Evaluation of Conversational Skills for Commonsense

Baymurzina Dilyara
dilyara.baymurzina@phystech.edu

Kuratov Yurii
yurii.kuratov@phystech.edu

Kuznetsov Denis
kuznetsov.dp@phystech.edu

Kornev Daniel
danielko@deeppavlov.ai

Burtsev Mikhail
burtcev.ms@mipt.ru
Moscow Institute of Physics and Technology
Dolgoprudny, Russia

Abstract

Lack of commonsense is one of the most challenging problems in the field of conversational AI. Despite the recent significant progress in NLP driven by pre-trained language models, commonsense reasoning is still out of reach. We propose an approach to evaluate conversational commonsense usage. We use the approach to evaluate conversational skills of the socialbot during interaction with users. Analysis of data with joint manual and automatic annotations allowed us to identify automatic metrics tied to commonsense.We also develop two commonsense conversational skills that combine commonsense knowledge graphs completion model COMeT [6] and template-based approach.

Keywords: commonsense, common sense, conversational skills, commonsense evaluation
DOI: 10.28995/2075-7182-2021-20-1002-1011

Здравый смысл и его оценка в разговорных навыках диалоговых систем

Баймурзина Диляра
dilyara.baymurzina@phystech.edu

Куратов Юрий
yurii.kuratov@phystech.edu

Кузнецов Денис
kuznetsov.dp@phystech.edu

Корнев Данила
danielko@deeppavlov.ai

Бурцев Михаил
burtcev.ms@mipt.ru

Московский физико-технический институт
Долгопрудный, Россия

Аннотация

Одной из главных сложностей разработки разговорного искуственного интеллекта является внедрение здравого смысла. Несмотря на значительный прогресс в решении задач обработки естественного языка за счет использования предобученных моделей, диалоговые системы все еще не в состоянии рассуждать на основе здравого смысла. Мы предлагаем подход для оценки выраженности здравого смысла в диалоговых системах, сравнивая различные разговорные навыки в контексте диалога. Анализ диалогов с ручной и автоматической разметкой позволяет определить, какие из автоматических метрик коррелируют с проявлением здравого смысла. Мы также представляем два разговорных навыка, использующих модели предсказания по графу знаний здравого смысла для заполнения шаблонных реплик.

Ключевые слова: здравый смысл, оценка здравого смысла, диалоговые системы, разговорные навыки

1 Introduction

Modern pre-trained language models [10, 12, 4] significantly boosted scores on many natural language understanding benchmarks including ones designed to be especially hard for machines (for example WSC [16]). In spite of that, current state of the art solutions still fail to demonstrate robust commonsense reasoning in many cases as shown in [19] and [18]. There are several new tasks and datasets focused on evaluation of commonsense reasoning like WinoGrande [27] and ART [1]. These tasks are simple for humans (94% for WinoGrande and 91.4% for ART) but are still hard for the modern NLP systems employing large pre-trained language models.

This work explores possibilities of integrating commonsense knowledge in human-bot dialogues. Although there has not yet been any generally accepted formal definition of commonsense knowledge, these characteristics of commonsense knowledge are identified in [32]: (1) *sharing* - commonsense knowledge is possessed and shared by people; (2) *fundamentality* - people understand commonsense knowledge so well they take it for granted; (3) *implicitness* - usually people do not talk or write full commonsense knowledge explicitly since others also know it, and short reference is enough in many cases; (4) *large-scale* - commonsense knowledge has a tremendously large scale in both amount and diversity; (5) *open-domain* - commonsense knowledge covers all aspects of our daily life rather than a specific domain, and (6) *default* - commonsense knowledge are default assumptions about typical cases in everyday life, so most of them are feasible rather than definitely correct. In this paper we consider commonsense knowledge based on these characteristics. However, this definition is more characteristic for human-to-human conversations, and the human perception of the collocator's commonsense in a human-to-bot dialog differs significantly. The real users often doubt the socialbot's commonsense knowledge, and try to find proofs during the conversation. Therefore, in spite of the *implicitness* characteristic, we distinguish explicit commonsense demonstration as direct utilization of the knowledge in the responses (e.g. causal relationships expressions or utilization of object properties). Moreover, we propose conversational skills which aims to explicitly demonstrate commonsense knowledge as much as possible to convince the user of the socialbot's commonsense understanding. We define commonsense at the level of single utterance and at the level of the dialogue. The system is expected to produce utterances that are in agreement with current context, relevant and consistent to the dialogue history. More details on proposed definition of commonsense and overview of the data annotation is given in Section 3.

