Language Resources in the World of Neural Networks

Pisa, 24-1-2020

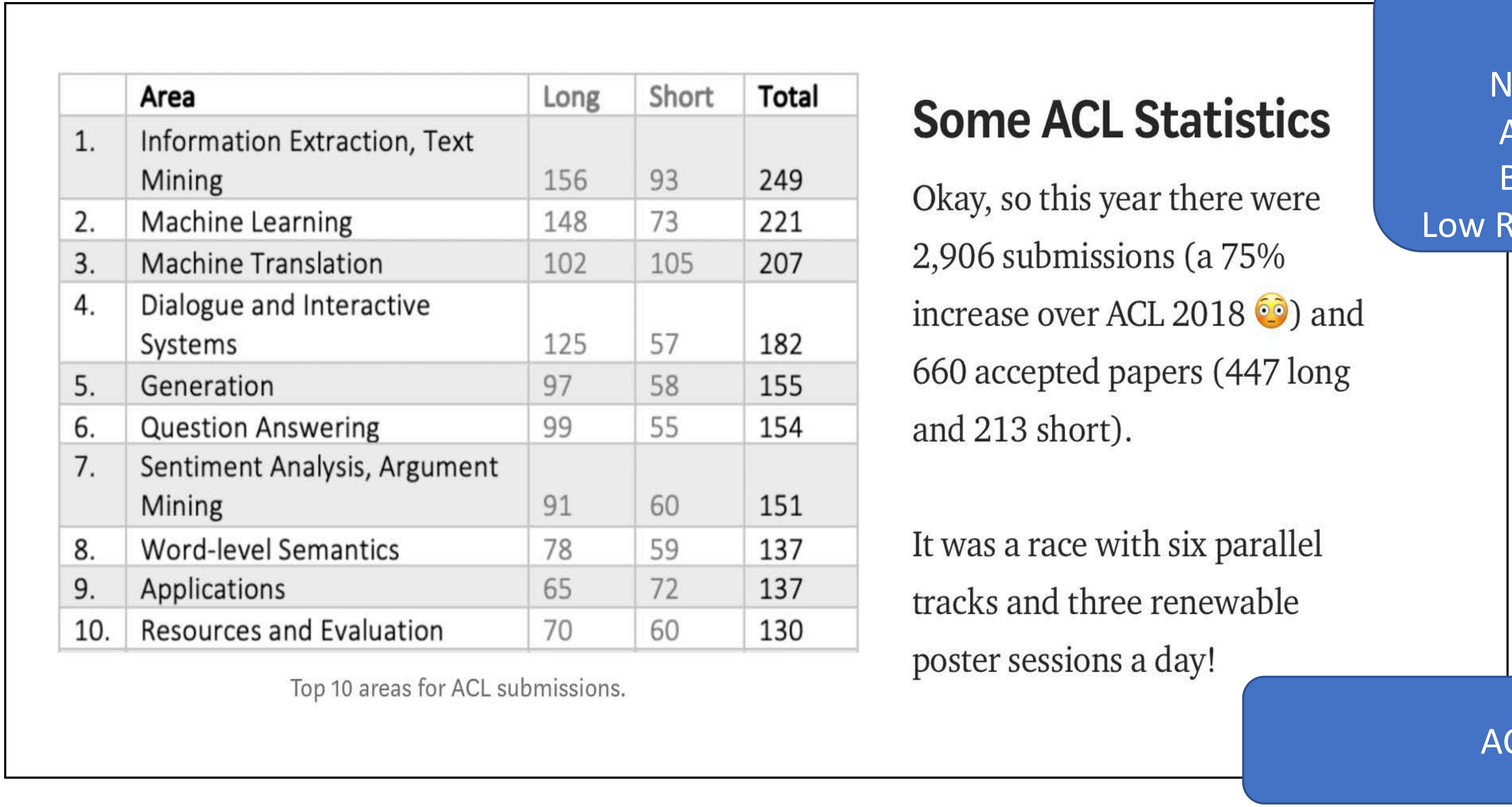
Núria Bel

Catedrática de Tecnologías del Lenguaje

Universitat Pompeu Fabra

https://www.upf.edu/web/nuria-bel

Neural Networks and Deep Learning have impacted NLP in many ways …



	Area	Long	Short	Total
1.	Information Extraction, Text Mining	156	93	249
2.	Machine Learning	148	73	221
3.	Machine Translation	102	105	207
4.	Dialogue and Interactive Systems	125	57	182
5.	Generation	97	58	155
6.	Question Answering	99	55	154
7.	Sentiment Analysis, Argument Mining	91	60	151
8.	Word-level Semantics	78	59	137
9.	Applications	65	72	137
10.	Resources and Evaluation	70	60	130

Top 10 areas for ACL submissions.

Some ACL Statistics

Okay, so this year there were 2,906 submissions (a 75% increase over ACL 2018 😳) and 660 accepted papers (447 long and 213 short).

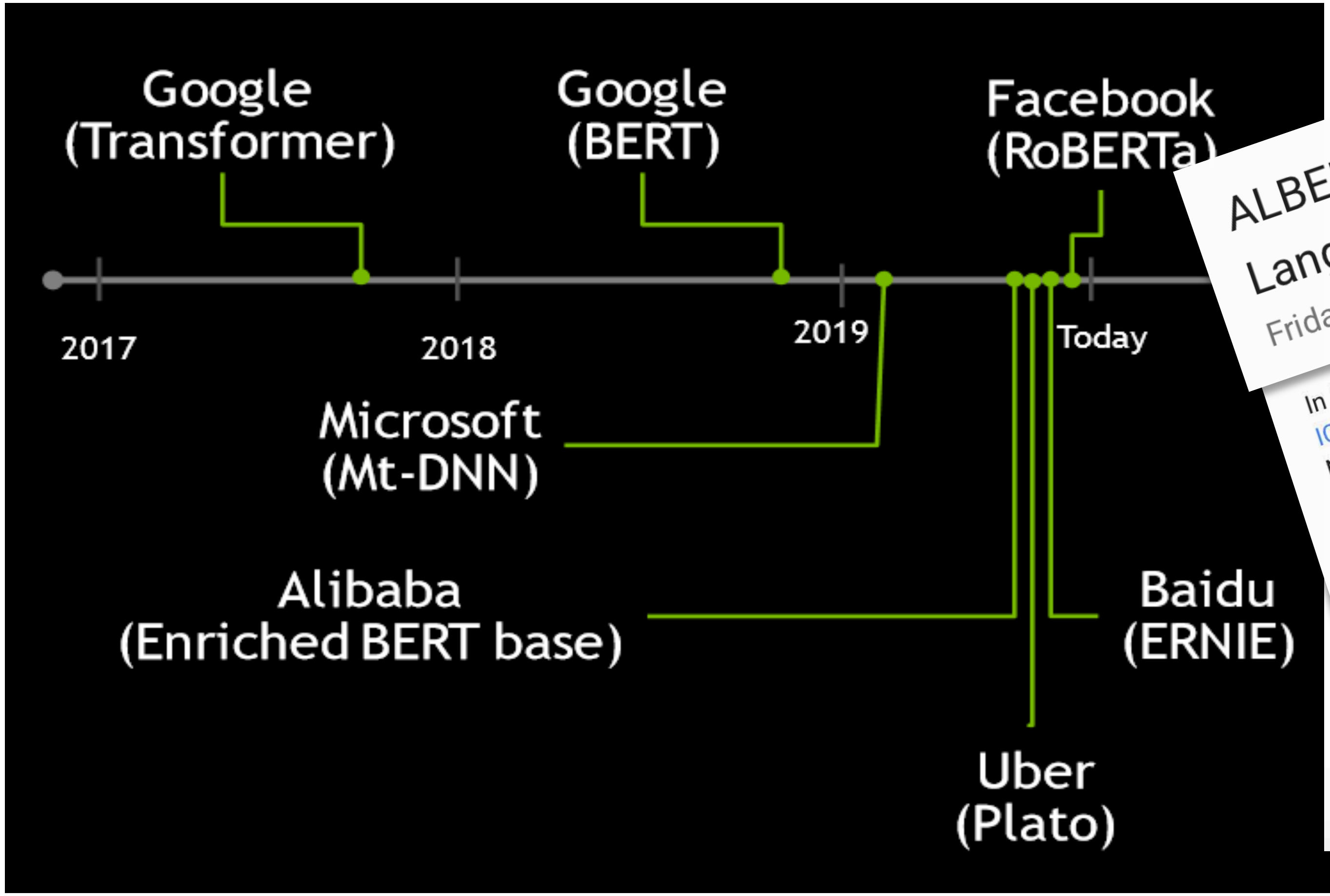
It was a race with six parallel tracks and three renewable poster sessions a day!

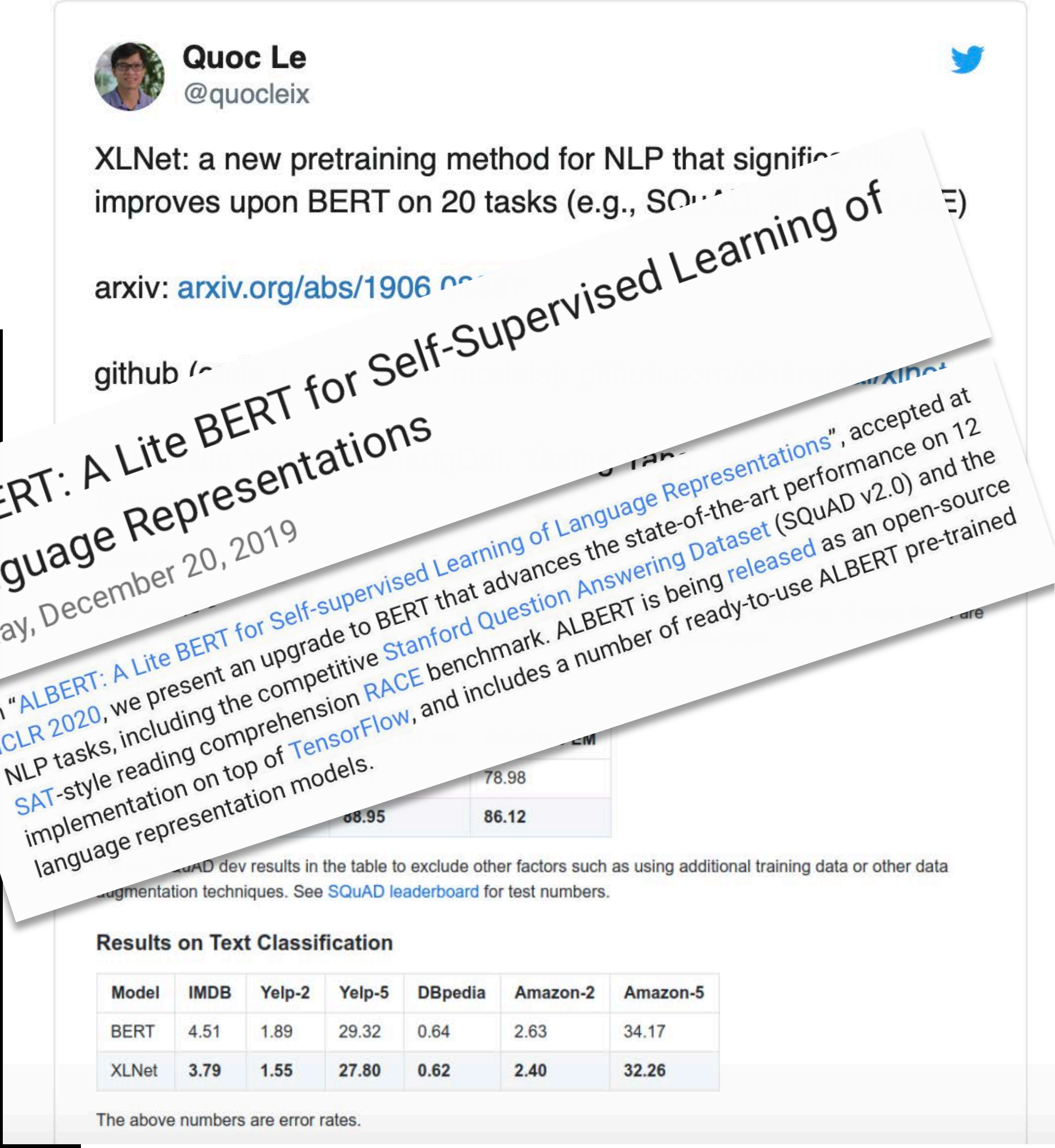
Keywords

Neural	174
Attention	58
BERT	27
Low Resources	10

ACL 2019

Fast development and achievements!!





Quoc Le
@quocleix

XLNet: a new pretraining method for NLP that signific
improves upon BERT on 20 tasks (e.g., SQ E)

arxiv: arxiv.org/abs/1906

github

ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations

Friday, December 20, 2019

In "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", accepted at ICLR 2020, we present an upgrade to BERT that advances the state-of-the-art performance on 12 NLP tasks, including the competitive Stanford Question Answering Dataset (SQuAD v2.0) and the SAT-style reading comprehension RACE benchmark. ALBERT is being released as an open-source implementation on top of TensorFlow, and includes a number of ready-to-use ALBERT pre-trained language representation models.

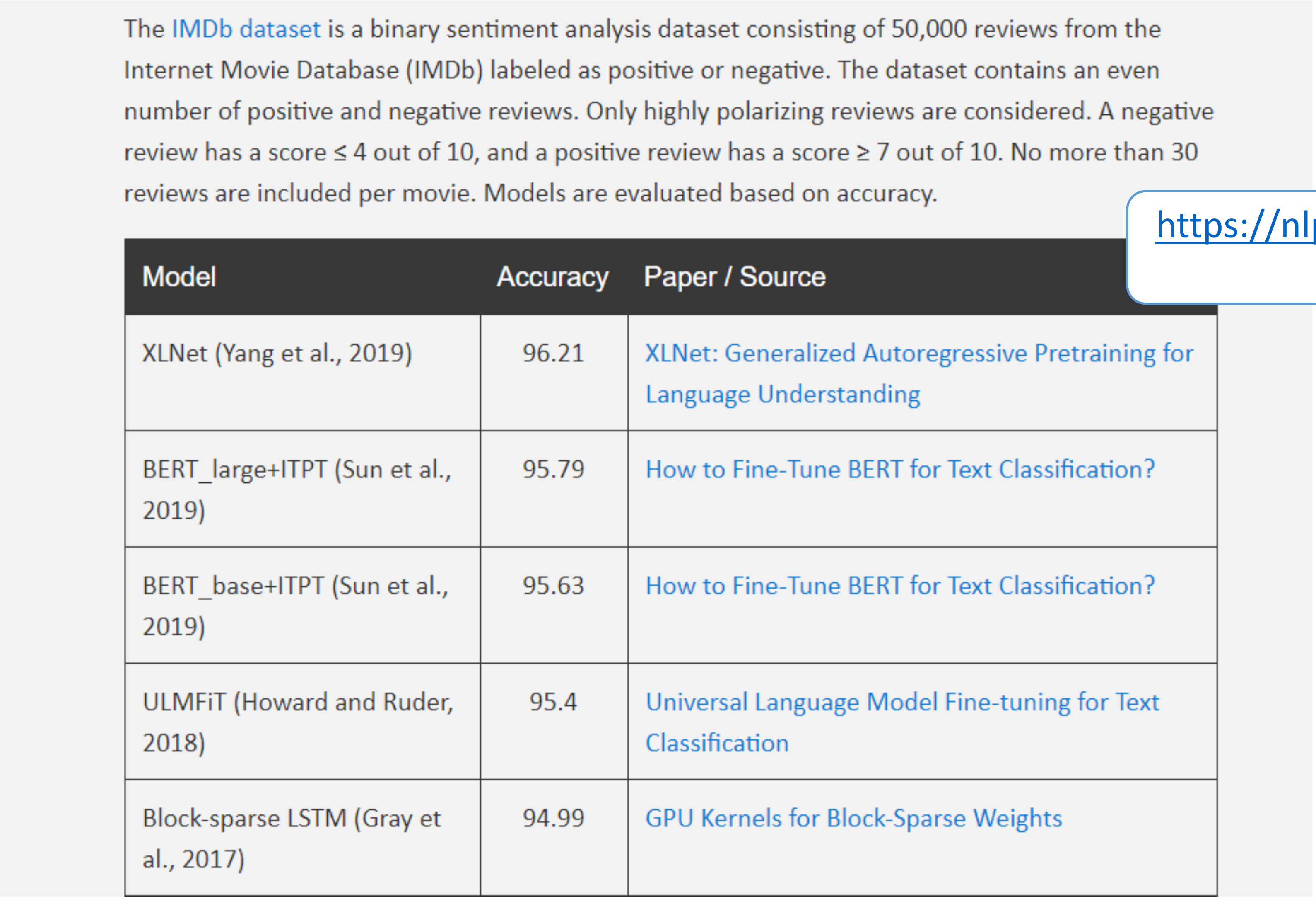
AD dev results in the table to exclude other factors such as using additional training data or other data augmentation techniques. See SQuAD leaderboard for test numbers.

Results on Text Classification

Model	IMDB	Yelp-2	Yelp-5	DBpedia	Amazon-2	Amazon-5
BERT	4.51	1.89	29.32	0.64	2.63	34.17
XLNet	**3.79**	**1.55**	**27.80**	**0.62**	**2.40**	**32.26**

The above numbers are error rates.

Awsome improvements in many tasks!

