IIC 3800 Tópicos en CC
NLP

https://github.com/marcelomendoza/IIC3800

- EVALUACIÓN -

GLUE: General Language Understanding Evaluation

CoLA (Corpus of Linguistic Acceptability): Esta es una tarea de clasificación binaria en la que los sistemas deben predecir si una oración en inglés es gramaticalmente correcta o no.

SST-2 (Stanford Sentiment Treebank): Esta es otra tarea de clasificación binaria en la que los sistemas deben predecir el sentimiento de una oración, ya sea positivo o negativo.

MRPC (Microsoft Research Paraphrase Corpus): En esta tarea, los sistemas deben determinar si dos oraciones son paráfrasis la una de la otra.

STS-B (Semantic Textual Similarity Benchmark): Aquí, los sistemas deben predecir qué tan similar es el significado de dos oraciones en una escala de 0 a 5.

QQP (Quora Question Pairs): En esta tarea, los sistemas deben determinar si dos preguntas son semánticamente equivalentes.

MNLI (Multi-Genre Natural Language Inference): Esta tarea requiere que los sistemas determinen la relación entre una oración de premisa y una oración de hipótesis: si la hipótesis es un entailment, una contradicción, o neutra.

QNLI (Question Natural Language Inference): Esta es una versión de la tarea de Natural Language Inference, en la que se presenta a los sistemas una pregunta y un párrafo, y deben determinar si la respuesta a la pregunta se puede inferir del párrafo.

RTE (Recognizing Textual Entailment): Similar a MNLI, pero las parejas de oraciones provienen de varios sets de datos anteriores.

WNLI (Winograd NLI): Esta tarea está diseñada para evaluar la capacidad de un sistema para la comprensión de la referencia de pronombres, es decir, a qué se refiere un pronombre en un texto.

