

Natural Language Processing with Deep Learning

CS224N



The Future of Deep Learning + NLP
Kevin Clark

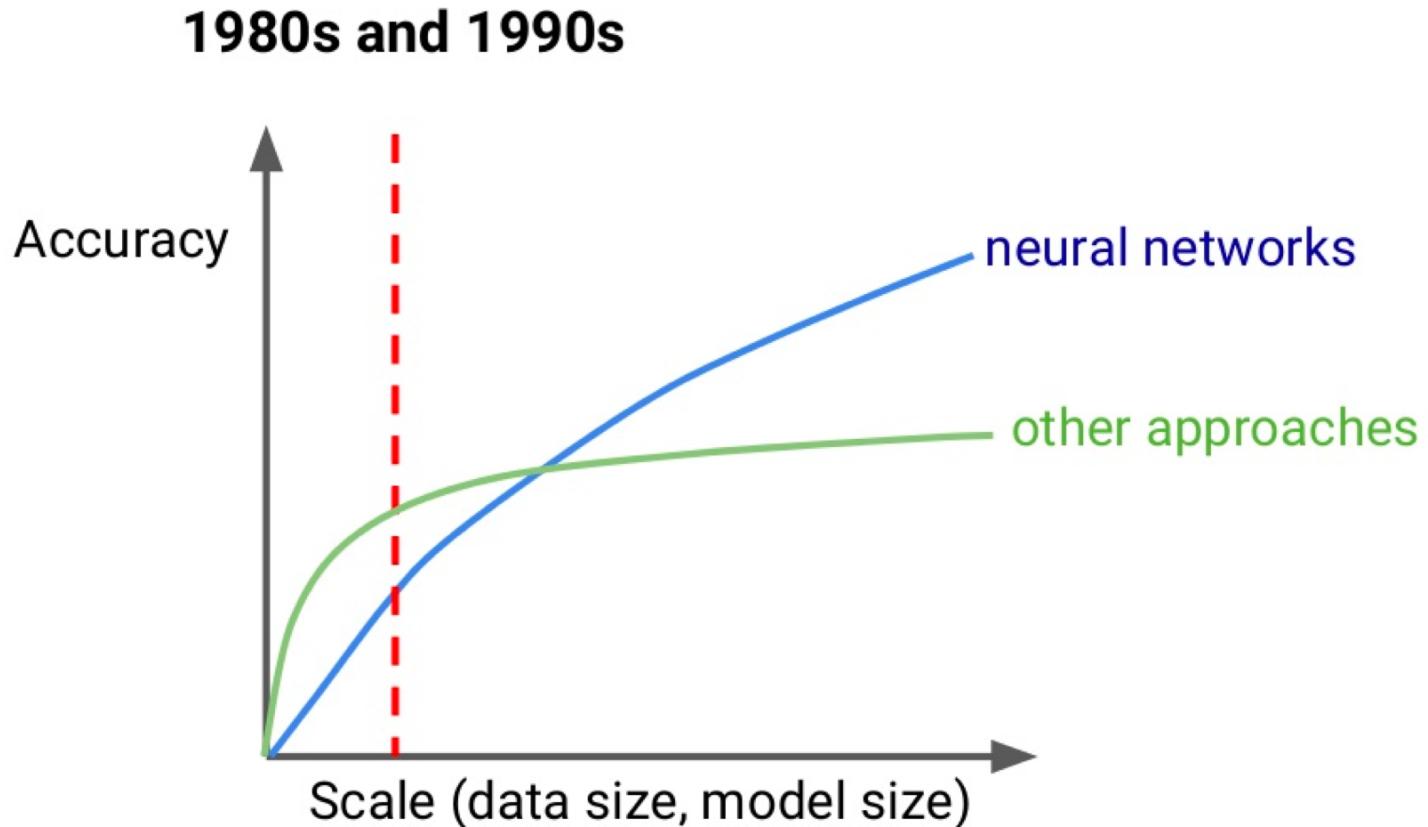
Deep Learning for NLP 5 years ago

- No Seq2Seq
- No Attention
- No large-scale QA/reading comprehension datasets
- No TensorFlow or Pytorch
- ...

Future of Deep Learning + NLP

- **Harnessing Unlabeled Data**
 - Back-translation and unsupervised machine translation
 - Scaling up pre-training and GPT-2
- **What's next?**
 - Risks and social impact of NLP technology
 - Future directions of research

Why has deep learning been so successful recently?



NLP Datasets

- Even for English, most tasks have 100K or less labeled examples.
- And there is even less data available for other languages.
 - There are thousands of languages, hundreds with > 1 million native speakers
 - <10% of people speak English as their first language
- Increasingly popular solution: use **unlabeled** data.

Using Unlabeled Data for Translation

Machine Translation Data

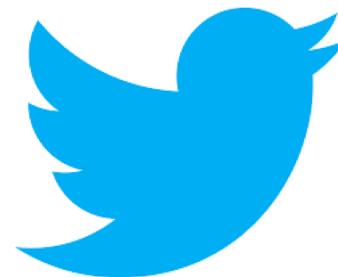
- Acquiring translations required human expertise
 - Limits the size and domain of data



TED

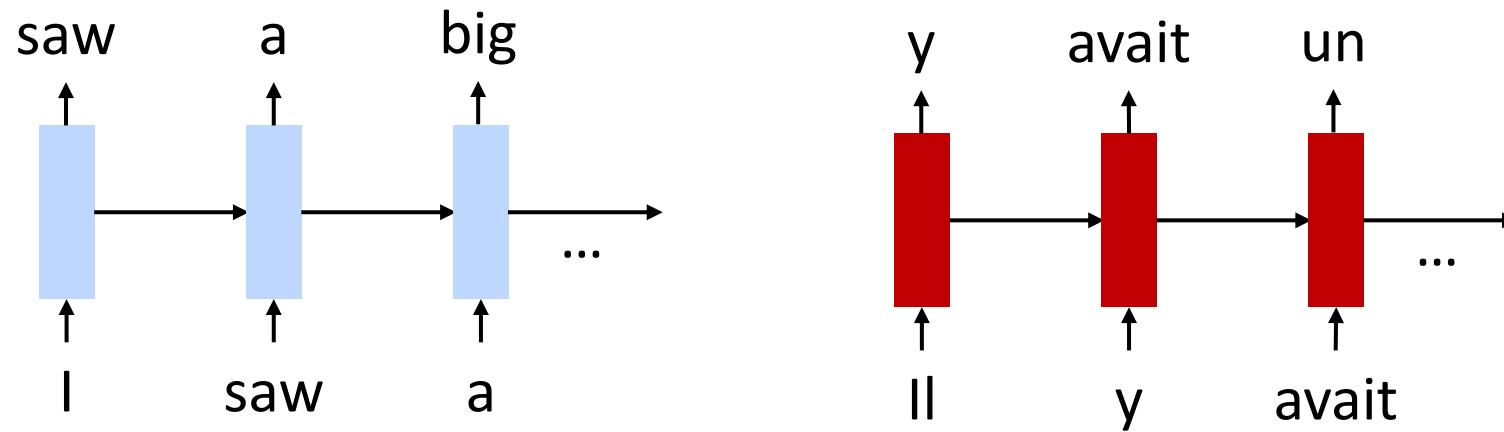
PS

- Monolingual text is easier to acquire!

