Competiciones y Campañas de Evaluación

Yoan Gutiérrez Vázquez

ygutierrez@dlsi.ua.es

Universitat d'Alacant
Universidad de Alicante
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UJI UNIVERSITAT JAUME I
valgrAI

INDICE

1. Benchmarks. ¿Qué es?
2. Marco genérico
3. Campañas de Evaluación
4. Benchmarks e infraestructuras de evaluación
5. Aplicaciones Específicas
6. Bibliografía

INDICE

Benchmark. ¿Qué es?

Definiciones

Economía y finanzas:
El benchmark es un punto de referencia utilizado para medir el rendimiento de una inversión.

Se trata de un indicador financiero utilizado como herramienta de comparación para evaluar el rendimiento de una inversión. [1]

Informática:
Una prueba de rendimiento o comparativa (en inglés benchmark) es una técnica utilizada para medir el rendimiento de un sistema o uno de sus componentes…[2] [3]

[1] https://economipedia.com/definiciones/benchmark.html
[2] (https://es.wikipedia.org/wiki/Benchmark_(informática)
[3] https://dac.cs.vt.edu/research-project/semi-supervised-learning-cyberbullying-harassment-patterns-social-media/

Benchmark. ¿Qué es?

Ventajas:

- **Marco común de comparación** bajo los **mismos criterios**, **recursos** y **estándares**
- **Disponibilidad de datasets** para el desarrollo de problema a abordar
- **Disponibilidad de métricas comunes** de evaluación
- **Disponibilidad colectiva de conocimiento** sobre cómo abordar el problema y ejemplos de iniciación.
- **Disponibilidad de sistemas base** (**baselines**) como recurso básico de comparación

Desventajas:

- Obligatoriedad de **adecuar y adaptar nuestros sistemas** y tecnologías al **marco común**
- **Limitaciones** de nuestros **sistemas** al **objetivo del reto**
- Es preciso **medir nuestro sistema** con las **métricas ya establecidas en el reto** lo cual puede no ser una forma fiel de medir su calidad

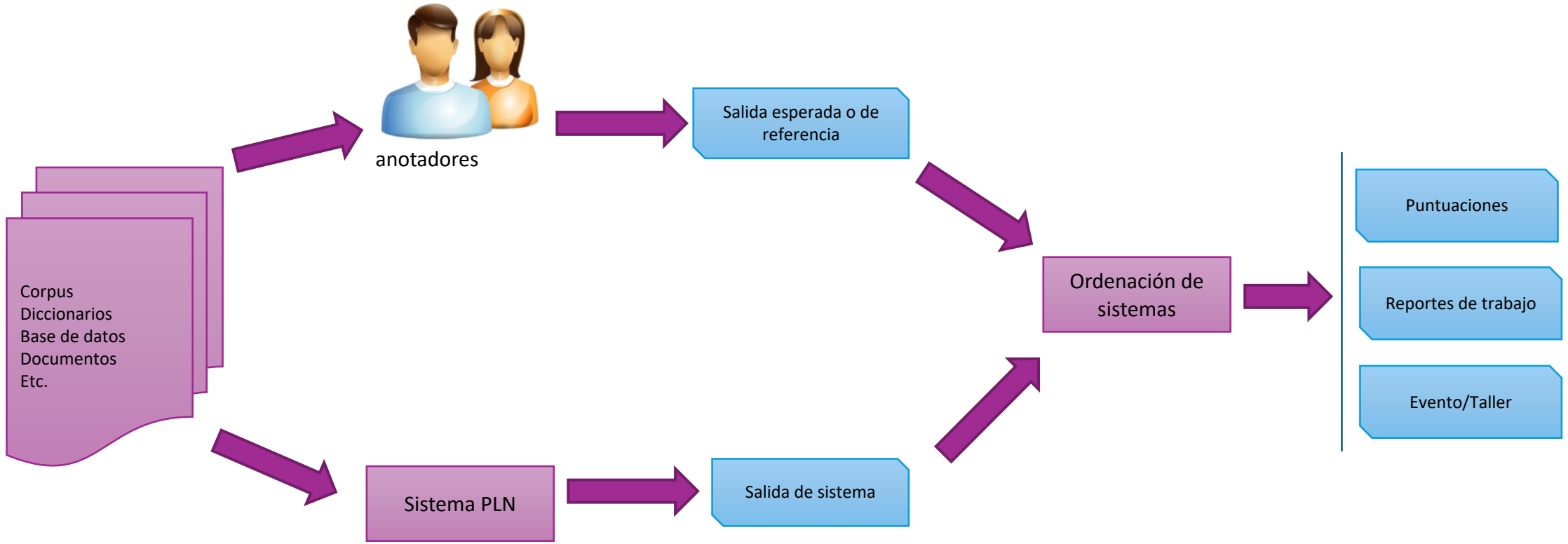
experto en procesamiento del lenguaje natural