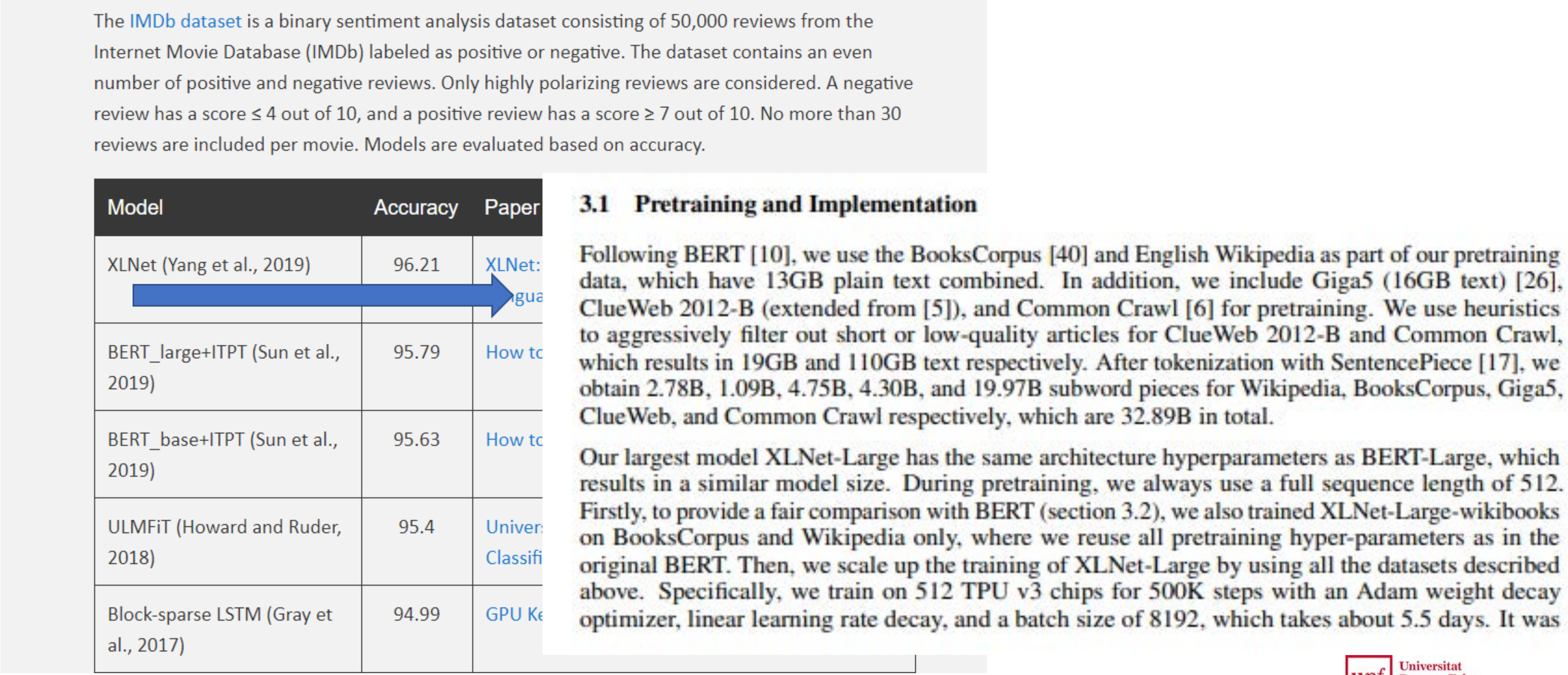


The IMDb dataset is a binary sentiment analysis dataset consisting of 50,000 reviews from the Internet Movie Database (IMDb) labeled as positive or negative. The dataset contains an even number of positive and negative reviews. Only highly polarizing reviews are considered. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. No more than 30 reviews are included per movie. Models are evaluated based on accuracy.

Model	Accuracy	Paper / Source
XLNet (Yang et al., 2019)	96.21	XLNet: Generalized Autoregressive Pretraining for Language Understanding
BERT_large+ITPT (Sun et al., 2019)	95.79	How to Fine-Tune BERT for Text Classification?
BERT_base+ITPT (Sun et al., 2019)	95.63	How to Fine-Tune BERT for Text Classification?
ULMFiT (Howard and Ruder, 2018)	95.4	Universal Language Model Fine-tuning for Text Classification
Block-sparse LSTM (Gray et al., 2017)	94.99	GPU Kernels for Block-Sparse Weights

https://nlpprogress.com/english/sentiment_analysis.html

Tons of data, HPC and during many days!!



The IMDb dataset is a binary sentiment analysis dataset consisting of 50,000 reviews from the Internet Movie Database (IMDb) labeled as positive or negative. The dataset contains an even number of positive and negative reviews. Only highly polarizing reviews are considered. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. No more than 30 reviews are included per movie. Models are evaluated based on accuracy.

Model	Accuracy	Paper
XLNet (Yang et al., 2019)	96.21	XLNet: ngua
BERT_large+ITPT (Sun et al., 2019)	95.79	How to
BERT_base+ITPT (Sun et al., 2019)	95.63	How to
ULMFiT (Howard and Ruder, 2018)	95.4	Univers Classifi
Block-sparse LSTM (Gray et al., 2017)	94.99	GPU Ke

3.1 Pretraining and Implementation

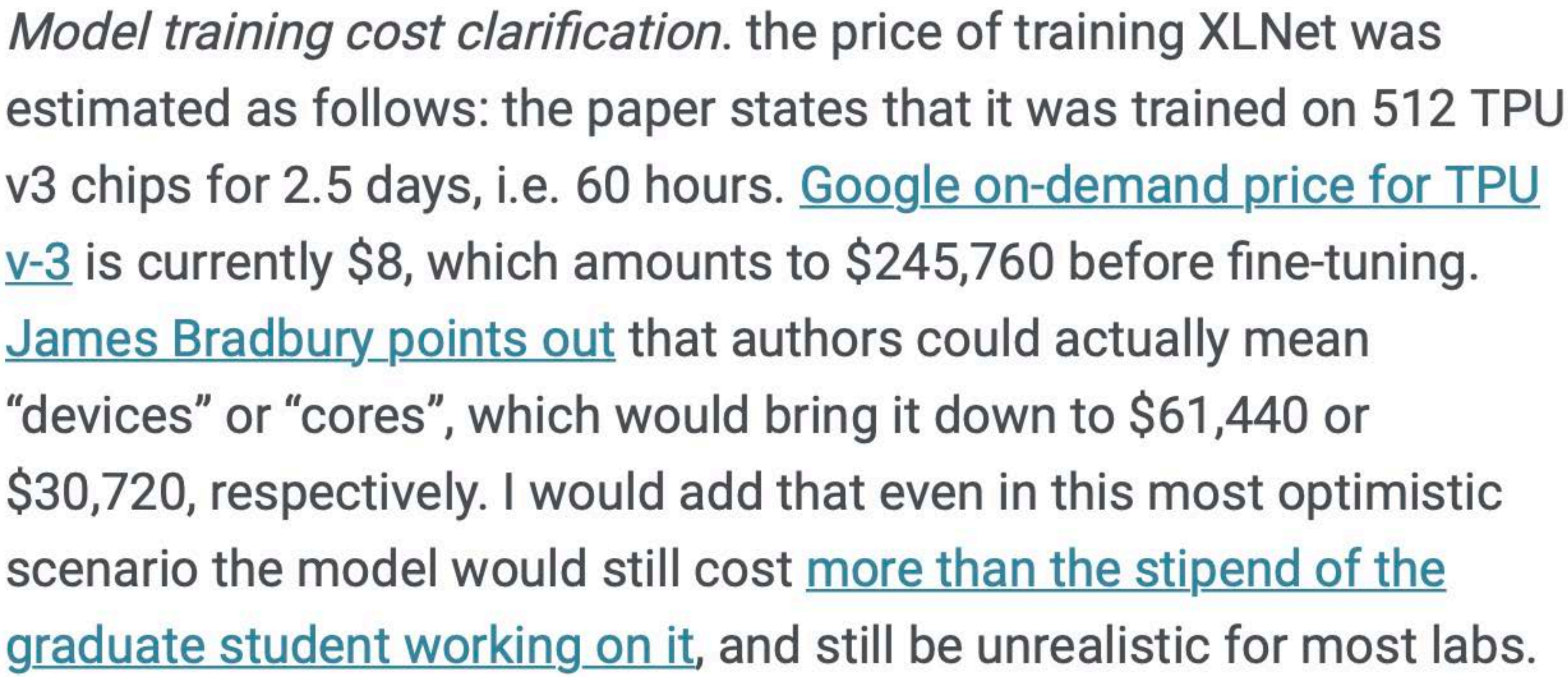
Following BERT [10], we use the BooksCorpus [40] and English Wikipedia as part of our pretraining data, which have 13GB plain text combined. In addition, we include Giga5 (16GB text) [26], ClueWeb 2012-B (extended from [5]), and Common Crawl [6] for pretraining. We use heuristics to aggressively filter out short or low-quality articles for ClueWeb 2012-B and Common Crawl, which results in 19GB and 110GB text respectively. After tokenization with SentencePiece [17], we obtain 2.78B, 1.09B, 4.75B, 4.30B, and 19.97B subword pieces for Wikipedia, BooksCorpus, Giga5, ClueWeb, and Common Crawl respectively, which are 32.89B in total.

Our largest model XLNet-Large has the same architecture hyperparameters as BERT-Large, which results in a similar model size. During pretraining, we always use a full sequence length of 512. Firstly, to provide a fair comparison with BERT (section 3.2), we also trained XLNet-Large-wikibooks on BooksCorpus and Wikipedia only, where we reuse all pretraining hyper-parameters as in the original BERT. Then, we scale up the training of XLNet-Large by using all the datasets described above. Specifically, we train on 512 TPU v3 chips for 500K steps with an Adam weight decay optimizer, linear learning rate decay, and a batch size of 8192, which takes about 5.5 days. It was

High cost on computation …

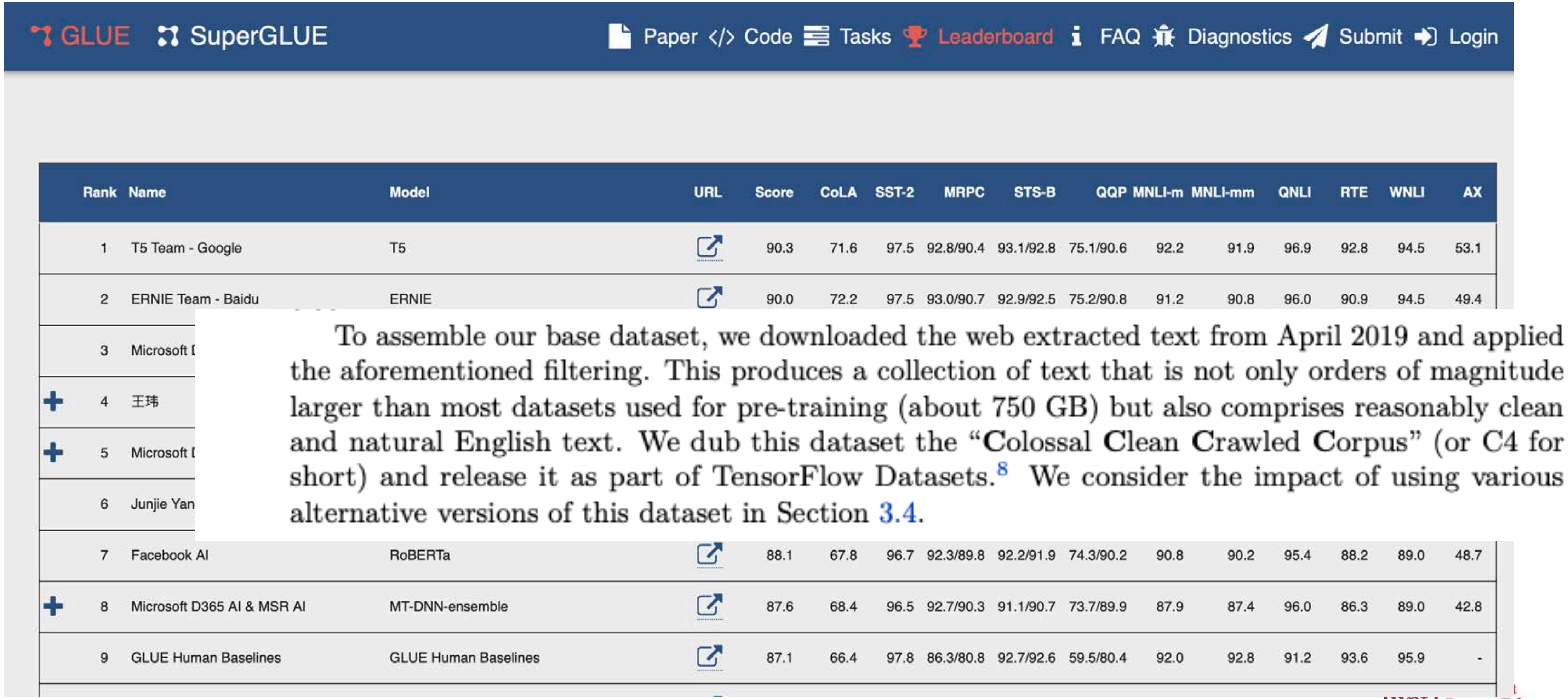
Anna Rogers

Thinking aloud: computational linguistics, cognition, AI and NLP



Model training cost clarification. the price of training XLNet was estimated as follows: the paper states that it was trained on 512 TPU v3 chips for 2.5 days, i.e. 60 hours. Google on-demand price for TPU v-3 is currently $8, which amounts to $245,760 before fine-tuning. James Bradbury points out that authors could actually mean “devices” or “cores”, which would bring it down to $61,440 or $30,720, respectively. I would add that even in this most optimistic scenario the model would still cost more than the stipend of the graduate student working on it, and still be unrealistic for most labs.

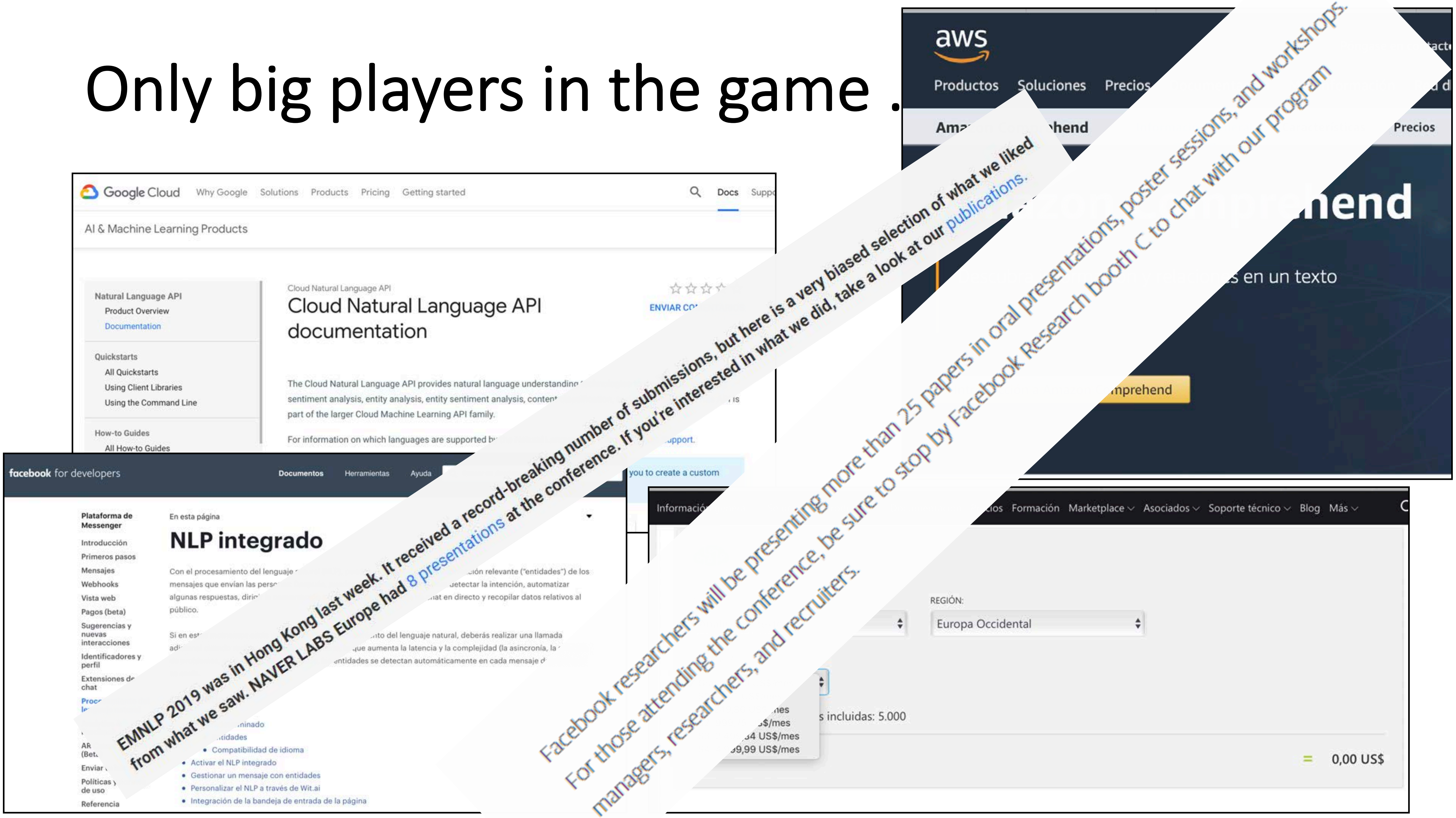
Tons of words …. clean and selected?



Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
2	ERNIE Team - Baidu	ERNIE		90.0	72.2	97.5	93.0/90.7	92.9/92.5	75.2/90.8	91.2	90.8	96.0	90.9	94.5	49.4
3	Microsoft [														
4	王玮														
5	Microsoft [														
6	Junjie Yan														
7	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
8	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
9	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

To assemble our base dataset, we downloaded the web extracted text from April 2019 and applied the aforementioned filtering. This produces a collection of text that is not only orders of magnitude larger than most datasets used for pre-training (about 750 GB) but also comprises reasonably clean and natural English text. We dub this dataset the "Colossal Clean Crawled Corpus" (or C4 for short) and release it as part of TensorFlow Datasets.[8] We consider the impact of using various alternative versions of this dataset in Section 3.4.

Only big players in the game .



Google Cloud · Why Google · Solutions · Products · Pricing · Getting started · Docs · Supp

AI & Machine Learning Products

Natural Language API
Product Overview
Documentation

Quickstarts
All Quickstarts
Using Client Libraries
Using the Command Line

How-to Guides
All How-to Guides

Cloud Natural Language API

Cloud Natural Language API documentation

ENVIAR CO

The Cloud Natural Language API provides natural language understanding sentiment analysis, entity analysis, entity sentiment analysis, conten is part of the larger Cloud Machine Learning API family.

For information on which languages are supported b upport.

you to create a custom

facebook for developers · Documentos · Herramientas · Ayuda

Plataforma de Messenger
Introducción
Primeros pasos
Mensajes
Webhooks
Vista web
Pagos (beta)
Sugerencias y nuevas interacciones
Identificadores y perfil
Extensiones de chat
Proc
AR (Bet
Enviar
Políticas de uso
Referencia

En esta página

NLP integrado

Con el procesamiento del lenguaje ción relevante ("entidades") de los mensajes que envían las pers detectar la intención, automatizar algunas respuestas, dirig hat en directo y recopilar datos relativos al público.

