Deep Learning:
NLP tasks, Benchmarking Datasets & Evaluation

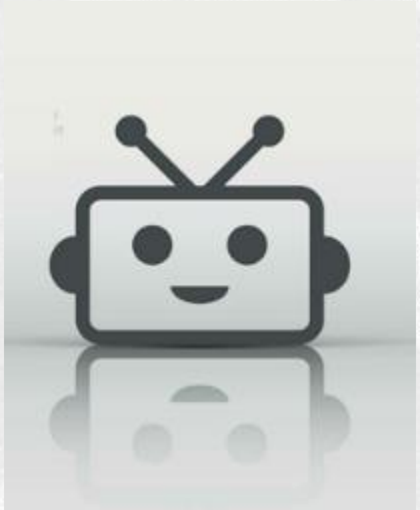
Roberto Basili,
Deep Learning 2022/2023

Outline

- NLP tasks
  - Classification tasks
  - Textual Inference tasks
  - Sentiment Analysis and Social Media analytics
- GLUE
  - Datasets
  - Benchmarks
- SQUAD
- Multimodal Tasks

NLP tasks

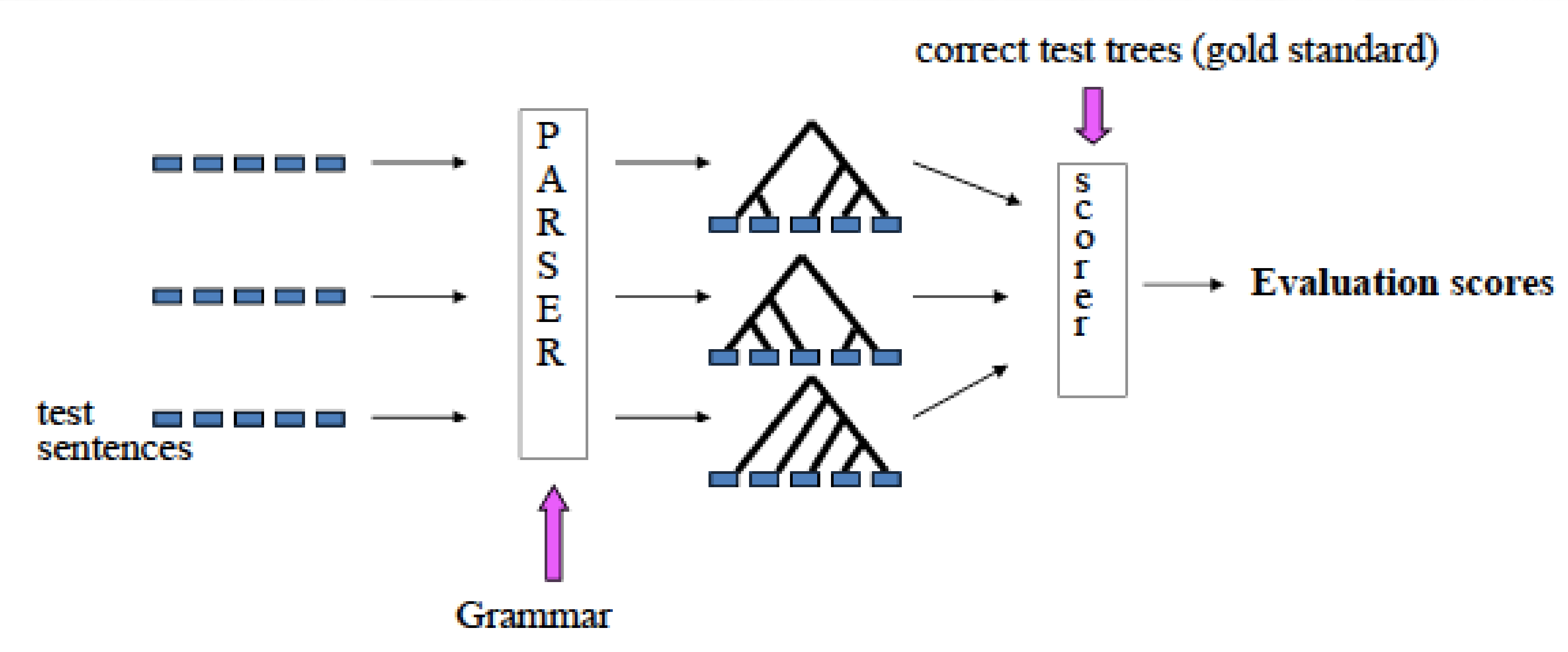
- Language Processing models, such as Deep Learning or Large Language Models, makes sense only in view of a number of tasks where they must show performances in line with human "natural" behaviours



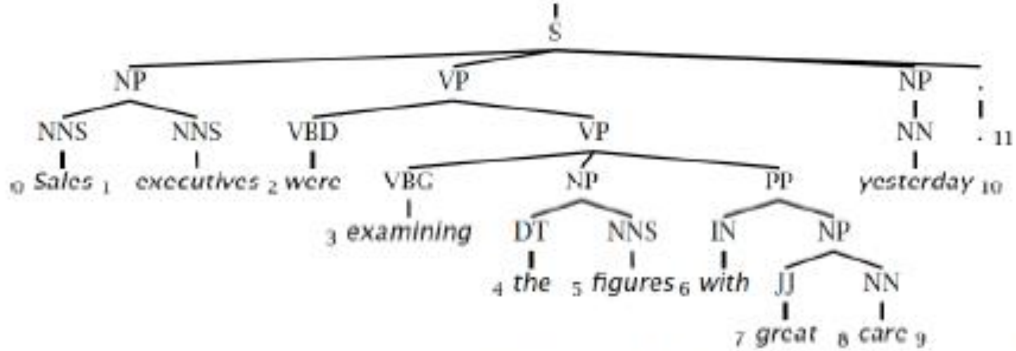
Traditional NLP tasks

- Parsing: The task of mapping one sentence into its grammatically explicit counterpart, based on
  - Trees, e.g. Constituent-based representations for CFGs
  - Graphs, e.g. UD in Dependency graphs
  - Relational (i.e. tabular) forms
- Metrics:
  - Accuracy
  - Bracketed Accuracy

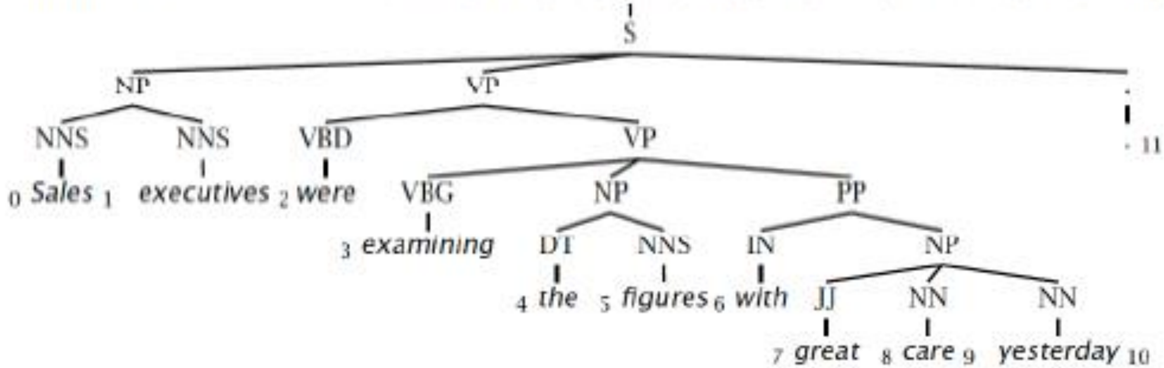
Parsing: Evaluation



Gold standard brackets: **S-(0:11)**, **NP-(0:2)**, VP-(2:9), VP-(3:9), **NP-(4:6)**, PP-(6-9), NP-(7,9), NP-(9:10)

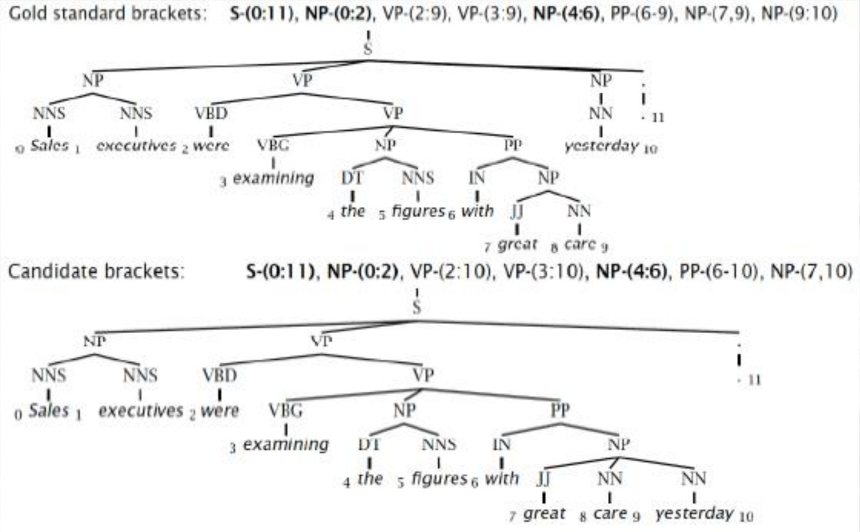


Candidate brackets: **S-(0:11)**, **NP-(0:2)**, VP-(2:10), VP-(3:10), **NP-(4:6)**, PP-(6-10), NP-(7,10)



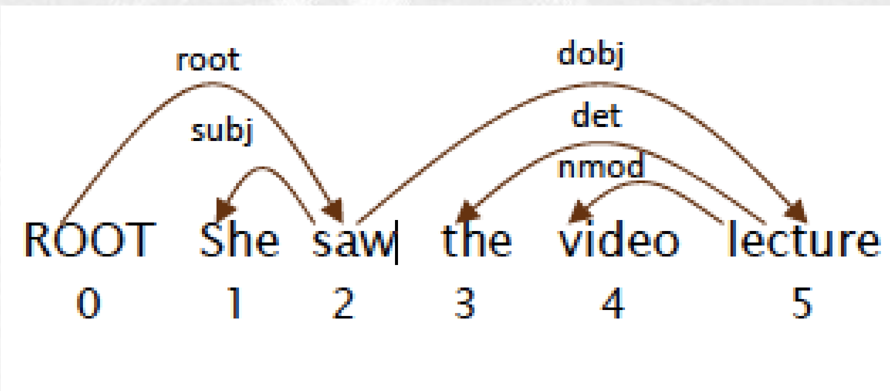
Labeled P/R/F

- Gold brackets:
  - **S(0:11), NP(0:2)**, VP(2:9), VP(3:9), **NP (4:6)**, PP (6:9), NP (7,9), NP (9:10).
- Candidate brackets:
  - **S(0:11), NP(0:2)**, VP(2:10), VP(3:10) **NP(4:6)**, PP (6:10), NP (7:10)



- Parseval measures
  - Labeled Precision: P=3/7=42.9%
  - Labeled Recall: R=3/8=37.5%
  - F=40.0%

Parsing:
Dependency formalisms



Gold			
1	She	2	subj
2	saw	0	root
3	the	5	det
4	video	5	nmod
5	lecture	2	dobj

Parsed			
1	She	2	subj
2	saw	0	root
3	the	4	det
4	video	5	vmod
5	lecture	2	iobj

- Measures
  - Unlabeled Attachment Score (UAS)
  - Labeled Attachment Score (LAS)
  - Label Accuracy (LA)

For the sentence:

She saw the video lecture

- UAS: 4/5 = 80%
- LAS: 2/5 = 40%
- LA: 3/5 = 60%

Link Grammar		Conexor FDG	
Name	*Description*	*Name*	*Description*
`Bs`	Singular external object of relative clause	`cc`	Coordination
`Ds`	Singular determiner	`det`	Determiner
`Js`	Singular object of a preposition	`ins`	*<not documented>*
`MVp`	Verb-modifying preposition	`main`	Main element
`O^`	Object	`mod`	General post-modifier
`R`	Relative clause	`obj`	Object
`RS`	Part of subject-type relative clause	`pcomp`	Prepositional complement
`Ss`	Singular subject	`subj`	Subject
`Wd`	Declarative sentence		

Table 1: Some of the dependency types used by Link Grammar and Conexor FDG.

