Augmenting End-to-End Dialog Systems with Commonsense Knowledge

Tom Young[1], **Erik Cambria**[2], **Iti Chaturvedi**[2], **Minlie Huang**[3], **Hao Zhou**[3], **Subham Biswas**[2]
[1]School of Information and Electronics, Beijing Institute of Technology, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Dept. of Computer Science and Technology, Tsinghua University, China
TOMYOUNG903@GMAIL.COM, CAMBRIA@NTU.EDU.SG, ITI@NTU.EDU.SG,
AIHUANG@TSINGHUA.EDU.CN, TUXCHOW@GMAIL.COM, BISWAS.SUBHAM1@GMAIL.COM

Abstract

Building dialog agents that can converse naturally with humans is a challenging yet intriguing problem of artificial intelligence. In open-domain human-computer conversation, where the conversational agent is expected to respond to human responses in an interesting and engaging way, commonsense knowledge has to be integrated into the model effectively. In this paper, we investigate the impact of providing commonsense knowledge about the concepts covered in the dialog. Our model represents the first attempt to integrating a large commonsense knowledge base into end-to-end conversational models. In the retrieval-based scenario, we propose the Tri-LSTM model to jointly take into account message and commonsense for selecting an appropriate response. Our experiments suggest that the knowledge-augmented models are superior to their knowledge-free counterparts in automatic evaluation.

Introduction

In the past few years, data-driven approaches to building conversation models have been made possible by the proliferation of social media conversation data and computing power. By relying on a large number of message-response pairs, the Seq2Seq framework attempts to produce an appropriate response based solely on the message itself, without any memory module. During evaluation, such models are thus defined only by fixed parameters learned during training. In natural human conversation, however, people respond to each other's utterances in a meaningful way not only by paying attention to the latest utterance of the conversational partner itself, but also by recalling relevant information about the concepts covered in the utterance and integrating it into their responses (in a way that suits the context). Such information may contain personal experience, recent events, commonsense knowledge and more (Fig. 1). As a result, we speculate that a conversational model with a "memory look-up" module can mimic human conversations more closely.

In open-domain human-computer conversation, where the model is expected to respond to human utterances in an interesting and engaging way, commonsense knowledge has to be integrated into the model effectively. In artificial intelligence, commonsense knowledge is the set of background information that an individual is intended to know or assume and the ability to use it when appropriate (Minsky 1986; Cambria et al. 2009; Tran, Cambria, and Hussain 2016). Due to the vastness of such knowledge, we speculate that this goal is better suited by employing an external memory module containing such knowledge than forcing the model to encode it in model parameters as in traditional methods. Hence, in this paper we investigate augmenting end-to-end dialog systems with commonsense knowledge as external memory.

The rest of the paper is organized as follows: Section 2 briefly discusses related work on conversational models and commonsense knowledge; Section 3 introduces our main model and supplementary baselines; Section 4 describes dataset, analysis, and experiments in detail; finally, Section 5 concludes the paper and discusses future work.

Related Work

Conversational models

Data-driven conversational models generally fall into two categories: retrieval-based methods (Lowe et al. 2015b; 2016a; Zhou et al. 2016), that select a response from a predefined repository and generation-based methods (Ritter, Cherry, and Dolan 2011; Serban et al. 2016; Vinyals and Le 2015) that employ an encoder-decoder framework where the message is encoded into a vector representation and, hence, fed to the decoder to generate the response. The latter is more natural without the need of the response repository yet suffers from generating dull or vague responses and generally needs a great amount of training data.

