COMMONSENSE MINING AS KNOWLEDGE BASE COMPLETION? A STUDY ON THE IMPACT OF NOVELTY

Stanisław Jastrzębski, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio, Jackie Chi Kit Cheung

MILA Université de Montréal

Uniwersytet Jagielloński w Krakowie

McGill

Abstract

- We analyze whether knowledge base completion models can be used to mine commonsense
- We propose "novelty" with respect to the training set as an important factor in evaluation, and use it to show a simpler model outperforms state-of-the-art

Mining Commonsense

- Many NLP tasks require commonsense knowledge but collecting and organizing it is difficult
- Commonsense knowledge bases (e.g. ConceptNet) represent commonsense knowledge as relational triples: *("pen", "UsedFor", "writing")*

We look at automatically mining commonsense knowledge by improving the coverage of commonsense knowledge bases

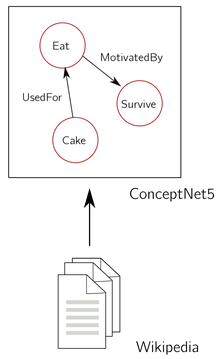
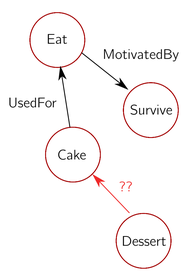


Fig. 1: Visualization of the ConceptNet graph

Mining as Knowledge Base Completion

- A common way of improving the coverage of knowledge bases is through knowledge base completion (KBC)
- **Li et al** approached commonsense mining as a KBC task. Their method mines candidate triples from Wikipedia and reranks the triples with a KBC model in order to extend ConceptNet



"Dessert", "IsA", "Cake") → {0

Fig. 2: Visualization of the ConceptNet graph

The goal is to evaluate the potential of the recent systems for mining commonsense and how well they truly understand commonsense

Models for KBC

All our models take (h, r, t) triples as inputs, where h and t are sequences of words representing concepts and r is a relation. We embed h and t by embedding each word and summing over the sequence to get $\mathbf{h}$ and $\mathbf{t}$ and we embed the relation to get $\mathbf{r}$

- **Bilinear**$(\mathbf{h}, r, \mathbf{t}) = \mathbf{h}^{\mathbf{T}}\mathbf{M}_r\mathbf{t}$,
- **DNN**$(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \mathbf{W_2}\phi(\mathbf{W_1}[\mathbf{h}, \mathbf{r}, \mathbf{t}] + \mathbf{b_1}) + \mathbf{b_2}$,
- **Factorized**$(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \alpha\langle\mathbf{Ah} + \mathbf{b_1}, \mathbf{Bt} + \mathbf{b_2}\rangle + \beta\langle\mathbf{Ar} + \mathbf{b_1}, \mathbf{Bt} + \mathbf{b_2}\rangle + \gamma\langle\mathbf{Ar} + \mathbf{b_1}, \mathbf{Bh} + \mathbf{b_2}\rangle$
- **Prototypical**$(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \beta\langle\mathbf{Ar} + \mathbf{b_1}, \mathbf{Bt} + \mathbf{b_2}\rangle + \gamma\langle\mathbf{Ar} + \mathbf{b_1}, \mathbf{Bh} + \mathbf{b_2}\rangle$

ConceptNet and Wikipedia Setup

Models are trained using 100k triples from ConceptNet5 that were extracted from the OMCS corpus

ConceptNet5 - Completion task considers two ways to split the dataset: a random split, and confidence-based split using triples with the highest confidence scores as a test set

Wikipedia - Mining commonsense task is based on a set of 1.7 M extracted candidate triples from Wikipedia by **Li et al.** The extracted triples are ranked using a KBC model, and the top of the ranking is manually evaluated

ConceptNet and Wikipedia Results Mismatch

- **We find that the knowledge base completion task is a poor indicator of performance on Wikipedia**
- Factorized and Prototypical models achieve a similar or worse score compared to DNN on the KBC task, their mining performance on the top 100 triples is better than both DNN and the Bilinear model

Novelty \ Model	DNN	Factorized	Prototypical
Entire	**0.892**	0.890	0.794
$\leq 33\%$	**0.950**	0.922	0.911
$(33\%, 66\%]$	**0.920**	0.898	0.839
$\geq 66\%$	0.720	**0.821**	0.574

Tab. 1: F1 scores on Li et al. confidence-based test set. F1 score is reported on each bucket (based on the percentile of triple novelty) and the entire test set

	Bilinear	Factorized	Prototypical	DNN
Wikipedia	2.04	**2.61**	2.55	2.5

Tab. 2: Average human-assigned score (from 1 to 5) of the top 100 Wikipedia triples ranked by baselines compared to DNN and Bilinear from Li et al

Novelty Explains the Mismatch

Dataset \ Novelty	1	2	3	4	5
Wikipedia	14%	5%	17%	8%	44%
Confident	65%	22%	4%	4%	2%
Random	21%	10%	16%	3%	29%

In Category 1 we find *("egg", "IsA", "food")*, which has a close analog in the training set: *("egg", "IsA", "type of food")*. In Category 3 (*"different relation and related word"*) we find *("floor", "UsedFor", "walk on")*, which has a corresponding triple in the training set *("floor", "UsedFor", "stand on")*

Novelty-Binned Setup

- To approximate novelty, we use the distance between the word embeddings of triples' heads and tails
  $d(a, b) = ||\mathbf{h}(a) - \mathbf{h}(b)||_2 + ||\mathbf{t}(a) - \mathbf{t}(b)||_2$
- **We split the Wikipedia and ConceptNet5 datasets into 3 buckets based on novelty**: 33% and 66% quantiles of distance to the training set

Novelty-Binned Results

- **Model performance is directly tied to test set novelty**
- Factorized model usually outperforms DNN on all buckets despite being a simpler model

Novelty \ Model	DNN	Factorized	Prototypical
Entire	0.809	**0.822**	0.755
$\leq 33\%$	**0.883**	0.874	0.866
$(33\%, 66\%]$	0.809	**0.812**	0.758
$\geq 66\%$	0.725	**0.731**	0.674

Tab. 4: F1 scores on random split. F1 score is reported on each bucket (based on the percentile of triple novelty) and the entire set

Novelty \ Model	DNN	Factorized	Prototypical
$\leq 33\%$	2.47	**2.58**	2.33
$(33\%, 66\%]$	2.34	**2.41**	2.24
$\geq 66\%$	1.41	**2.26**	1.63

Tab. 5: Novelty-based evaluation of quality of mined triples from Wikipedia dataset, triples are scored by humans on a scale from 1 to 5

Conclusions

- We show that previous evaluation metrics for KBC are insufficient and propose "novelty" as a correction for dataset biases
- We critically assessed models for mining commonsense knowledge, and found that a simpler model outperforms state-of-the-art when correcting for dataset biases with novelty
- Future work could focus on developing new regularization techniques to better generalize to novel triples

Acknowledgments: SJ was supported by Grant No. DI 2014/016644 from Ministry of Science and Higher Education, Poland. Work at MILA was funded by NSERC, CIFAR, and Canada Research Chairs