Reading Group 1@19-20

Highlights from ACL and EMNLP

Outline

1. **Towards Language Agnostic Universal Representations.** Armen Aghajanyan, Xia Song, Saurabh Tiwary. ACL 2019.

2. **Language Models as Knowledge Bases?** Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, Sebastian Riedel. EMNLP 2019.

3. **COMET: Commonsense Transformers for Automatic Knowledge Graph Construction.** Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, Yejin Choi. ACL 2019.

4. **AMR Dependency Parsing with a Typed Semantic Algebra.** Jonas Groschwitz, Matthias Lindemann, Meaghan Fowlie, Mark Johnson, Alexander Koller. ACL 2018.

Towards Language Agnostic Universal Representations

Armen Aghjanyan, Xia Song and Saurabh Tiwary. In Proceedings of ACL 2019.

- Create word and sentence **representations** that are **independent** from the input **language**.
- Follow the **weak linguistic influence theory** which brings forward the **P**rinciples a**N**d **P**arameters (PnP) argument and the **Universal Grammar** (UG) hypothesis.
- **Are universal features of the languages learnable within a statistical framework?**

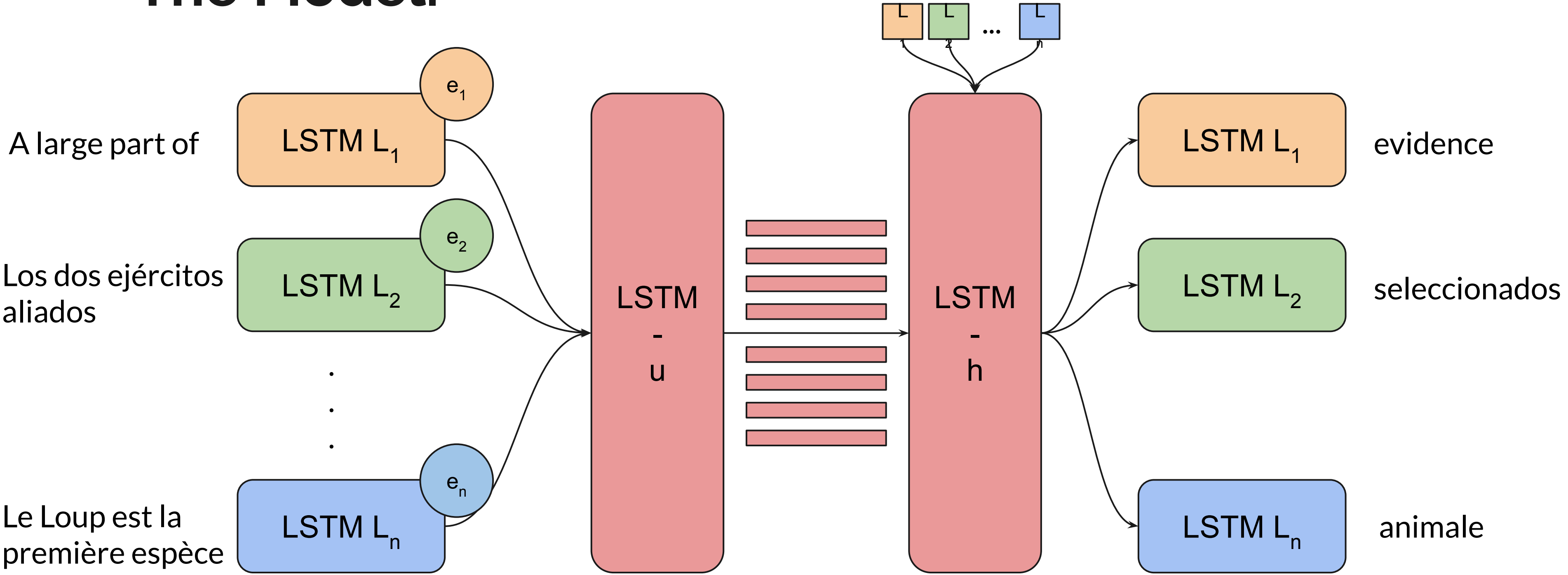
The High-Level Idea.

- Under the UG hypothesis, a language model for a specific language is composed by a language-specific part and a set of universal rules.

- Hence, the probability P(w_i | w_0, … w_(i-1)) should factorized in two components: a universal and a language-specific one.

- The language-agnostic component should be shared across all the languages.

The Model.



L

L

...

L

e_1

A large part of

LSTM L_1

e_2

Los dos ejércitos aliados

LSTM L_2

.
.
.

e_n

Le Loup est la première espèce

LSTM L_n

LSTM - u

LSTM - h

LSTM L_1

evidence

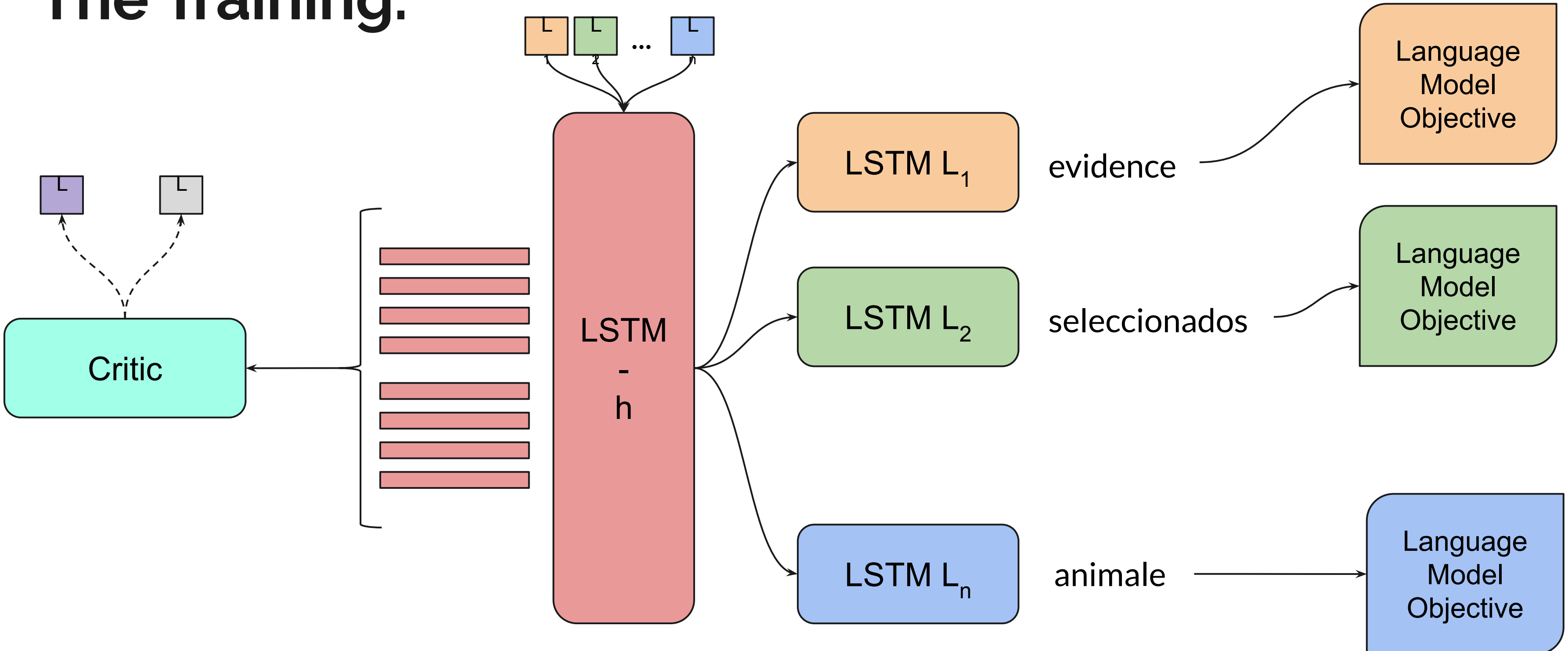
LSTM L_2

seleccionados

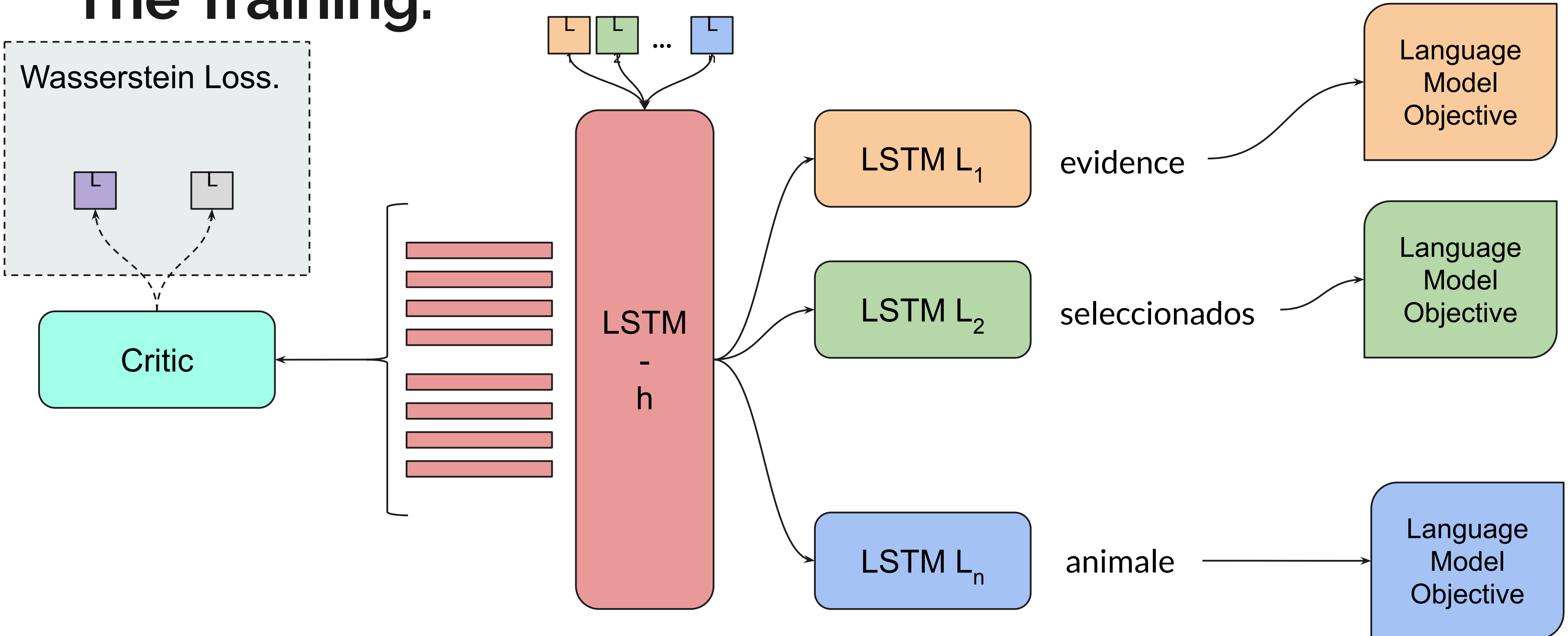
LSTM L_n

animale

The Training.



The Training.



Wasserstein Loss.

L

L

Critic

L_1

L_2

...

