A Persona Based Neural Conversation Model

Jiwei Li, Michel Galley, Chris Brockett,
Georgios P. Spithourakis, Jianfeng Gao, Bill Dolan

Association for Computational Linguistics, 2016

Problems Addressed

1. The major issue with all the data-driven models seen till now is their propensity to choose a response with greatest likelihood - which leads to inconsistent responses.

Proposed Solution

1. The proposed solution defines a *PERSONA* - a character that the AI system must play during conversations.
    - The persona is a composite of elements of identity - background facts, language behaviour, interaction style etc. - of a user.
    - The persona is adaptive - an agent may present different facets to different speakers depending on interaction.
2. The personas are incorporated in the form of embeddings, which are trained on top of human-human conversation data.
3. Two persona models, both built on top of a seq2seq framework, are explored:
    - A single speaker *SPEAKER* model, which integrates a speaker-level vector representation at the target part of seq2seq model.
    - A dyadic speaker *SPEAKER-ADDRESSEE* model, which encodes the interaction patterns of two interlocutors by constructing an interaction representation from their individual embeddings and adding it to the seq2seq model.

Notation

- Let $M = m_1, m_2, \ldots, m_I$ denote the input message, and $R = r_1, r_2, \ldots, r_J, EOS$ denote the response to M.
- r_t denotes a word token that is associated with a K dimensional distinct word embedding e_t. V is the vocabulary size.

Speaker Model

- The attributes encoded in the embedding aren't explicitly annotated. The model manages to cluster users along traits based on responses alone.
- Each speaker $i \in [1, N]$ is associated with a user level representation $v_i \in \mathbb{R}^{k \times 1}$.
- The source LSTM encodes the message into a fixed size vector h_S. For each step at the target side, hidden units are obtained by combining the representation produced by target LSTM at the previous time step, the word representations at the current time step, and the speaker embedding v_i. This ensures that the speaker information is injected into the hidden layer at each time step.
- v_i is shared across all conversations that involve speaker i, and it is learnt by backpropagation.

Claims

- This model helps *infer* answers to questions even if evidence isn't present in training set. This is important as not all attributes of every user are covered by training set.
- Since speaker representations are learnt based on conversational content by different speakers, and since similar responses tend to have similar embeddings which lie close by in the vector space, the training data of such similar speakers tend to increase generalization ability of the model.

Speaker-Addressee Model

- Speaking style and content also depends heavily on the identity of the addressee, along with the speaker.
- An interactive representation $V_{ij} \in \mathbb{R}^{k \times 1}$ where $V_{ij} = tanh(W_1 \cdot v_1 + W_2 \cdot v_2), W_1, W_2 \in \mathbb{R}^{K \times K}$ is incorporated into the LSTM at each time step in the target.

Problems

- Obtaining a large scale dataset in which each speaker is involved in a conversation with a large variety of people is problematic.

Claims

- The generalization ability ensures that at test time, even if two speakers $iandj$ were never involved in the training data, two speakers $i'andj'$ that are respectively close in embeddings might have been, which will help model $i's$ response towards j.

Decoding and Reranking

4. For decoding, N-best lists are generated with a beam size of 200, where a maximum length of 20 is set for generated candidates.
5. To prevent small, generic responses, the N-best list is reranked using the scoring function:

   $\log(p(R|M, v)) + \lambda \log(p(M|R)) + \gamma|R|$, *where* λ *and* γ are the associated penalty weights.

   - λ and γ are optimized over N-best lists of response candidates generated from the development set using MERT by optimizing BLEU.
   - An inverse seq2seq model is trained by swapping responses and messages to compute $p(M|R)$. No speaker information was considered in this.

Experiments

1. The datasets used are the Twitter Persona Dataset, Twitter Sordoni Dataset, and Television Series Transcripts.

   Twitter Persona Dataset

   - Training data for the Speaker Model was extracted from the Twitter FireHose for the six-month period beginning January 1, 2012.
   - The sequences were limited to those where the responders had engaged in at least 60 (and at most 300) 3-turn conversational interactions during the period.
   - This yielded a set of 74,003 users who took part in a minimum of 60 and a maximum of 164 conversational turns (average: 92.24, median: 90). The dataset extracted using responses by these "conversationalists" contained 24,725,711 3-turn sliding-window (context-message-response) conversational sequences.
   - In addition, we sampled 12000 3-turn conversations from the same user set from the Twitter FireHose for the three-month period beginning July 1, 2012, and set these aside as development, validation, and test sets (4000 conversational sequences each).
   - Development, validation, and test sets for this data are single-reference, which is by design as multiple reference responses would typically require acquiring responses from different people, which would confound different personas.

