Maestría en Inteligencia Artificial Aplicada (MNA)

Métricas y bases de datos en NLP:

GLUE, superGLUE, BLEU, ROUGE

Procesamiento de Lenguaje Natural (NLP)

Luis Eduardo Falcón Morales

Métricas en NLP

- La arquitectura Transformer nos ha brindado una manera de generar modelos de última generación en el área de procesamiento de lenguaje natural.

- Los modelos pre-entrenados por su parte nos dan acceso a estas grandes arquitecturas para poder implementar o adaptar a nuestros casos de uso.

- Sin embargo, ¿cómo medir el desempeño de alguno de estos modelos, o dónde conseguir bases de datos que nos permitan entranar o hacer ajuste de parámetros?

- Veamos a continuación la manera en que se ha venido resolviendo estos retos.

Métricas

Para medir el desempeño en el área de Procesamiento de Lenguaje Natural seguiremos utilizando las métricas usuales dentro del área de aprendizaje automático, pero también se agregaran otras que veremos a continuación:

- Accuracy
- Precisión*/Recall*
- F1-score
- Correlación de Pearson o Spearman
- Correlación de Matthew
- BLEU score
- ROUGE score

Bases de Datos

Para poder medir el desempeño de un modelo y después compararlo con algún otro, se requiere contar con corpus de referencia. Existe una gran variedad, pero mencionamos en particular los siguiente:

- Wikipedia
- Bibliotecas
- Parlamento europeo
- …
- GLUE
- superGLUE

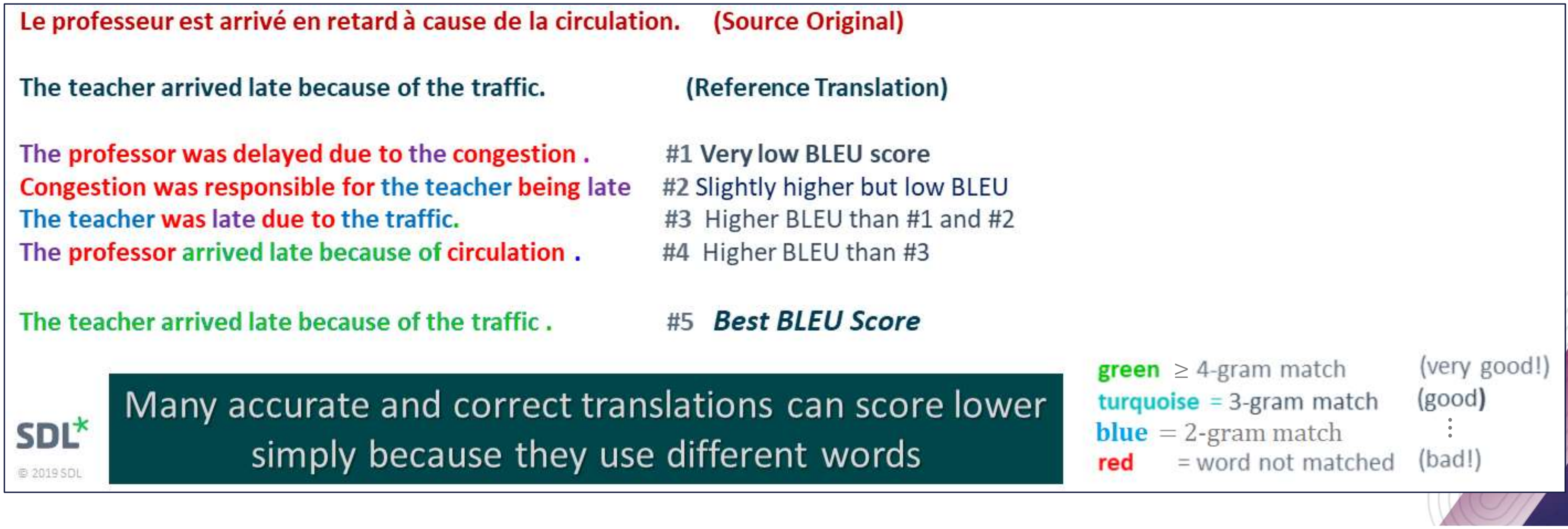
https://aclanthology.org/P02-1040.pdf

BLEU score
Bilingual Evaluation Understudy

El BLEU score es el método mediante el cual se evalúa la calidad de la traducción de un texto de un idioma a otro mediante una Machine-Translation (MT).

La calidad de la traducción se basa en la comparación, *n-grama* por *n-grama*, hecha por humanos. Este valor subjetivo y de palabra por palabra, es una de las mayores críticas que recibe dicha métrica.

La métrica está normalizada de 0 a 1, siendo 1 la de mayor calidad o similaridad con la de un humano.



Le professeur est arrivé en retard à cause de la circulation. (Source Original)

The teacher arrived late because of the traffic. (Reference Translation)

The professor was delayed due to the congestion . #1 Very low BLEU score
Congestion was responsible for the teacher being late #2 Slightly higher but low BLEU
The teacher was late due to the traffic. #3 Higher BLEU than #1 and #2
The professor arrived late because of circulation . #4 Higher BLEU than #3

The teacher arrived late because of the traffic . #5 ***Best BLEU Score***

Many accurate and correct translations can score lower simply because they use different words

green	≥ 4-gram match	(very good!)
turquoise	= 3-gram match	(good)
		⋮
blue	= 2-gram match	
red	= word not matched	(bad!)

SDL*
© 2019 SDL

ROUGE score
Recall-Oriented Understudy for Gisting Evaluation

https://aclanthology.org/W04-1013/

Se apoya en las métricas Precision y Recall para evaluar el enunciado predicho.
También se aplica usando *n-gramas*.
Lo mismo que con BLEU, tampoco toma en cuenta el contexto o palabras similares/sinónimos.