Modern dialogue systems usually combine some or all of the three most common approaches: (1) rule-based (e.g., slot-based scripted scenarios), (2) retrieval (e.g., tf-idf, ConveRT [8]), and (3) generative (e.g., seq2seq [25]). The rule-based conversational skills have an advantage of explicit incorporation of the partial commonsense knowledge in the form of the pre-defined scripts. This makes possible to produce responses consistent with the commonsense but only for a small subset of the scripted situations in the dialogue.

Since almost every dialogue on common topics goes beyond scripted situations an undesired system behavior might happen frequently. Unlike a rule-based approach, generative and retrieval models can not guarantee semantic coherence and non-contradiction of the response utterance to the dialogue history. On the other hand, they are not limited by the number of the scripted situations where they can give plausible responses. The middle ground between rigid scripts and too flexible data driven dialogue models lies in the combination of high quality commonsense data accumulated in the knowledge bases and trainable models to fit relevant knowledge into the dialogue context.

In this work, we use ATOMIC [2] and ConceptNet [24] knowledge graphs completion model COMeT [6] to build two commonsense conversational skills. These skills combine rule-based and knowledge graph-based methods. They are developed to express and argue about the system's opinions, as well as to demonstrate rudimentary understanding of causal relationships.

Commonsense Conversational skills were created for the socialbot that participated in the Alexa Prize 2019 competition. The socialbot is an open domain dialogue system that should be capable to converse on the topic of user's choice. Our socialbot is implemented with open-source DeepPavlov Agent framework[1] and consists of multiple skills. The high-level dialogue logic is orchestrated by a Skill Selector,

---

[1] `https://github.com/deepmipt/dp-agent`

which chooses a subset of skills to produce response hypotheses, and a Response Selector, which defines what final response should be sent to the user. Participation in the competition provided us with a unique opportunity to test our solution with a large pool of Amazon Alexa users.

The main contributions of this paper are: (1) we propose an annotation scheme to capture commonsense in a dialogue; (2) we analyze automatic metrics for their relevance for commonsense recognition and evaluate a number of retrieval and rule-based skills for the commonsense; (3) we implement two commonsense conversational skills that rely on knowledge graphs completion model COMeT.

2 Conversational Skills

In this section we describe proposed Commonsense Conversational skills as well as some other skills selected for comparison. The socialbot itself has multi-skill structure, and the final response is decided by Response Selector with the help of hand-written heuristics and empirical formula that combines skills' confidences and CoBot Conversation Evaluator [26] scores (described in Subsection 4). One can find more details of the socialbot structure in [9].

2.1 Proposed Commonsense Conversational Skills

Common casual conversation often contains mentions of different activities and their discussion. Although some of these activities can be discussed throughout the focus on the subject (e.g. "play piano", "study geography"), others are not directly related to some rare enough subjects (e.g. "go swimming", "get tired"), and, finally, some are composed by verbs and subjects together (e.g. "train a dog", "hangout with friends"). Here comes a motivation to build following commonsense conversational skills: (1) able to keep the conversation on wide variety of human activities; (2) able to talk about human activities not only based on information from knowledge graphs but also in terms of feelings, motivation, consequences. All these aspects can be summarized as exercise of commonsense, and thanks to COMeT commonsense prediction models [6] can be retrieved for all action expressions. For both skills, target activities are extracted using a complex method including morphological analysis, entity extraction, words and n-grams frequencies. Entities for expressing socialbot's personal opinion are extracted using either morphological analysis or named entity recognition modules [9].

`Activity Discussion Skill` expresses *motivation of the socialbot to better understand human world.* The skill enlists user's help for an explanation of the common human activities which can be either extracted from the user's utterance or picked up from the predefined list of different hobbies. The socialbot is making assumptions about characteristics of different human activities and causal relationships which demonstrates the socialbot's understanding of commonsense knowledge but still leaves a room for mistakes. Discussion of activities consists of an explanation request, several clarification questions, and a user's opinion request. The skill has template-based replies with *causal relationships* and *understanding of human emotions* slots which are filled by COMeT ATOMIC model [6]. For example, for the activity *"practice yoga"* the COMeT ATOMIC model generates *"to be healthy"* as an "xIntent" aspect. Therefore, the skill can produce a response *"Do humans practice yoga [to be healthy]?"* filling slots with the activity name and predicted intent. The example of the full dialog with the skill is presented in Figure 1.