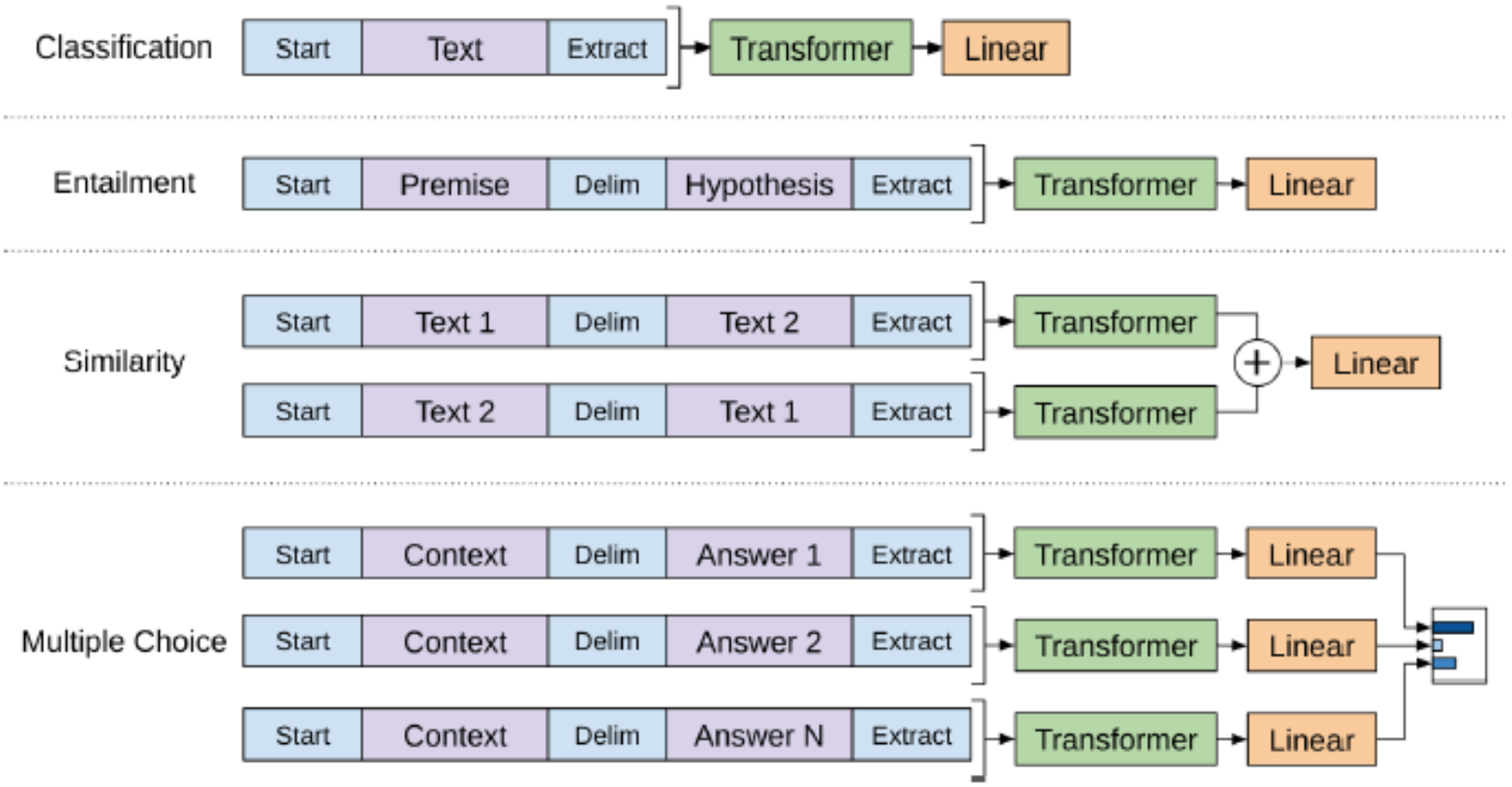
👉 https://gluebenchmark.com/

GLUE: General Language Understanding Evaluation

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

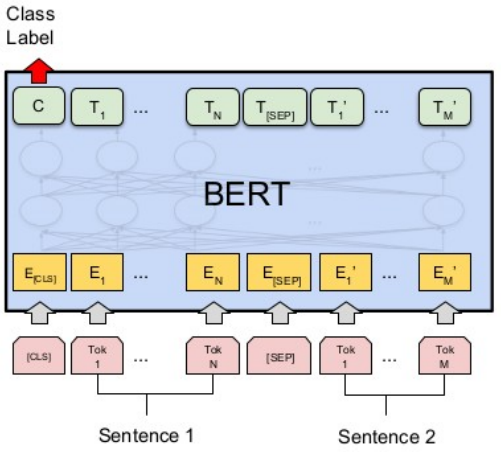
GLUE: General Language Understanding Evaluation

Fine-tuning:

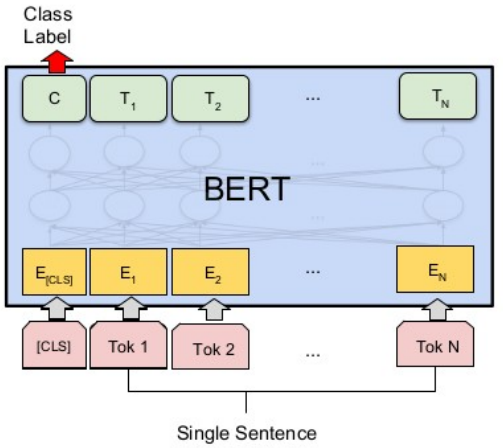


GLUE: General Language Understanding Evaluation

Codificación para GLUE (fine tuning), también se evalúa en SquAD y CoNLL NER:

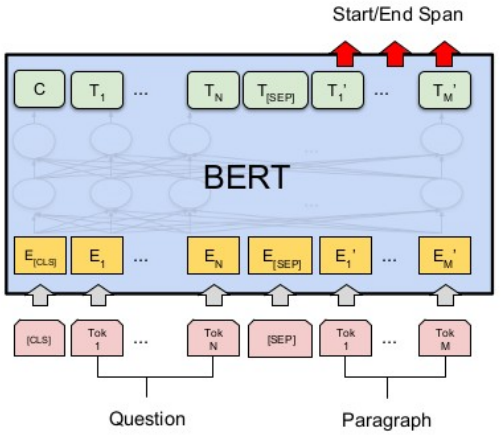


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

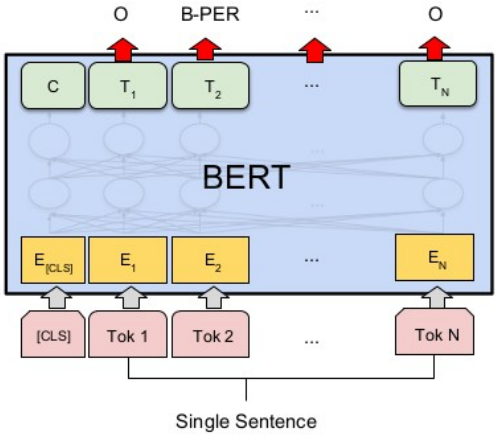


(b) Single Sentence Classification Tasks: SST-2, CoLA

Predecir el fragmento del párrafo que responde a la pregunta



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

GLUE: General Language Understanding Evaluation

Resultados (GLUE):

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

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. BlackboxNLP@EMNLP2018: 353-355

SuperGLUE

BoolQ (BoolQ boolean questions): Los participantes deben responder preguntas de sí o no basadas en un párrafo corto de un artículo de Wikipedia.

CB (CommitmentBank): Se evalúa la capacidad de un modelo para predecir la postura de un hablante hacia una afirmación que ha hecho, basándose en la evidencia proporcionada.

COPA (Choice of Plausible Alternatives): Los modelos tienen que seleccionar la causa o el efecto más plausible de una situación dada.

MultiRC (Multi-Sentence Reading Comprehension): Los modelos deben responder preguntas sobre un párrafo, donde cada pregunta tiene múltiples respuestas correctas.

WiC (Word-in-Context): Esta tarea evalúa si un modelo puede entender el significado de una palabra en dos oraciones diferentes.

WSC (Winograd Schema Challenge): Es una tarea de resolución de la correferencia de pronombres que requiere una comprensión profunda del sentido común y del contexto.

RTE (Recognizing Textual Entailment): Los modelos deben determinar si una oración es verdadera (entails), es falsa (contradicts) o es neutral (neither) en base a una oración previa.

SQuAD v1.1 (Stanford Question Answering Dataset): Es una tarea de respuesta a preguntas basada en la comprensión de un párrafo.

👉 https://super.gluebenchmark.com/

SuperGLUE

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman: SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. NeurIPS2019: 3261-3275

SuperGLUE

Model Metrics	Avg	BoolQ Acc.	CB F1/Acc.	COPA Acc.	MultiRC $F1_a$/EM	ReCoRD F1/EM	RTE Acc.	WiC Acc.	WSC Acc.	AX_b MCC	AX_g GPS Acc.
Most Frequent	47.1	62.3	21.7/48.4	50.0	61.1 / 0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/ 50.0
CBoW	44.3	62.1	49.0/71.2	51.6	0.0 / 0.4	14.0/13.6	49.7	53.0	65.1	-0.4	100.0/ 50.0
BERT	69.0	77.4	75.7/83.6	70.6	70.0 / 24.0	72.0/71.3	71.6	**69.5**	**64.3**	23.0	97.8 / 51.7
BERT++	**71.5**	79.0	**84.7/90.4**	73.8	70.0 / 24.1	72.0/71.3	79.0	**69.5**	**64.3**	38.0	99.4 / 51.4
Outside Best	-	**80.4**	- / -	**84.4**	**70.4*/24.5***	**74.8/73.0**	**82.7**	-	-	-	- / -
Human (est.)	89.8	89.0	95.8/98.9	100.0	81.8*/51.9*	91.7/91.3	93.6	80.0	100.0	77.0	99.3 / 99.7

ROUGE (Recall-oriented understudy for Gisting Evaluation)

Métrica para evaluación de resúmenes automáticos.

