

Evaluation of Conversational Skills for Commonsense

Baymurzina Dilyara Kuratov Yurii Kuznetsov Denis
Kornev Daniel Burtsev Mikhail

Moscow Institute of Physics and Technology

June 2021

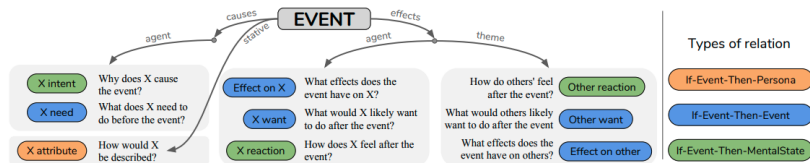
Commonsense definition

- *sharing* - commonsense knowledge is possessed and shared by people;
- *fundamentality* - people understand commonsense knowledge so well they take it for granted;
- *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;
- *large-scale* - commonsense knowledge has a tremendously large scale in both amount and diversity;
- *open-domain* - commonsense knowledge covers all aspects of our daily life rather than a specific domain;
- *default* - commonsense knowledge are default assumptions about typical cases in everyday life, so most of them are feasible rather than definitely correct.

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.

ATOMIC¹ Knowledge Graph

We use ATOMIC¹ KG completion model COMeT².



ATOMIC¹ Knowledge Graph (image from the paper).

[1] 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.

[2] 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>.

ConceptNet¹ Knowledge Graph

We also use ConceptNet¹ KG completion model COMeT².

The screenshot shows the ConceptNet interface for the term 'people'. At the top left is the 'en people' logo with the text 'An English term in ConceptNet 5.8'. Below this is a source attribution: 'Sources: Open Mind Common Sense contributors, DBpedia 2015, Verbosity players, German Wiktionary, English Wiktionary, French Wiktionary, and Open Multilingual WordNet. View this term in the API'. On the right side, there are navigation links for 'Documentation', 'FAQ', 'Chat', and 'Blog'. The main content area is divided into four columns: 'people is capable of...', 'Properties of people', 'Related terms', and 'people has...'. Each column contains a list of related concepts, each with a small 'en' icon and a right-pointing arrow. For example, under 'people is capable of...', the items are 'talk to each other', 'believe in a god', 'catch colds', and 'forget things'. Under 'Properties of people', the items are 'stupid', 'ignorant', 'weird', and 'kind'. Under 'Related terms', the items are 'crowd', 'person', 'population', and 'party'. Under 'people has...', the items are 'feelings', 'dogs as pets', 'emotions', and 'brown hair'.

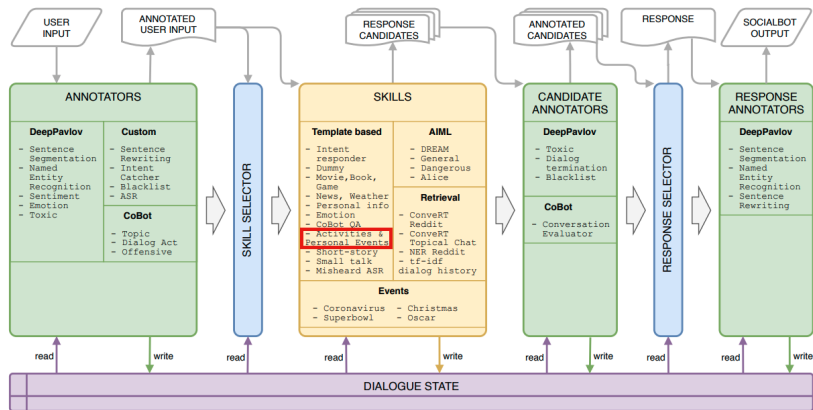
ConceptNet¹ KG demo¹.

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

[2] 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>.

¹<https://conceptnet.io/c/en/people>

DREAM in Alexa Prize Challenge



DREAM Socialbot Architecture, proposed commonsense skills are highlighted.