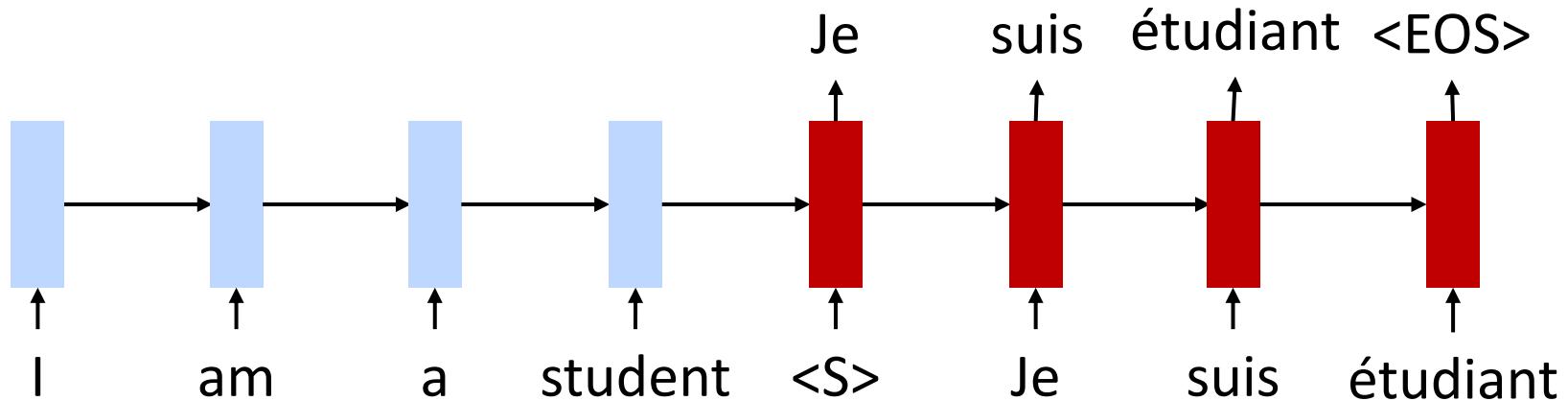


Pre-Training

1. Separately Train Encoder and Decoder as Language Models

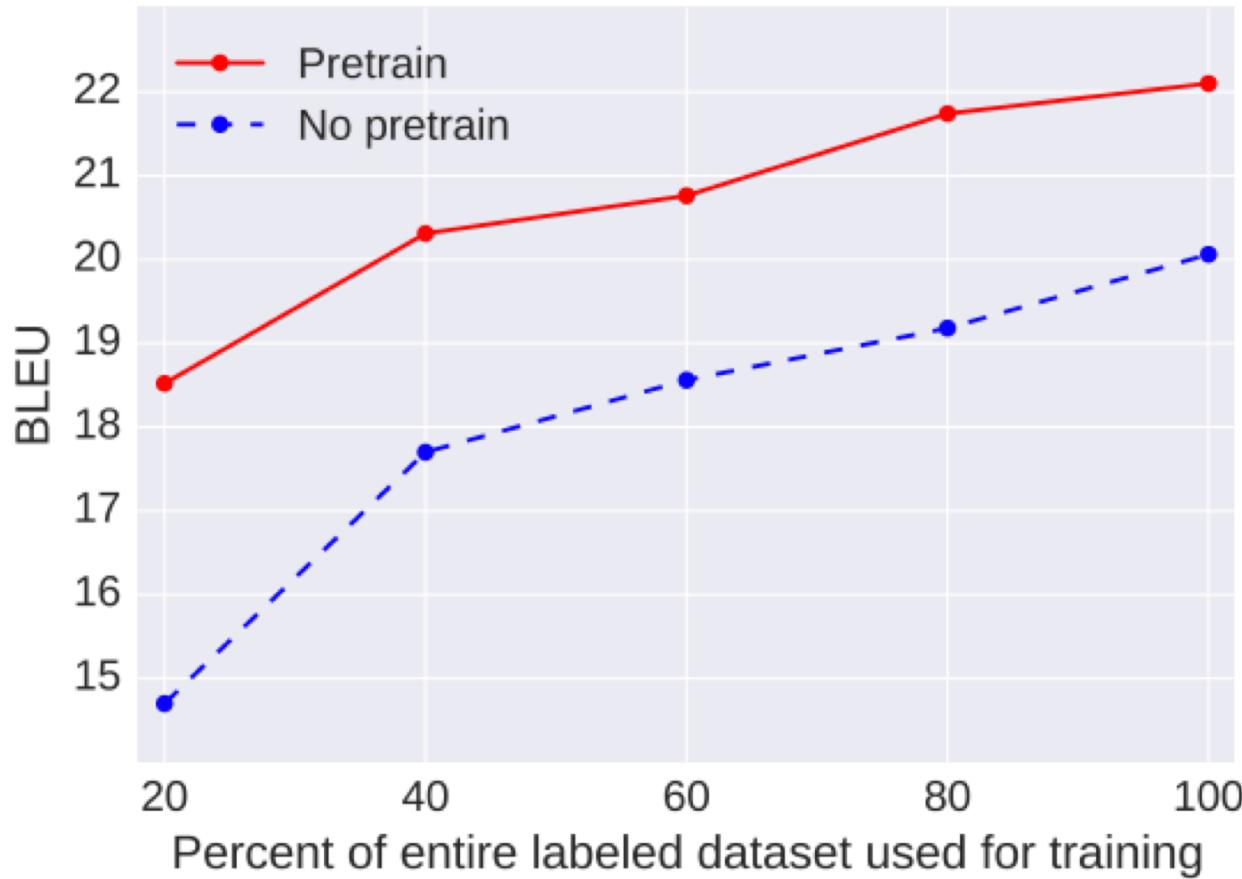


2. Then Train Jointly on Bilingual Data



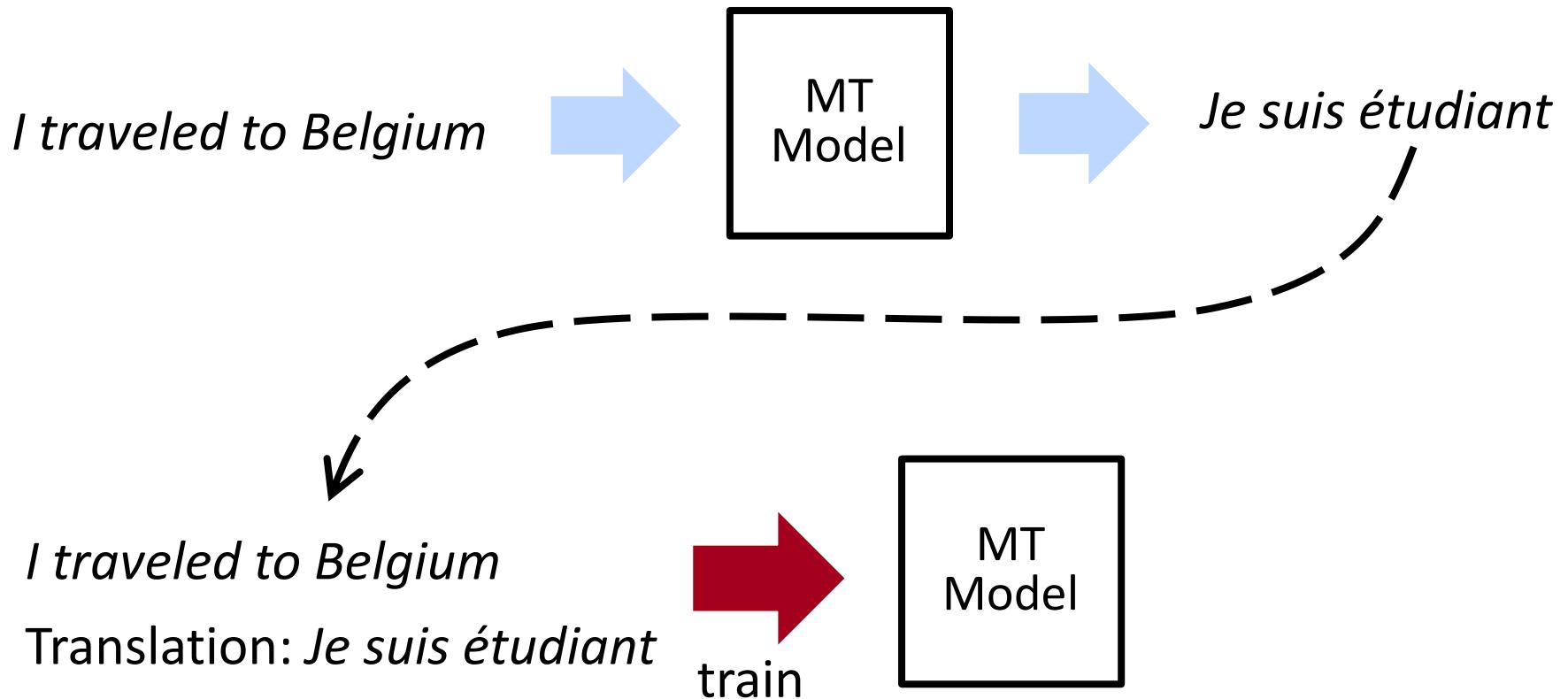
Pre-Training

- English -> German Results: 2+ BLEU point improvement



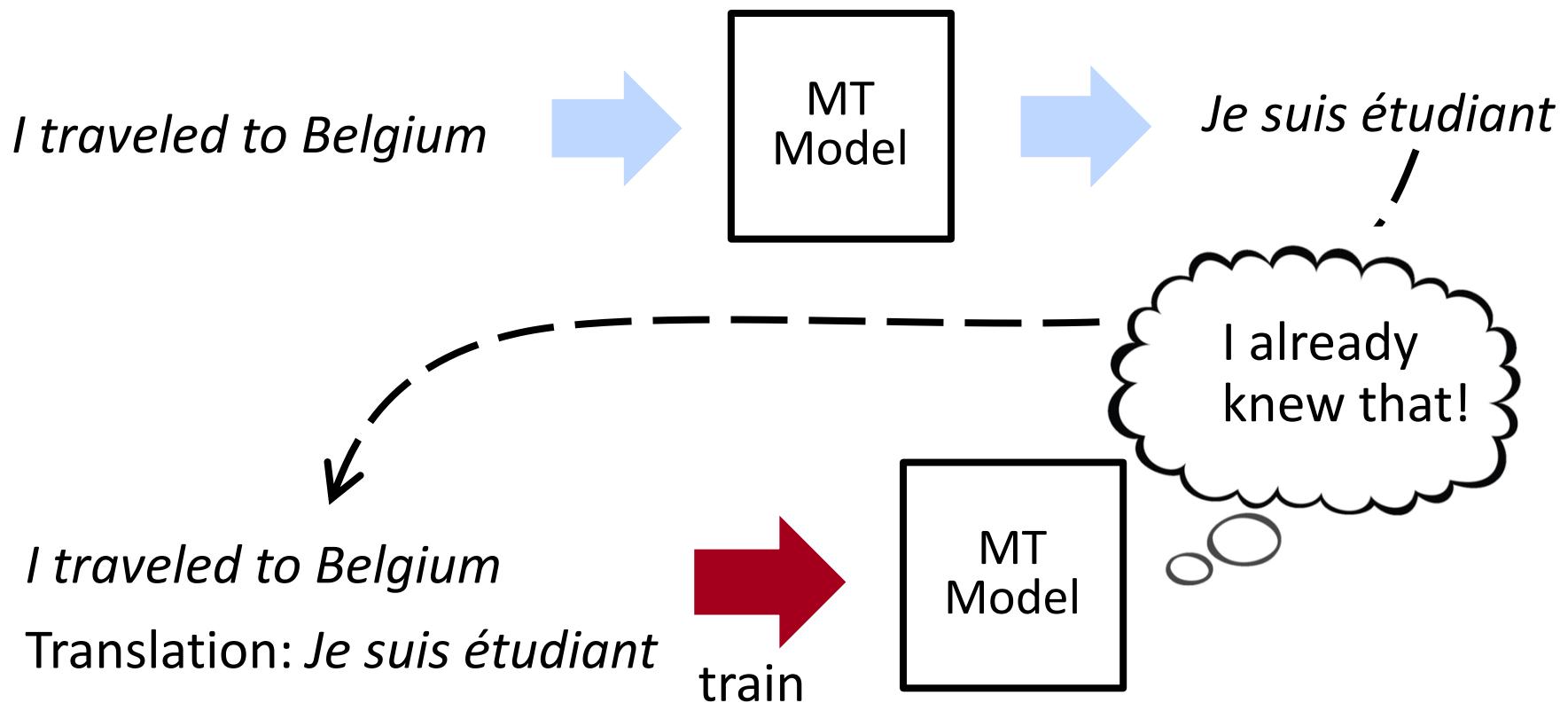
Self-Training

- Problem with pre-training: no “interaction” between the two languages during pre-training
- Self-training: label unlabeled data to get noisy training examples



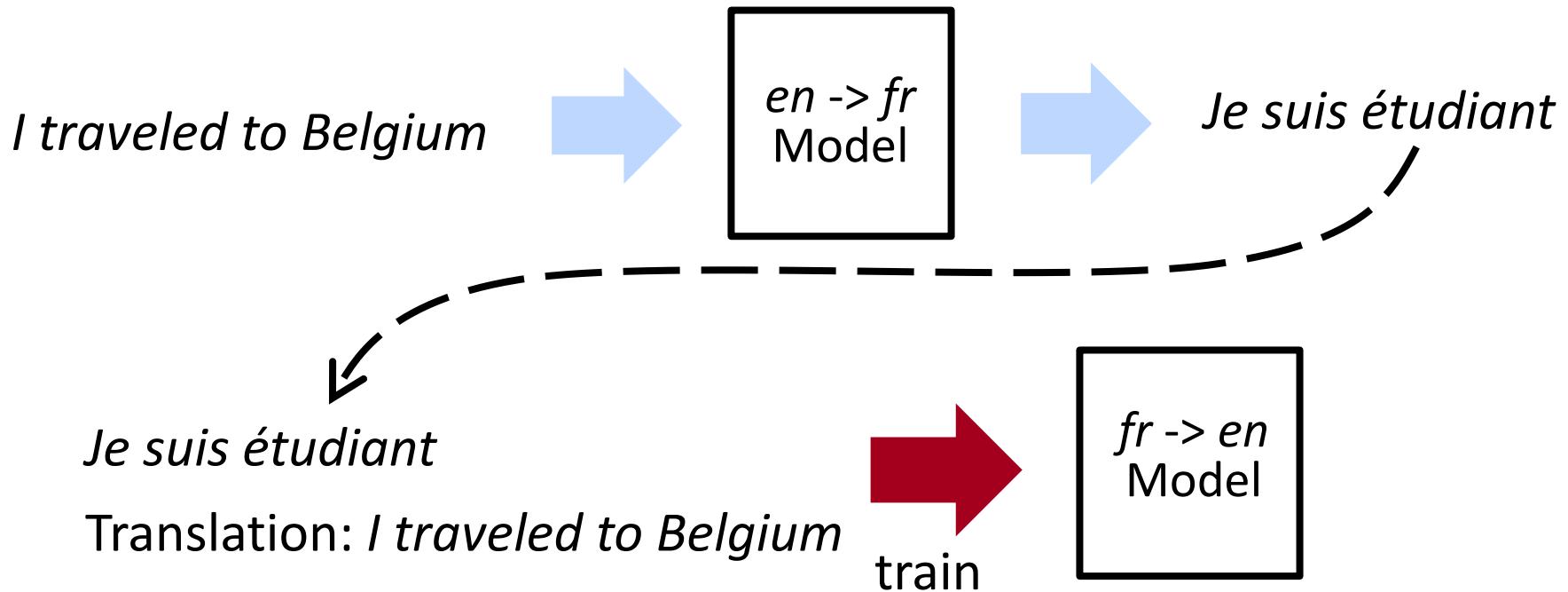
Self-Training

- Circular?



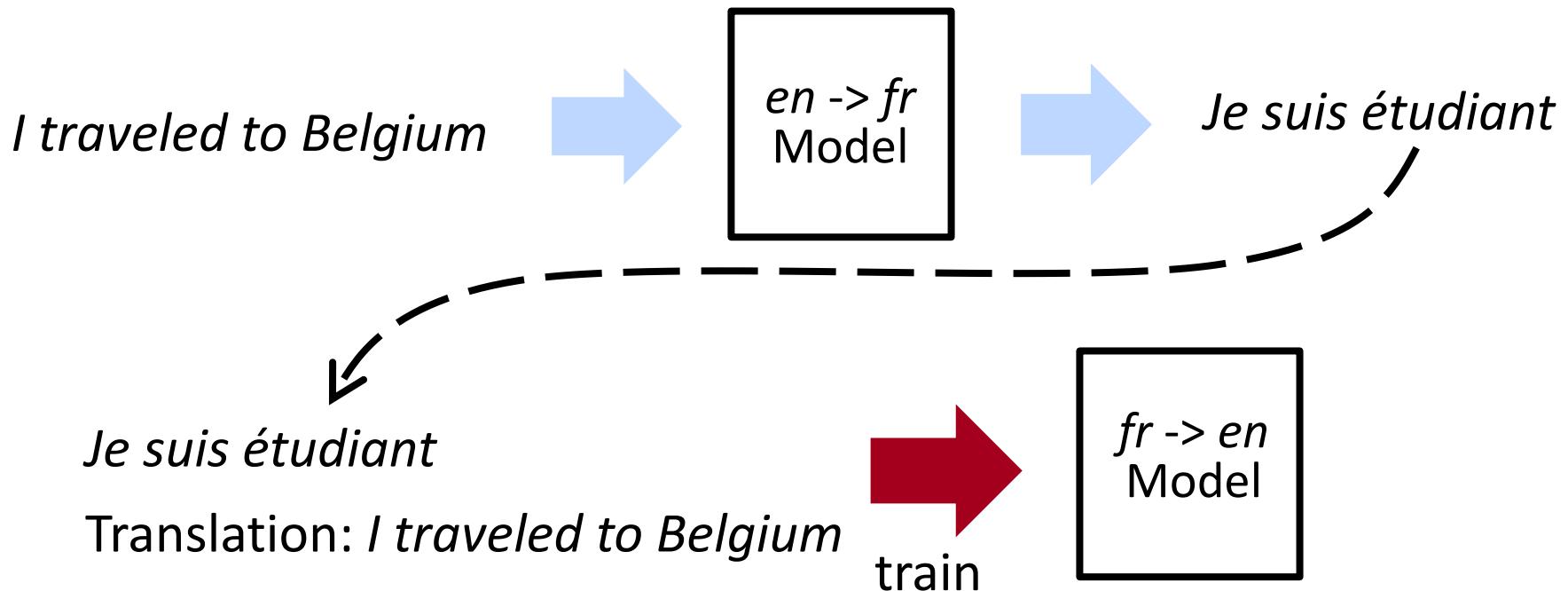
Back-Translation

- Have two machine translation models going in opposite directions (*en* -> *fr*) and (*fr* -> *en*)



Back-Translation

- Have two machine translation models going in opposite directions ($en \rightarrow fr$) and ($fr \rightarrow en$)