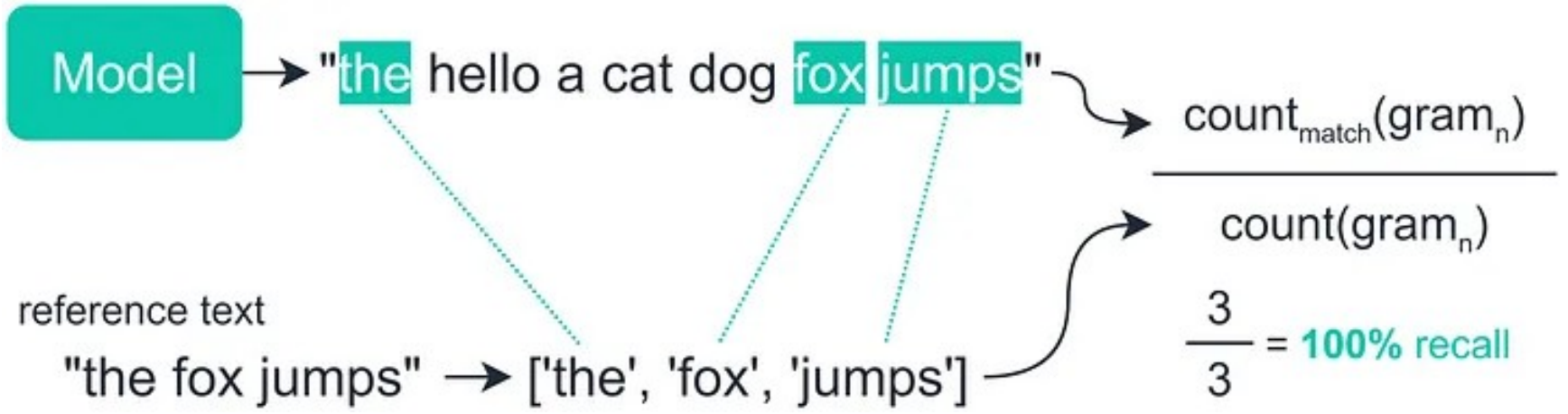
ROUGE-N: Mide la superposición de N-gramas entre el resumen generado y el de referencia. N puede ser cualquier número entero, aunque típicamente se usa 1 (para palabras individuales, es decir, unigramas) y 2 (para pares de palabras, es decir, bigramas). Esta es una medida de precisión y recuperación.

ROUGE-L: Considera la superposición de secuencias más largas de palabras, utilizando la subsecuencia común más larga (LCS, por sus siglas en inglés). Esto ayuda a capturar relaciones más largas entre las palabras y puede ser menos sensible a pequeñas diferencias en la formulación exacta.

ROUGE-S: Mide la superposición de skip-grams.

ROUGE (Recall-oriented understudy for Gisting Evaluation)