Campañas de evaluación más populares

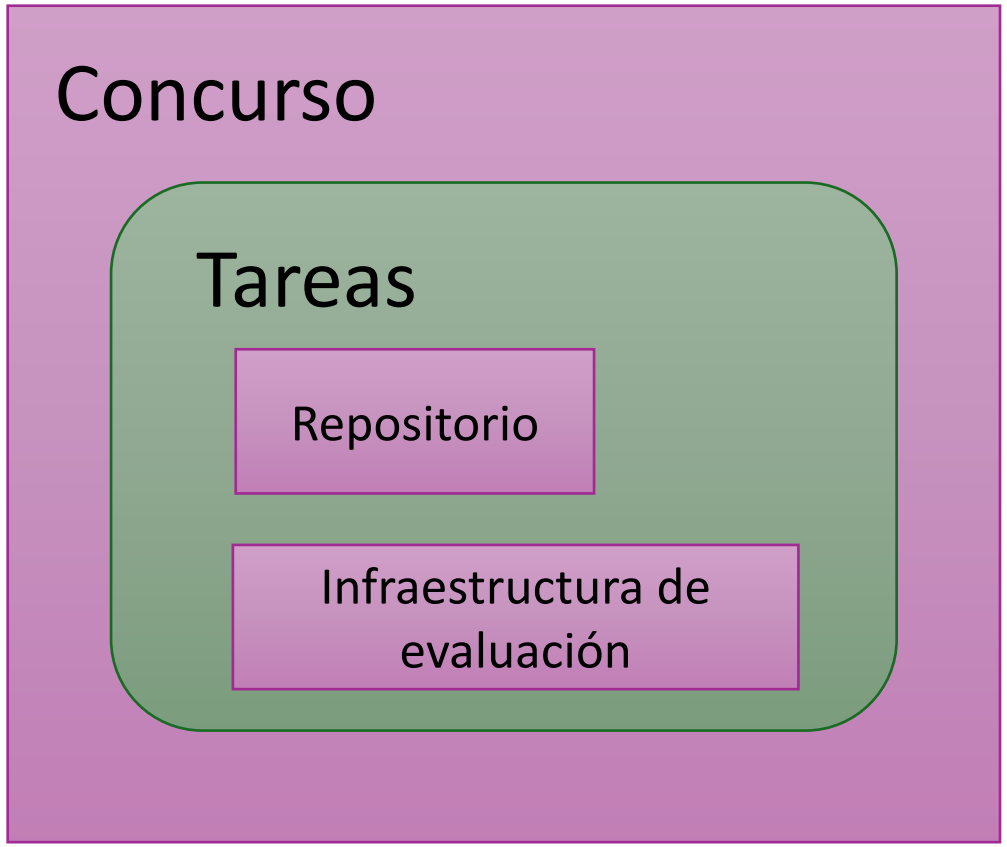
- Semeval
- CLEF
- TREC
- TAC
- IberLEF
- otros.

INDICE

Marco genérico de un concurso



Infraestructura de concursos



Repositorios:

- Github,
- Gitlab,
- Servidores online,
- Otros..

Infraestructura de Evaluación:

- Colab,
- Kaggle,
- Otras...

INDICE

Campañas de evaluación

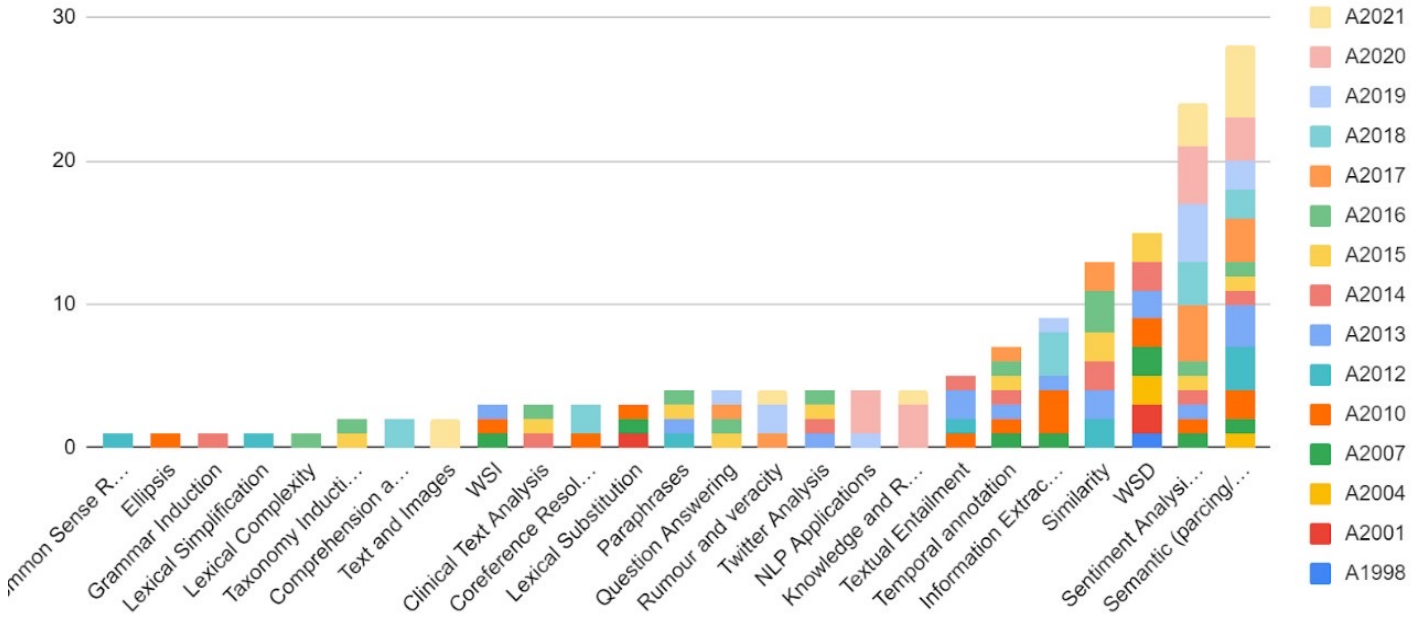
- Semeval
- CLEF
- TAC
- IBERLEF
- …

Campañas de evaluación. SenseEval/SemEval

Ediciones:

- Senseval-1 (1998)
- Senseval-2 (2001)
- Senseval-3 (2004)
- SemEval-2007 (Senseval-4) (2007)
- SemEval-2010 (2010)
- SemEval-2012 (2012)
- SemEval-2013 (2013)
- …
- Semeval-2020 (2020)
- Semeval-2021 (2021)
- Semeval-2022 (2022)
- Semeval-2023 (2023)

Campañas de evaluación. Semeval



Datasets and Corpora

Este marco de evaluación como la mayoría **proporciona** a los investigadores **datasets y corpus** de entrenamiento y evaluación.

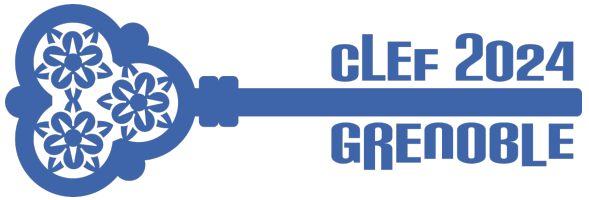
Primero se llamó **SensEval** (Sens de sentido o significado y Eval de evaluación) y luego **SemEval** (Sem de **semántica**).

