CS224N : Lecture 15 - Add Knowledge to Language Models

What does a LM know?

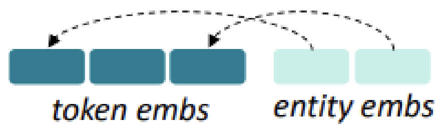
- Takeaway: predictions generally make sense (e.g. the correct types), but are not all factually correct.

- Why might this happen?
  - Unseen facts: some facts may not have occurred in the training corpora at all
  - Rare facts: LM hasn't seen enough examples during training to memorize the fact
  - Model sensitivity: LM may have seen the fact during training, but is sensitive to the phrasing of the prompt
    - Correctly answers "x was *made* in y" templates but not "x was *created* in y"

- The inability to *reliably* recall knowledge is a key challenge facing LMs today!
  - Recent works have found LMs can recover *some* knowledge, but have a way to go.

Techniques to add knowledge to LMs

Techniques to add knowledge to LMs



Add pretrained entity embeddings

- ERNIE
- KnowBERT

Use an external memory

- KGLM
- kNN-LM

Modify the training data

- WKLM
- ERNIE (another!), salient span masking

Evaluating knowledge in LMs

LAnguage Model Analysis (LAMA) Probe [Petroni et al., EMNLP 2019]

- How much relational (commonsense and factual) knowledge is already in off-the-shelf language models?
  - Without any additional training or fine-tuning

- Manually constructed a set of cloze statements to assess a model's ability to predict a missing token. *Examples:*

The theory of relativity was developed by [MASK].
The native language of Mammootty is [MASK].
Ravens can [MASK].
You are likely to find a overflow in a [MASK].

LAnguage Model Analysis (LAMA) Probe [Petroni et al., EMNLP 2019]

- Generate cloze statements from KG triples and question-answer pairs
- Compare LMs to supervised relation extraction (RE) and question answering systems
- **Goal:** evaluate knowledge present in existing pretrained LMs (this means they may have different pretraining corpora!)

Mean precision at one (P@1)

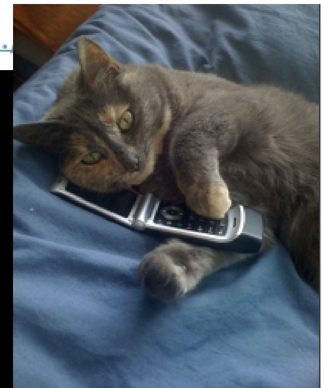
BERT struggles on N-to-M relations

Corpus	DrQA	RE baseline	fairseq-fconv	Transformer-XL	ELMo	ELMo (5.5B)	BERT-base	BERT-large
Google-RE	-	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	-	**33.8**	8.9	18.3	4.7	7.1	31.1	32.2
ConceptNet	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	**37.5**	-	3.6	3.9	1.6	4.3	14.1	17.4

LMs are NOT finetuned!

49

LAnguage Model Analysis (LAMA) Probe [Petroni et al.



You can try out examples to assess knowledge in popular LMs: https://github.com/facebookresearch/LAMA

The cat is on the [MASK].

```
bert:

| Top10 predictions
0       phone          -2.345
1       floor          -2.630
2       ground         -2.968
3       couch          -3.387
4       move           -3.649
5       roof           -3.651
6       way            -3.718
7       run            -3.757
8       bed            -3.802
9       left           -3.965

--------------------------------------------------------------------------
index   token          log_prob    prediction      log_prob    rank@1000
--------------------------------------------------------------------------

1       The            -5.547      .               -0.607      14
2       cat            -0.367      cat             -0.367      0
3       is             -0.019      is              -0.019      0
4       on             -0.001      on              -0.001      0
5       the            -0.002      the             -0.002      0
6       [MASK]         -14.321     phone           -2.345      -1
7       .              -0.002      .               -0.002      0
```

[1] Example courtesy of the authors at link above.

50

A More Challenging Probe: LAMA-UnHelpful Names (LAMA-UHN)

[Poerner et al., EMNLP 2020]

- Key idea: Remove the examples from LAMA that can be answered without relational knowledge
- Observation: BERT may rely on surface forms of entities to make predictions
  - String match between subject and object
  - "Revealing" person name
    - Name can be a (possibly incorrect) prior for native language, place of birth, nationality, etc.
- BERT's score on LAMA drops ~8% with LAMA-UHN
  - Knowledge-enhanced model E-BERT score drops only <1%

Native language of French-speaking actors according to BERT

Person Name	BERT
Jean Marais	French
Daniel Ceccaldi	Italian
Orane Demazis	Albanian
Sylvia Lopez	Spanish
Annick Alane	English

--