`Personal Event Discussion Skill` has two modes: (1) expressing opinion on different objects, and (2) discussing user's actions to demonstrate the ability to trace causal relationships and to establish emotional common ground with the user. First mode enables reacting to user's activities in terms of intents, feelings, effects, and consequences using COMeT ATOMIC [6]. For example, if user says *"I will go to the theater next weekend."*, the skill extracts action "go to the theater" taking into account future tense, then a follow up question might be composed in the following way *"Did you [buy the tickets]?"* filling a template with a prediction of the COMeT ATOMIC model on an "xNeed" aspect. The second mode of `Personal Event Discussion Skill` is designed to express a reasoned opinion on a wide variety of objects (not necessarily events) depending on the sentiment of the selected commonsense assertion from COMeT ConceptNet [6]. For example, for the requested opinion on cats, the skill fills "SymbolOf" slot in template-based response with the prediction of the COMeT ConceptNet model "love", and produces the following: *"I love cats so much because cat is a sign of [love]."*. The example of the dialog with the both modes of the skill is presented in Figure 1.

Human: I've *jumped with a parachute*!
Bot: What does *jumping with a parachute* mean? [explanation request]
Human: Yeah. It's a kind of risky activity.
Bot: Is it feels [scared] (xAttr) to *jump with a parachute*? [clarification]
Human: Definitely.
Bot: Why someone would need [to buy a parachute] (xNeed)? [clarification]
Human: No, you may rent it.
Bot: Great! Thank you. So, did you like *jumping with a parachute*?

Human: Do you like *musicals*?
Bot: Well, I like *musicals*. Makes me want [to learn to sing] (CausesDesire). [opinion]
Human: Cool. I *went to the cinema* to see it on weekends.
Bot: Well. Did you want [to be entertained] (xWant) when *went to the cinema*? [clarification]
Human: Kind of.
Bot: So, hope you feel quite [excited] (xReact) now. [comment]
Human: Aha. Thanks.

Figure 1: Dialog Samples with `Activity Discussion Skill` (on the left) and `Personal Event Discussion Skill` (on the right). Dialogues are not real, templated utterances are also rewritten.

2.2 Other Skills

Artificial Intelligence Markup Language (AIML) is an XML dialect for creating dialogue agents. AIML is well-documented, widely used and easy to start language to implement chatbots. In our system we use Program Y[2] framework. `AIML Chit-Chat` is based on Template-y bot[3] that already implements various rules. We curated and updated these rules to add a greeting scenario, several general conversation templates, as well as jokes.

`Alice` is an open-source AIML chatbot[4]. It has a comprehensive set of grammars and was especially helpful at the beginning of the competition.

`Movie Skill` provides responses expressing the socialbot's movie preferences. The skill produces scripted dialogue focused on a movie extracted from the user utterance. If `Movie Skill` detects a non-popular in terms of number of votes on IMDb movie title, skill clarifies whether the extracted title is correct. The movie dialogue script includes opinion exchange, questions about movie genre, cast or characters, and some interesting facts about discussed movie.

`CoBotQA` answers factoid questions as well as provides facts about extracted noun phrases and named entities for "fact about" requests. It is implemented on top of the remote Q&A CoBot service which works with a plain text. Output from Q&A CoBot service is limited to 1-2 sentences and augmented with small opinion-like phrases. Q&A CoBot service itself can also give conversational responses.

`ConveRT Reddit` Retrieval Skill uses a ConveRT [8] encoder to build efficient representations for sentences. The model retrieves candidate responses by ranking response-context pairs by cosine similarity of the corresponding embeddings. Context is created by concatenation of utterances in the dialogue history. Model was pre-trained on 654M input-response pairs. We fine-tuned it on 80k Reddit comments that were used as a retrieval base for the skill.

3 User experience analysis

We introduce a labelling scheme to capture commonsense in a dialogue. Commonsense knowledge is usually categorized as being implicit, e.g., usually people do not talk or write commonsense knowledge explicitly since others also know it. Moreover, commonsense knowledge is possessed and shared by people in their conversations, and is seen as default assumptions about typical cases in everyday life.