The use of an external memory module in NLP tasks has received considerable attention recently, such as in question answering (Weston et al. 2015) and language modeling (Sukhbaatar et al. 2015). It has also been employed on dialog modeling in several limited settings. With Memory Networks, (Dodge et al. 2015) used a set of fact triples about movies as long-term memory when modeling reddit dialogs, movie recommendation and factoid question answering. Similarly in a restaurant reservation setting, (Bordes and Weston 2016) provided local restaurant information to the conversational model during training and eval-

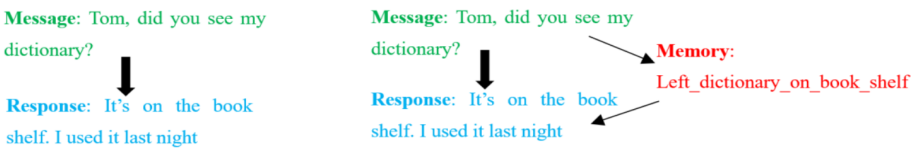

Figure 1: Left: In traditional dialog modeling, the response is determined solely by the message. (Arrows denote dependencies) Right: The responder recalls relevant information from memory about the message; memory and message jointly determine the response. As in the illustrated example, the responder model retrieves the event "Left_dictionary_on_book_shelf" from memory, which, along with the message, triggers a meaningful response.

uation. Researchers have also proposed several methods to incorporate knowledge into the Seq2Seq framework. (Xing et al. 2016) incorporated the topic words of the message obtained from a pre-trained LDA model into the context vector through a joint attention mechanism. (Ghazvininejad et al. 2017) mined FoodSquare tips to be searched by an input message in the food domain and encoded such tips into the context vector through one-turn hop. The Tri-LSTM model we propose in this work shares similarities with (Lowe et al. 2015a), which encoded unstructured textual knowledge with RNN. Our work distinguishes itself from previous research in that we consider a heterogeneous commonsense knowledge base in the open-domain retrieval-based dialog setting.

Commonsense knowledge

Several commonsense knowledge bases have been constructed during the past decade, such as ConceptNet (Speer and Havasi 2012) and SenticNet (Cambria, Olsher, and Rajagopal 2014). The aim is to give a foundation of real-world knowledge to a variety of AI applications. Typically a commonsense knowledge base can be seen as a *semantic network* where *concepts* are nodes in the graph and *relations* are edges. Each $< concept1, relation, concept2 >$ triple is termed an *assertion*. Based on the Open Mind Common Sense project(Singh et al. 2002), ConceptNet not only contains objective facts such as "Paris is the capital of France" that are constantly true, but also captures informal relations between common concepts that are part of everyday knowledge such as "A dog is a pet". This feature of ConceptNet is desirable in our experiments, because the ability to recognize the informal relations between common concepts is necessary in the open-domain conversation setting we are considering in this paper.

Model description

Task definition

In this work, we concentrate on integrating commonsense knowledge into retrieval-based conversational models, because they are easier to evaluate (Liu et al. 2016; Lowe et al. 2016a) and generally take a lot less data to train. We leave the generation-based scenario to future work.

Message (*context*) c and *response* r are a sequence of tokens from vocabulary V. Given c and a set of response candidates $[r_1, r_2, r_3..., r_K] \in R$, the model chooses the most appropriate response $\hat{r}$ according to:

$$\hat{r} = \arg\max_{r \in R} \quad f(c, r), \tag{1}$$

where $f(c, r)$ is a function measuring the "compatibility" of c and r. The model is typically trained on $< message, response, label >$ triples with cross entropy loss, where $label$ is binary indicating whether $response$ is positive or negative.

The Recall@k method is used for evaluating retrieval-based conversational models (Lowe et al. 2016b). The model is asked to rank a total of N responses containing one positive response and $N-1$ negative responses. If the ranking of the positive response is not larger than k, Recall@k is positive for that instance.

