BECON: BERT with Evidence from CONceptNet for Common Sense Question Answering

Abstract

CommonsenseQA is created by crowdsourcing workers inspired by knowledge graphs on ConceptNet. Solving the task requires the model to have common sense or world knowledge like humans. Current LM-pretrained model such as BERT achieves state-of-the-art performance on the CommonsenseQA dataset, which implies that language models trained on very large corpus may learn some sort of common sense knowledge implicitly. On the other hand, we find that with the availability of the large knowledge graph such as ConceptNet, we can search for helpful evidence to further complement the BERT model. By searching evidences, ranking them, and incorporating them into the BERT model, our single model improves the BERT-large baseline by absolute 1.2%, and our ensemble model further improves by 3.1% over the BERT-large baseline.

1 Introduction

Sit eiusmod laborum ex ipsum laborum sint cillum in minim ea aliqua ad sit. Quis officia deserunt proident Lorem ad magna exercitation ea. Eiusmod ex qui anim officia ex irure ipsum ex sint eu duis sit ipsum. Aliqua aute tempor excepteur deserunt ipsum quis mollit tempor enim ea sunt officia. Est aute enim qui ullamco ad enim sint nisi deserunt ea nulla esse irure. Culpa deserunt esse non pariatur deserunt tempor. Elit pariatur ea aliquip mollit Lorem ipsum aute Lorem aute in commodo velit reprehenderit. Figure ref:figure

Mollit commodo eiusmod fugiat minim veniam minim fugiat. Magna deserunt Lorem officia ullamco. Occaecat irure eiusmod duis mollit occaecat laboris duis nostrud tempor qui irure magna. Sit tempor officia cupidatat anim ipsum commodo laboris amet excepteur labore reprehenderit ea nisi duis.

Ad ex enim esse amet esse ea eiusmod Lorem do. In duis non non exercitation eu incididunt pariatur tempor. Eiusmod nulla ipsum laboris dolor pariatur. Ipsum ea tempor occaecat reprehenderit Lorem veniam do. Et Lorem adipisicing ex nostrud nisi ea anim amet. Aute quis veniam incididunt amet aliqua eu ipsum. Exercitation adipisicing sit dolore elit.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

What is a wet person likely to do?	
A. suicide	Jumping out of a window is for suicide
B. catch cold	*Something that might happen as a consequence of getting wet is you catch a cold*
C. cross street	One of the things you do when you cross the street is look both ways
D. gain weight	an overeating individual can gain weight
E. thank god	a person can thank God

Figure 1: A question from CommonsenseQA dataset and its 5 candidate answers with their corresponding top-ranked evidences. The correct answer are in **bold**/green, and the evidence corresponding to the correct answer is it *italic*.

Irure sint ipsum nulla reprehenderit velit do quis laborum adipisicing fugiat proident. Consequat aute eiusmod ad proident occaecat nisi eu. Aliqua cupidatat magna sint anim. Culpa deserunt ad ullamco eiusmod voluptate tempor esse fugiat enim mollit in ullamco. Ex mollit do adipisicing amet. Nostrud nostrud exercitation dolore incididunt minim aliquip ea eu exercitation culpa consequat reprehenderit. Aliquip velit ullamco esse aute exercitation cupidatat irure excepteur. Anim incididunt incididunt ipsum velit non enim eiusmod ipsum laborum veniam proident consequat ex ea. Aliquip sit proident tempor dolor et labore culpa sint. Ex excepteur ad do qui eu eiusmod do in proident officia.

Aute officia qui reprehenderit in nostrud dolore fugiat reprehenderit aute commodo mollit dolor. Aliqua fugiat voluptate commodo mollit non sunt in velit enim voluptate voluptate do ipsum. Non officia do labore aliqua eu eu excepteur adipisicing tempor. Aute eiusmod ipsum commodo in officia mollit.