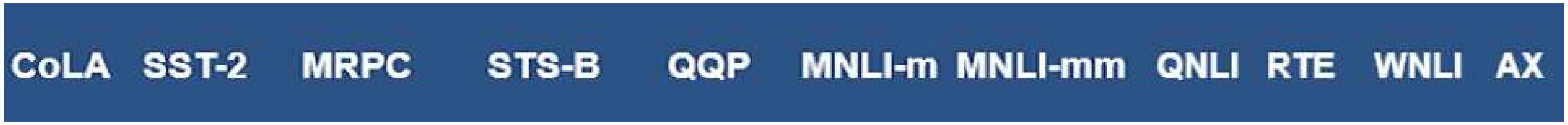
Predicción del modelo: Un pájaro comiendo semillas

Referencia: (basado en humanos) Pájaro comiendo semillas

$$Precision = \frac{Total\ de\ intersección\ de\ n\_gramas}{longitud\ enunciado\ del\ modelo\ de\ prediccion\ en\ n\_gramas} = \frac{3}{4}$$

$$Recall = \frac{Total\ de\ intersección\ de\ n\_gramas}{longitud\ del\ enunciado\ de\ referencia} = \frac{3}{3}$$

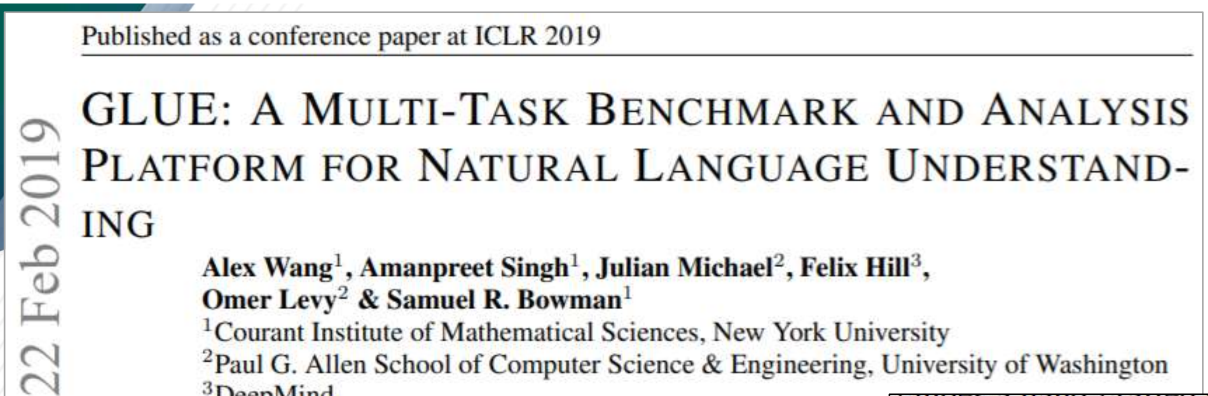
Bases de Datos



CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX

GLUE

General Language Understanding Evaluation



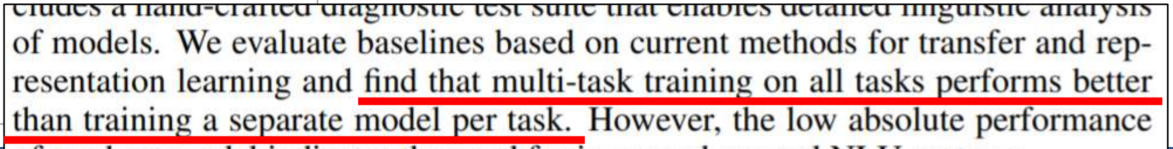
Published as a conference paper at ICLR 2019

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
{alexwang,amanpreet,bowman}@nyu.edu
{julianjm,omerlevy}@cs.washington.edu
felixhill@google.com

