4.31	43.0	80.6	95.8
CVAE-D	**18.1**	**2.85**	**36.3**	**5.20**	**59.2**	**94.2**	**99.8**

measure character diversity by calculating the proportion of distinctive [1,4]-grams12 in the generated poems, where final distinctness values are normalized to [0,100].

<table>
<tr><td></td><td></td><td></td><td>Ovrall</td><td></td></tr>
<tr><td>mutual RL</td><td>Sun
18 EMNLP</td><td>Base
Memory
huamn</td><td>(Auto:)
reward score
diversity and innovation
ty-idf distribution
topic distribution
(hunam:)
Fluency
Coherence
Meaning
Overall Quality</td><td><table><tr><th>Models</th><th>Bigram Ratio</th><th>Jaccard</th></tr><tr><td>Base</td><td>0.126</td><td>0.214</td></tr><tr><td>Mem</td><td>0.184</td><td>0.183</td></tr><tr><td>MRL</td><td>0.181</td><td>0.066</td></tr><tr><td>GT</td><td>0.218</td><td>0.006</td></tr><tr><td>SRL</td><td>0.133</td><td>0.146</td></tr><tr><td>LMRL</td><td>0.178</td><td>0.085</td></tr><tr><td>GMRL</td><td>0.186</td><td>0.075</td></tr><tr><td>MRL</td><td>0.181</td><td>0.066</td></tr></table></td></tr>
<tr><td>Salient-clue</td><td>Sun
18 CoNLL</td><td>planning
polishing
seq2seqPG(2017 Sun)
human</td><td>BLEU
(human:)</td><td><table><tr><th></th><th>Models</th><th>Wujue</th><th>Qijue</th></tr><tr><td rowspan="4">Different Models</td><td>Planning</td><td>0.460</td><td>0.554</td></tr><tr><td>iPoet</td><td>0.502</td><td>0.591</td></tr><tr><td>seq2seqPG</td><td>0.466</td><td>0.620</td></tr><tr><td>SC</td><td>0.532</td><td>0.669</td></tr><tr><td rowspan="5">Different Strategies of SC</td><td>naive-TopK-SDU</td><td>0.442</td><td>0.608</td></tr><tr><td>naive-SSal-SDU</td><td>0.471</td><td>0.610</td></tr><tr><td>tfidf-SSal-SDU</td><td>0.533</td><td>0.648</td></tr><tr><td>tfidf-SSal-SSI</td><td>0.530</td><td>0.667</td></tr><tr><td>tfidf-SSal-SSI-intent</td><td>0.532</td><td>0.669</td></tr></table>Table 1: BLEU evaluation results. The scores are calculated by the multi-bleu.perl script.</td></tr>
<tr><td>Unsupervised</td><td>Sun
18 EMNLP</td><td>seq2seq
polishing
memory
human</td><td>(all human)
Fluency
Coherence
Meaningfulness
Poeticness</td><td></td></tr>
<tr><td>CVAE + hybrid dec</td><td>Waterloo
18 IJCAI</td><td>AS2S</td><td>PPL
NNL(KL)
BLEU-1 BLEU-2 BLEU-3 BLEU-4
RES (Rhythm Score Evaluation)</td><td><table><tr><th>Approach</th><th>PPL</th><th>NNL(KL)</th><th>BLEU-1</th><th>BLEU-2</th><th>BLEU-3</th><th>BLEU-4</th><th>RES</th></tr><tr><td>Ground truth</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>0.8975</td></tr><tr><td>AS2S</td><td>55.5255</td><td>4.0168(-)</td><td>0.4296</td><td>0.3596</td><td>0.3045</td><td>0.2640</td><td>0.5046</td></tr><tr><td>AS2S+AW2V</td><td>54.1752</td><td>3.9922(-)</td><td>0.4319</td><td>0.3625</td><td>0.3076</td><td>0.2669</td><td>0.5076</td></tr><tr><td>CVAE</td><td>52.1154</td><td>3.9535(0.0083)</td><td>0.4366</td><td>0.3605</td><td>0.3046</td><td>0.2637</td><td>0.5137</td></tr><tr><td>CVAE+AW2V</td><td>52.0516</td><td>3.9522(0.0126)</td><td>0.4382</td><td>0.3630</td><td>0.3073</td><td>0.2663</td><td>0.5149</td></tr><tr><td>CVAE-HD+AW2V</td><td>51.7236</td><td>3.9459(0.0163)</td><td>0.4405</td><td>0.3644</td><td>0.3084</td><td>0.2672</td><td>0.5245</td></tr></table></td></tr>
</table>

multi-bleu.perl 脚本计算bleu