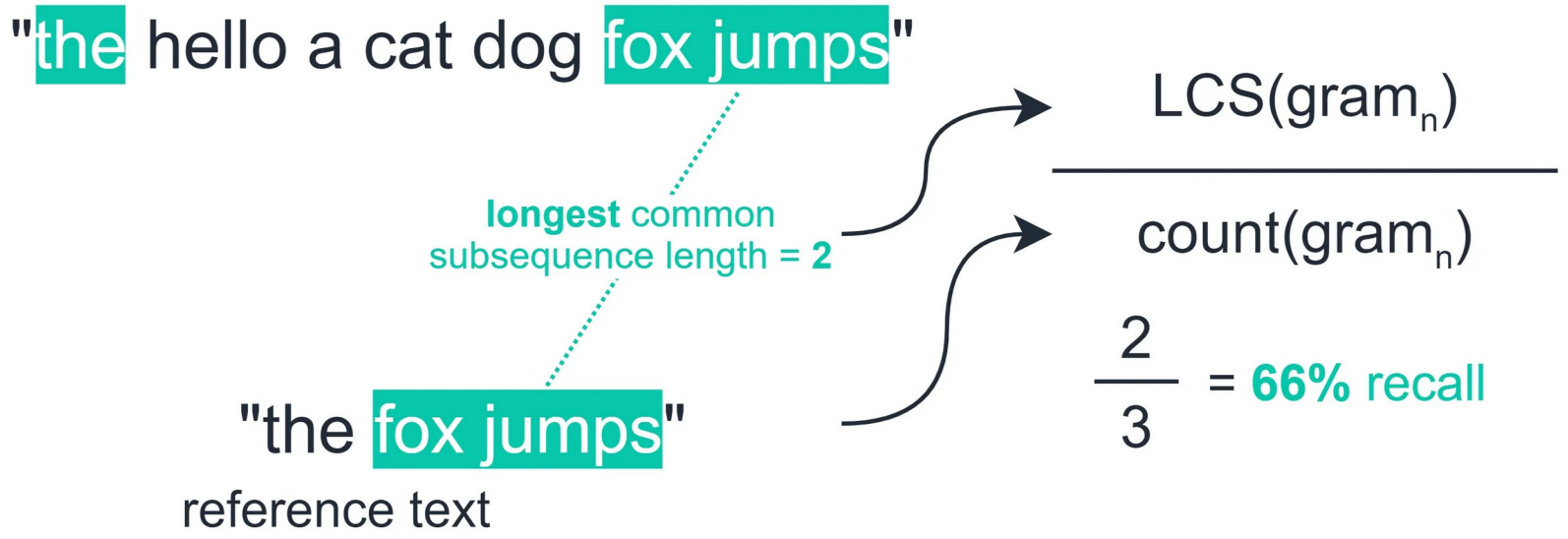
ROUGE-N



ROUGE-1: 1-grams

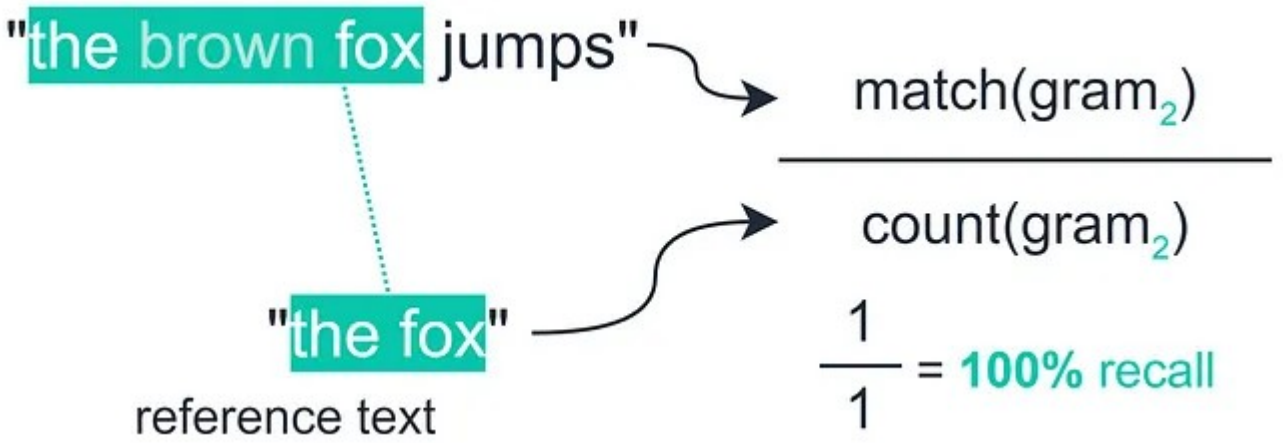
ROUGE (Recall-oriented understudy for Gisting Evaluation)

ROUGE-L



ROUGE (Recall-oriented understudy for Gisting Evaluation)

ROUGE-S



Skip bi-gram

BLEU score (Bilingual Evaluation Understudy)

BLEU (Bilingual Evaluation Understudy) es una métrica que se utiliza para evaluar la calidad de las traducciones producidas por sistemas de traducción automática.

BLEU: % de MT output n-grams que coinciden con el texto de referencia.