The origin of the generated conversational model

张璐
2019.09.08

Data-Driven Response Generation in Social Media

Goal

- Train a **phrased-based statistical machine translation model** to generate response

Data

- 1.3 million conversations scraped from Twitter

- First two utterances from each conversation

Model

Response Generation as Translation

- Ensure r is an appropriate response to s → $p(s|r)$, $p(r|s)$
- Ensure r is a well-formed response → n-gram language model

Challenge

Lexical Repetition

- Penalize lexical similarity with Jaccard similarity

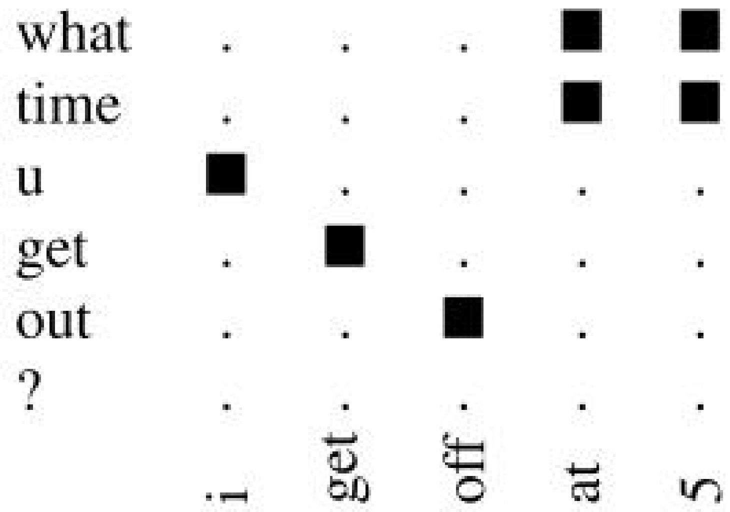
$$\phi_{\text{lex}}(s,t) = J(s,t)$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Challenge

Word Alignment

- Fi lter out
  co
- p–



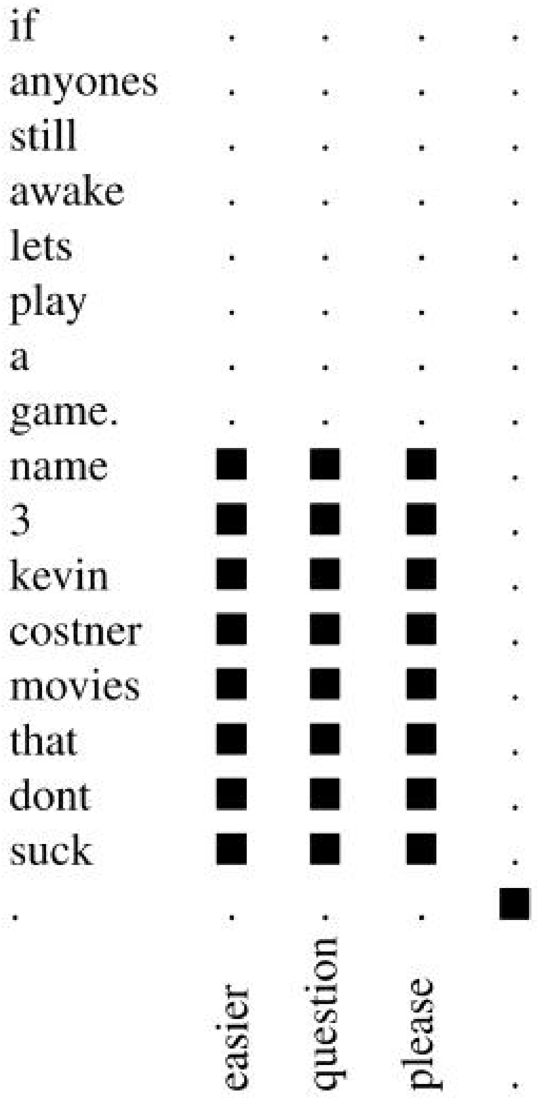
	i	get	off	at	5
what	.	.	.	■	■
time	.	.	.	■	■
u	■	.	.	.	.
get	.	■	.	.	.
out	.	.	■	.	.
?	.	.	.	.	.