Si en est nto del lenguaje natural, deberás realizar una llamada adi que aumenta la latencia y la complejidad (la asincronía, la entidades se detectan automáticamente en cada mensaje d

- minado
- tidades
  - Compatibilidad de idioma
- Activar el NLP integrado
- Gestionar un mensaje con entidades
- Personalizar el NLP a través de Wit.ai
- Integración de la bandeja de entrada de la página

aws · Productos · Soluciones · Precios

Ama hend · Precios

hend

s en un texto

mprehend

Informació · cios · Formación · Marketplace · Asociados · Soporte técnico · Blog · Más

REGIÓN: Europa Occidental

nes
$/mes
34 US$/mes
99,99 US$/mes

s incluidas: 5.000

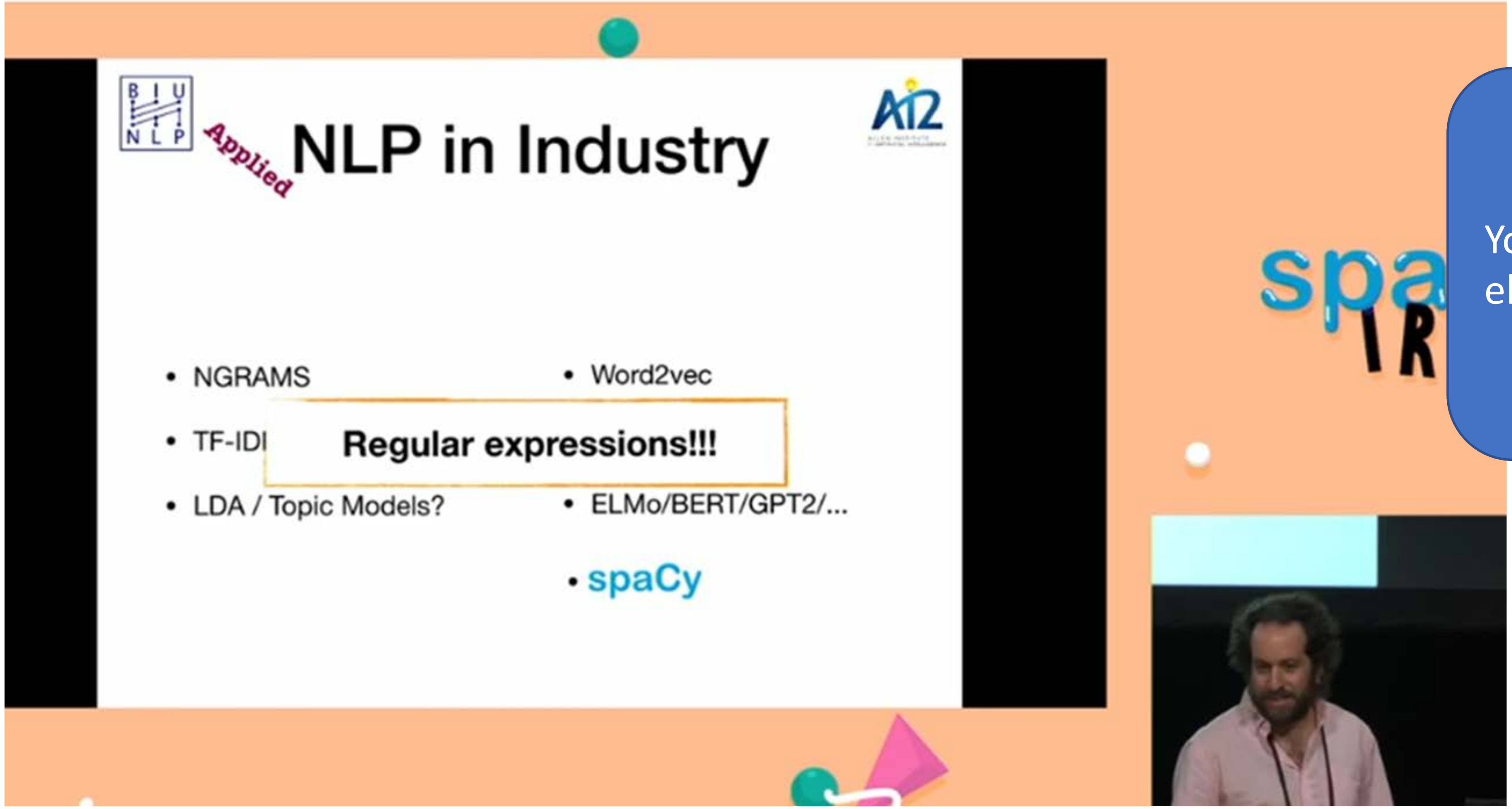
= 0,00 US$

EMNLP 2019 was in Hong Kong last week. It received a record-breaking number of submissions, but here is a very biased selection of what we liked from what we saw. NAVER LABS Europe had 8 presentations at the conference. If you're interested in what we did, take a look at our publications.

Facebook researchers will be presenting more than 25 papers in oral presentations, poster sessions, and workshops. For those attending the conference, be sure to stop by Facebook Research booth C to chat with our program managers, researchers, and recruiters.

But, still a lot to do with old technics …

- https://www.youtube.com/watch?v=e12danHhlic



Applied NLP in Industry

- NGRAMS
- TF-IDI
- LDA / Topic Models?
- Word2vec
- ELMo/BERT/GPT2/...
- spaCy

Regular expressions!!!

Yoav Goldberg: The missing elements in NLP (spaCy IRL 2019)

Research Topics

- For low resourced languages,
- Multilingual BERT has been proposed ..
- A transfer technology has been proposed, …
- Also for downstream tasks for domains with particular terminology: health, law, … ?
- …

Obvious problem is how to cope with low-resource cases?

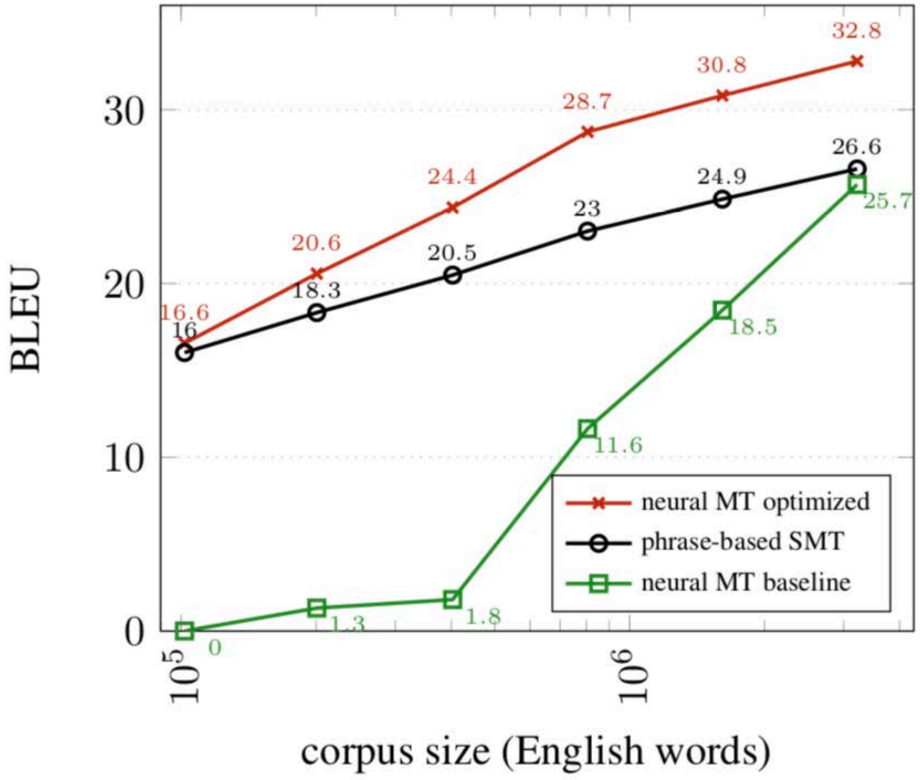
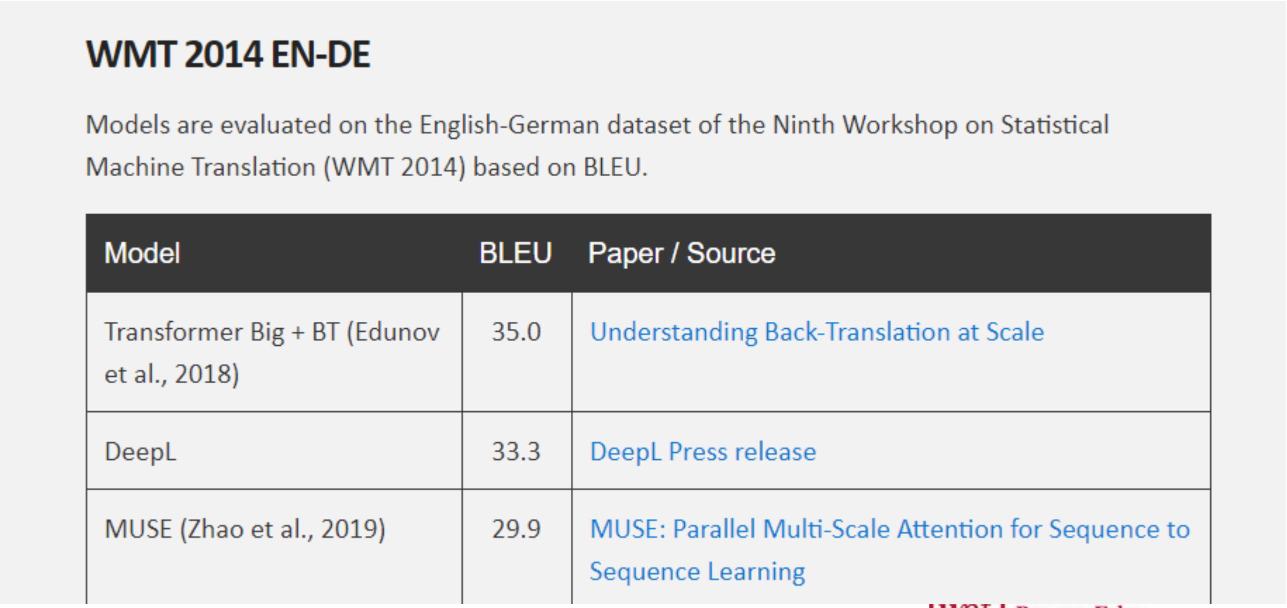


Figure 2: German→English learning curve, showing BLEU as a function of the amount of parallel training data, for PBSMT and NMT.

Sennrich and Zhang, 2019, Revisiting Low-Resource Neural Machine Translation: A Case Study

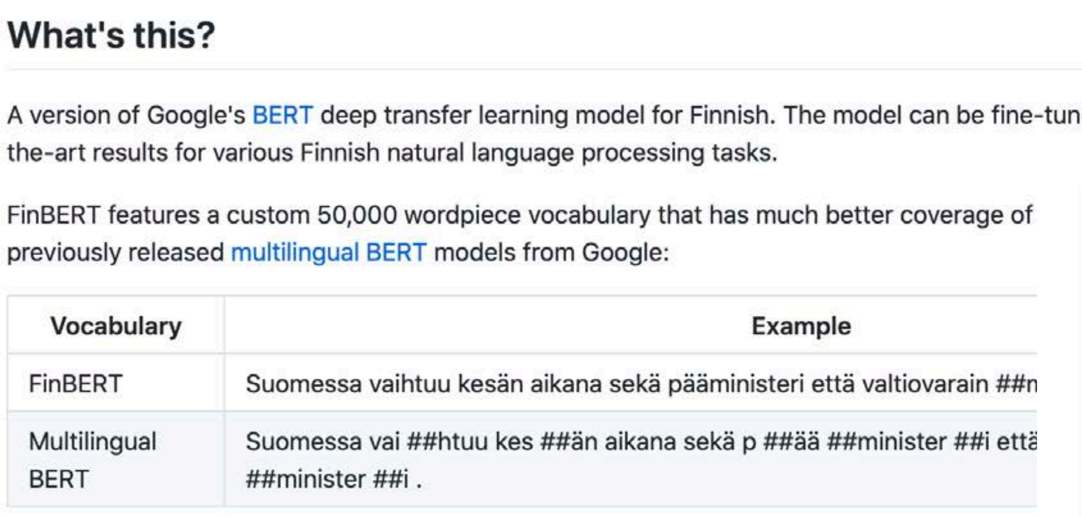


WMT 2014 EN-DE

Models are evaluated on the English-German dataset of the Ninth Workshop on Statistical Machine Translation (WMT 2014) based on BLEU.

Model	BLEU	Paper / Source
Transformer Big + BT (Edunov et al., 2018)	35.0	Understanding Back-Translation at Scale
DeepL	33.3	DeepL Press release
MUSE (Zhao et al., 2019)	29.9	MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning

Is "Multilingual BERT" a solution?

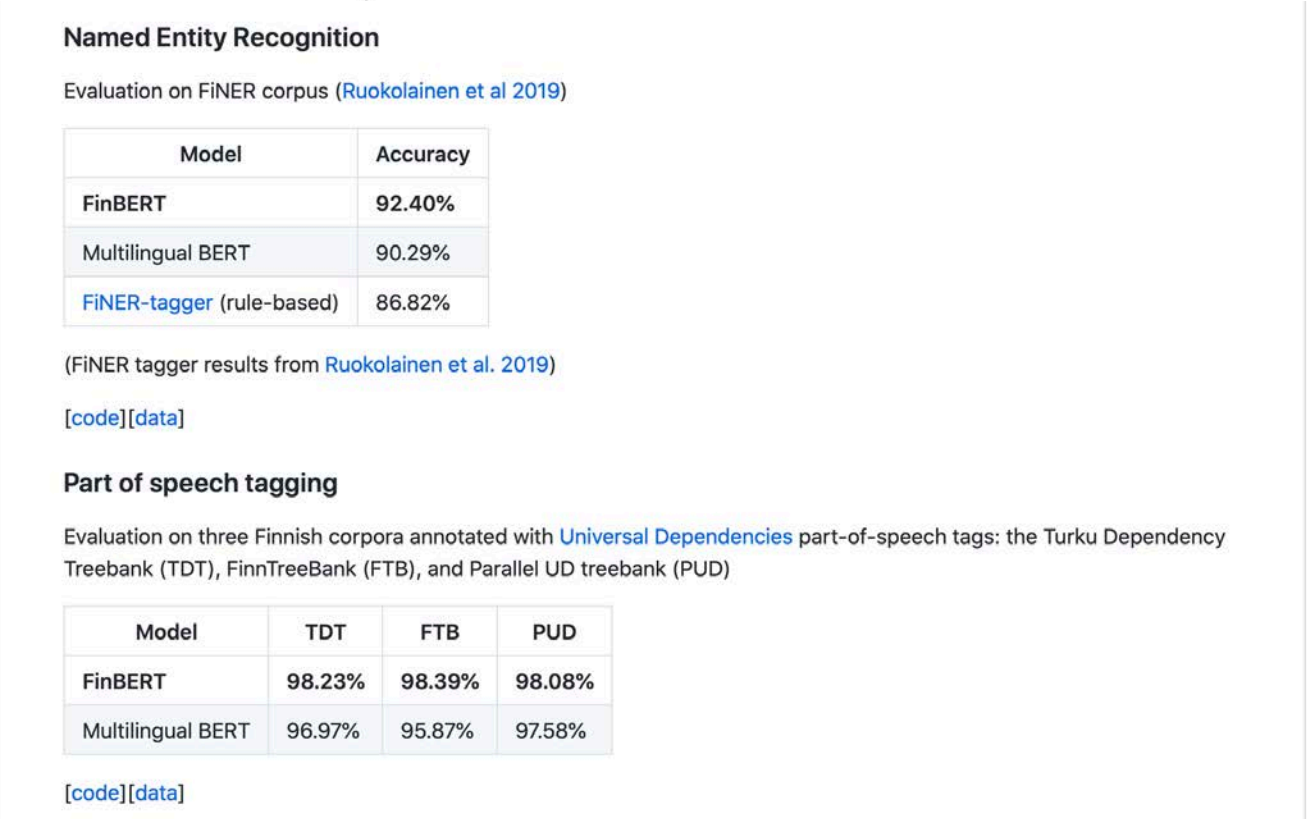


What's this?

A version of Google's BERT deep transfer learning model for Finnish. The model can be fine-tuned to achieve state-of-the-art results for various Finnish natural language processing tasks.

FinBERT features a custom 50,000 wordpiece vocabulary that has much better coverage of previously released multilingual BERT models from Google:

Vocabulary	Example
FinBERT	Suomessa vaihtuu kesän aikana sekä pääministeri että valtiovarain ##n
Multilingual BERT	Suomessa vai ##htuu kes ##än aikana sekä p ##ää ##minister ##i että ##minister ##i .



Named Entity Recognition

Evaluation on FiNER corpus (Ruokolainen et al 2019)

Model	Accuracy
FinBERT	92.40%
Multilingual BERT	90.29%
FiNER-tagger (rule-based)	86.82%

(FiNER tagger results from Ruokolainen et al. 2019)

[code][data]

Part of speech tagging

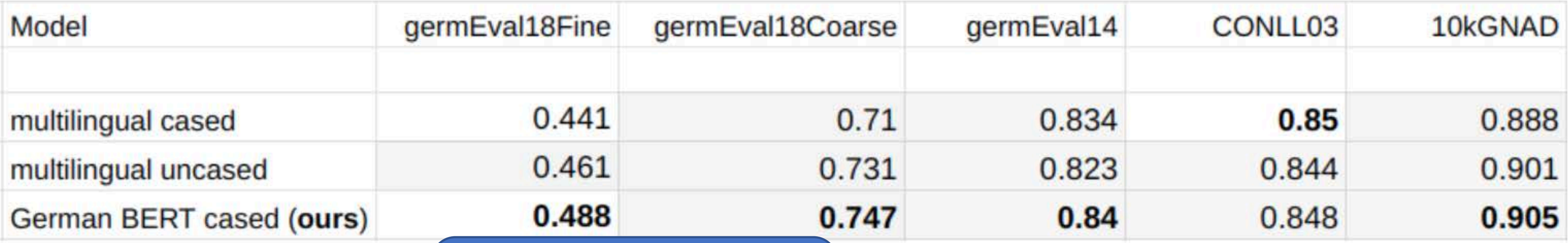
Evaluation on three Finnish corpora annotated with Universal Dependencies part-of-speech tags: the Turku Dependency Treebank (TDT), FinnTreeBank (FTB), and Parallel UD treebank (PUD)

Model	TDT	FTB	PUD
FinBERT	98.23%	98.39%	98.08%
Multilingual BERT	96.97%	95.87%	97.58%

[code][data]

FinBERT has been pre-trained on over 3 billion tokens (24B characters) of Finnish text drawn from news, online discussion, and internet crawls
https://github.com/TurkuNLP/FinBERT

German BERT has been trained with 1.826.856.564 words (approx.) and a single cloud TPU v2 with standard settings.