L_n

LSTM

-

h

LSTM L_1

LSTM L_2

LSTM L_n

evidence

seleccionados

animale

Language Model Objective

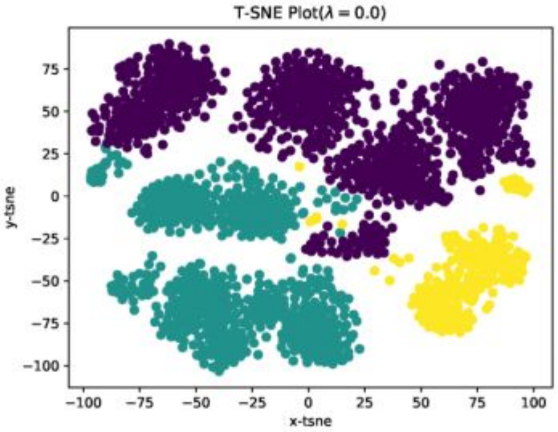
Language Model Objective

Language Model Objective

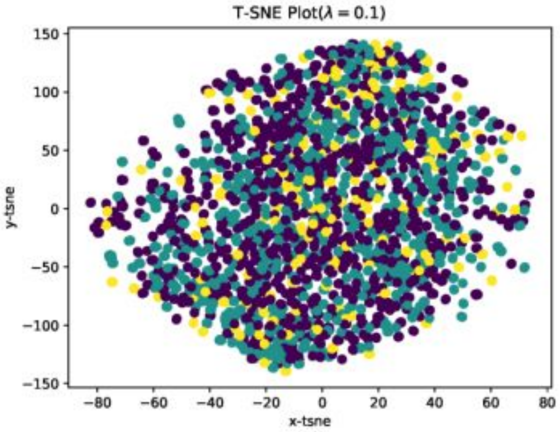
Research Questions.

1. Is the adversarial training needed or simply training N language models is already enough?
2. Are language-specific rules learned while also learning universal features?
3. Are commonalities between languages informative enough to be exploited?

Experiments - Is the adversarial training needed?



Without Adversarial Loss



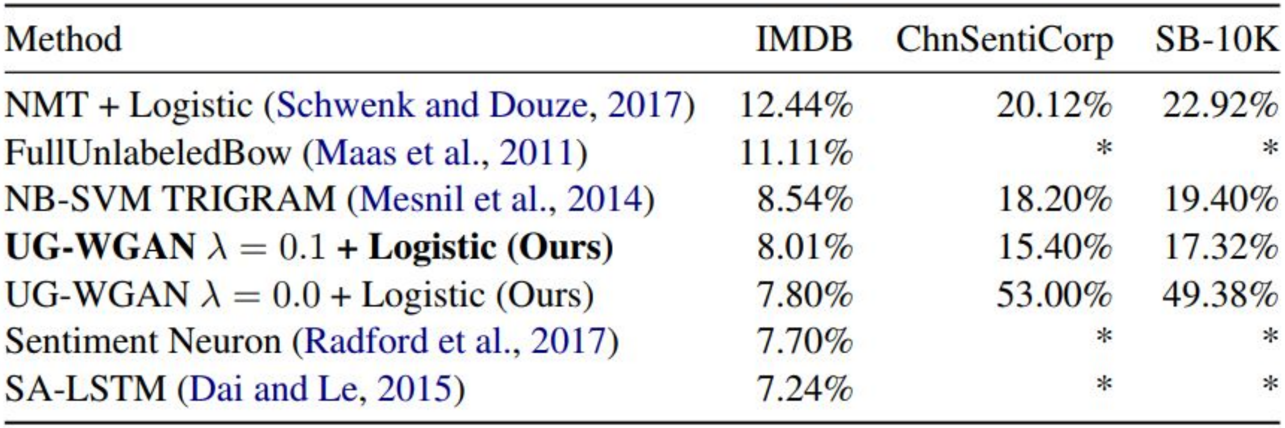
With Adversarial Loss

Experiments Sentiment Analysis.

- UG-WGAN: trained on English, Chinese and German on Wikipedia dumps.
- Reference Model: Stacked LSTM fed with UG representations and trained on IMDB in English only.
- Test Sets: zero-shot setting on Chinese and German (ChnSentiCorp, German SB-10K).
- Comparison systems: $e_1, \dots, e_n$ and the encoders trained with the NMT framework by Klein et al. 2017: Open-Source Toolkit for Neural Machine Translation.

Results.

Error rates for each model and dataset.



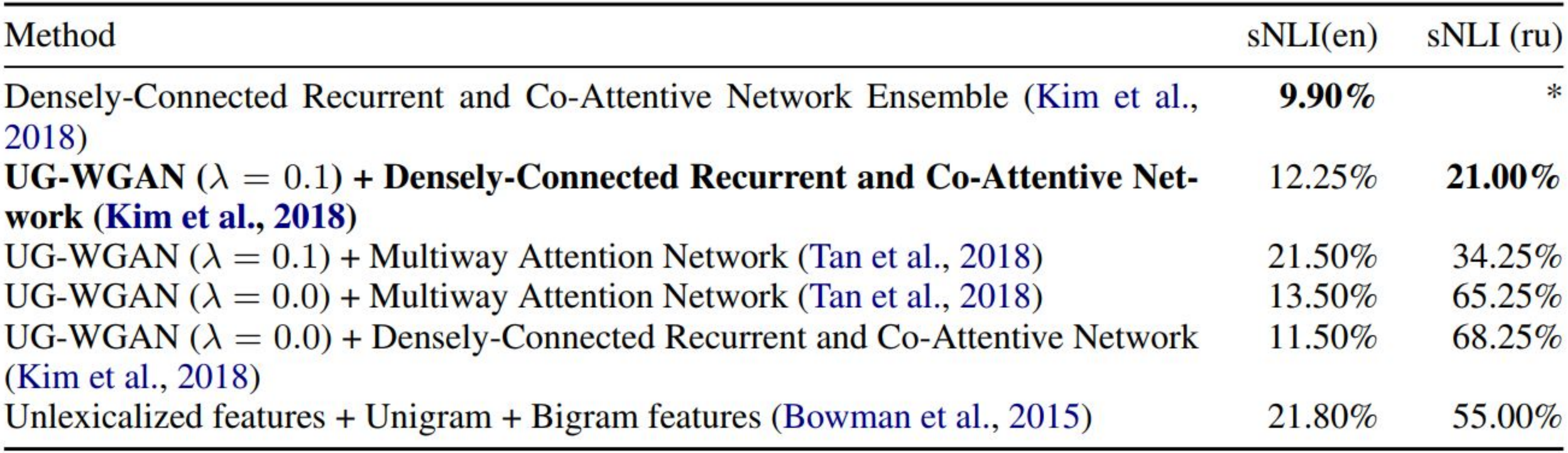
Method	IMDB	ChnSentiCorp	SB-10K
NMT + Logistic (Schwenk and Douze, 2017)	12.44%	20.12%	22.92%
FullUnlabeledBow (Maas et al., 2011)	11.11%	*	*
NB-SVM TRIGRAM (Mesnil et al., 2014)	8.54%	18.20%	19.40%
UG-WGAN $\lambda = 0.1$ **+ Logistic (Ours)**	8.01%	15.40%	17.32%
UG-WGAN $\lambda = 0.0$ + Logistic (Ours)	7.80%	53.00%	49.38%
Sentiment Neuron (Radford et al., 2017)	7.70%	*	*
SA-LSTM (Dai and Le, 2015)	7.24%	*	*

Experiments NLI.

- UG-WGAN: trained on English and Russian Wikipedia dumps.
- Reference Models: DCRCAN (Kim et al. 2018 - Semantic sentence matching with densely-connected recurrent and co-attentive information.) and MAN (Tan et al. 2018 - Multiway attention networks for modeling sentence pairs.) fed with UG embeddings.
- Test Set: Standard Stanford-NLI dataset (Bowman et al. 2015 - A large annotated corpus with for learning natural language inference.) and 400 manually translated examples in Russian.
- DCRCAN and Baseline by Bowman et al. 2015.

Results.

Error rates for each model and dataset.



Method	sNLI(en)	sNLI (ru)
Densely-Connected Recurrent and Co-Attentive Network Ensemble (Kim et al., 2018)	**9.90%**	*
UG-WGAN ($\lambda = 0.1$) **+ Densely-Connected Recurrent and Co-Attentive Network (Kim et al., 2018)**	12.25%	**21.00%**
UG-WGAN ($\lambda = 0.1$) + Multiway Attention Network (Tan et al., 2018)	21.50%	34.25%
UG-WGAN ($\lambda = 0.0$) + Multiway Attention Network (Tan et al., 2018)	13.50%	65.25%
UG-WGAN ($\lambda = 0.0$) + Densely-Connected Recurrent and Co-Attentive Network (Kim et al., 2018)	11.50%	68.25%
Unlexicalized features + Unigram + Bigram features (Bowman et al., 2015)	21.80%	55.00%

Take-Home Message.

- Wasserstein Loss can be effectively used to learn unified representation of tokens across languages.
- The learnt representations show promising results on two different datasets, however, more experiments are needed in order to asses the real scalability power of the model.
- The approach is interesting but it would be interesting to study how to remove the need of data in other languages (even for the pretraining step) and other constraints to unify the representations.

Language Models as Knowledge Bases?

Fabio Petroni, Tim Rocktaschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller and Sebastian Riedel. In Proceedings of EMNLP 2019.

- The authors challenge two (masked) language models, i.e., ELMo and BERT to answer questions from various datasets and try to answer the following questions:

- How much relational knowledge do language models store?
- How does this differ for different type of knowledge such as facts about entities, common sense and general question answering?
- How do they compare to symbolic knowledge bases that were automatically extracted from texts?

The Lama Probe

- LAMA: LAnguage Model Analysis, is a framework to test the factual and commonsense knowledge contained in a Language Model.
- It comprises facts in the form of either Subject Relation Object triples or Question-Answer pairs.
- It can automatically convert the triples in cloze statements that can be used to query the language models.
- The evaluation is carried out by measuring the ranking of the correct token (answer / object) with respect ranking given as output by the language models.

The LAMA Probe.

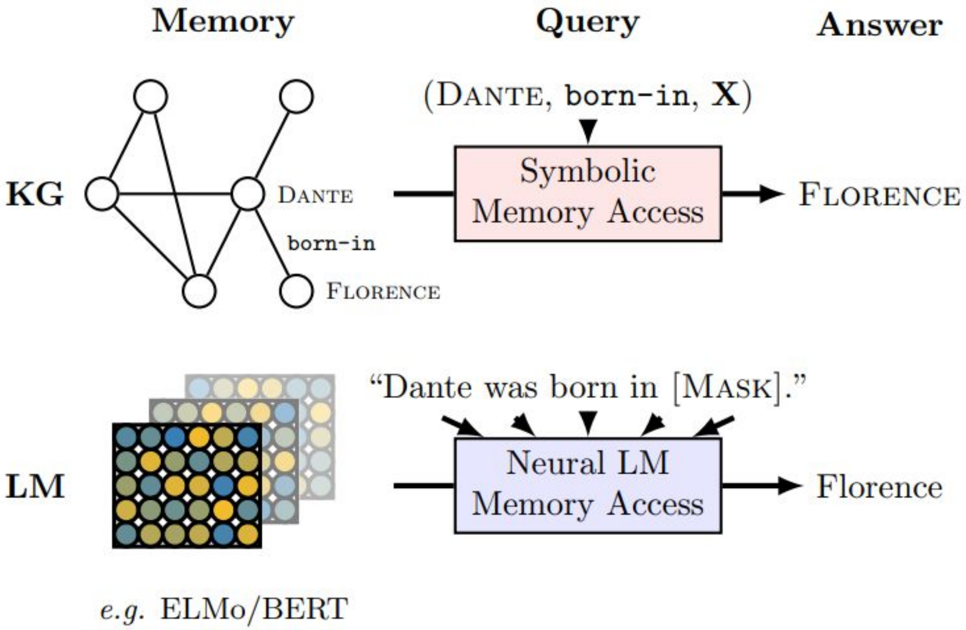
- LAMA: LAnguage Model Analysis, is a framework to test the factual and commonsense knowledge contained in a Language Model.