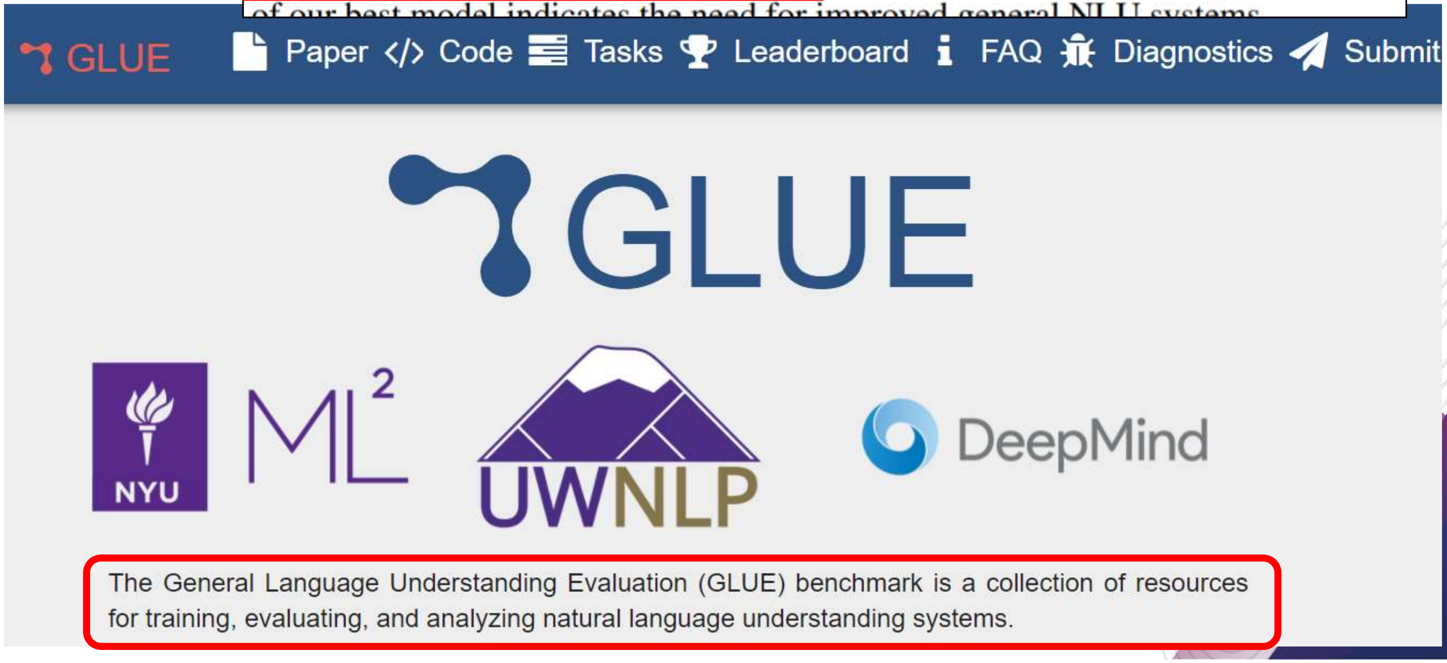
CL] 22 Feb 2019

https://arxiv.org/abs/1804.07461



cludes a hand-crafted diagnostic test suite that enables detailed linguistic analysis of models. We evaluate baselines based on current methods for transfer and representation learning and find that multi-task training on all tasks performs better than training a separate model per task. However, the low absolute performance of our best model indicates the need for improved general NLU systems.

https://gluebenchmark.com/



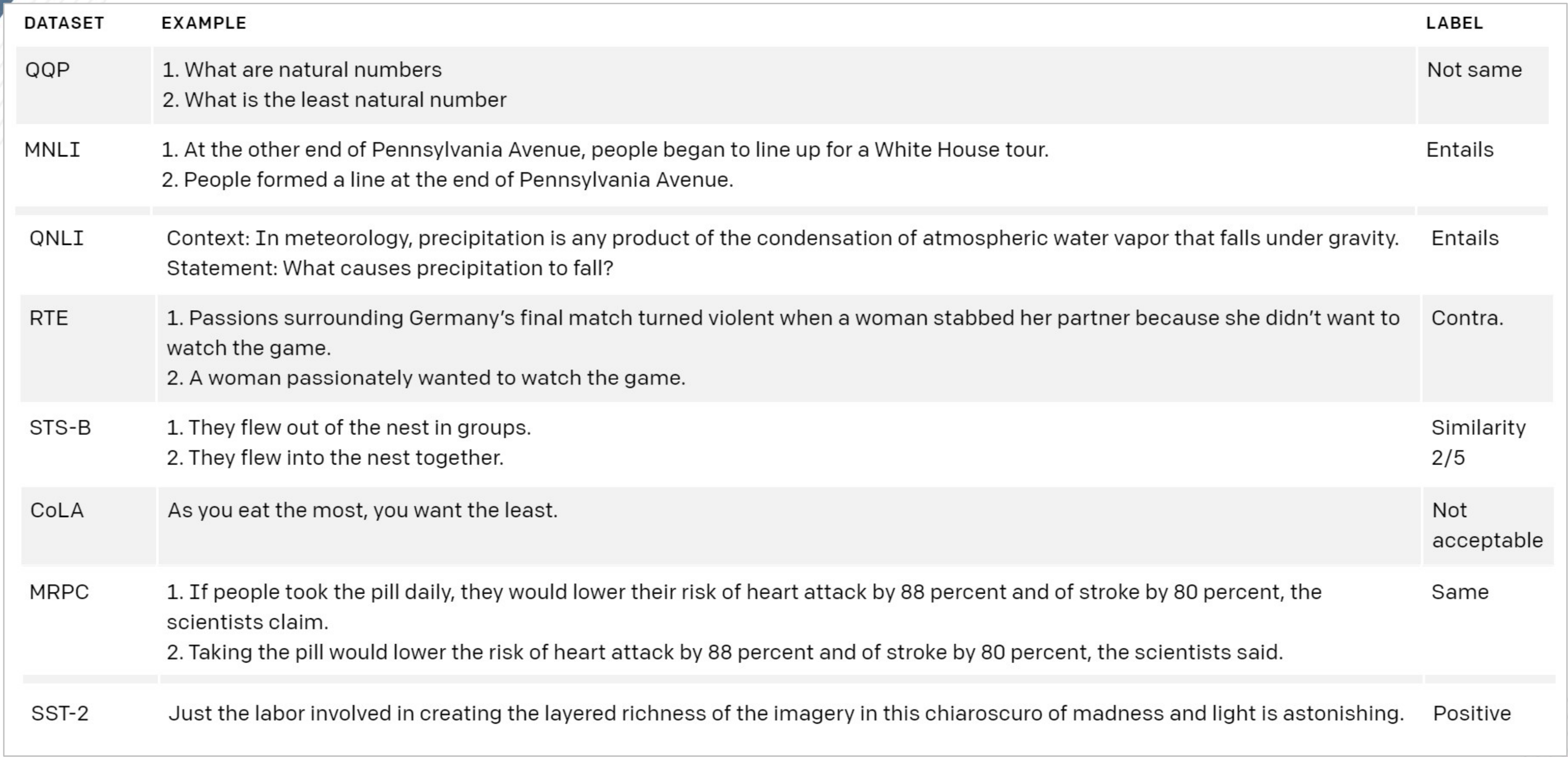
The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems.