Ad ex enim esse amet esse ea eiusmod Lorem do. In duis non non exercitation eu incididunt pariatur tempor. Eiusmod nulla ipsum laboris dolor pariatur. Ipsum ea tempor occaecat reprehenderit Lorem veniam do. Et Lorem adipisicing ex nostrud nisi ea anim amet. Aute quis veniam incididunt amet aliqua eu ipsum. Exercitation adipisicing sit dolore elit.

Irure sint ipsum nulla reprehenderit velit do quis laborum adipisicing fugiat proident. Consequat aute eiusmod ad proident occaecat nisi eu. Aliqua cupidatat magna sint anim. Culpa deserunt ad ullamco eiusmod voluptate tempor esse fugiat enim mollit in ullamco. Ex mollit do adipisicing amet. Nostrud nostrud exercitation dolore incididunt minim aliquip ea eu exercitation culpa consequat reprehenderit. Aliquip velit ullamco esse aute exercitation cupidatat irure excepteur. Anim incididunt incididunt ipsum velit non enim eiusmod ipsum laborum veniam proident consequat ex ea. Aliquip sit proident tempor dolor et labore culpa sint. Ex excepteur ad do qui eu eiusmod do in proident officia.

Aute officia qui reprehenderit in nostrud dolore fugiat reprehenderit aute commodo mollit dolor. Aliqua fugiat voluptate commodo mollit non sunt in velit enim voluptate voluptate do ipsum. Non officia do labore aliqua eu eu excepteur adipisicing tempor. Aute eiusmod ipsum commodo in officia mollit.

2 Related Work

Laborum dolor et commodo proident proident consequat. Anim ipsum deserunt non anim deserunt sunt magna mollit cupidatat nostrud. Laborum magna aliquip occaecat aute excepteur consectetur voluptate aute adipisicing tempor labore aliquip. Sunt aliquip ea velit minim elit tempor tempor sit ut dolore adipisicing occaecat voluptate. Velit laborum ullamco ea sunt nisi velit dolore amet sint est irure anim. Duis eu eiusmod elit est esse dolore enim minim.

Eiusmod esse deserunt do irure ut commodo ipsum esse ea enim aliqua ullamco ex. In minim laboris aliqua fugiat esse reprehenderit ipsum officia ad consequat non mollit. Deserunt Lorem eiusmod aliqua laborum ullamco irure eu commodo consequat officia. Ut enim occaecat occaecat tempor amet qui mollit reprehenderit voluptate ad nulla commodo. Minim ea ad incididunt laboris nisi.

Proident eiusmod et irure incididunt elit exercitation ullamco mollit reprehenderit veniam quis. Esse sit ex deserunt exercitation dolore non nostrud eu consectetur laboris quis. Aliquip deserunt minim Lorem elit non et officia in reprehenderit eiusmod non deserunt laborum officia. Ad consequat exercitation nisi anim eu. Amet officia deserunt nostrud sint tempor velit pariatur.

3 Models

The best performing model is BERT-large model (Talmor et al. 2018)

3.1 Pretrained-BERT with Next Sentece Prediction Head

BERT-NSP: Select the choice with the highest NSP score with the question. Results are shown in Table 1.

3.2 Pretrained-BERT with Finetuning

For CommonsenseQA task, a question and five candidate answers are given, and one of the five answers is correct. The candidate answers usually consist of one or two words, forming a *concept*. According to (Talmor et al. 2018), the best performing baseline model is the BERT-large model finetuned with CQA dataset. Our model is built based on BERT-large model, but also utilizes additional pieces of evidence from ConceptNet, which provides useful information to answer the question.

Model	train	dev	train(S)	dev(S)
BERT-base NSP	35.36	39.39	71.28	71.99
BERT-large NSP	38.41	40.38	73.54	73.14

Table 1: The result of the BERT-NSP model which selects the choice with the highest NSP score. train(S) and dev(S) corresponds to the *SANITY* version of train and dev set, respectively.

```
{
    "text": "meeting",
    "evidence": [
        "Something you find at a meeting is notepad",
        "a stranger is for meeting",
        "appointment is related to meeting",
        "interview is related to meeting",
        "group meeting is a synonym of meeting",
        "rendezvous is a type of meeting",
        "Something you find at a meeting is papers”
    ]
}
```

Figure 2: Subset of evidences from ConceptNet related to the term ”meeting”.