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

Activity Discussion Skill

PersonX *jumps with a parachute*. COMeT ATOMIC:

- xAttr: scared
- xNeed: to buy a parachute

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** to *jump with a parachute*? [clarification]

Human: Definitely.

Bot: Why someone would need **to buy a parachute**? [clarification]

Human: No, you may rent it.

Bot: Great! Thank you. So, did you like *jumping with a parachute*?

Personal Event Discussion Skill

Musicals. COMeT ConceptNet:

- CausesDesire: to learn to sing

PersonX *go to the cinema*. COMeT ATOMIC:

- xWant: to be entertained
- xReact: excited

Human: Do you like *musicals*?

Bot: Well, I like *musicals*. Makes me want to **learn to sing**. [opinion]

Human: Cool. I *went to the cinema* to see it on weekends.

Bot: Well. Did you want **to be entertained** when
went to the cinema? [clarification]

Human: Kind of.

Bot: So, hope you feel quite **excited** now. [comment]

Human: Aha. Thanks.

Skills to Compare with

- AIML Chit-Chat is based on Template-y bot¹.
- Alice is an open-source AIML chatbot².
- Movie Skill produces scripted dialogue focused on a movie extracted from the user utterance. The skill uses IMDb database.
- CoBotQA answers factoid questions as well as provides facts. The skill uses the remote Q&A CoBot service which works with a plain text.
- ConveRT Reddit Retrieval Skill uses a ConveRT³ encoder fine-tuned on 80k Reddit comments. The skill retrieves from those 80k Reddit comments.

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

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

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

Commonsense Markup

Level	Label	Example
Phrase Level	Explicit	It's rainy outside, don't forget an umbrella; You can pet a cat
	Implicit	I like unicorns
	No Sense	I like braavawqera
	Undefined	Good; Yes; Aha
Context Level	Explicit	- What do you think about unicorns? - They are unreal. (Unicorns are unreal.)
	Implicit	- What do you think about unicorns? - I like them.
	No Sense	- What do you think about unicorns? - Braavawqera. (Do you listen to music?)
	Undefined	- What do you think about unicorns? - Amazing. (Okay.)

Automatic metrics

- Stanford Sentiment Treebank based on English Conversational BERT¹.
- Toxic Classifier based on English Conversational BERT¹ trained on Kaggle Toxic Comment Classification Challenge² dataset.
- CoBot Conversation Evaluator is the remote service which predicts whether candidate response is interesting, comprehensible, on-topic, engaging and erroneous³.
- AllenNLP Textual Entailment models based on RoBERTa⁴ to get MNL⁵ and SNLI⁶ annotations.

[1] http://docs.deeppavlov.ai/en/master/features/pretrained_vectors.html#downloads

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

[3] Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators / Sanghyun Yi, Rahul Goel, Chandra Khatri et al. <https://www.aclweb.org/anthology/W19-8608>.

[4] demo.allennlp.org/textual-entailment

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

[6] A large annotated corpus for learning natural language inference / Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D. Manning // Proceedings of the 2015 EMNLP.

Commonsense Markup

We have annotated $100 \times 7 \times 3 = 4200$ contexts each of which has 2 labels – phrase and context level.

Inter-annotator agreement Kappa is 0.414.

	Phrase Level			
	Explicit	Implicit	No sense	Undefined
CoBotQA	0.77	0.18	0	0.05
Movie Skill	0.65	0.34	0	0.013
ConveRT Reddit	0.41	0.46	0.013	0.12
Activity Discussion	0.37	0.54	0.05	0.033
Personal Event Discussion	0.21	0.61	0.08	0.1
Alice	0.2	0.6	0.03	0.17
AIML Chit-Chat	0.21	0.59	0.023	0.17

Phrase level distribution of commonsense types for different skills.

	Context Level			
	Explicit	Implicit	No sense	Undefined
CoBotQA	0.57	0.17	0.19	0.064
Movie Skill	0.49	0.33	0.15	0.023
ConveRT Reddit	0.35	0.44	0.13	0.077
Activity Discussion	0.33	0.49	0.12	0.06
Personal Event Discussion	0.27	0.35	0.28	0.1
Alice	0.097	0.54	0.22	0.15
AIML Chit-Chat	0.17	0.63	0.14	0.06

Context level distribution of commonsense types for different skills.