https://wp.nyu.edu/ml2/

https://nlp.washington.edu/

https://www.deepmind.com/

https://openai.com/blog/language-unsupervised/

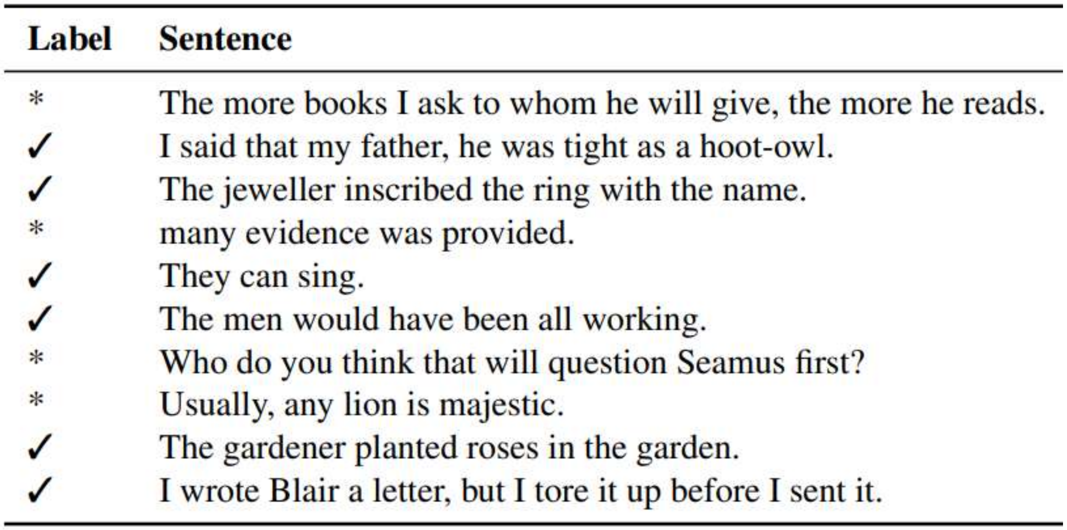


DATASET	EXAMPLE	LABEL
QQP	1. What are natural numbers 2. What is the least natural number	Not same
MNLI	1. At the other end of Pennsylvania Avenue, people began to line up for a White House tour. 2. People formed a line at the end of Pennsylvania Avenue.	Entails
QNLI	Context: In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. Statement: What causes precipitation to fall?	Entails
RTE	1. Passions surrounding Germany's final match turned violent when a woman stabbed her partner because she didn't want to watch the game. 2. A woman passionately wanted to watch the game.	Contra.
STS-B	1. They flew out of the nest in groups. 2. They flew into the nest together.	Similarity 2/5
CoLA	As you eat the most, you want the least.	Not acceptable
MRPC	1. If people took the pill daily, they would lower their risk of heart attack by 88 percent and of stroke by 80 percent, the scientists claim. 2. Taking the pill would lower the risk of heart attack by 88 percent and of stroke by 80 percent, the scientists said.	Same
SST-2	Just the labor involved in creating the layered richness of the imagery in this chiaroscuro of madness and light is astonishing.	Positive

CoLA

https://nyu-mll.github.io/CoLA/

The Corpus of Linguistic Acceptability



Label	Sentence
*	The more books I ask to whom he will give, the more he reads.
✓	I said that my father, he was tight as a hoot-owl.
✓	The jeweller inscribed the ring with the name.
*	many evidence was provided.
✓	They can sing.
✓	The men would have been all working.
*	Who do you think that will question Seamus first?
*	Usually, any lion is majestic.
✓	The gardener planted roses in the garden.
✓	I wrote Blair a letter, but I tore it up before I sent it.

(✓= acceptable, *=unacceptable)

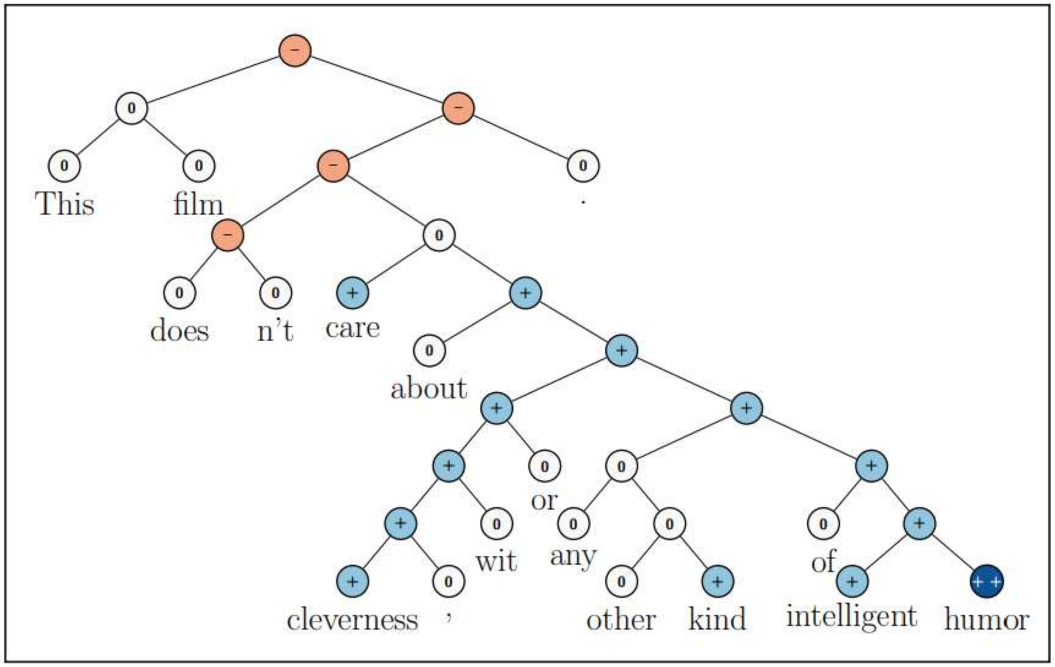
https://arxiv.org/abs/1805.12471

CoLA in its full form consists of 10,657 sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors.