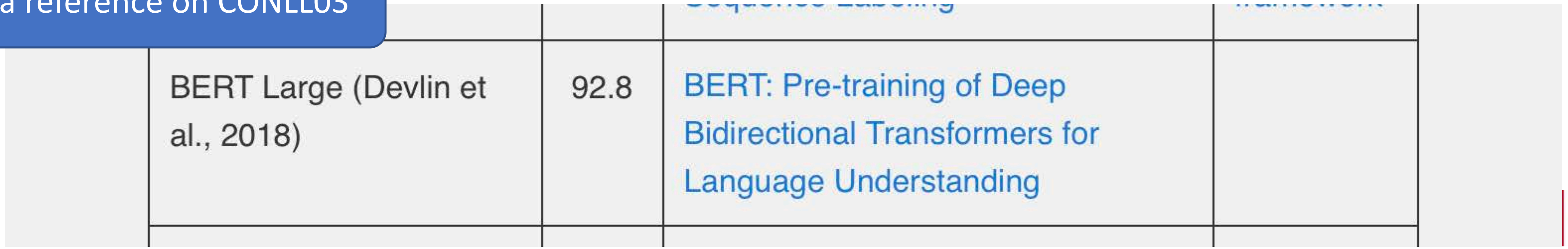
Model	germEval18Fine	germEval18Coarse	germEval14	CONLL03	10kGNAD
multilingual cased	0.441	0.71	0.834	**0.85**	0.888
multilingual uncased	0.461	0.731	0.823	0.844	0.901
German BERT cased **(ours)**	**0.488**	**0.747**	**0.84**	0.848	**0.905**

multiclass and binary sentiment classification

NER

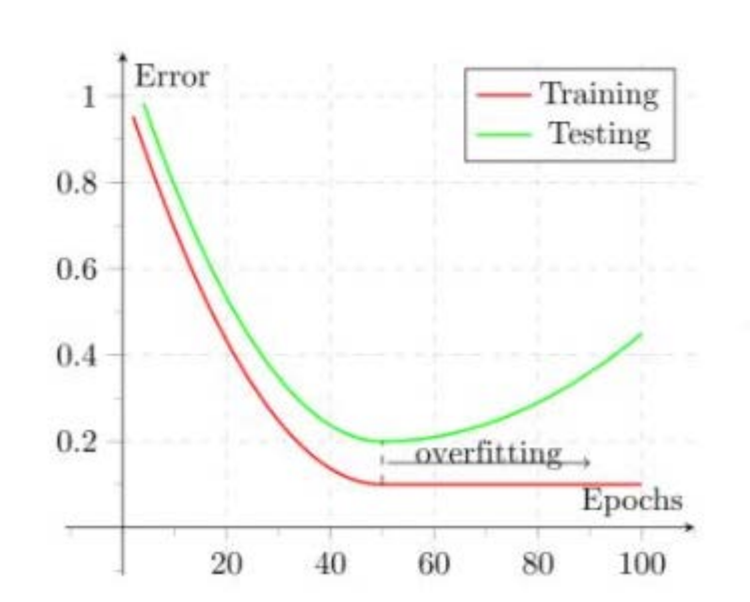
Doc classification

As a reference on CONLL03



		Sequence Labeling	framework
BERT Large (Devlin et al., 2018)	92.8	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	

What DL systems learn?



- DL are said to be able to select, on their own, the features that are relevant for a task on hand;
- DL systems are trained as a Language Model (LM) or predictors of parts of sentences and as a Classifier for specific tasks.
- For LM, DL mostly use lots of raw data and build dense, real number-valued and distributed vectors;
- DL systems for downstream tasks use previous pre-trained vectors and task specific annotated data and learn to relate them to cases of the task (millions of parameters or weights for the relations)

Is there a problem with Deep Learning?

1) DL only works well with tones of raw data. Is this amount of data sufficient and necessary?

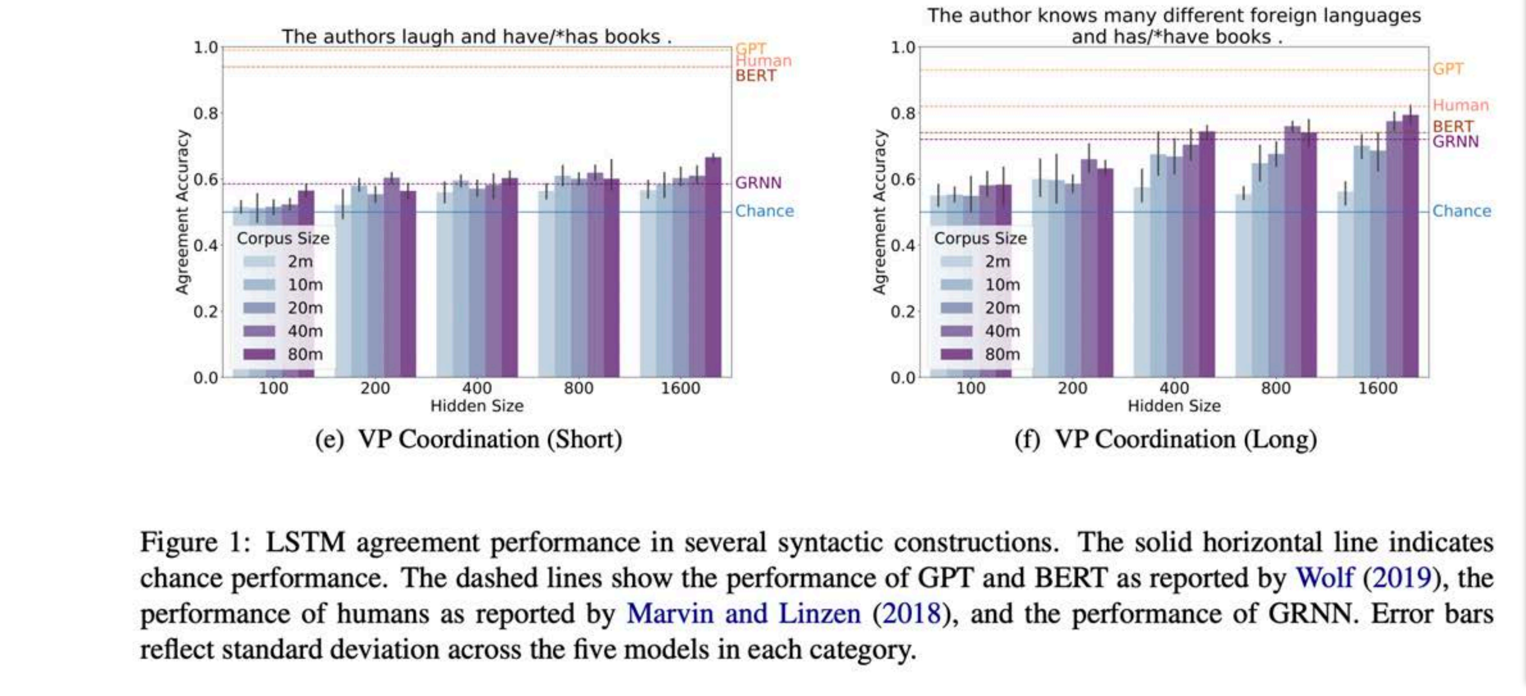
2) Why? does DL learn abstract rules? Does it generalize?

Language Resources are language specific information used to generalize, aren’t they needed anymore?

1) Is data enough?

Problems with word prediction in complex syntactic sentences cannot be overcomed by the model capacity or an increased corpus size. They conclude that reliable and data-efficient learning of syntax is likely to require external supervision signals or a stronger inductive bias.

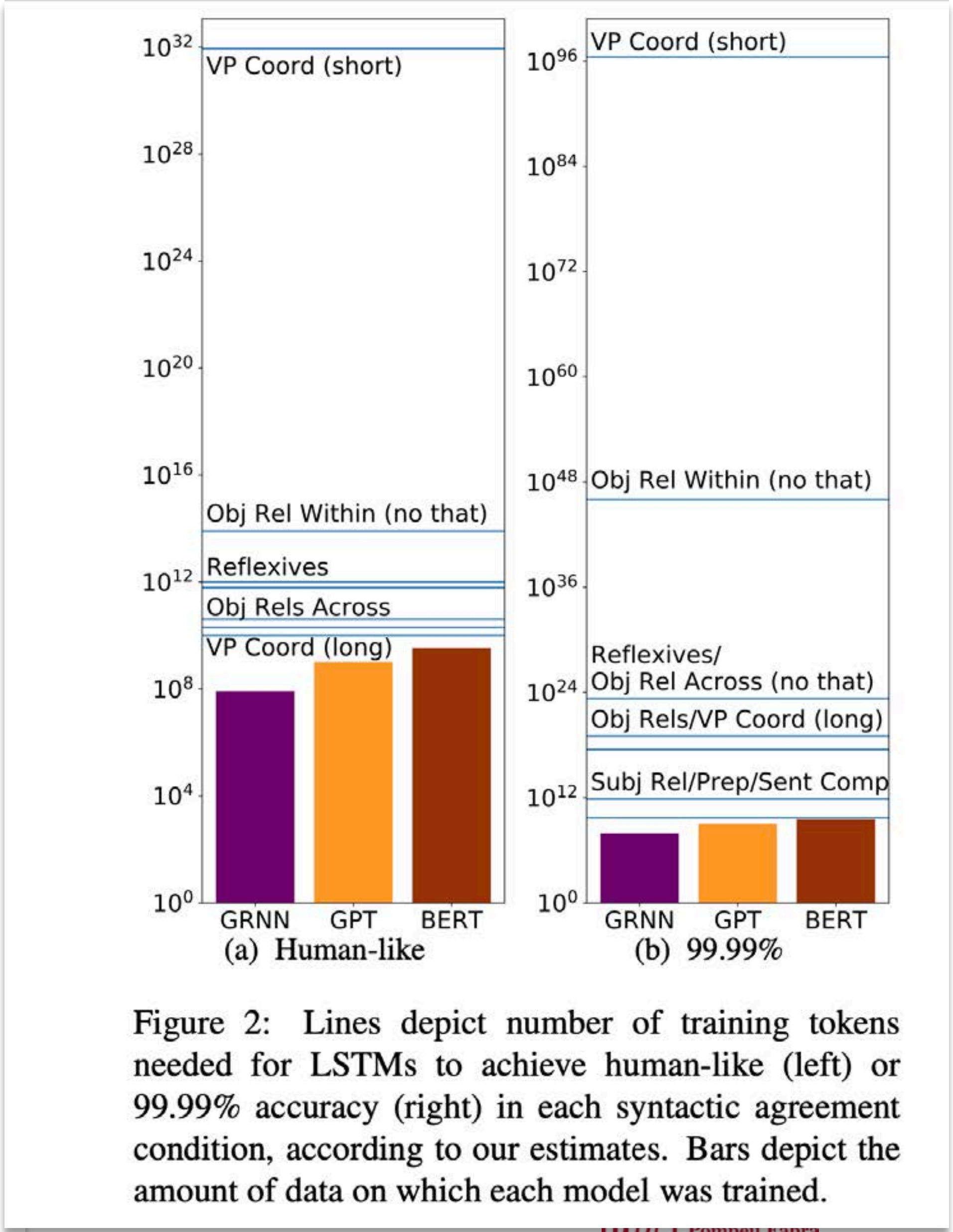
Schijndel, M.; Mueller, A. and Linzen, T. (2019) Quantity doesn't buy quality syntax with neural language models. EMNLP 2019



(e) VP Coordination (Short)

(f) VP Coordination (Long)

Figure 1: LSTM agreement performance in several syntactic constructions. The solid horizontal line indicates chance performance. The dashed lines show the performance of GPT and BERT as reported by Wolf (2019), the performance of humans as reported by Marvin and Linzen (2018), and the performance of GRNN. Error bars reflect standard deviation across the five models in each category.

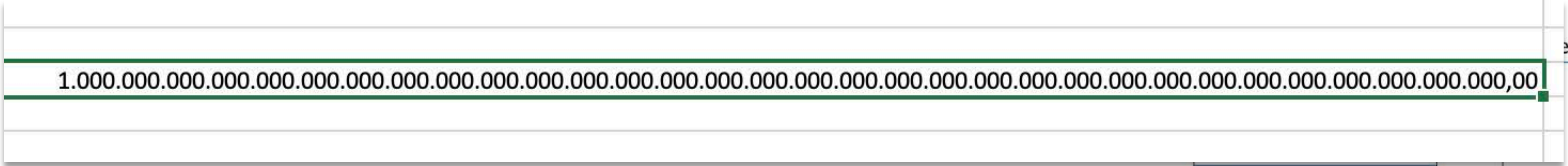


(a) Human-like

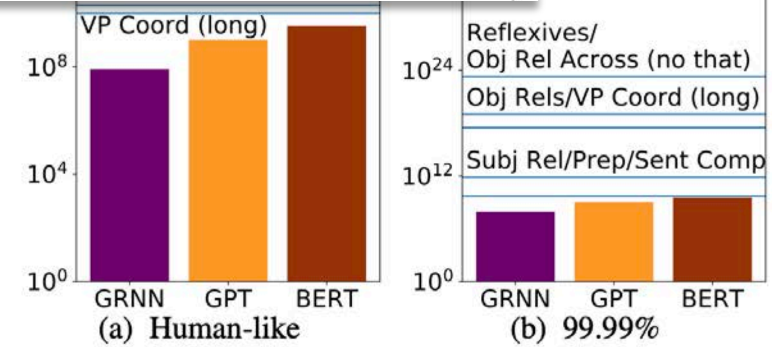
(b) 99.99%

Figure 2: Lines depict number of training tokens needed for LSTMs to achieve human-like (left) or 99.99% accuracy (right) in each syntactic agreement condition, according to our estimates. Bars depict the amount of data on which each model was trained.

Is data enough? Is it possible?
Tokens required: visualize the figure!



1.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000,00



(a) Human-like

(b) 99.99%

Figure 2: Lines depict number of training tokens needed for LSTMs to achieve human-like (left) or 99.99% accuracy (right) in each syntactic agreement condition, according to our estimates. Bars depict the amount of data on which each model was trained.

Tons of data and more for domain adaptation Yogatama et al. 2019

- Number of additional training examples for a domain task, that is: how much information from generic training is reused for a domain task.
- 400,000

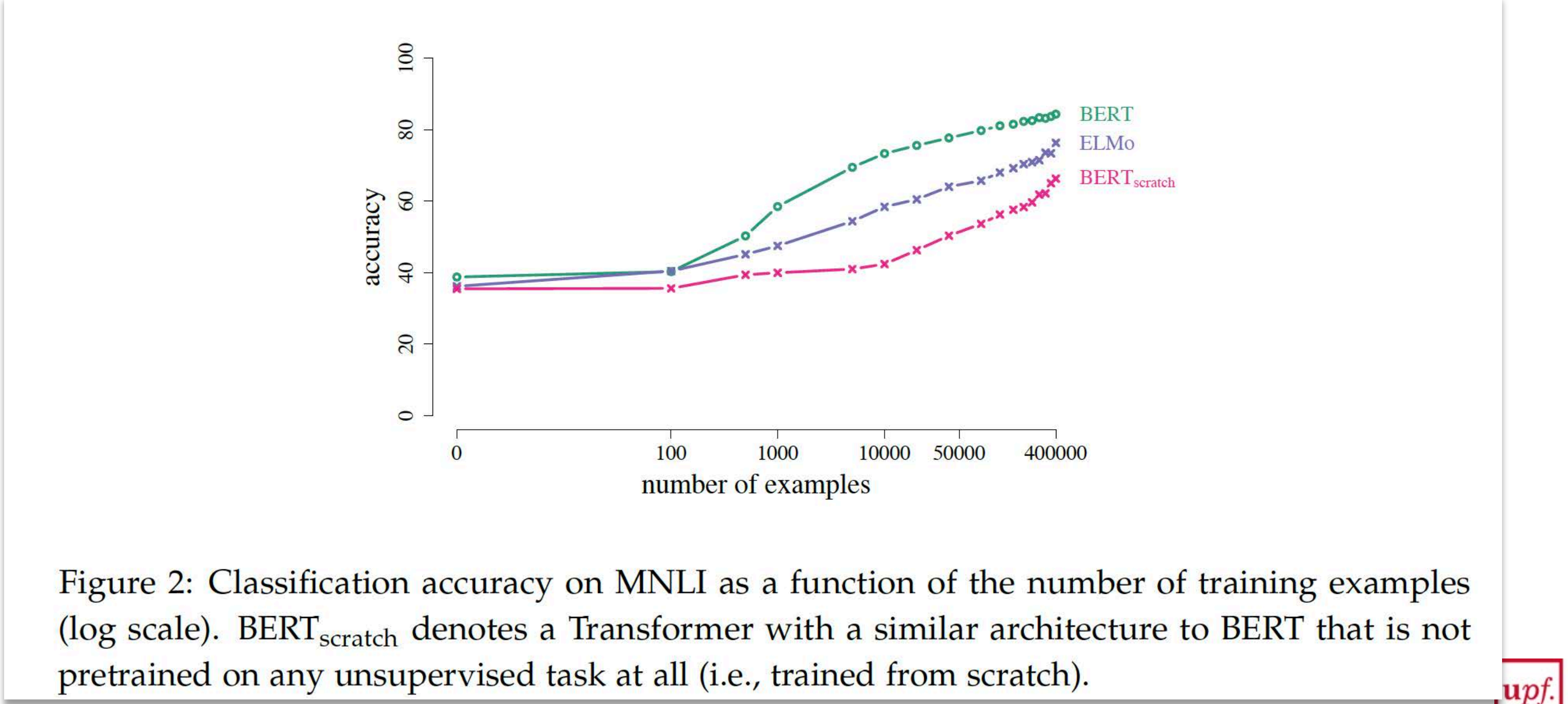
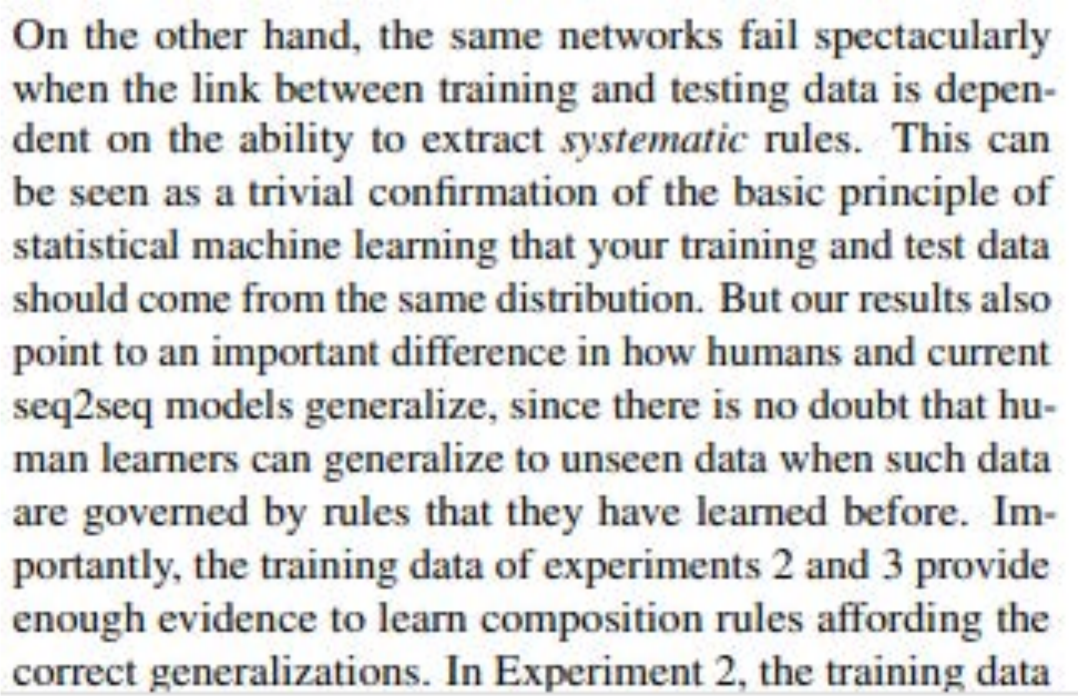


