

Introducción al Procesamiento del Lenguaje Natural

Marta Guerrero Nieto. Coordinadora en el Instituto de
Ingeniería del conocimiento www.iic.uam.es

Análisis semántico

- Es la extracción y la formalización del **significado de un texto** y su conversión a un valor computacional.
- Los proyectos de PLN más demandados consisten en un enriquecimiento semántico.
- Es el análisis más avanzado (objetivo último del PLN)
- Hay algunas tareas que requieren cierto nivel de procesamiento previo (split, token, POS)
- Algunas tareas más destacadas son **análisis del sentimiento y detección de entidades**

Análisis del sentimiento

- Opinion mining/ sentiment analysis
 - Identificar, clasificar y cuantificar valoraciones o información subjetiva.
 - Analizar información de cliente, encuestas de satisfacción, redes sociales, etc.
-
- **Objetivo:** determinar la actitud, reacción emocional, estado afectivo de un hablante hacia un tema, evento, convección,...
 - Análisis de la opinión
 - Análisis de emociones
 - Análisis de la concienciación

Dos aproximaciones

➤ **Aproximación léxica y capas de análisis**

□ Se construye más rápido, se necesita conocimientos lingüístico expertos, se apoya en las otras capas de análisis lingüístico. Generaliza peor.

➤ **Aproximación corpus y aprendizaje automático**

□ Se construye más lento, se necesita mucho texto anotado (corpus anotado), no necesita de otras capas. Mayor generalización

Aproximaciones de análisis del sentimiento

□ Aproximación léxica

- **Morir** – palabra connotaciones
 - ✓ **Me muero** por ese vestido
 - ✓ **Ha muerto** una persona en un accidente a la salida de un colegio
 - ✓ Quiero que **mueras** de forma lenta
- **Riesgo**- palabra connotaciones
 - ✓ La prima de **riesgo** está en los 120 puntos
 - ✓ Comer fruta reduce el **riesgo** de padecer enfermedades
 - ✓ Eres un **riesgo** para la sociedad
- **Increíble**- palabra connotaciones
 - ✓ Es **increíble** que nadie le toque un pelo a ese farsante
 - ✓ El equipo de futbol es **increíble**.

Definición de criterios de análisis del sentimiento

C1.-Neutro

Reglas positivas

C1.P1. Se considera neutro aquellos comentarios que en su interpretación global no denotan ninguna valoración, esto es, no se puede decir que sea positivo ni negativo. Son comentarios habitualmente que describen una situación pero no la valoran.

- **Ejemplo:** Esta tarde participamos en la mesa redonda sobre "Tecnologías emergentes y privacidad" [#MañanaEmpiezaHoy](#) en [@feriademadrid](#)

C1.P2. En algunas ocasiones, no hay una dominancia clara en el comentario que expresa positividad o negativas, en estos casos se podrá poner neutro.

Ejemplo: te han echo tan mierda que cuando alguien te dice algo bonito te piensas q te están vacilando

C2.- Positivo

Reglas positivas

C2.P1. Se considera positivo aquellos comentarios que en su interpretación global son positivos.

- **Ejemplo:** No tengo palabras para describir lo que sentí y lo que siento, estuviste increíble en absolutamente todo ,te quedó super bonito. Annieeee Diosss mioo eres I-N-C-R-E-Í-B-L-E.

C3.- Negativo

Reglas positivas

C3.P1. Se considera negativo aquellos comentarios que en su interpretación global son negativos.

- **Ejemplo:** Y con este termino! No quiero saber nunca mas nada de [@somosyoigo](#) [@masmovil](#) Que os vaya bonito! Aunque con el trato que daís es difícil

Aproximaciones de análisis del sentimiento

□ Aproximación de machine learning

- Me caes muy bien -Tienes que jugar más partidas al lol con Russel y conmigo - Por qué tan Otako, deja de ser otako -Haber si me muero **[Positivo]**
- Quiero mogollón a @AlbaBenito99 pero sobretodo por lo rápido que contesta a los wasaps **[Positivo]**
- @toNi_end seria mejor que dejaran de emitir esa basura ya hay que evolucionar para bien y eso **[Negativo]**
- #TweetLikeThe2000s Esto si que era una buena serie y lo demás es tontería **[Positivo]**
- Pesimo servicio del banco de VENEZUELA su pagina caida durante dias .Hasta cuando ese atroz servicio que perjudica a millones .Incapacidad ,saboteo .Corrijan renuncien si no pueden **[Negativo]**

Análisis de emociones



Seguir

Aun estoy alucinando con lo innecesarias que han sido las valoraciones a Ana de la tipa esa. PERO QUE COÑO TE CREES?!

#OTGala11

16:38 - 15 ene. 2018

Ira,
enfado



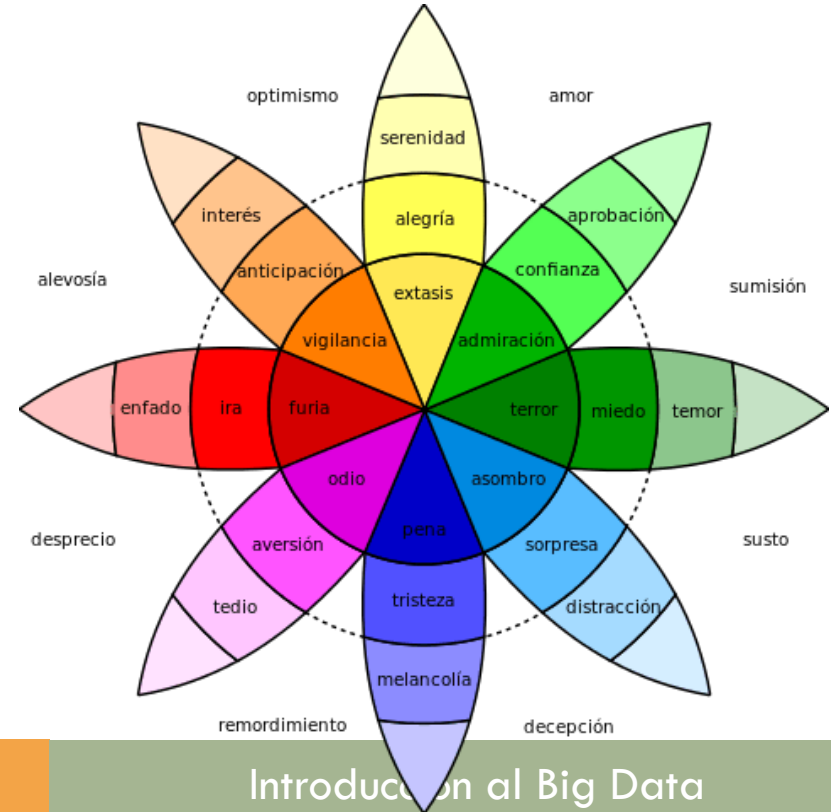
Seguir

Es la noticia del año!!! Muy emocionado y orgulloso, no hay nadie mejor que @RobertoLealG para @OT_Oficial Gracias @tinetr #RoberOT