SST-2

https://nlp.stanford.edu/sentiment/index.html

The Stanford Sentiment Treebank



https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

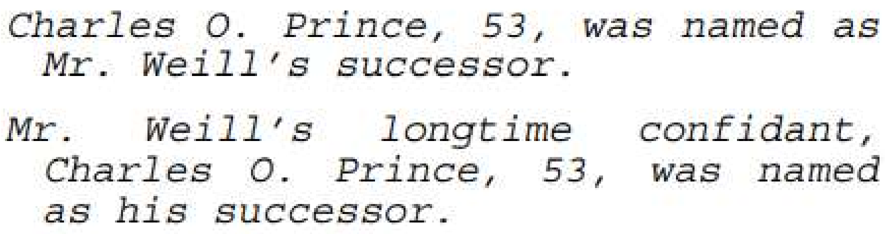
Sentiment analysis for movie reviews. Our new deep learning model actually builds up a representation of whole sentences based on the sentence structure. It computes the sentiment based on how words compose the meaning of longer phrases.

It includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences.

MRPC

Microsoft Research Paraphrase Corpus

https://www.microsoft.com/en-us/download/details.aspx?id=52398



Charles O. Prince, 53, was named as Mr. Weill's successor.

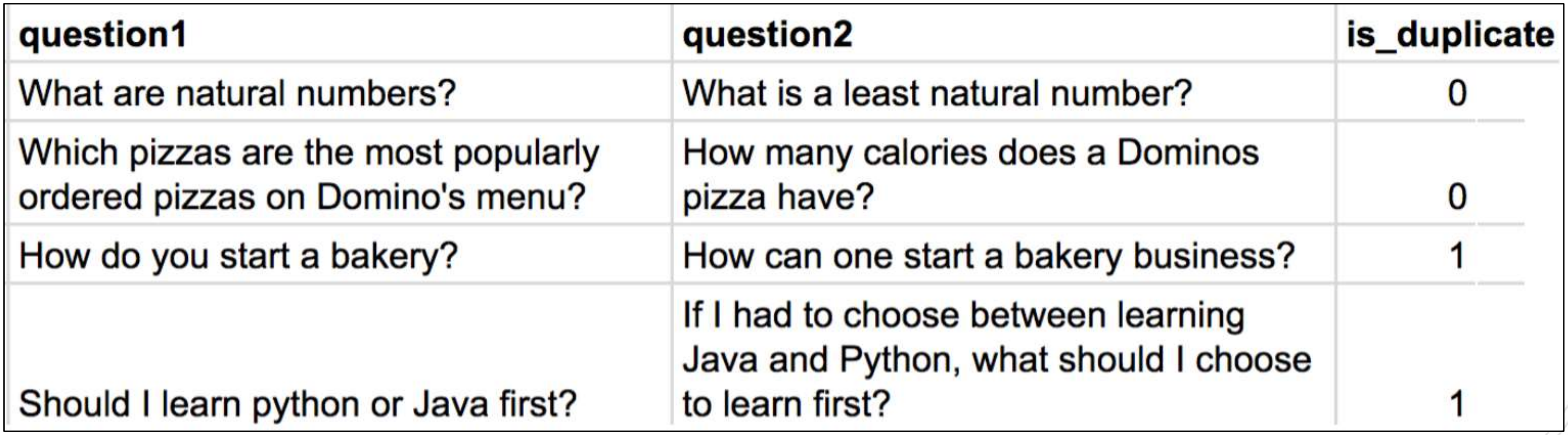
Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

A text file containing 5800 pairs of sentences which have been extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship.

https://aclanthology.org/I05-5002.pdf

QQP

Quora Question Pairs



question1	question2	is_duplicate
What are natural numbers?	What is a least natural number?	0
Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
How do you start a bakery?	How can one start a bakery business?	1
Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

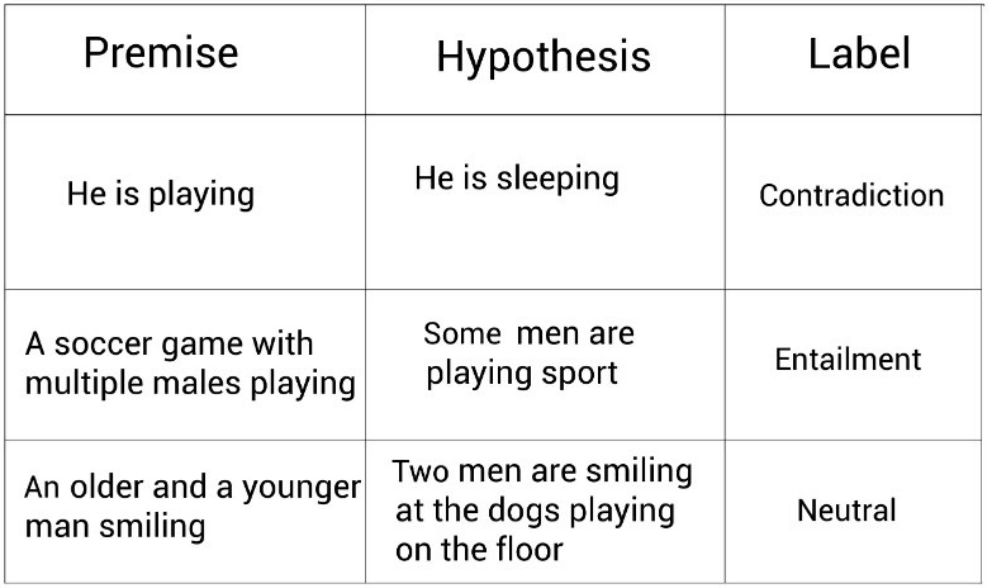
https://arxiv.org/pdf/1702.03814v3.pdf

https://www.quora.com/profile/Ricky-Riche-2/First-Quora-Dataset-Release-Question-Pairs

MNLI

https://cims.nyu.edu/~sbowman/multinli/

Multi-Genre NaturaL Inference (matched/mismatched)