Relation	*Description*
`SUBJ(head, dependent, initial_gr)`	Subject
`OBJ(head, dependent)`	Object
`XCOMP(type, head, dependent)`	Clausal complement without an overt subject
`MOD(type, head, dependent)`	Modifier

Table 2: Grammatical relations used in the intrinsic evaluation.

Link Grammar		Conexor FDG	
Name	*Description*	*Name*	*Description*
`Bs`	Singular external object of relative clause	`cc`	Coordination
`Ds`	Singular determiner	`det`	Determiner
`Js`	Singular object of a preposition	`ins`	*<not documented>*
`MVp`	Verb-modifying preposition	`main`	Main element
`O^`	Object	`mod`	General post-modifier
`R`	Relative clause	`obj`	Object
`RS`	Part of subject-type relative clause	`pcomp`	Prepositional complement
`Ss`	Singular subject	`subj`	Subject
`Wd`	Declarative sentence		

Table 1: Some of the dependency types used by Link Grammar and Conexor FDG.

Relation	*Description*
`SUBJ(head, dependent, initial_gr)`	Subject
`OBJ(head, dependent)`	Object
`XCOMP(type, head, dependent)`	Clausal complement without an overt subject
`MOD(type, head, dependent)`	Modifier

Table 2: Grammatical relations used in the intrinsic evaluation.

Traditional NLP tasks: NERC

- Recognition of specific types of entities in free text
  - News Domain: people, locations, dates, organizations,
  - Medical Domain: names of Body Parts, Chemicals, Pharmaceuticals, Dosages, …
  - Banking: Organisations, Legal Entities, Process types, Organizational Units, Account details, Dates, …
  - …

Traditional NLP Tasks: Document Classification

- Given a text T (a document, a title or a paragraph)
- Determine the (topical, editorial, pragmatic, …) category C that characterize T
  - Multilabel, if more than one category can be assigned to T

Natural Language Inference: Textual Entailment

- Given
  - a text (usually referred to as a premise P
  - a sentence H (hypothesis)
- FIND: the logical relationship between H and P:
  - Entailment
  - Independence
  - Contradiction

P[a]	A senior is waiting at the window of a restaurant that serves sandwiches.	Relationship
H[b]	A person waits to be served his food.	Entailment
	A man is looking to order a grilled cheese sandwich.	Neutral
	A man is waiting in line for the bus.	Contradiction

[a]**P**, Premise.
[b]**H**, Hypothesis.

Natural Language Inference

GLUE | MultiNLI | SNLI | QNLI | MRPC | WinoGrande

XNLI | SICK | ANLI | PAWS

See all 77 natural language inference datasets

Subtasks

Answer Generation | Visual Entailment | Cross-Lingual Natural Language Inference

Sentiment Analysis

- Recognition of the subjective position of the speaker/writer about some FOCUS OF THE DISCOURSE
- Different tasks
  - Subjectivity Recognition (*John* *is* *ugly* / *tall*)
  - Polarity Detection (*John* *is* *fantastic* / *terrible*)
  - Aspect-based classification
    - Recognition of different aspects of the judgment
    - *The tool* *is* *very fast* *but* *socially dangerous*
    - EFFICIENCY *VS.* *APPLICABILITY*
- Aimed at large scale text analysis for aggregate information

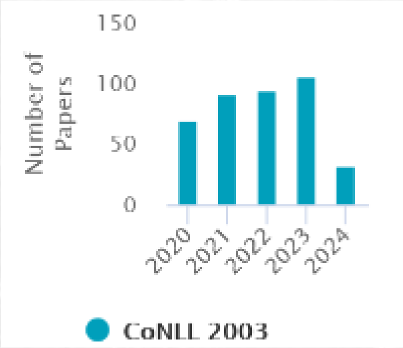
NLP tasks & Benchmarking

- The different tasks inspired the development of large-scale data sets as reference benchmarking resources able to
  - Focus on specific linguistic phenomena and models
  - Formally define the corresponding tasks
  - Develop training data
  - Define performance metrics for the tasks
  - Study the evolutionary impact of state-of-the-art methodologies in a competitive (and thus selective) setting
- Objective: Evolutionary Selection of Optimal models of the different application tasks

Datasets

- CoNLL 2003, NERC
- Groningen Meaning Bank (2018), Semantic Parsing
- **GLUE** (2019), a collection of datasets inspired by different tasks
- **Winogrande** (2019)
- **SQUAD** (2017), question answering
- DialoGLUE (2020), dialog
- **WikiSQL** (2018), Automatic SQL Code generation
- **WikiHow** (2018), Text Summarization

CoNLL2003: NERC

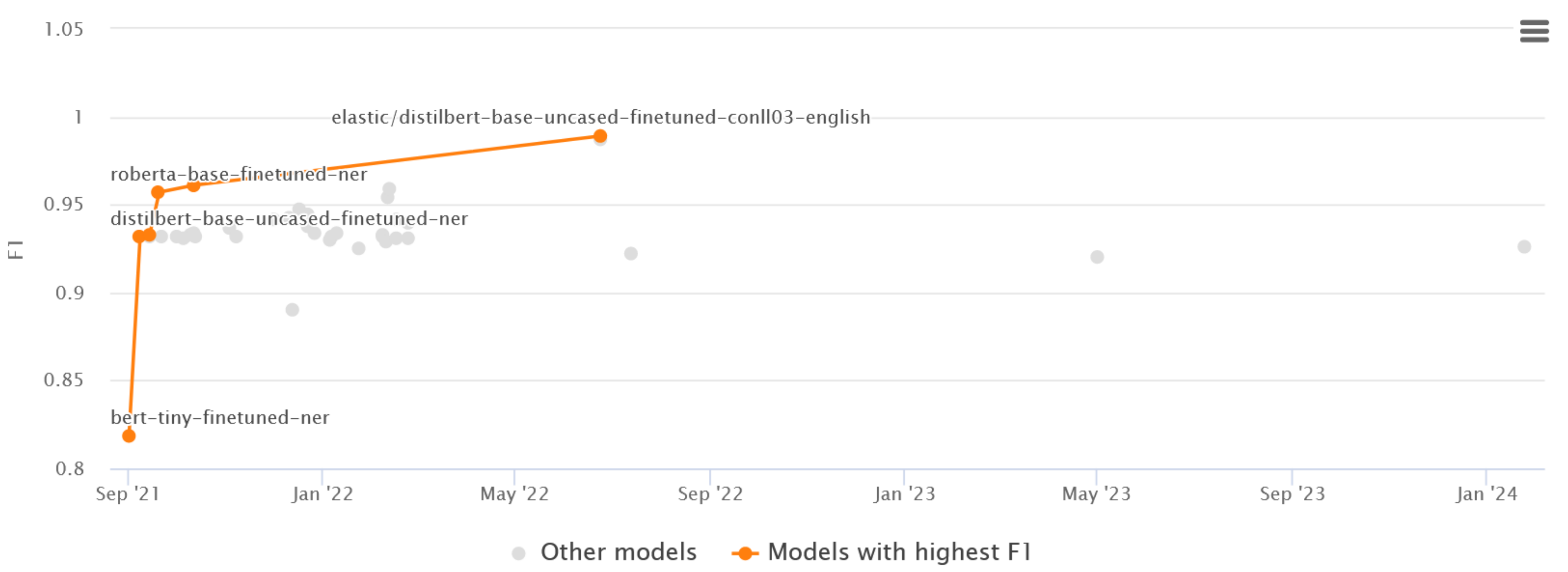


- Named entity recognition dataset released as a part of CoNLL-2003 shared task:
- Language-independent named entity recognition task.
- The data consists of 8 files covering 2 languages: English and German.
- For each of the languages there is a **training file**, a **development file**, a **test file** and a **large file with unannotated data**.

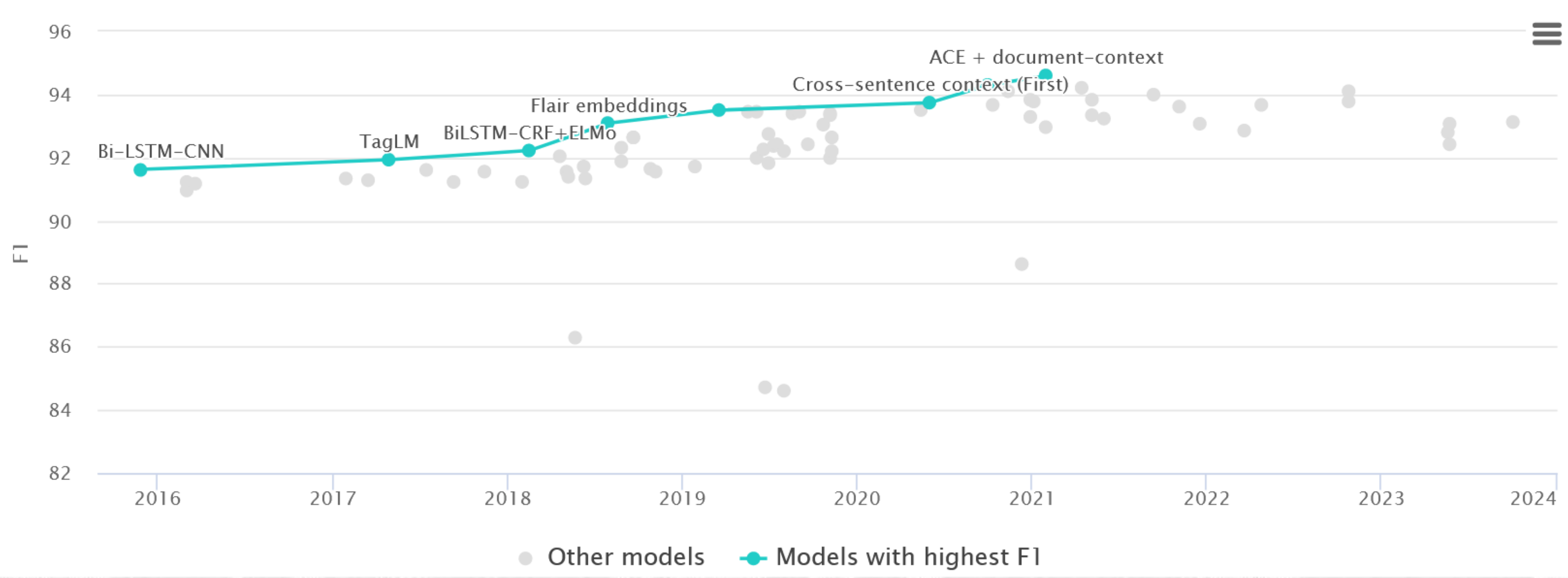
CoNLL 2003: English Data

English data	Articles	Sentences	Tokens	LOC	MISC	ORG	PER
Training set	946	14,987	203,621	7140	3438	6321	6600
Development set	216	3,466	51,362	1837	922	1341	1842
Test set	231	3,684	46,435	1668	702	1661	1617