- Factual triples:
  Obama, born-in, Hawaii.
  iPod, produced-by, Apple.

- Question-Answer pairs:
  Who developed the theory of relativity? Einstein.
  Where is Rome?

- Conversion patterns:
  [S] was born in [O] -- manually built pattern for each type of relations.
  Rome is in __? -- Question-answers were rewritten in a cloze style.

The LAMA Probe.



The LAMA Probe - Knowledge Bases.

- **Google-RE**: selection of ~ 60K manually extracted triples from Wikipedia for 3 relations (place of birth, date of birth and place of death).

- **T-REx** (Elsahar et al., LREC 2018): selection of ~ 41K triples for 41 different relations.

- **ConceptNet** (Speer and Havasi LREC 2012): Multilingual knowledge base which encodes commonsense knowledge. They considered 16 distinct relations from ConceptNet.

- **SQuAD** (Rajpurkar et al. EMNLP 2016): selection of 305 context-insensitive question-answer pairs that have been manually converted in cloze-style questions.

The LAMA Probe - Baselines & Metrics.

- **Freq**: It outputs the word that most frequently appears in relation with the input subject relation.

- **RE** (Sorokin and Gurevych EMNLP 2017)**:** An LSTM-based model trained on Wikipedia sentences tagged with Wikidata triples to create a knowledge graph based. At test time, the KB is queried with subject and relation and the highest ranked object is returned.

- **DrQA** (Chen et al. CoRR 2017)**:** A two-steps approach to answer open-domain questions. It uses TF-IDF to retrieve the most relevant documents for a question and then a neural reading-comprehension model is used to extract the answers.

- **Precision@k:** It is 1 when the correct object is ranked in the top *k* results, 0 otherwise.

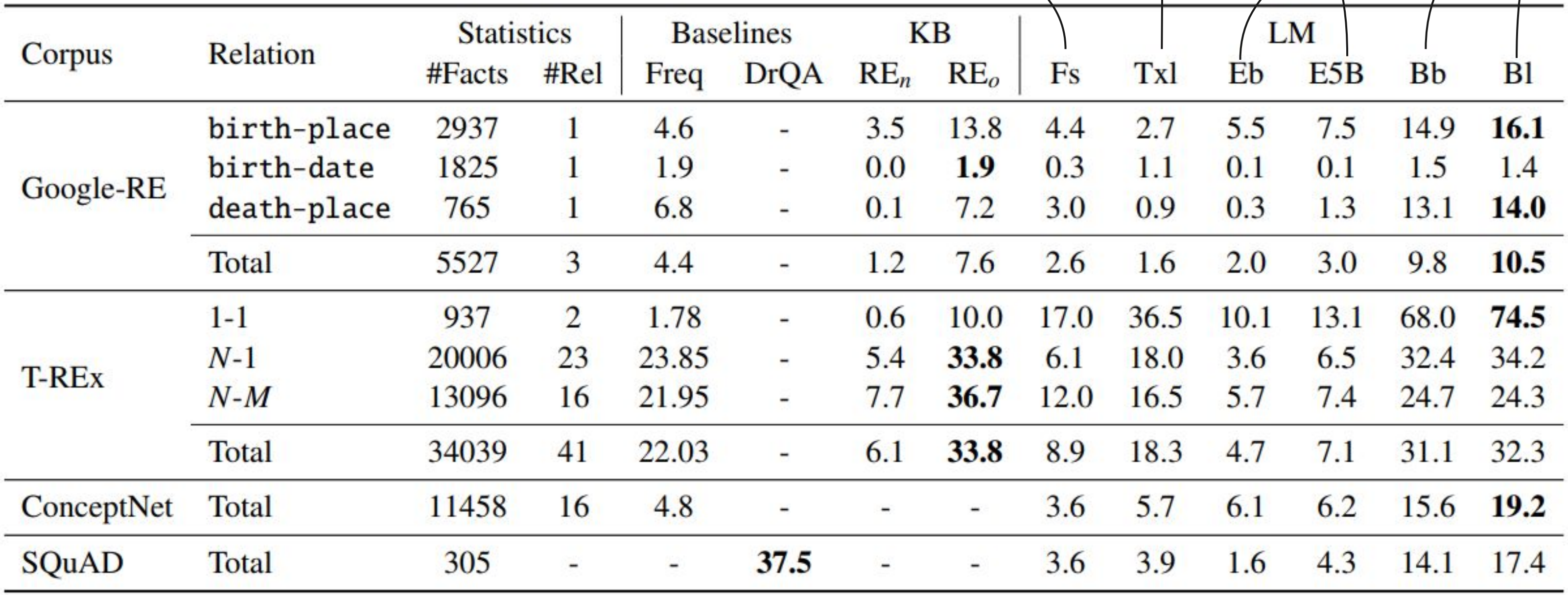
Results.

Fairseq-fconv
(Dauphi et al. ICML 2017)

Transformer-XL
(Dai et al. CoRR 2019)

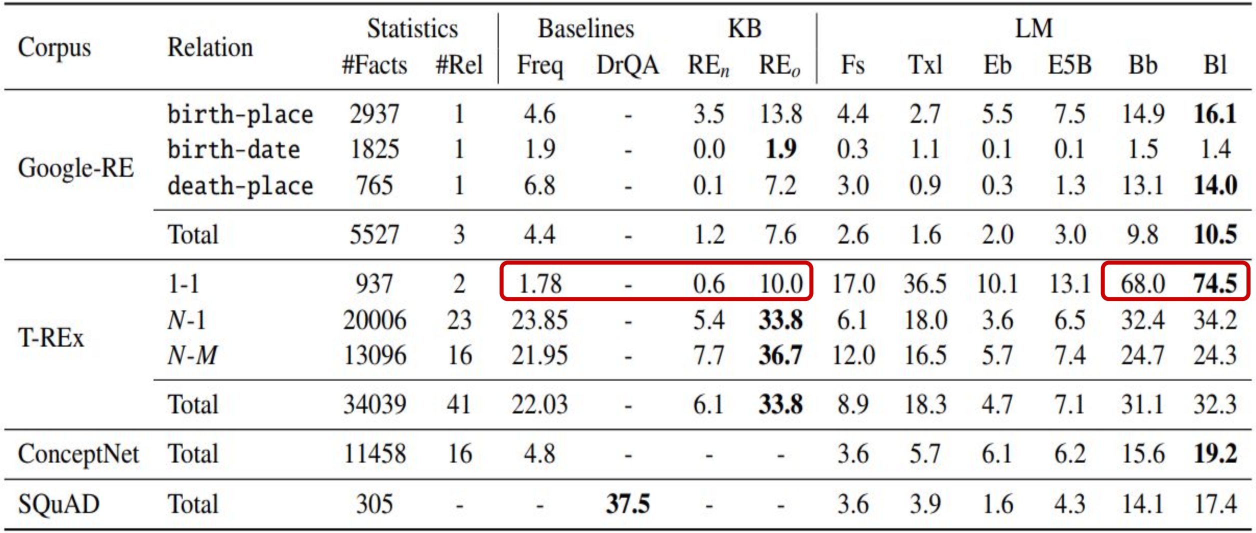
ELMo/ELMo Wiki
(Peters et al. NAACL 2018)

BERT-base/BERT-large
(Devlin et al. NAACL 2019)



Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

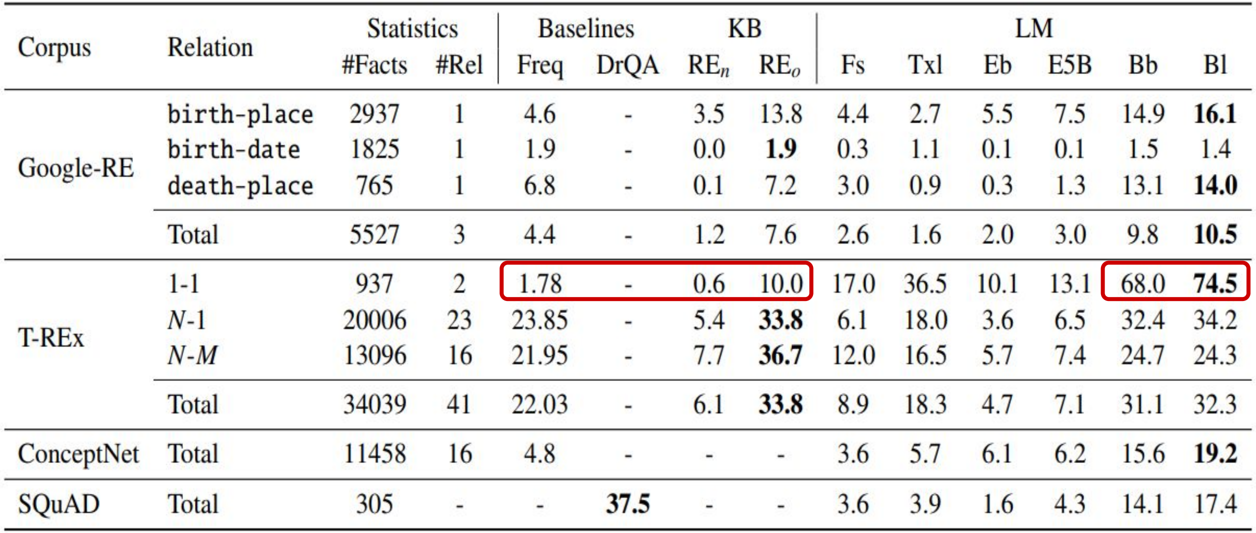
Results.



Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE$_n$	RE$_o$	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

- Note that Eb, E5b,Bb and Bl saw test sentences during training and in fact are particularly good in 1-1 relations.
- Authors found out that the object being in the training data correlate positively with high performance.
- BERT has been found to provide prediction for the masked token that, when not correct, are of the same type.