Euforia,
enorgullecerse



Modelo psicológico de emociones humanas PUTCHIK



Ejercicio 4: Análisis del sentimiento

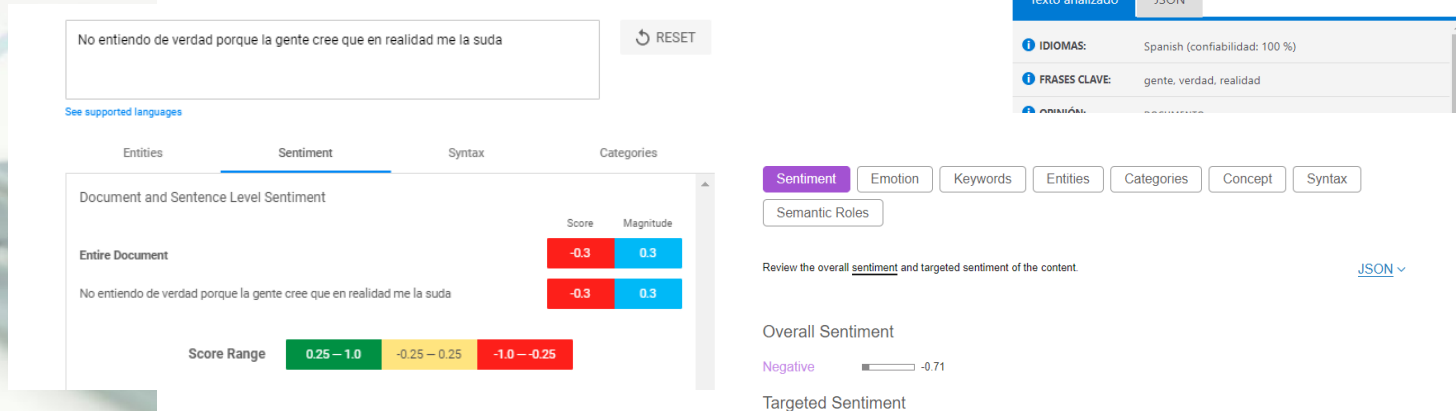
Anota la opinión (positiva, negativa y neutra) de las siguientes frases. Primero anota de forma individual y luego contrasta con un compañero.

- 1. @jonoro96 te mandaria a comprarte un burro, pero no creo que hayan tiendas abiertas ahora
- 2. @hywzz la voz de María al final me mata JAJAJAJAJAJ quilla y yo que
- 3. @misscatingcat Bueno, al menos has acabado la noche con una hermosa vista Ami tambien me gusta verlos por la noche ^^
- 4. Aunque pensaba que no podría hacerlo hasta el mes de octubre, el lunes próximo volveré a ver a mi familia de #Mataró *.* #felicidad
- 5. Gracias por los ánimos! He dormido un poco más, me he tomado una pastilla y ha bajado a 36 y medio, pero aún me noto rara
- 6. @angeleta1973 @elespanolcom No he visto su DNI o Pasaporte pero presume de serlo aunque es muy,demasiado,peculiar y poco caritativa
- 7. @prxttytimebomb @emimartinez_98 Si es que odio a muerte a la nueva Harley que quieres
- 8. Confirmao tengo el pie roto, he notado como el puto hueso se movía
- 9. @lmg muchísimas gracias Luis Tú también eres un ejemplo de constancia y buen trabajo, espero que os vaya todo genial!
- 10. Tengo una adicción muy seria al burger King necesito ayuda profesional
- 11. @fonzeff @malenko_ @girotix Eso me temo, mucho calor y mucha gente por todos lados, lo malo de viajar en agosto
- 12. @Jordicanal64 Pues si no fichan "lo que sea" además de un ridículo espantoso igual QSF nos deja plantados...esperemos que no

Ejercicio 5: Análisis del sentimiento

- **Compara el resultado con las siguientes herramientas comerciales (textos en slide siguiente)**

- Watson: <https://natural-language-understanding-demo.mybluemix.net>
- Google: <https://cloud.google.com/natural-language/>
- Meaning Cloud: <https://www.meaningcloud.com/es/demos/demo-analitica-textos>



Ejercicio 5: Análisis del sentimiento

□ **Compara el resultado con las diferentes herramientas comerciales de análisis de opinión:**

1. España no está despoblada de cultura, es riquísima. Pero no parece interesar, su valor es inestimable.
2. Dejar caer Bankia es la mejor opción y no tiene coste para los ciudadanos.
3. Casi ningún medio ha dicho q el ABC repartió propaganda de HAZTE OÍR el DOMINGO ¡QUÉ RARO! Y pedían dinero...
4. @pepephone buenas tardes! Se puede cambiar una tarifa dentro de la parte personal de la web, si es así, cómo? Muchas gracias. Un saludo.
5. No pueden ser +cutres y +cobardes y bajunos! resulta q no le dejan ni escribir ? Son una pandilla de impresentables y no lo podemos permitir
6. Creo que #yoigo me engaña. No puede ser que esto vaya tan lento el día 2 con la infinita
7. Pues el AVE y mi bolsillo son enemigos eternos
8. A partir del día 11. La Escuela está en obras. Han cortado el teléfono. Desgraciadamente el personal de la Escuela no puede hacer más de lo que hace. Lo sentimos.
9. Mejoras que pedíamos a gritos :) Muy bien!
10. El Populismo mediático se supera. El País ya cuenta con sección de lectores disconformes con Podemos. Cartas al director en portada.

Detección de entidades

- Tarea de PLN de los 90
- Recoge dos tareas: identifica entidad y distingue el tipo de entidad
- Detección de entidades (**NER**): personas, organizaciones y localizaciones (**3 etiquetas**)
- (**7 etiquetas**): Direcciones, colectivos, monedas, fechas, personas, organizaciones, localizaciones
- Muy ligado al idioma y dominio
- Se entrenan modelos con corpus anotados y gazzeters (diccionarios léxicos).

