Universidad Rey Juan Carlos

NLP con modelos de transformers

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

27/04/22

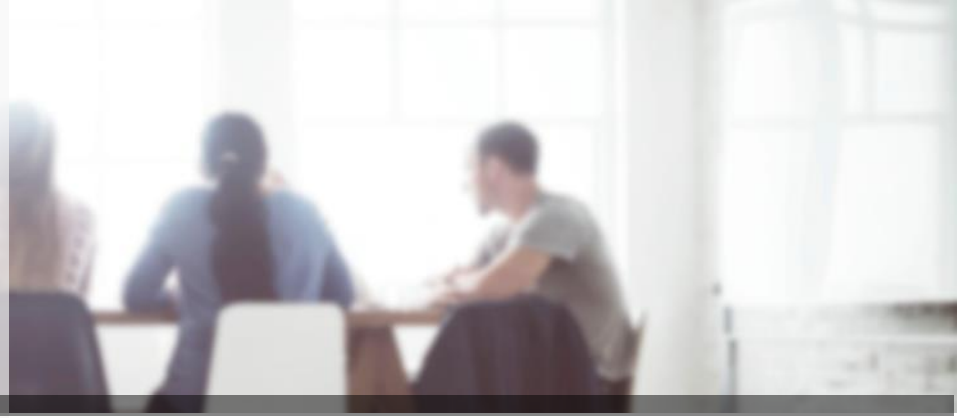
Índice

1. Introducción al NLP
2. Modelos clásicos
3. Transformers
4. ¿NLP en español?
5. Ejemplo práctico

Índice

Introducción al NLP

¿Qué es el NLP?



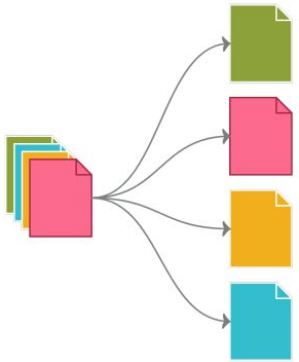
Análisis automático de cualquier tipo de texto en diferentes idiomas con el fin de **descubrir, clasificar, organizar o buscar contenido no explícito** para agilizar tareas donde se manejan grandes volúmenes de datos de forma manual.

- Análisis del sentimiento y emociones (redes sociales, encuestas, prensa…)
- Detección y anonimización de entidades (personas, empresas, dnis, calles...)
- Recuperación de información en texto y audio (ámbito jurídico)
- Clasificación automática de reclamaciones
- Descubrir tendencias en documentos (redes sociales, blogs…)
- Filtrado de spam

Introducción al NLP

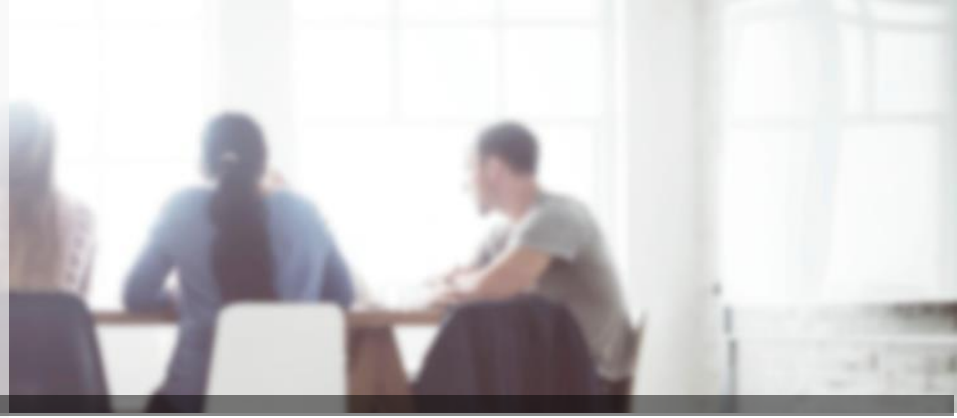
Problemas típicos de NLP

A parte de los típicos problemas del machine learning (clasificación, regresión y clustering), el NLP (sobre todo el moderno) afronta una gran variedad de problemas.