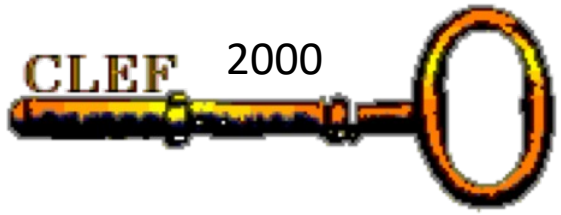
- Pues porque Senseval-1 (1998) y Senseval-2 (2001) se centraron en **tareas de desambiguación semántica** y
- ya **luego** fueron **incorporándo se** otros tipos de **tareas de PLN** lo que dió lugar a un cambio de nombre SemEval que pudiera ampliar el marco de representación de dicho nombre.

Campañas de evaluación. CLEF

CLEF(inicios año 2000) obtiene sus siglas en inglés de **Conference and Labs of the Evaluation Forum**, y formalmente **conocido** como **Cross-Language Evaluation Forum.**

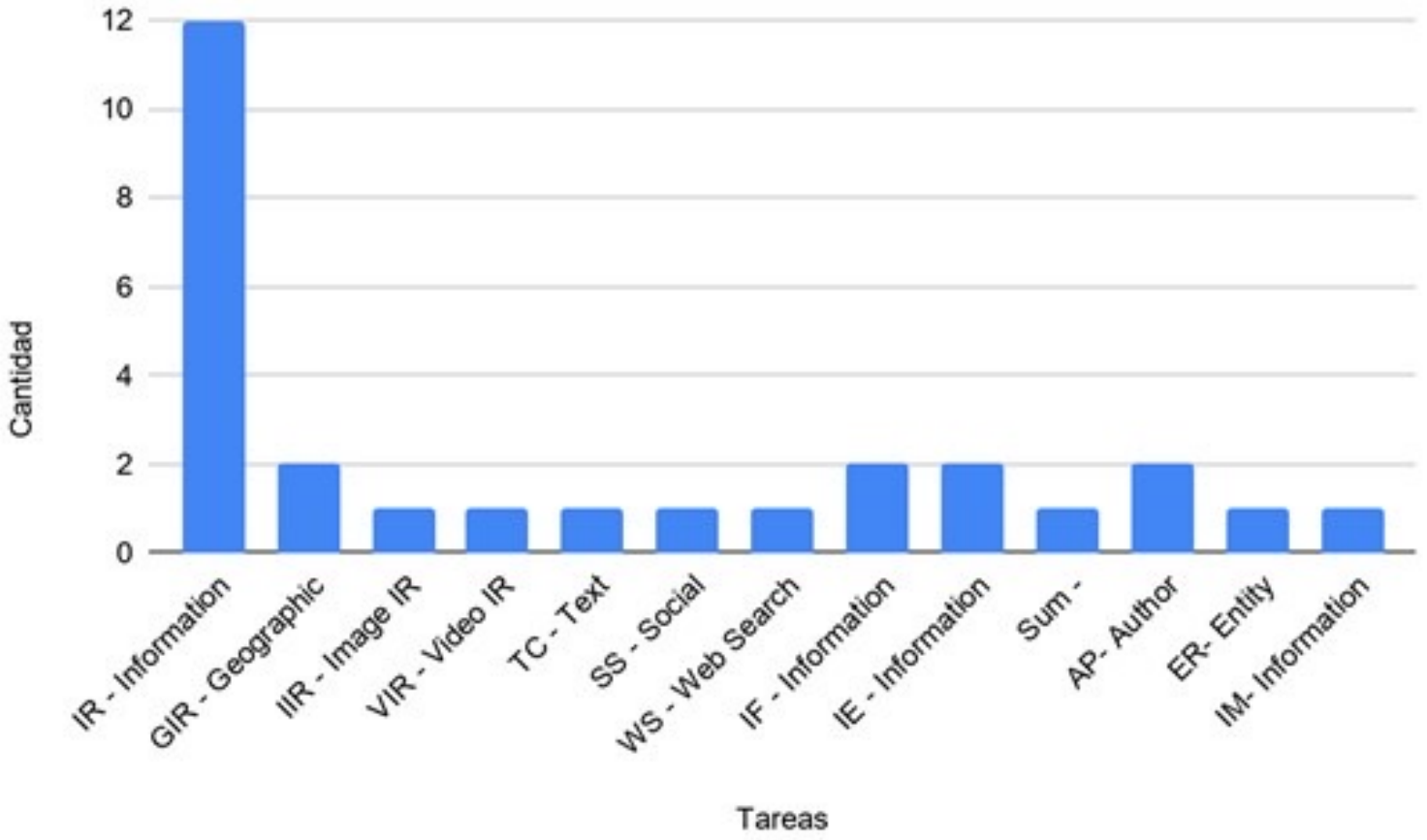
Es un **organismo auto-organizado** cuya misión principal es **promover la investigación**, la innovación y el **desarrollo de sistemas de acceso a la información** con énfasis en la información **multilingüe** y **multimodal** con varios niveles de estructura.





http://www.clef-initiative.eu/

Campañas de evaluación. CLEF



Listado detallado de tareas: https://jaspock.github.io/mtextos2324/bloque3_t2_subaplicaciones-benchmarks.html#clef