Definición de las entidades

- ❑ **Localización.** Se asocian habitualmente con puntos geográficos definidos pero qué pasa con algunas localizaciones de interés turístico?
 - ❑ Camp Nou
 - ❑ La puerta de brandenburgo
- ❑ **Persona.** Se asocia habitualmente con un nombre de persona pero qué pasa con los personajes de ficción o cargos/nombramientos:
 - ❑ El duque de Barcelona
 - ❑ Papá Noel /Batman
- ❑ **Organizaciones.** Se asocian habitualmente con empresas pero qué pasa con instituciones o marcas muy conocidas?:
 - ❑ Iphone 6
 - ❑ Museo del Prado

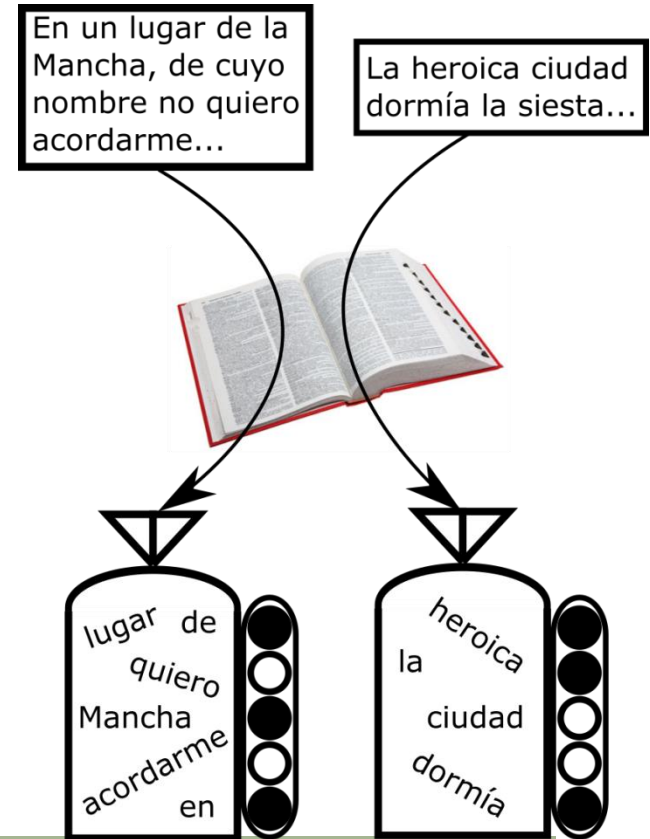
Ejercicio 6: Detección de entidades

- **Compara el resultado con las diferentes herramientas comerciales de detección de entidades. Son las mismas herramientas que ejercicio 5.**

1. Casado mantendrá la guerra total al Gobierno en base a sondeos que constatan el regreso de votantes que estaban en Vox
2. Investigadores españoles rebajan la esperanza de que el verano pare los pies a la COVID-19
3. Un día luchando contra el coronavirus en el Hospital de La Princesa de Madrid
4. La macrocausa fallida del agujero de 130 millones de la Fórmula 1 en Valencia: batalla de la Generalitat contra Anticorrupción para que no cierre dos piezas más
5. La Xunta aprovecha el estado de alarma para acelerar la tramitación ambiental de proyectos contestados socialmente
6. Cómo limpiar fotos y vídeos de WhatsApp sin borrar chats
7. Facebook y Twitter llevan un año gestionando en secreto peticiones para borrar contenidos de otros usuarios
8. Madrid pretende que el Gobierno financie sus menús de Telepizza y el Ministerio le pide un proyecto más "saludable"
9. Volkswagen, obligada a pedir perdón por una publicidad acusada de racista
10. Reino Unido y xenofobia: "Es evidente que el Brexit ha tenido una influencia en cómo los jóvenes tratan el tema de la inmigración"
11. España, único país de Europa donde decenas de curas se saltan la ley para officiar misas con público
12. Santiago Cantera, prior del Valle de los Caídos: "Lo que realmente molesta no es Franco, es la comunidad benedictina"

Extracción de características del texto: Bag of Words

- Las palabras que se utilizan en un texto dan **mucha información**
 - **Temática del texto:** vocabulario específico al tema, tecnicismos, ...
 - **Autor:** nivel cultural, región, género, edad, ..
 - Sentimientos del autor (levemente): emociones, opinión sobre el tema tratado...
- Bag of Words
 - Crear un **diccionario** con todas las palabras de un corpus referencia: N
 - Para cada texto a procesar, construir un **vector binario** de N entradas que refiere a cada palabra del diccionario, y toma el valor:
 - 1 si la palabra correspondiente del diccionario está presente en el texto
 - 0 si la palabra correspondiente no está presente en el texto



Problemas de bag of words

	de	gallina	huevos	la	los
los huevos de la gallina	X	X	X	X	X
la gallina de los huevos	X	X	X	X	X
	baño	el	en	me	río
me baño en el río	X	X	X	X	X
me río en el baño	X	X	X	X	X

- Bag of Words desecha la estructura del texto original, perdiendo significado
 - Relaciones sustantivo – modificador
 - Relaciones de causalidad
 - Imposible realizar desambiguaciones de palabras con varios significados

n-gramas

- Bag of Words o frecuencias de conjuntos de n tokens adyacentes
- Normalmente menos costoso que n-gramas de caracteres
 - En un diccionario de W palabras, la gran mayoría de las posibles $O(W^n)$ combinaciones de tokens nunca aparece en el corpus
- Permite mantener información parcial del contexto de cada token
 - Si esto es importante, puede mejorar a Bag of Words

	de la	de los	gallina de	huevos de	la gallina	los huevos
los huevos de la gallina	X			X	X	X
la gallina de los huevos		X	X		X	X

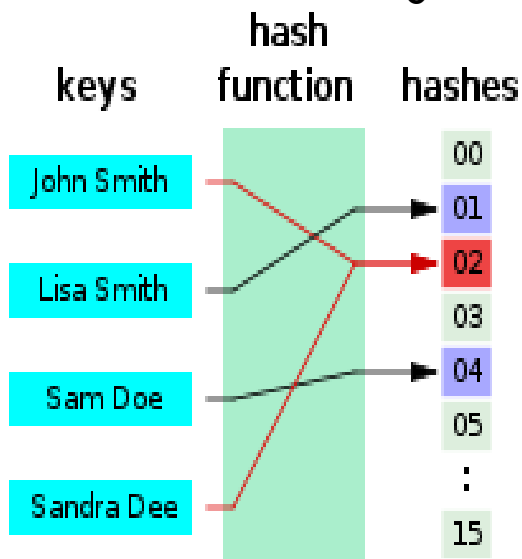
Un problema: palabras fuera del vocabulario

- Un problema importante de los métodos basados en diccionario (bag of words, n-gramas) las representaciones vectoriales se construyen en base a un diccionario fijo de palabras.
- Si más adelante nos encontramos con palabras antes no vistas, no tenemos forma de codificarlas correctamente como vectores, por lo que es necesario recalcular el diccionario.