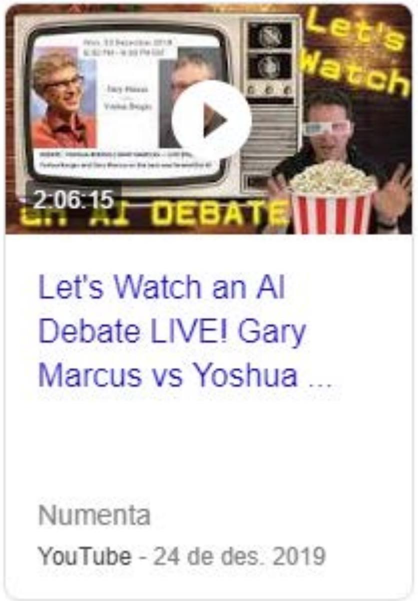
Figure 2: Classification accuracy on MNLI as a function of the number of training examples (log scale). BERT$_{\text{scratch}}$ denotes a Transformer with a similar architecture to BERT that is not pretrained on any unsupervised task at all (i.e., trained from scratch).

2) Learning the rule vs. Learning the distribution of the dataset



On the other hand, the same networks fail spectacularly when the link between training and testing data is dependent on the ability to extract *systematic* rules. This can be seen as a trivial confirmation of the basic principle of statistical machine learning that your training and test data should come from the same distribution. But our results also point to an important difference in how humans and current seq2seq models generalize, since there is no doubt that human learners can generalize to unseen data when such data are governed by rules that they have learned before. Importantly, the training data of experiments 2 and 3 provide enough evidence to learn composition rules affording the correct generalizations. In Experiment 2, the training data

Lake and Baroni (2018)
Currently DL cannot work compositionally!



Let's Watch an AI Debate LIVE! Gary Marcus vs Yoshua ...

Numenta
YouTube - 24 de des. 2019

G. Marcus' example about rule learning

1010 => 1010
1110 => 1110
1000 => 1000
1001 => 1000

New research lines

1. Do DL systems select linguistically motivated features?
2. Are they able to generalize, i.e. forgetting what is not relevant for creating abstract categories?

- Why are these questions relevant for the Language Resources community?

Because LR were used to gain in generalization: linguistic abstract categories to support language understanding tasks.

Methods

Probing tasks or diagnostic classifiers

Language understanding tasks

Techniques to visualize information

Probing tasks or diagnosis classifiers

- Adi et al. (2017) trained classifiers with word embeddings -- CBOW and LSTM autoencoder – and predicted characteristics of the senetences like length, word content and word order.

Data: 1M sentences from Wikipedia and ad-hoc data.

- Conneau et al. (2018) introduced ten new tasks organized by the type of linguistic properties. They got sentence representations from different encoders and used them to train specialized classifiers for the probing tasks and confirmed that embeddings capture features like tense and number, depth of syntactic structure, etc.

Datos: Toronto Book Corpus, training with 100k sentences and 10k sentences for validating each task.

Linzen et al. (2016) measured whether LM's predictions, obtained with a LSTM, reflect a correct analysis of sentence structure (syntax) by comparing the probability to sentences differing only in gramaticality.

Data: 1.35M Wikipedia (9% for training)

http://tallinzen.net/projects/lstm_agreement

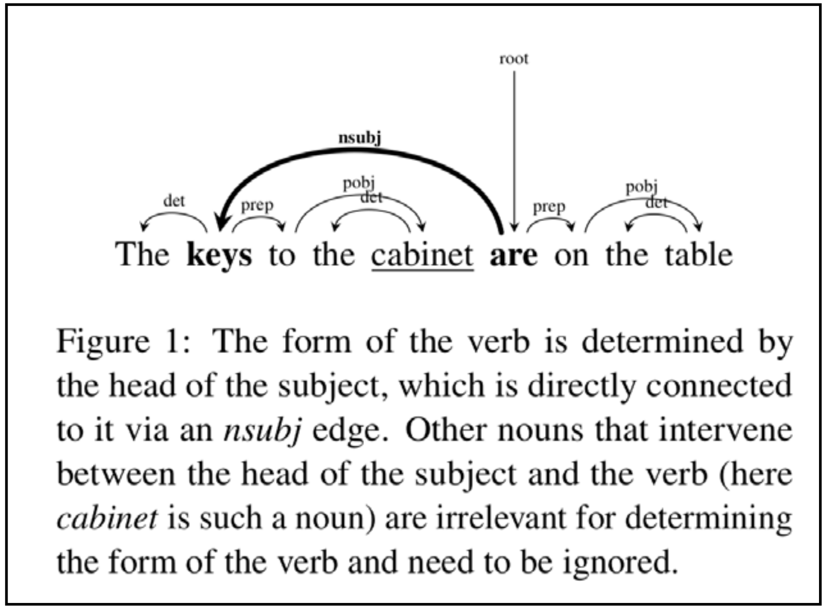
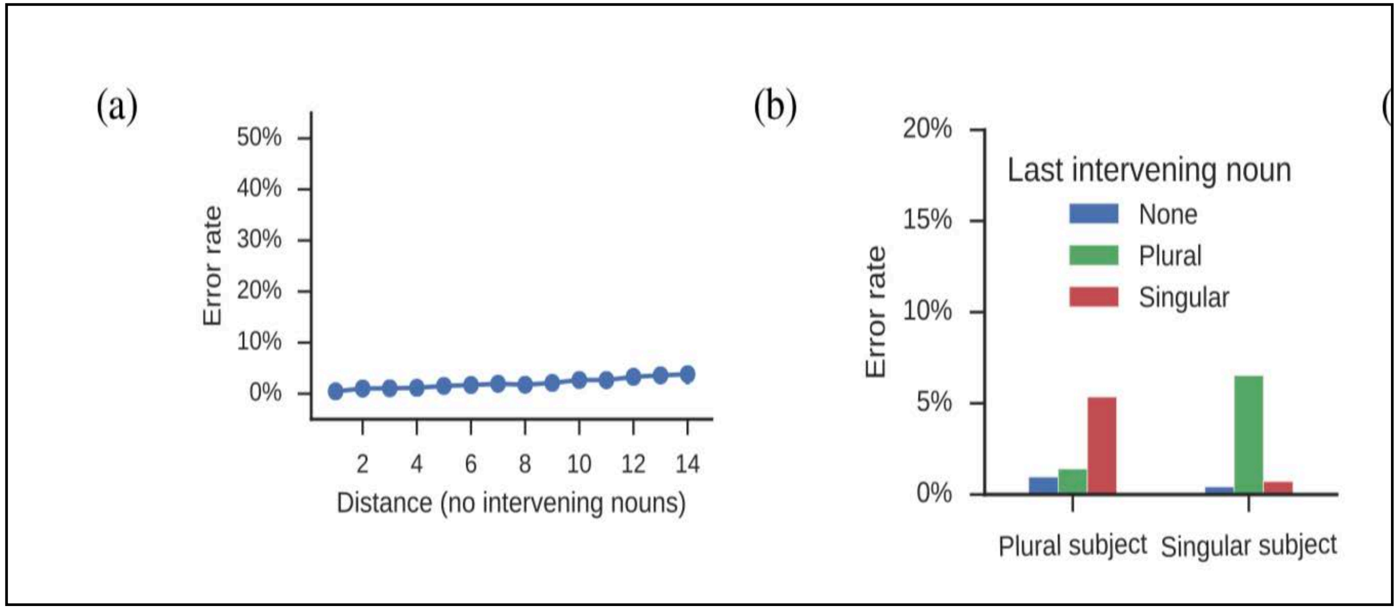


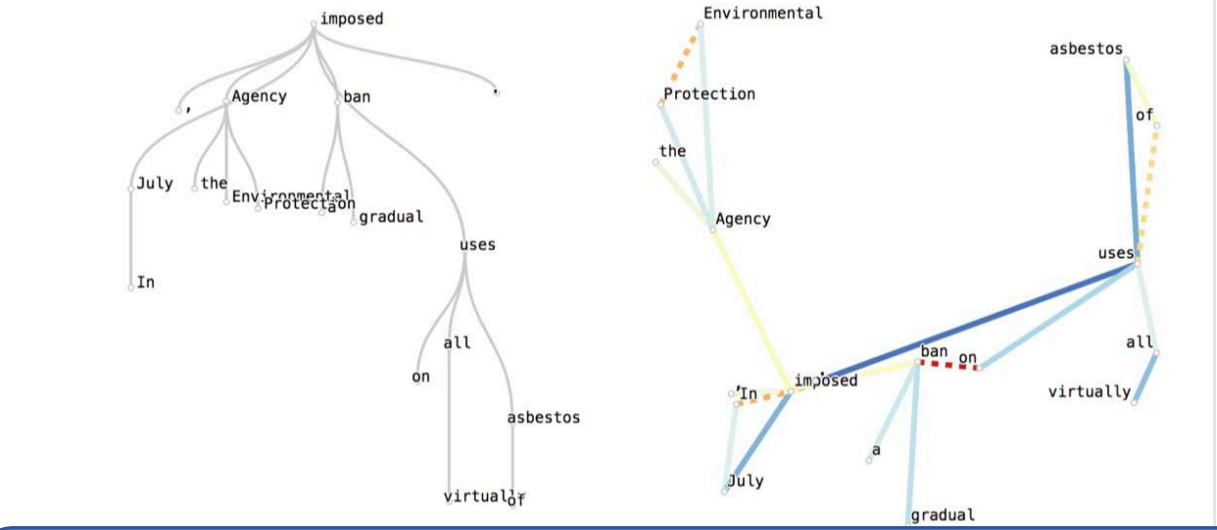
Figure 1: The form of the verb is determined by the head of the subject, which is directly connected to it via an *nsubj* edge. Other nouns that intervene between the head of the subject and the verb (here *cabinet* is such a noun) are irrelevant for determining the form of the verb and need to be ignored.

Accuracy decreases when distracting nouns are between the head of the subject and the verb, specially in relative clauses (Marvin and Linzen, 2018)

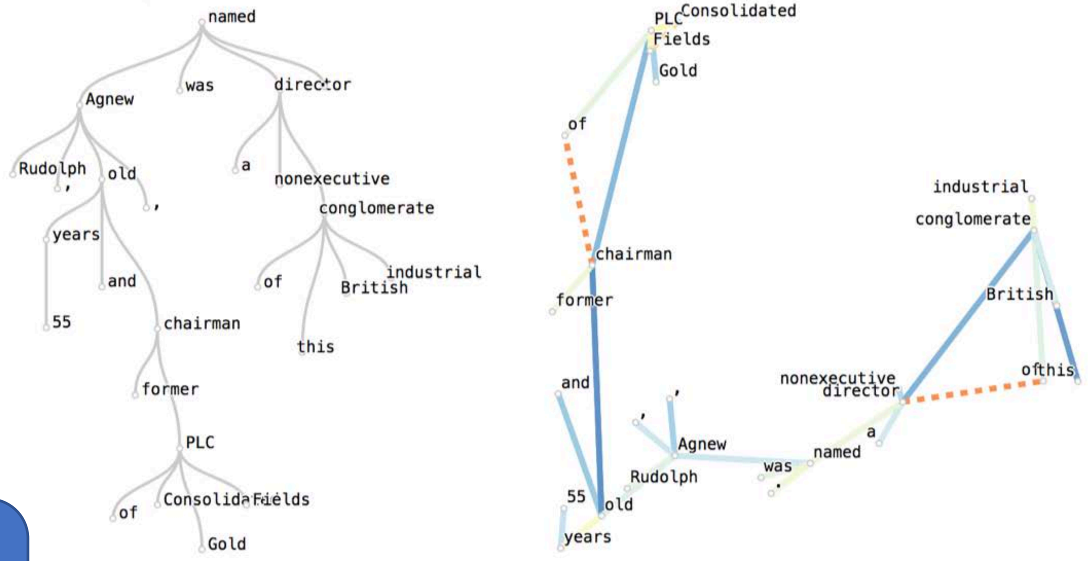


Probing tasks: Classifying relations by the distance^2 between word pairs Coenen et al. (2019)

"In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos."



"Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate."



Labelled embeddings with BERT-base used to train a L2 regularized linear multiclass classifier to predict the type of dependency relation between two tokens.
Data: 3.1 million dependency relations got from Penn Treebank

Probing tasks: Datasets to test the actual performance of bilingual lexicon induction

Czarnowska et al. (2019) introduce a new resource: 40 morphologically complete dictionaries for 5 Slavic and 5 Romance languages which contain the inflectional paradigm of every word they hold. They provide an upper bound for the performance on the generalization task.

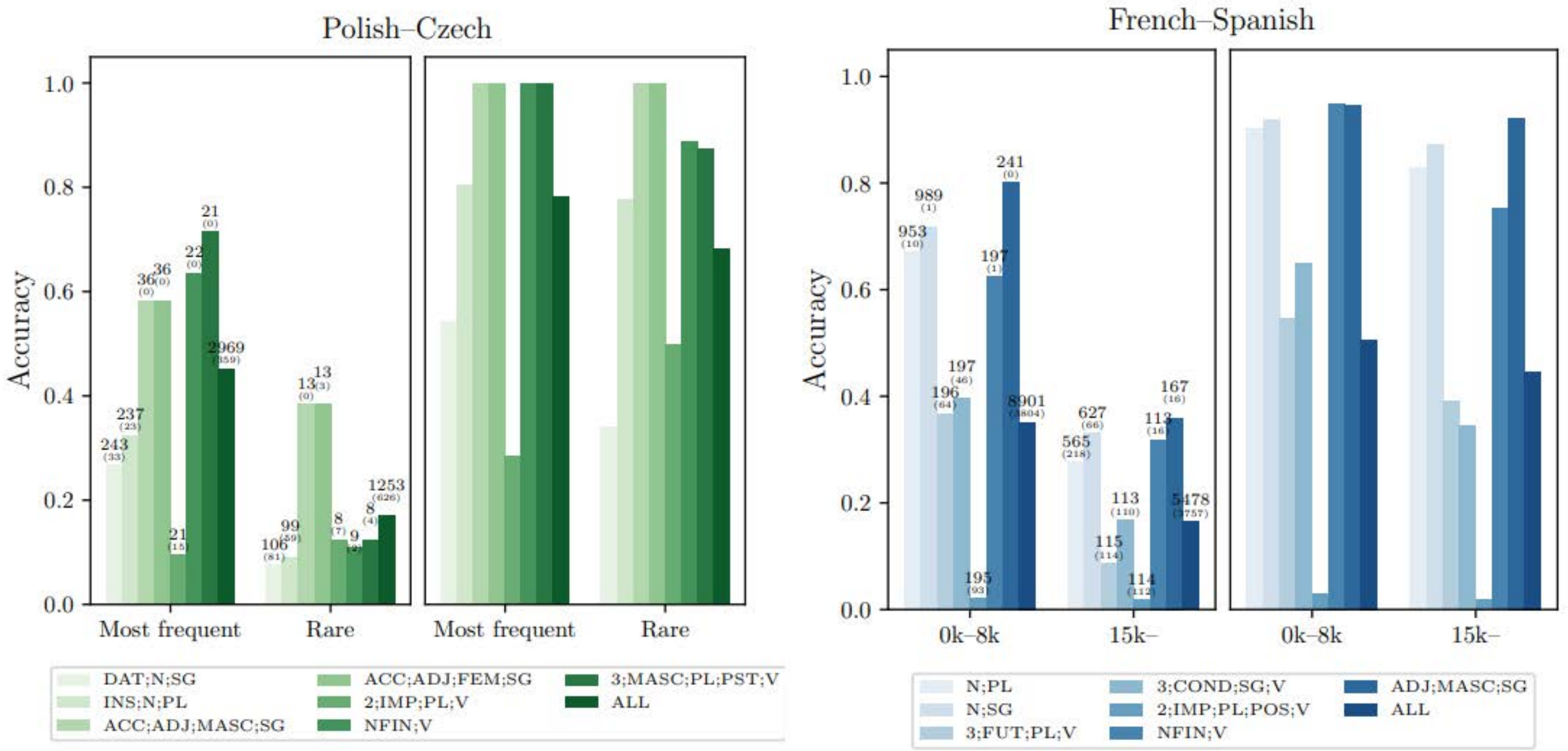


Figure 3: The performance on the standard BLI task (left side of the graphs) and the controlled for lexeme BLI (right side) for words pairs belonging to the most frequent paradigms and the infrequent paradigms. The numbers above the bars are dictionary sizes and the number of out-of-vocabulary forms in each dictionary (bracketed).