Dual-LSTM encoder

As a version of recurrent neural network, a long short-term memory (LSTM) network (Hochreiter and Schmidhuber 1997) is good at handling long-term dependencies and can be used to map an utterance to its last hidden state as fixed-size embedding representation. The kth token in an utterance is first embedded into $e_k \in \mathcal{R}^d$ using a word embedding matrix, where d is the word embedding dimension. Thus, the hidden representation h_k at time step k for the utterance is defined by:

$$\begin{aligned} i_k &= \sigma(W_i \cdot [h_{k-1}, e_k]) \\ f_k &= \sigma(W_f \cdot [h_{k-1}, e_k]) \\ o_k &= \sigma(W_o \cdot [h_{k-1}, e_k]) \\ l_k &= tanh(W_l \cdot [h_{k-1}, e_k]) \\ c_k &= f_k \cdot c_{k-1} + i_k \cdot l_k \\ h_k &= o_k \cdot tanh(c_k) \end{aligned} \tag{2}$$

where W_i, W_f, W_o, $W_l \in \mathcal{R}^{D \times (D+d)}$. An input gate, a memory gate and an output gate, denoted as i_k, f_k and o_k, are used to update cell state c_k and hidden state h_k iteratively. D is the dimension of hidden state h_k. σ denotes the sigmoid function.

Dual-LSTM encoder(Lowe et al. 2015b) represents c and r as fixed-size embeddings $\vec{c}$ and $\vec{r}$ with the last hidden states of the same LSTM. The "compatibility" function of c and r is thus defined by:

$$f(c, r) = \sigma(\vec{c}^T W \vec{r}), \tag{3}$$

where matrix $W \in \mathcal{R}^{D \times D}$ is learned during training.

Commonsense knowledge retrieval

In this paper, we assume that a commonsense knowledge base is composed of assertions A about concepts C. Each as-

sertion $a \in A$ takes the form of a triple $< c_1, r, c_2 >$, where $r \in R$ is a *relation* between c_1 and c_2, such as *IsA*, *CapableOf*, etc. c_1, c_2 are concepts in C. The relation set R is typically much smaller than C. c can be a single word ("dog", "book") or a multi-word phrase such as "take_a_stand", "eiffel_tower"[1]. To improve the speed of query, We build a hash table H out of A where every concept c is a key and a list of all assertions in A concerning c, i.e., $c = c_1$ or $c = c_2$, is the value.

Our goal is to retrieve commonsense knowledge about every concept covered in the message. We define A_c as the set of commonsense assertions concerned with message c. To recover concepts in message c, we use a simple n-gram model ($n \leq N$)[2]. Every n-gram x in c is considered a potential concept[3]. If x exists in H, the corresponding value, i.e., all assertions in A concerning the concept, is added to A_c (Fig. 3).

Tri-LSTM encoder

Our main approach to integrating commonsense knowledge into the conversational model involves using another LSTM for encoding all assertions a in A_c. Each a originally in the form of $< c_1, r, c_2 >$, is transformed into a common sentence by chunking c_1, c_2, concepts which are potentially multi-word phrases, into $[c_{11}, c_{12}, c_{13}...]$ and $[c_{21}, c_{22}, c_{23}...]$. Thus, $a = [c_{11}, c_{12}, c_{13}..., r, c_{21}, c_{22}, c_{23}...]$. We add R to vacabulary V, that is, each r in R will be treated like any regular word in V during encoding. We decide not to use each concept c as unit for encoding a because C is typically too large (¿1M). a is encoded as embedding representation $\vec{a}$ using another LSTM. Note that this encoding scheme is suitable for any natural utterances containing commonsense knowledge[4] in addition to well-structured commonsense assertions. We define the *match score* of an assertion a and response r as:

$$m(a, r) = \vec{a}^T W_a \vec{r} \tag{4}$$

where $W_a \in \mathcal{R}^{D \times D}$ is learned during training. Commonsense assertions A_c associated with a message is usually large (¿100 in our experiment). We observe that in a lot of cases of human communication, response r can be seen as triggered by certain perception or understanding of message c defined by one or more assertions in A_c, as illustrated in Fig. 3.