CoNLL 2003: Results
(token level)



CoNLL 2023: F1

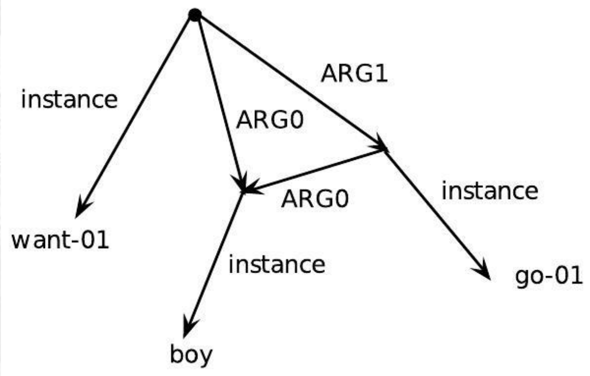


ACE model (2021): Wang et al., "**Automated Concatenation of Embeddings for Structured Prediction**", Proc. ACL 2021

Groningen Meaning Bank

- Groningen Meaning Bank is a semantic resource that anyone can edit and that integrates various semantic phenomena, including predicate-argument structure, scope, tense, thematic roles, animacy, pronouns, and rhetorical relations.

The boy wants to go



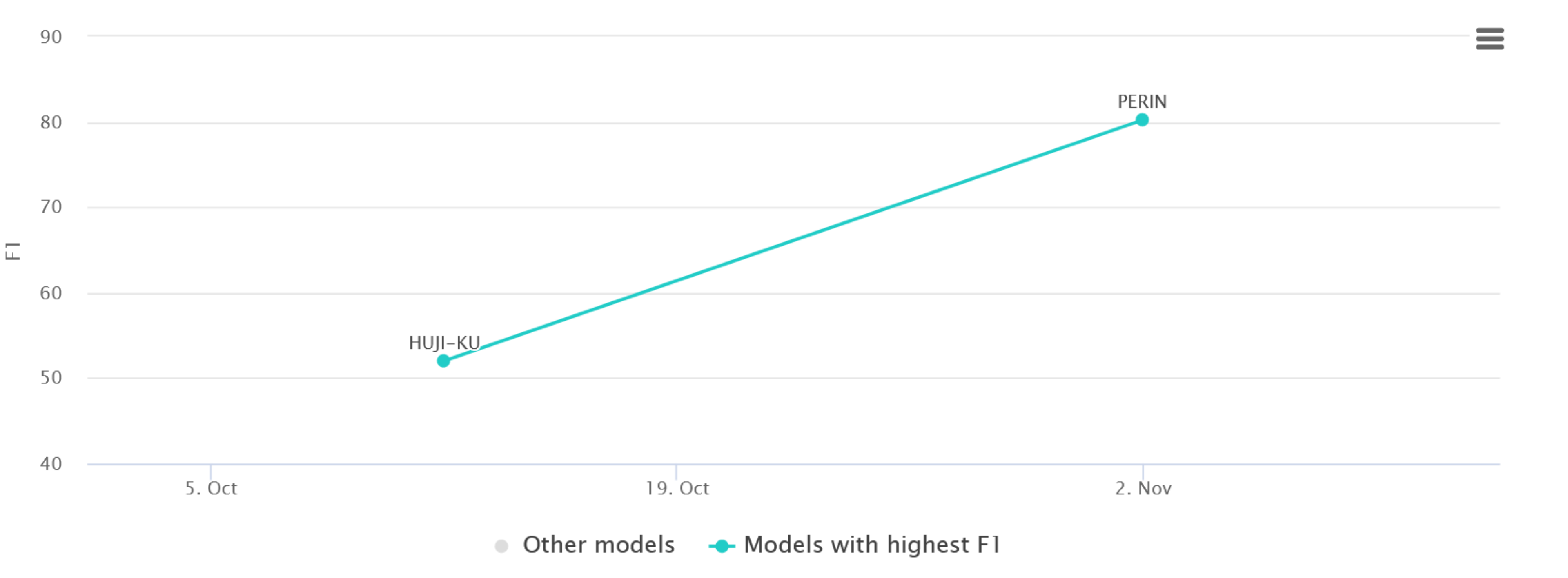
- Parallel effort: AMR Bank
  - Abstract Meaning Representation for Sembanking, Banarescu et al., 2021, 7th Linguistic Annotation Workshop, pages 178–186, Sofia, Bulgaria, August 8-9, 2013.

GMB: tagset

ANA	PRO	pronoun
	DEF	definite
	HAS	possessive
	REF	reflexive
	EMP	emphasizing
ACT	GRE	greeting
	ITJ	interjection
	HES	hesitation
	QUE	interrogative
ATT	QUA	quantity
	UOM	measurement
	IST	intersective
	REL	relation
	RLI	rel. inv. scope
	SST	subsective
	PRI	privative
	INT	intensifier
	SCO	score
LOG	ALT	alternative
	EXC	exclusive
	NIL	empty
	DIS	disjunct./exist.
	IMP	implication
	AND	conjunct./univ.
	BUT	contrast
COM	EQA	equative
	MOR	comparative pos.
	LES	comparative neg.
	TOP	pos. superlative
	BOT	neg. superlative
	ORD	ordinal
DEM	PRX	proximal
	MED	medial
	DST	distal
DIS	SUB	subordinate
	COO	coordinate
	APP	appositional
MOD	NOT	negation
	NEC	necessity
	POS	possibility
ENT	CON	concept
	ROL	role
NAM	GPE	geo-political ent.
	PER	person
	LOC	location
	ORG	organisation
	ART	artifact
	NAT	natural obj./phen.
	HAP	happening
	URL	url
EVE	EXS	untensed simple
	ENS	present simple
	EPS	past simple
	EFS	future simple
	EXG	untensed prog.
	ENG	present prog.
	EPG	past prog.
	EFG	future prog.
	EXT	untensed perfect
	ENT	present perfect
	EPT	past perfect
	EFT	future perfect
	ETG	perfect prog.
	ETV	perfect passive
	EXV	passive
TNS	NOW	present tense
	PST	past tense
	FUT	future tense
TIM	DOM	day of month
	YOC	year of century
	DOW	day of week
	MOY	month of year
	DEC	decade
	CLO	clocktime

Table 1: Semantic tags used in this paper.

AMR Bank Parsing Task



WikiSQL

- WikiSQL is a collection of questions, corresponding SQL queries, and SQL tables.
- A single example in WikiSQL contains a table, a SQL query, and the NL question corresponding to the SQL query.
- Namely, WikiSQL is the largest hand-annotated semantic parsing dataset to date - it is an order of magnitude larger than other datasets that have logical forms, either in terms of the number of examples or the number of tables.
- The queries in WikiSQL span over a large number of tables and hence presents an unique challenge: the model must be able to not only generalize to new queries, but to new table schema.

Table: CFLDraft

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Ottawa Renegades	L.P. Ladouceur	DT	California
30	Toronto Argonauts	Frank Hoffman	DL	York
...	...	...	...	...

Question:

How many CFL teams are from York College?

SQL:

```
SELECT COUNT CFL Team FROM
CFLDraft WHERE College = "York"
```

Result:

2

Figure 2: An example in WikiSQL. The inputs consist of a table and a question. The outputs consist of a ground truth SQL query and the corresponding result from execution.

WikiSQL: details

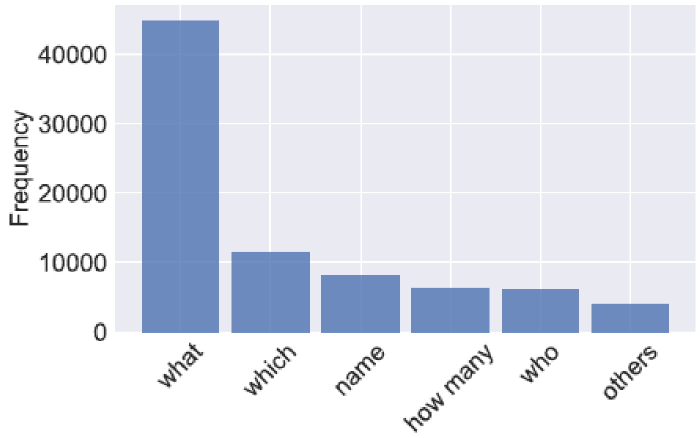


Figure 4: Distribution of questions in WikiSQL.

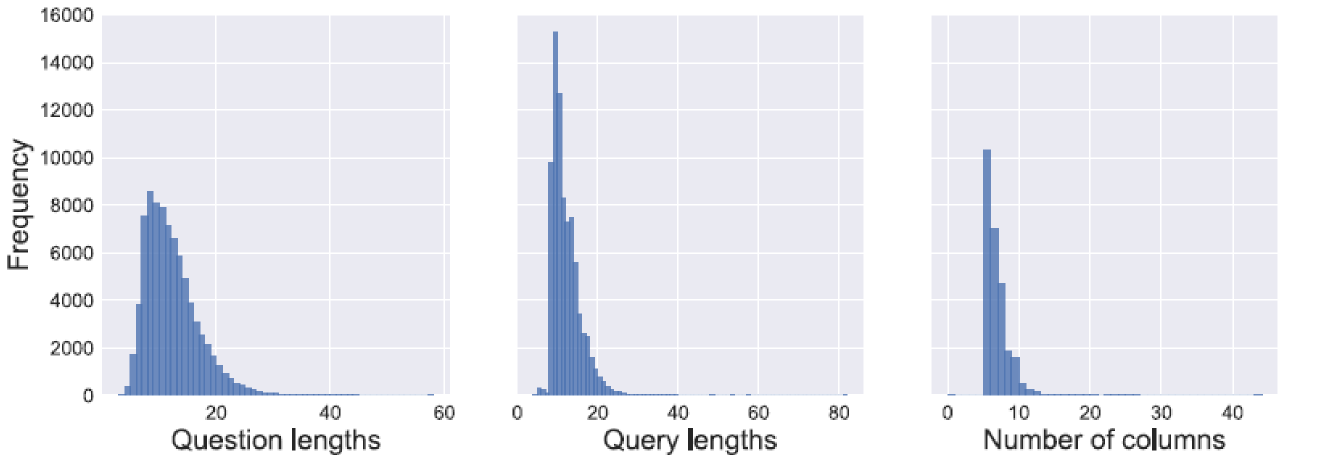


Figure 5: Distribution of table, question, query sizes in WikiSQL.