- No longer circular
- Models never see “bad” translations, only bad inputs

Large-Scale Back-Translation

- 4.5M English-German sentence pairs and 226M monolingual sentences

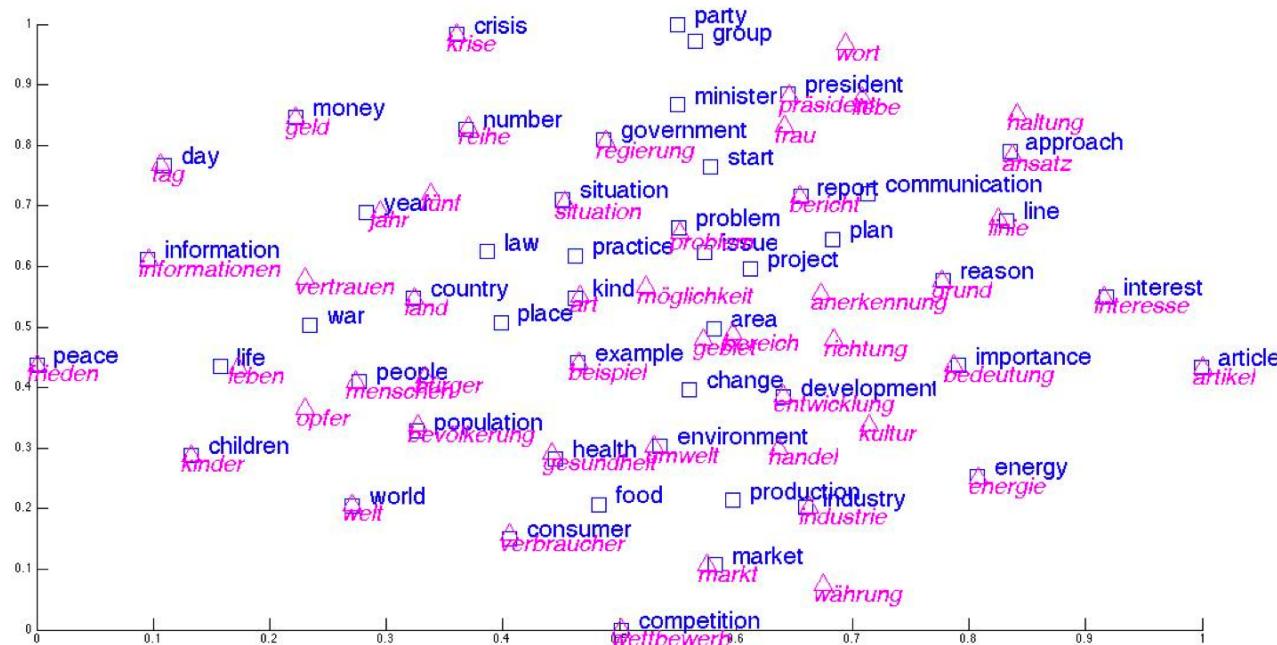
Citation	Model	BLEU
Shazeer et al., 2017	Best Pre-Transformer Result	26.0
Vaswani et al., 2017	Transformer	28.4
Shaw et al, 2018	Transformer + Improved Positional Embeddings	29.1
Edunov et al., 2018	Transformer + Back-Translation	35.0

What if there is no Bilingual Data?

Unsupervised Word Translation

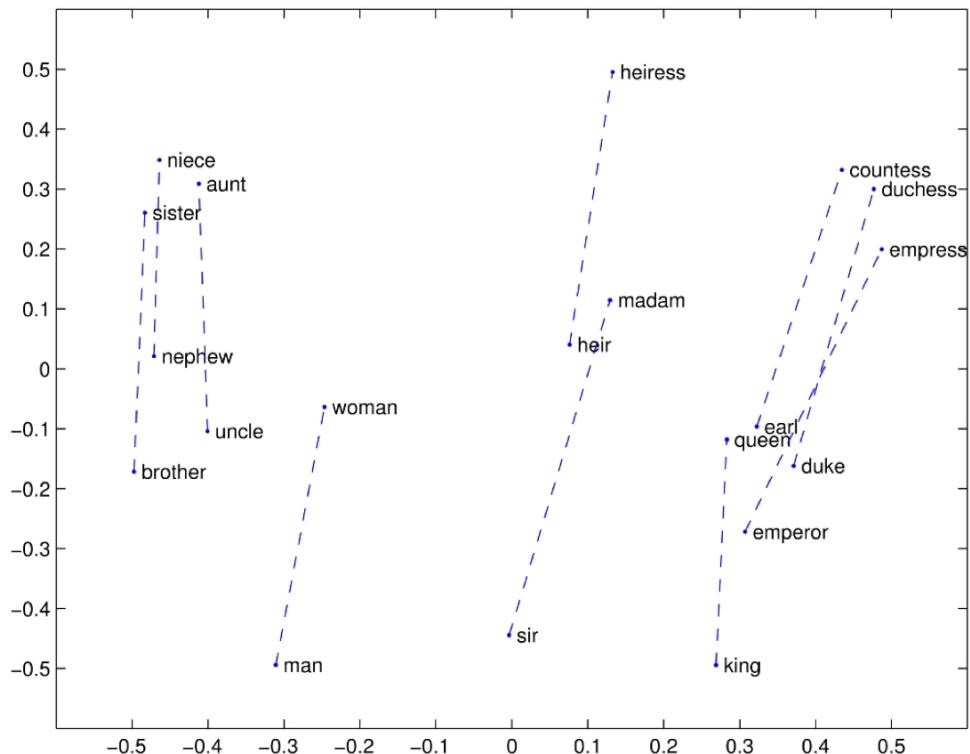
Unsupervised Word Translation

- *Cross-lingual word embeddings*
 - Shared embedding space for both languages
 - Keep the normal nice properties of word embeddings
 - But also want words close to their translations
- Want to learn from monolingual corpora



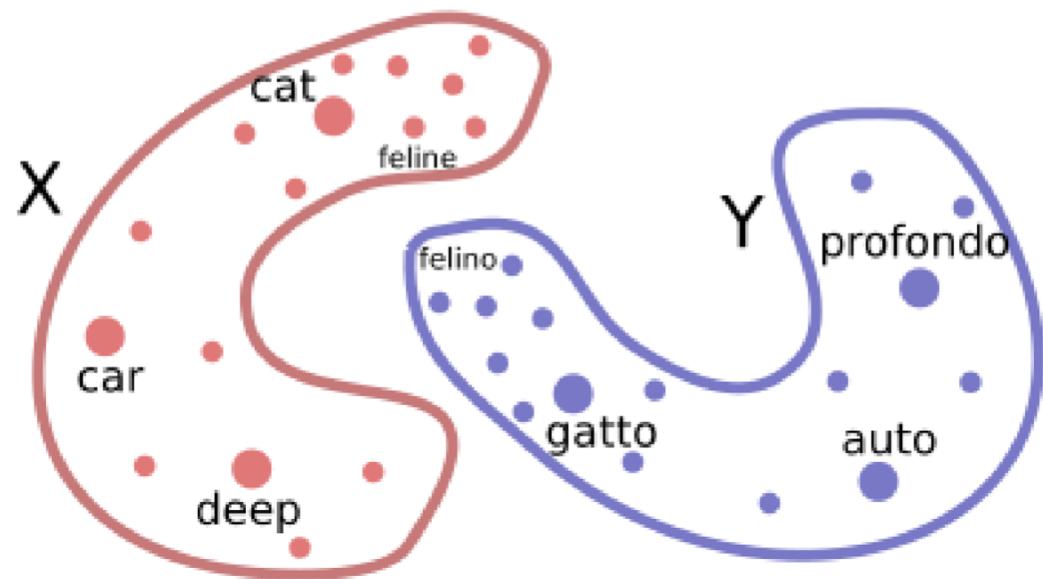
Unsupervised Word Translation

- Word embeddings have a lot of structure
- Assumption: that structure should be similar across languages



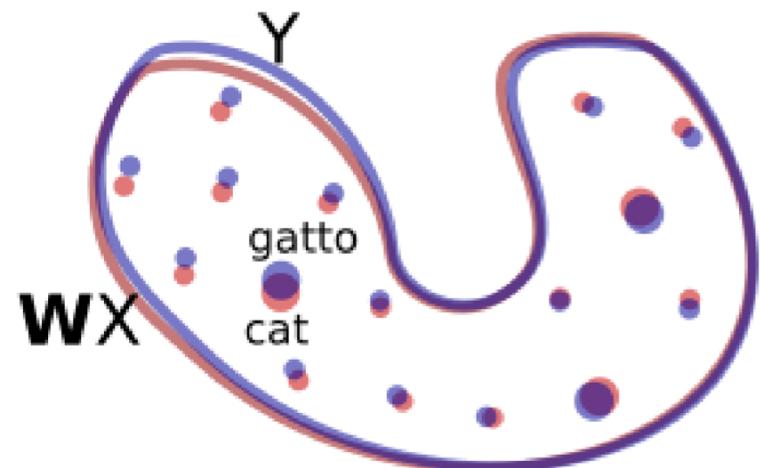
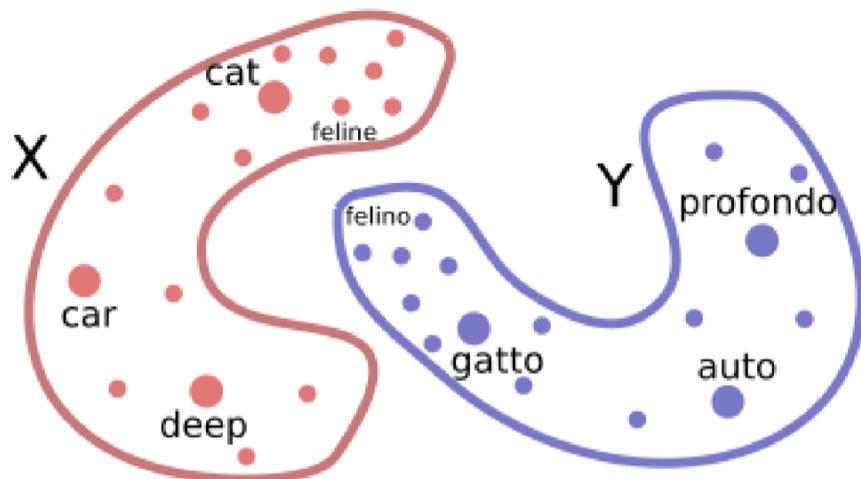
Unsupervised Word Translation

- Word embeddings have a lot of structure
- Assumption: that structure should be similar across languages



Unsupervised Word Translation

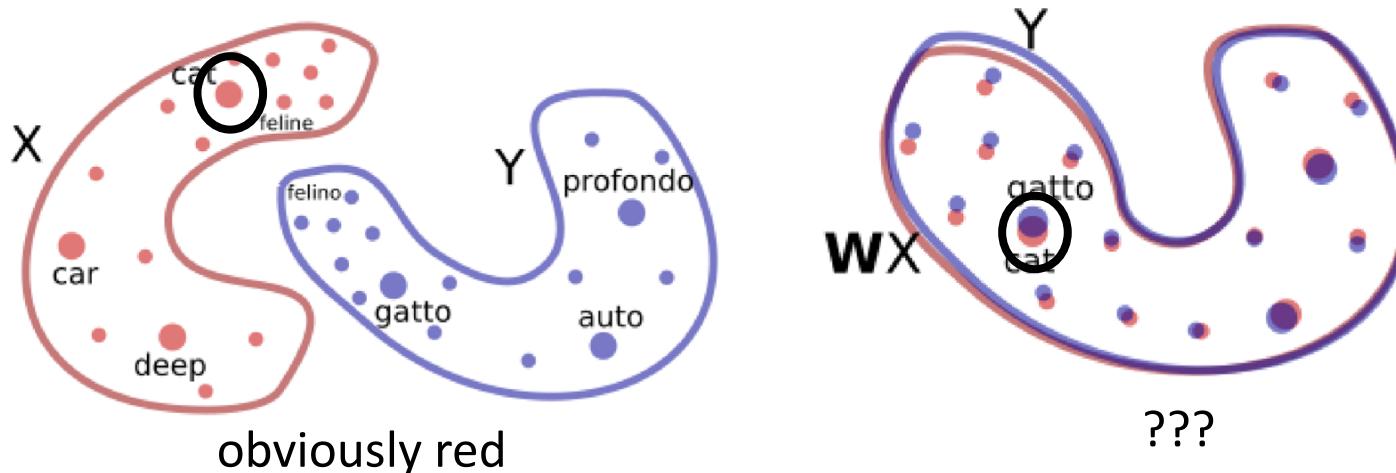
- First run word2vec on monolingual corpora, getting words embeddings X and Y
- Learn an (orthogonal) matrix W such that $WX \sim Y$



Unsupervised Word Translation

- Learn W with *adversarial training*.
- Discriminator: predict if an embedding is from Y or it is a transformed embedding Wx originally from X .
- Train W so the Discriminator gets “confused”

Discriminator predicts: is the circled point red or blue?