---

[2] `github.com/keiffster/program-y`

[3] `github.com/keiffster/program-y/wiki/Available-Bots`

[4] `github.com/sld/convai-bot-1337/tree/master/ALICEChatAPI`

Therefore, we see it useful to differentiate use of commonsense in a casual dialogue by the level of its usage.

Different skills composing the socialbot can generate responses that demonstrate or rely on commonsense knowledge but do not fit the context. In this case, from the human interlocutor's point of view the socialbot either contradicts the commonsense or gives a meaningless response. Therefore, we distinguish *the phrase* and the *context level* of commonsense presence. For the both *phrase* and dialogue *context* levels we differentiate responses as expressing *explicit commonsense*, *implicit commonsense*, *no sense*, or being *undefined*.

We identify utterance as an *explicit commonsense* if it explicitly contains statements that reiterate commonsense knowledge or default assumptions about typical cases in everyday life. For example, *"It's rainy outside, don't forget an umbrella"*, or *"It feels so magical to see unicorns in a dream"*, or *"You can pet a cat"*. In all of these cases default assumptions are explicitly stated within the utterance. We classify meaningful statements that do not state default assumptions but are referencing them as an *implicit commonsense*. For example, phrase *"I like unicorns"* does not reiterate default assumptions (like *"unicorns don't exist"*) but it is meaningful to both interlocutors because of the shared commonsense knowledge (e.g., a concept of Unicorns that only exists in a fairy tales world). The *no sense* class includes meaningless responses and phrases that contradict commonsense. For example, *"I like braavawqera"* is meaningless cause while phrase explicitly states interlocutor's relation to "braavawqera", still "braavawqera" is not a part of a dialogue context nor commonsense knowledge. For the cases when it is not possible to determine whether commonsense knowledge is required to give the answer we use the *undefined* class.

The socialbot reply is considered as expressing *explicit commonsense in a context* if it relates to the context, and includes commonsense on the phrase level or complements the context to evidently express commonsense knowledge. Therefore, for a context *"What do you think about unicorns?"*, both responses *"They are unreal"* or *"Unicorns are unreal"* express explicit commonsense. While the last utterance expresses explicit commonsense on phrase level, the utterance "They are unreal" is related to implicit commonsense on phrase level but when taking into account the context, and the fact that "they" implies "unicorns", we can classify this case as explicit commonsense in context. Another examples illustrating the case when the socialbot utterance complements the context to the explicit commonsense are following *"What is the color of the sky?"* - *"It's blue"* and *"I studied history in college"* - *"You have to be very smart"*. *Implicit commonsense in context* corresponds to the response which is appropriate but do not include explicit commonsense knowledge or reasoning by itself, e.g. *"What do you think about unicorns?"* - *"I like them"*. If the response does not fit the context or contradicts commonsense as is or within the context, we consider it to be *no sense* in the context. Class *undefined* in context corresponds to cases when even the context can't help to understand whether the response was meaningful or not.

We have collected 100 samples per each skill from Section 2. Three assessors annotated each sample with 2 labels: one for the phrase, and another for the context level commonsense. Our resulting dataset is a combined set of these annotations, $100 \times 7 \times 3$ contexts each of which has 2 labels (4200 annotated samples in total). Inter-annotator agreement Kappa is 0.414.

4 Automatic Metrics

Assume that the following automatic metrics can be useful for commonsense detection: sentiment and toxicity of user response, conversation evaluation and natural language inference predictions on the socialbot response.

Sentiment classifier indicates if user utterance is positive, negative, or neutral. The classifier on top of English Conversational BERT[5]was trained on Stanford Sentiment Treebank dataset [21] with five classes: very positive, positive, neutral, negative, and very negative. During inference, very positive (negative) labels are assigned to positive (negative).

Toxic classifier identifies whether an utterance contains insults, threats, obscene words, identity hate, sexual explicit talk, or other toxicity manifestations. The classification head on top of English Conversational BERT-model[5] was trained on Kaggle Toxic Comment Classification Challenge[6] dataset.

Response Selector in the socialbot chooses the final response using skills confidences and conversation evaluation scores by CoBot. CoBot Conversation Evaluator is trained on the Alexa Prize data from previous competitions and predicts whether candidate response is interesting, comprehensible, on-topic, engaging and erroneous [26]. CoBot Conversation Evaluator was provided to participants as an existing remote service.

We also used AllenNLP Textual Entailment models[7] based on RoBERTa [22] to get MNLI [29] and SNLI [31] annotations.

