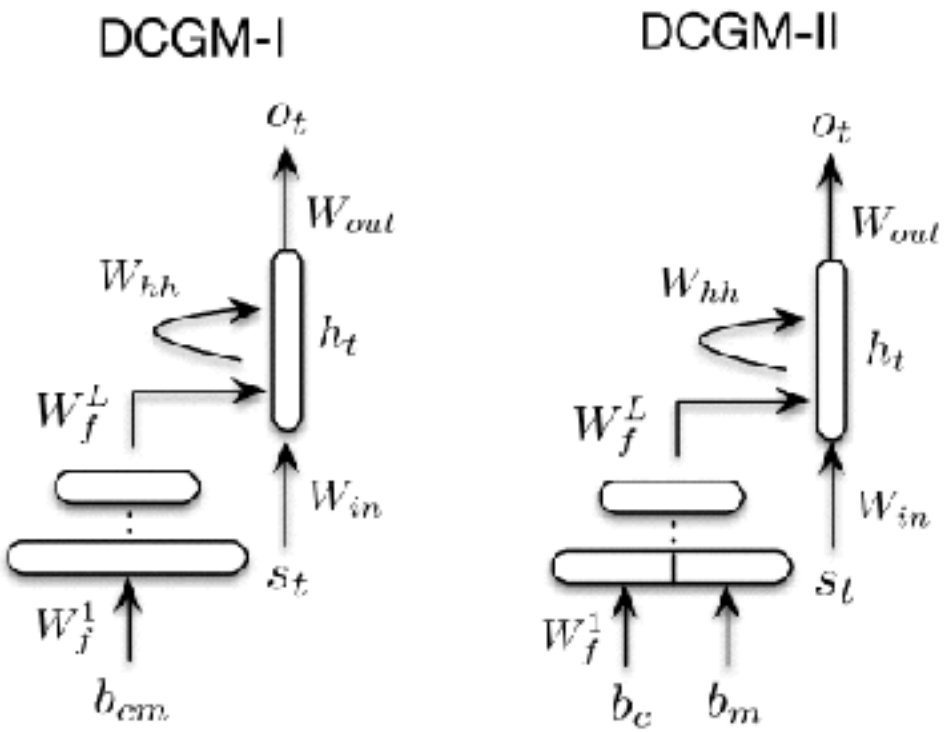
Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models