Campañas de evaluación. TAC. *Text Analysis Conference*

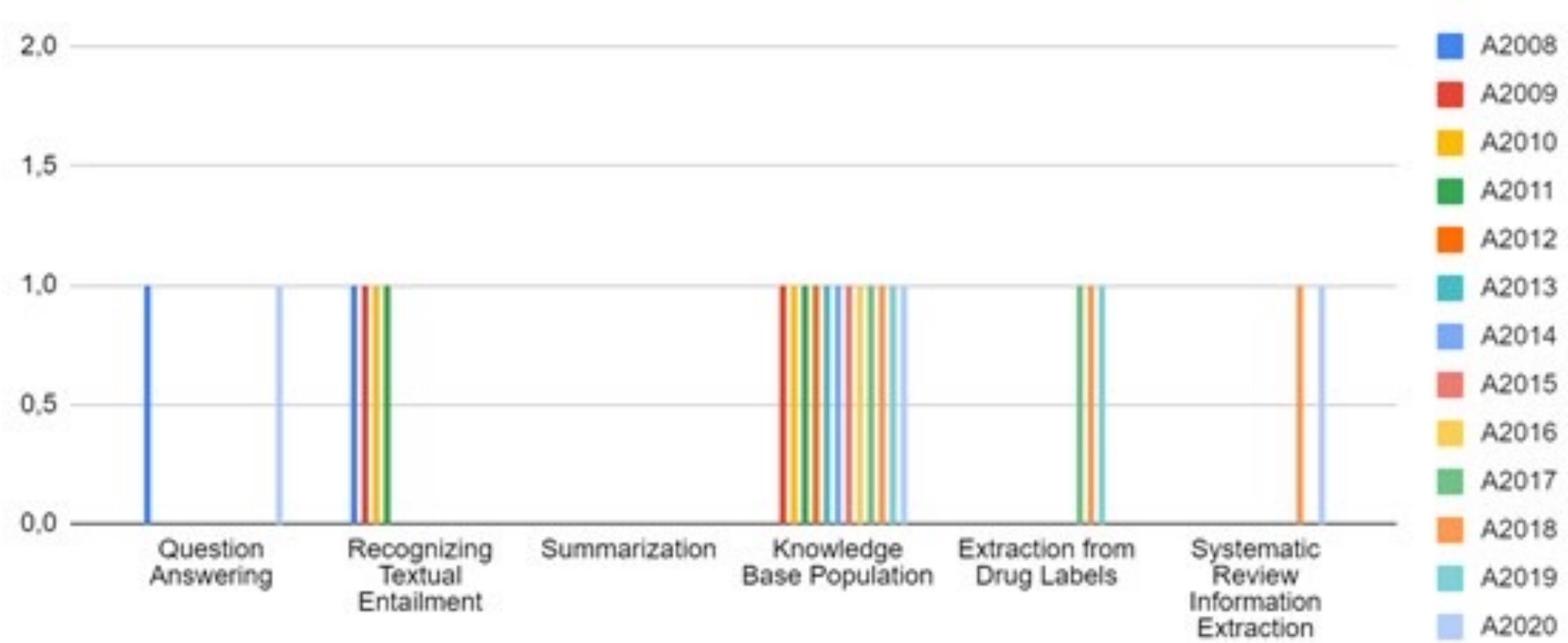
- La **Conferencia de Análisis de Texto** (TAC) se conforma por una serie de **talleres de evaluación** organizados para fomentar la investigación en el **procesamiento del lenguaje natural** y **aplicaciones relacionadas**.
- **Proporciona** una gran **colección de pruebas, procedimientos de evaluación comunes** y un foro para que las organizaciones compartan sus resultados.

Campañas de evaluación. TAC. *Text Analysis Conference*

TAC comprende conjuntos de **tareas conocidas como "pistas**(*tracks*)**"**, cada una de las cuales se **centra en un subproblema particular** del PLN.

Los *tracks* de TAC se centran en las **tareas del usuario final**, pero también incluyen **evaluaciones** de componentes situadas dentro del contexto de estas.

Campañas de evaluación. TAC. *Text Analysis Conference*



Campañas de evaluación. IberLEF: *Iberian Languages Evaluation Forum*

- IberLEF es una **campaña de evaluación** comparativa de sistemas de **Procesamiento de Lenguaje Natural** en **español y otras lenguas ibéricas**.
- Su objetivo es alentar a la comunidad investigadora a **organizar tareas competitivas** de **procesamiento**, **comprensión** y **generación de textos** para definir nuevos desafíos de investigación y establecer nuevos resultados de vanguardia en esos idiomas.

Campañas de evaluación. IberLEF: *Iberian Languages Evaluation Forum*

Ediciones

- TASS 2012-2020: http://tass.sepln.org/
- IberEval 2017: https://sepln2017.um.es/ibereval.html
- IberEval 2018: https://sites.google.com/view/ibereval-2018
- Iberlef2019: https://sites.google.com/view/iberlef-2019
- Iberlef2020: https://sites.google.com/view/iberlef2020/home
- Iberlef2021: https://sites.google.com/view/iberlef2021
- Iberlef2022: https://sites.google.com/view/iberlef2022
- Iberlef2023: https://sites.google.com/view/iberlef-2023
- Iberlef2024: https://sites.google.com/view/iberlef-2024

INDICE

Benchmarks e infraestructuras de evaluación. CodaLab

CodaLab [cuadernos de trabajo, concursos]

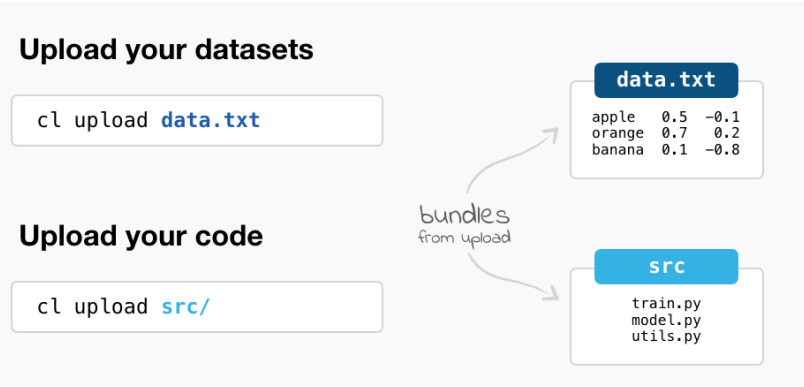
- CodaLab es una **plataforma de código abierto** que proporciona un **ecosistema** para realizar **investigación computacional** de una manera más eficiente, **reproducible y colaborativa**.