5 Results

We show resulting distributions of commonsense types for considered skills on phrase and context levels in Figure 2 and Figure 3. Knowledge-based `CobotQA` and `Movie Skill` have the highest fraction of *explicit commonsense* as expected. General domain skills `Alice` and `AIML Chit-Chat` have a prevalence of *implicit commonsense*. `Activity Discussion Skill` is similar to retrieval skill. `Personal Event Discussion Skill` is the only one among all the considered skills which has higher fraction of *explicit commonsense* on context level than phrase's one. This observation illustrates that compared to others `Personal Event Discussion Skill` provides the largest number of relevant answers that might lacking explicit commonsense on phrase level, still complement the context to explicit demonstration of commonsense.

In Figure 4 we present correlation of the commonsense metrics and automatic metrics. Meaningless contexts (*no sense* and *undefined* in context labels) are well characterized by high level of toxicity of user responses, low "Comprehensible" score by CoBot Conversation Evaluator [26], and neutral in terms of MNLI labels. *No sense* phrases are not comprehensible while *no sense* in context responses correspond to "Neutral" in terms of NLI.

Positive sentiment of user reaction is inherent for *explicit commonsense* on both phrase and context levels while also has significant negative correlation with *no sense* in context responses.

CoBot conversation evaluation parameters such as "OnTopic", "Interesting", and "Erroneous" can be used for differentiation of *explicit* and *implicit commonsense* on both phrase and context level. "Entailment" in terms of NLI also distinguishes *explicit* and *implicit commonsense* on context level because *explicit commonsense* implies bringing some new information in responses. Responses with *implicit commonsense* could be well characterized as not-including commonsense, but that are relevant to the context. It is reflected in positive correlation with "Entailment" and "Contradiction".

6 Related Work

There are several approaches for integrating commonsense knowledge into conversational systems. Some aspects of commonsense could be added to generative neural networks by conditioning them on external information like dialogue acts [23], persona [20], and dialogue features (e.g., simplicity, staying on topic, asking questions) [28]. Neural response models could be conditioned on information extracted from knowledge bases [5, 15, 30] or textual sources [13] to force model to use this knowledge in its answers. Some works are especially focused on the usage of commonsense KBs as a source of external information to improve retrieval [3] and generative [7] neural response models. Another way to control dialogue consistency, which is an integral part of commonsense in a dialogue setting, is to use models trained on Dialogue NLI dataset to re-rank response candidates [11].

Half-year after the original paper of current authors [9] was published describing the conversational skills utilizing COMeT commonsense prediction models, the article [17] presented an approach of expansion of socialbot's persona sentences using commonsense knowledge bases. The utilized commonsense improved generation of persona grounded responses.

---

[5] `http://docs.deeppavlov.ai/en/master/features/pretrained\_vectors.html\#downloads`

[6] `kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview`

[7] `demo.allennlp.org/textual-entailment`

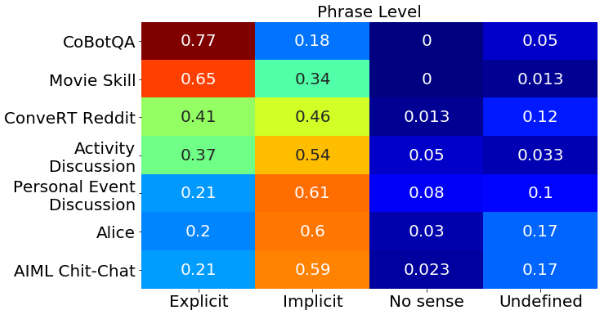

Figure 2: Phrase level distribution of commonsense types for different skills.

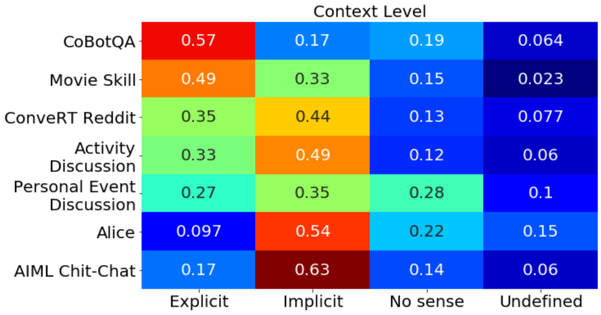

Figure 3: Context level distribution of commonsense types for different skills.