Our assumption is that A_c is helpful in assigning higher scores to an appropriate response r. However, usually very few assertions in A_c are related to a particular response r in the open-domain setting. As a result, we define the *match score* of A_c and r as

$$m(A_c, r) = \max_{a \in A_c} \; m(a, r), \tag{5}$$

that is, we only consider the commonsense assertion a with the highest match score with r, as most of A_c are not relevant to r. Incorporating $m(A_c, r)$ into the Dual-LSTM encoder, our Tri-LSTM encoder model is thus defined as:

$$f(c, r) = \sigma(\vec{c}^T W \vec{r} + m(A_c, r)), \tag{6}$$

i.e., we use simple addition to supplement c with A_c, without introducing a mechanism for any further interaction between c and A_c. This simple approach favors response selection and proves effective in practice (Section 4). The intuition we are trying to capture here is that an appropriate response r should not only be compatible with c, but also related to certain memory recall triggered by c as captured by $m(A_c, r)$. In our case, the memory is commonsense knowledge about the world. In cases where $A_c = \emptyset$, i.e., no commonsense knowledge is recalled, $m(A_c, r) = 0$ and the model degenerates to dual-LSTM encoder.

Comparison Approaches

Supervised word embeddings We follow (Bordes and Weston 2016; Dodge et al. 2015) and use supervised word embeddings as another baseline. Word embeddings are most well-known in the context of unsupervised training on raw text as in (Mikolov et al. 2013), yet they can also be used to score (c, r) pairs. The embedding vectors are trained directly for this goal. The bag-of-word embeddings of c, r are $\vec{c} = A\hat{c}, \vec{r} = B\hat{r}$, where $\hat{x}$ is the bag-of-word representation of utterance x and $A, B \in \mathcal{R}^{d \times |V|}$ are word embedding matrices for c and r. In this setting the "compatibility" function of c and r is defined as:

$$f(c, r) = \vec{c}^T \vec{r} \tag{7}$$

With retrieved commonsense assertions A_c, we embed each $a \in A_c$ with $C \in \mathcal{R}^{d \times |V|}$ to $\vec{a}$ and have:

$$f(c, r) = \vec{c}^T \vec{r} + \max_{a \in A_c} \; \vec{a}^T \vec{r}. \tag{8}$$

This linear model differs from Tri-LSTM encoder in that it represents an utterance with its bag-of-word embedding instead of recurrent neural networks.

Memory Networks Memory Networks (Sukhbaatar et al. 2015; Weston, Chopra, and Bordes 2014) are a recent class of models that perform language understanding by incorporating a memory component. It performs attention over memory to retrieve all relevant information that may help with the task. In our dialog modeling setting, we use A_c as the memory component. Our implementation of Memory Networks, similar to (Bordes and Weston 2016; Dodge et al. 2015), differs from supervised word embeddings in Section 3.5.1 in only one aspect: how to treat multiple entries in memory.

In memory networks, output memory representation $\vec{o} = \sum_i p_i \vec{a}_i$, where $\vec{a}_i$ is the bag-of-word embedding of $a_i \in A_c$ and p_i is the attention signal over memory A_c calculated by

[1] We overload notation c with $concept$ in a knowledge base and $message$ ($context$) in conversation models; we overload notation r with $relation$ in a knowledge base and $response$ in conversation models.

[2] More sophisticated methods such as $concept\ parser$(Rajagopal et al. 2013) are also possible. Here we chose n-gram for speed. N is set to 5.

[3] For unigrams, we exclude a set of stopwords. Both the original version of every word and stemmed version are considered.

[4] Termed *surface text* in ConceptNet.

