Entity Resolution in Open-domain Conversations

Mingyue Shang*, Tong Wang*, Mihail Eric, Jiangning Chen, Jiyang Wang
Matthew Welch, Tiantong Deng, Akshay Grewal, Han Wang, Yue Liu
Imre Kiss, Yang Liu, Dilek Hakkani-Tur
Amazon Alexa
{myshang, tonwng, mihaeric, cjiangni, welcmtt, jiyangw, gracdeng,
aksgrewa, wnghn, lyu, ikiss, yangliud, hakkanit}@amazon.com

Abstract

In recent years, incorporating external knowledge for response generation in open-domain conversation systems has attracted great interest. To improve the relevance of retrieved knowledge, we propose a neural entity linking (NEL) approach. Different from formal documents such as news, conversational utterances are informal and multi-turn, which makes it more challenging to disambiguate the entities. Therefore, we present a context-aware named entity recognition model (NER) and entity resolution (ER) model to utilize dialogue context information. We conduct NEL experiments on three open-domain conversation datasets and validate that incorporating context information improves the performance of NER and ER models. Furthermore, we verify that using knowledge sentences identified based on NEL benefits the neural response generation model.

1 Introduction

Building an informative open-domain conversational agent that can naturally interact with humans has been one of recent scientific research topics. Inspired by the development of neural networks, neural generation based conversation systems have made great progress (Sutskever et al., 2014; Vinyals and Le, 2015; Li et al., 2017; Wolf et al., 2019a; Zhou et al., 2020). However, one issue in such approaches is that the neural models often produce universal and less informative responses (Huang et al., 2020). To address this issue, previous work proposed to incorporate external information into the response generation models, such as topics (Xing et al., 2017) and emotions (Zhou et al., 2018a). One line of research investigates the use of external knowledge to enrich the information of the responses (Ghazvininejad et al., 2018; Young et al., 2018; Dinan et al., 2018; Gopalakrishnan et al., 2019; Meng et al., 2020). Most existing studies retrieve relevant knowledge from a knowledge base using the entities and noun phrases in the input text. Thus, correctly identifying these entities is crucial to find the relevant knowledge for a given dialog context. This typically involves two subtasks: given a user utterance, the system first identifies any named entities it contains (NER task) and then performs entity resolution (ER) to disambiguate the mentioned entities using a knowledge base. Both NER and ER (or NEL) have been well explored in previous studies and demonstrated to perform highly for news or well written text. However, for open domain spoken conversations and human-bot dialog, performance suffers due to ASR errors, incomplete or ungrammatical sentences from users, difference of spoken and written style, and less training data for such tasks.

In this paper, we propose to use neural entity linking (NEL) technologies that leverage both utterance-level and dialog-level context to retrieve relevant knowledge. As shown in the example in Figure 1, dialogues often contain multiple turns and information is dispersed throughout each turn. Thus, a single turn of interaction may be insufficient for entity disambiguation. Therefore, we leverage previous utterances in the dialogue as the context information and propose context-aware models to better solve the NER and ER tasks in open-domain conversation systems. When recognizing and disambiguating entities in a given utterance, we encode dialog context, and adopt the attention mechanism to extract the information related to the current utterance. To verify the effectiveness of context-aware models, in addition to the intrinsic evaluations, i.e., NER and ER standalone performance, we conduct an extrinsic evaluation where NER and ER results are integrated in a knowledge grounded neural response generation model in an open domain conversation system and response quality is evaluated. Our major contributions can