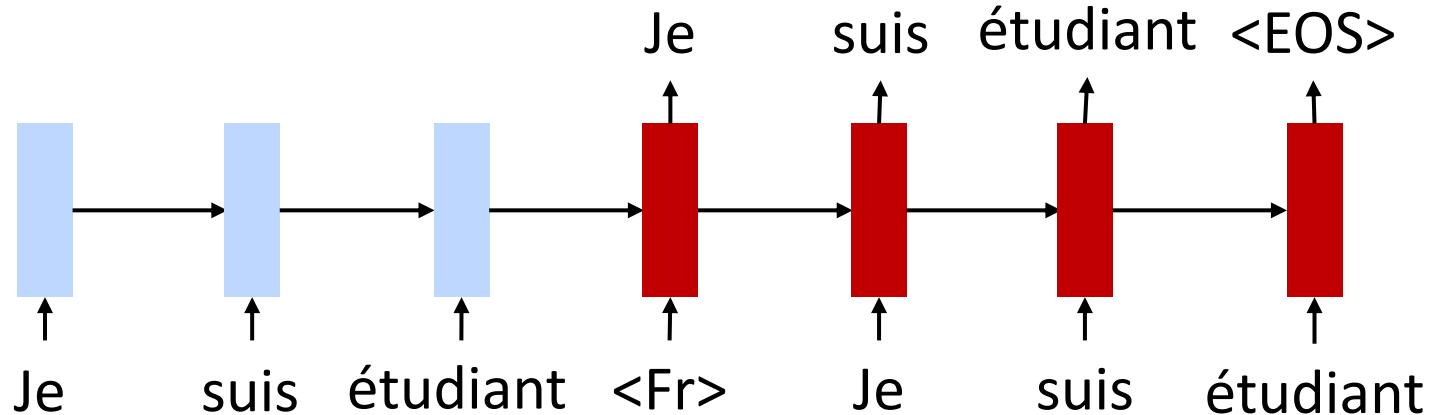
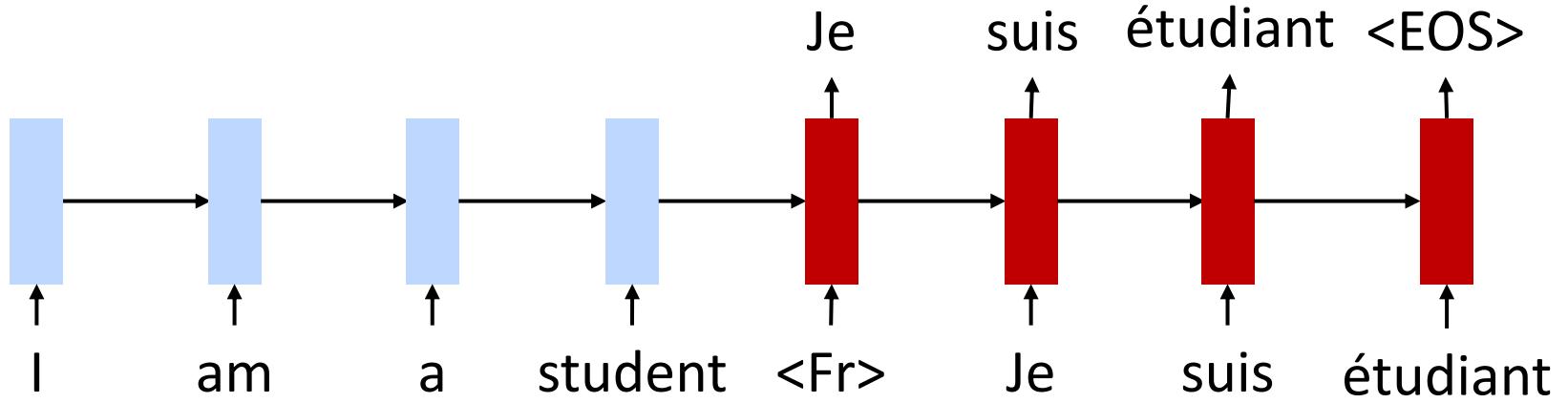


- Other tricks can be used to further improve performance, see [Word Translation without Parallel Data](#)

Unsupervised Machine Translation

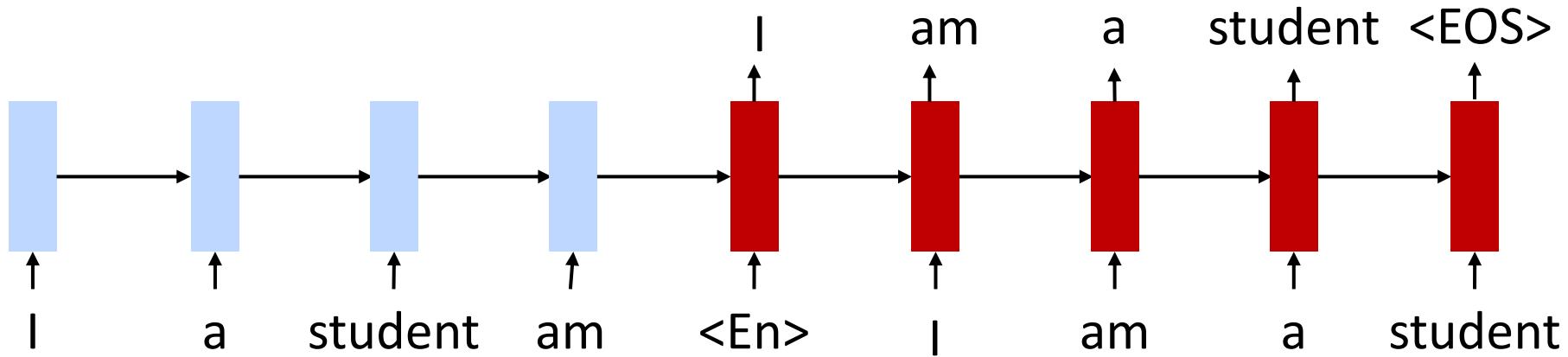
Unsupervised Machine Translation

- Model: **same** encoder-decoder used for both languages
 - Initialize with cross-lingual word embeddings



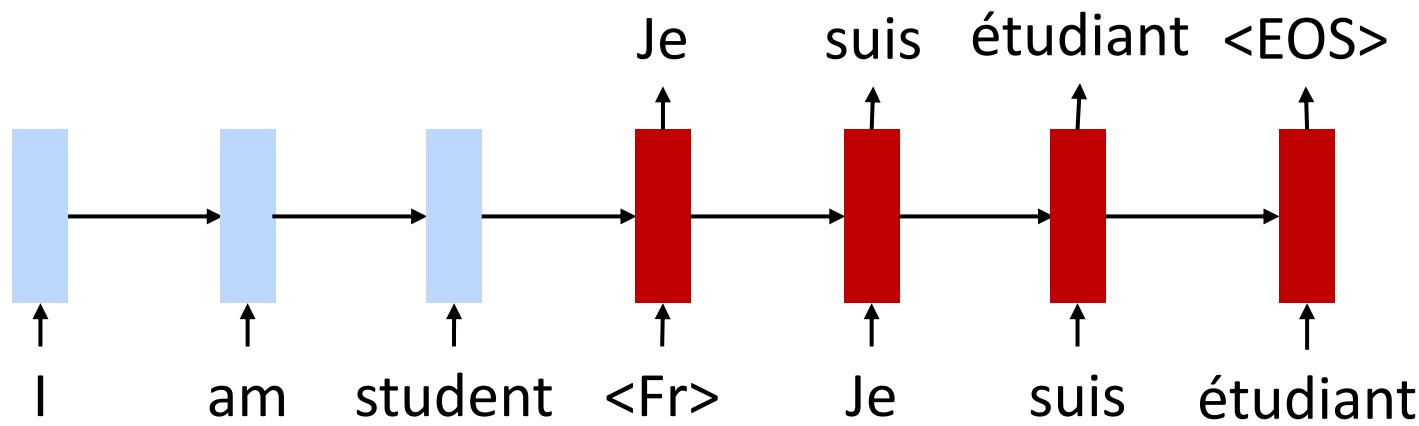
Unsupervised Neural Machine Translation

- Training objective 1: de-noising autoencoder



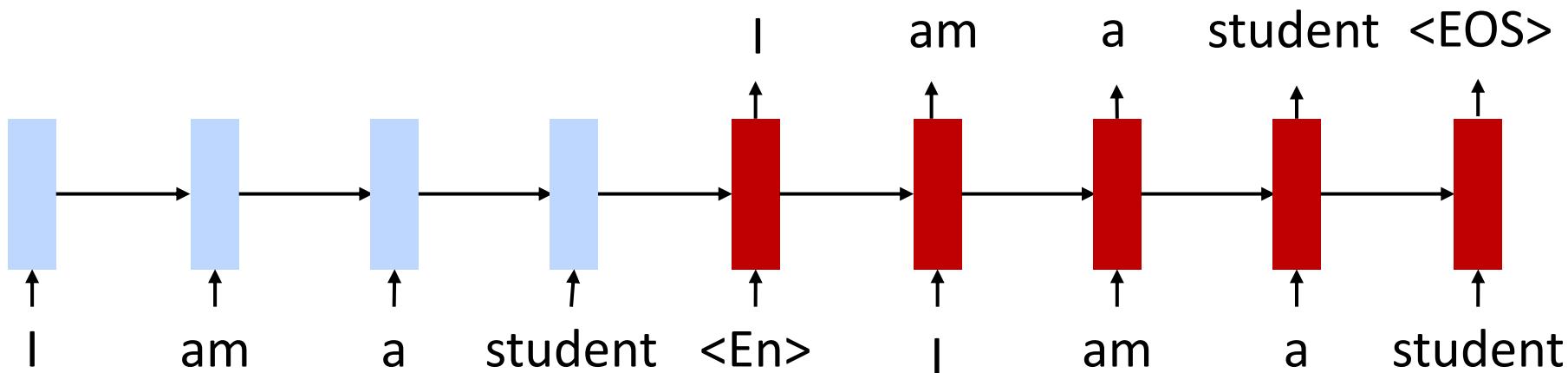
Unsupervised Neural Machine Translation

- Training objective 2: back translation
 - First translate *fr* -> *en*
 - Then use as a “supervised” example to train *en* -> *fr*



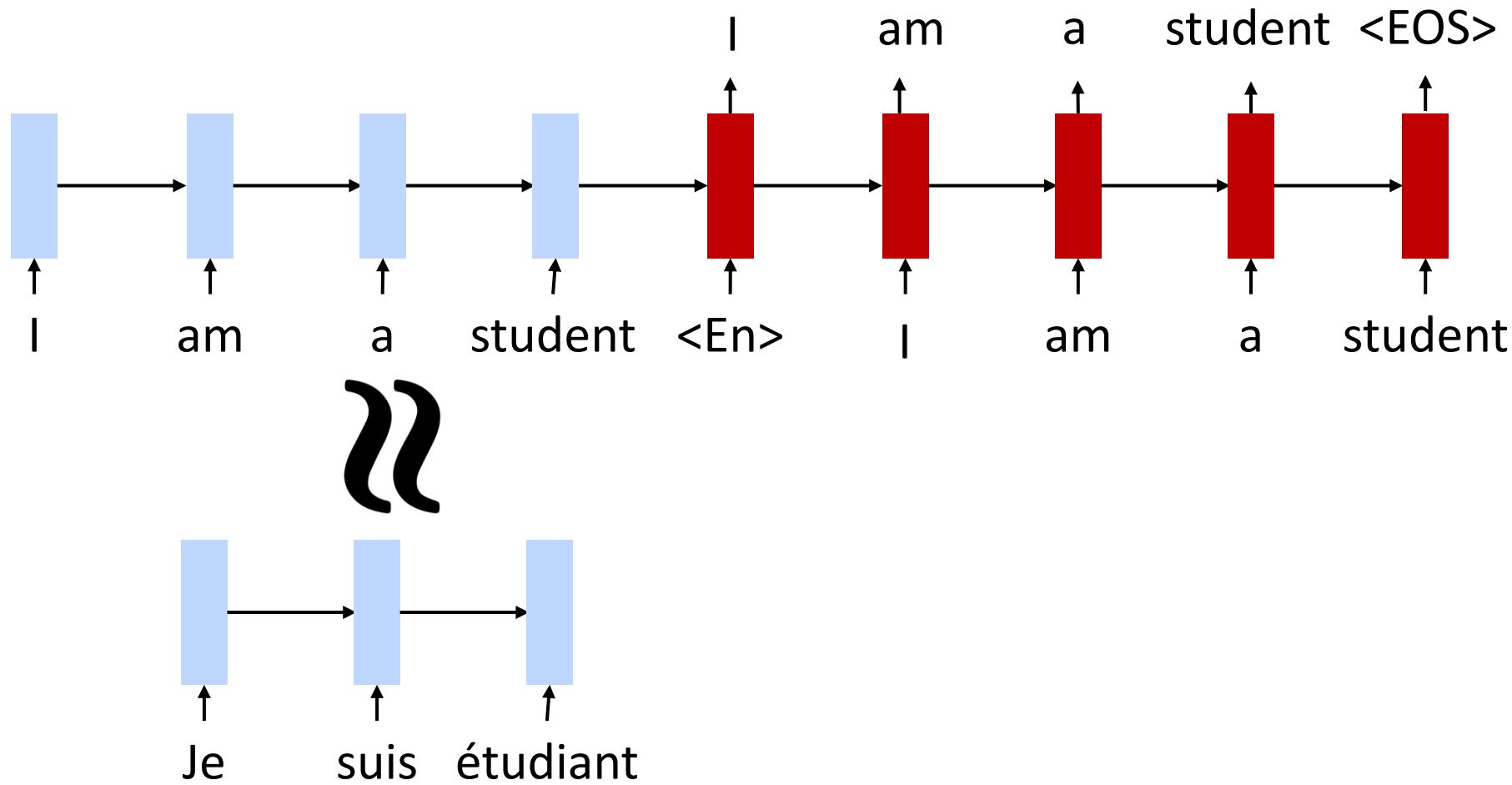
Why Does This Work?