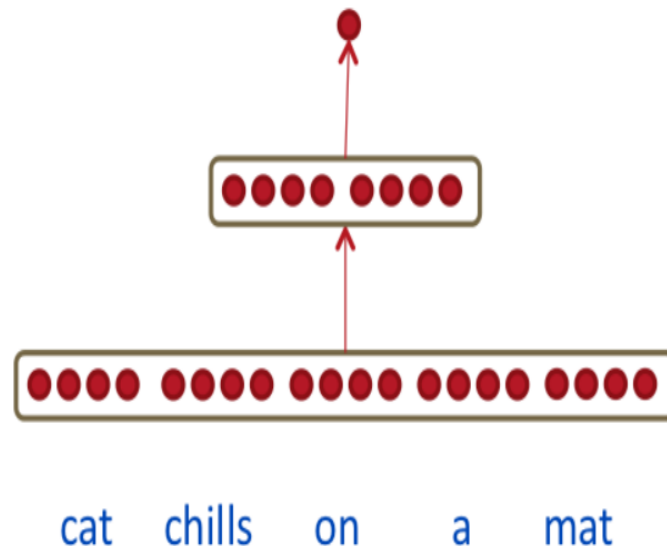
	de la	de los	gallina de	huevos de	la gallina	los huevos
los huevos de la gallina	X			X	X	X
la gallina de los huevos		X	X		X	X
el zorro rojo						

Vectorización de palabras

- No supervisado/ corpus grandes de texto no anotado
- Vectorización es una representación numérica de cada palabra (300 números). Cada palabra tiene un embedding

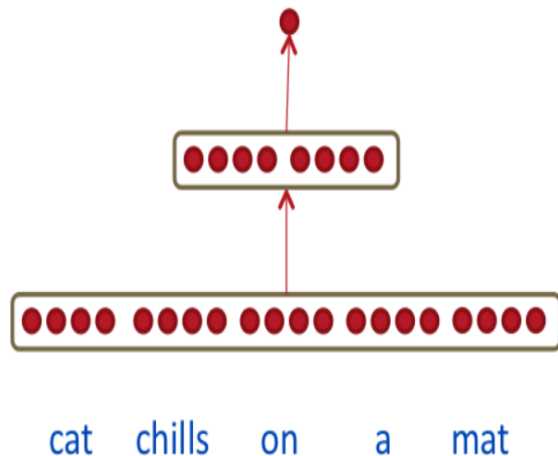


Hashing Vectorizer



word2vec

Word2vec



word2vec

Modelo de machine learning, 2013.

transforma cada palabra de un texto en un vector de números reales de longitud fija. Estos vectores codifican información sintáctica y semántica muy útil para modelos de clasificación o similitud.

Se entrena una red neuronal que aprende a predecir palabras del corpus en base a sus vecinas. Como patrones negativos se usan frases con cambios aleatorios en las palabras.

Vi un **gato** negro corriendo → +

[1.2, 3.4, -1.2, 6.7, 5.4, 3.7, 9.1, -0.3, 1.7, -7.4]

Vi un **seta** negro corriendo → -

<https://code.google.com/archive/p/word2vec/>

Word2vec

¿Ves cómo las palabras "man" y "woman" son más similares entre sí que cualquiera de ellas con "king"?

"king"



"Man"



"Woman"

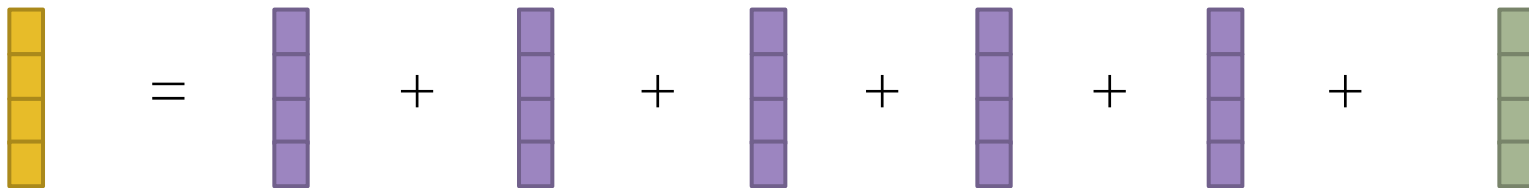


<http://jalammar.github.io/illustrated-word2vec/>

Fasstext

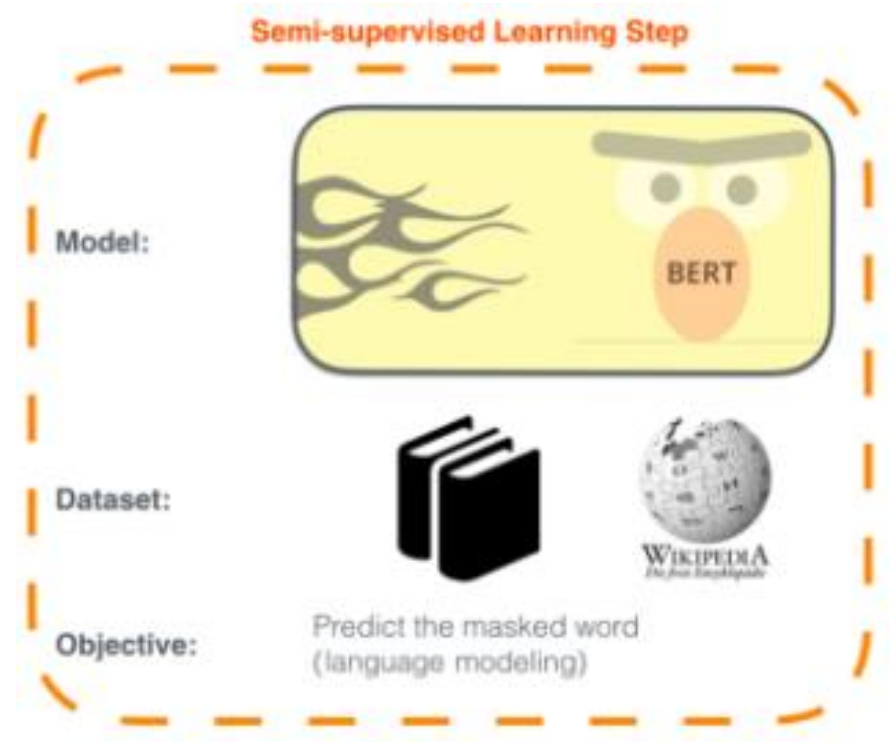
- Representa mejor que los anteriores la estructura interna de las lenguas, sobre todo en aquellas que tienen muchas morfología, como el español.
- El vector representa un token como la suma de todos los subtokens y hay uno especial que representa todo el token completo.

where = <wh + whe + her + ere + re> + <where>



Bert: modelos pre-entrenados

- Es una red neuronal basada en la arquitectura transformer. El modelo se entrena con un corpus de 3.300 millones de palabras (800m de palabras de BooksCorpus y Wikipedia (2.500M de palabras).
- Bert es capaz de descubrir relaciones y secuencias dentro de la frase para una entidad, es donde el concepto transformer marca la diferencia.