Twitter Sordoni Dataset

- To obtain a point of comparison with prior state-of-the-art work the baseline (non-persona) LSTM model is measured against prior work on the dataset of (Sordoni et al., 2015), which is called the Twitter Sordoni Dataset.
- Only the test-set portion, which contains responses for 2114 context and messages, is used.
- Since the Sordoni dataset offers up to 10 references per message, while the Twitter Persona dataset has only 1 reference per message, BLEU scores cannot be compared across the two Twitter datasets (BLEU scores on 10 references are generally much higher than with 1 reference).

Television Series Transcripts

- For the dyadic Speaker-Addressee Model, scripts from the American television comedies Friends and The Big Bang Theory, available from Internet Movie Script Database (IMSDb) are used.
- 13 main characters from the two series in a corpus of 69,565 turns were collected. The corpus was split into training/development/testing sets, with development and testing sets each of about 2,000 turns.

OpenSubtitles Database

- A standard seq2seq model is first trained using the OpenSubtitles (OSDb) dataset, and then it is adapted to the TV series dataset.
- Since this dataset doesn't specify the speaker of every line, following Vinyals et al. (2015), a simplifying assumption that each line of subtitle constitutes a full speaker turn is made.

2. Two types of models are trained: the first one is an RNN Encoder–Decoder, and the other is the proposed model, referred to as RNNsearch.

3. Each model is trained twice: first with the sentences of length up to 30 words (RNNencdec-30, RNNsearch-30) and then with the sentences of length up to 50 word (RNNencdec-50, RNNsearch-50).

4. The encoder and decoder of the RNNencdec have 1000 hidden units each. The encoder of the RNNsearch consists of forward and backward recurrent neural networks (RNN) each having 1000 hidden units. Its decoder has 1000 hidden units.

5. In both cases, a multilayer network with a single maxout hidden layer is used to compute the conditional probability of each target word.

6. A minibatch SGD algorithm together with Adadelta is used to train each model. Each SGD update direction is computed using a minibatch of 80 sentences. Each model was trained for approximately 5 days.

7. Once a model is trained, a beam search is used to find a translation that approximately maximizes the conditional probability.

Results

- The model is also found to be sensitive to the identity of the addressee, generating words specifically targeted at that addressee (e.g., her name).

System	BLEU
MT baseline (Ritter et al., 2011)	3.60%
Standard LSTM MMI (Li et al., 2016)	5.26%
Standard LSTM MMI (our system)	5.82%
Human	*6.08%*

Table 2: BLEU on the Twitter Sordoni dataset (10 references). We contrast our baseline against an SMT baseline (Ritter et al., 2011), and the best result (Li et al., 2016) on the established dataset of (Sordoni et al., 2015). The last result is for a human oracle, but it is not directly comparable as the oracle BLEU is computed in a leave-one-out fashion, having one less reference available. We nevertheless provide this result to give a sense that these BLEU scores of 5-6% are not unreasonable.

Model	Standard LSTM	Speaker Model
Perplexity	47.2	42.2 (−10.6%)

Table 3: Perplexity for standard SEQ2SEQ and the Speaker model on the Twitter Persona development set.

Model	Objective	BLEU
Standard LSTM	MLE	0.92%
Speaker Model	MLE	1.12% (+21.7%)
Standard LSTM	MMI	1.41%
Speaker Model	MMI	1.66% (+11.7%)

Table 4: BLEU on the Twitter Persona dataset (1 reference), for the standard SEQ2SEQ model and the Speaker model using as objective either maximum likelihood (MLE) or maximum mutual information (MMI).

Model	Standard LSTM	Speaker Model	Speaker-Addressee Model
Perplexity	27.3	25.4 (−7.0%)	25.0 (−8.4%)

Table 5: Perplexity for standard SEQ2SEQ and persona models on the TV series dataset.

Model	Standard LSTM	Speaker Model	Speaker-Addressee Model
MLE	1.60%	1.82% (+13.7%)	1.83% (+14.3%)
MMI	1.70%	1.90% (+10.6%)	1.88% (+10.9%)

Table 6: BLEU on the TV series dataset (1 reference), for the standard SEQ2SEQ and persona models.