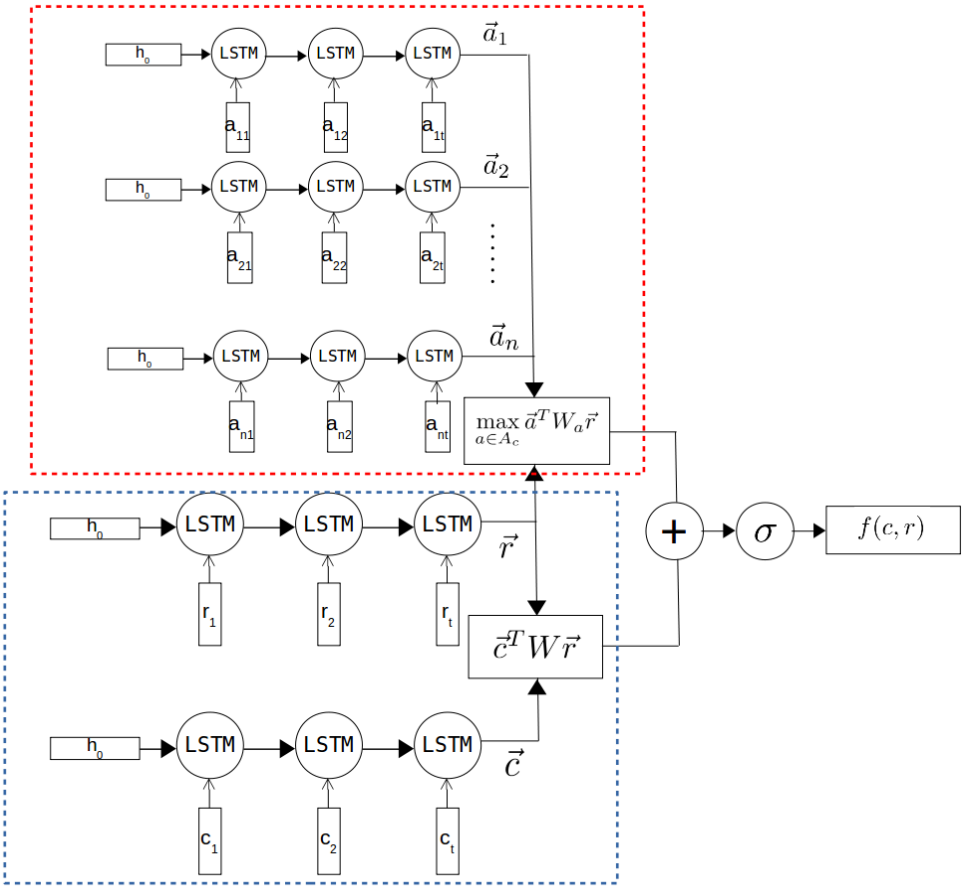

Figure 2: Tri-LSTM encoder. We use LSTM to encode message, response and commonsense assertions. LSTM weights for message and response are tied. The lower box represents Dual-LSTM encoder. The upper box is the memory module encoding all commonsense assertions.

$p_i = softmax(\vec{c}^T \vec{a_i})$. The "compatibility" function of c and r is defined as :

$$f(c, r) = (\vec{c} + \vec{o})^T \vec{r} = \vec{c}^T \vec{r} + (\sum_i p_i \vec{a}_i)^T \vec{r} \quad (9)$$

In contrast to supervised word embeddings in Section 3.5.1, attention over memory is determined by message c. This mechanism is originally designed to retrieve information from memory that is relevant to the context, which in our setting is already achieved in commonsense knowledge retrieval (Section 3.3). As speculated in Section 3.4, the attention over multiple memory entries is better determined by response r in our setting. We empirically prove this point in Section 4.

Experiments

Twitter Dialog Dataset

As far as we are aware of, there is currently no well-established open-domain response selection benchmark dataset available, although certain Twitter datasets have been used in the response generation setting (Li et al. 2015; 2016). We thus evaluate our method against state-of-the-art approaches in the response selection task on Twitter dialogs.

1.4M Twitter (status, response) pairs are used for our experiments. They were extracted over the 5-month period from February through July in 2011. 1M Twitter (status, response) pairs are used for training. With the original response as ground truth, we construct 1M (message, response, label=1) triples as positive instances. Another 1M negative instances (message, response, label=0) are constructed by replacing the ground truth response with a random response in the training set.