C

C(

(

$\overline{N!C(}$



	easier	question	please	.
if	.	.	.	.
anyones	.	.	.	.
still	.	.	.	.
awake	.	.	.	.
lets	.	.	.	.
play	.	.	.	.
a	.	.	.	.
game.	.	.	.	.
name	■	■	■	.
3	■	■	■	.
kevin	■	■	■	.
costner	■	■	■	.
movies	■	■	■	.
that	■	■	■	.
dont	■	■	■	.
suck	■	■	■	.
.	.	.	.	■

$$\frac{\neg t)!}{C(\neg s,\neg t)!}$$

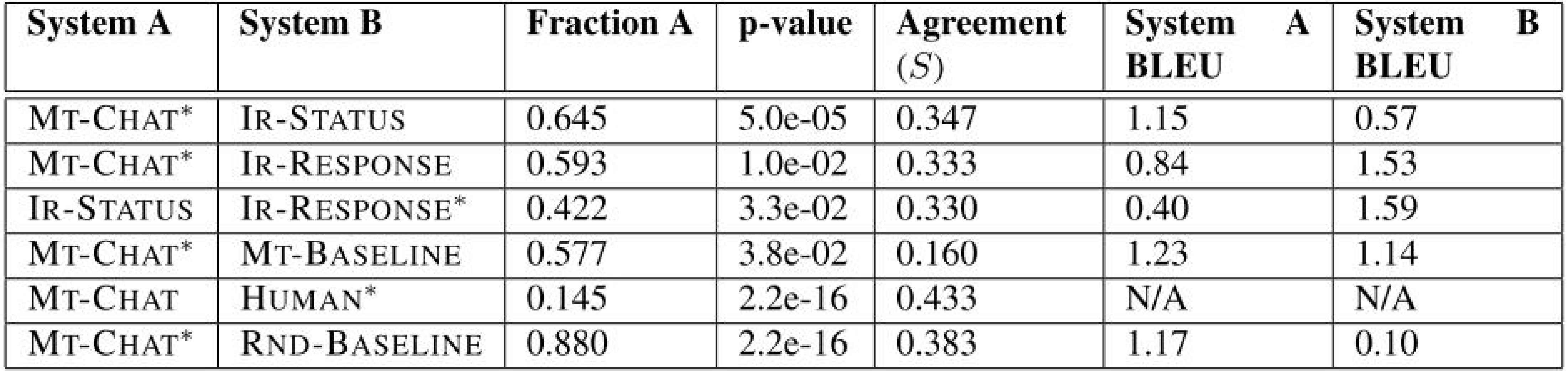
Experiments

- Phase-table: 5M phrases with highest association according the Fisher Exact Test statistic
- Language model: 1.3M responses from the training data, along with roughly 1M replies collected using Twitter’ s streaming API
- Use the default feature weights provided by Moses

Result

- Human evaluation
- Test set: 200 tweets for each experiment
- Pairwise comparisons
- An appropriate response: on the same topic, make sense

Result



System A	System B	Fraction A	p-value	Agreement (S)	System A BLEU	System B BLEU
MT-CHAT*	IR-STATUS	0.645	5.0e-05	0.347	1.15	0.57
MT-CHAT*	IR-RESPONSE	0.593	1.0e-02	0.333	0.84	1.53
IR-STATUS	IR-RESPONSE*	0.422	3.3e-02	0.330	0.40	1.59
MT-CHAT*	MT-BASELINE	0.577	3.8e-02	0.160	1.23	1.14
MT-CHAT	HUMAN*	0.145	2.2e-16	0.433	N/A	N/A
MT-CHAT*	RND-BASELINE	0.880	2.2e-16	0.383	1.17	0.10

- P-value: exact Binomial significance test
- S coefficient: a measure of agreement between annotators for each experiment
- Automatic Evaluation: BLEU grades an out- put candidate according to n-gram matches to one or more reference outputs