Modelos de lenguaje - BERT

1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:



Dataset:



Objective:

Predict the masked word
(language modeling)

2 - **Supervised** training on a specific task with a labeled dataset.

Supervised Learning Step

Classifier

75% Spam
25% Not Spam

Model:
(pre-trained
in step #1)



Dataset:

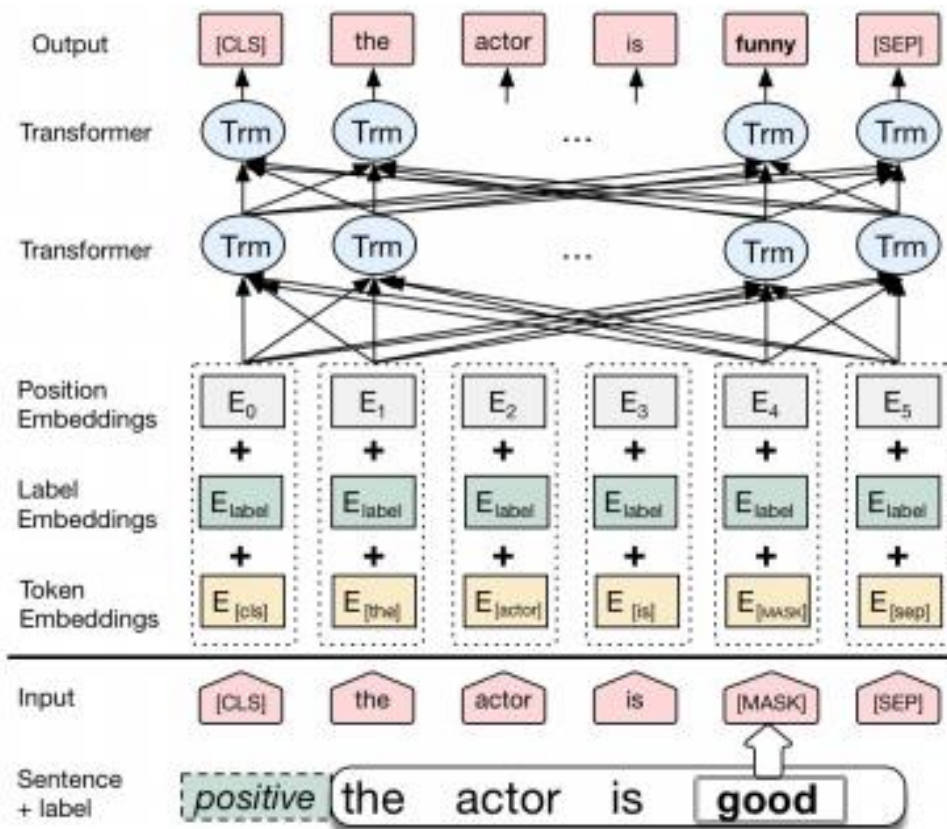
Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached...	Not Spam

Jay Allamar – The Illustrated BERT, ELMo and co - <https://jallamar.github.io/illustrated-bert/>
Devlin et al – BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
BERT pre-trained models: <https://github.com/google-research/bert>