张璐 2019.09

Dynamic-Context Generative Model I & II

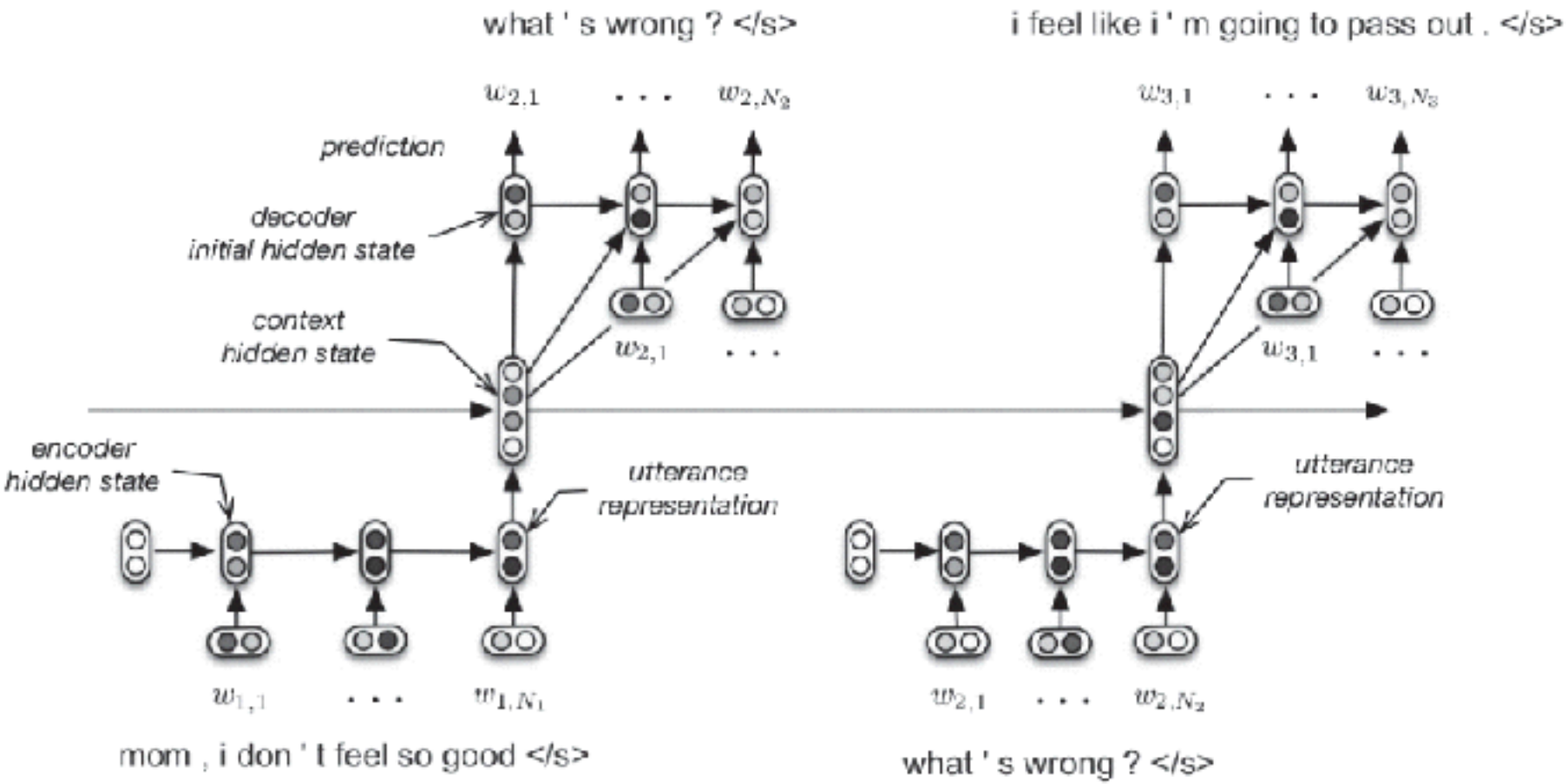


$$k_1 = b_{cm}^\top W_f^1$$
$$k_\ell = \sigma(k_{\ell-1}^\top W_f^\ell) \quad \text{for } \ell = 2, \cdots, L$$

$$k_1 = [b_c^\top W_f^1, b_m^\top W_f^1],$$
$$k_\ell = \sigma(k_{\ell-1}^\top W_f^\ell) \quad \text{for } \ell = 2, \cdots, L$$

Model

- **Hierarchical Recurrent Encoder-Decoder**



Model

- **Bidirectional HRED**
- **Encoder RNN → bidirectional RNN**
- Take the **concatenation of the last state of each RNN** as input to the **context RNN**
- Apply **L2 pooling** over the last state of each RNN, then take the **concatenation** of the two pooled states as input to the context RNN

$$\sqrt{1/N_m \sum_{n=1}^{N_m} h_n^2}$$

Model

- **Bootstrapping from Word Embeddings and Subtitles Q-A**
- **Initialize the word embeddings** E with **Word2Vec**
- **Pretrain** the model on a **large non-dialogue corpus**, which covers similar topics and types of interactions between interlocutors
- **{Q,A}** → **{U1=Q, U2=A}**

Evaluation

- **Perplexity**
- Lower perplexity is indicative of a better model
- **Word classification error/word error-rate**: defined as the number of words in the dataset the model has predicted incorrectly **divided** by the total number of words in the dataset

$$\exp\left(-\frac{1}{N_W}\sum_{n=1}^{N}\log P_\theta(U_1^n, U_2^n, U_3^n)\right)$$

Results

Model	Perplexity	Perplexity@U_3	Error-Rate	Error-Rate@U_3
Backoff N-Gram	64.89	65.05	-	-
Modified Kneser-Ney	60.11	54.75	-	-
Absolute Discounting N-Gram	56.98	57.06	-	-
Witten-Bell Discounting N-Gram	53.30	53.34	-	-
RNN	35.63 ± 0.16	35.30 ± 0.22	$66.34\% \pm 0.06$	$66.32\% = 0.08$
DCGM-I	36.10 ± 0.17	36.14 ± 0.26	$66.44\% \pm 0.06$	$66.57\% = 0.10$
HRED	36.59 ± 0.19	36.26 ± 0.29	$66.32\% \pm 0.06$	$66.32\% = 0.11$
HRED + Word2Vec	33.95 ± 0.16	33.62 ± 0.25	$66.06\% \pm 0.06$	$66.05\% = 0.09$
RNN + SubTle	27.09 ± 0.13	26.67 ± 0.19	$64.10\% \pm 0.06$	$64.07\% = 0.10$
HRED + SubTle	27.14 ± 0.12	26.60 ± 0.19	$64.10\% \pm 0.06$	$64.03\% = 0.10$
HRED-Bi. + SubTle	$\mathbf{26.81 \pm 0.11}$	$\mathbf{26.31 \pm 0.19}$	$\mathbf{63.93\% \pm 0.06}$	$\mathbf{63.91\% \pm 0.09}$

Results

Reference (U_1, U_2)	**MAP**	**Target** (U_3)
U_1: yeah , okay . U_2: well , i guess i ' ll be going now .	i ' ll see you tomorrow .	yeah .
U_1: oh . <continued_utterance> oh . U_2: what ' s the matter , honey ?	i don ' t know .	oh .
U_1: it ' s the cheapest . U_2: then it ' s the worst kind ?	no , it ' s not .	they ' re all good , sir .
U_1: <person> ! what are you doing ? U_2: shut up ! c ' mon .	what are you doing here ?	what are you that crazy ?

- The majority of the predictions were **generic**, such as I don't know
- Data scarcity, model may only learn have learned to predict the most **frequent utterances**
- The majority of tokens were punctuation marks and pronouns → exploring neural architectures which explicitly separate semantic structure from syntactic structure
- The context of a triple may be too short

Discuss

- Stochastic samples from the model produced more diverse dialogues
- Study models for full length dialogues
- More data