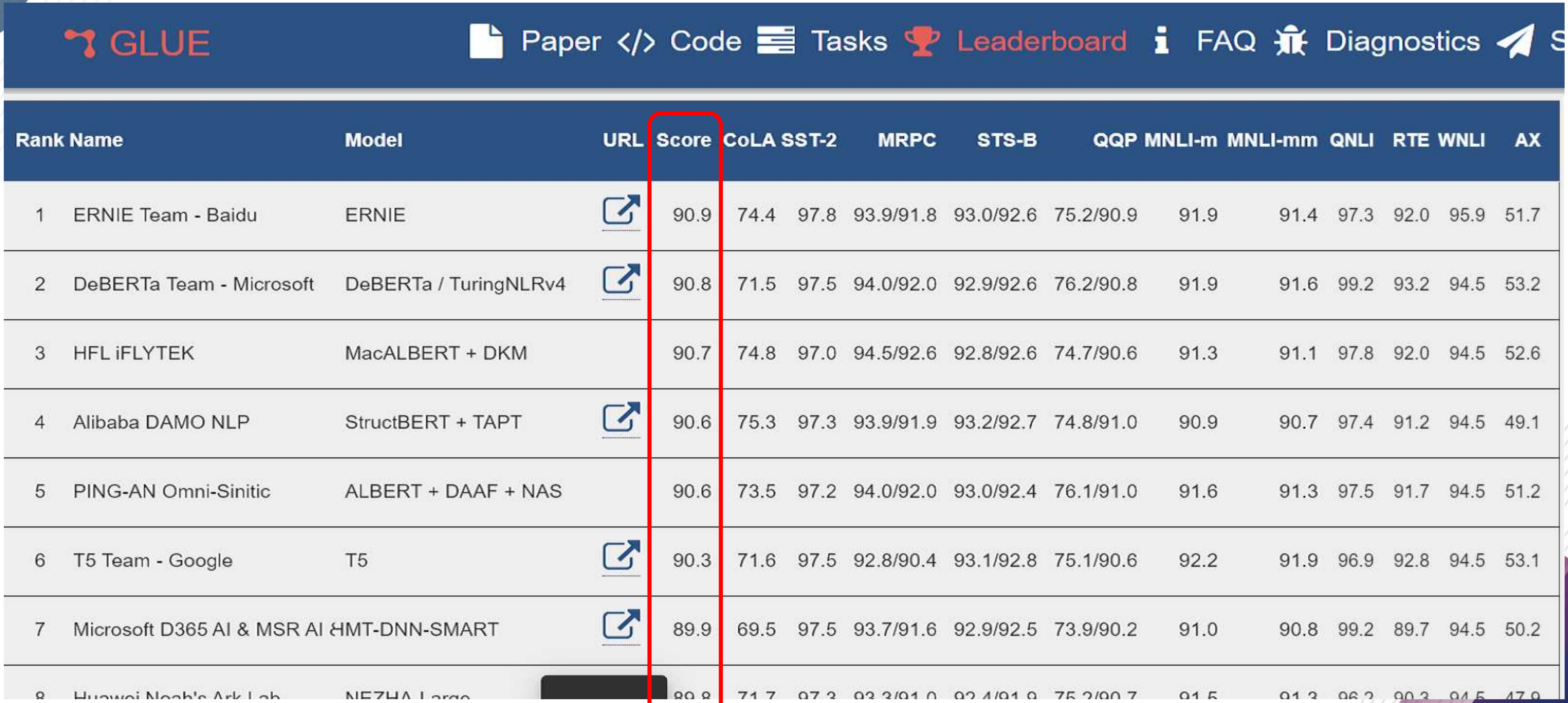


Premise	Hypothesis	Label
He is playing	He is sleeping	Contradiction
A soccer game with multiple males playing	Some men are playing sport	Entailment
An older and a younger man smiling	Two men are smiling at the dogs playing on the floor	Neutral

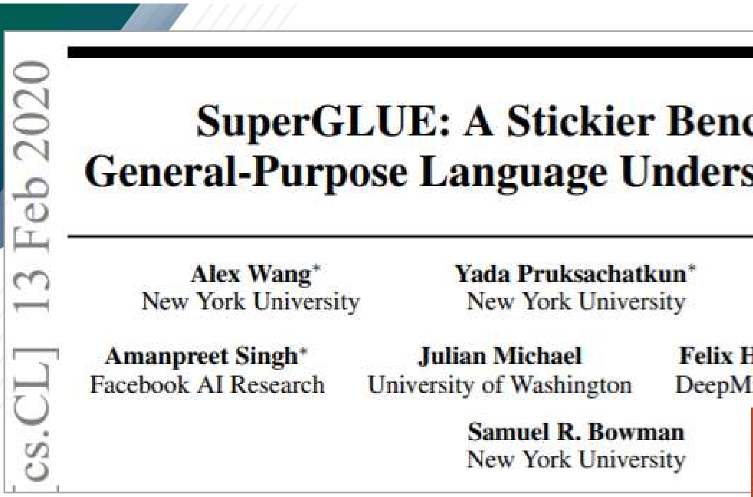
In this task, also known as recognizing textual entailment a model is presented with a pair of sentences and asked to judge the relationship between their meanings by picking a label from a small set: typically ENTAILMENT, NEUTRAL, and CONTRADICTION.

https://cims.nyu.edu/~sbowman/multinli/paper.pdf

En un año aproximadamente este conjunto de datos dejó de ser retador…



Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	ERNIE Team - Baidu	ERNIE	↗	90.9	74.4	97.8	93.9/91.8	93.0/92.6	75.2/90.9	91.9	91.4	97.3	92.0	95.9	51.7
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	↗	90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
3	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
4	Alibaba DAMO NLP	StructBERT + TAPT	↗	90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	90.7	97.4	91.2	94.5	49.1
5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
6	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
7	Microsoft D365 AI & MSR AI &	MT-DNN-SMART	↗	89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
8	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.3	94.5	47.9