Discuss

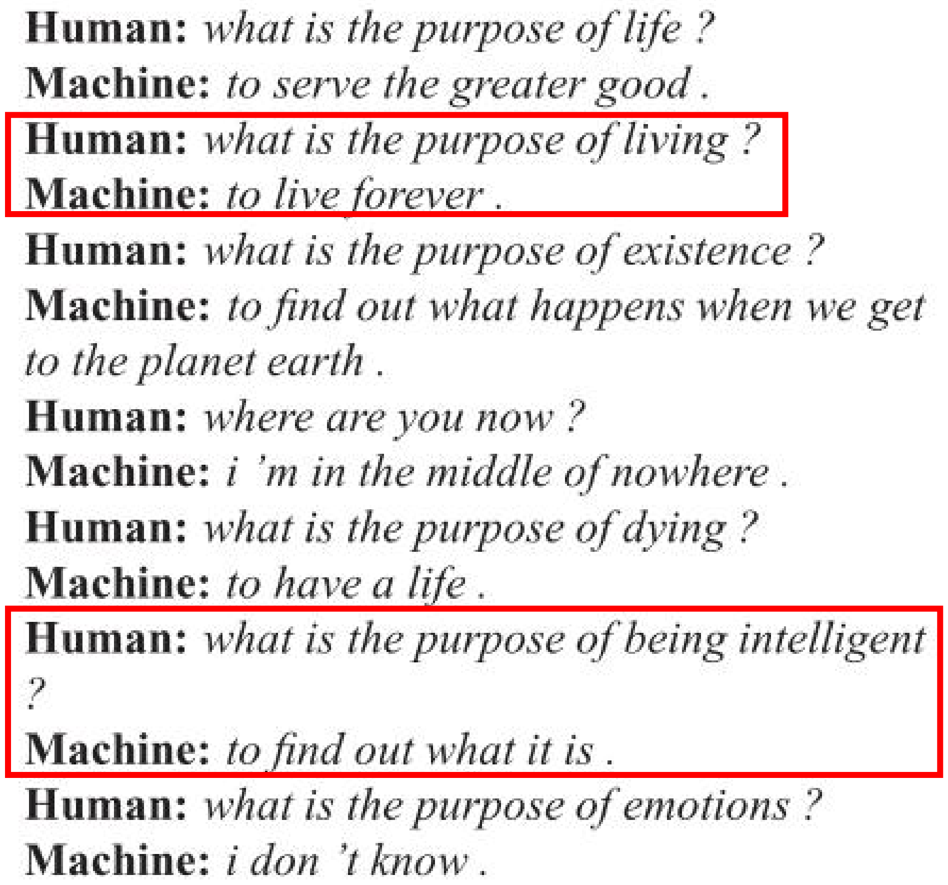
Much room for improvement

- Design appropriate word alignment and decoding algorithm account for the selective nature of response in dialogue
- Simply add more training data

- One to one vs one to many
- Unaligned words/phrases

A Neural Conversational Model

Conversation 4: Philosophical Q&A



Human: *what is the purpose of life ?*
Machine: *to serve the greater good .*
Human: *what is the purpose of living ?*
Machine: *to live forever .*
Human: *what is the purpose of existence ?*
Machine: *to find out what happens when we get to the planet earth .*
Human: *where are you now ?*
Machine: *i 'm in the middle of nowhere .*
Human: *what is the purpose of dying ?*
Machine: *to have a life .*
Human: *what is the purpose of being intelligent ?*
Machine: *to find out what it is .*
Human: *what is the purpose of emotions ?*
Machine: *i don 't know .*

outline

- Goal
- Dataset
- Model
- Result
- Discuss

Goal

- Train an **end-to-end data-driven** conversational model

Dataset 1

A closed-domain IT helpdesk troubleshooting dataset

- Typical interactions are 400 words long
- Turn taking is clearly signaled
- Training Set: 30M tokens   Validation Set: 3M tokens
- Process: remove common names, numbers, and full URLs