For tuning and evaluation, we used 20K (status, response) pairs that constitute the validation set (10K) and test set (10K). They are selected by a criterion that encourages interestingness and relevance: both the status and response have to be at least 3 tokens long and contain at least one non-stopword. For every status, at a least one concept has to be found in the commonsense knowledge base.[5] For each instance, we collect another 9 random responses from else-

[5]We found that 73% of all twitter statuses satisfy this condition in our dataset. We concede that our current approach to integrating commonsense knowledge is only effective to this portion of data

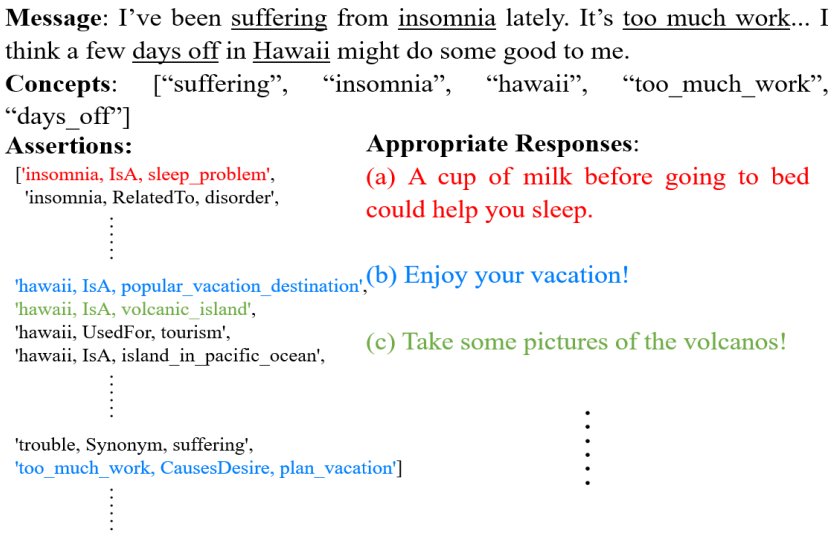

Figure 3: We find all n-grams in the message in the hash table H. In the illustrated case, five concepts are found in the knowledge base. All assertions associated with the five concepts constitute A_c. We show 3 appropriate responses for this single message. Each of them is associated with (same color) only one or two commonsense assertions, which is a paradigm in open-domain conversation and provides ground for our max-pooling strategy. It's also possible that an appropriate response is not relevant to any of the common assertions in A_c at all, in which case our current method makes no difference.

where to constitute the response candidates.

Preprocessing of the dataset includes normalizing hashtags, "@User", URLs, emoticons. Vacabulary V is built out of the training set with 5 as minimum word frequency, containing 62535 words and an extra $< UNK >$ token representing all unknown words.

ConceptNet

In our experiment, ConceptNet 5[6] is used as the commonsense knowledge base. Preprocessing of this knowledge base includes removing assertions containing non-English characters or any word outside vocabulary V. A total of 46 relations are added to V.

The resulting hash table H contains 1.4M concepts. 0.8M concepts are single words, 0.43M are bi-grams and the other 0.17M are tri-grams or more. Each concept is associated with an average of 4.3 assertions. More than half of the concepts are associated with only one assertion.

An average of 2.8 concepts can be found in ConceptNet 5 for each status in our Twitter Dialog Dataset, yielding an average of 150 commonsense assertions (the size of A_c). Unsurprisingly, common concepts with more assertions associated are favored in actual human conversations.

It's worth noting that ConceptNet 5 is also noisy due to uncertainties in the constructing process, where 15.5% of all entries are considered "false" or "vague" by human evaluators (Speer and Havasi 2012). Our max-pooling strategy used in Tri-LSTM encoder and supervised word embeddings is partly designed to alleviate this weekness.

and offers no performance boost to the rest 27%.