- Cross lingual embeddings and shared encoder gives the model a starting point



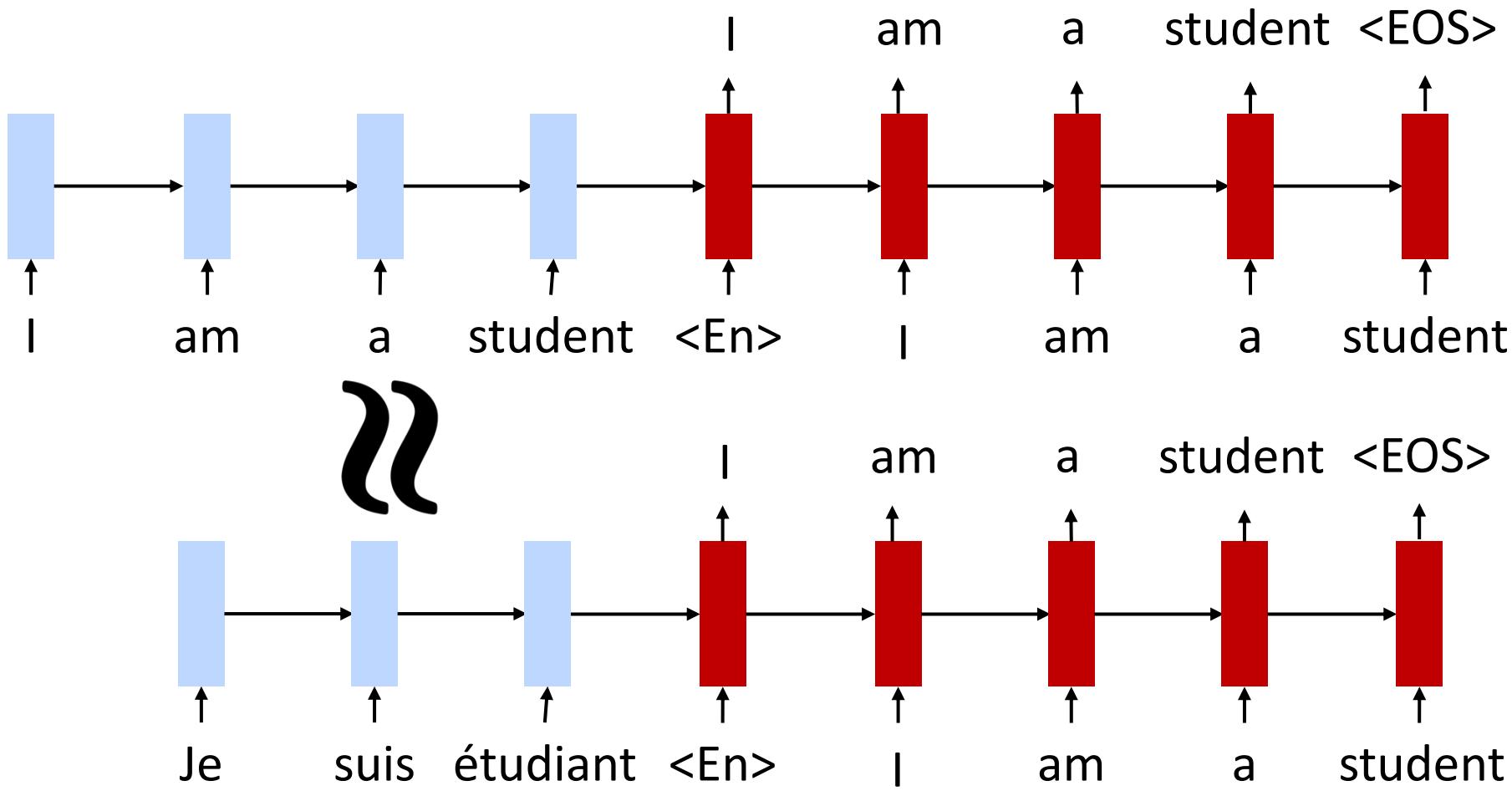
Why Does This Work?

- Cross lingual embeddings and shared encoder gives the model a starting point



Why Does This Work?

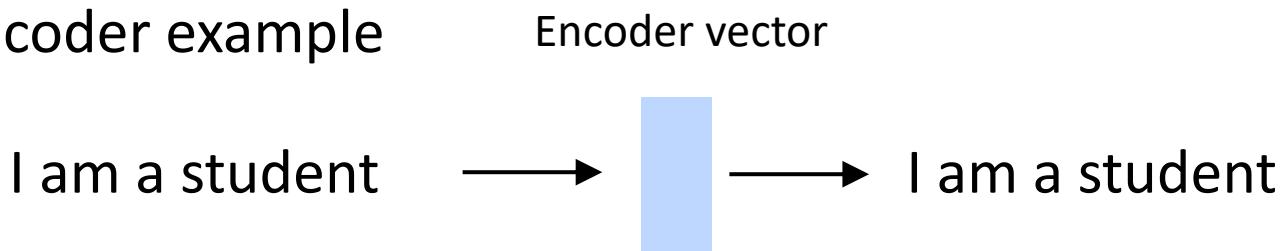
- Cross lingual embeddings and shared encoder gives the model a starting point



Why Does This Work?

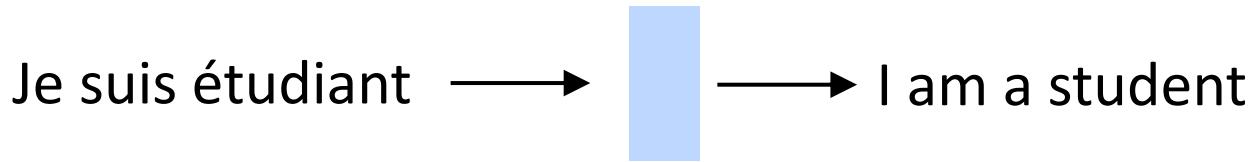
- Objectives encourage language-agnostic representation

Auto-encoder example



Encoder vector

Back-translation example



Why Does This Work?

- Objectives encourage language-agnostic representation

Auto-encoder example

I am a student

Encoder vector



I am a student

Back-translation example

Je suis étudiant

Encoder vector

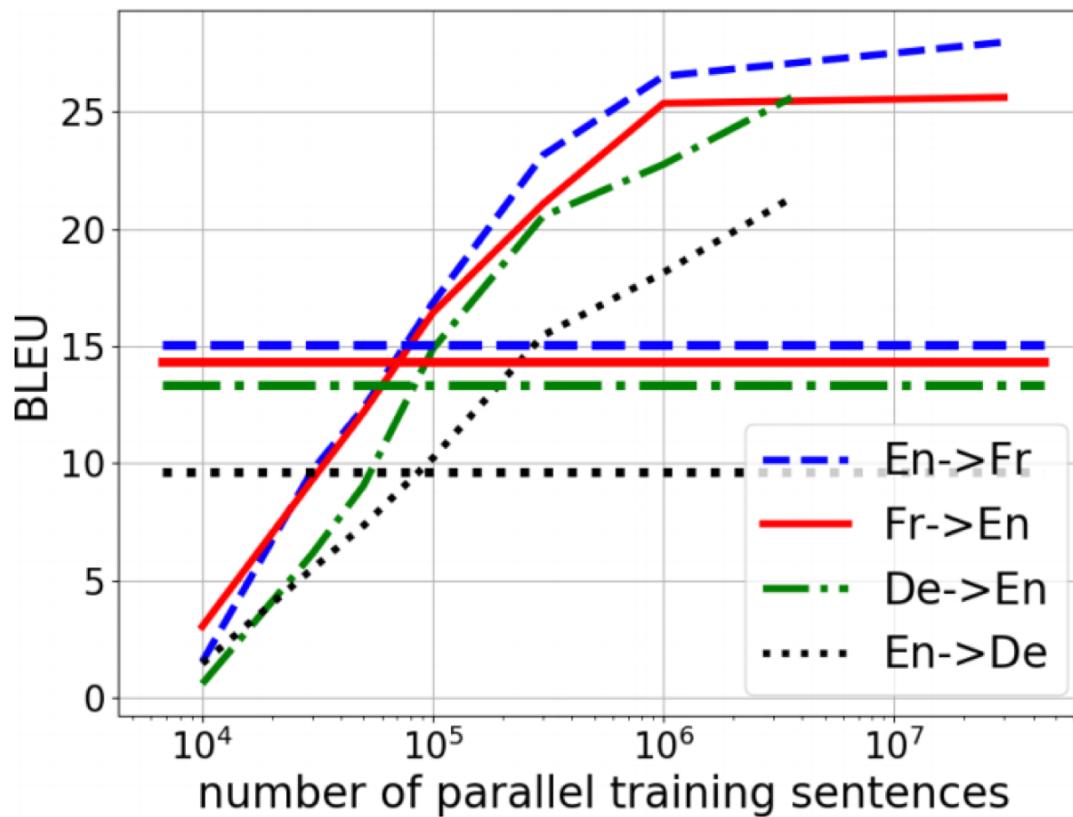


I am a student

**need to be
the same!**

Unsupervised Machine Translation

- Horizontal lines are unsupervised models, the rest are supervised



Attribute Transfer

- Collector corpora of “relaxed” and “annoyed” tweets using hashtags
- Learn un unsupervised MT model

Relaxed ↔ Annoyed

Relaxed	Sitting by the Christmas tree and watching Star Wars after cooking dinner. What a nice night	❤️🎄🌟
Annoyed	Sitting by the computer and watching The Voice for the second time tonight. What a horrible way to start the weekend	😡🤬🤬
Annoyed	Getting a speeding ticket 50 feet in front of work is not how I wanted to start this month	😔
Relaxed	Getting a haircut followed by a cold foot massage in the morning is how I wanted to start this month	😊

Male ↔ Female

Male	Gotta say that beard makes you look like a Viking...
Female	Gotta say that hair makes you look like a Mermaid...
Female	Awww he's so gorgeous 😍 can't wait for a cuddle. Well done 🥺 xxx
Male	Bro he's so f***ing dope can't wait for a cuddle. Well done bro

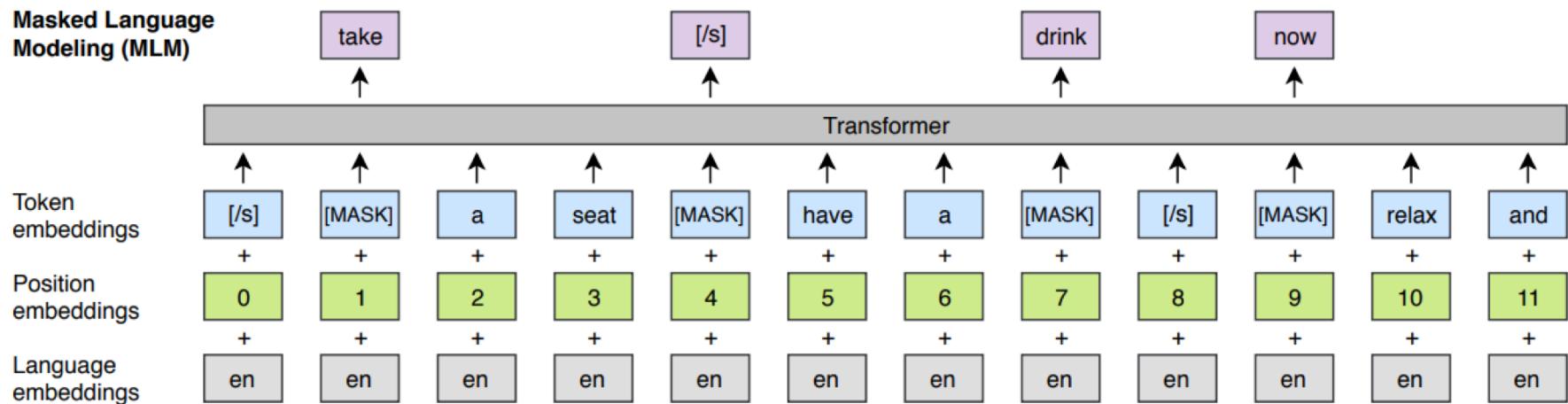
Not so Fast

- English, French, and German are fairly similar
- On very different languages (e.g., English and Turkish)...
 - Purely unsupervised word translation doesn't work very well. Need *seed dictionary* of likely translations.
 - Simple trick: use identical strings from both vocabularies
- UNMT barely works

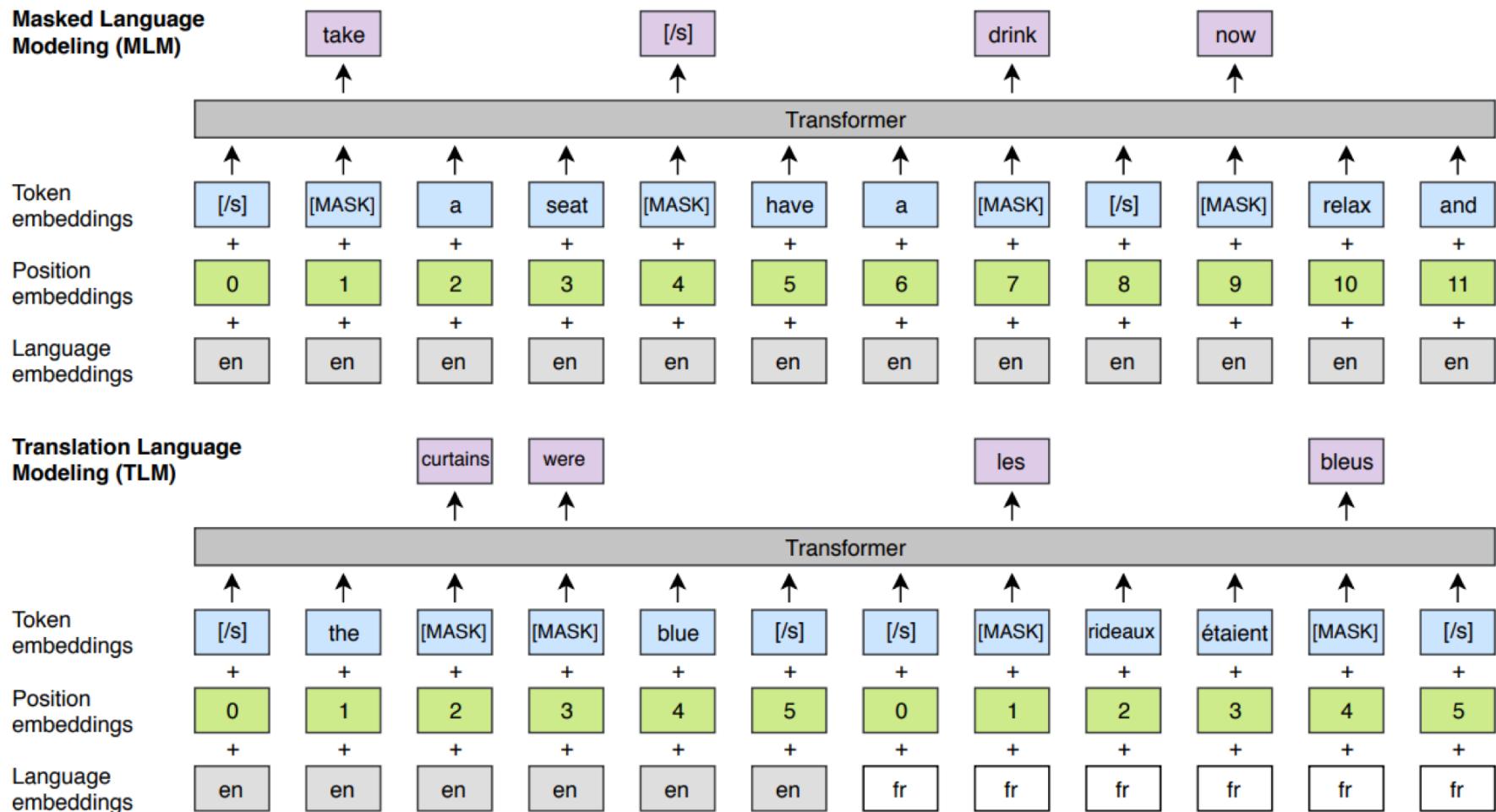
System	English-Turkish BLEU
Supervised	~20
Word-for-word unsupervised	1.5
UNMT	4.5