*The first two authors have equal contribution

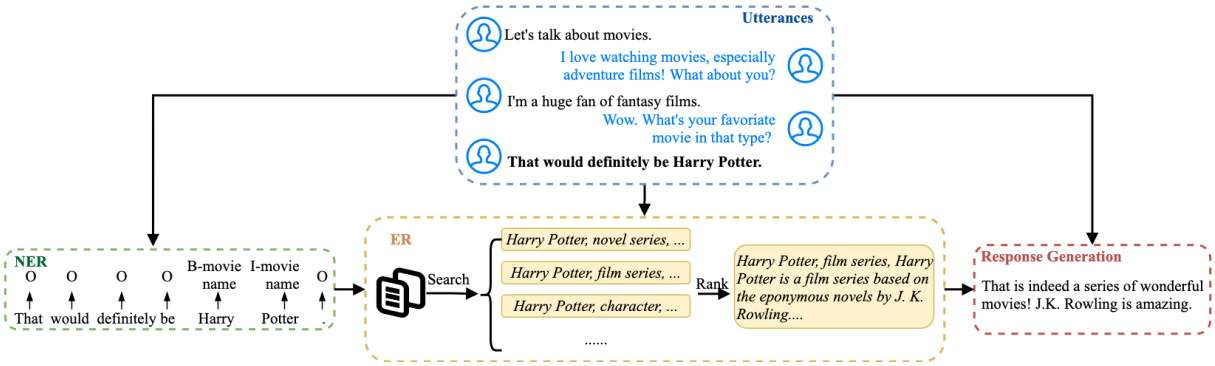

Figure 1: An example dialog illustrating the pipeline of NER, ER, and response generation. The bold sentence in the utterances is the current utterance and the previous utterances are the context. The current utterance and its context are fed to the NER module to identify the entity mentions. Then the ER module takes the entity mentions and all the sentences as input to resolve the entity. The response generation module produces an output based on the knowledge entity information and the dialog input.

be summarized as follows:

- We propose neural network based context-aware models for NER and ER respectively in open domain conversations.
- Experimental results on different conversation datasets show that our proposed context-aware NER and ER models outperform other state-of-the-art models that do not use context information.
- In an end2end evaluation, we demonstrate that incorporating ER information improves quality of neural response generation models in open domain conversations.

2 Related Work

2.1 Open-domain Conversation System

Inspired by the availability of conversational data and the prosperity of neural networks, building open-domain conversation systems by data-driven approaches has achieved great progress. Previous methods can be roughly divided into two categories, retrieval-based (Zhang et al., 2018; Wu et al., 2019; Tao et al., 2019) and generation-based (Vinyals and Le, 2015; Li et al., 2017; Asghar et al., 2018; Tao et al., 2018). Chen et al. (2017) point out that conventional sequence-to-sequence methods tend to generate trivial responses that lack information and diversity. To address this issue, a line of research proposes to incorporate external knowledge into the generation process. Most of the work in this line retrieves knowledge based on a search or retrieval step first, and followed by further reranking of retrieved relevant knowledge snippets (Ghazvininejad et al., 2018; Young et al., 2018; Zhou et al., 2018b; Gopalakrishnan et al., 2019; Zhao et al., 2020). In our work, we propose neural entity recognition and linking to identify and resolve entities more accurately in order to obtain more relevant knowledge for knowledge grounded response generation.

2.2 Neural Entity Linking

NEL typically involves two tasks: recognizing named entities in a given text and then disamgibuating the entity mentions according to the knowledge base (KB). Researchers have shown great success in NER with the help of Convolutional Neural Networks (CNNs), Bidirectional Recurrent Neural Networks (Bi-RNNs), and attention mechanisms along with a CRF decoder (Chiu and Nichols, 2016; Akbik et al., 2018; Ghaddar and Langlais, 2018; Jiang et al., 2019; Baevski et al., 2019; Yamada et al., 2020). Deep neural networks (DNNs) are also dominant in entity resolution tasks. They are used to calculate the semantic similarity between the recognized entity mentions and the entities in the KB (Yamada et al., 2016; Ganea and Hofmann, 2017; Sil et al., 2018; Raiman and Raiman, 2018). However, previous NEL work has mainly focused on news or formal documents, which is different from open-domain dialogues in many aspects. Sentences in open-domain dialogues are more informal, making it more difficult to recognize and disambiguate entities. In addition, since conversations are multi-turn, the semantic information in the current utterance is ambiguous and context needs to be considered. In this paper, we investigate NEL in open-domain conversational data and propose context-aware NER and ER models.