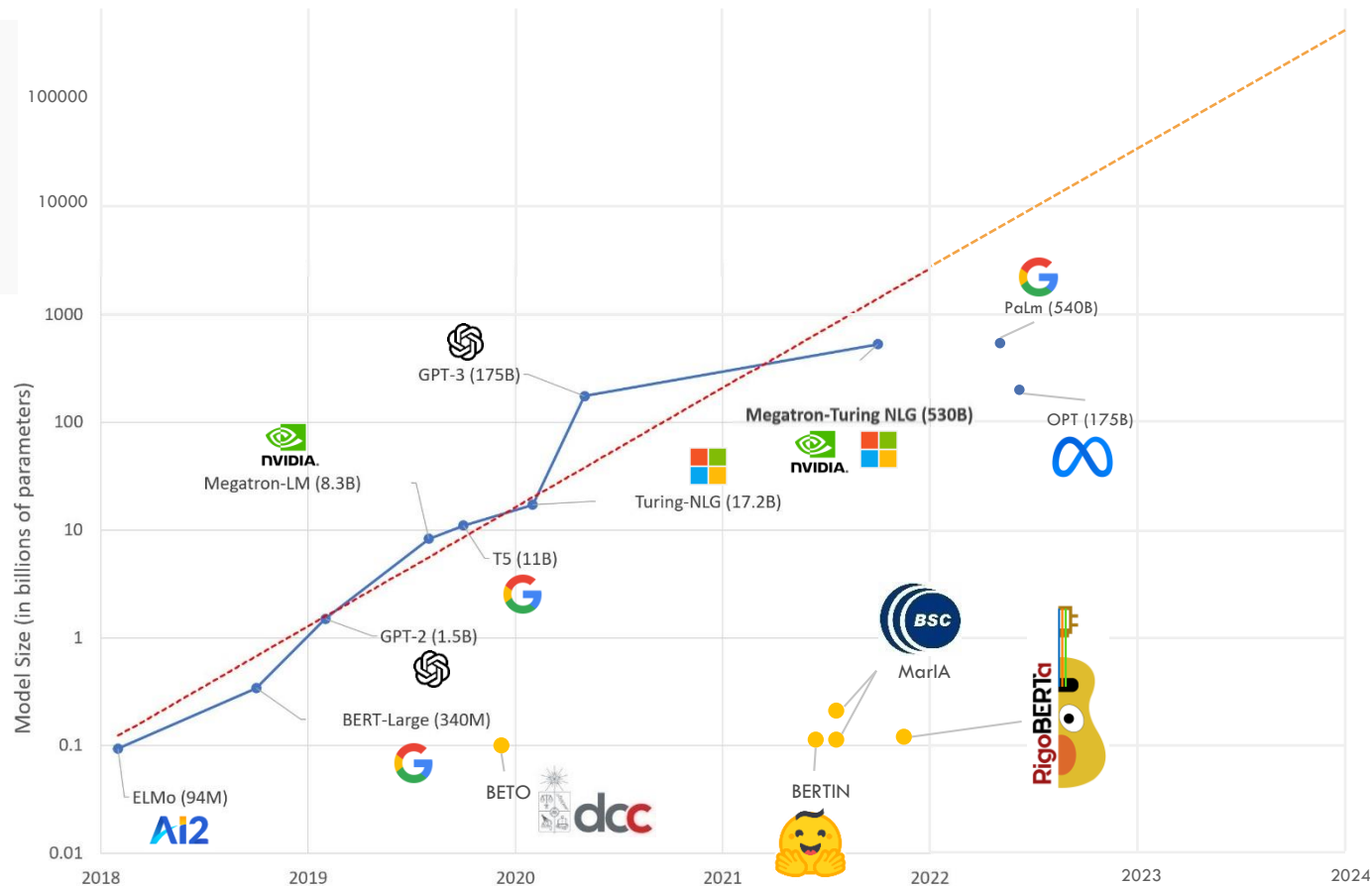
Bert II

Se calcula de forma bidireccional
(de derecha a izquierda
y de izquierda a derecha)

Se añade una etiqueta MASK
a una palabra, que el modelo
tiene que completar.



Modelos del lenguaje



<https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/>

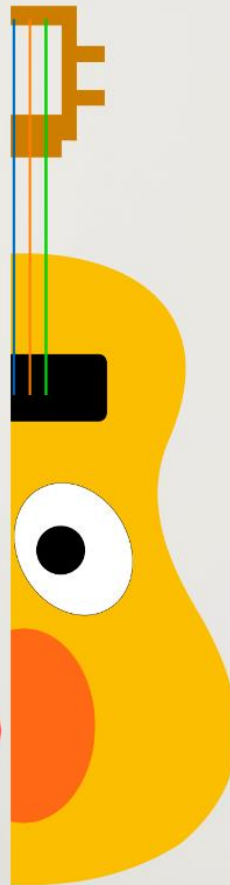
Big Data y Data Science

Introducción al Big Data

www.iic.uam.es

iic
instituto
de ingeniería
del conocimiento
www.iic.uam.es

RigoBERTa



Un nuevo modelo del
lenguaje español:

RigoBERTa



RigoBERTa

Fuentes



149GB de crawls de webs en español. 26 mil millones de palabras.



OSCAR

69GB de noticias de medios de prensa en español.



Porción española del corpus mC4. 433 mil millones de palabras.



Spanish Unnannotated Corpora. 3 mil millones de palabras

RigoBERTa – Resultados del benchmark



	Dataset	BETO	BERTIN	MarIA	RigoBERTa
NER	CANTEMISTNER	89.9%	79.5%	92.3%	93.3% ★
NER	CAPITEL	87.0%	86.5%	87.8% ★	87.4%
NER	CONLL2002	89.6%	90.1% ★	89.9%	89.5%
Anonymize	MEDDOCAN	84.7%	72.2%	84.1%	85.0% ★
NER	MEDDOPROF1	80.5%	71.0%	80.7%	83.1% ★
NER	MEDDOPROF2	81.8%	44.2%	78.5%	86.4% ★
Classification	MLDOC	95.4%	94.4%	95.6% ★	95.6% ★
Paraphrasing	PAWS-X	89.7%	90.1%	88.9%	91.0% ★
NER	PHARMACONER	61.4%	47.1%	57.1%	70.0% ★
QA	SQAC	76.2%	75.0%	86.6%	89.7% ★
QA	SQUADES	75.6%	70.0%	81.8%	85.4% ★
Sentiment	TASS2020	46.1%	46.1%	47.3% ★	46.7%
Entailment	XNLI	81.7%	79.4%	81.6%	83.4% ★
TOTALS		76.5%	69.6% x1	77.3% x3	79.8% ★ x10

Modelo Beto

<https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased>

Hosted inference API

Fill-Mask

Mask token

Mi nombre es [MASK] y vivo en Nueva York.

Computation time on cpu: cached.

[UNK]

David

Alex

Michael

Mary

</> JSON Output API API Endpoint

Hosted inference API

Fill-Mask

Mask token: [MASK]

Vivo en [MASK], una ciudad hermosa, con Sierra Nevada

Compute

Computation time on cpu: 0.079 s

Vivo en, una ciudad hermosa, con Sierra Nevada

0.201

Granada

0.090

Monterrey

0.060

Salamanca

0.059

Madrid

0.042

</> JSON Output API API Endpoint

Maximize

0.005

Maximize

Modelo BSC- tarea NER

BSC-TeMU/**roberta-base-bne-capitel-ner** like 1

Token Classification PyTorch Transformers bne capitel es arxiv:1907.11692 arxiv:2107.07253 apache-2.0 roberta national library of spain spanish bne capitel ner

Model card Files and versions Train Deploy Use in Transformers

Spanish RoBERTa-base trained on BNE finetuned for CAPITEL Named Entity Recognition (NER) dataset.

RoBERTa-base-bne is a transformer-based masked language model for the Spanish language. It is based on the [RoBERTa](#) base model and has been pre-trained using the largest Spanish corpus known to date, with a total of 570GB of clean and deduplicated text processed for this work, compiled from the web crawlings performed by the [National Library of Spain \(Biblioteca Nacional de España\)](#), from 2009 to 2019.

Original pre-trained model can be found here: <https://huggingface.co/BSC-TeMU/roberta-base-bne>

Hosted inference API

Token Classification

Me llamo Wolfgang y vivo en Berlin

Compute

Computation time on cpu: cached

Me llamo Wolf **S_PER** gang y vivo en Berlin **S_LOC**

JSON Output Maximize

<https://huggingface.co/BSC-TeMU/roberta-base-bne-capitel-ner?text=Me+llamo+Wolfgang+y+vivo+en+Berlin>

Generación de texto



OpenAI generador de noticias falsas



- GPT-2: generador de noticias falsas
- OpenAI (Elon Musk)
- Desarrolla el texto a partir de una semilla de una o dos frases escritas por un humano
- Algoritmo liberado el 05/11/2019 junto con un detector
<https://openai.com/blog/gpt-2-1-5b-release/>
- Último gran hito del PLN

<https://www.youtube.com/watch?v=XMJ8VxgUzTc#action=share>

Sistemas de pregunta respuesta (Question Answering)

El Stanford Question Answering Dataset (SQuAD) es un dataset de comprensión lectora de preguntas y Respuestas para la lengua inglesa.