Cross-Lingual BERT

Cross-Lingual BERT



Cross-Lingual BERT



Cross-Lingual BERT

Unsupervised MT Results

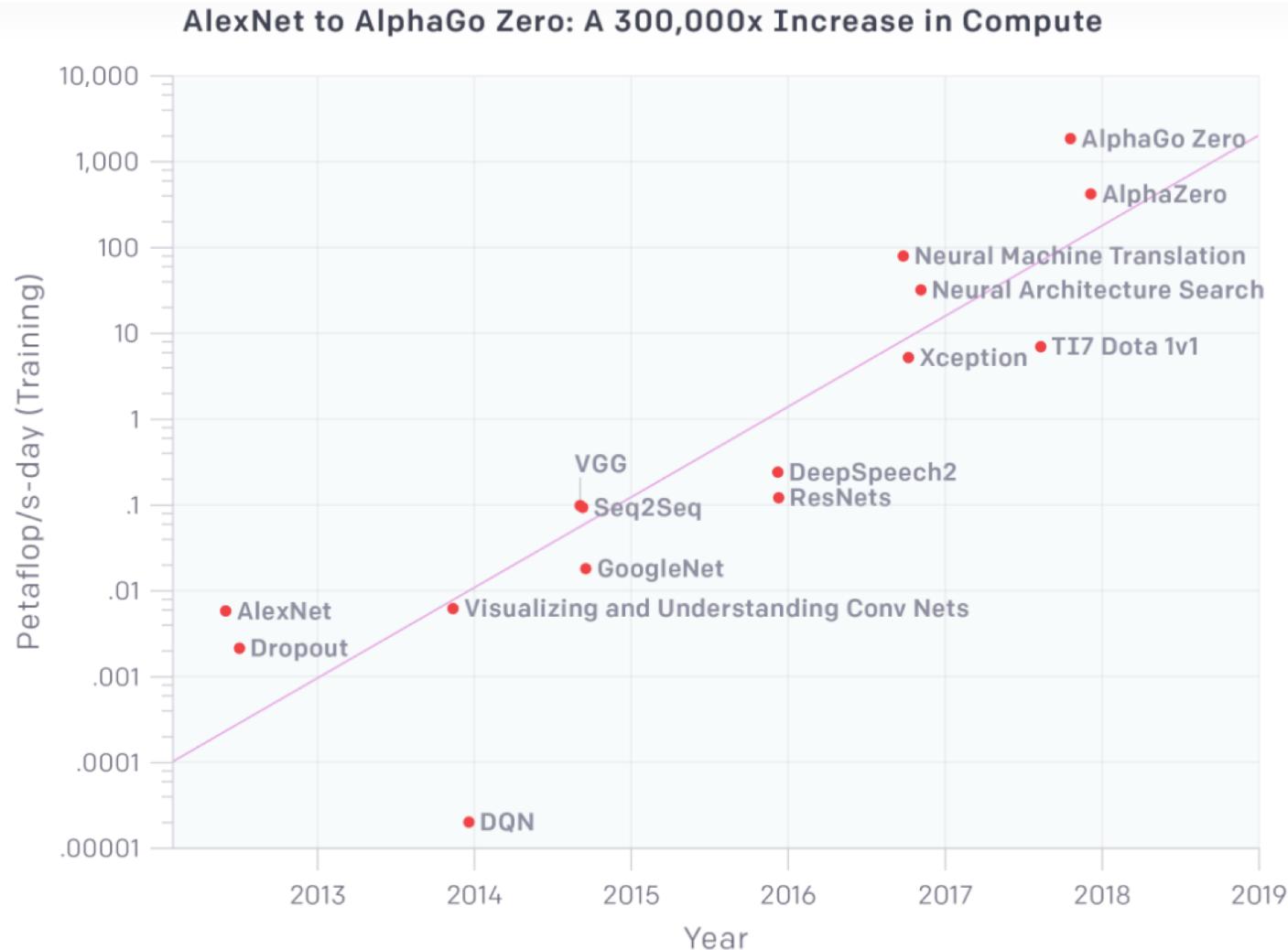
Model	En-Fr	En-De	En-Ro
UNMT	25.1	17.2	21.2
UNMT + Pre-Training	33.4	26.4	33.3
Current supervised State-of-the-art	45.6	34.2	29.9

Huge Models and GPT-2

Training Huge Models

Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B

This is a General Trend in ML



GPT-2

- Just a really big Transformer LM
- Trained on 40GB of text
 - Quite a bit of effort going into making sure the dataset is good quality
 - Take webpages from reddit links with high karma

So What Can GPT-2 Do?

- Obviously, language modeling (but very well)!
- Gets state-of-the-art perplexities on datasets it's not even trained on!

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	21.8
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50	75.20
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37	55.72
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05	44.575
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48	42.16

So What Can GPT-2 Do?

- **Zero-Shot Learning:** no supervised training data!
 - Ask LM to generate from a prompt
- **Reading Comprehension:** <context> <question> A:
- **Summarization:** <article> TL;DR:
- **Translation:**

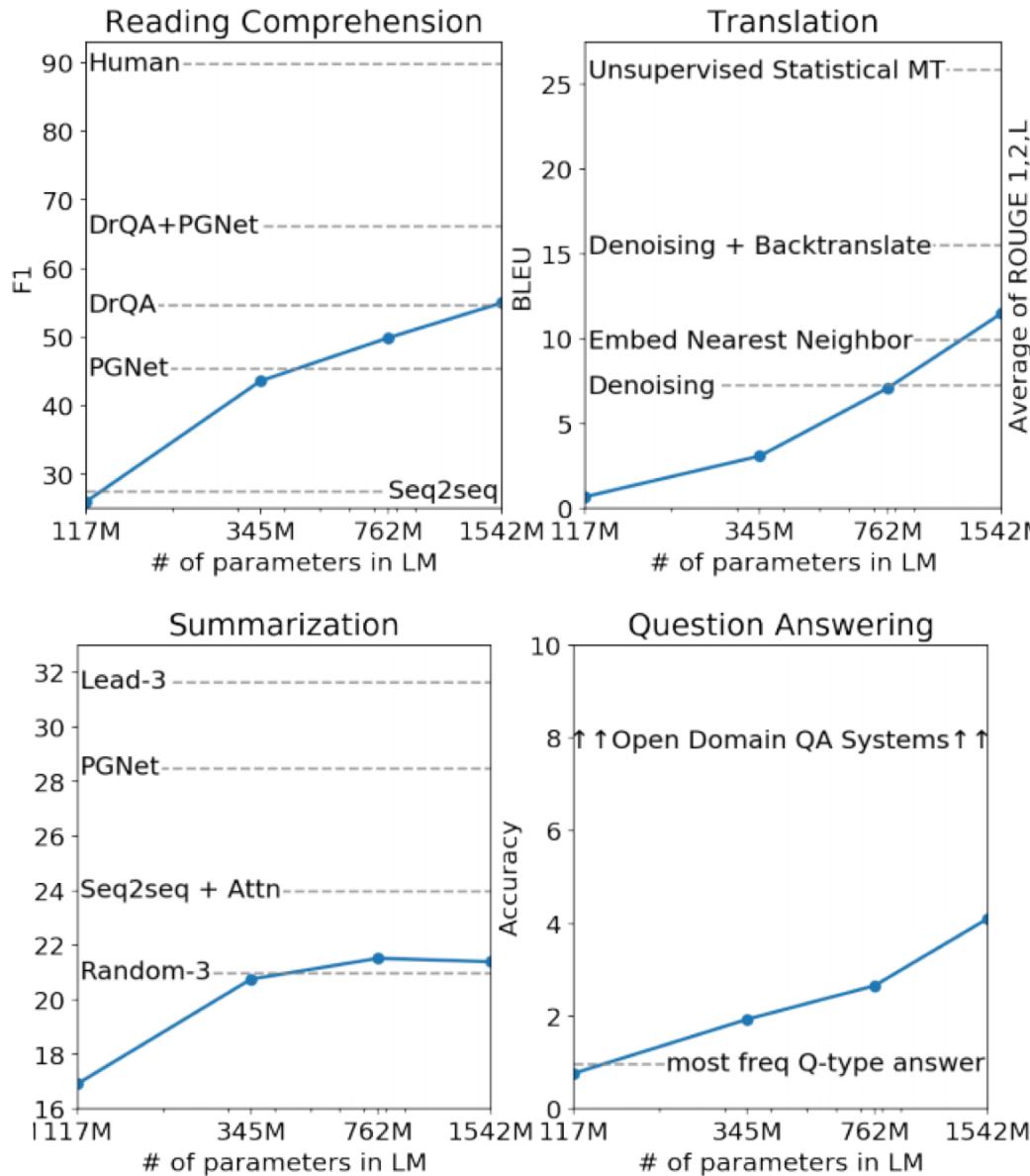
<English sentence1> = <French sentence1>

<English sentence 2> = <French sentence 2>

.....

<Source sentence> =
- **Question Answering:** <question> A:

GPT-2 Results



How can GPT-2 be doing translation?

- It's just given a big corpus of text that's almost all English

How can GPT-2 be doing translation?

- It's just given a big corpus of text that's almost all English

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool]**.

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "**Lie lie and something will always remain.**"

"I hate the word '**perfume**','" Burr says. 'It's somewhat better in French: '**parfum**'.

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre côté? -Quel autre côté?**", which means "**- How do you get to the other side? - What side?**".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

"Brevet Sans Garantie Du Gouvernement", translated to English: "**Patented without government warranty**".

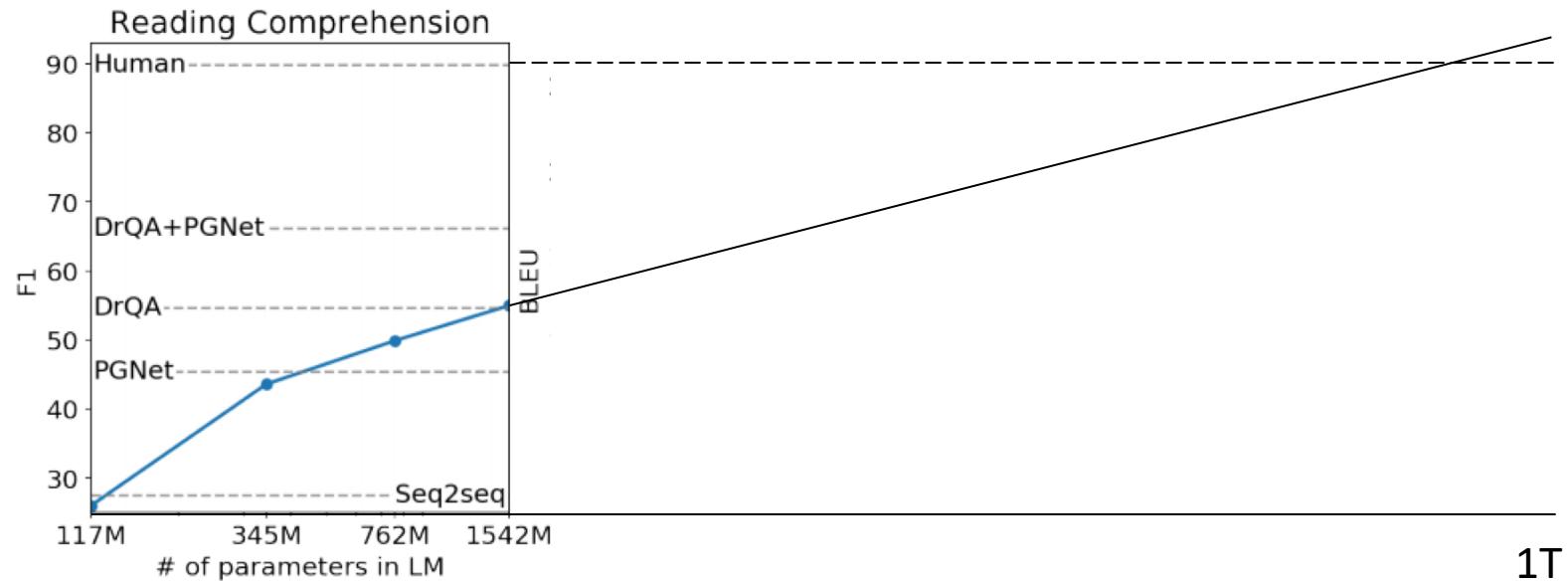
GPT-2 Question Answering

- Simple baseline: 1% accuracy
- GPT-2: ~4% accuracy
- Cherry-picked most confident results

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%

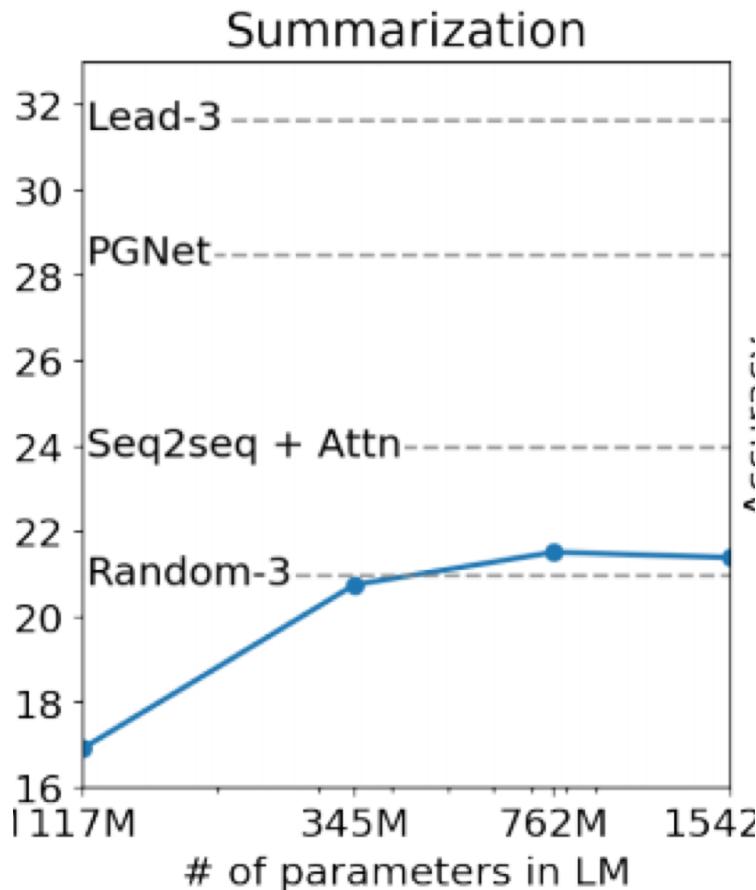
What happens as models get even bigger?