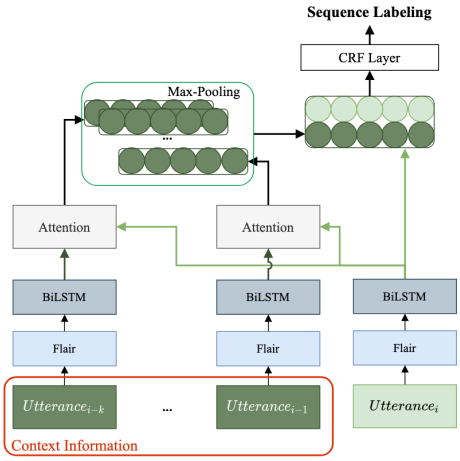

Figure 2: Context-aware NER model: information from previous k utterances is used while performing NER on utterance i.

3 Methodology

3.1 Problem Formulation

Our problem can be formulated as follows. Given an open-domain dialogue until a time point $D = c_i, x_i$, where x_i is the current utterance, we define the utterance context $c_i = \{u_1, \ldots, u_k\}$ as the list of utterances prior to x_i, and k is the size of the context. For each x_i given c_i, an NER model is applied to detect entity mentions in the form of BIO labels.[1] Then for each predicted entity mention, y_j, a query is formulated to search a knowledge base to get a list of candidate entities, $\{e_1, \ldots, e_m\}$, where m is the size of the returned entities from the search. An ER model is then used to rank the entities and identify the most relevant entity, e_t. Finally, a response, r_i, is generated based on c_i, x_i, and knowledge sentences obtained from the linked entities e_t. Note a knowledge ranking algorithm is applied when there are multiple knowledge sentences corresponding to e_t or there are multiple entity mentions in x_i. Figure 1 overviews the pipeline of generating responses with NER and ER modules.

3.2 Context-Aware Named Entity Recognition Model

Figure 2 gives the overall architecture of the context-aware NER model. Following the framework presented by Chiu and Nichols (2016), we employ a bi-directional, long short-term memory (Bi-LSTM) model to extract word features and a conditional random field (CRF) to predict the NER labels.

[1]These labels are widely used for NER and indicate a token is *Begin*, *Inside*, or *Outside* an entity mention, respectively.

Suppose we have an utterance $x_i = \{w_1^i, \ldots, w_T^i\}$, where T is the length of x_i and w_t is the t-th token. After converting each token in x_i to its vector representation through a word embedding table[2], the Bi-LSTM layer encodes the sentence into hidden states h_t^i, which are the concatenation of $\overrightarrow{h}_t^i$ from the forward LSTM and $\overleftarrow{h}_t^i$ from the backward LSTM. The CRF layer then takes the hidden states as input to predict the label probability.

As discussed earlier, as opposed to news or documents, recognizing and disambiguating the named entities in conversational utterances requires consideration of the context information. Therefore, we employ another Bi-LSTM layer to encode the context utterances from the previous turns,

$$s_t^j = [\overrightarrow{s}_t^j; \overleftarrow{s}_t^j] \tag{1}$$

where $\overrightarrow{s}_t^j$ is the forward hidden state of the t-th token in the context utterance u_j and $\overleftarrow{s}_t^j$ is the backward hidden state.

We use an attention mechanism to model the different impact of the previous utterances in the context:

$$\text{Attention} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, V refer to the query, key, and value, respectively. Here, the key and value are the context sentences, and the query is the current utterance. To aggregate the context information, a max-pooling operation is performed on the dimension of sentences. Then, the context vector is concatenated with the sentence vector, and then is supplied as the input of the CRF layer.

3.3 Context-aware Entity Resolution Model

Our entity resolution model contains two steps: coarse-grained candidate selection and fine-grained candidate ranking.

Candidate selection At this stage we retrieve relevant entities from the KB. We create an Elasticsearch (Gormley and Tong, 2015) index with the entity labels and apply both an exact and a Levenshtein distance based fuzzy match to obtain candidate entities. For each entity mention, we take the top 10 search results, ranked by Elasticsearch, as the candidates for the subsequent reranking step.

[2]Here we adopt the stacked embedding released by Flair (Akbik et al., 2018).