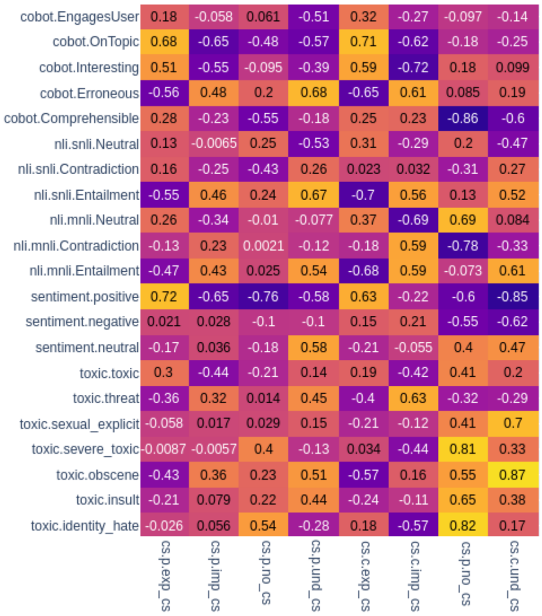

Figure 4: Correlation map of commonsense (phrase *cs.p* and context *cs.c* levels, *exp*licit and *imp*licit, *und*efined, *no* common*s*ense) and automatic metrics: sentiment, toxicity, NLI, and CoBot Conversation Evaluator.

We evaluated commonsense by manual annotation of dialogues. There have been attempts to automatically evaluate already present commonsense knowledge in pre-trained language models on the task of prediction masked objects, which are part of commonsense triplets within the sentence [14].

7 Conclusion

We propose annotation scheme of commonsense types for phrase and context levels. Although we use manual labelling of commonsense in this study, it is not appropriate for the production system, so we also analysed correlations of commonsense in the annotated data with the following automatic metrics: sentiment classification, toxicity detection, natural language inference, and CoBot conversation evaluation. We found out that some of the metrics correlate with commonsense: (1) "Comprehensible" by CoBot conversation evaluation, "Neutral" in terms of MNLI, and toxicity level of user's reaction can help to determine replies which are contrary to commonsense or do not fit the context; (2) Positive sentiment of user reaction is inherent for *explicit commonsense* on both phrase and context level while also have significant negative correlation with meaningless in context responses; (3) CoBot conversation evaluation ”OnTopic”, ”Interesting” and ”Erroneous” parameters, "Entailment" and "Contradiction" in terms of NLI are helpful to distinguish explicit and implicit commonsense types.

We implemented two different conversational skills combining commonsense KGs completion models and template-based approach. These skills show higher number of explicit commonsense than rule-based open domain skills while their implicit commonsense is at the same level as for a retrieval one. One of the proposed skills has the highest number of responses without explicit commonsense on phrase level and with explicit commonsense in context, while also has the highest number of meaningless responses. It means that responses are tightly tied with context, but may fail because of misused templates or KGs completion models erroneous predictions.

References

[1] Abductive Commonsense Reasoning / Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya et al. // International Conference on Learning Representations. — 2020. — Access mode: `https://openreview.net/forum?id=Byg1v1HKDB`.

[2] Atomic: An atlas of machine commonsense for if-then reasoning / Maarten Sap, Ronan Le Bras, Emily Allaway et al. // Proceedings of the AAAI Conference on Artificial Intelligence. — Vol. 33. — 2019. — P. 3027–3035.

[3] Augmenting end-to-end dialogue systems with commonsense knowledge / Tom Young, Erik Cambria, Iti Chaturvedi et al. // Thirty-Second AAAI Conference on Artificial Intelligence. — 2018.

[4] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: `https://www.aclweb.org/anthology/N19-1423`.

[5] Bordes Antoine, Boureau Y-Lan, Weston Jason. Learning end-to-end goal-oriented dialog // arXiv preprint arXiv:1605.07683. — 2016.

[6] COMET: Commonsense Transformers for Automatic Knowledge Graph Construction / Antoine Bosselut, Hannah Rashkin, Maarten Sap et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4762–4779. — Access mode: `https://www.aclweb.org/anthology/P19-1470`.

[7] Commonsense Knowledge Aware Conversation Generation with Graph Attention. / Hao Zhou, Tom Young, Minlie Huang et al. // IJCAI. — 2018. — P. 4623–4629.