Probing tasks (LM probability assignment)

- Gulordava et al. (2018) creating comparable grammatical but non-sense sentences like “Green ideas sleep…” making sure that **only syntactic information** is considered.

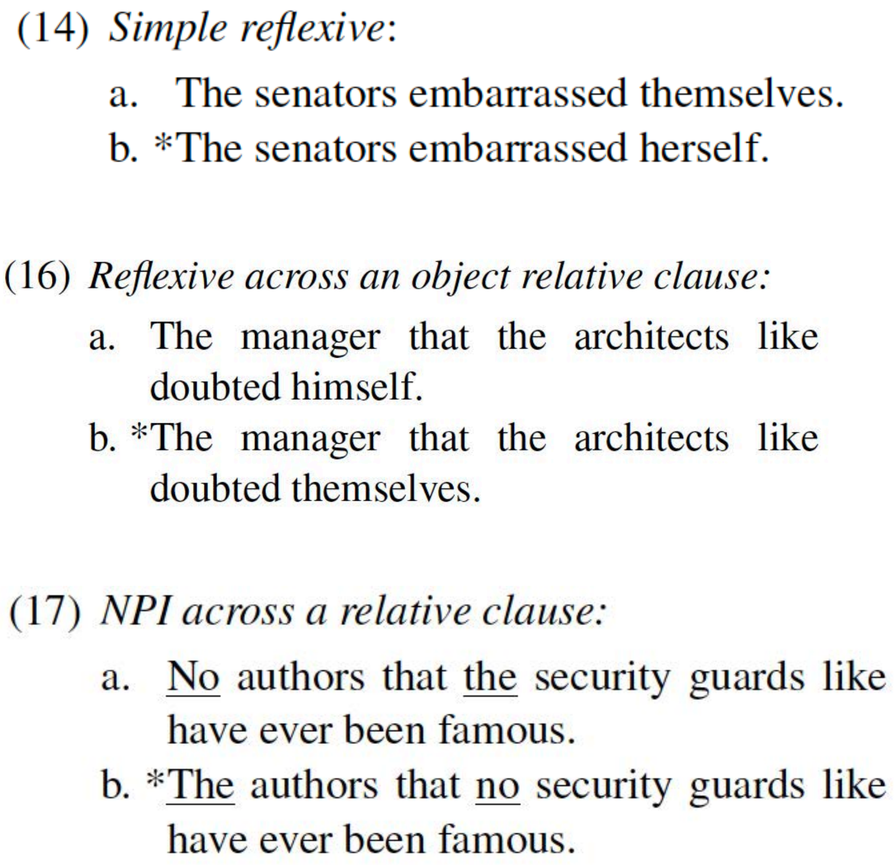
Data: training with Wikipedia and validating with Universal Dependencies Treebank data. https://github.com/facebookresearch/colorlessgreenRNNs.

- Marvin and Linzen (2018) check whether a ML-LSTM assigns less probability to a ungramatical sentence than to a gramatical one for agreement, reflexive anaphora and negative polarity elements.

Data: automatically building a resource made of sentence pairs: gramatical and ungrammatical. https://github.com/BeckyMarvin/LM_syneval.

Probing tasks (LM)

Marvin & Linzen (2018)



(14) *Simple reflexive*:

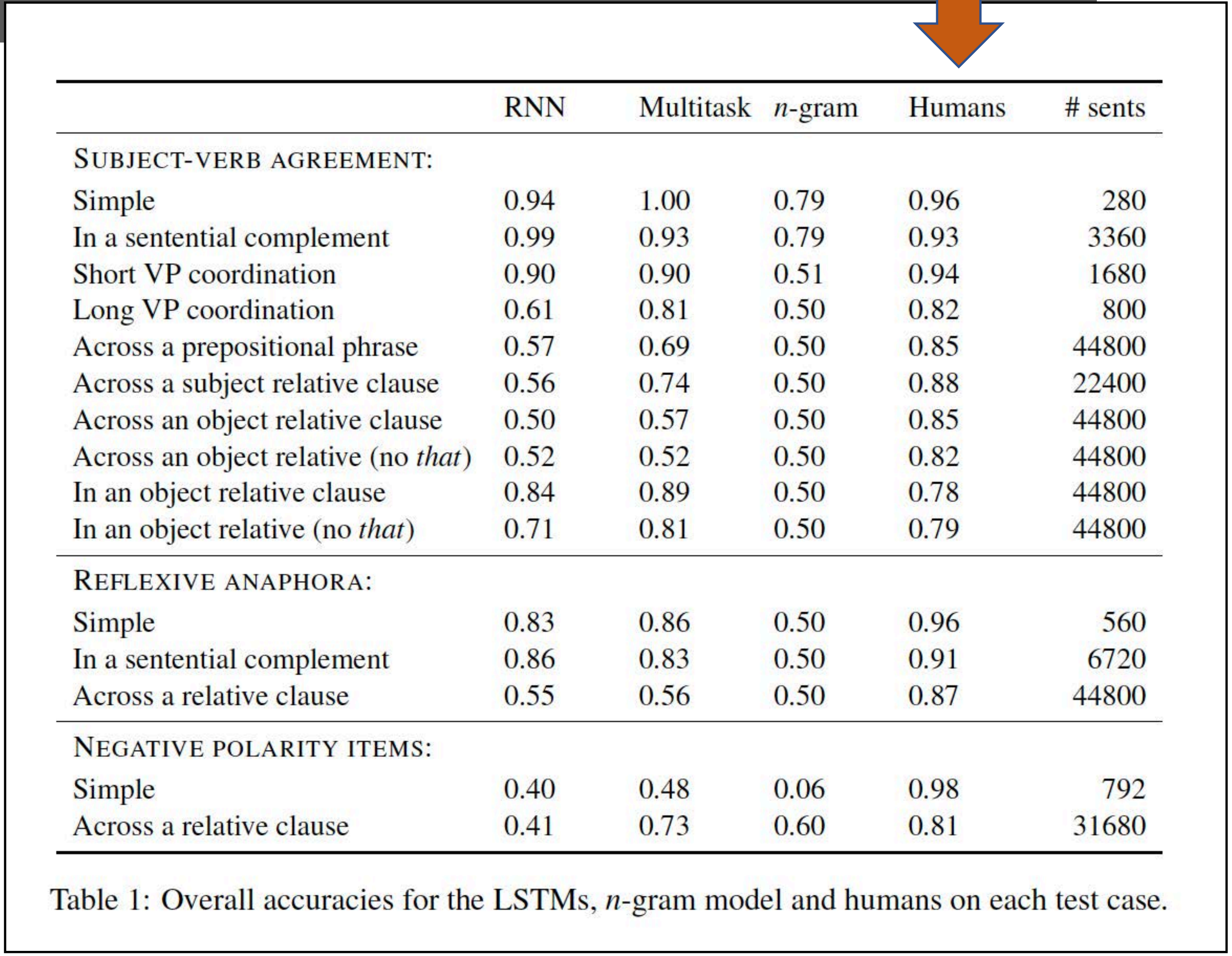
a. The senators embarrassed themselves.
b. *The senators embarrassed herself.

(16) *Reflexive across an object relative clause:*

a. The manager that the architects like doubted himself.
b. *The manager that the architects like doubted themselves.

(17) *NPI across a relative clause:*

a. No authors that the security guards like have ever been famous.
b. *The authors that no security guards like have ever been famous.



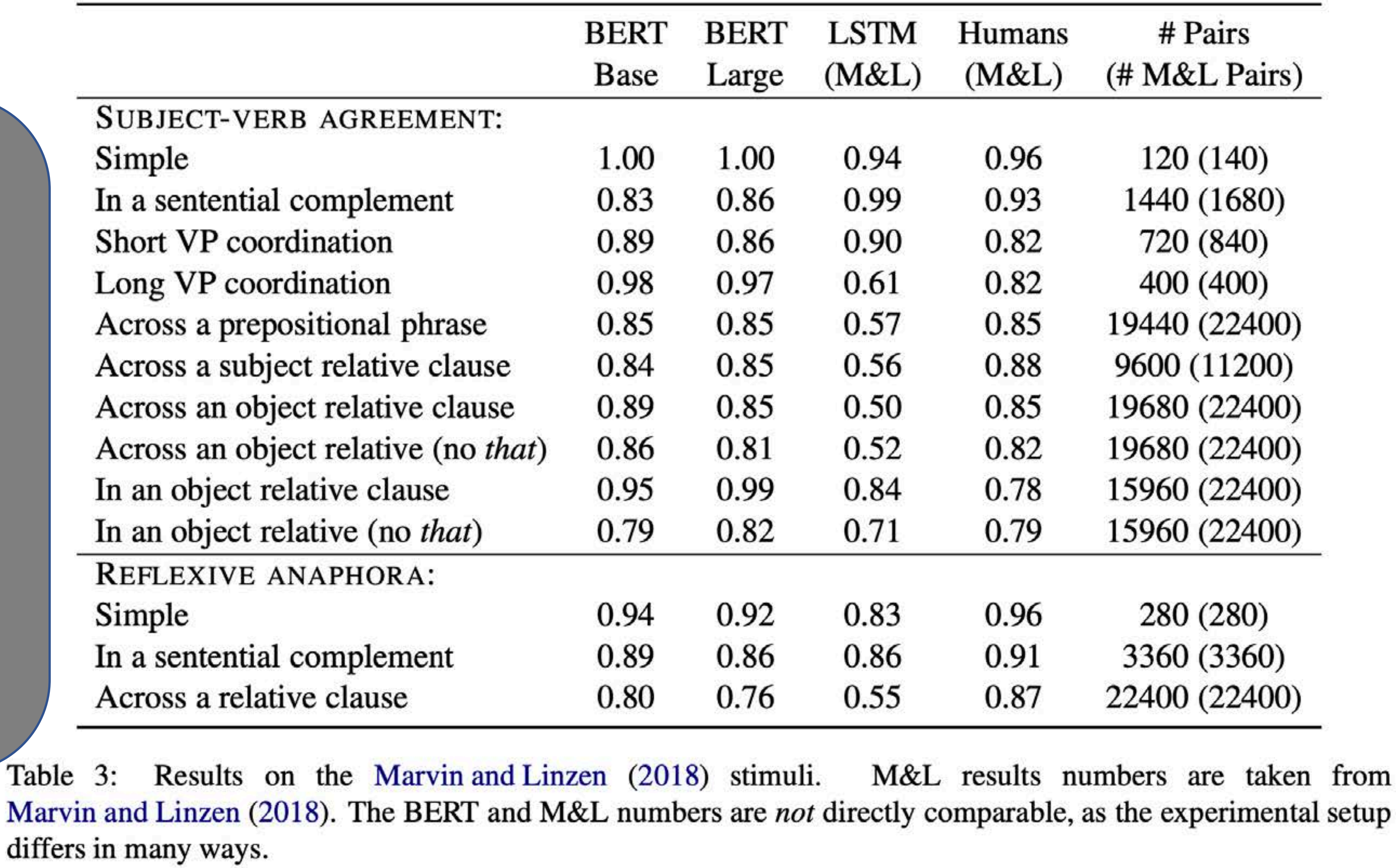
	RNN	Multitask	*n*-gram	Humans	# sents
SUBJECT-VERB AGREEMENT:					
Simple	0.94	1.00	0.79	0.96	280
In a sentential complement	0.99	0.93	0.79	0.93	3360
Short VP coordination	0.90	0.90	0.51	0.94	1680
Long VP coordination	0.61	0.81	0.50	0.82	800
Across a prepositional phrase	0.57	0.69	0.50	0.85	44800
Across a subject relative clause	0.56	0.74	0.50	0.88	22400
Across an object relative clause	0.50	0.57	0.50	0.85	44800
Across an object relative (no *that*)	0.52	0.52	0.50	0.82	44800
In an object relative clause	0.84	0.89	0.50	0.78	44800
In an object relative (no *that*)	0.71	0.81	0.50	0.79	44800
REFLEXIVE ANAPHORA:					
Simple	0.83	0.86	0.50	0.96	560
In a sentential complement	0.86	0.83	0.50	0.91	6720
Across a relative clause	0.55	0.56	0.50	0.87	44800
NEGATIVE POLARITY ITEMS:					
Simple	0.40	0.48	0.06	0.98	792
Across a relative clause	0.41	0.73	0.60	0.81	31680

Table 1: Overall accuracies for the LSTMs, *n*-gram model and humans on each test case.

Probing tasks (LM) with BERT

Goldberg, Y. Assessing BERT’s Syntactic Abilities. (2019) with Marvin & Linzen (2018).

. Less data
. BERT sees the whole sentence
. Trained with larger corpus

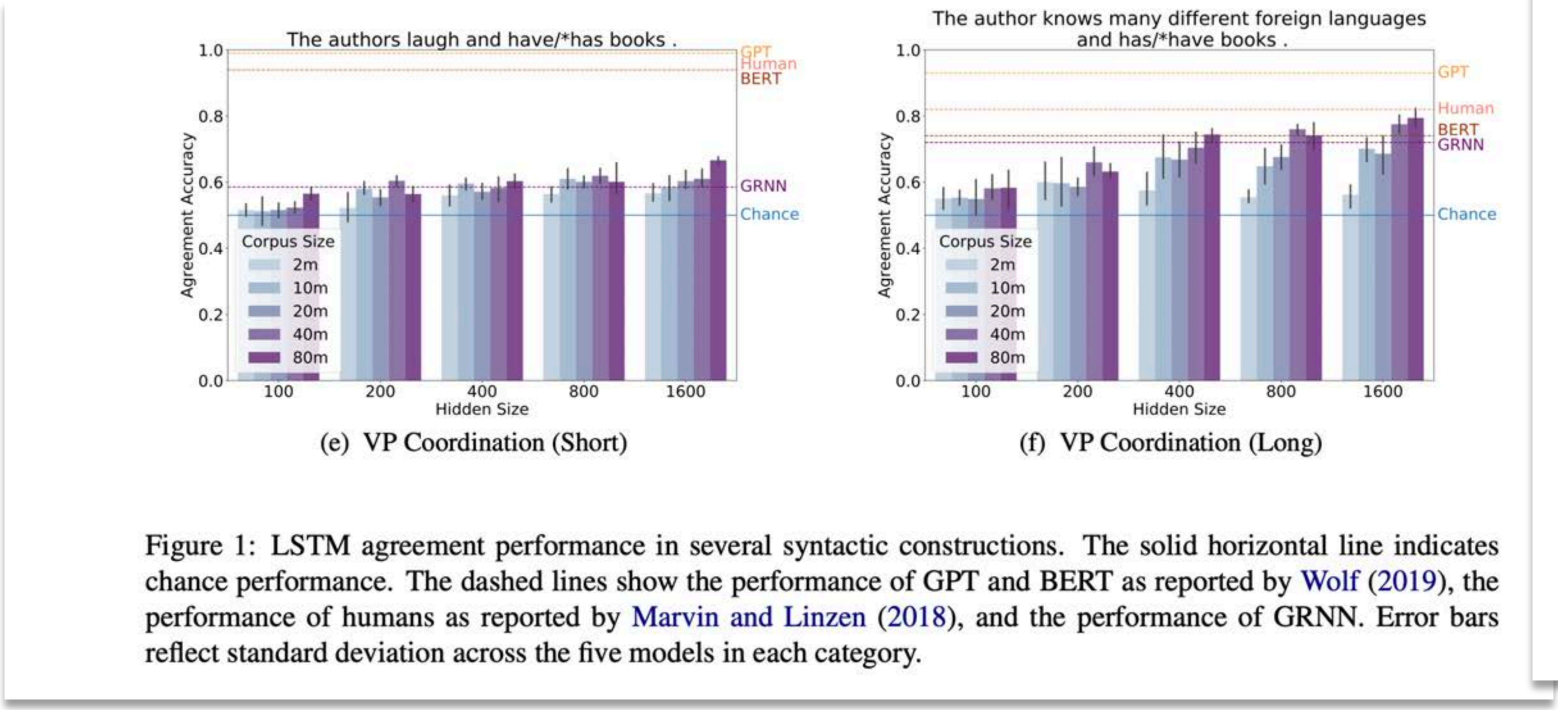


	BERT Base	BERT Large	LSTM (M&L)	Humans (M&L)	# Pairs (# M&L Pairs)
SUBJECT-VERB AGREEMENT:					
Simple	1.00	1.00	0.94	0.96	120 (140)
In a sentential complement	0.83	0.86	0.99	0.93	1440 (1680)
Short VP coordination	0.89	0.86	0.90	0.82	720 (840)
Long VP coordination	0.98	0.97	0.61	0.82	400 (400)
Across a prepositional phrase	0.85	0.85	0.57	0.85	19440 (22400)
Across a subject relative clause	0.84	0.85	0.56	0.88	9600 (11200)
Across an object relative clause	0.89	0.85	0.50	0.85	19680 (22400)
Across an object relative (no *that*)	0.86	0.81	0.52	0.82	19680 (22400)
In an object relative clause	0.95	0.99	0.84	0.78	15960 (22400)
In an object relative (no *that*)	0.79	0.82	0.71	0.79	15960 (22400)
REFLEXIVE ANAPHORA:					
Simple	0.94	0.92	0.83	0.96	280 (280)
In a sentential complement	0.89	0.86	0.86	0.91	3360 (3360)
Across a relative clause	0.80	0.76	0.55	0.87	22400 (22400)

Table 3: Results on the Marvin and Linzen (2018) stimuli. M&L results numbers are taken from Marvin and Linzen (2018). The BERT and M&L numbers are *not* directly comparable, as the experimental setup differs in many ways.

RNN and Transformers (non-recurrent self attention) learn to predict upcoming words well, on average, but in sintactically complex sentences.