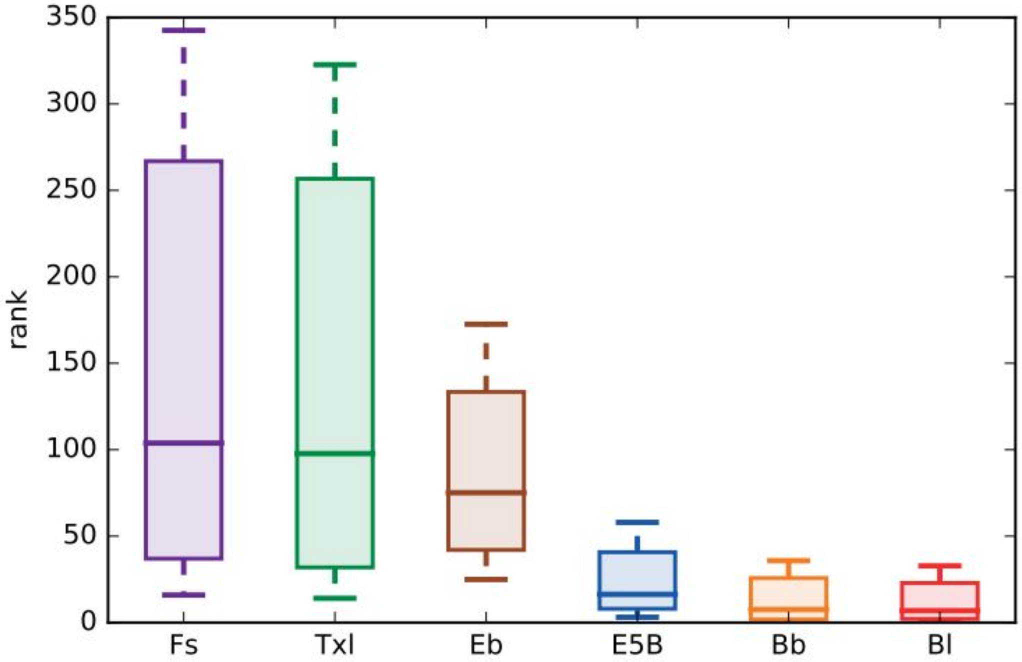
Results.



Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE$_n$	RE$_o$	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

- Note that Eb, E5b,Bb and Bl saw test sentences during training and in fact are particularly good in 1-1 relations.
- Authors found out that the object being in the training data correlate positively with high performance.
- BERT has been found to provide prediction for the masked token that, when not correct, are of the same type.

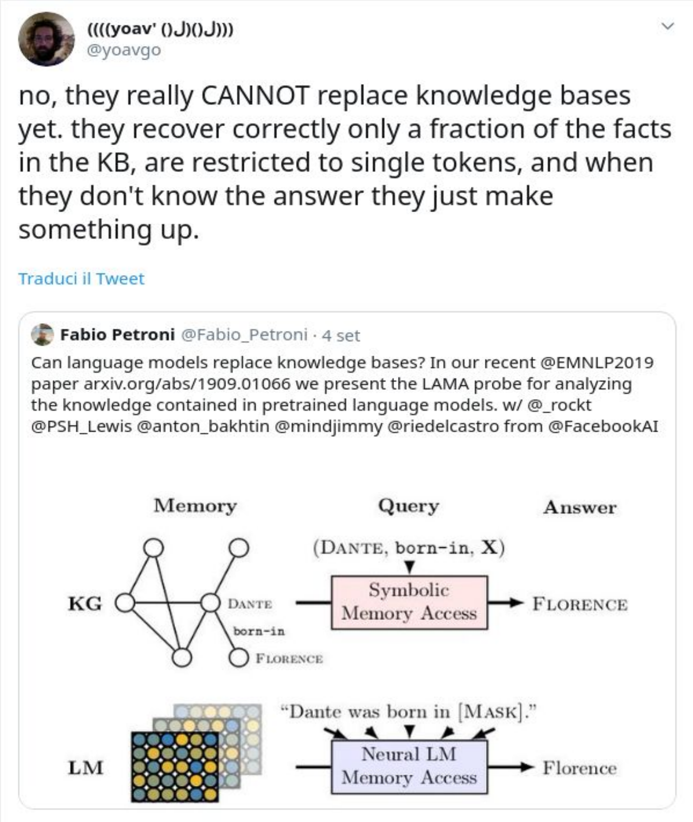
Results.



- When providing the question in different ways the models change their prediction.
- Bl and Bb showed the lowest variance in ranking the right object in their outputs.
- The best performing models had seen the test sentences during training.
- FS and Txl saw less Wikipedia sentence during training.

Take-Home Message.

- Language models showed promising results in retrieving information provided during training.
- BERT, especially, was the best LM performing across the boards.
- Still it is not clear the best way to query LM and the experiments were performed considering single token subject/relation/object only.
- It might be beneficial to fine-tune a LM to accept a specific query language to retrieve the information.
- Still the authors haven not proved that the models are actually understanding either the subjects/objects nor the relation between them. On the contrary, the probably only rely on statistics gathered during training.



((((yoav' ()J)()J)))
@yoavgo

no, they really CANNOT replace knowledge bases yet. they recover correctly only a fraction of the facts in the KB, are restricted to single tokens, and when they don't know the answer they just make something up.

Traduci il Tweet

Fabio Petroni @Fabio_Petroni · 4 set

Can language models replace knowledge bases? In our recent @EMNLP2019 paper arxiv.org/abs/1909.01066 we present the LAMA probe for analyzing the knowledge contained in pretrained language models. w/ @_rockt @PSH_Lewis @anton_bakhtin @mindjimmy @riedelcastro from @FacebookAI

More discussion @ https://bit.ly/31TOU4s

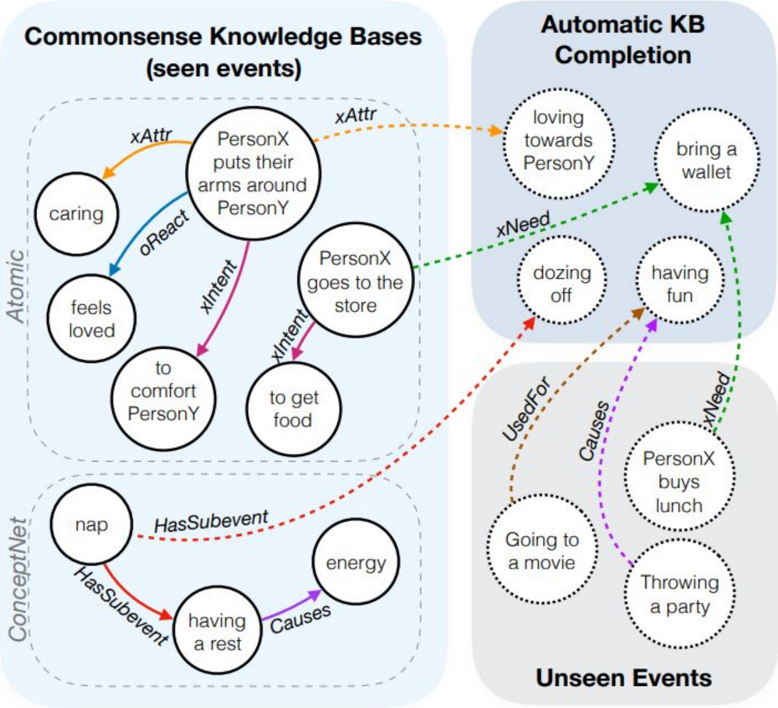
COMET: Commonsense Transformers for Automatic Knowledge Graph Construction.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz and Yejin Choi. In Proceedings of ACL 2019.

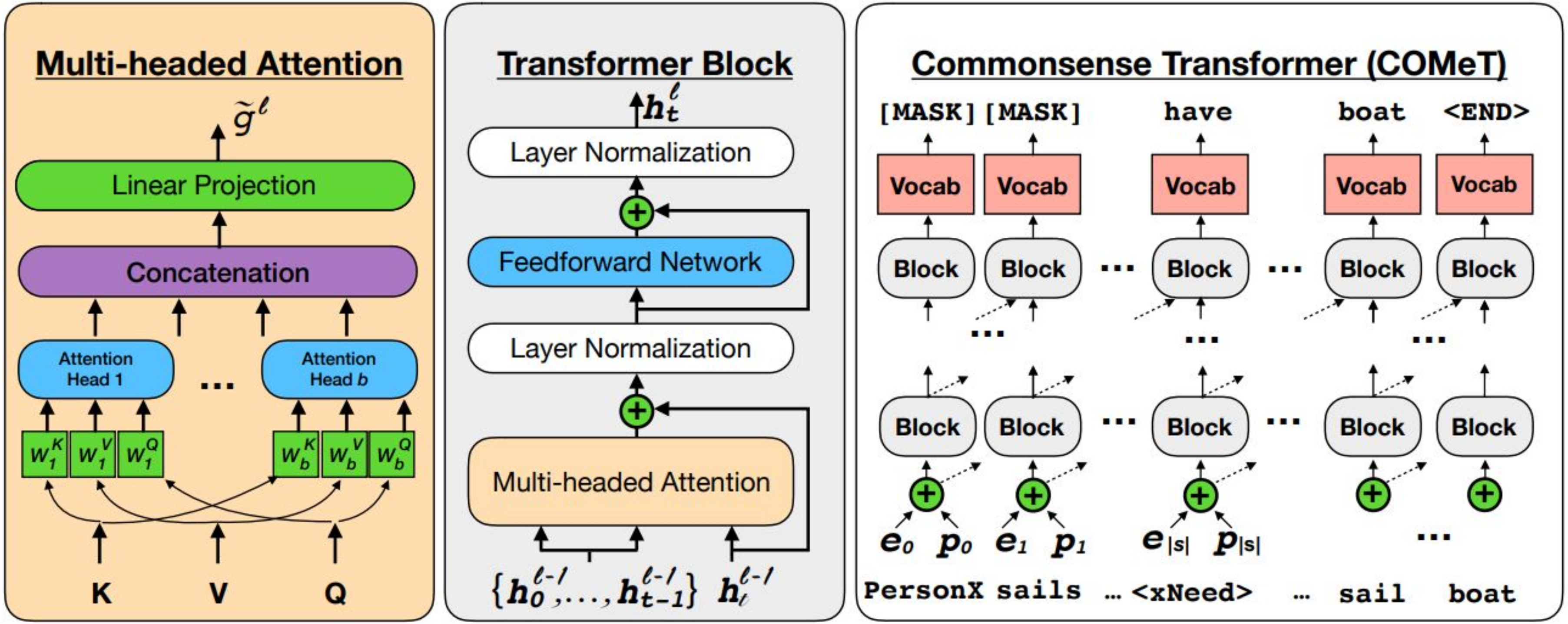
- Commonsense knowledge bases encode practical knowledge of day-to-day life that is common to a large group of people.

- The automatic construction of commonsense KB is usually framed as the problem of extracting triples from a given text.

- COMET is a transformer-based model which can generate *subject,relation,object* triples that can be used to enrich an existing KB.



COMET ☄ - A GPT-Based Transformer.



COMET ☄ - Training.

- COMET expects as input a sentence that represents a *subject-relation* couple and has to predict its corresponding *object*.

- Given a triple s="take a nap", r="Causes", o="have energy", COMET converts it in the following input:
- "*take a nap [MASK] causes [MASK] have energy*".

- The model is initialised with the pretrained weights of the GPT language model.