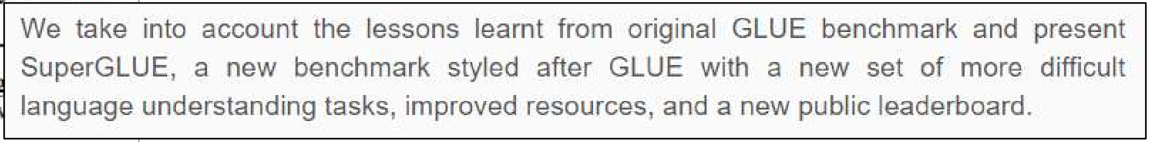


cs.CL] 13 Feb 2020

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

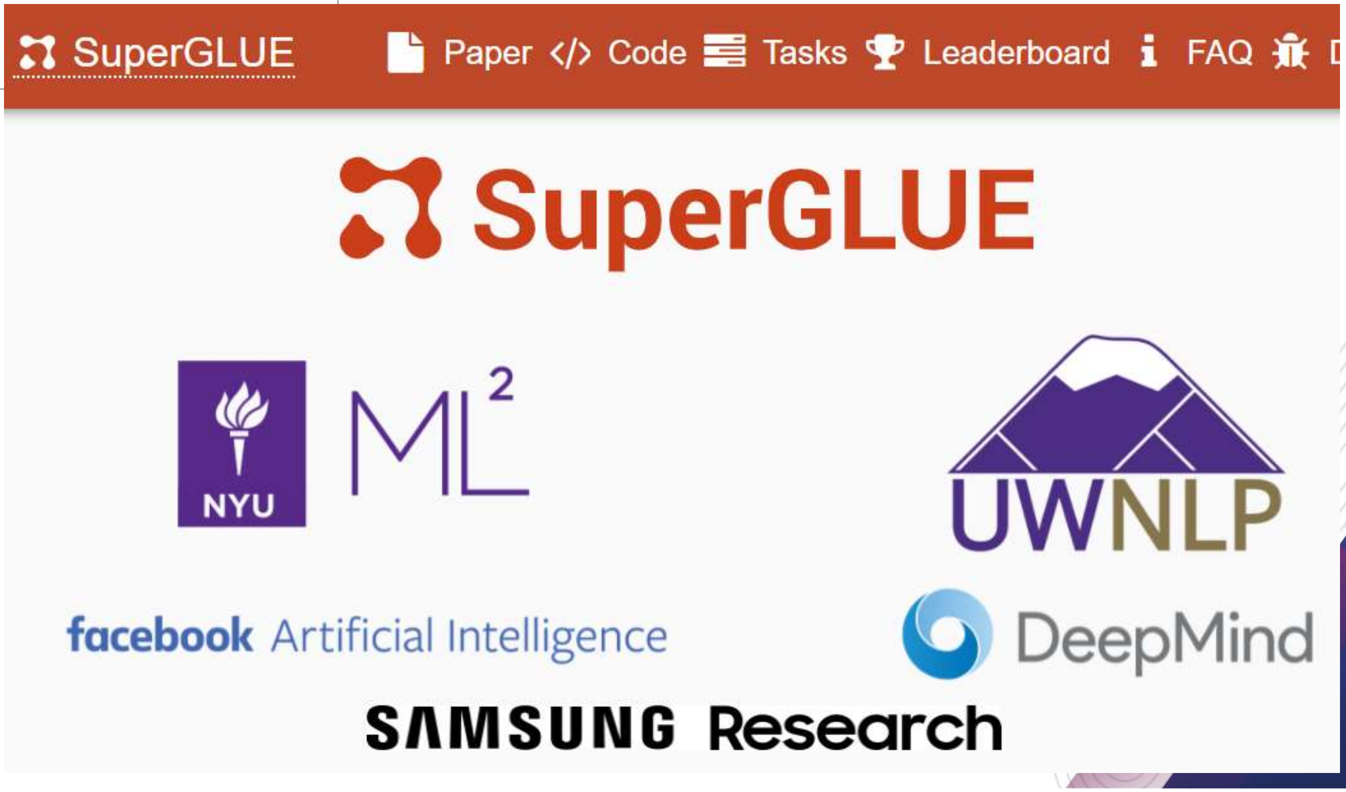
Alex Wang* New York University
Yada Pruksachatkun* New York University
Nikita Nang New York Univ
Amanpreet Singh* Facebook AI Research
Julian Michael University of Washington
Felix Hill DeepMind
Omer Levy Facebook AI Research
Samuel R. Bowman New York University

https://arxiv.org/abs/1905.00537



We take into account the lessons learnt from original GLUE benchmark and present SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard.

https://super.gluebenchmark.com/



https://ai.facebook.com/

https://research.samsung.com/

Choose No if the two sentences are not exact paraphrases and mean different things. For example,

> *"This work caused him to trigger important reflections on the practices of molecular genetics and genomics at a time when this was not considered ethical ."*
>
> *"This work led him to trigger ethical reflections on the practices of molecular genetics and genomics at a time when this was not considered important ."*

"In 2010 Ella Kabambe was not the official Miss Malawi; this was Faith Chibale, but Kabambe represented the country in the Miss World pageant. At the 2012 Miss World, Susan Mtegha pushed Miss New Zealand, Collette Lochore, during the opening headshot of the pageant, claiming that Miss New Zealand was in her space."