Dataset 2

An open-domain movie transcript dataset: OpenSubtitles dataset

- Noisy and Large
- Training Set: 62M sentence/923M tokens Validation set: 26M sentence/395M tokens
- Remove XML tags and non-conversational text

Model

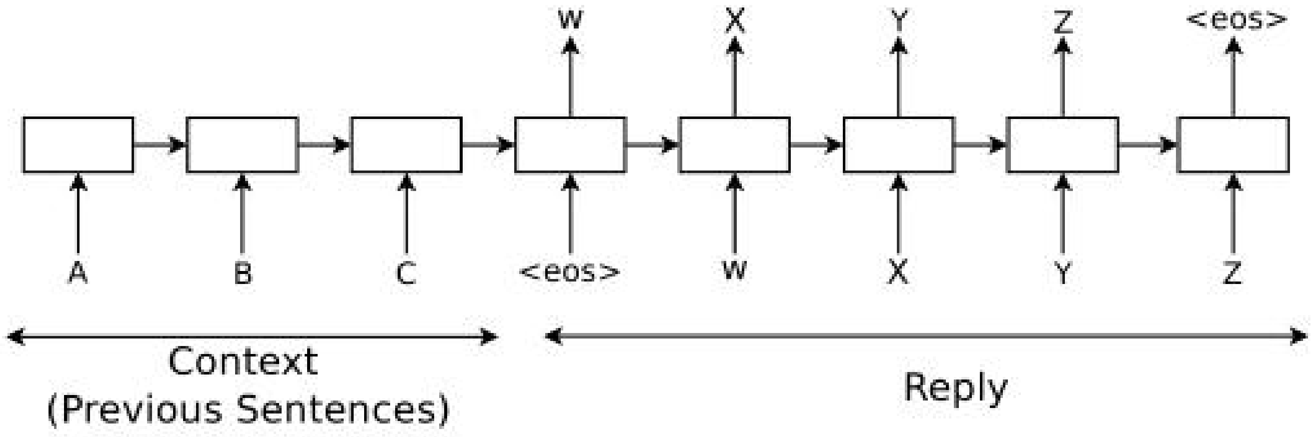


Figure 1. Using the *seq2seq* framework for modeling conversations.

- Maximize the cross entropy of the correct sequence given its context
- a “greedy” inference approach

Sequence to Sequence Learning with Neural Network

- LSTM performs well on long sentences
- LSTM learns better when the source sentences are reversed (NMT)

Experiments 1

IT Helpdesk Troubleshooting experiments

- A single layer LSTM with 1024 memory cells
- Stochastic gradient descent and gradient clipping
- Vocabulary: most common 20K words including special tokens indicating turn taking and actor

Experiments 2

OpenSubtitles experiments

- A two-layered LSTM with 4096 memory cells
- AdaGrad and gradient clipping
- To speed up the softmax, project the memory cells to 2048 linear units before feeding the information to the classifier
- Vocabulary: most frequent 100K words

Result 1

IT Helpdesk Troubleshooting experiments

- Perplexity: 8   n-gram model:18

OpenSubtitles experiments

- Perplexity: 17   smoothed 5-gram model: 28
- adding the soft attention mechanism did not significantly improve the perplexity on neither training or validation sets

Result 2

Human Evaluation

- Four different human were asked to pick which of the two bots they preferred
- Allowed to declare a tie if both answers were of equal quality
- If three out four human judges agreed, record the score

Model	Win	Lose	Tie	Disagreement	Total
NCM	97	60	20	23	200
CleverBot	60	97	20	23	200

Discuss

Highlight

- Generalize to new questions
- Remember facts, understand contexts, perform common sense reasoning

Drawbacks

- Give simple, short, sometimes unsatisfying answers to questions
- the objective function being optimized **does not capture the actual objective achieved through human communication**, which is typically longer term and based on exchange of information rather than next step prediction.
- The lack of a model to ensure **consistency** and general **world knowledge** is another obvious limitation of a purely unsupervised model.
- Open research problem of **designing a good metric** to quick measure the quality of a conversational model