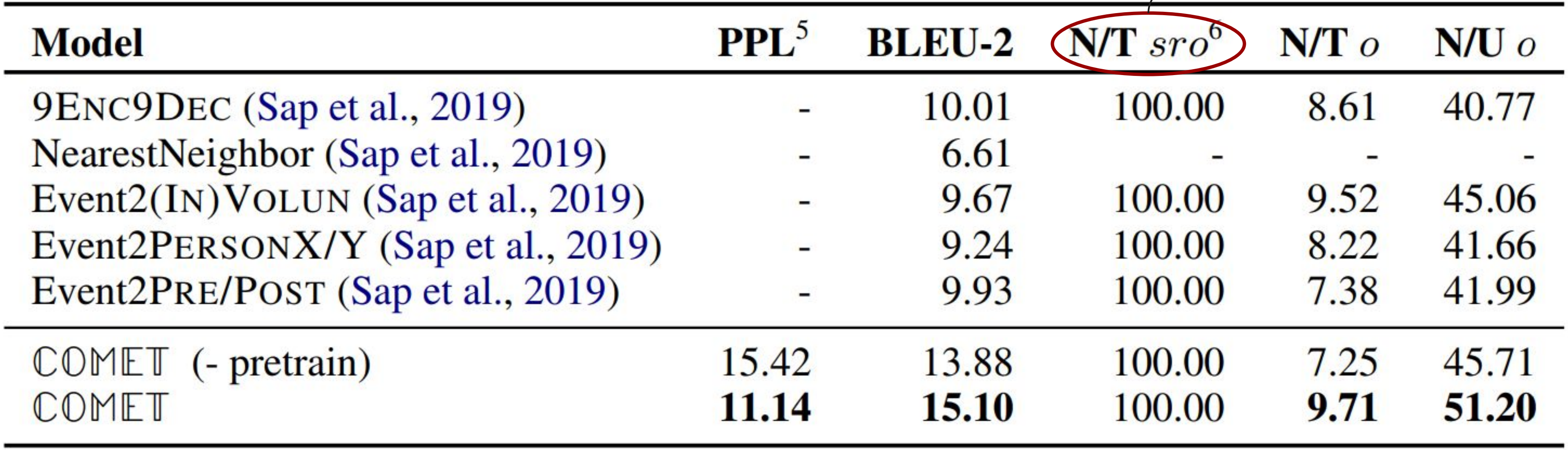
- The authors experimented with two different knowledge bases: ConceptNet (Robyn Speer et al. AAAI 2017) and ATOMIC (Sap et al. AAAI 2019).

COMET ☄ - Baselines & Evaluation Metrics.

- Sap et al. AAAI 2019 models: LSTM-based models that given a subject and a relation it produces an object as output and a BiLSTM-based model by Saito et al. CCNLL 2017.
- BLEU-2: compares the produced output and the gold one.
- Average perplexity on the gold sentences.
- Portion of generated tuples not in the training set (% N/T sro)
- Portion of generated tuples with a new object (% N/T o)
- Human evaluation with Amazon Mechanical Turk.
    - Manually evaluated 100 random events from the test set.
    - Each event is associated with 10 possible object candidate sampled with beam search on top of the model predictions.
    - Each event + object is evaluated by 5 persons.

COMET - ATOMIC Results.

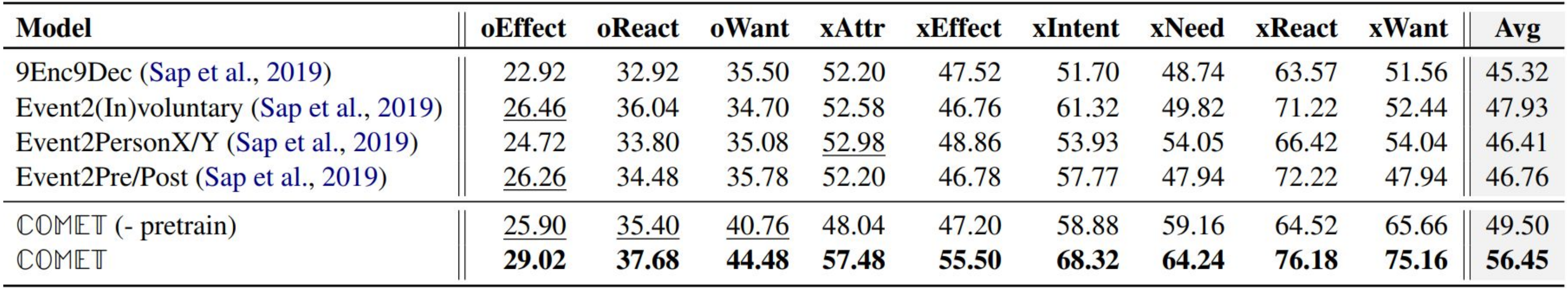
By design of the test set all the tuples are different from those in the training set.



Model	PPL[5]	BLEU-2	N/T *sro*[6]	N/T *o*	N/U *o*
9Enc9Dec (Sap et al., 2019)	-	10.01	100.00	8.61	40.77
NearestNeighbor (Sap et al., 2019)	-	6.61	-	-	-
Event2(In)Volun (Sap et al., 2019)	-	9.67	100.00	9.52	45.06
Event2PersonX/Y (Sap et al., 2019)	-	9.24	100.00	8.22	41.66
Event2Pre/Post (Sap et al., 2019)	-	9.93	100.00	7.38	41.99
COMET (- pretrain)	15.42	13.88	100.00	7.25	45.71
COMET	**11.14**	**15.10**	100.00	**9.71**	**51.20**

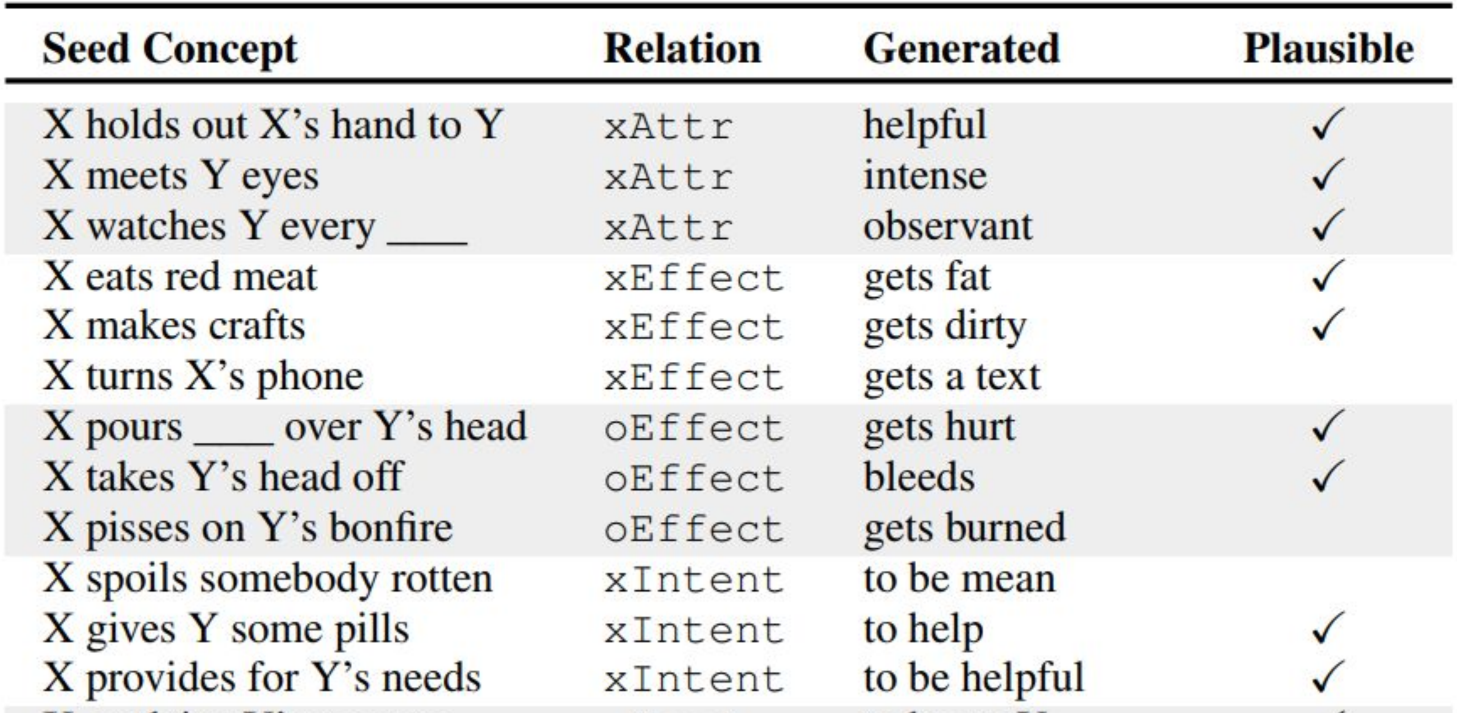
COMET - ATOMIC Human Evaluation Results.

Relations.



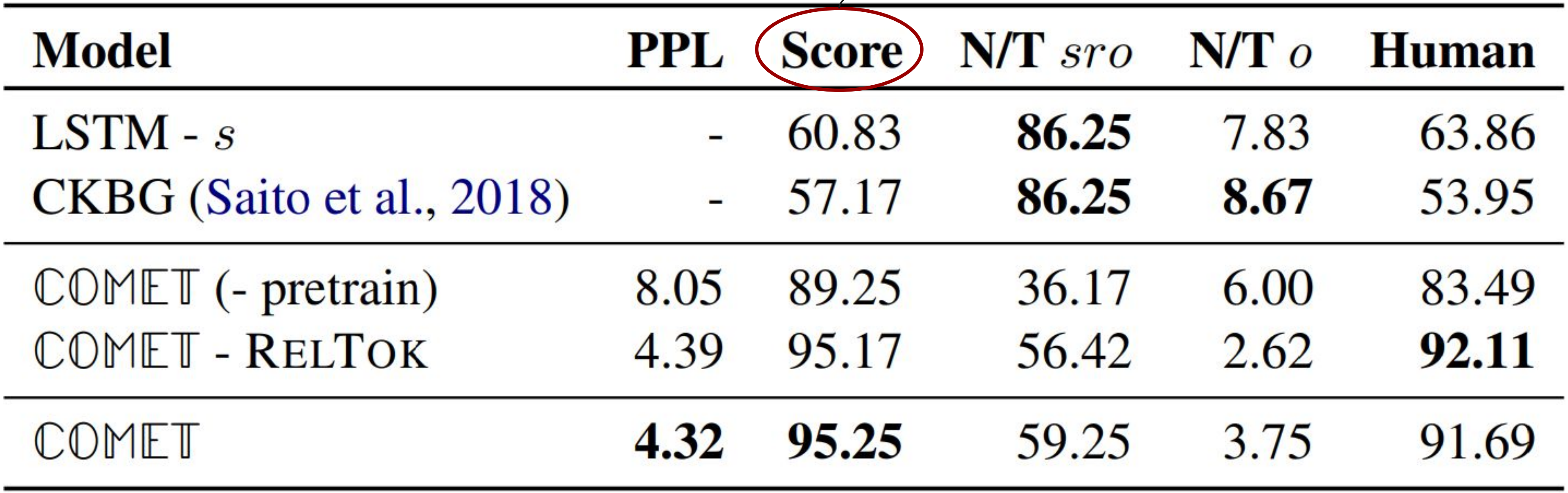
Model	oEffect	oReact	oWant	xAttr	xEffect	xIntent	xNeed	xReact	xWant	Avg
9Enc9Dec (Sap et al., 2019)	22.92	32.92	35.50	52.20	47.52	51.70	48.74	63.57	51.56	45.32
Event2(In)voluntary (Sap et al., 2019)	26.46	36.04	34.70	52.58	46.76	61.32	49.82	71.22	52.44	47.93
Event2PersonX/Y (Sap et al., 2019)	24.72	33.80	35.08	52.98	48.86	53.93	54.05	66.42	54.04	46.41
Event2Pre/Post (Sap et al., 2019)	26.26	34.48	35.78	52.20	46.78	57.77	47.94	72.22	47.94	46.76
COMET (- pretrain)	25.90	35.40	40.76	48.04	47.20	58.88	59.16	64.52	65.66	49.50
COMET	**29.02**	**37.68**	**44.48**	**57.48**	**55.50**	**68.32**	**64.24**	**76.18**	**75.16**	**56.45**

COMET - ATOMIC Novel Triples Annotated by Humans.