  - **Hojas de trabajo**
  - **Concursos**

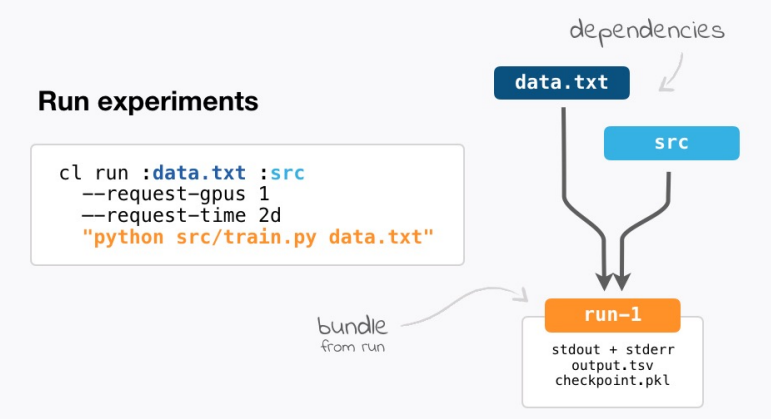
Benchmarks e infraestructuras de evaluación. CodaLab

Hojas de trabajo: permiten capturar **líneas de investigación** complejas de una manera **reproducible** y crear **"documentos ejecutables"**. Se puede utilizar **cualquier formato** de **datos** o **lenguaje** de programación.

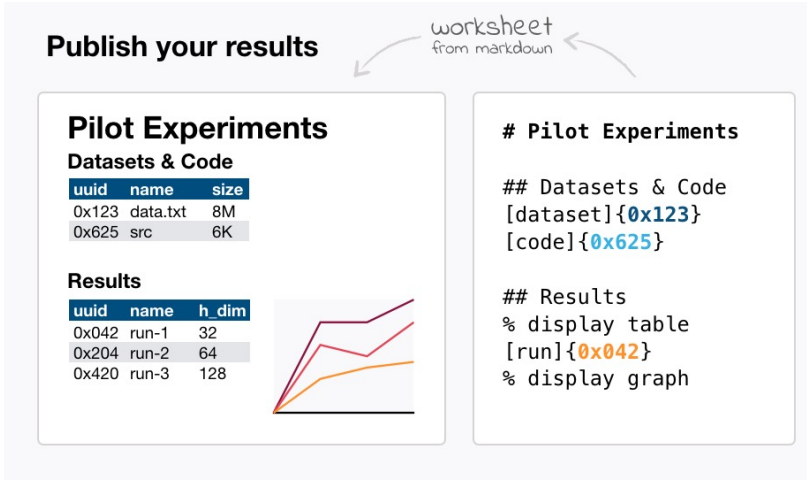
1.



2.

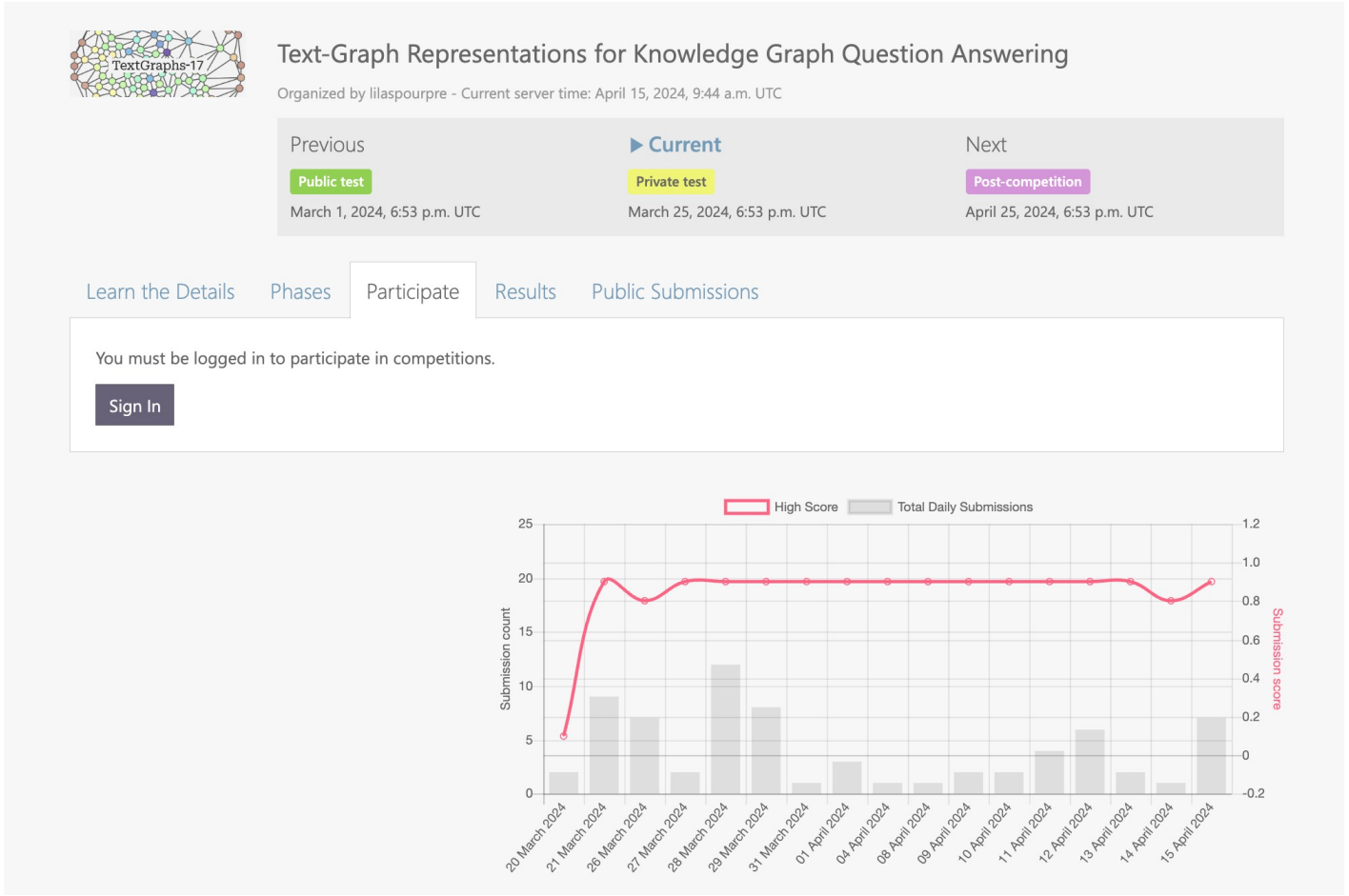


3.