[6]https://conceptnet.io. Downloading can be found at github.com/commonsense/conceptnet5/wiki/Downloads

Parameter Settings

In all our models excluding TF-IDF (Ramos and others 2003), we initialize word embeddings with pretrained GloVe embedding vectors(Pennington, Socher, and Manning 2014). The size of hidden units in LSTM models is set to 256 and the word embedding dimension is 100. We use Stochastic Gradient Descent (SGD) for optimizing with batch size of 64. We set 0.001 as fixed training rate.

Results and Analysis

The main results for TF-IDF, word-embeddings, Memory Networks and LSTM models are summarized in Table 1. We observe that:

(1) LSTMs perform better at modeling dialogs than models based on word embeddings on our dataset, as shown by the comparison between Tri-LSTM and WE.

(2) Integrating commonsense knowledge into conversational models boosts model performance, as Tri-LSTM outperforms Dual-LSTM by a certain margin.

(3) Max-pooling over all commonsense assertions depending on response r is a better method for utilizing commonsense knowledge than attention over memory in our setting, as demonstrated by the gain of performance of WE over MN.

We also analyze samples from the test set to gain an insight on how commonsense knowledge supplements the message itself in the response selection tasks by comparing Tri-LSTM encoder and Dual-LSTM encoder. As illustrated in Table 2, instances 1,2 represent cases where commonsense assertions as an external memory module provide certain clues that the other model failed to capture. For example in instance 2, Tri-LSTM selects the response "...improve

Table 1: Models performance on Recall@k with 10 candidates. WE stands for supervised word embeddings (Section 3.5.1). MN stands for Memory Networks (Section 3.5.2). * indicates models with commonsense knowledge integrated. The TF-IDF model is trained following (Lowe et al. 2015b). Human performance is obtained by averaging the performance of 3 volunteers on a total of 900 instances.

Recall@k	TF-IDF	WE*	MN*	Dual-LSTM	Tri-LSTM*	Human
Recall@1	32.6%	73.5%	72.1%	73.6%	**77.5%**	87.0%
Recall@2	47.3%	84.0%	83.6%	85.6%	**88.0%**	-
Recall@5	68.0%	95.5%	94.2%	95.9%	**96.6%**	-

Table 2: Case studies for the impact of commonsense assertions. "Activated Assertion" is a commonsense assertion entry in A_c produced by max-pooling. $\diamondsuit$ indicates correct selection. All 4 instances displayed are taken from the test set.

Instance	Message	Response selected by Dual-LSTM
1	i was helping my brother with his chinese.	did yoga help?
2	bonjour madame, quoi de neuf.	yeah me too !
3	help what colour shoes can i wear with my dress to the wedding?	very pale pink or black.$^\diamondsuit$
4	helping mum paint my bedroom.	shouldn't it be your mum helping you? what color are you going for ?$^\diamondsuit$

Instance	Response selected by Tri-LSTM	Activated Assertion (total size of A_c)
1	the language sounds interesting! i really gotta learn it !$^\diamondsuit$	*chinese, IsA, human_language (755)*
2	loool . you can stick with english , its all good unless you want to improve your french .$^\diamondsuit$	*bonjour, IsA, hello_in_french (9)*
3	very pale pink or black.$^\diamondsuit$	*pink, RelatedTo, colour (1570)*
4	shouldn't it be your mum helping you? what color are you going for ?$^\diamondsuit$	*paint, RelatedTo, household_color (959)*

your french" to message "bonjour madame" based on a retrieved assertion "$bonjour, IsA, hello\_in\_french$", while dual-LSTM selects an irrelevant response.

Unsurprisingly, Dual-LSTM is also able to select the correct response in some cases where certain commonsense knowledge is necessary, as illustrated in instance 3. Both models select "... pink or black" in response to message "...what colour shoes...", even though dual-LSTM doesn't have access to a helpful assertion "$pink, RelatedTo, colour$". Informally speaking, such cases suggest that to some extent, Dual-LSTM (models with no memory) is able to encode certain commonsense knowledge in model parameters (such as word embeddings) in an implicit way. In other cases like instance 4, the message itself is enough for the selection of the correct response, where both models do equally well.