Seed Concept	Relation	Generated	Plausible
X holds out X's hand to Y	xAttr	helpful	✓
X meets Y eyes	xAttr	intense	✓
X watches Y every ___	xAttr	observant	✓
X eats red meat	xEffect	gets fat	✓
X makes crafts	xEffect	gets dirty	✓
X turns X's phone	xEffect	gets a text	
X pours ___ over Y's head	oEffect	gets hurt	✓
X takes Y's head off	oEffect	bleeds	✓
X pisses on Y's bonfire	oEffect	gets burned	
X spoils somebody rotten	xIntent	to be mean	
X gives Y some pills	xIntent	to help	✓
X provides for Y's needs	xIntent	to be helpful	✓

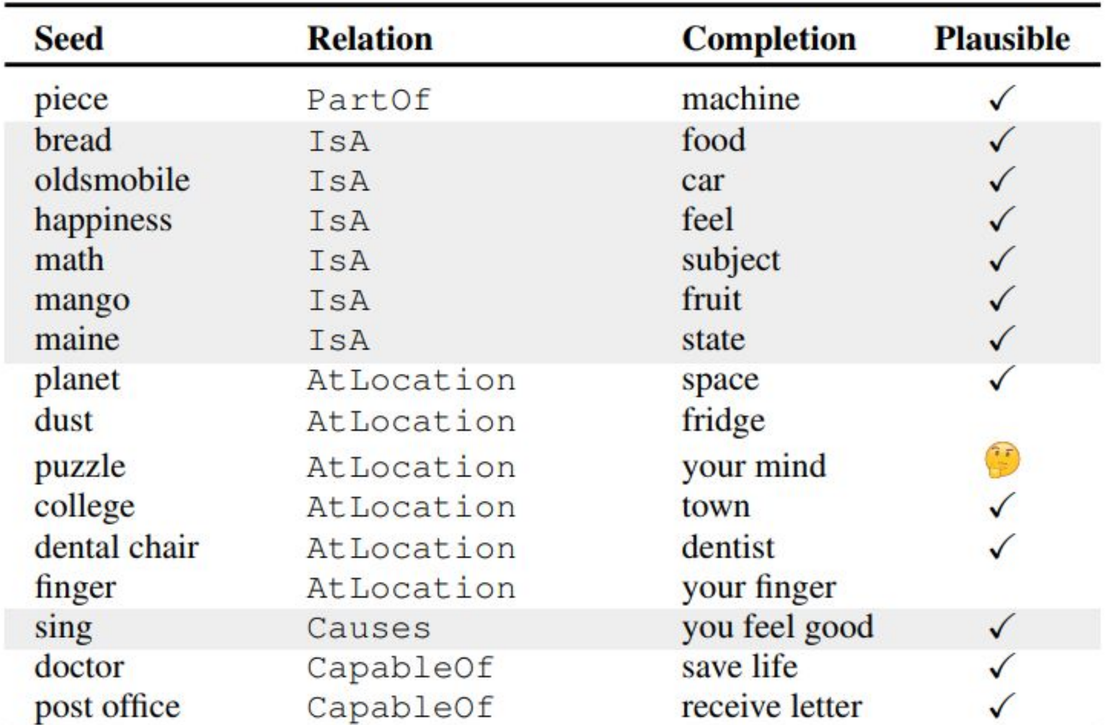
COMET - ConceptNet Results.



Model	PPL	Score	N/T sro	N/T o	Human
LSTM - s	-	60.83	**86.25**	7.83	63.86
CKBG (Saito et al., 2018)	-	57.17	**86.25**	**8.67**	53.95
COMET (- pretrain)	8.05	89.25	36.17	6.00	83.49
COMET - RELTOK	4.39	95.17	56.42	2.62	**92.11**
COMET	**4.32**	**95.25**	59.25	3.75	91.69

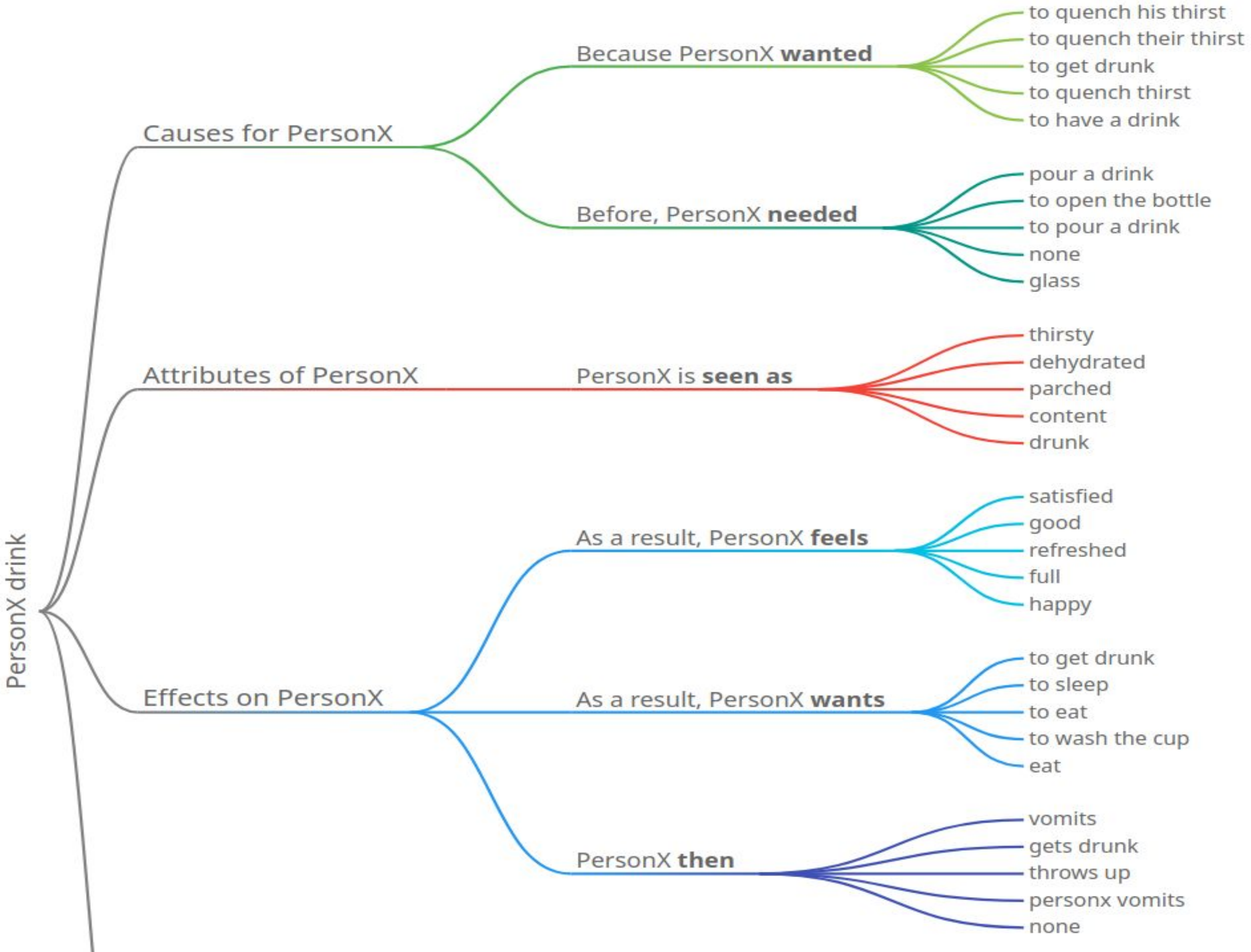
Evaluated by the automatic scorer of Li et al. ACL 2016

COMET - ConceptNet Novel Triples Annotated by Humans.

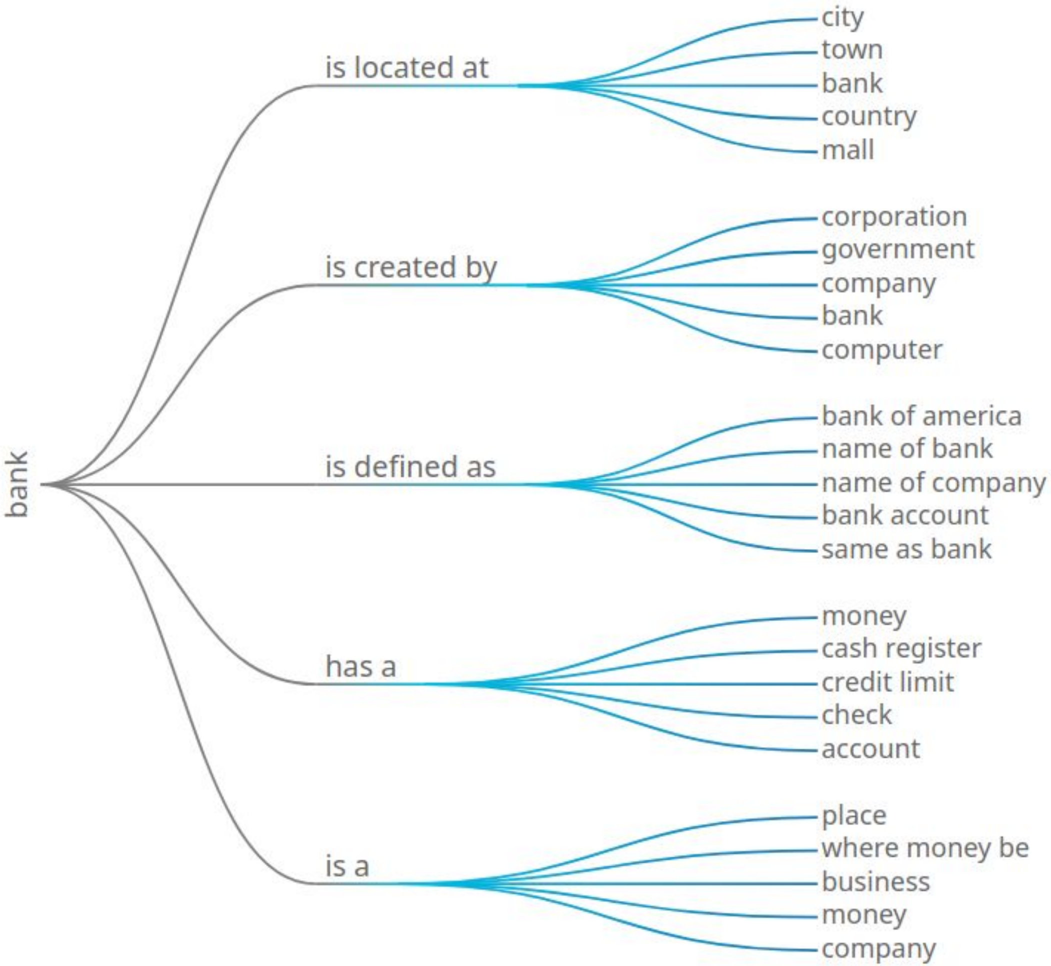


Seed	Relation	Completion	Plausible
piece	PartOf	machine	✓
bread	IsA	food	✓
oldsmobile	IsA	car	✓
happiness	IsA	feel	✓
math	IsA	subject	✓
mango	IsA	fruit	✓
maine	IsA	state	✓
planet	AtLocation	space	✓
dust	AtLocation	fridge	
puzzle	AtLocation	your mind	🤔
college	AtLocation	town	✓
dental chair	AtLocation	dentist	✓
finger	AtLocation	your finger	
sing	Causes	you feel good	✓
doctor	CapableOf	save life	✓
post office	CapableOf	receive letter	✓