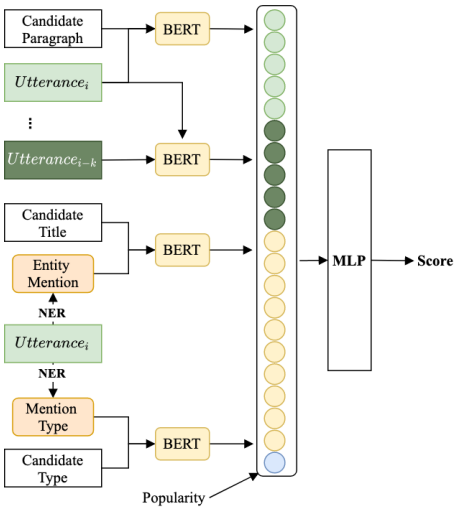

Figure 3: Context-aware ER reranking.

Reranking At this stage the candidate entities are re-ranked based on the match scores from our context-aware model. We propose to compute the relevance score from the entity, utterance and session levels. The structure of the multi-level re-ranking model is shown in Figure 3.

Entity-Level Matching: This considers the candidate entity's label and type attributes, and matches with the entity mention and the predicted type, respectively.

Utterance-Level Matching: This measures the matching degree between the candidate entity's description and the current utterance based on sentence-level semantic information.

Session-Level Matching: This treats the context and current utterance as a conversation session, and computes its match score with the candidate entity's description.

For each matching level, we first concatenate the representations from the entity candidate in the KB and the dialog side, and then employ BERT (Devlin et al., 2018) to get their representations. v_{label}, v_{type}, $v_{utterance}$, $v_{session}$ represent the output of BERT corresponding to the mention label and type (entity-level), utterance-level, and session-level, respectively. We also define the popularity of an entity based on the number of views in the last 60 days, represented as v_p. All these features are concatenated and then fed into an MLP layer to predict the ranking score:

$$\begin{aligned} v &= [v_{label}; v_{type}; v_{utterance}; v_{session}; v_p] \\ s &= \text{MLP}(v) \end{aligned} \tag{3}$$

To train this model, we minimize the pair-wise hinge-loss, defined as:

$$l_r = max(0, \sigma + s^- - s^+) \tag{4}$$

where s^+ is the ranking score of the ground-truth entity and s^- is the ranking score of a negative entity sampled from candidates other than the ground-truth. σ is a constant margin and is set to 0.5.

3.4 Response Generation Model

Given the linked entities, we employ a transformer-based response generation model that is trained to leverage the context of a dialogue along with the knowledge relevant at a given turn. More specifically, we first fine-tune a GPT2-medium model using the Wizard of Wikipedia (WOW) dataset (Dinan et al., 2018). WOW is a suitable dataset for fine-tuning as it involves knowledge-grounded conversations dealing with Wikipedia articles, a data source we are using for entity linking in this work.

The GPT2 generation model is fine-tuned in a matter consistent with (Wolf et al., 2019b; Gopalakrishnan et al., 2020). During generation, we are provided a dialogue context, $C = \{c_1, c_2, ..., c_{i-1}\}$ containing utterances before c_i. We use our linked entities to query the relevant Wikipedia articles, and use the first paragraph of the returned articles, giving us a collection of knowledge sentences, $K = \{k_1, k_2, ..., k_n\}$.

Next, we truncate each knowledge sentence with more than 64 tokens and provide a concatenated input consisting of the dialogue context and the knowledge sentences. We then sample from the language model, one token at a time, using nucleus sampling to form our generated system response.

4 Experiment Setup

4.1 Datasets

We rely on Wikipedia and Wiki data[3] to build the knowledge base for this task. We built a Knowledge Graph (KG) containing over 6M entities including attributes such as Wiki ID, title, type, and introduction. To perform NEL on conversational data, we collect a **M**ulti-turn **O**pen-domain **C**onversation Dataset (MOC) and ask crowd worker annotators to first annotate NER labels (entity mention and type), and then give ER labels – the ground truth Wikidata ID. Different from the entity labels in regular NER tasks, we define 50 entity types across 8 popular domains in open-domain conversations

[3]https://www.wikidata.org/wiki/,https://www.wikipedia.org/

including Fashion, Politics, Books, Sports, Music, Science/Technology, Game, Video/Movies. In addition, we created a synthetic dataset that contains ambiguous entities that can only be understood through dialog context. For example, in the utterance "I like Harry Potter", the model needs to understand the context of the utterance to figure out if the user is referring to the movie or the book. We also randomly selected some conversations from Wizard of Wikipedia (WoW), which is a collection of open-domain dialogues grounded on Wikipedia knowledge (Dinan et al., 2018). The statistics of the datasets we used are shown in Table 1.