- For several tasks performance seems to increase with $\log(\text{model size})$



What happens as models get even bigger?

- But trend isn't clear



GPT-2 Reaction

GPT-2 Reaction

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter:

GPT-2 Reaction

Just wanted to give you all a heads up, our lab found an amazing breakthrough in language understanding. but we also worry it may fall into the wrong hands. so we decided to scrap it and only publish the regular *ACL stuff instead. Big respect for the team for their great work.

10:08 AM - 15 Feb 2019

118 Retweets 782 Likes



29

118

782



Posted by u/astonished_crofty 25 days ago 2

625
↓

[Discussion] Should I release my MNIST model or keep it closed source fearing malicious use?

Discussion

Today I trained a 23064 layer ResNet and it got 99.6% accuracy on MNIST. I would love to share the model but I fear it being used maliciously. What if it is used to read documents by the Russians? What are your thoughts?

GPT-2 Reaction

Some arguments for release:

Some arguments against:

GPT-2 Reaction

Some arguments for release:

- This model isn't much different from existing work
- Not long until these models are easy to train
 - And we're already at this point for images/speech
- Photoshop
- Researchers should study this model to learn defenses
- Dangerous PR Hype
- Reproducibility is crucial for science
- ...

Some arguments against:

- Danger of fake reviews, news comments, etc.
 - Already done by companies and governments
- Precedent
 - Event if this model isn't dangerous, later ones will be even better
- Smaller model is being released
-

GPT-2 Reaction

- Should NLP experts be the ones making these decisions?
 - Experts on computer security?
 - Experts on technology and society?
 - Experts on ethics?
- Need for more interdisciplinary science
- Many other examples of NLP with big social ramifications, especially with regards to bias/fairness

High-Impact Decisions

- Growing interest in using NLP to help with high-impact decision making
 - Judicial decisions
 - Hiring
 - Grading tests
- Plus side: can quickly evaluate a machine learning system for some kinds of bias
- However, machine learning reflects or even amplifies bias in training data
 - ...which could lead to the creation of even more biased data

High-Impact Decisions

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ



Intelligent Machines

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

High-Impact Decisions



Ben Zimmer  @bgzimmer · 2 Jul 2018

This gobbledegook earns a perfect grade from the GRE's automated essay scoring system. Algorithms writing for algorithms. npr.org/2018/06/30/624...

"History by mimic has not, and presumably never will be precipitously but blithely ensconced. Society will always encompass imaginativeness; many of scrutinizations but a few for an amanuensis. The perjured imaginativeness lies in the area of theory of knowledge but also the field of literature. Instead of entralling the analysis, grounds constitutes both a disparaging quip and a diligent explanation."

51

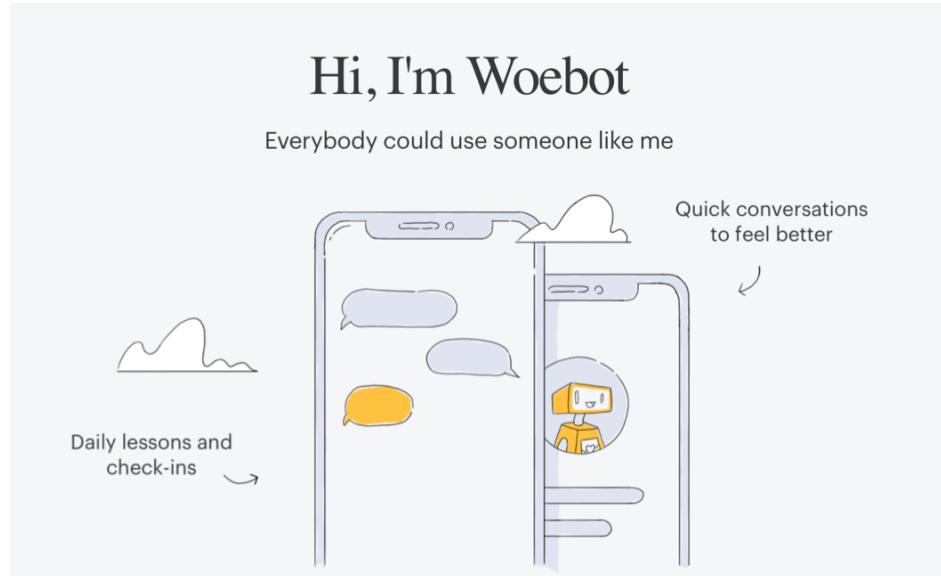
636

1.1K



Chatbots

- Potential for positive impact

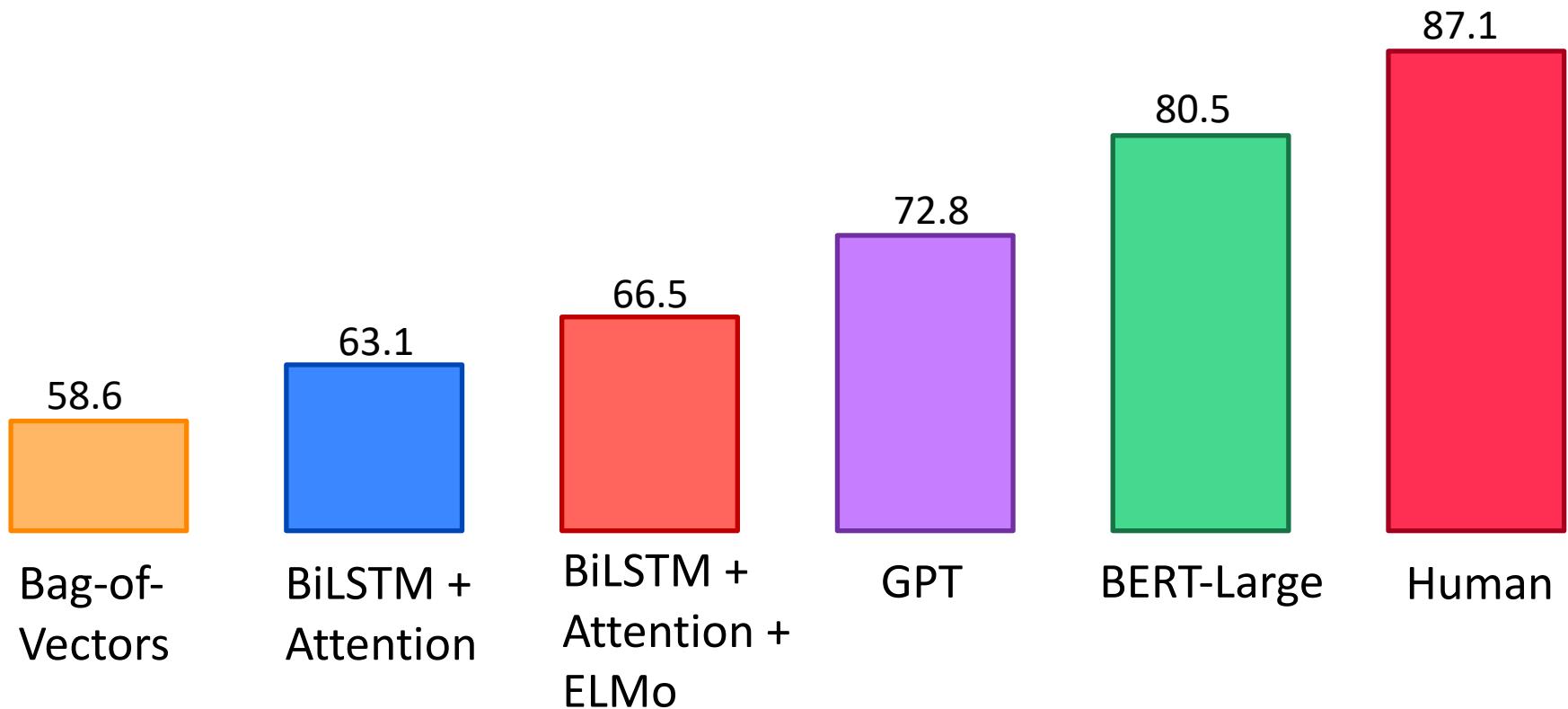


- But big risks

**AI ROBOTS LEARNING RACISM,
SEXISM AND OTHER PREJUDICES
FROM HUMANS, STUDY FINDS**

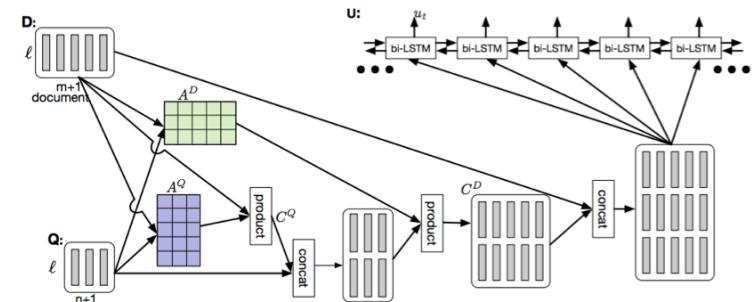
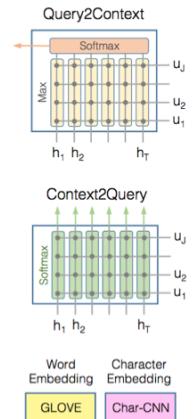
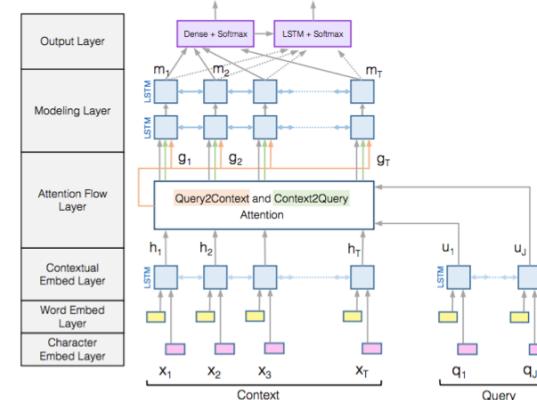
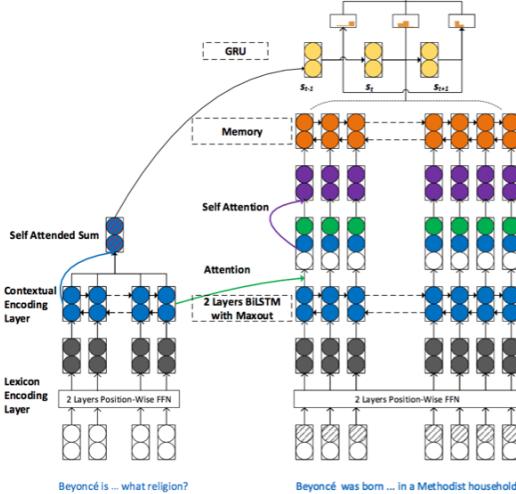
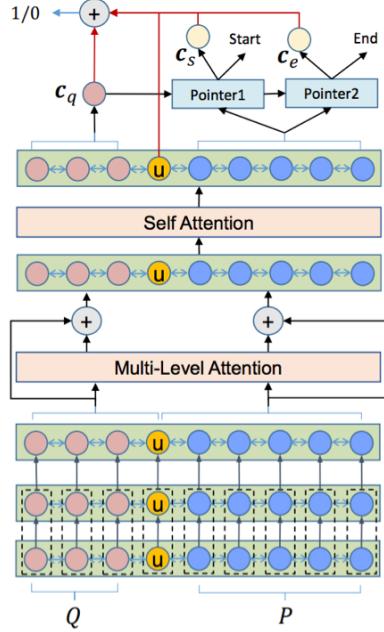
**What did BERT “solve” and what
do we work on next?**