COMET


PersonX drink
Causes for PersonX
Because PersonX **wanted**
to quench his thirst
to quench their thirst
to get drunk
to quench thirst
to have a drink
Before, PersonX **needed**
pour a drink
to open the bottle
to pour a drink
none
glass
Attributes of PersonX
PersonX is **seen as**
thirsty
dehydrated
parched
content
drunk
Effects on PersonX
As a result, PersonX **feels**
satisfied
good
refreshed
full
happy
As a result, PersonX **wants**
to get drunk
to sleep
to eat
to wash the cup
eat
PersonX **then**
vomits
gets drunk
throws up
personx vomits
none

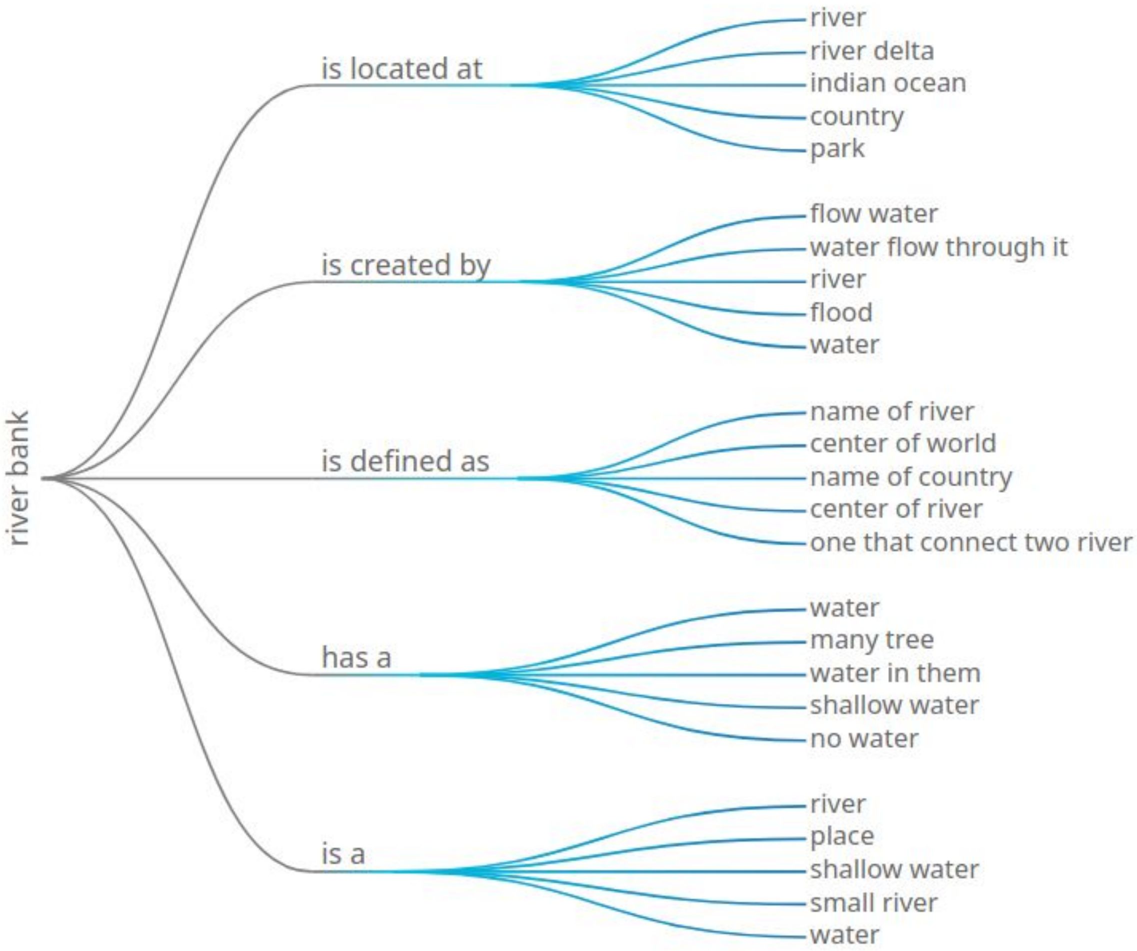
COMET



bank

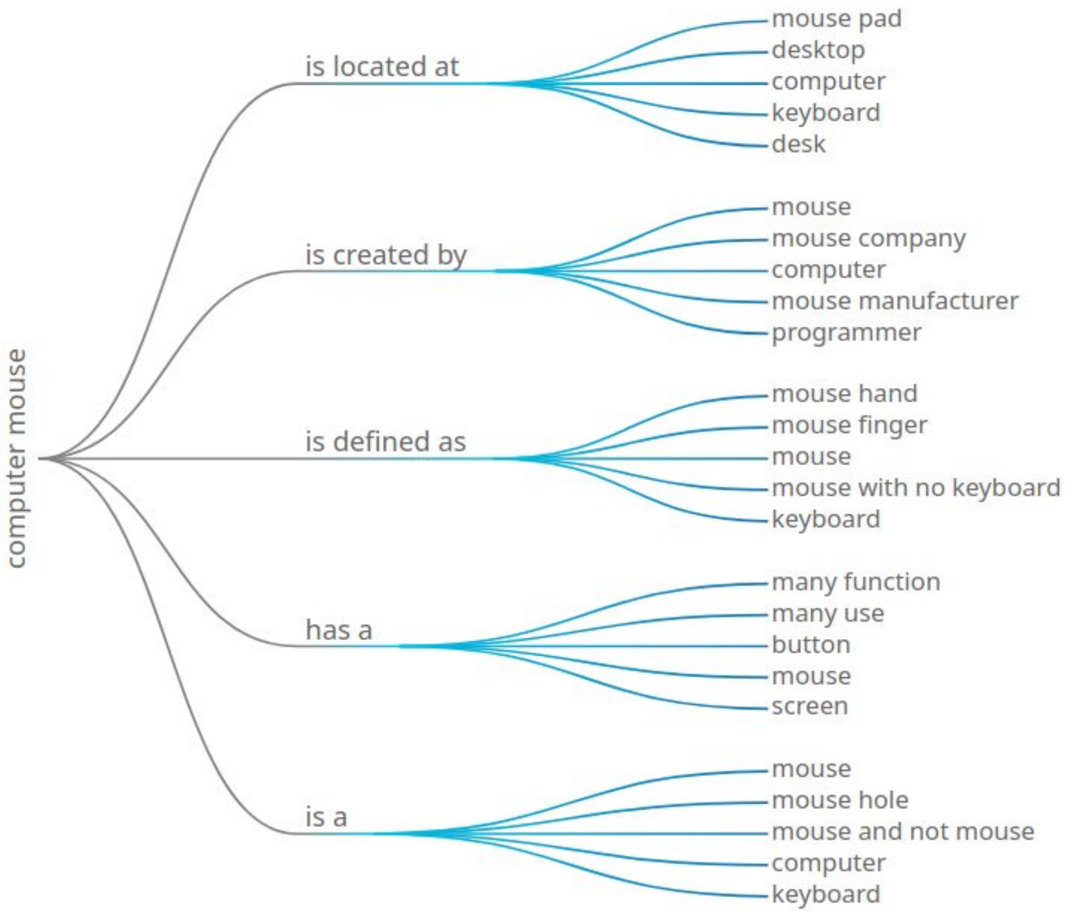
- is located at
  - city
  - town
  - bank
  - country
  - mall
- is created by
  - corporation
  - government
  - company
  - bank
  - computer
- is defined as
  - bank of america
  - name of bank
  - name of company
  - bank account
  - same as bank
- has a
  - money
  - cash register
  - credit limit
  - check
  - account
- is a
  - place
  - where money be
  - business
  - money
  - company

COMET



- river bank
  - is located at
    - river
    - river delta
    - indian ocean
    - country
    - park
  - is created by
    - flow water
    - water flow through it
    - river
    - flood
    - water
  - is defined as
    - name of river
    - center of world
    - name of country
    - center of river
    - one that connect two river
  - has a
    - water
    - many tree
    - water in them
    - shallow water
    - no water
  - is a
    - river
    - place
    - shallow water
    - small river
    - water

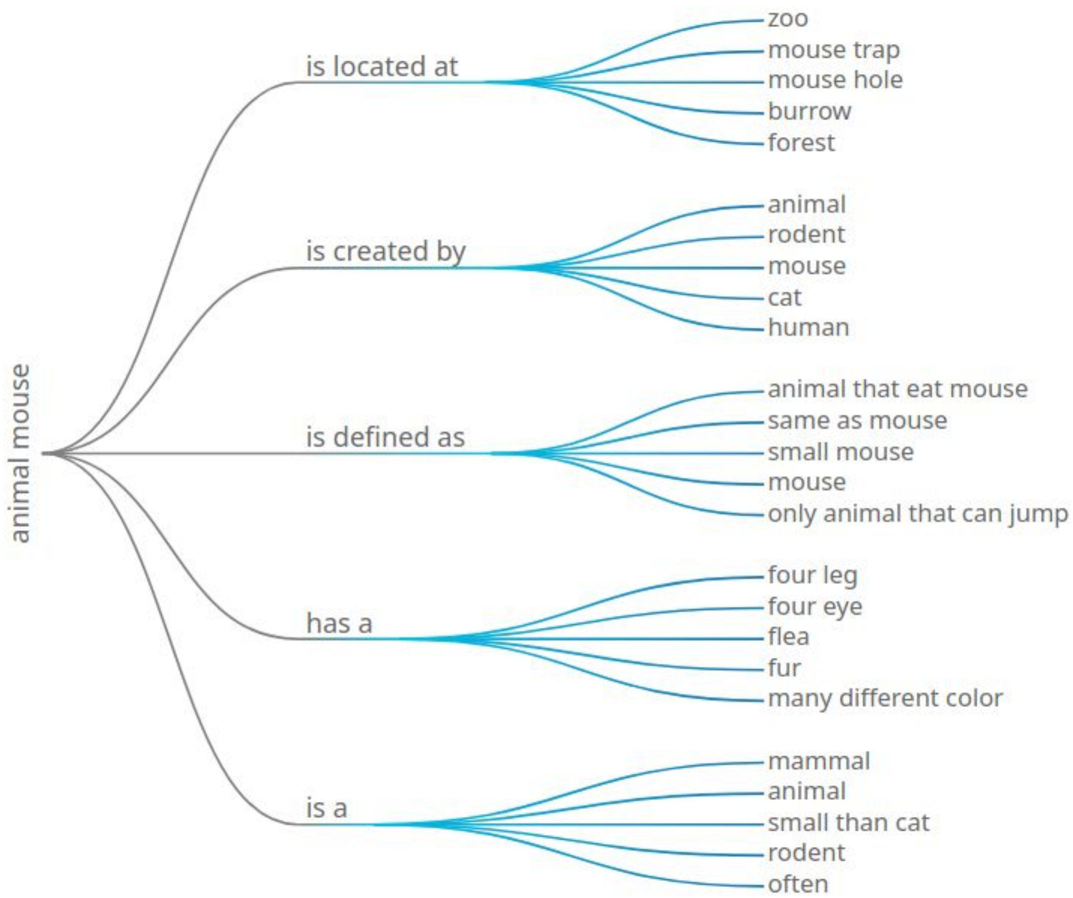
COMET



computer mouse

- is located at
  - mouse pad
  - desktop
  - computer
  - keyboard
  - desk
- is created by
  - mouse
  - mouse company
  - computer
  - mouse manufacturer
  - programmer
- is defined as
  - mouse hand
  - mouse finger
  - mouse
  - mouse with no keyboard
  - keyboard
- has a
  - many function
  - many use
  - button
  - mouse
  - screen
- is a
  - mouse
  - mouse hole
  - mouse and not mouse
  - computer
  - keyboard

COMET


animal mouse
is located at
zoo
mouse trap
mouse hole
burrow
forest
is created by
animal
rodent
mouse
cat
human
is defined as
animal that eat mouse
same as mouse
small mouse
mouse
only animal that can jump
has a
four leg
four eye
flea
fur
many different color
is a
mammal
animal
small than cat
rodent
often

Take-Home Message. ☄️

- Demo available at https://mosaickg.apps.allenai.org.
- Pretrained Neural Language Models showed to encode commonsense knowledge that can be used to generate new relations when fine tuned on a set known triples.
- Can we apply a similar approach to semantic graphs? Would pretrained LM be still that useful to capture this kind of information?
- Interesting direction for OpenIE, i.e., where the set of relations is not given.
- What about fake news detection or fact checking?