Dataset	Train	Validation	Test
MOC	5,962	662	1,111
Synthetic	8,150	905	2,896
WoW	1,948	216	540

Table 1: Number of utterances of the open-domain conversation data sets used in this study.

4.2 Model Setup

All models are implemented in Pytorch (Paszke et al., 2017). For the NER model, we initialize the word embedding with stacked embeddings, including Flair embeddings (Akbik et al., 2018) and FastText embeddings (Bojanowski et al., 2017). The sizes of the word embeddings and hidden state are 300 and 256, respectively. We adopt the SGD optimizer with an initial learning rate of 0.1 and decay rate of 0.5. The batch size is set to 16 and the maximum training epoch is set to 15 with an early stopping strategy. For the ER model, we use Adam as the optimizer and set the learning rate to 0.0005. The hidden size is 762 and the batch size is 8. The maximum sentence length in all the experiments is set to 128.

5 Results and Analysis

5.1 NER Results

The performance of the NER models is evaluated using precision, recall and F-1. We consider both the span of an entity and its type. Table 2 shows the results of NER models on three datasets. To compare with our context-aware NER model, we use Flair as the baseline, which is a state-of-the-art NER model on benchmarks in several domains (Akbik et al., 2018). It shows that our context-aware model achieves the best performance on most metrics. In particular, we observe the largest gain of our model using contextual information on the synthetic dataset. This is because that data was created to contain more ambiguous entities and thus requires dialog context to determine entity types.

Model	Dataset	P	R	F-1
Flair	MOC	-	-	-
Flair w/ context		- 0.1	2.7	1.2
Flair	Synthetic	-	-	-
Flair w/ context		16.0	17.7	16.9
Flair	WoW	-	-	-
Flair w/ context		0.7	1.8	1.2

Table 2: Results of NER models (relative gains compared to Flair in %).

5.2 ER Results

For the ER task, we evaluate the recall@n values (n = 1, 3, 5), which measures the ranking ability of the models. We compare our model with the following two baselines:

Search. After performing entity retrieval through Elasticsearch, we rank the candidate entities based on their popularity, i.e., the number of views in last 60 days.

Ranking. Similar to our method, here we only use entity and utterance-level matching scores, without dialog context in the ranking model.

Table 3 shows the ER results when ground-truth NER is provided as input. We can see that a ranking model can significantly improve the top entity relevance over the search baseline on all the three datasets. Compared to the non-context ranking model, our proposed context-aware model could further improve the results, especially for R@1.

Model	Dataset	R@1	R@3	R@5
Search	MOC	-	-	-
Rank		64.5	29.4	2.8
Rank w/ context		**65.0**	**29.7**	**2.9**
Search	Synthetic	-	-	-
Rank		82.9	**22.2**	9.4
Rank w/ context		**91.0**	21.9	**10.0**
Search	WoW	-	-	-
Rank		82.1	28.8	**11.2**
Rank w/ context		**89.1**	**29.2**	**11.2**

Table 3: Results of ER models (relative gains compared to baseline search in %) using ground-truth NER information.

5.3 End-to-end NEL Results

In Section 5.2, the input of the ER task is the ground-truth NER results. In the practical scenario, the input is the prediction of the NER models.