Introducción al NLP

Problemas típicos de NLP



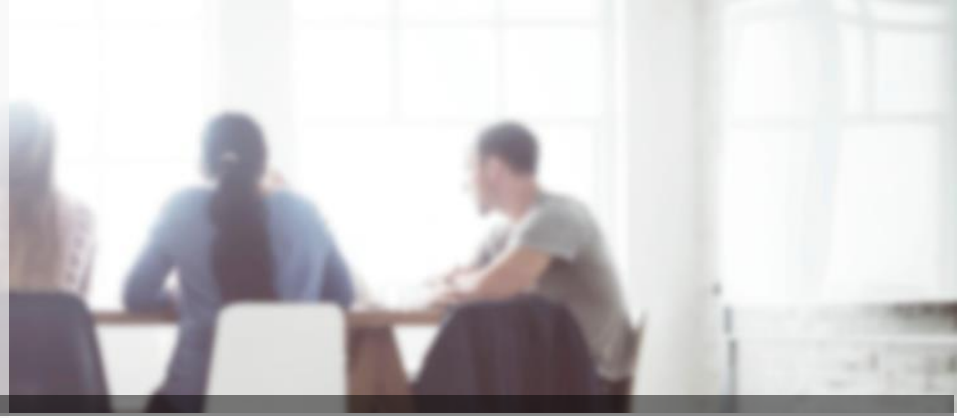
Trata de detectar entidades nombradas o ciertas **palabras pertenecientes a algún campo semántico concreto**. Se puede usar para anonimizar documentos, obtener insights que agregar para un análisis posterior, etc. Tarea supervisada, en esencia es clasificación de tokens.

Named Entity Recognition (NER):

Albert Einstein PER Albert Einstein was born in Ulm LOC in Germany LOC on March 14, 1879. Six weeks later the family moved to Munich LOC , where he later on began his schooling at the Luitpold Gymnasium ORG . In 1896 he entered the Swiss Federal Polytechnic School ORG in Zurich LOC to be trained as a teacher in physics and mathematics.

Introducción al NLP

Problemas típicos de NLP



A partir de un texto base, generar la continuación. Generar fake news, responder preguntas, etc. Esta tarea es no supervisada o semi-supervisada y **es la base de muchas otras tareas**.

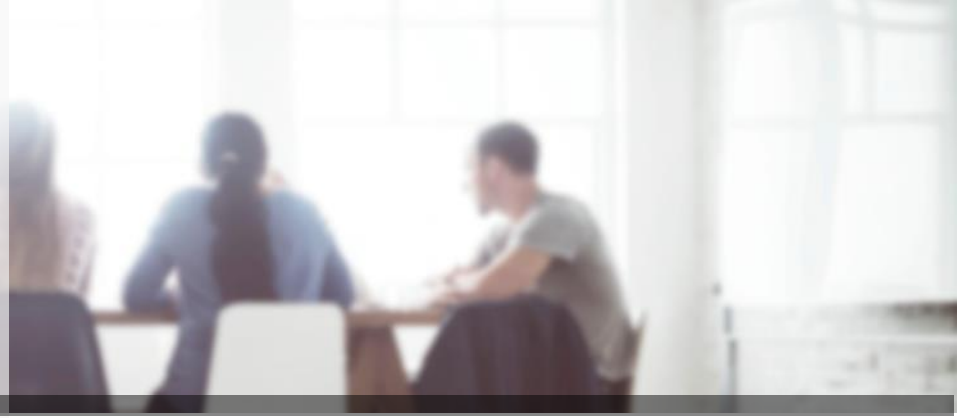
Generación de texto:

Playground

El otro día iba en el metro y me encontré una moneda.

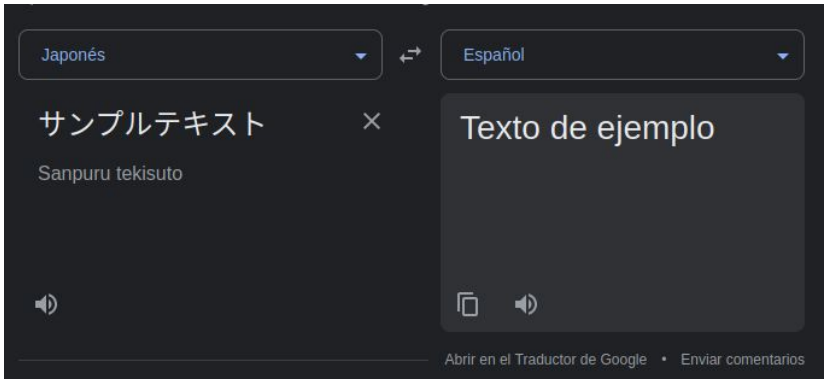
Introducción al NLP

Problemas típicos de NLP



De las aplicaciones más antiguas. Dado un texto en un idioma, obtener el “equivalente” en otro idioma.