Bibliography

1. David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. Luke, I am your father: dealing with out-of-domain requests by using movies subtitles. In Intelligent Virtual Agents, pages 13–21. Springer.

2. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proc. of the International Conference on Learning Representations (ICLR).

3. Rafael E Banchs and Haizhou Li. 2012. IRIS: a chat oriented dialogue system based on the vector space model. In Proc. of the ACL 2012 System Demonstrations, pages 37–42.

4. Yun-Nung Chen, Wei Yu Wang, and Alexander Rudnicky. 2013. An empirical investigation of sparse log-linear models for improved dialogue act classification. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8317–8321. IEEE.

5. Werner Deutsch and Thomas Pechmann. 1982. Social interaction and the development of definite descriptions. Cognition, 11:159–184.

6. Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. BLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In Proc. of ACL-IJCNLP, pages 445–450, Beijing, China, July.

7. Jianfeng Gao, Xiaodong He,Wen-tau Yih, and Li Deng. 2014. Learning continuous phrase representations for translation modeling. In Proc. of ACL, pages 699–709, Baltimore, Maryland.

8. Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

9. Alfred Kobsa. 1990. User modeling in dialog systems: Potentials and hazards. AI & society, 4(3):214–231.

10. Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. IEEE Transactions on Speech and Audio Processing, 8(1):11–23.

11. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Proc. of NAACL-HLT.

12. Grace I Lin and Marilyn A Walker. 2011. All the world's a stage: Learning character models from film. In Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE).

13. Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In Proc. of ACL, pages 11–19, Beijing, China, July.

14. Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, Mirna Adriani, and Satoshi Nakamura. 2014.

15. Developing non-goal dialog system based on examples of drama television. In Natural Interaction with Robots, Knowbots and Smartphones, pages 355–361. Springer.

16. Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

17. Alice H Oh and Alexander I Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems-Volume 3, pages 27–32.

18. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. of ACL, pages 311–318.

19. Roberto Pieraccini, David Suendermann, Krishna Dayanidhi, and Jackson Liscombe. 2009. Are we there yet? research in commercial spoken dialog systems. In Text, Speech and Dialogue, pages 3–13. Springer.

20. Adwait Ratnaparkhi. 2002. Trainable approaches to surface natural language generation and their application to conversational dialog systems. Computer Speech & Language, 16(3):435–455.

21. Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 583– 593.

22. Jost Schatztnann, Matthew N Stuttle, Karl Weilhammer, and Steve Young. 2005. Effects of the user model on simulation-based learning of dialogue strategies. In Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on, pages 220– 225.

23. Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proc. of AAAI. Lifeng Shang, Zhengdong Lu, and Hang Li. 2015.

24. Neural responding machine for short-text conversation. In ACL-IJCNLP, pages 1577–1586. Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Meg Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In Proc. of NAACL-HLT.

25. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems (NIPS), pages 3104–3112.

26. J¨org Tiedemann. 2009. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In Recent advances in natural language processing, volume 5, pages 237–248.

27. Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In Proc. of ICML Deep Learning Workshop.

28. Wolfgang Wahlster and Alfred Kobsa. 1989. User models in dialog systems. Springer.

29. Marilyn A Walker, Rashmi Prasad, and Amanda Stent. 2003. A trainable generator for recommendations in multimodal dialog. In INTERSPEECH.

30. Marilyn A Walker, Ricky Grant, Jennifer Sawyer, Grace I Lin, Noah Wardrip-Fruin, and Michael Buell. 2011. Perceived or not perceived: Film character models for expressive nlg. In Interactive Storytelling, pages 109–121. Springer.

31. Marilyn A Walker, Grace I Lin, and Jennifer Sawyer. 2012. An annotated corpus of film dialogue for learning and characterizing character style. In LREC, pages 1373–1378.

32. William Yang Wang, Ron Artstein, Anton Leuski, and David Traum. 2011. Improving spoken dialogue understanding using phonetic mixture models. In FLAIRS Conference.

33. Tsung-Hsien Wen, Milica Gasic, Nikola Mrk˘si´c, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In Proc. of EMNLP, pages 1711–1721, Lisbon, Portugal, September. Association for Computational Linguistics.

34. Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model. CoRR, abs/1510.08565.

35. Steve Young, Milica Ga˘si´c, Simon Keizer, Franc¸ois Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. Computer Speech & Language, 24(2):150–174