Compositional Semantic Parsing Across Graphbanks.

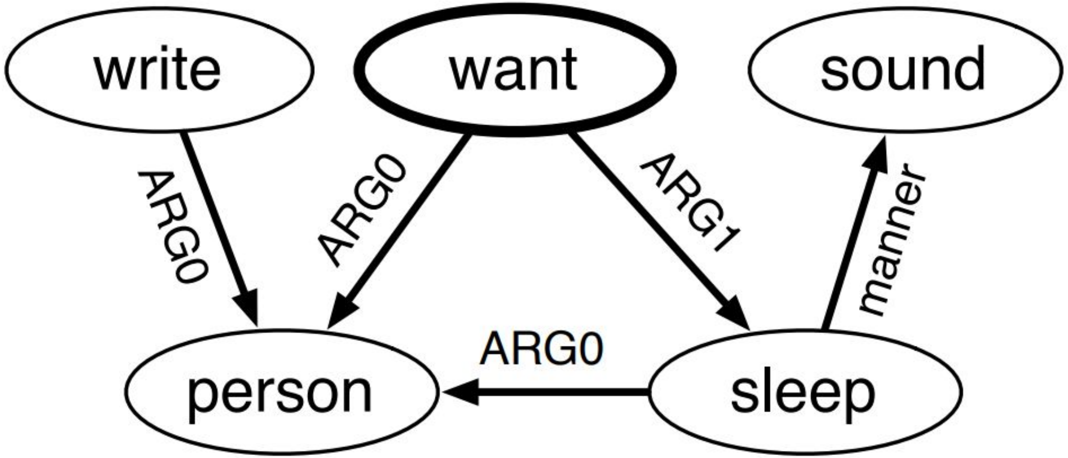
Matthias Linemann, Jonas Groschwitz and Alexander Koller. ACL 2019.

- Based on the AM algebra (Groshwitz et al. IWCS 2017) to represent a sentence in small tree representations that can be then composed to form a complete AMR graph.
- Exploit a type system to ensure correctness of the built AMR graph.
- Exploit existing neural techniques for inducing the super tags (atomic trees) and building a semantically enriched dependency tree that have a 1-1 mapping to an AMR representation.
- The approach can be applied to multiple graphbanks.

The Apply-Modify (AM) Algebra. (Groschwitz et al. IWCS 2017, Groschwitz et al. ACL 2018).

The writer wants to sleep soundly.

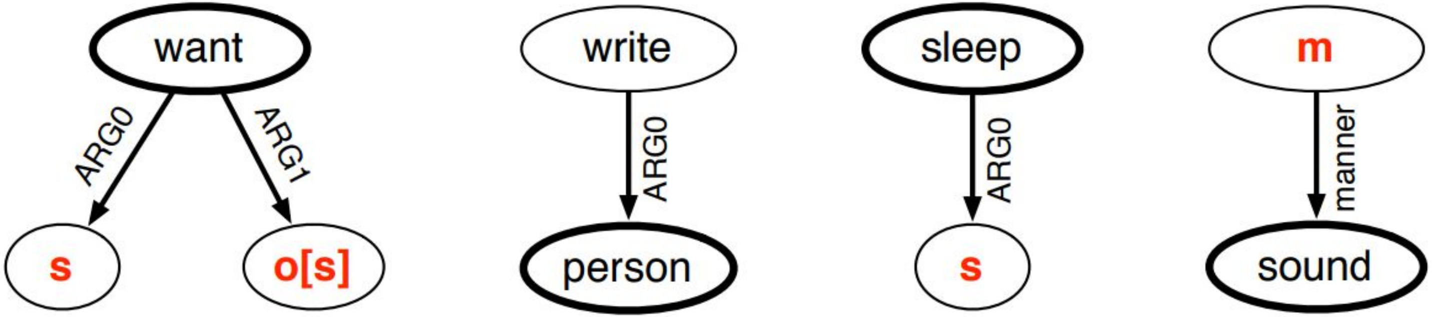
AMR:



The Apply-Modify (AM) Algebra. (Groschwitz et al. IWCS 2017, Groschwitz et al. ACL 2018).

The writer wants to sleep soundly.

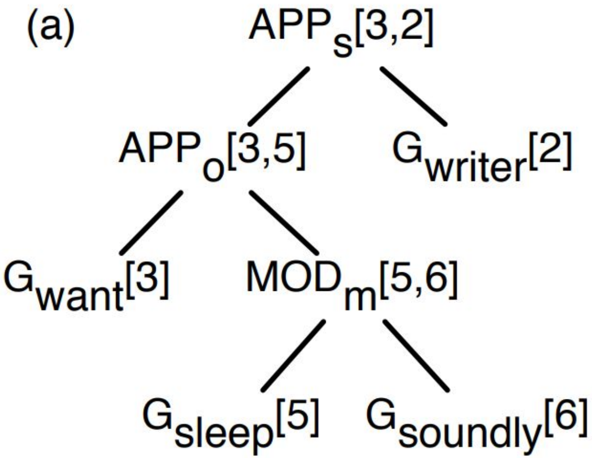
as-graphs:



The Apply-Modify (AM) Algebra. (Groschwitz et al. IWCS 2017, Groschwitz et al. ACL 2018).

The writer wants to sleep soundly.

indexed AM term:



From AMR to AM Terms.

- Authors divided this task in 3 steps:
- Break each AMR up into atomic as-graphs and identify their roots: assume 1 root per atomic graph and exploit hand-crafted rules.
- Assign sources and annotations to build as-graphs out of atomic graphs: add annotations to each leaf node (if necessary) based on the structure of the graph and on the incoming edges.
- Combine them into indexed AM terms: Exploit a dependency parser to combine the as-graphs.

Training.

- The graphbank is first used to create a set of AM terms.

- A BiLSTM-model is trained in order to predict for each token in a sentence the AM term it should be associated with.

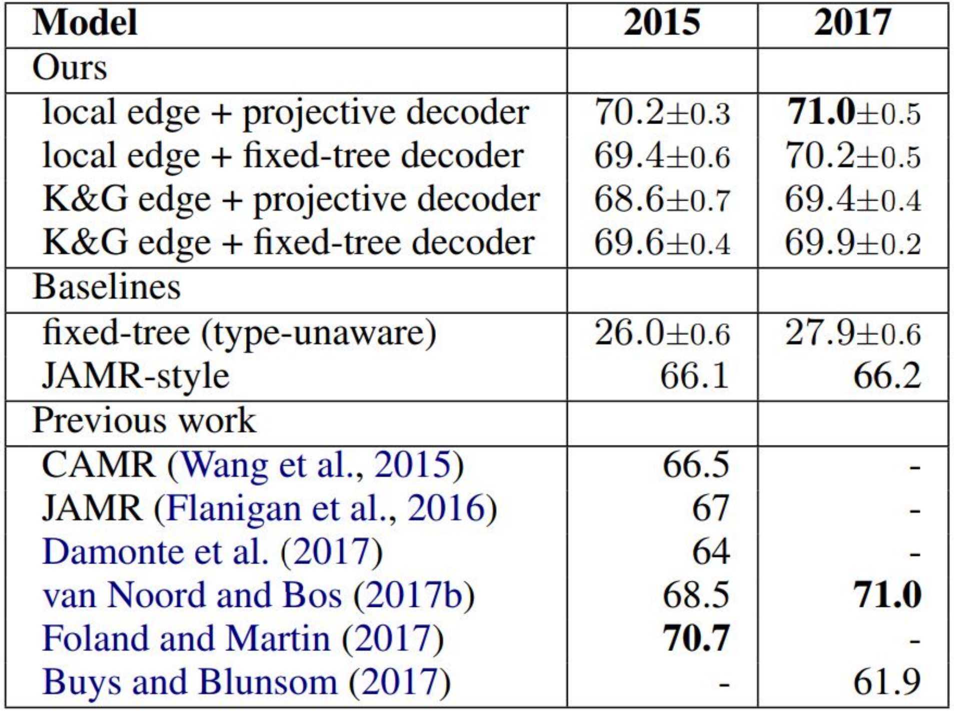
- A dependency parser algorithm is used to combine the atomic trees into a well-formed AM dependency tree.

- The AM dependency tree can be uniquely evaluated into a AMR graph representing the input sentence.

Evaluation.

- Training data: LDC datasets with AMR annotations from 2015 and 2017.
- Evaluation Metric: Smatch, i.e., the maximum f-score obtainable via a one-to one matching of variables between the two AMRs (Cai and Knight ACL 2013).
- Competitors: Various ARM parser from the past, inter alia, Foland and Marting 2017 and Buys and Blunsom 2017.
- Baselines: 1) Type-unaware fixed-tree baseline which discard the type information when merging the AM terms, 2) JAMR-style baseline which does not have any explicit information about which is the root node in a AM term.

Results.



Model	**2015**	**2017**
Ours		
local edge + projective decoder	70.2±0.3	**71.0**±0.5
local edge + fixed-tree decoder	69.4±0.6	70.2±0.5
K&G edge + projective decoder	68.6±0.7	69.4±0.4
K&G edge + fixed-tree decoder	69.6±0.4	69.9±0.2
Baselines		
fixed-tree (type-unaware)	26.0±0.6	27.9±0.6
JAMR-style	66.1	66.2
Previous work		
CAMR (Wang et al., 2015)	66.5	-
JAMR (Flanigan et al., 2016)	67	-
Damonte et al. (2017)	64	-
van Noord and Bos (2017b)	68.5	**71.0**
Foland and Martin (2017)	**70.7**	-
Buys and Blunsom (2017)	-	61.9

Take-Home Message:

- The paper showed an interesting and effective method for explicitly represent the compositional structure of AMR.

- This explicit compositionality potentially allows to scale more easily over different domains as long as one can build a large inventory of atomic graphs.

- Still the method rely on hand-crafter rules to generate as-graphs from full AMR graphs and the bank of as-graphs can only be generated from AMR graphs and not from raw sentences.