GLUE Benchmark Results



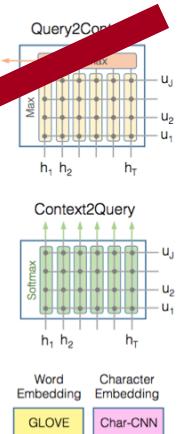
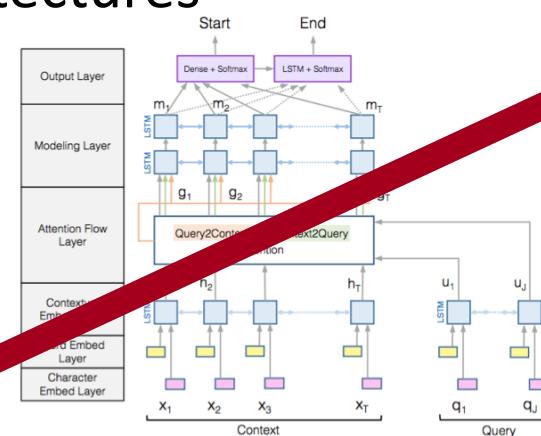
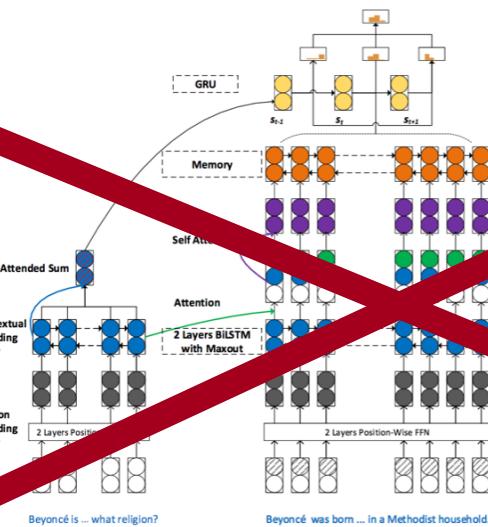
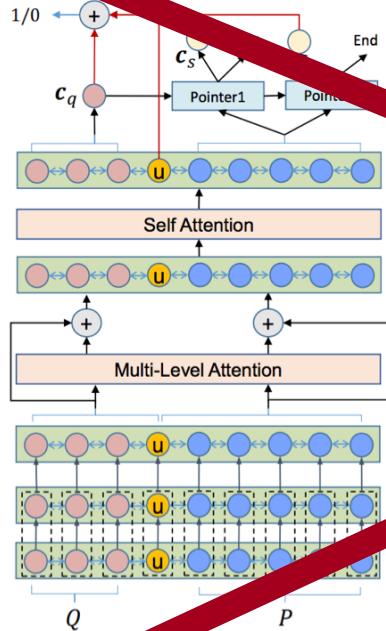
The Death of Architecture Engineering?

Some SQuAD NN Architectures



The Death of Architecture Engineering?

Some SQuAD NN Architectures



Attention Is All You Need

The Death of Architecture Engineering?

- 6 months of research on architecture design, get 1 F1 point improvement
- ... Or just make BERT 3x bigger, get 5 F1 points
- Top 20 entrants on the SQuAD leaderboard all use BERT

	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) Google AI Language https://github.com/google-research/bert	86.673	89.147
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) Google AI Language https://github.com/google-research/bert	85.150	87.715
3 Jan 15, 2019	BERT + MMFT + ADA (ensemble) Microsoft Research Asia	85.082	87.615
4 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) Google AI Language https://github.com/google-research/bert	84.292	86.967
5 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) Layer 6 AI NLP Team	83.469	86.043
5 Dec 21, 2018	PAML+BERT (ensemble model) PINGAN GammaLab	83.457	86.122
5 Dec 13, 2018	BERT finetune baseline (ensemble) Anonymous	83.536	86.096
6 Mar 04, 2019	SemBERT (ensemble model) Shanghai Jiao Tong University	83.243	85.821
6 Jan 14, 2019	BERT + MMFT + ADA (single model) Microsoft Research Asia	83.040	85.892
7 Jan 10, 2019	BERT + Synthetic Self-Training (single model) Google AI Language https://github.com/google-research/bert	82.972	85.810

Harder Natural Language Understanding

- Reading comprehension...
 - On longer documents or multiple documents
 - That requires multi-hop reasoning
 - Situated in a dialogue
- Key problem with many existing reading comprehension datasets: *People writing the questions see the context*
 - Not realistic
 - Encourages easy questions

Multi-Task Learning

- Another frontier of NLP is getting one model to perform many tasks. GLUE and DecaNLP are recent examples.
- Multi-task learning yields improvements on top of BERT

Rank	Name	Model	URL	Score
1	GLUE Human Baselines	GLUE Human Baselines		87.1
2	王玮	ALICE large (Alibaba DAMO NLP)		83.0
3	Microsoft D365 AI & MSR AI	MT-DNNv2 (BigBird)		83.0
4	Jason Phang	BERT on STILTs		82.0
5	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hid		80.5

BERT + Multi-task

Low-Resource Settings

- Models that don't require lots of compute power (can't use BERT)!
 - Especially important for mobile devices
- Low-resource languages
- Low-data settings (few shot learning)
 - Meta-learning is becoming popular in ML.

Interpreting/Understanding Models

- Can we get explanations for model predictions?
- Can we understand what models like BERT know and why they work so well?
- Rapidly growing area in NLP
- Very important for some applications (e.g., healthcare)

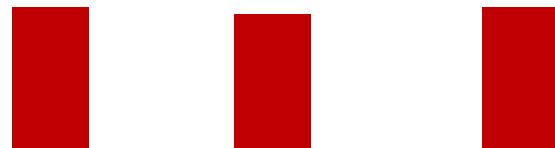
Diagnostic/Probing Classifiers

- Popular technique to see what linguistic information models “know”
- Diagnostic classifier takes representations produced by a model (e.g., BERT) as input and do some task

DET NNP VBD



Diagnostic
Classifier



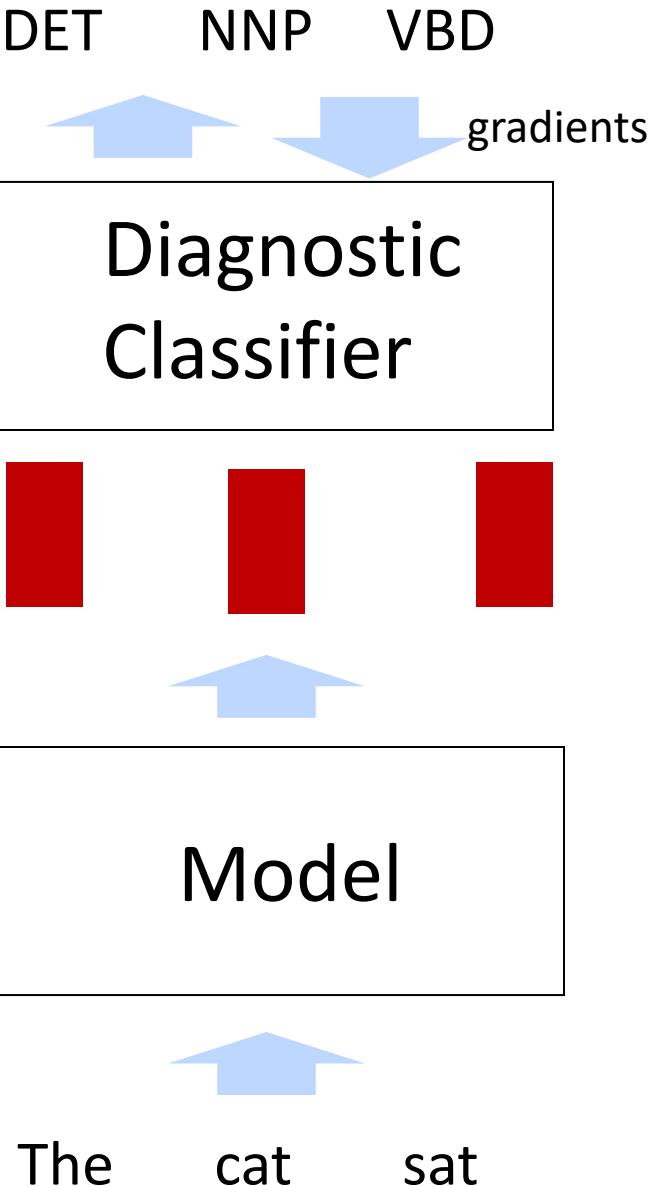
Model



The cat sat

Diagnostic/Probing Classifiers

- Popular technique to see what linguistic information models “know”
- Diagnostic classifier takes representations produced by a model (e.g., BERT) as input and do some task
- Only the diagnostic classifier is trained



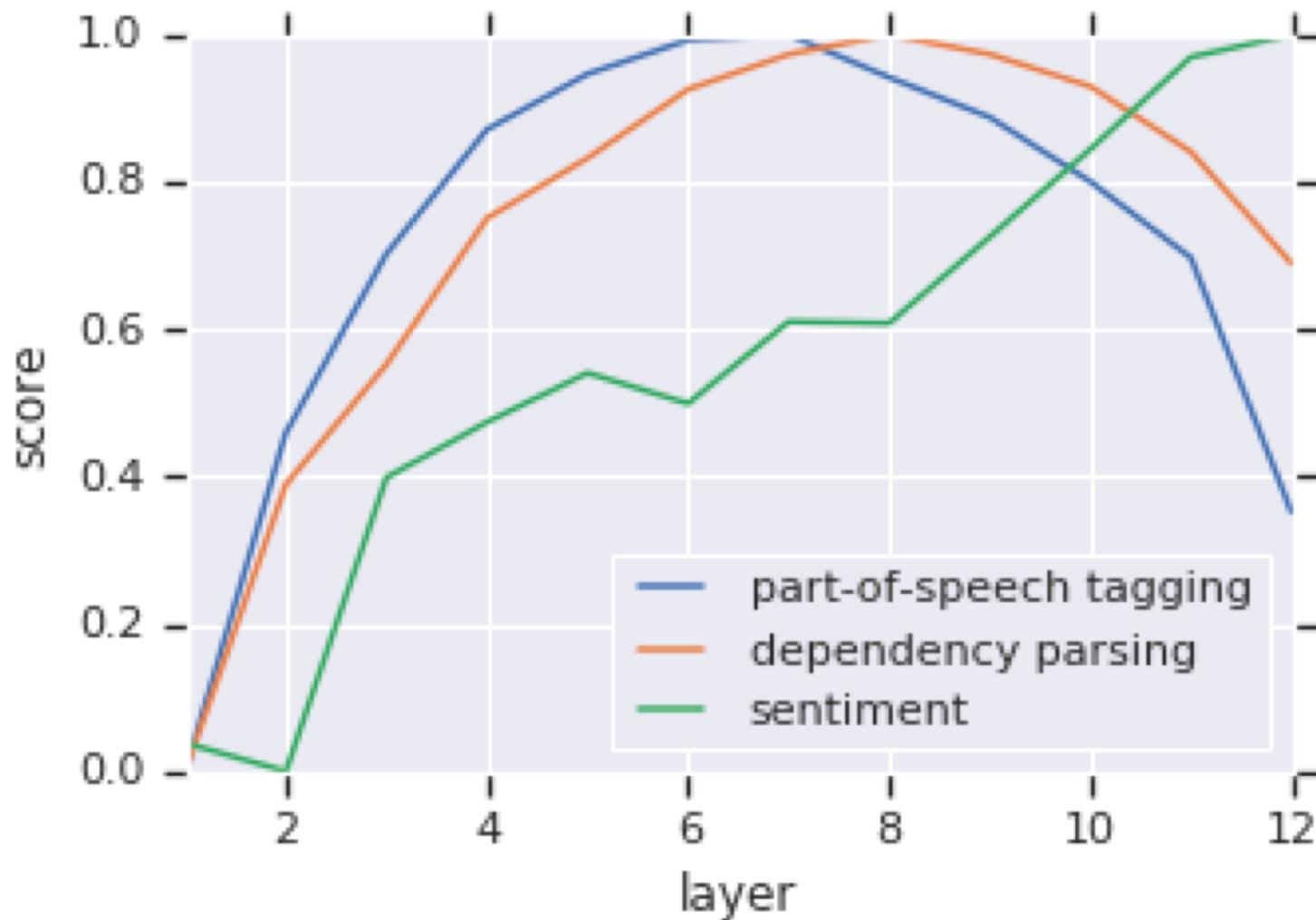
Diagnostic/Probing Classifiers

- Diagnostic classifiers are usually very simple (e.g., a single softmax). Otherwise they could learn to do the tasks without looking at the model representations
- Some diagnostic tasks

POS	The important thing about Disney is that it is a global [brand] ₁ . → NN (Noun)
Constit.	The important thing about Disney is that it [is a global brand] ₁ . → VP (Verb Phrase)
Depend.	[Atmosphere] ₁ is always [fun] ₂ → nsubj (nominal subject)
Entities	The important thing about [Disney] ₁ is that it is a global brand. → Organization
SRL	[The important thing about Disney] ₂ [is] ₁ that it is a global brand. → Arg1 (Agent)
SPR	[It] ₁ [endorsed] ₂ the White House strategy... → {awareness, existed_after, ... }
Coref. ^O	The important thing about [Disney] ₁ is that [it] ₂ is a global brand. → True
Coref. ^W	[Characters] ₂ entertain audiences because [they] ₁ want people to be happy. → True Characters entertain [audiences] ₂ because [they] ₁ want people to be happy. → False
Rel.	The [burst] ₁ has been caused by water hammer [pressure] ₂ . → Cause-Effect(e_2, e_1)

Diagnostic/ Probing Classifiers: Results

- Lower layers of BERT are better at lower-level tasks



NLP in Industry

- NLP is rapidly growing in industry as well. Two particularly big areas:
- Dialogue
 - Chatbots
 - Customer service
- Healthcare
 - Understanding health records
 - Understanding biomedical literature



Conclusion

- Rapid progress in the last 5 years due to deep learning.
- Even more rapid progress in the last year due to larger models, better usage of unlabeled data
 - Exciting time to be working on NLP!
- NLP is reaching the point of having big social impact, making issues like bias and security increasingly important.