Benchmarks e infraestructuras de evaluación. CodaLab

Concursos (competiciones): Estos sirven para **reunir** a la **comunidad** científica para **abordar los problemas** informáticos y de datos más desafiantes de la actualidad. Puedes **ganar premios** y también crear tu propia competencia.

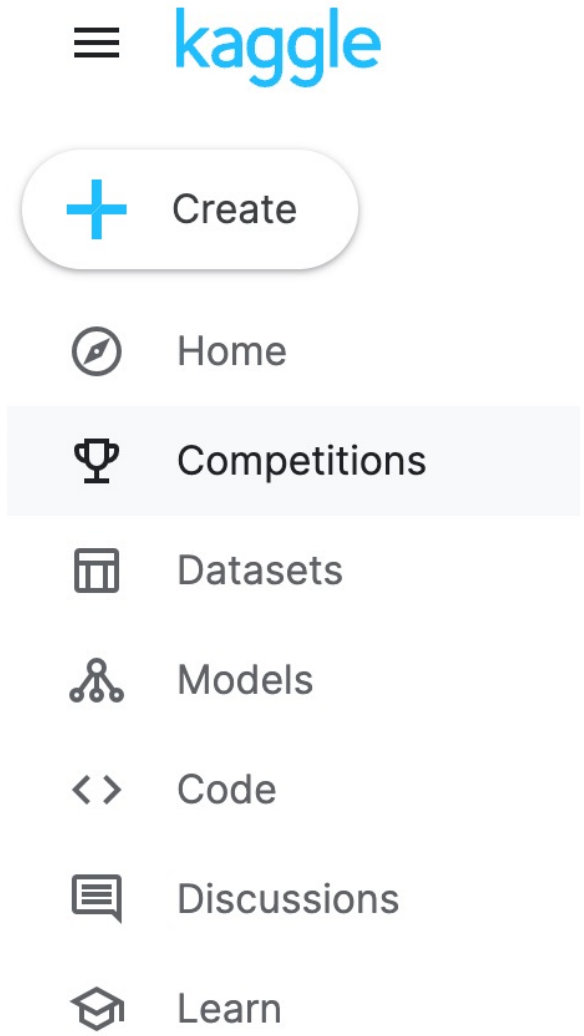


https://codalab.lisn.upsaclay.fr/competitions/18214

Benchmarks e infraestructuras de evaluación. Kaggle

Kaggle [concursos, conjunto de datos, códigos fuente]

- Entidad **subsidiaria de Google LLC**, permite a los usuarios **encontrar y publicar conjuntos de datos, explorar y construir modelos** en un entorno de ciencia de datos basado en la web, trabajar **organizar concursos** para resolver desafíos de ciencia de datos.



https://www.kaggle.com/

Benchmarks e infraestructuras de evaluación. GLUE

GLUE [tabla de rankings, conjunto de datos, códigos fuente]

- El **marco de referencia** de **evaluación** de comprensión del lenguaje general (GLUE) es una **colección de recursos** para **entrenar, evaluar y analizar sistemas de comprensión del lenguaje natural**.

Benchmarks e infraestructuras de evaluación. GLUE

- Un **marco de referencia** (**Benckmark**) de **nueve tareas** de comprensión del lenguaje de oraciones o pares de oraciones basadas en conjuntos de datos existentes establecidos y seleccionados para cubrir una amplia gama de tamaños de conjuntos de datos, géneros de texto y grados de dificultad.
- Un **conjunto de datos** (**datasets**) de diagnóstico diseñado para evaluar y analizar el rendimiento del modelo con respecto a una amplia gama de fenómenos lingüísticos que se encuentran en el lenguaje natural, y

GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

https://gluebenchmark.com/tasks

Benchmarks e infraestructuras de evaluación. GLUE

- Una **tabla de ranking pública** (**leaderboard**) para **reflejar** el **desempeño** en el Benckmark y un **tablero para visualizar el desempeño** de los **modelos** en el conjunto de diagnóstico.

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	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
8	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
9	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
10	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
11	Microsoft D365 AI & MSR AI & GATECH	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
12	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
13	LG AI Research	ANNA		89.8	68.7	97.0	92.7/90.1	93.0/92.8	75.3/90.5	91.8	91.6	96.0	91.8	95.9	51.8
14	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
15	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
16	David Kim	2digit LANet		89.3	71.8	97.3	92.4/89.6	93.0/92.7	75.5/90.5	91.8	91.6	96.4	91.1	88.4	54.6
17	倪仕文	DropAttack-RoBERTa-large		88.8	70.3	96.7	92.6/90.1	92.1/91.8	75.1/90.5	91.1	90.9	95.3	89.9	89.7	48.2
18	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
19	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3

https://gluebenchmark.com/leaderboard

Benchmarks e infraestructuras de evaluación

- El **formato** Benckmark GLUE es **independiente del modelo**, por lo que **cualquier sistema capaz de procesar oraciones y pares de oraciones y producir las predicciones correspondientes es elegible para participar**.
- Las tareas recogidas en el marco de GLUE actualmente ofrecen **rendimientos cercanos al nivel de humanos no expertos**, lo que sugiere un **margen limitado** para **futuras** investigaciones.

Benchmarks e infraestructuras de evaluación

SuperGLUE [tabla de rankings, conjunto de datos, códigos fuente]

- **SuperGLUE**, un **nuevo Benckmark** con el **estilo de GLUE** con un **nuevo conjunto de tareas** de comprensión del idioma **más difíciles**, **recursos mejorados** y **una nueva tabla de clasificación pública**.

Benchmarks e infraestructuras de evaluación

SuperGLUE [tabla de rankings, conjunto de datos, códigos fuente]

- **SuperGLUE**, un **nuevo Benckmark** con el **estilo de GLUE** con un **nuevo conjunto de tareas** de comprensión del idioma **más difíciles**, **recursos mejorados** y **una nueva tabla de clasificación pública**.