WikiSQL

100
80
60
40
20
EXECUTION ACCURACY
NL2SQL-RULE
TypeSQL+TC (Yu et al., 2018)+
STAMP+RL (Sun et al., 2018)+
PT-MAML (Huang et al., 2018)
Seq2SQL (Zhong et al., 2017)
Oct '17
Jan '18
Apr '18
Jul '18
Oct '18
Jan '19
Apr '19
Jul '19
Oct '19
Other models
Models with highest Execution Accuracy

WikiSQL:

Content Enhanced BERT-based Text-to-SQL Generation (Guo & Gao, 2019)

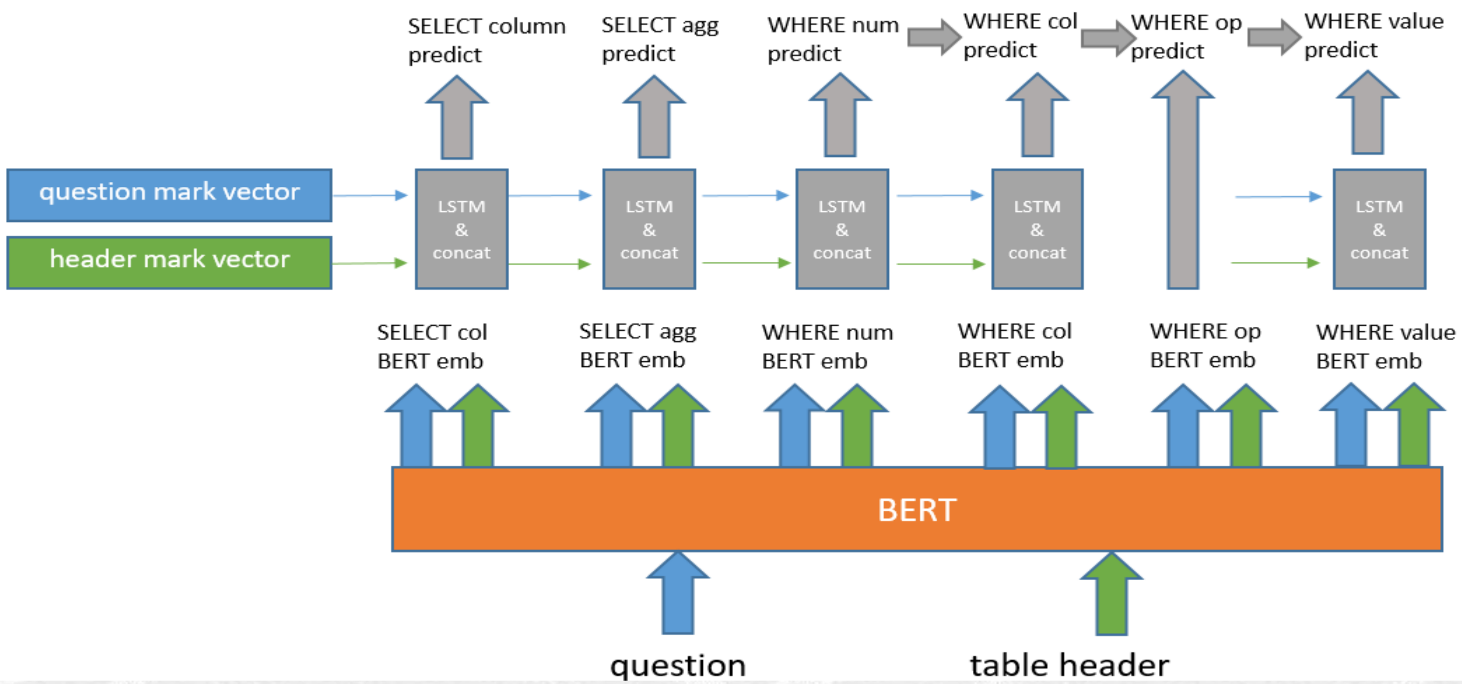
- This are example data istances

Table:							
Player	No.	Nationality	Position	Years in Toronto	School/Club Team		Question:
Antoniio Lang	21	United States	Guard-Forward	1999-2000	Duke		Who is the player that wears No 42
Voshon Lenard	2	United States	Guard	2002-2003	Minnesota		SQL:
Martin Lewis	32	United States	Guard-Forward	1996-1997	Butler CC		SELECT Player WHERE No. = 42
Brad Lohaus	33	United States	Guard-Forward	1996-1996	Iowa		Answer:
Art Long	42	United States	Guard-Forward	2002-2003	Cincinnati		Art Long

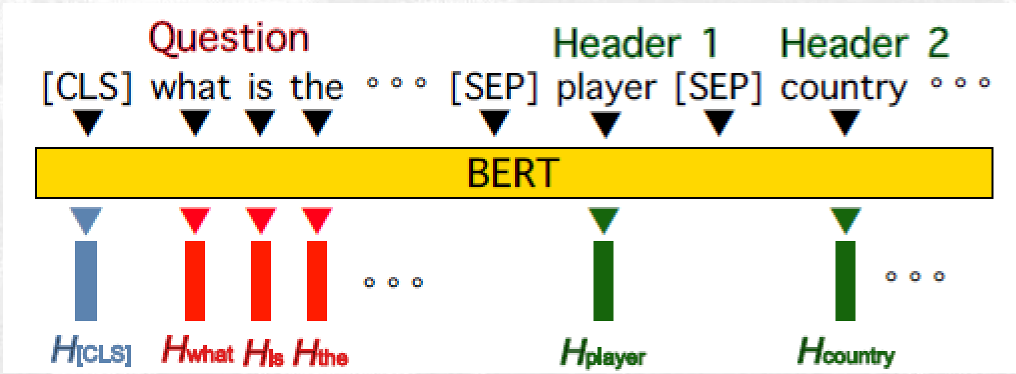
WikiSQL

Given the question tokens $w_1, w_2, ..., w_n$ and the table header $h_1, h_2, ..., h_n$, we follow the BERT convention and concat the question tokens and table header for BERT input. The detail encoding is below:

$$[CLS], w_1, w_2, ..., w_n, [SEP], h_1, [SEP], h_2, [SEP], ..., h_n, [SEP]$$



WikiSQL: NL2SQL



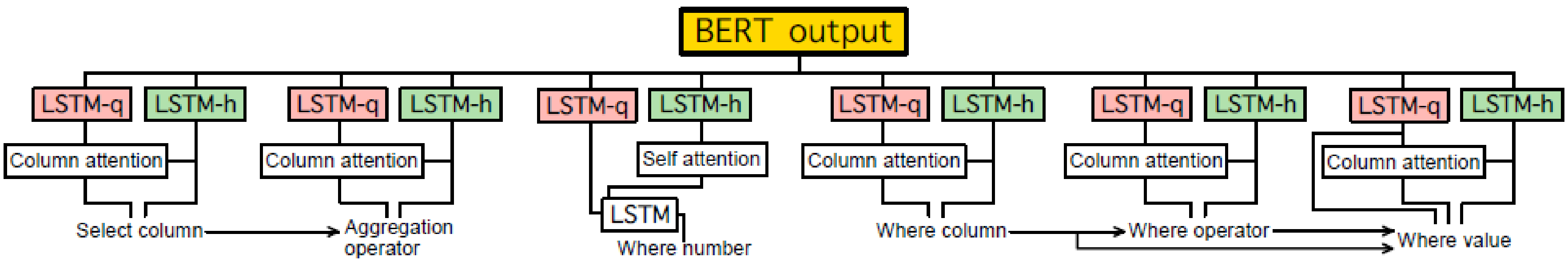
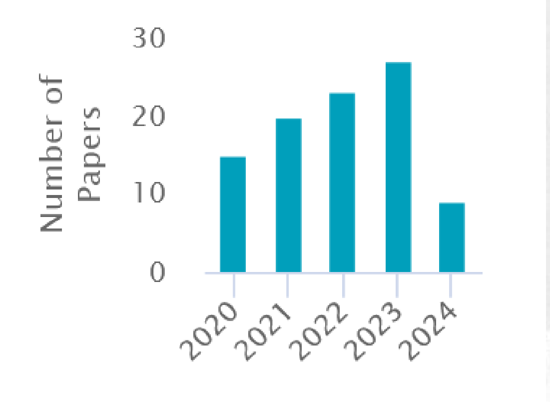


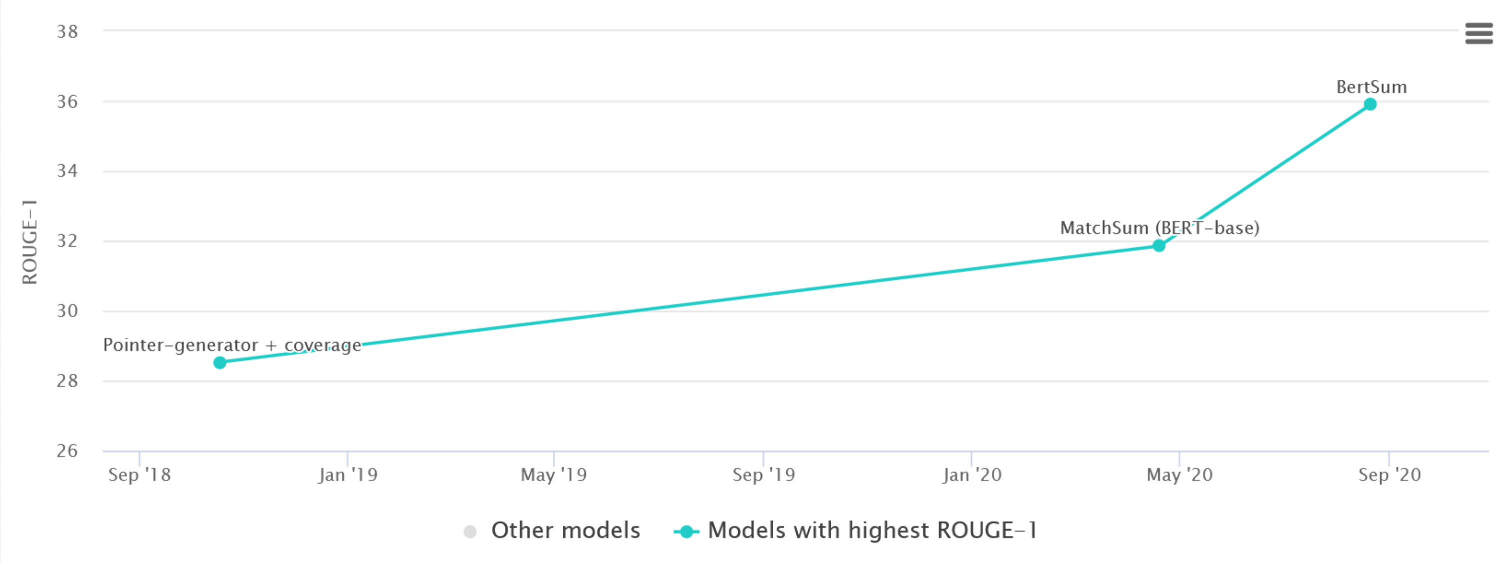
Figure 3: The illustration of NL2SQL LAYER (Section 3.2). The outputs from table-aware encoding layer are encoded again with LSTM−q (question encoder) and LSTM−h (header encoder).