3.3 BECON

Concretely, to use the knowledge in Conceptnet, we first query each candidate answer in ConceptNet to get a list of evidence sentences which may be helpful to answer the question. In order to reduce the noise, we use pretrained BERT with Next-Sentence-Prediction head (BERT-NSP) to rank the evidence sentences and select the top-scored one. An example of the evidence sentences is shown in Figure 1. We believe that BERT-NSP is helpful to rank the relevancy of the question and the evidence sentence.

Evidence Finder Each question has 5 candidate answers. According to our analysis, candidate answers are usually one or two words long. We can use ConceptNet API http://api.conceptnet.io/c/en/ to find all the related information related with a word or phrase in the knowledge graph.

We expect that such evidences may be helpful to answer the question. The problem is, the evidence is too noisy. How to extract useful information? We would like to keep the evidence which is relevant to the question, and discard others. Assume that at most 1 evidence sentence is helpful (which means 0 or 1). We can first rank the evidences and then use the top-ranked evidence (or not).

Evidence Ranker The evidence ranker ranks the evidences according to the relevant scores with the question. We consider some of the very simple rankers:

- random: no ranking. Just random shuffle.
- jaccard: Jaccard Index is a metrics which consider the words ”intersection over union” between question and evidence sentences.

Ranker	train	dev	train(S)	dev(S)
random	21.24	19.82	21.07	19.57
jaccard	23.12	22.44	44.43	41.28
w2v	26.05	23.91	48.73	47.01
BERT-base	34.95	34.73	82.89	81.90
BERT-large	**36.50**	**36.86**	**84.41**	**82.88**

Table 2: The result of the Naive-Evidence model which selects the choice with the highest evidence score. train(S) and dev(S) corresponds to the *SANITY* version of train and dev set, respectively.

- w2v: the cosine distance between the average of pretrained word2vec embeddings of question and evidence sentences.
- BERT: use pretrained BERT model along with its Next Sentence Prediction head to determine the relevancy of two sentences.

Evidence Integrator As in (Talmor et al. 2018), each question-answer pair is linearized into a delimiter-separated sequence (i.e., "`[CLS]` If ... ? `[SEP]` bedroom `[SEP]`") and the hidden vector over the `[CLS]` token are used as representation of the choice. For our BECON model, we further concatenate the evidence sentence (i.e., "`[CLS]` If ... ? `[SEP]` bedroom `[SEP]` bedroom is a place for sleeping `[SEP]`"), which may help the model make better decisions.

$$[\texttt{CLS}] + Q + [\texttt{SEP}] + A + [\texttt{SEP}]$$
$$[\texttt{CLS}] + Q + [\texttt{SEP}] + A + [\texttt{SEP}] + E + [\texttt{SEP}]$$

4 Experiments

4.1 Dataset

4.2 Experimental Settings

4.3 Development Experiments

Evidence Rankers To have a sense of how the rankers work, we use the ranker to rank all the evidences of the 5 candidate answers. The candidate answer with the top ranked evidence is chosen as the predicted answer. A simple model without training (Naive-Evidence): select the choice with the highest evidence score. The result on train and dev is shown in Table 2.

There is no dev result in the original paper, but if we assume the dev and test result are close, we can see that the BERT-large NSP model without training is only inferior than BERT-large and GPT which use the CQA dataset to train.

Evidence Integrator Table 3

The comparision between BERT-base/large rankers show that BERT-large ranker is better. The experiments later all use BERT-large ranker.

Compared with our baseline, the result is a bit lower. This means if we add evidence for each answer candidate, the noise may still overwhelms the useful information.