*Does **her** refer to option A or B below?*

A *Susan Mtegha*

B *Collette Lochore*

C *Neither*

Choose Insincere if you believe the person asking the question was not really seeking an answer but was being inflammatory, extremely rhetorical, or absurd. For example,

> *"How do I sell Pakistan? I need lots of money so I decided to sell Pakistan any one wanna buy?"*
>
> *"If Hispanics are so proud of their countries, why do they move out?"*
>
> *"Why Chinese people are always not welcome in all countries?"*

Choose Yes if the tag is applicable and accurately describes the selected word or phrase. For example,

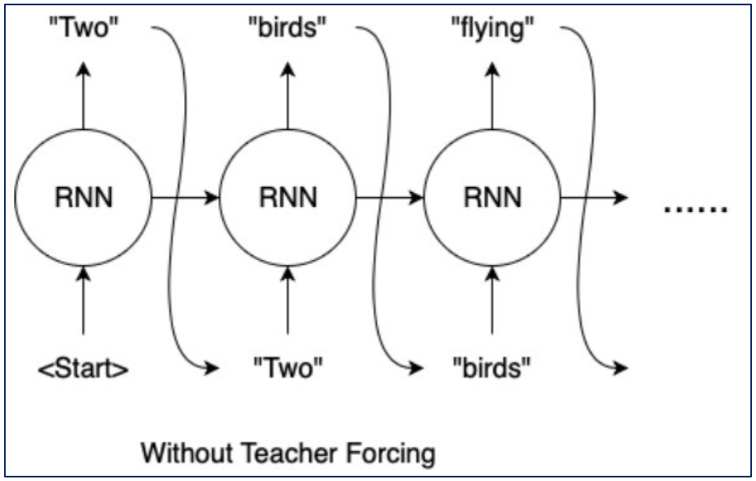
> *"Spain was the gold line." **It** started out with zero gold in 1937, and by 1945 it had 65.5 tons.*
>
> *Tag: nation*

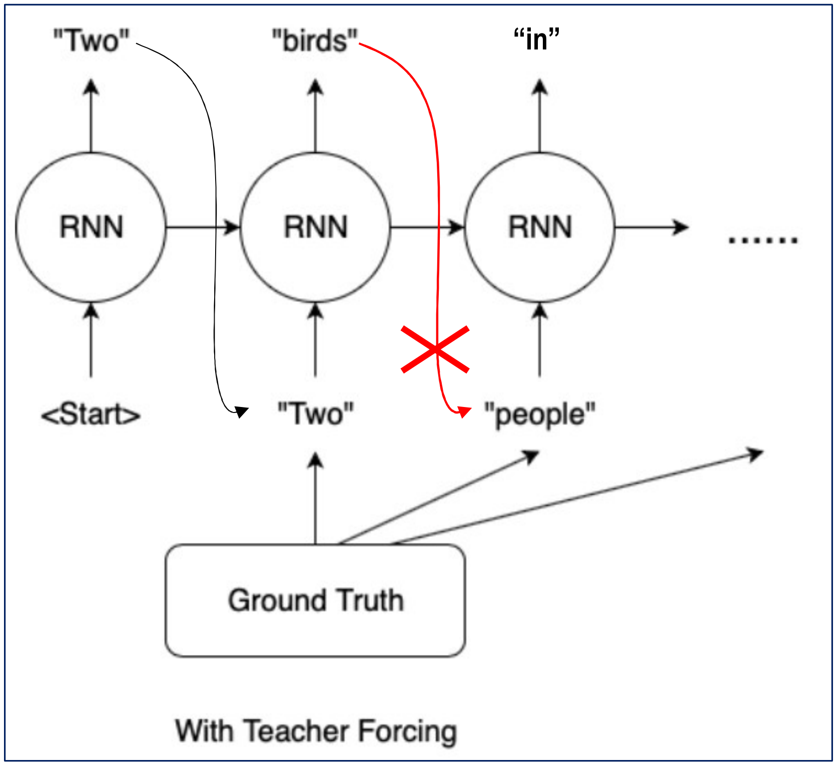
Choose No if the tag is not applicable and does not describes the selected word or phrase. For example,

> ***Iraqi museum workers** are starting to assess the damage to Iraq's history.*
>
> *Tag: organism*

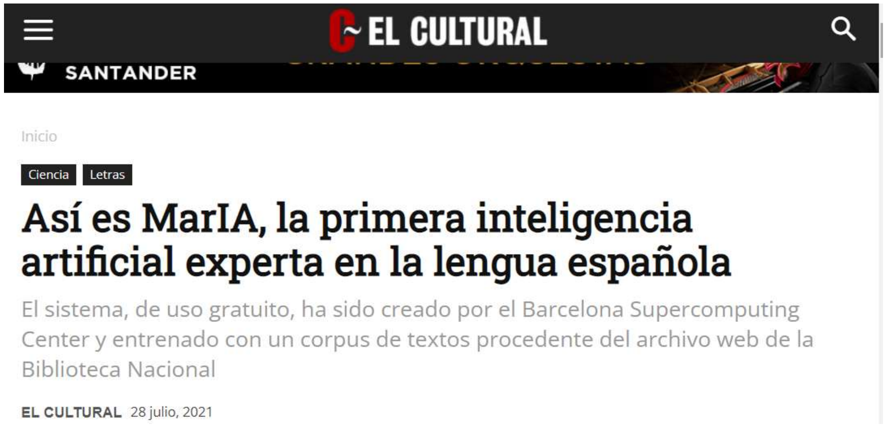
Ground-truth:
Two people in the classroom

Teacher Forcing & Exposure Bias





"Exposure bias" es la discrepancia que se genera entre el modelo con datos de entrenamiento usando Teacher Forcing y el modelo usado con datos de prueba sin Teacher Forcing.



Inicio

Ciencia Letras

Así es MarIA, la primera inteligencia artificial experta en la lengua española

El sistema, de uso gratuito, ha sido creado por el Barcelona Supercomputing Center y entrenado con un corpus de textos procedente del archivo web de la Biblioteca Nacional

EL CULTURAL 28 julio, 2021

https://elcultural.com/asi-es-maria-la-primera-inteligencia-artificial-experta-en-la-lengua-espanola

Tecnológico de Monterrey