SQuAD1.1 (Rajpurkar et al. 2016)

- 107.758 preguntas
- Preguntas “contestables”: todas tienen respuesta
- Modelo de regresión logística F1 51% (sobre el baseline 20%).
- Se apoya en árboles de constituyentes y de dependencias.

https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Doctor_Who.html

Ranking de acierto con SQUAD

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) Google Research & TTIC https://arxiv.org/abs/1909.11942	89.731	92.215
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) PINGAN Omni-Sinitic	88.592	90.859
2 Jul 26, 2019	UPM (ensemble) Anonymous	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) Shanghai Jiao Tong University & CloudWalk https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) Shanghai Jiao Tong University & CloudWalk https://arxiv.org/abs/1908.05147	87.238	90.071
5 Jul 26, 2019	UPM (single model) Anonymous	87.193	89.934

Generación de resumen

<https://huggingface.co/spaces/hackathon-pln-es/BioMedIA>

Pregúntame sobre BioMedicina o temas relacionados. Puedes simplemente preguntarme aquí y darle al botón verde de abajo que pone Enviar.

¿qué es la aspirina?

Sube un audio con tu respuesta aquí si quieres.

Coloque el audio aquí
- o -
Haga click para cargar

Graba aquí un audio con tu pregunta.

Record from microphone

Minimum size for the answer 50

El acetaminofén es un analgésico, y es un antiinflamatorio. El paracetamol es un ácido acetilsalicílico, un antihistamínico y un antihistamínico.

El paracetamol, también conocido como acetaminofén, es un fármaco con propiedades analgésicas (que combate el dolor) y antipiréticas (que combate la fiebre), disponible en el mercado en tabletas, jarabe, gotas o suposí[...]

Es una sustancia sólida blanca, cristalina, constituida por ácido acetilsalicílico, que se usa como analgésico y antipirético. Analgésicos para el alivio del dolor de cabeza, dolores musculares y articulares. La aspiri[...]

El paracetamol (DCI) o acetaminofén es un fármaco con propiedades analgésicas, sin propiedades antiinflamatorias clínicamente significativas. Actúa inhibiendo la síntesis de prostaglandinas, mediadores celulares respon[...]

El ácido acetilsalicílico es una sustancia química sintetizada en 1897 por un joven químico alemán, Felix Hoffman. Su origen se encuentra en el reino vegetal, ya que diferentes especies vegetales, como el sauce blanco [...]

Inicio » Cáncer » La aspirina podría sumarse a la batalla contra el cáncer Tomar una aspirina al día, además de ayudar a las personas con problemas del corazón, también podría ser bueno para combatir y reducir el cánc[...]

Respuesta en audio.



0:00 / 0:16



Generación de resumen

https://huggingface.co/GiordanoB/mT5_multilingual_XLSum-sumarizacao-PTBR

⚡ Hosted inference API ⓘ

Text2Text Generation

Una Italia empobrecida y desilusionada se entrega al experimento de un Gobierno de la ultraderecha
La desconexión de las zonas suburbanas, el envejecimiento del electorado y la falta de mensaje de la izquierda más allá del miedo a la extrema derecha son algunas claves que explican la victoria de Giorgia Meloni

Compute

ctrl+Enter

0.0

Computation time on cpu: 5.040 s

La victoria de Giorgia Meloni en las elecciones presidenciales de Italia es un experimento de un Gobierno de la ultraderecha.

</> JSON Output

Maximize

- Resumen extractivo (extracción de frases y ordenación. Hay distintas técnicas para ver qué frases son más importantes)
- Resumen abstractivo/generativo (utiliza extracción de frases, luego puede hacer un resumen inicial, que luego simplifica, o utiliza la comprensión de frases para hacer un resumen definitivo)

Generación de Texto

Muy poco desarrollada en español

GPT- NEO modelos disponibles online

Se introduce un texto corto y el modelo te genera un texto nuevo

<https://huggingface.co/blog/few-shot-learning-gpt-neo-and-inference-api>

API Token

api_org_YDfgSDFs_replace_with_your_token

Task

Write your own prompt

End Sequence

###

Token Length

75

Temperature

1

Example prompt:

El presidente del gobierno decretará el estado de alerta de coronavirus. El comunicado dio a conocer que el exjefe de Estado, Pedro Sánchez, había dado el alta al canciller Jorge Fernández. Este martes pasado, la autoridad sanitaria confirmó el estado de alerta en España.

Más información sobre modelos pre-entrenados y transformers

AI Workshop: Curso Práctico de NLP de cero a cien con Hugging Face

- <https://www.youtube.com/playlist?list=PLBILcz47fTtPspj9QDm2E0oHLe1p67tMz>

Reconocimiento de voz

- Servicio consolidado
- Multitud de idiomas, inglés idioma principal
- +20 dialectos español
- Gran acierto en conversaciones con buena acústica
- Transcripción por identificación de locutores

¿A qué esperas para convertir tu voz en texto?

Selecciona un idioma y haz clic en la opción para comenzar a grabar.

Input type
☒ Microphone ☐ File upload


Language
Español (España) ▼

Speaker diarization **BETA**
Off ▼

Speakers
1 speaker ▼

Punctuation
☒

Show JSON ▼

 START NOW

<https://cloud.google.com/speech-to-text>

Ejemplo de transcripción automática (Orbita Laika)



■ Speaker 1 ■ Speaker 2

“ en la conversación está hoy que los inviernos son los de antes fíjate qué verano nos tenemos pasamos el invierno al verano es una sensación que la gente tiene absolutamente decir estamos viviendo como esos veranos el tiempo veraniego se está alargando ahora tenemos días de verano en octubre días de verano en marzo en abril inviernos que te quedan reducidos a la mínima expresión aunque hace frío a veces está te voy a dar un gato por nota a aburrir mucho a los amigos no se escuchan los ven ahora mismo en España la duración media del verano es cerca de un mes mayor que hace 30 años un mes más de verano un grano alargado que empieza a percibir mucho antes y termina mucho después en estos últimos 30 años la cosa se ha ido dilatando y expandiéndose hasta prácticamente no tiene efecto directo sobre los productos agrícolas españoles notablemente y si te parece mira ”

Ejercicio 7: Speech2text

Transcripción automática

- Vamos a probar distintos servicios.
- Primera opción: Servicio de Google. Transcribe un vídeo de menos de 1 minuto (*) con el servicio de Google. : <https://cloud.google.com/speech-to-text>, realiza las pruebas tanto en la identificación de locutores como sin ellos, así como la calidad de la transcripción.