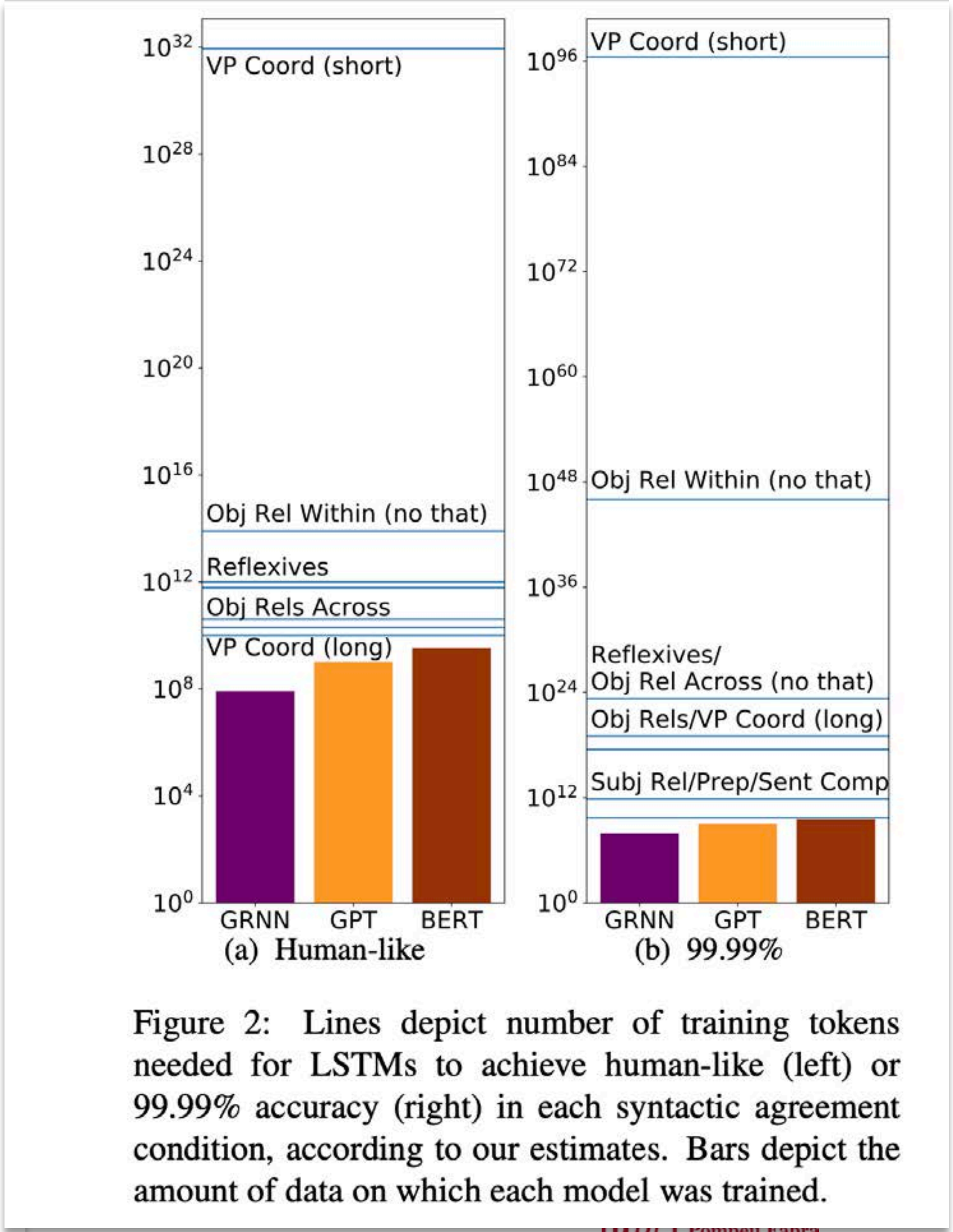
Schijndel, M.; Mueller, A. and Linzen, T. (2019) Quantity doesn't buy quality syntax with neural language models. EMNLP 2019



(e) VP Coordination (Short)

(f) VP Coordination (Long)

Figure 1: LSTM agreement performance in several syntactic constructions. The solid horizontal line indicates chance performance. The dashed lines show the performance of GPT and BERT as reported by Wolf (2019), the performance of humans as reported by Marvin and Linzen (2018), and the performance of GRNN. Error bars reflect standard deviation across the five models in each category.



(a) Human-like

(b) 99.99%

Figure 2: Lines depict number of training tokens needed for LSTMs to achieve human-like (left) or 99.99% accuracy (right) in each syntactic agreement condition, according to our estimates. Bars depict the amount of data on which each model was trained.

Natural Language Understanding tasks: NL Inference

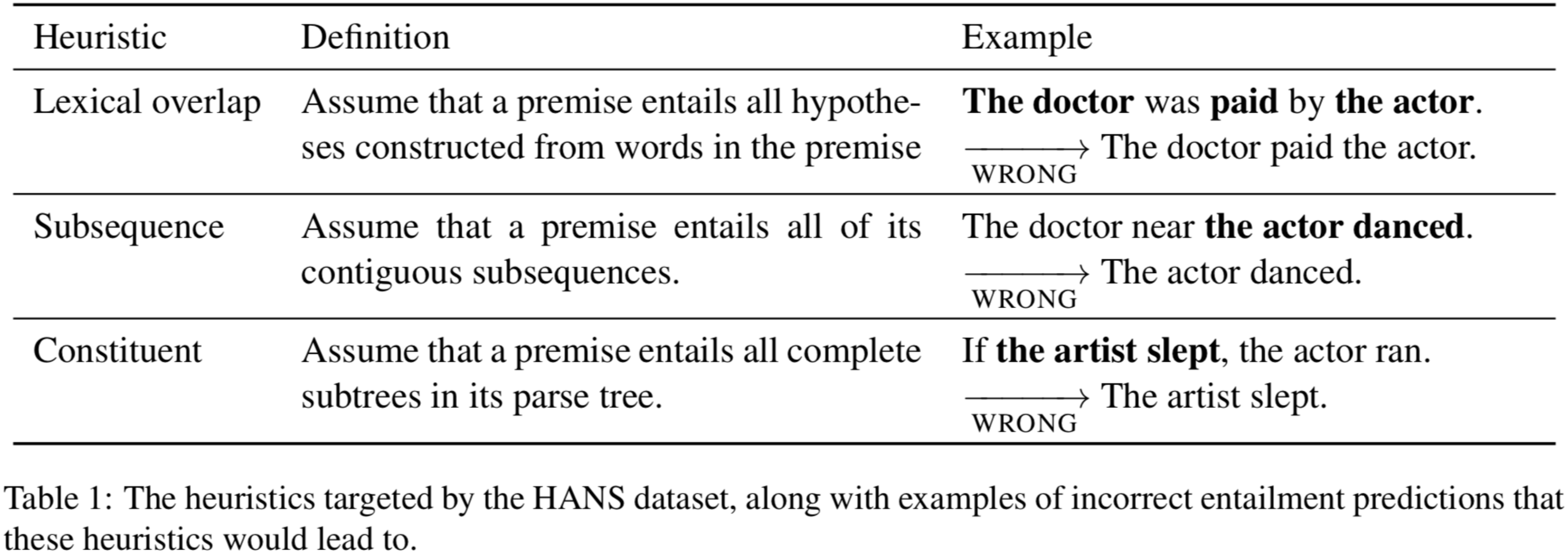
- Wang et al. (2018) evaluate NLI with similar sentences that have different inferences (GLUE platform).

Data: Developed “Test-suites”: 550 sentence pairs (reddit, facebook, etc.) with different linguistic phenomena: universal quantification, negation and double negation, correference, etc.

- McCoy et al. (2019) built datasets to evaluate whether a NLI system is learning superficial heuristics, for instance the number of words that are repeated in the premise and the hypothesis.

Data: HANS, a dataset to check specific cases of possible heuristics. 10.000 samples for each heuristic.

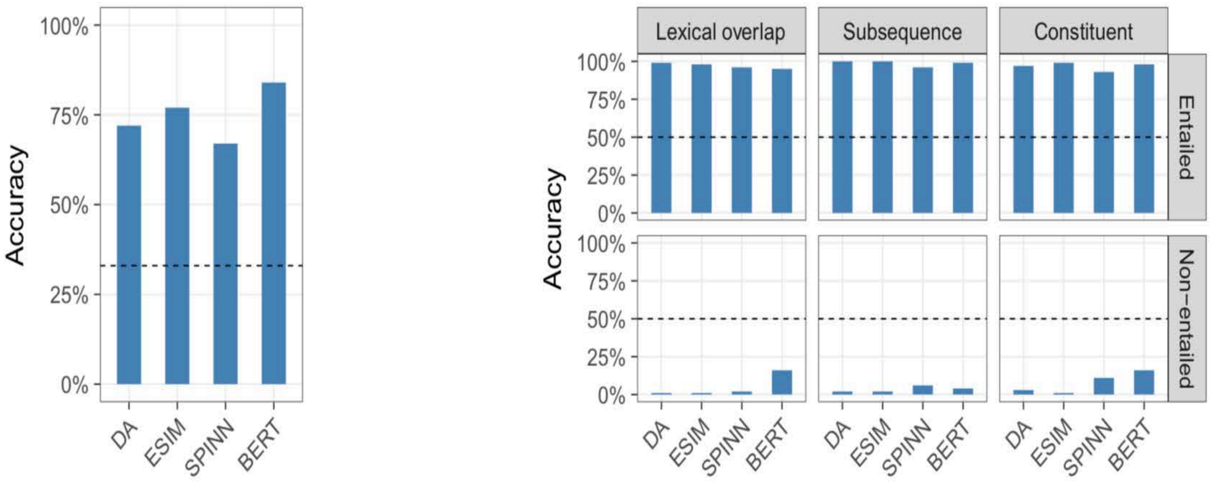
Examples from McCoy et al. (2019)



Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

Table 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

McCoy et al. (2019) Results



(a)

(b)

Figure 1: (a) Accuracy on the MNLI test set. (b) Accuracies on the HANS evaluation set, which has six subcomponents, each defined by its correct label and the heuristic it addresses. Dashed lines show chance performance. All models behaved as we would expect them to if they had adopted the heuristics targeted by HANS. That is, they nearly always predicted *entailment* for the examples in HANS, leading to near-perfect accuracy when the true label is *entailment*, and near-zero accuracy when the true label is *non-entailment*. Exact results are in Appendix G.

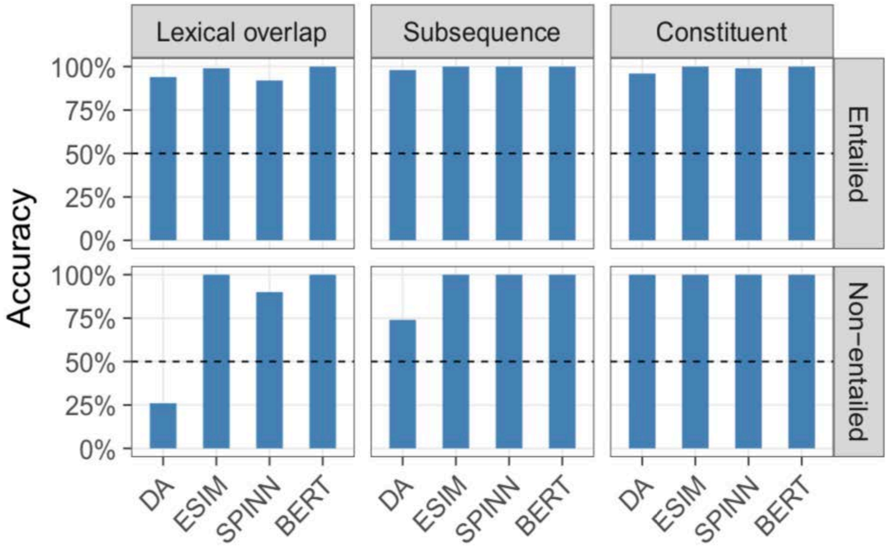
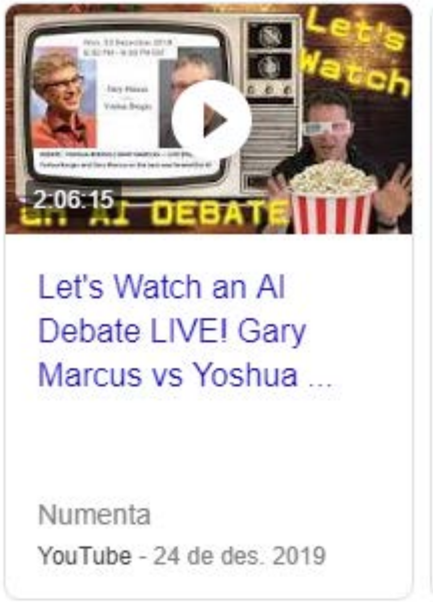


Figure 2: HANS accuracies for models trained on MNLI plus examples of all 30 categories in HANS.

Conclusions about DL capacities?

- **Not yet … sorry!**
- **Thinking … has BERT learnt the rule?**
- **Better wait for Bengio's "concious module?"**



Let's Watch an AI Debate LIVE! Gary Marcus vs Yoshua ...

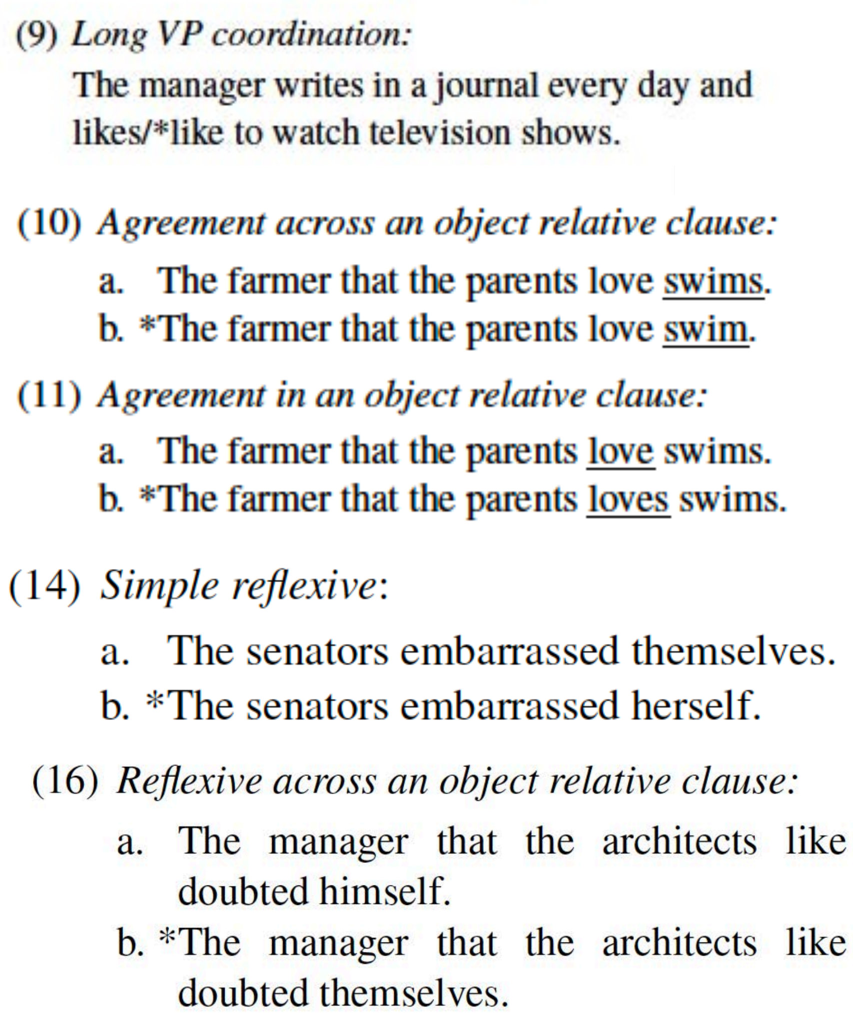
Numenta

YouTube - 24 de des. 2019

G. Marcus' example about rule learning

1010 => 1010
1110 => 1110
1000 => 1000
1001 => 1000

I’m surprised!



(9) *Long VP coordination:*

The manager writes in a journal every day and likes/*like to watch television shows.

(10) *Agreement across an object relative clause:*

a. The farmer that the parents love swims.

b. *The farmer that the parents love swim.

(11) *Agreement in an object relative clause:*

a. The farmer that the parents love swims.

b. *The farmer that the parents loves swims.

(14) *Simple reflexive*:

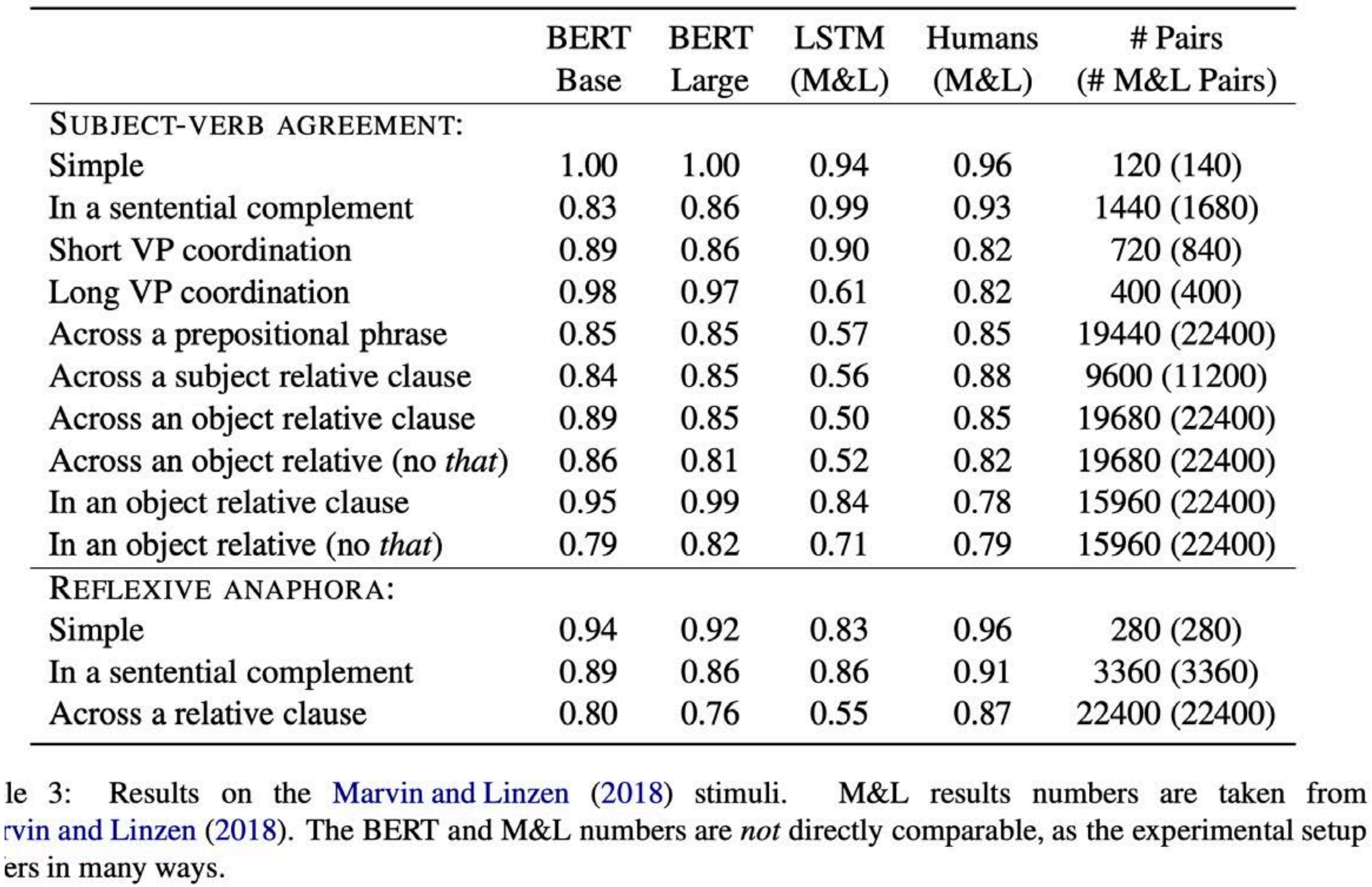
a. The senators embarrassed themselves.

b. *The senators embarrassed herself.