Hwang, Wonseok, et al. "A comprehensive exploration on wikisql with table-aware word contextualization." *arXiv preprint arXiv:1902.01069* (2019). https://arxiv.org/abs/1902.01069

WikiHow



- **WikiHow** is a dataset of more than 230,000 **article and summary pairs** extracted and constructed from an online knowledge base written by different human authors.
- The articles span a wide range of topics and represent high diversity styles.



WIkiHow

- Wikihow dataset: a large scale text dataset containing over 200,000 single document summaries.
- Wikihow is a consolidated set of recent “How To” instructional texts compiled from wikihow.com, ranging from topics such as “How to deal with coronavirus anxiety” to “How to play Uno”.
- These articles vary in size and topic but are structured to instruct the user. The first sentences of each paragraph within the article are concatenated to form a summary.

WikiHow: examples

How to Help Save Rivers

Method 1 Reducing Your Water Usage

1 **Take quicker showers to conserve water.** One easy way to conserve water is to cut down on your shower time. Practice cutting your showers down to 10 minutes, then 7, then 5. Challenge yourself to take a shorter shower every day.

2 **Wait for a full load of clothing before running a washing machine.** Washing machines take up a lot of water and electricity, so running a cycle for a couple of articles of clothing is inefficient. Hold off on laundry until you can fill the machine.

3 **Turn off the water when you're not using it.** Avoid letting the water run while you're brushing your teeth or shaving. Keep your hoses and faucets turned off as much as possible. When you need them, use them sparingly.

...

Method 2 Using River-Friendly Products

1 **Select biodegradable cleaning products.** Any chemicals you use in your home end up back in the water supply. Choose natural soaps or create your own cleaning and disinfecting agents out of vinegar, baking soda, lemon juice, and other natural products. These products have far less of a negative impact if they reach a river.

2 **Choose recycled products instead of new ones.** New products take way more water to make than recycled products. Reuse what you already own when possible. If you need to buy something, opt for products made out of recycled paper or other reused material.

...

Article 1:
One easy way to conserve water is to cut down on your shower time. Practice cutting your showers down to 10 minutes, then 7, then 5. Challenge yourself to take a shorter shower every day. Washing machines take up a lot of water and electricity, so running a cycle for a couple of articles of clothing is inefficient. Hold off on laundry until you can fill the machine. Avoid letting the water run while you're brushing your teeth or shaving. Keep your hoses and faucets turned off as much as possible. When you need them, use them sparingly.
...

Summary 1:
Take quicker showers to conserve water. Wait for a full load of clothing before running a washing machine. Turn off the water when you're not using it.
...

Article 2:
Any chemicals you use in your home end up back in the water supply. Choose natural soaps or create your own cleaning and disinfecting agents out of vinegar, baking soda, lemon juice, and other natural products. These products have far less of a negative impact if they reach a river. New products take way more water to make than recycled products. Reuse what you already own when possible. If you need to buy something, opt for products made out of recycled paper or other reused material.
....

Summary 2:
Select biodegradable cleaning products. Choose recycled products instead of new ones.
...

Figure 2: An example of our new dataset: WikiHow summary dataset, which includes +200K summaries. The bold lines summarizing the paragraph (shown in red boxes) are extracted and form the summary. The detailed descriptions of each step (except the bold lines) will form the article. Note that the articles and the summaries are truncated and the presented texts are not in their actual lengths.

BERTSum (Liu&Lapata, 2019)

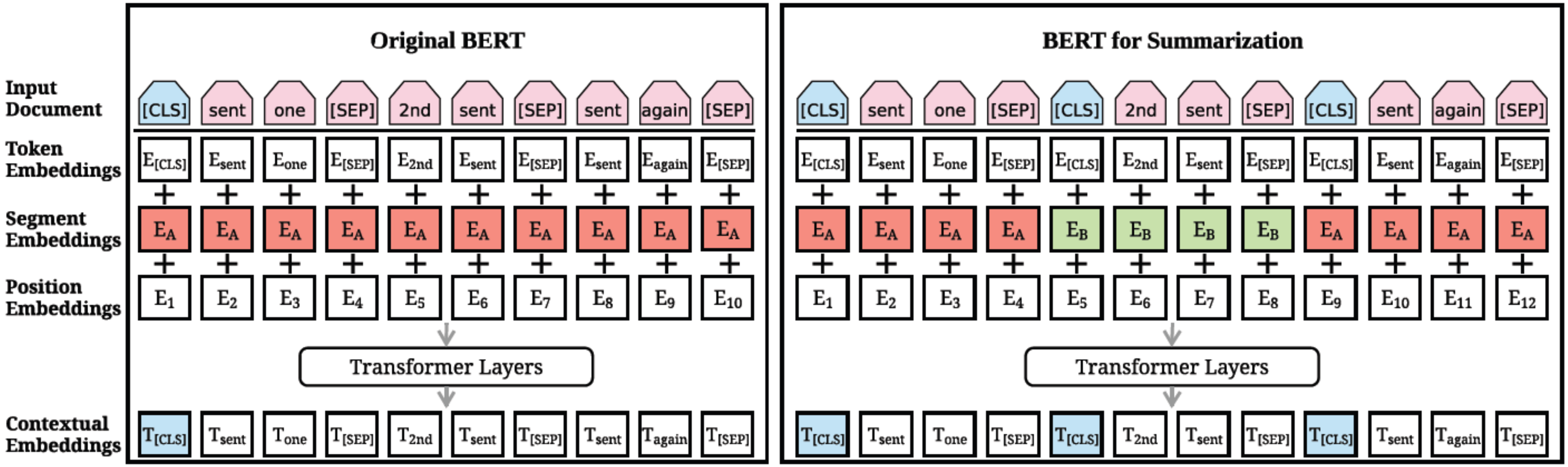
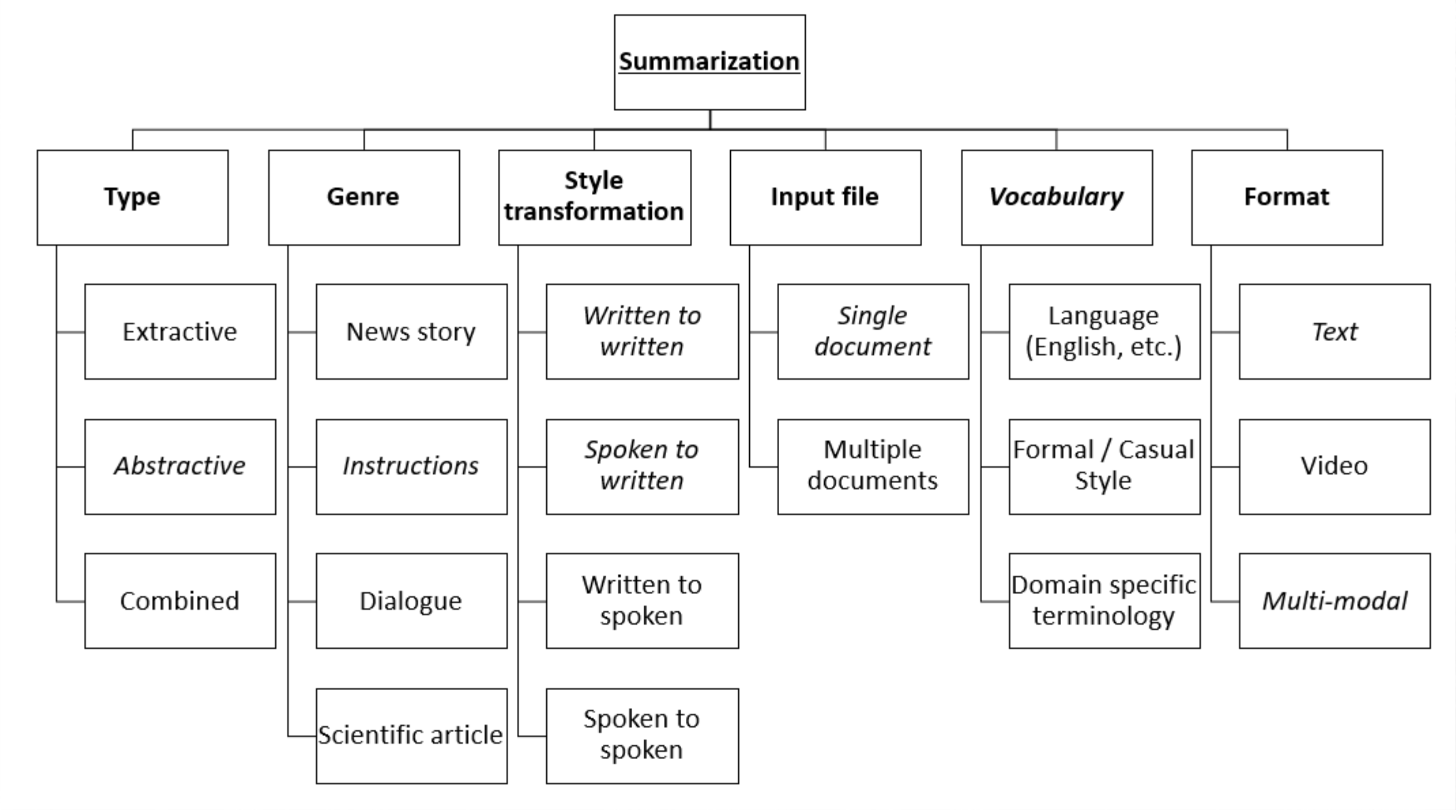


Figure 1: Architecture of the original BERT model (left) and BERTSUM (right). The sequence on top is the input document, followed by the summation of three kinds of embeddings for each token. The summed vectors are used as input embeddings to several bidirectional Transformer layers, generating contextual vectors for each token. BERTSUM extends BERT by inserting multiple [CLS] symbols to learn sentence representations and using interval segmentation embeddings (illustrated in red and green color) to distinguish multiple sentences.

Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). ACL.

WikiHow vs. How2



How2 Data

Auto-Generated Transcripts

00:52	Staining wood is not hard,
00:54	but what's key is the prep.
00:56	And I'm not one for details,
00:58	and this is very detail oriented,
01:00	but I'm gonna make it happen,
01:01	and I'm gonna show you guys how you can do it too.
01:03	Couple of key things you need.
01:05	First of all, the stain ingredients, I call them.
01:08	If you're gonna be using pine, or a soft wood,
01:11	you need a pre-stain conditioner.
01:13	This is really important.
01:14	I skipped this the first time I did it,

How To Stain Wood - GardenFork

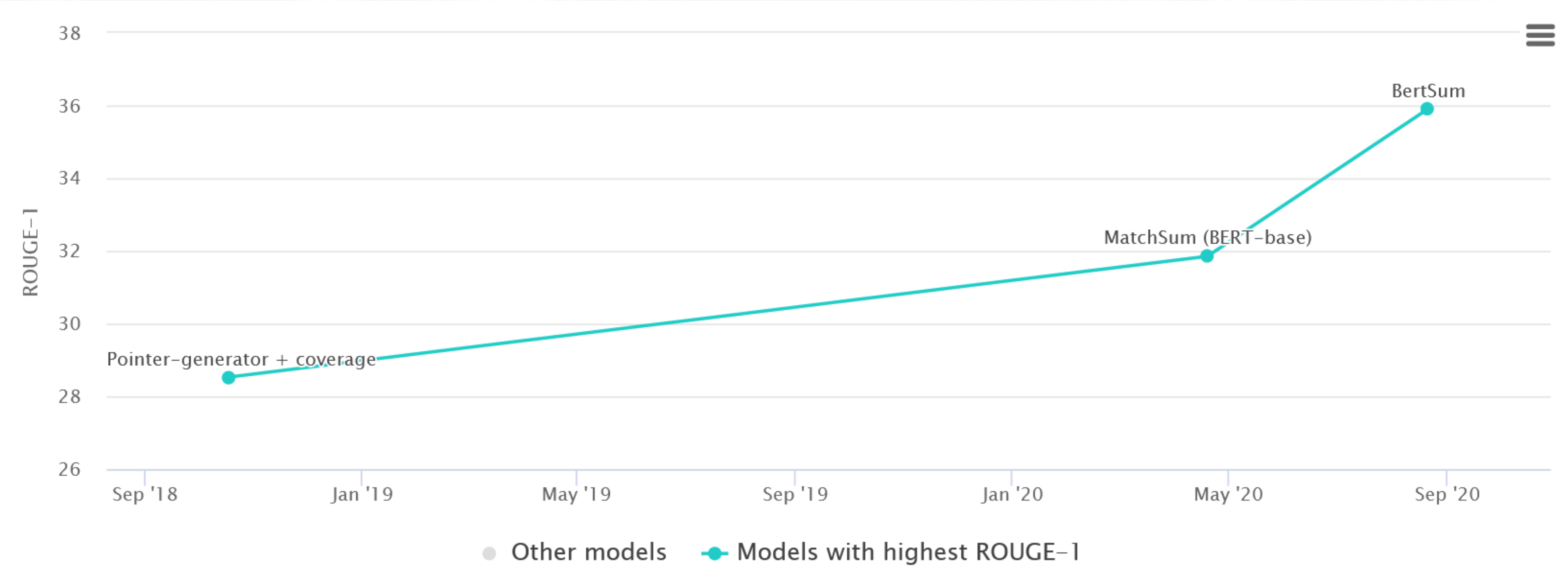
78,859 views • Mar 1, 2016

857 32 SHARE SAVE

Machine Generated Summary:
watch as a seasoned professional demonstrates how to prepare wood for staining in this free online video about home care. get tips for staining wood in the home of a professional carpenter. find out how to stain wood furniture for a home improvement.

https://www.youtube.com/watch?v=BFCJPkabNSo

WikiHow: Results



Text-Generation oriented Metrics: ROUGE

- ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing.

- ROUGE (std)
  (usually averaged across sentences)

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

- ROUGE-L (Longest Common Subsentence)

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (2)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (3)$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (4)$$

Other content-oriented metrics:

- Fluency: Does the text have a natural flow and rhythm?
- Usefulness: Does it have enough information to make a user decide whether they want to spend time watching the video?
- Succinctness: Does the text look concise or do does it have redundancy?
- Consistency: Are there any non sequiturs - ambiguous, confusing or contradicting statements in the text?
- Realisticity: Is there anything that seems far-fetched and bizarre in words combinations and doesn’t look "normal"?
- All grading options are in 0-5 range

Content-based metrics

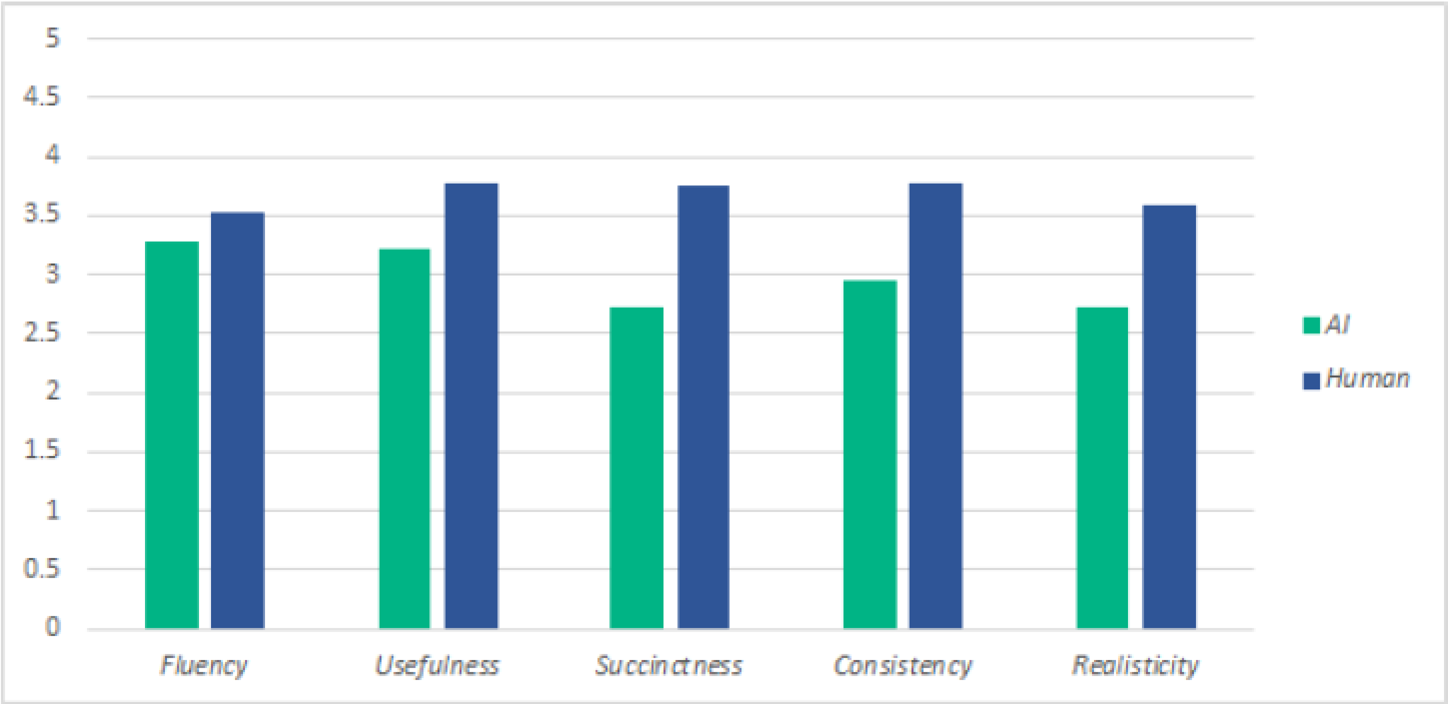
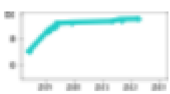
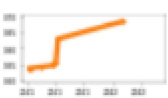
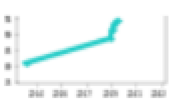


Figure 8: Quality assessment of generated summaries

GLUE

GLUE: overall view

Trend	Task	Dataset Variant	Best Model	Paper	Code
	**Natural Language Inference**	RTE	Vega v2 6B		
	**Text Classification**	GLUE	distilbert-base-uncased-finetuned-sst-2-english		
	**Semantic Textual Similarity**	MRPC	MT-DNN-SMART		
	**Linguistic Acceptability**	CoLA	En-BERT + TDA + PCA		
	**Natural Language Inference**	QNLI	ALBERT		

Glue: Single Sentence Tasks

- **CoLA** The Corpus of Linguistic Acceptability (Warstadt et al., 2018) consists of English acceptability judgments drawn from books and journal articles on linguistic theory. Each example is a sequence of words annotated with whether it is a grammatical English sentence.
  - **Metrics**: we use Matthews correlation coefficient (Matthews, 1975) as the evaluation metric, which evaluates performance on unbalanced binary classification and ranges from -1 to 1, with 0 being the performance of uninformed guessing.
- **SST-2 The Stanford Sentiment Treebank** (Socher et al., 2013) consists of sentences from movie reviews and human annotations of their sentiment. The task is to predict the sentiment of a given sentence. We use the two-way (positive/negative) class split and only sentence-level labels.

GLUE:
SIMILARITY AND PARAPHRASE TASKS

- **MRPC The Microsoft Research Paraphrase Corpus** (Dolan & Brockett, 2005) is a corpus of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent.
  - Classes are imbalanced (68% positive), Metrics: accuracy, F1 score.
- **QQP** The Quora Question Pairs data set is a collection of question pairs from the community question-answering website Quora. The task is to determine whether a pair of questions are semantically equivalent. As in MRPC, the class distribution in QQP is unbalanced (63% negative).
  - Standard test set are used, for which private labels have been made available. The test set has a different label distribution than the training set.
- **STS-B The Semantic Textual Similarity Benchmark** (Cer et al., 2017) is a collection of sentence pairs drawn from news headlines, video and image captions, and natural language inference data.

GLUE: Inference Tasks

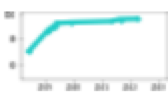
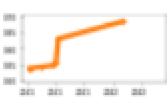
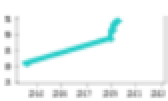
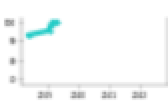
- **MNLI The Multi-Genre Natural Language Inference Corpus** (Williams et al., 2018) is a crowdsourced collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). The premise sentences are gathered from ten different sources, including transcribed speech, fiction, and government reports.