- Segunda opción: Probar Whisper <https://huggingface.co/spaces/openai/whisper>



```
!pip install git+https://github.com/openai/whisper.git
```

```
[2] !whisper input.mp3 --model medium
```

- Cortar vídeos: Puede hacerse de distintas maneras, algunas sugerencias para conseguir un vídeo de menos de 1 minuto: Descargar vídeo desde URL: <https://videocyborg.com/>. Cortar vídeos: <https://clideo.com/>

<https://openai.com/blog/whisper/>

Chatbot

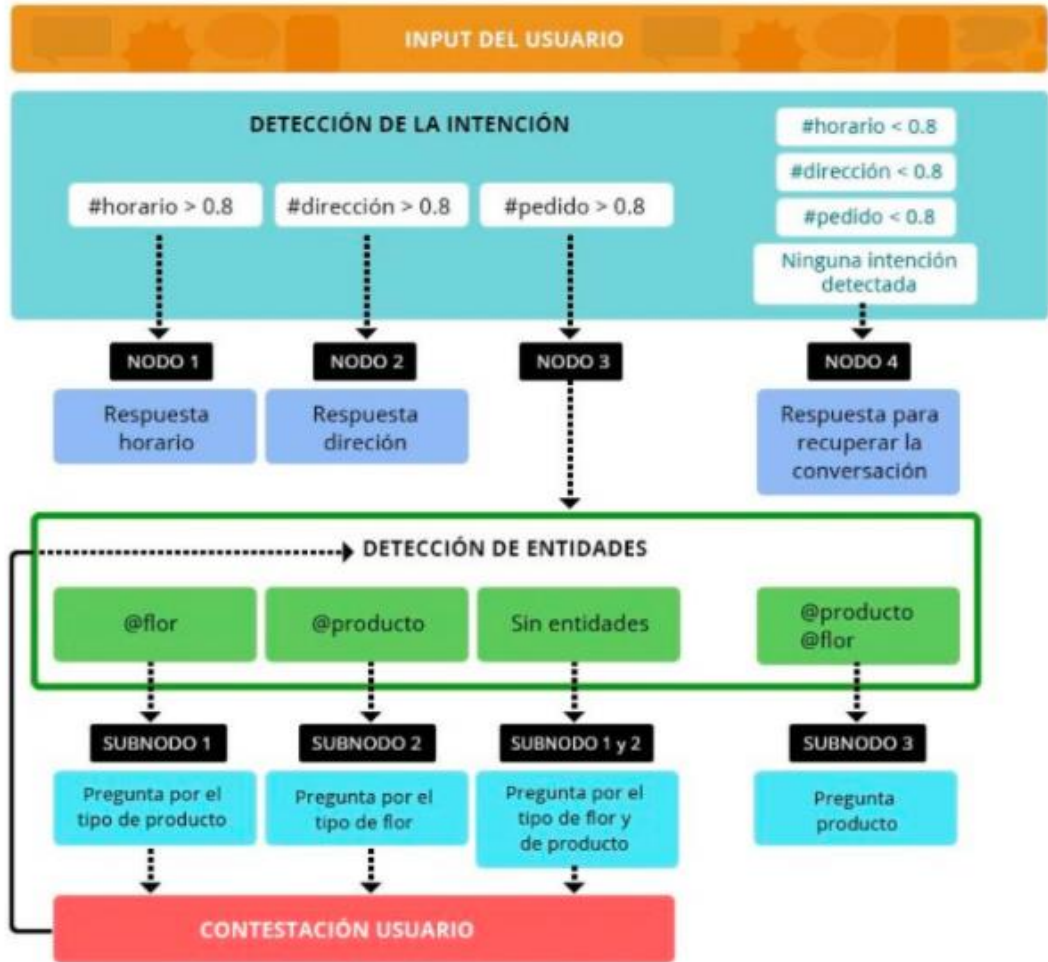
Un chatbot es un software conversacional entrenado para dar respuestas automáticas ante las entradas del usuario y cuyo propósito principal es simular la conversación con una persona real.

Se consigue construyendo un árbol de diálogo, y se entrenan las intenciones que tiene el usuario

Intenciones: comprar un móvil -> Quiero comprar un móvil

Me gustaría comprar un móvil

¿Tenéis móviles para ofrecerme?



Chatbot

<https://msfcovid19.org/>



RESIDENCIAS E INSTITUCIONES HOSPITALES Y SERVICIOS DE SALUD COORDINADORES Y AUTORIDADES FOF

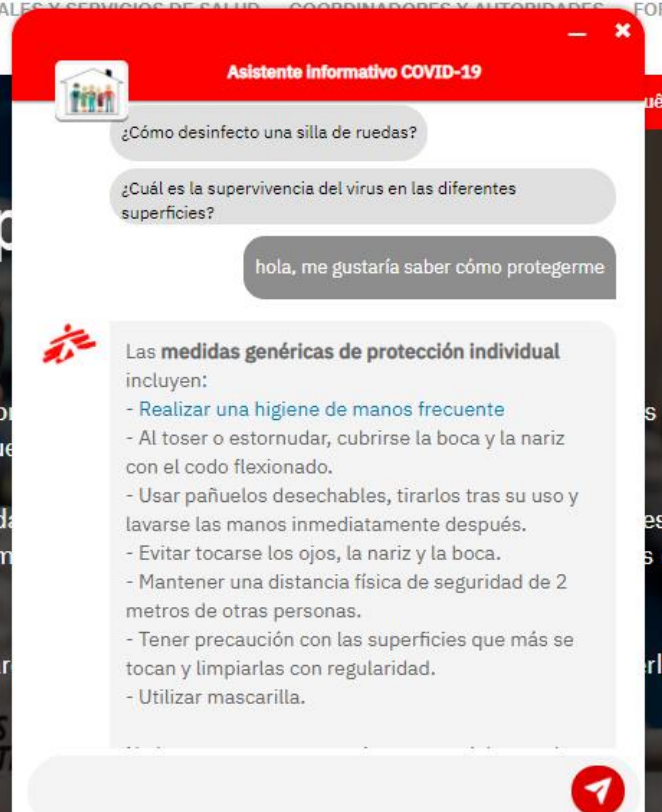
Recursos para atención en p pandemia de COVID-19

A través de este sitio web, la organización humanitaria internacional Médicos Sin Fronteras ofrece recursos para instituciones y trabajadores que se encuentran en la primera línea de respuesta al nuevo coronavirus.

Los visitantes encontrarán aquí guías técnicas, protocolos y formaciones especializadas sobre los mecanismos de respuesta a la pandemia y personal de residencias para adultos migrantes.

Los mismos materiales también pueden ser descargados de la aplicación para celular disponible aún sin conexión.

Contacto: msfe-covid19-latam@msf.org



!!Muchas gracias!!