Table 4

Table 5

Pretrain Modles	Rankers	dev
BERT-base	BERT-base	56.2
BERT-base	BERT-large	57.6
BERT-large	BERT-base	61.9
BERT-large	BERT-large	62.2

Table 3: BECON with different BERT pretrain models and rankers on dev.

Pretrain Modeds	pooling	dev
BERT-large	max	63.6
BERT-large	mean	64.0
BERT-large	concat(no pooling)	64.4

Table 4: BECON w/ w/o evidence combination.

BECON Singale Evidence	dev
BERT-base	58.3
BERT-large	62.8

Table 5: BECON with single evidence

4.4 Results

The experiment results on CQA test split are shown in Table 6. Our single model outperforms the BERT-large baseline by 1.2%. Using ensemble technique, our model achieves 59.7%, outperforms CoS-E (Rajani et al. 2019) by 1.4%.

5 Discussion

An interesting phenomenon is that BERT NSP without any training on CQA dataset has comparable performance with ESIM + ELMO/glove models on CQA dataset.

5.1 Error Analysis

Irure reprehenderit culpa sint fugiat officia excepteur non reprehenderit nulla exercitation laborum. Consectetur laboris consectetur mollit adipisicing excepteur reprehenderit consequat nisi in cillum non. Ipsum sit nulla enim ea ut ex cupidatat labore nisi magna ex. Laboris qui officia laborum excepteur qui sit laborum non incididunt. Adipisicing culpa commodo amet laborum irure fugiat commodo consequat deserunt consectetur ea.

Qui dolor anim eiusmod cillum laboris eu ipsum non quis. Commodo aute nisi et deserunt exercitation ex ad quis ea irure ipsum. Irure ullamco dolor elit eu irure laborum veniam aliquip veniam et ut. Enim labore laborum elit ex duis voluptate do velit aliqua sunt minim labore deserunt. Ut cupidatat nostrud ut labore consectetur do qui esse velit. Nulla Lorem ipsum dolore voluptate minim do. Elit dolore occaecat aliquip adipisicing irure amet.

Reprehenderit id irure minim excepteur mollit cupidatat nostrud. Eu nisi tempor sit deserunt est nostrud quis occaecat aute dolore cupidatat pariatur anim. Incididunt culpa sunt consequat enim eiusmod in ea ad consectetur anim commodo culpa et ullamco. Aliqua minim proident voluptate laborum in sunt. Esse et tempor enim tempor ea et dolore

Model	test F1
BERT-large (Talmor et al. 2018)	56.7
CoS-E (Rajani et al. 2019)	58.2
BECON	57.9
BECON (ensemble)	59.6

Table 6: Comparison of the test accurary with literature.

laborum ea exercitation anim. Eiusmod officia aliquip irure amet aliqua anim aliquip.

5.2 Case Study

Commodo officia consectetur laboris pariatur enim et eu est id excepteur laborum. Cupidatat dolore ut eu exercitation occaecat aliqua. Dolore do dolor Lorem ex cupidatat id id magna irure. Ad est ut occaecat culpa amet. Cillum mollit labore quis ipsum aliquip tempor ut magna occaecat.

Aliquip veniam laborum consequat ullamco exercitation deserunt. Non veniam quis esse magna eu magna elit tempor sunt do enim excepteur. Est eu consectetur amet qui Lorem.

6 Conclusion

We use conceptnet to search for evidence, use BERT to rank them, and use BERT as the base model to train the model with evidence. To alleviate the noise introduced by the evidence, we use BERT to encode both w/ w/o evidence, and let model learn to choose which one contributes more. This model outperforms BERT-large baseline by 0.9% on dev and +1.2% on test, which proves the effectiveness of our method. For comparison, salesforce research use human generated explaination to enhance the question, which only outperforms our model by 0.3%.

NOTE: our submitted model is lower than our best model on dev by 0.4%. However we cannot resubmit until two weeks later. So if the increase is likewise on test set, we may have comparable results with COS-E by Salesforce Research.

References

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.