[8] ConveRT: Efficient and Accurate Conversational Representations from Transformers / Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić et al. // arXiv preprint arXiv:1911.03688. — 2019.

[9] DREAM technical report for the Alexa Prize 2019 / Yuri Kuratov, Idris Yusupov, Dilyara Baymurzina et al. // Alexa Prize Proceedings. — 2020.

[10] Deep Contextualized Word Representations / Matthew Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — 2018. — P. 2227–2237.

[11] Dialogue Natural Language Inference / Sean Welleck, Jason Weston, Arthur Szlam, Kyunghyun Cho // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — P. 3731–3741.

[12] Improving language understanding by generative pre-training / Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever // URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf. — 2018.

[13] Incorporating Unstructured Textual Knowledge Sources into Neural Dialogue Systems / Ryan Lowe, Nissan Pow, Laurent Charlin et al. // Machine Learning for Spoken Language Understanding and Interaction, NIPS 2015 Workshop. — 2015. — Access mode: `http://slunips2015.wixsite.com/slunips2015/accepted-papers`.

[14] Language Models as Knowledge Bases? / Fabio Petroni, Tim Rocktäschel, Sebastian Riedel et al. // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 2463–2473. — Access mode: `https://www.aclweb.org/anthology/D19-1250`.

[15] Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings / He He, Anusha Balakrishnan, Mihail Eric, Percy Liang // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2017. — P. 1766–1776.

[16] Levesque Hector, Davis Ernest, Morgenstern Leora. The winograd schema challenge // Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning. — 2012.

[17] Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions / Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley // arXiv preprint arXiv:2010.03205. — 2020.

[18] Marcus Gary. GPT-2 and the Nature of Intelligence // The Gradient. — 2020.

[19] Niven Timothy, Kao Hung-Yu. Probing Neural Network Comprehension of Natural Language Arguments // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4658–4664. — Access mode: `https://www.aclweb.org/anthology/P19-1459`.

[20] Personalizing Dialogue Agents: I have a dog, do you have pets too? / Saizheng Zhang, Emily Dinan, Jack Urbanek et al. // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2018. — P. 2204–2213.

[21] Recursive deep models for semantic compositionality over a sentiment treebank / Richard Socher, Alex Perelygin, Jean Wu et al. // Proceedings of the 2013 conference on empirical methods in natural language processing. — 2013. — P. 1631–1642.

[22] RoBERTa: A robustly optimized BERT pretraining approach. arXiv 2019 / Y Liu, M Ott, N Goyal et al. // arXiv preprint arXiv:1907.11692.

[23] Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems / Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić et al. // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — Lisbon, Portugal : Association for Computational Linguistics, 2015. — Sep. — P. 1711–1721. — Access mode: `https://www.aclweb.org/anthology/D15-1199`.

[24] Speer Robyn, Chin Joshua, Havasi Catherine. Conceptnet 5.5: An open multilingual graph of general knowledge // Thirty-First AAAI Conference on Artificial Intelligence. — 2017.

[25] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. — 2014. — P. 3104–3112.

[26] Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators / Sanghyun Yi, Rahul Goel, Chandra Khatri et al. // Proceedings of the 12th International Conference on Natural Language Generation. — Tokyo, Japan : Association for Computational Linguistics, 2019. — Oct.–Nov. — P. 65–75. — Access mode: `https://www.aclweb.org/anthology/W19-8608`.

[27] WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale / Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, Yejin Choi // ArXiv. — 2019. — Vol. abs/1907.10641.

[28] What makes a good conversation? How controllable attributes affect human judgments / Abigail See, Stephen Roller, Douwe Kiela, Jason Weston // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — 2019. — P. 1702–1723.

[29] Williams Adina, Nangia Nikita, Bowman Samuel R. A broad-coverage challenge corpus for sentence understanding through inference // arXiv preprint arXiv:1704.05426. — 2017.

[30] A knowledge-grounded neural conversation model / Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang et al. // Thirty-Second AAAI Conference on Artificial Intelligence. — 2018.

[31] A large annotated corpus for learning natural language inference / Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D. Manning // Proceedings of the 2015 Conference on Em-

pirical Methods in Natural Language Processing (EMNLP). — Association for Computational Linguistics, 2015.

[32] A survey of commonsense knowledge acquisition / Liang-Jun Zang, Cong Cao, Ya-Nan Cao et al. // Journal of Computer Science and Technology. — 2013. — Vol. 28, no. 4. — P. 689–719.