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+	6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9

Benchmarks e infraestructuras de evaluación. Huggingface 🤗

- Huggingface 🤗 [conjunto de datos, código fuente, modelos]
- Hugging Face es una **empresa emergente líder en el PLN** con más de mil empresas que utilizan sus bibliotecas de código abierto (específicamente: la **biblioteca Transformers**) en producción. La biblioteca Transformer basada en Python expone las API para usar rápidamente **arquitecturas NLP** como: **BERT** (Google, 2018)

Benchmarks e infraestructuras de evaluación. Huggingface 🤗

Proporciona:

- miles de **modelos previamente entrenados** para realizar múltiples tareas en más de 100 idiomas.
- **API para descargar y usar** rápidamente esos **modelos previamente entrenados**
- **Espacios de trabajo y Endpoints**
- Está **respaldado/integrado** por librerías como PyTorch y TensorFlow

https://huggingface.co/datasets/yelp_review_full

https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions

https://huggingface.co/timpal0l/mdeberta-v3-base-squad2

https://huggingface.co/spaces/ehristoforu/dalle-3-xl-lora-v2

Benchmarks e infraestructuras de evaluación. Extreme

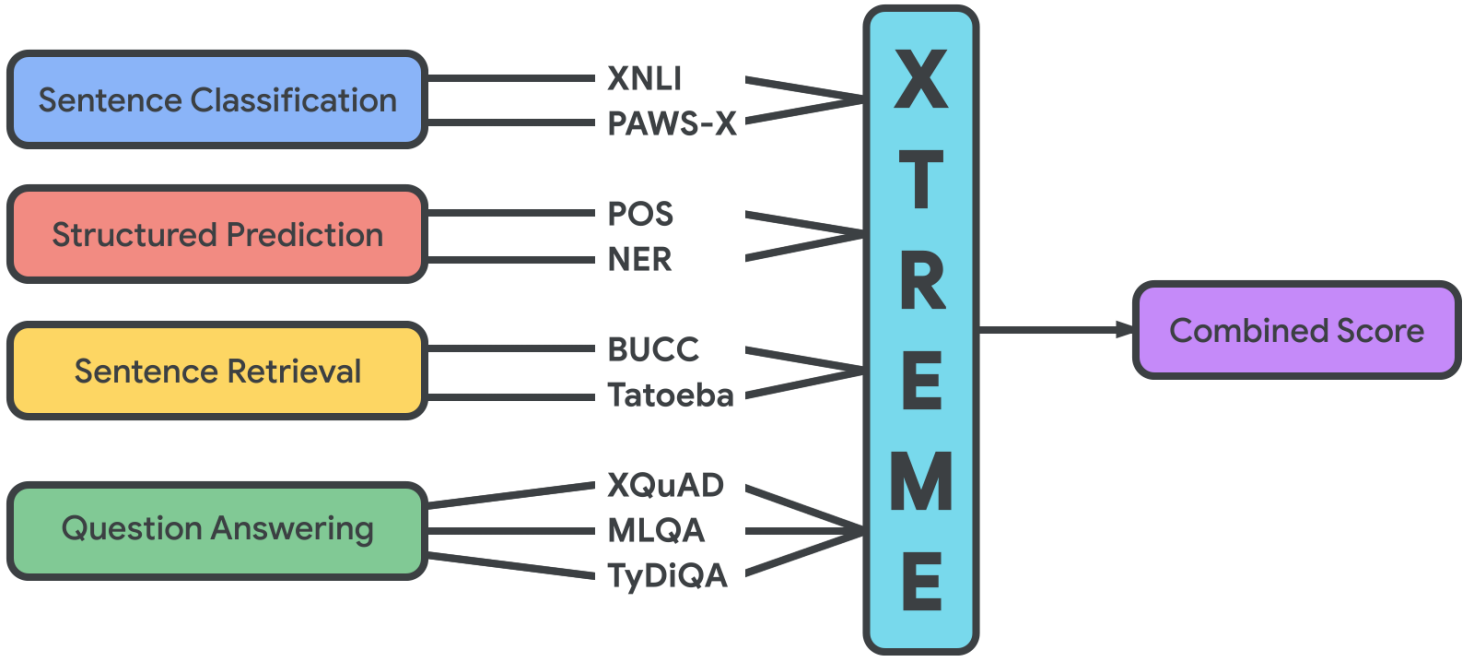
Extreme [tabla de rankings, conjunto de datos, códigos fuente. papers, modelos]

- **TRansfer Evaluation of Multilingual Encoders** ([Extreme]) es un benckmark para la **evaluación** de la capacidad de **generalización** entre **idiomas de modelos multilingües** previamente entrenados.

https://github.com/google-research/xtreme

Benchmarks e infraestructuras de evaluación. Extreme

- Cubre **40 idiomas** tipológicamente diversos (que abarcan 12 familias de idiomas) e incluye **nueve tareas** que colectivamente requieren razonamiento sobre diferentes niveles de sintaxis y semántica.



https://github.com/google-research/xtreme

INDICE

1. Benchmarks. ¿Qué es?
2. Marco genérico
3. Campañas de Evaluación
4. Benchmarks e infraestructuras de evaluación
5. Aplicaciones Específicas
6. Bibliografía

Benchmarks: Aplicaciones del PLN

- ***Information Retrieval*** (Recuperación de Información)
- ***Information Extraction*** (Extracción de Información)
- ***Question Answering*** (QA, Respuestas a preguntas)
- ***Text Clasification*** (Clasificación textual)
- ***Natural Language Generation*** (NLG, Generación de lenguaje natural)
- ***Text Sumarisation*** (Generación de resúmenes)
- ***Machine Translation*** (Traducción Automática)

Benchmarks: Aplicaciones del PLN. Recuperación de Información

Benchmarcks

- Papers With Code: https://paperswithcode.com/task/information-retrieval
- Kaggle: https://www.kaggle.com/search?q=%22information+retrieval%22+in%3Acompetitions
- TREC: https://trec.nist.gov/

Repositorios de Código

- Kaggle: https://www.kaggle.com/search?q=%22information+retrieval%22+in%3Anotebooks