- **QNLI The Stanford Question Answering Dataset** (Rajpurkar et al. 2016) is a question-answering dataset consisting of question-paragraph pairs, where one of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the corresponding question (written by an annotator). We convert the task into sentence pair classification by forming a pair between each question and each sentence in the corresponding context, and filtering out pairs with low lexical overlap between the question and the context sentence. The task is to determine whether the context sentence contains the answer to the question. We call the converted dataset QNLI (Question-answering NLI)

- **RTE The Recognizing Textual Entailment (RTE) datasets** come from a series of annual textual entailment challenges. Combine the data from RTE1 (Dagan et al., 2006), RTE2 (Bar Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009).4 Examples are constructed based on news and Wikipedia text. We convert all datasets to a two-class split, where for three-class datasets we collapse neutral and contradiction into not entailment, for consistency.

- **WNLI The Winograd Schema Challenge** (Levesque et al., 2011) is a reading comprehension task in which a system must read a sentence with a pronoun and select the referent of that pronoun from a list of choices.

GLUE: overall view

Trend	Task	Dataset Variant	Best Model	Paper	Code
	**Natural Language Inference**	RTE	Vega v2 6B		
	**Text Classification**	GLUE	distilbert-base-uncased-finetuned-sst-2-english		
	**Semantic Textual Similarity**	MRPC	MT-DNN-SMART		
	**Linguistic Acceptability**	CoLA	En-BERT + TDA + PCA		
	**Natural Language Inference**	QNLI	ALBERT		

Winogrande

Winogrande: motivation

- The Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011), is a benchmark for commonsense reasoning,
- Includes aset of 273 expert-crafted pronoun resolution problems originally designed to be unsolvable for statistical models that rely on selectional preferences or word associations.
- Recent advances in neural language models have already reached around 90% accuracy on variants of WSC.
- Have these models have truly acquired robust commonsense capabilities?
- Are they only related to spurious biases in the datasets (i.e. overestimation of the true capabilities of machine commonsense.

Winogrande: the dataset

- WinoGrande, a large-scale dataset of 44k problems, inspired by the original WSC
- Adjusted to improve both the scale and the complexity of the dataset.
- Key steps:
  - a carefully designed crowdsourcing procedure, followed by
  - systematic bias reduction using a novel AfLite algorithm that generalizes human-detectable word associations to machine-detectable embedding associations.
- State-of-the-art methods on WinoGrande is 59.4-79.1%, which are 15-35% below human performance of 94.0%, depending on the amount of the training data allowed.
- Implications:
  - demonstrate the effectiveness of WinoGrande when used as a resource for transfer learning.
  - raise a concern that we are likely to be overestimating the true capabilities of machine commonsense across all these benchmarks.
  - emphasize the importance of algorithmic bias reduction in existing and future benchmarks to mitigate such overestimation.

Winogrande: examples

		Twin sentences	Options (answer)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

Table 1: WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices. The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)-(3) are drawn from WSC (Levesque, Davis, and Morgenstern 2011) and (4) from DPR (Rahman and Ng 2012)). Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable since the word "predators" is more often associated with the word "lions", compared to "zebras"

Winogrande: elicitation

- Data Bias Reduction

Algorithm 1: AFLITE

Input: dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, ensemble size n, training set size m, cutoff size k, filtering threshold τ
Output: dataset $\mathcal{D}'$

$\mathcal{D}' = \mathcal{D}$
while $|\mathcal{D}'| > m$ **do**
    `// Filtering phase`
    forall $e \in \mathcal{D}'$ **do**
        Initialize the ensemble predictions $E(e) = \emptyset$
    for *iteration* $i : 1..n$ **do**
        Random partition $(\mathcal{T}_i, \mathcal{V}_i)$ of $\mathcal{D}'$ s.t. $|\mathcal{T}_i| = m$
        Train a linear classifier $\mathcal{L}$ on $\mathcal{T}_i$
        forall $e = (\mathbf{x}, y) \in \mathcal{V}_i$ **do**
            Add $\mathcal{L}(\mathbf{x})$ to $E(e)$
    forall $e = (\mathbf{x}, y) \in \mathcal{D}'$ **do**
        $score(e) = \frac{|\{p \in E(e) \text{ s.t. } p = y\}|}{|E(e)|}$
    Select the top-k elements $\mathcal{S}$ in $\mathcal{D}'$ s.t. $score(e) \geq \tau$
    $\mathcal{D}' = \mathcal{D}' \setminus \mathcal{S}$
    if $|\mathcal{S}| < k$ **then**
        break
return $\mathcal{D}'$

Winogrande: debiased sent's

	Twin sentences	Options (answer)
✗	The monkey loved to play with the balls but ignored the blocks because he found **them** *exciting*.	**balls** / blocks
	The monkey loved to play with the balls but ignored the blocks because he found **them** *dull*.	balls / **blocks**
✗	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *weak*.	**William** / Jason
	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *strong*.	William / **Jason**
✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *less* time to get ready for school.	**Robert** / Samuel
	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *more* time to get ready for school.	Robert / **Samuel**
✓	The child was screaming after the baby bottle and toy fell. Since the child was *hungry*, **it** stopped his crying.	**baby bottle** / toy
	The child was screaming after the baby bottle and toy fell. Since the child was *full*, **it** stopped his crying.	baby bottle / **toy**

Table 2: Examples that have *dataset-specific* bias detected by AFLITE (marked with ✗). The words that include (dataset-specific) polarity bias (§3) are highlighted (positive and negative). For comparison, we show examples selected from WINOGRANDE$_{debiased}$ (marked with ✓).

Winogrande: early results

Methods	dev acc. (%)	test acc.(%)
WKH	49.4	49.6
Ensemble LMs	53.0	50.9
BERT	65.8	64.9
RoBERTa	**79.3**	**79.1**
BERT (local context)	52.5	51.9
RoBERTa (local context)	52.1	50.0
BERT-DPR⋆	50.2	51.0
RoBERTa-DPR⋆	59.4	58.9
Human Perf.	94.1	94.0

Table 3: Performance of several baseline systems on WINOGRANDE$_{debiased}$ (dev and test). The star (⋆) denotes that it is zero-shot setting (e.g., BERT-DPR⋆ is a BERT model fine-tuned with the DPR dataset and evaluated on WINOGRANDE$_{debiased}$.)

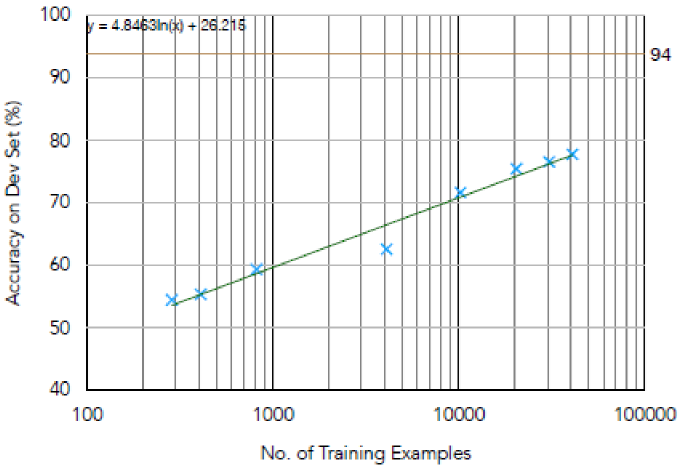


Figure 2: Learning curve on the dev set of WINOGRANDE. Each point on the plot is the best performance for a given number of randomly selected training examples, computed over ten random seeds.

Training size	dev acc. (%)	test acc.(%)
XS (160)	51.5	50.4
S (640)	58.6	58.6
M (2,558)	66.9	67.6
L (10,234)	75.8	74.7
XL (40,938)	79.3	79.1

Table 4: Performance of RoBERTa with different training sizes.

SQuAD

SQUAD

- The **Stanford Question Answering Dataset** (**SQuAD**) is a collection of question-answer pairs derived from Wikipedia articles.
- The correct answers of questions can be any sequence of tokens in the given text.
- Produced by humans through crowdsourcing (more diverse than some other question-answering datasets).
- SQuAD 1.1 contains 107,785 question-answer pairs on 536 articles.
- SQuAD2.0 (open-domain SQuAD, SQuAD-Open), the latest version, combines the 100,000 questions in SQuAD1.1 with over 50,000 un-answerable questions written adversarially by crowdworkers in forms that are similar to the answerable ones.

SQUAD 1.1

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

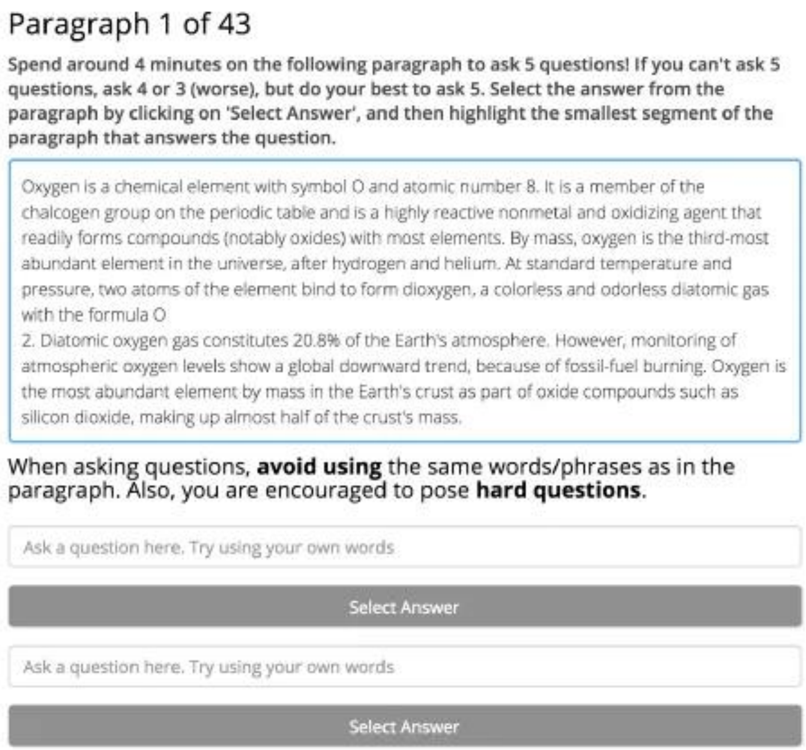


Figure 2: The crowd-facing web interface used to collect the dataset encourages crowdworkers to use their own words while asking questions.