(16) *Reflexive across an object relative clause:*

a. The manager that the architects like doubted himself.

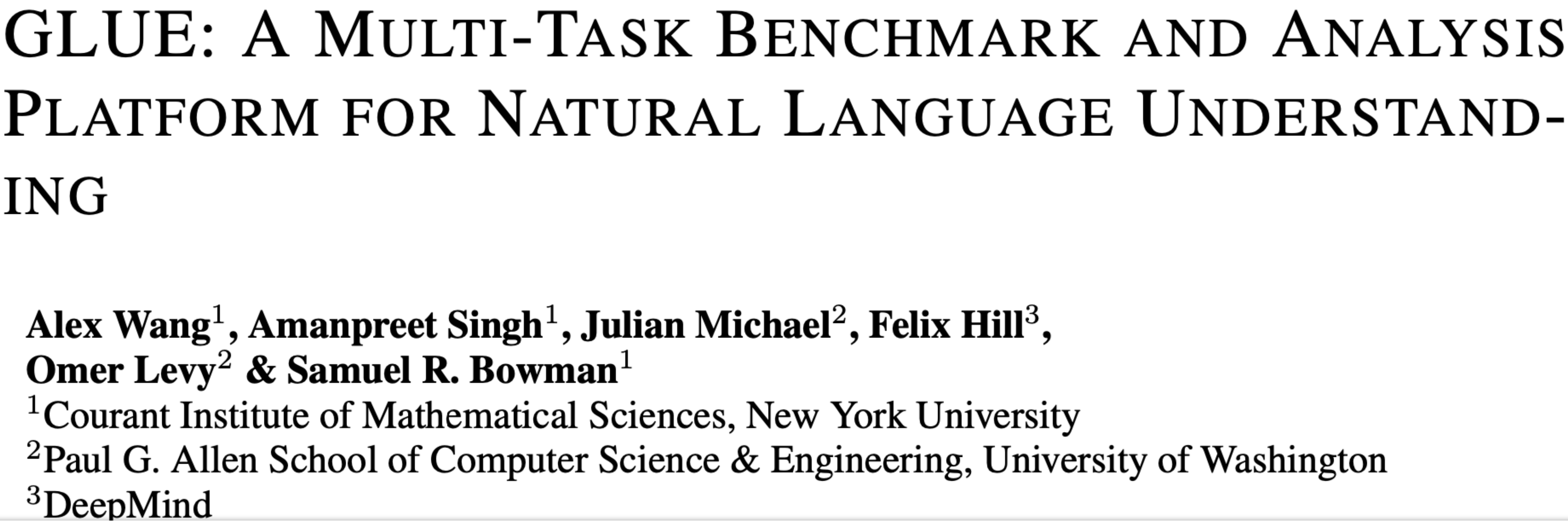
b. *The manager that the architects like doubted themselves.



	BERT Base	BERT Large	LSTM (M&L)	Humans (M&L)	# Pairs (# M&L Pairs)
SUBJECT-VERB AGREEMENT:					
Simple	1.00	1.00	0.94	0.96	120 (140)
In a sentential complement	0.83	0.86	0.99	0.93	1440 (1680)
Short VP coordination	0.89	0.86	0.90	0.82	720 (840)
Long VP coordination	0.98	0.97	0.61	0.82	400 (400)
Across a prepositional phrase	0.85	0.85	0.57	0.85	19440 (22400)
Across a subject relative clause	0.84	0.85	0.56	0.88	9600 (11200)
Across an object relative clause	0.89	0.85	0.50	0.85	19680 (22400)
Across an object relative (no *that*)	0.86	0.81	0.52	0.82	19680 (22400)
In an object relative clause	0.95	0.99	0.84	0.78	15960 (22400)
In an object relative (no *that*)	0.79	0.82	0.71	0.79	15960 (22400)
REFLEXIVE ANAPHORA:					
Simple	0.94	0.92	0.83	0.96	280 (280)
In a sentential complement	0.89	0.86	0.86	0.91	3360 (3360)
Across a relative clause	0.80	0.76	0.55	0.87	22400 (22400)

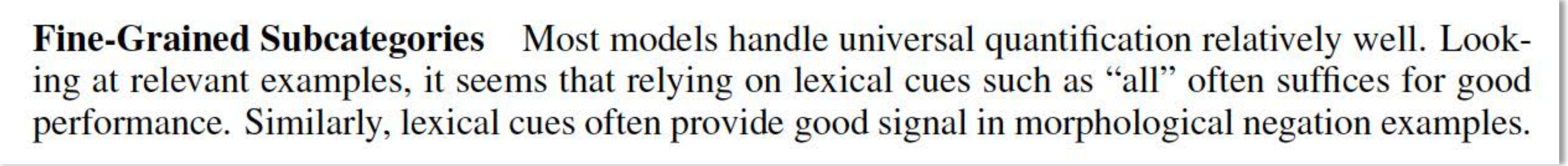
le 3: Results on the Marvin and Linzen (2018) stimuli. M&L results numbers are taken from rvin and Linzen (2018). The BERT and M&L numbers are *not* directly comparable, as the experimental setup ers in many ways.

Linguistic view? The role of ‘stop words’



GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

**Alex Wang[1], Amanpreet Singh[1], Julian Michael[2], Felix Hill[3],
Omer Levy[2] & Samuel R. Bowman[1]**
[1]Courant Institute of Mathematical Sciences, New York University
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington
[3]DeepMind



Fine-Grained Subcategories Most models handle universal quantification relatively well. Looking at relevant examples, it seems that relying on lexical cues such as “all” often suffices for good performance. Similarly, lexical cues often provide good signal in morphological negation examples.

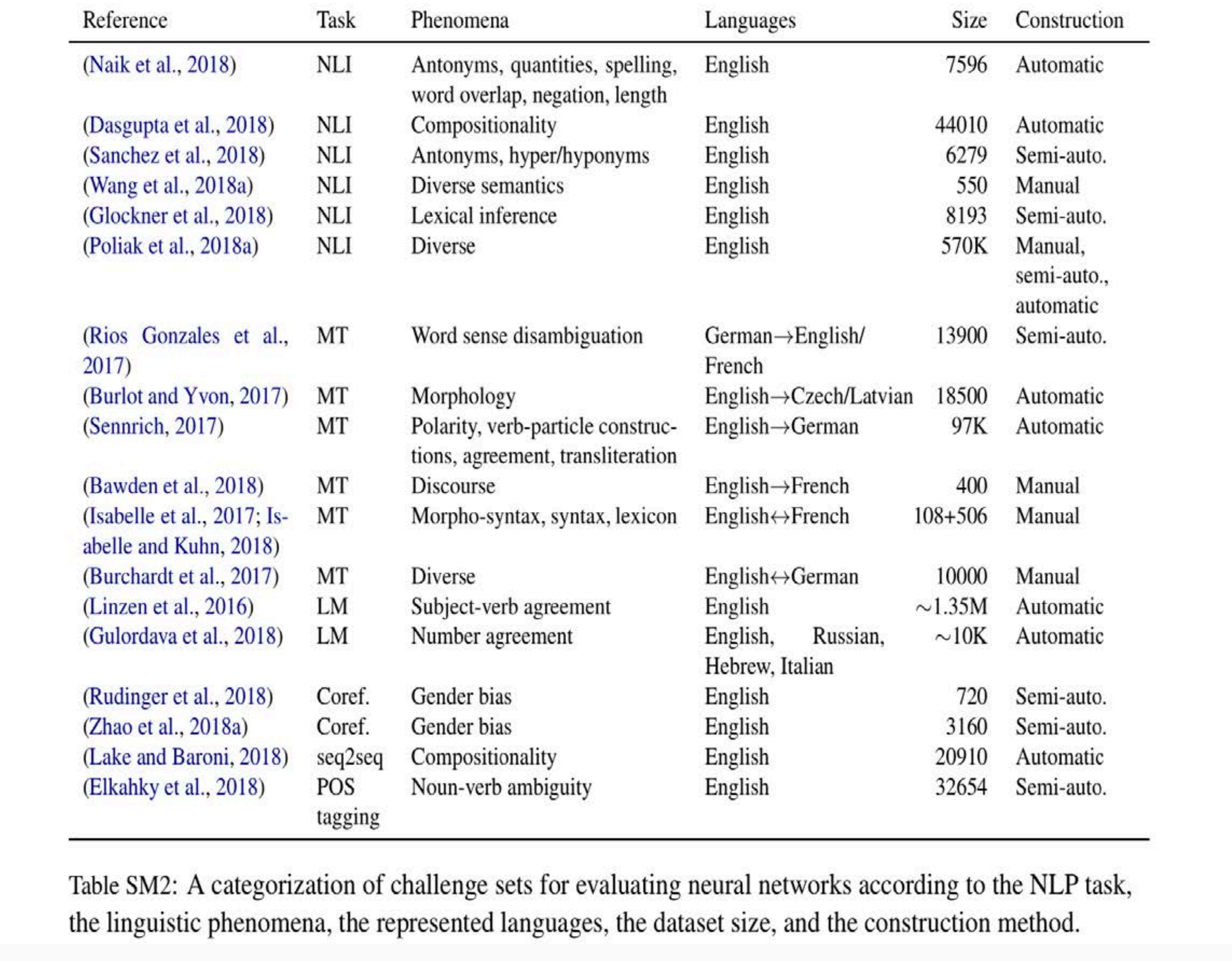
Some open questions …

- Traditional linguistically motivated questions about the expected capabilities of a linguistic processor could be expressed as:
- Does the method capture relations between words, that is, do the method capture linguistic structures? How different are the representations of the following sentences?
  - The dog chased the cat
  - The dog that we saw yesterday chased the cat
  - The dog chased the cat that we saw yesterday
- Does the method capture phrase and sentence meaning relations beyond lexical similarity? Are similarities between active, passive, interrogative or cleft sentences represented consistently?
  - The dog chased the cat
  - The cat was chased by the dog
  - Did the dog chased the cat?
  - It was the dog that chased the cat!
  - It was the cat that was chased by the dog!
- Does the method represent differently well-formed and ill-formed sentences?
  - The girl walked along the path
  - *The girl walking along the path
  - The girl walking along the path was my daughter

But good news for LR’s: datasets wanted! More languages!!!

- Datasets (LRs) to evaluate (validate) linguistic capacities of DL systems in different languages
- Datasets (LRs) that must be annotated accurately to provide insights about different linguistic phenomena
- Datasets (LRs) to train systems, in particular adversarial data to bias or unbias and improve results
- …

Data sets survey (Belinkov and Glass, 2019)



Reference	Task	Phenomena	Languages	Size	Construction
(Naik et al., 2018)	NLI	Antonyms, quantities, spelling, word overlap, negation, length	English	7596	Automatic
(Dasgupta et al., 2018)	NLI	Compositionality	English	44010	Automatic
(Sanchez et al., 2018)	NLI	Antonyms, hyper/hyponyms	English	6279	Semi-auto.
(Wang et al., 2018a)	NLI	Diverse semantics	English	550	Manual
(Glockner et al., 2018)	NLI	Lexical inference	English	8193	Semi-auto.
(Poliak et al., 2018a)	NLI	Diverse	English	570K	Manual, semi-auto., automatic
(Rios Gonzales et al., 2017)	MT	Word sense disambiguation	German→English/ French	13900	Semi-auto.
(Burlot and Yvon, 2017)	MT	Morphology	English→Czech/Latvian	18500	Automatic
(Sennrich, 2017)	MT	Polarity, verb-particle constructions, agreement, transliteration	English→German	97K	Automatic
(Bawden et al., 2018)	MT	Discourse	English→French	400	Manual
(Isabelle et al., 2017; Isabelle and Kuhn, 2018)	MT	Morpho-syntax, syntax, lexicon	English↔French	108+506	Manual
(Burchardt et al., 2017)	MT	Diverse	English↔German	10000	Manual
(Linzen et al., 2016)	LM	Subject-verb agreement	English	∼1.35M	Automatic
(Gulordava et al., 2018)	LM	Number agreement	English, Russian, Hebrew, Italian	∼10K	Automatic
(Rudinger et al., 2018)	Coref.	Gender bias	English	720	Semi-auto.
(Zhao et al., 2018a)	Coref.	Gender bias	English	3160	Semi-auto.
(Lake and Baroni, 2018)	seq2seq	Compositionality	English	20910	Automatic
(Elkahky et al., 2018)	POS tagging	Noun-verb ambiguity	English	32654	Semi-auto.

Table SM2: A categorization of challenge sets for evaluating neural networks according to the NLP task, the linguistic phenomena, the represented languages, the dataset size, and the construction method.

Research lines!!

- Assess the capacities of DL systems to reduce the dependency of tones of data,
- Designing linguistic tasks and linguistic insights (like the role of grammatical words) and selecting/building datasets for specific experiments.
- How quickly producing focussed evaluation datasets?
- Can synthetic data be used for better training? Unbiass …

How to generate synthetic data?

- Kuhnle y Copestake (2018) used a large symbolic computational grammar (ERG) to generate evaluation data
- Ribeiro et al. (2018) used paraphrases generation techniques with NMT systems in order to check hypersensibility: all the paraphrases must deliver the same results.

References

- Yossi Adi, Elnat Kermany, Yonatan Belinkov, Ofer Lavi, Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. Proceedings of the ICLR 2017.
- Yonatan Belinkov, James Glass. Analysis Methods in Neural Language Processing: A Survey.Transactions of the Association for Computational Linguistics (TACL) 2019
- Coenen et al. (2019). Visualizing the Geometry of BERT. https://arxiv.org/pdf/1906.02715.pdf
- Alexis Conneau and Douwe Kiela, SentEval: An Evaluation Toolkit for Universal Sentence Representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC2018, Miyazaki, Japan, May 7-12, 2018.
- Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, Marco Baroni. «What you can cram into a single \$\&!#* vector: Probing sentence embeddings for linguistic properties.» *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 2018.
- Robin Cooper , Dick Crouch , Jan Van Eijck , Chris Fox , Josef Van Genabith , Jan Jaspars , Hans Kamp , David Milward , Manfred Pinkal , Massimo Poesio , Steve Pulman , Ted Briscoe , Holger Maier , Karsten Konrad. Using the framework. Fracas Consortium. 1996. https://nlp.stanford.edu/~wcmac/downloads/fracas.xml [Consultado el 24-10-2018]
- Czarnowska, P.; Ruder, S.; Grave, E.; C 2019otterell, R.; and Copestake, A. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In EMNLP 2019
- Goldberg, Y. Assessing BERT's Syntactic Abilities. https://arxiv.org/pdf/1901.05287.pdf
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, Marco Baroni. «Colorless Green Recurrent Networks Dream Hierarchically.» *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* Association for Computational Linguistics, 2018.
- Robin Jia y Percy Liang. Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328, 2017. Publicado en Proceedings of the EMNLP 2018.
- Ákos Kádár, Grzegorz Chrupała, Afra Alishahi. «Representation of linguistic form and function in recurrent neural networks» *Computational Linguistics, 43, 4*, 2017.
- Alexander Kuhnle and Ann Copestake. "Deep learning evaluation using deep linguistic processing." 2018. https://arxiv.org/pdf/1706.01322 . Consultado el 24-09-2018]
- Jiwei Li, Will Monroe and Dan Jurafsky. «Understanding Neural Networks through Representation Erasure.» http://arxiv.org/abs/1612.08220. 2017. [Consultado el 24-09-2018]
- Jiwei Li, Xinlei Chen, Eduard Hovy and Dan Jurafsky. "Visualizing and Understanding Neural Models in NLP." *Proceedings of NAACL-HLT 2016.* San Diego, California: Association for Computational Linguistics, 2016.
- Tal Linzen, Emmanuel Dupoux and Yoav Goldberg. «Assessing the ability of LSTMs to learn Syntax-Sensitive Dependencies.» *Transactions of the Association for Computational Linguistics, vol. 4*, 2016.
- Nelson F. Liu, Omer Levy, Roy Schwartz, Chenhao Tan, Noah A. Smith, LSTMs Exploit Linguistic Attributes of Data, Proceedings of the ACL 2018 RepL4NLP workshop, 2018. http://arxiv.org/abs/1805.11653.
- Rebecca Marvin and Tal Linzen, Targeted Syntactic Evaluation of Language Models, EMNLP, 2018.
- Marco Tulio Ribeiro Marco Tulio; Sameer Singh,; Carlos Guestrin,. Semantically Equivalent Adversarial Rules for Debugging NLP Models. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
- R. Thomas McCoy, Ellie Pavlick and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel R. Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." CoRR abs/1804.07461. 2018.

ありがとう
Tak
ďakujem
Sağ ol
Salamat
спасибо
Grazie
благодаря
Dankie
þakka þér
Danke
ačiū
paldies
matur nuwun
děkuji
Dziękuję
aitäh
obrigado
Thank you!
Tack
Grazzi
謝謝
ju faleminderit
Mauruuru
Merci
Gracias
Dankon
eskerrik asko
kiitos
terima kasih
hvala
Na gode
gràcies
Grazas
köszönöm