Commonsense Markup Conclusions

- 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 *ConveRT Reddit* 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.

Correlation of Commonsense and Automatic Metrics

cobot.EngagesUser	0.18	-0.058	0.061	-0.51	0.32	-0.27	-0.097	-0.14
cobot.OnTopic	0.68	-0.65	-0.48	-0.57	0.71	-0.62	-0.18	-0.25
cobot.Interesting	0.51	-0.55	-0.095	-0.39	0.59	-0.72	0.18	0.099
cobot.Erroneous	-0.56	0.48	0.2	0.68	-0.65	0.61	0.085	0.19
cobot.Comprehensible	0.28	-0.23	-0.55	-0.18	0.25	0.23	-0.86	-0.6
nli.snli.Neutral	0.13	-0.0065	0.25	-0.53	0.31	-0.29	0.2	-0.47
nli.snli.Contradiction	0.16	-0.25	-0.43	0.26	0.023	0.032	-0.31	0.27
nli.snli.Entailment	-0.55	0.46	0.24	0.67	-0.7	0.56	0.13	0.52
nli.mnli.Neutral	0.26	-0.34	-0.01	-0.077	0.37	-0.69	0.69	0.084
nli.mnli.Contradiction	-0.13	0.23	0.0021	-0.12	-0.18	0.59	-0.78	-0.33
nli.mnli.Entailment	-0.47	0.43	0.025	0.54	-0.68	0.59	-0.073	0.61
sentiment.positive	0.72	-0.65	-0.76	-0.58	0.63	-0.22	-0.6	-0.85
sentiment.negative	0.021	0.028	-0.1	-0.1	0.15	0.21	-0.55	-0.62
sentiment.neutral	-0.17	0.036	-0.18	0.58	-0.21	-0.055	0.4	0.47
toxic.toxic	0.3	-0.44	-0.21	0.14	0.19	-0.42	0.41	0.2
toxic.threat	-0.36	0.32	0.014	0.45	-0.4	0.63	-0.32	-0.29
toxic.sexual_explicit	-0.058	0.017	0.029	0.15	-0.21	-0.12	0.41	0.7
toxic.severe_toxic	-0.0087	-0.0057	0.4	-0.13	0.034	-0.44	0.81	0.33
toxic.obscene	-0.43	0.36	0.23	0.51	-0.57	0.16	0.55	0.87
toxic.insult	-0.21	0.079	0.22	0.44	-0.24	-0.11	0.65	0.38
toxic.identity_hate	-0.026	0.056	0.54	-0.28	0.18	-0.57	0.82	0.17
	cs.p.exp_cs	cs.p.imp_cs	cs.p.no_cs	cs.p.und_cs	cs.c.exp_cs	cs.c.imp_cs	cs.p.no_cs	cs.c.und_cs

Conclusions on Correlation of Commonsense

- Meaningless contexts are well characterized by high level of toxicity of user responses, low "Comprehensible" score by CoBot Conversation Evaluator, and neutral in terms of MNLI labels.
- *No sense* phrases has low "Comprehensible" score by CoBot.
- Positive sentiment of user reaction is inherent for *explicit commonsense* on both phrase and context levels.
- Positive sentiment of user reaction 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.

Conclusion

- We propose annotation scheme of commonsense types for phrase and context levels.
- We 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 can be useful to determine presence and type of commonsense:
 - "Comprehensible "OnTopic", "Interesting" and "Erroneous" by CoBot;
 - "Neutral "Entailment"and "Contradiction" in terms of NLI,
 - toxicity level of user's reaction;
 - positive sentiment of user reaction.
- We implemented two different conversational skills combining commonsense KGs completion models and template-based approach.
 - with higher number of explicit commonsense than rule-based open domain skills;
 - one them has the highest number of responses without explicit commonsense on phrase level and with explicit commonsense in context.