Conclusions

In this paper, we emphasized the role of memory in conversational models. In an open-domain chit-chat setting, we experimented with commonsense knowledge as external memory and proposed a method of using LSTM to encode commonsense assertions to supplement response selection. In the other research line of response generation, such knowledge can potentially be used to condition the decoder in favor of more interesting and relevant responses.

Although the gains presented by our new method is not spectacular in the traditional Recall@k evaluation framework, our view represents an attempt at integrating a large heterogeneous knowledge base that potentially describes the world into conversational models as a memory component. Our future work includes extending the commonsense knowledge with common knowledge, e.g., to extend the knowledge base coverage by linking named entities to commonsense knowledge concepts (Cambria et al. 2014), and developing a better mechanism for utilizing such knowledge instead of the simple max-pooling scheme used in this paper.

References

Bordes, A., and Weston, J. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Cambria, E.; Hussain, A.; Havasi, C.; and Eckl, C. 2009. Common sense computing: From the society of mind to digital intuition and beyond. In Fierrez, J.; Ortega, J.; Esposito, A.; Drygajlo, A.; and Faundez-Zanuy, M., eds., *Biometric ID Management and Multimodal Communication*, volume 5707 of *Lecture Notes in Computer Science*. Berlin Heidelberg: Springer. 252–259.

Cambria, E.; Song, Y.; Wang, H.; and Howard, N. 2014. Semantic multi-dimensional scaling for open-domain sentiment analysis. *IEEE Intelligent Systems* 29(2):44–51.

Cambria, E.; Olsher, D.; and Rajagopal, D. 2014. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*.

Dodge, J.; Gane, A.; Zhang, X.; Bordes, A.; Chopra, S.; Miller, A.; Szlam, A.; and Weston, J. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.

Ghazvininejad, M.; Brockett, C.; Chang, M.; Dolan, B.; Gao, J.; Yih, W.; and Galley, M. 2017. A knowledge-grounded neural conversation model. *CoRR* abs/1702.01932.

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

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Lowe, R.; Pow, N.; Charlin, L.; Pineau, J.; and Serban, I. V. 2015a. Incorporating unstructured textual knowledge sources into neural dialogue systems. In *Machine Learning for Spoken Language Understanding and Interaction, NIPS 2015 Workshop*.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015b. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016a. On the evaluation of dialogue systems with next utterance classification. *arXiv preprint arXiv:1605.05414*.

Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016b. On the evaluation of dialogue systems with next utterance classification. *CoRR* abs/1605.05414.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Minsky, M. 1986. *The Society of Mind*. New York: Simon and Schuster.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, 1532–1543.

Rajagopal, D.; Cambria, E.; Olsher, D.; and Kwok, K. 2013. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *Proceedings of the 22nd International Conference on World Wide Web*, 565–570. ACM.

Ramos, J., et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, 583–593. Association for Computational Linguistics.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Singh, P.; Lin, T.; Mueller, E.; Lim, G.; Perkins, T.; and Li Zhu, W. 2002. Open mind common sense: Knowledge acquisition from the general public. *On the move to meaningful internet systems 2002: CoopIS, DOA, and ODBASE* 1223–1237.

Speer, R., and Havasi, C. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, 3679–3686.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, 2440–2448.

Tran, H.-N.; Cambria, E.; and Hussain, A. 2016. Towards gpu-based common-sense reasoning: Using fast subgraph matching. *Cognitive Computation* 8(6):1074–1086.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W. 2016. Topic augmented neural response generation with a joint attention mechanism. *CoRR* abs/1606.08340.

Zhou, X.; Dong, D.; Wu, H.; Zhao, S.; Yan, R.; Yu, D.; Liu, X.; and Tian, H. 2016. Multi-view response selection for human-computer conversation. *EMNLP16*.