SQuAD Home page

The Stanford Question Answering Dataset

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

What is Southern California often abbreviated as?
Ground Truth Answers: SoCal SoCal SoCal

Despite being traditionall described as "eight counties", how many counties does this region actually have?
Ground Truth Answers: 10 counties 10 10

What is a major importance of Southern California in relation to California and the United States?
Ground Truth Answers: economic center major economic center economic center

What are the ties that best described what the "eight counties" are based on?
Ground Truth Answers: demographics and economic ties economic demographics and economic

The reasons for the las two counties to be added are based on what?
Ground Truth Answers: historical political divisions historical political divisions historical political divisions

SQUAD 1.1: statistics

Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855
WikiQA (Yang et al., 2015)	query logs	IR, sentence selection	3047
TREC-QA (Voorhees and Tice, 2000)	query logs + human editor	IR, free form	1479
CNN/Daily Mail (Hermann et al., 2015)	summary + cloze	RC, fill in single entity	1.4M
CBT (Hill et al., 2015)	cloze	RC, fill in single word	688K

Table 1: A survey of several reading comprehension and question answering datasets. SQuAD is much larger than all datasets except the semi-synthetic cloze-style datasets, and it is similar to TREC-QA in the open-endedness of the answers.

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Table 2: We automatically partition our answers into the following categories. Our dataset consists of large number of answers beyond proper noun entities.

SQUAD 1.1: Performance metrics

- **Exact match**: the percentage of predictions that match any one of the ground truth answers exactly.
- **(Macro-averaged) F1 score**: average overlap between the prediction and ground truth answer.
  - The prediction and ground truth are treated as bags of tokens, and their F1 is computed.
  - **The maximum F1 over all of the ground truth answers** is taken for a given question, and then averaged across all questions.

SQUAD 1.1: Performance

	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

Table 5: Performance of various methods and humans. Logistic regression outperforms the baselines, while there is still a significant gap between humans.

	F_1	
	Train	Dev
Logistic Regression	91.7%	51.0%
– Lex., – Dep. Paths	33.9%	35.8%
– Lexicalized	53.5%	45.4%
– Dep. Paths	91.4%	46.4%
– Match. Word Freq.	91.7%	48.1%
– Span POS Tags	91.7%	49.7%
– Match. Bigram Freq.	91.7%	50.3%
– Constituent Label	91.7%	50.4%
– Lengths	91.8%	50.5%
– Span Word Freq.	91.7%	50.5%
– Root Match	91.7%	50.6%

Table 6: Performance with feature ablations. We find that lexicalized and dependency tree path features are most important.

	Logistic Regression Dev F1	Human Dev F1
Date	72.1%	93.9%
Other Numeric	62.5%	92.9%
Person	56.2%	95.4%
Location	55.4%	94.1%
Other Entity	52.2%	92.6%
Common Noun Phrase	46.5%	88.3%
Adjective Phrase	37.9%	86.8%
Verb Phrase	31.2%	82.4%
Clause	34.3%	84.5%
Other	34.8%	86.1%

Table 7: Performance stratified by answer types. Logistic regression performs better on certain types of answers, namely numbers and entities. On the other hand, human performance is more uniform.

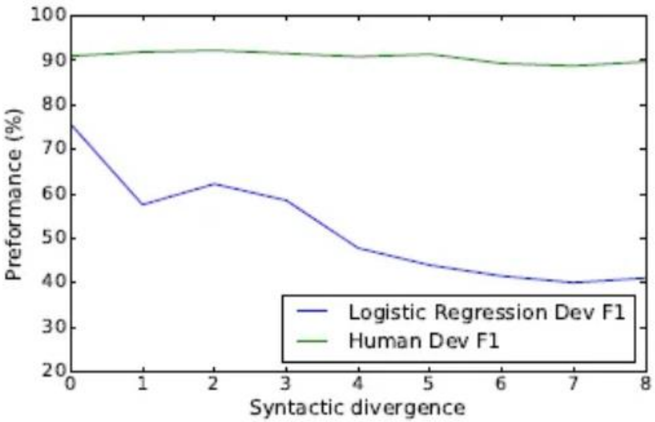
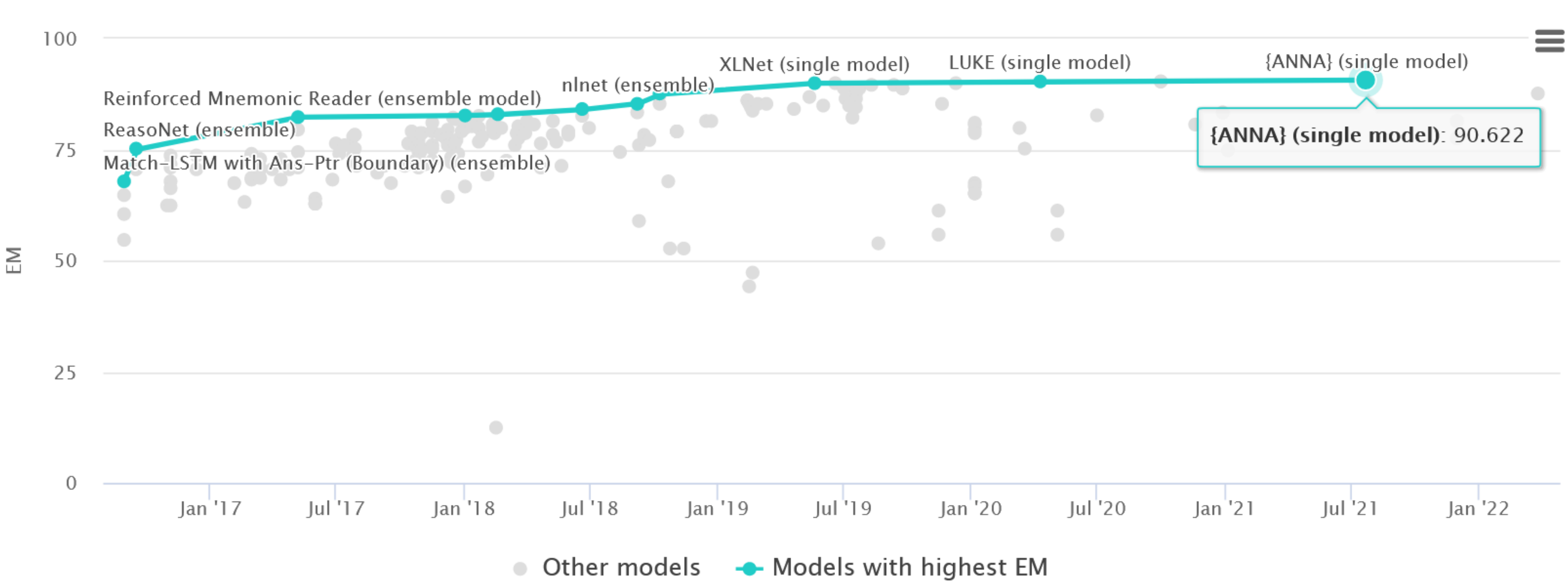


Figure 5: Performance stratified by syntactic divergence of questions and sentences. The performance of logistic regression degrades with increasing divergence. In contrast, human performance is stable across the full range of divergence.

SQUAD nowadays



SQUAD: SpanBERT training

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$

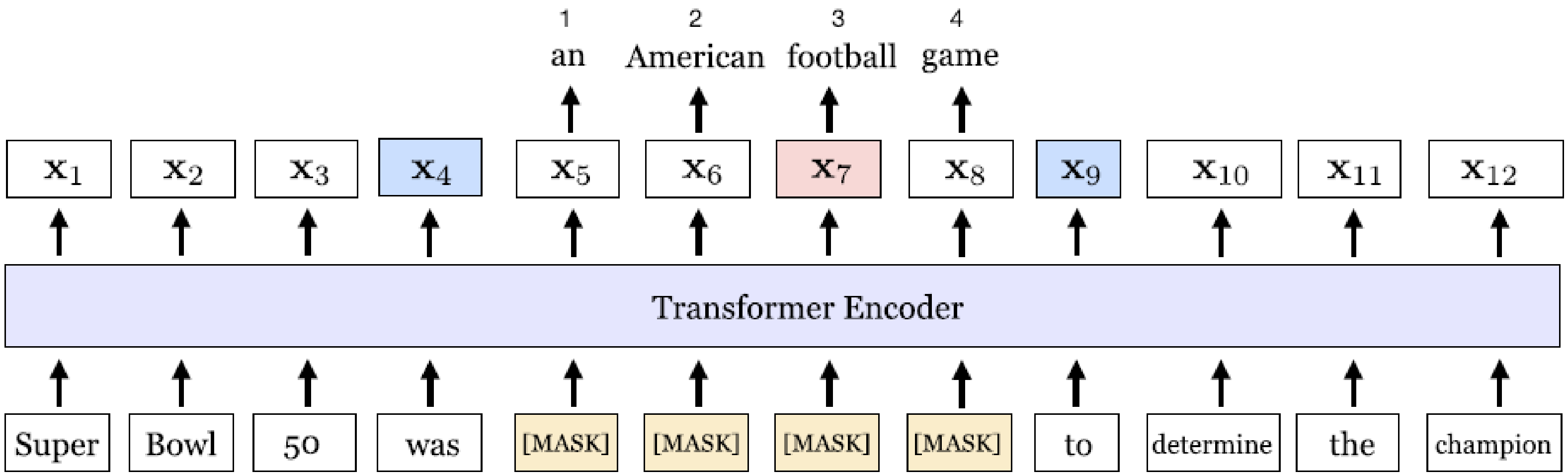


Figure 1: An illustration of SpanBERT training. The span *an American football game* is masked. The SBO uses the output representations of the boundary tokens, $\mathbf{x}_4$ and $\mathbf{x}_9$ (in blue), to predict each token in the masked span. The equation shows the MLM and SBO loss terms for predicting the token, *football* (in pink), which as marked by the position embedding $\mathbf{p}_3$, is the *third* token from x_4.

SpanBERT: Improving Pre-training by Representing and Predicting Spans, Joshi et al., Transactions of the Association for Computational Linguistics, vol. 8, pp. 64–77, 2020.

SpanBERT and SQUAD

35 Apr 13, 2019	SemBERT (ensemble) *Shanghai Jiao Tong University* https://arxiv.org/abs/1909.02209	86.166	88.886
35 Sep 29, 2019	BERTSP (single model) *NEUKG* http://www.techkg.cn/--please	85.838	88.921
35 Sep 22, 2020	RoBERTa-Large (ensemble model) *SAIL*	85.872	88.793
35 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
35 Jul 22, 2019	SpanBERT (single model) *FAIR & UW*	85.748	88.709
36 Sep 21, 2020	RoBERTa-Large (single model) *SAIL*	85.173	88.425

SQuAD Leaderboard
(May 2024)

What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5	Retro-Reader (ensemble) *Shanghai Jiao Tong University*	90.578	92.978

Papers

- TASKS & Datasets:
  - https://paperswithcode.com/area/natural-language-processing
  - https://paperswithcode.com/dataset/glue
  - https://paperswithcode.com/dataset/winogrande
- Papers:
  - Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy & Samuel R. Bowman, GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING, Porc. of ICLR 2019