Context	Utterance	Model	NER	ER	Entity Description
In the 1968 three of the genre most famous acts Led Zeppelin, Black Sabbath	I love led Zeppelin! they have really influenced many bands.	w/o context Search	led Zeppelin, person	led Zeppelin, band	English rock band
		w/o context Rank	led Zeppelin, person	Jason Bonham, human	English hard rock drummer (born 1966)
		w/ context Rank	led Zeppelin, person	led Zeppelin, band	English rock band
		Groundtruth	led Zeppelin, person	led Zeppelin, band	English rock band
Well, So what gaming platform do you prefer console or computer?	Uh I don't know what a Nintendo WII is	w/o context Search	Nintendo, device	Nintendo Switch, hybrid video game console	hybrid video game console developed by Nintendo
		w/o context Rank	Nintendo, device	Nintendo, business	Japanese multinational video game and consumer electronics company
		w/ context Rank	Nintendo WII, device	Wii, home video game console	seventh-generation home video game console by Nintendo
		Groundtruth	Nintendo WII, device	Wii, home video game console	seventh-generation home video game console by Nintendo

Table 4: Examples of NEL in open-domain conversations.

Therefore, we also evaluate the performance of end-to-end NEL, where the predictions of NER models are used for ER. For performance metrics, we compare the predicted entity with the ground-truth one, and compute precision, recall and F-1. The results are shown in Table 5. Here we observe again that a ranking model can significantly improve results, and the context model yields further gain.

NER	ER	Dataset	P	R	F-1
Flair	Search	MOC	-	-	-
Flair	Rank		62.1	59.7	60.6
w/ context	w/ context		**62.9**	**63.4**	**62.8**
Flair	Search	Synthetic	-	-	-
Flair	Rank		68.2	68.9	68.4
w/ context	w/ context		**71.0**	**71.0**	**71.0**
Flair	Search	WoW	-	-	-
Flair	Rank		62.0	62.0	62.1
w/ context	w/ context		**75.4**	**72.8**	**73.9**

Table 5: End-to-end experimental results (relative gains in % compared to the end-to-end model of Flair NER and baseline ER search).

5.4 Case Study

Table 4 shows NER and ER results for two example utterances along with their context. We can see when there is an ambiguity in the current utterance, our context-aware model can use context information to correctly recognize the entities and link them to the right entities in KB. In the first example, the named entity is correctly recognized by all the models, however, the model without context failed in the ER task because of insufficient information. In the second case, models without using context information recognize a wrong entity and then link it to a seemingly reasonable but not the most appropriate entity.

5.5 Response Generation Results

We generate outputs for 100 distinct conversational contexts in the WoW data set using using configurations: Baseline GPT2 and GPT2 with NEL. Here, we provide crowd-worker annotators the conversational context along with the generated response, without the associated knowledge extracted through linking. We then ask the workers to evaluate according to two metrics, appropriateness and informativeness, on an ordinal scale from 0-2.

Our results show that in the generated responses, GPT2 with NEL module is superior over baseline GPT2 on both the appropriateness and informativeness metrics, suggesting that our solution can better understand conversation context and is able to generate informative and appropriate responses.

Model	Appropr.	Inform.
GPT2	-	-
GPT2 w/ NEL	25.5	53.8

Table 6: Human evaluation of generated responses. (%, relative gains compared to GPT2)

6 Conclusion

In this paper, we investigate NEL in multi-turn open-domain conversations. Considering the characteristic of dialogs, where the meaning of the cur-

rent utterance often varies depending on the context, we design a context-aware NER model and an ER model. Experimental results on three datasets prove that using context information improves the entity recognition and resolution performance. Extrinsic evaluation on response generation also validates the effectiveness of the entity information.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*.

Abbas Ghaddar and Philippe Langlais. 2018. Robust lexical features for improved neural network named-entity recognition. *arXiv preprint arXiv:1806.03489*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Karthik Gopalakrishnan, Behnam Hedayatnia, Longshaokan Wang, Yang Liu, and Dilek Hakkani-Tür. 2020. Are Neural Open-Domain Dialog Systems Robust to Speech Recognition Errors in the Dialog History? An Empirical Study. In *INTERSPEECH*.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. " O'Reilly Media, Inc.".

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Yufan Jiang, Chi Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2019. Improved differentiable architecture search for language modeling and named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3576–3581.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Refnet: A reference-aware network for background based conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8496–8503.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019a. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, 45(1):163–197.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. *arXiv preprint arXiv:1601.01343*.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. *arXiv preprint arXiv:2010.08824*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.