Conjuntos de datos

- TREC: https://trec.nist.gov/data.html
- Kaggle: https://www.kaggle.com/search?q=%22information+retrieval%22+in%3Adatasets

Artículos

- TREC: https://trec.nist.gov/pubs.html

Benchmarks: Aplicaciones del PLN. Extracción de Información

Benchmarcks

- Papers With Code: https://paperswithcode.com/task/information-extraction
- Kaggle: https://www.kaggle.com/search?q=%22information+extraction%22+in%3Acompetitions

IberLEF (Español):

  - eHealth-KD: https://knowledge-learning.github.io/ehealthkd-2020/
  - Named Entity Recognition and Relation Extraction for Portuguese: https://www.inf.pucrs.br/linatural/wordpress/iberlef-2019/
  - Cantemist: https://temu.bsc.es/cantemist/

Repositorios de Código

- Kaggle: https://www.kaggle.com/search?q=%22information+extraction%22+in%3Anotebooks

Conjuntos de datos

- Kaggle: https://www.kaggle.com/search?q=%22information+extraction%22+in%3Adatasets

Artículos

- Papers With Code: https://paperswithcode.com/task/information-extraction

Benchmarks: Aplicaciones del PLN. QA, Respuestas a preguntas

Benchmarcks

- Papers With Code: https://paperswithcode.com/task/question-answering
- Kaggle: https://www.kaggle.com/search?q=Question+Answering+in%3Acompetitions
- CodaLab: https://competitions.codalab.org/competitions/?q=Question+Answering

Repositorios de Código

- Papers With Code: https://paperswithcode.com/task/question-answering
- Kaggle: https://www.kaggle.com/search?q=Question+Answering+in%3Anotebooks

Conjuntos de datos

- Huggingface: https://huggingface.co/datasets?filter=task_categories:question-answering
- Kaggle: https://www.kaggle.com/search?q=Question+Answering+in%3Adatasets

Artículos

- Papers With Code: https://paperswithcode.com/task/question-answering

Benchmarks: Aplicaciones del PLN. Clasificación textual

Benchmarcks

- Papers With Code: https://paperswithcode.com/task/information-extraction
- Kaggle: https://www.kaggle.com/search?q=%22information+extraction%22+in%3Acompetitions
- IberLEF (Español):
  - eHealth-KD: https://knowledge-learning.github.io/ehealthkd-2020/
  - Named Entity Recognition and Relation Extraction for Portuguese: https://www.inf.pucrs.br/linatural/wordpress/iberlef-2019/
  - Cantemist: https://temu.bsc.es/cantemist/

Repositorios de Código

- Kaggle: https://www.kaggle.com/search?q=%22information+extraction%22+in%3Anotebooks

Conjuntos de datos

- Kaggle: https://www.kaggle.com/search?q=%22information+extraction%22+in%3Adatasets

Artículos

- Papers With Code: https://paperswithcode.com/task/information-extraction

Benchmarks: Aplicaciones del PLN. Generación de lenguaje natural

Benchmarcks

- Papers With Code: https://paperswithcode.com/task/text-generation
- WebNLG: https://webnlg-challenge.loria.fr/
- NLP Progress: http://nlpprogress.com/english/data_to_text_generation.html

Repositorios de Código

- Papers With Code: https://paperswithcode.com/task/text-generation
- Kaggle: https://www.kaggle.com/search?q=%22text+generation%22+in%3Anotebooks

Conjuntos de datos

- RotoWire: https://github.com/harvardnlp/boxscore-data/blob/master/rotowire.tar.bz2
- Kaggle: https://www.kaggle.com/search?q=%22text+generation%22+in%3Adatasets

Artículos

- Papers With Code: https://paperswithcode.com/task/text-generation

Benchmarks: Aplicaciones del PLN. Generación de resúmenes

Benchmarcks

- Papers With Code: https://paperswithcode.com/search?q_meta=&q=summarisation
- LaySumm: https://ornlcda.github.io/SDProc/sharedtasks.html; https://competitions.codalab.org/competitions/25516

Repositorios de Código

- Papers With Code: https://paperswithcode.com/search?q_meta=&q=summarisation

Conjuntos de datos

- Huggingface: https://huggingface.co/datasets?filter=task_ids:summarization

Artículos

- Papers With Code: https://paperswithcode.com/search?q_meta=&q=summarisation

Benchmarks: Aplicaciones del PLN. Traducción Automática

Benchmarcks

- Papers With Code: https://paperswithcode.com/task/machine-translation
- Codalab: https://competitions.codalab.org/competitions/?q=machine+translation

Repositorios de Código

- Papers With Code: https://paperswithcode.com/task/machine-translation
- Kaggle: https://www.kaggle.com/search?q=machine+translation+in%3Anotebooks+tag%3Alanguages

Conjuntos de datos

- Huggingface: https://huggingface.co/datasets?filter=task_ids:machine-translation
- PapersWithCode: https://paperswithcode.com/datasets?q=machine+trasnlation&v=lst&o=match&mod=texts&page=1

Artículos

- Papers With Code: https://paperswithcode.com/task/machine-translation

INDICE

1. Benchmarks. ¿Qué es?
2. Marco genérico
3. Campañas de Evaluación
4. Benchmarks e infraestructuras de evaluación
5. Aplicaciones Específicas
6. Bibliografía

Bibliografía

[1] https://www.acl-bg.org/proceedings/2017/RANLP%202017/pdf/RANLP032.pdf

[2] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., … & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10), e0203794.

[3]https://dac.cs.vt.edu/research-project/semi-supervised-learning-cyberbullying-harassment-patterns-social-media/

[4] https://www.iic.uam.es/inteligencia/la-importancia-tener-ner/

[5] https://temu.bsc.es/codiesp/

[6] Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, Manuel Palomar: Fighting post-truth using natural language processing: A review and open challenges. Expert Syst. Appl. 141 (2020)

Competiciones y Campañas de Evaluación

Yoan Gutiérrez Vázquez

ygutierrez@dlsi.ua.es

Universitat d'Alacant
Universidad de Alicante
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UJI UNIVERSITAT JAUME I
valgrAI