Traducción automática:

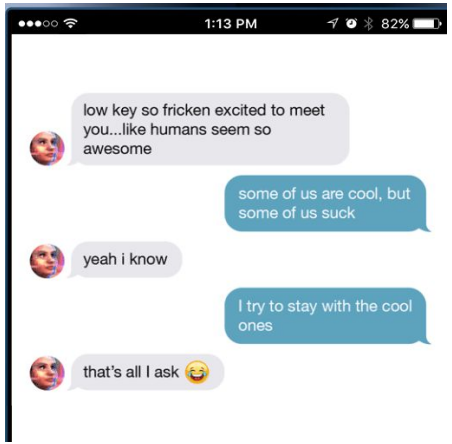


Introducción al NLP

Problemas típicos de NLP

Hay de dos tipos. Los “generales” y los que están programados con reglas (árboles de decisión). Muy mediáticos, pero siguen lejos de lo que los periodistas creen que son.

Chatbots:



Introducción al NLP

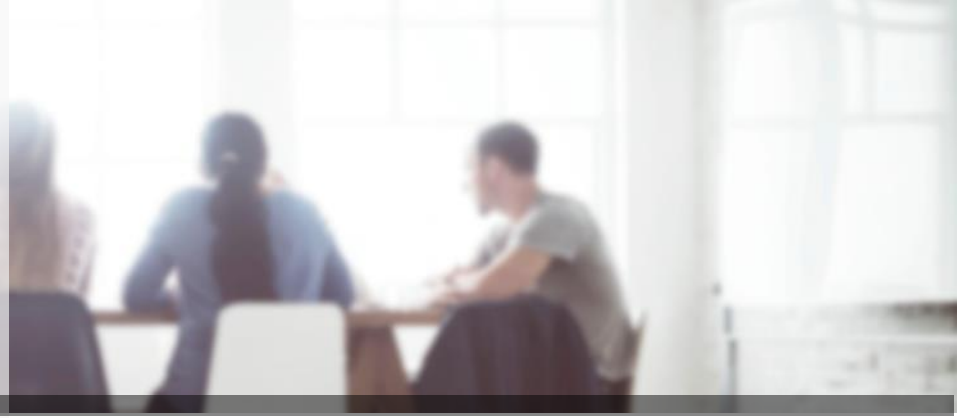
Problemas típicos de NLP

Dado un texto y una pregunta sobre el mismo, generar la respuesta. Muy útil para extracción de información de grandes bases de textos. (el ejemplo no solo hace eso)

Question answering:

Introducción al NLP

Problemas típicos de NLP



Dado un texto, generar su resumen. Estos pueden ser extractivos o abstractivos.

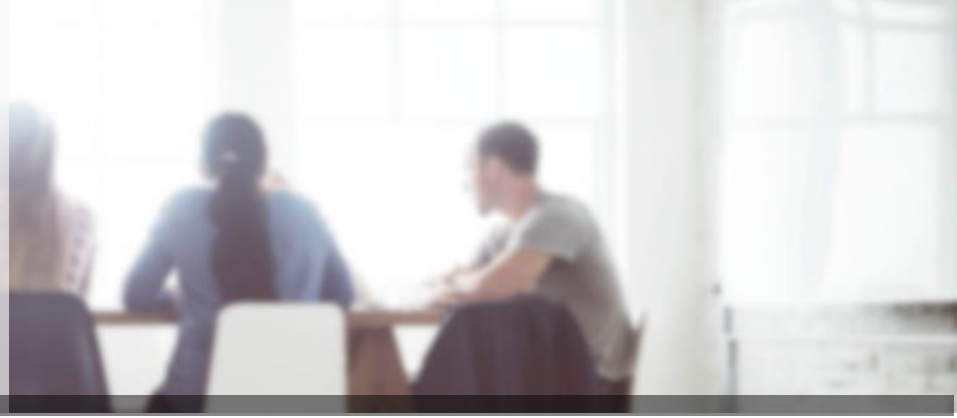
Resúmenes automáticos:

Hey everyone, obvious throwaway here, I'm a 22 year old recent cs graduate from Columbus, Ohio (Go bucks, fuck Michigan!!). I voted in my first election in 2016 for donald trump, as did all of my immediate family. I was brought up in a catholic household in suburban Ohio, and while we always leaned conservative, we tried to evaluate each candidate on an issue by issue basis, supporting and voting for a variety of candidates such as Obama (both times) and kasich. We were never really strong supporters of trump, including his behavior with women, though we mistakenly overlook it based on his unique and vaguely positive rhetoric on trade deals, unions, and "bringing the jobs", which resonated well here in Ohio and many other areas of the Midwest. We felt that Hillary did not make an attempt to reach out to folks here, considering many have been left behind here in Ohio as manufacturing jobs shut down and disappeared or moved overseas. I now realized that this was probably the worst mistake of our lifetime, trump has shown he is the most unprofessional, unethical, and unamerican president we have had in our lifetimes. None of the jobs are back, constant racist dog whistles, alienation of our allies, 200000 Americans dead from *willfull* malicious negligence, and now attempting to ruin the mail because he doesn't like amazon and mail in voting? In addition, reading more into the Democratic Party platform, a lot of it just makes sense. Medicare for all, a green new deal, and new infrastructure projects are all things that will net benefit american families with a better sense of security and good paying jobs around Ohio, the rust belt, and the rest of America. Of course we have disagreements, notably on guns, but we need a president who at least is willing to cooperate and listen to the other side. I can tell you that I, my mom, dad, and younger (18!) sister will not only be voting biden, but blue down the ticket until trumpism is dead from the Republican Party. We're sorry for this mess, we really are. We're trying our best to reach out to as many people as we can to vote and get their voice out there, especially since Ohio is barely in trumps column according to polls. I still think he will win here but just want to let you guys know we're not all stupid (anymore). Thanks for reading!

Tldr: I was an idiot in 2016. Im not one anymore.

Introducción al NLP

Problemas típicos de NLP



Dado un texto, generar su resumen. Estos pueden ser extractivos o abstractivos.

Resúmenes automáticos:

Hey everyone, obvious throwaway here, I'm a 22 year old recent cs graduate from Columbus, Ohio (Go bucks, fuck Michigan!!). I voted in my first election in 2016 for donald trump, as did all of my immediate family. I was brought up in a catholic household in suburban Ohio, and while we always leaned conservative, we tried to evaluate each candidate on an issue by issue basis, supporting and voting for a variety of candidates such as Obama (both times) and kasich. We were never really strong supporters of trump, including his behavior with women, though we mistakenly overlook it based on his unique and vaguely positive rhetoric on trade deals, unions, and "bringing the jobs", which resonated well here in Ohio and many other areas of the Midwest. We felt that Hillary did not make an attempt to reach out to folks here, considering many have been left behind here in Ohio as manufacturing jobs shut down and disappeared or moved overseas. I now realized that this was probably the worst mistake of our lifetime, trump has shown he is the most unprofessional, unethical, and unamerican president we have had in our lifetimes. None of the jobs are back, constant racist dog whistles, alienation of our allies, 200000 Americans dead from *willfull* malicious negligence, and now attempting to ruin the mail because he doesn't like amazon and mail in voting? In addition, reading more into the Democratic Party platform, a lot of it just makes sense. Medicare for all, a green new deal, and new infrastructure projects are all things that will net benefit american families with a better sense of security and good paying jobs around Ohio, the rust belt, and the rest of America. Of course we have disagreements, notably on guns, but we need a president who at least is willing to cooperate and listen to the other side. I can tell you that I, my mom, dad, and younger (18!) sister will not only be voting biden, but blue down the ticket until trumpism is dead from the Republican Party. We're sorry for this mess, we really are. We're trying our best to reach out to as many people as we can to vote and get their voice out there, especially since Ohio is barely in trumps column according to polls. I still think he will win here but just want to let you guys know we're not all stupid (anymore). Thanks for reading!

Tldr: I was an idiot in 2016. Im not one anymore.

Índice

Modelos clásicos

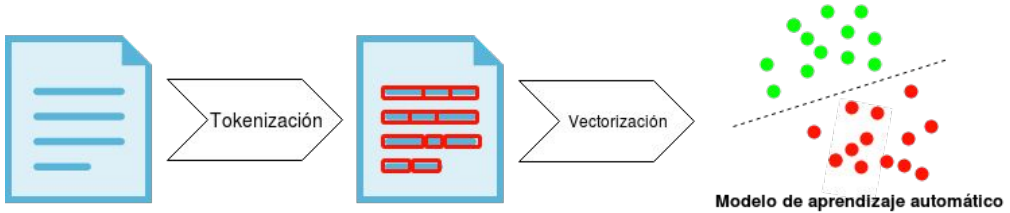
Modelos basados en diccionarios

Expertos lingüistas creaban diccionarios enormes de keywords para la tarea que se quisiera llevar a cabo, identificación de un sentimiento (clasificación), traducción, NER, etc. **Muy costoso** en tiempo y dinero, resultados malos.

Modelos clásicos

Approach con Machine Learning

Se necesita obtener una representación del texto que los modelos puedan entender. Solo aplica a problemas típicos, no a generación de texto, etc.



Modelos clásicos

Embeddings de texto

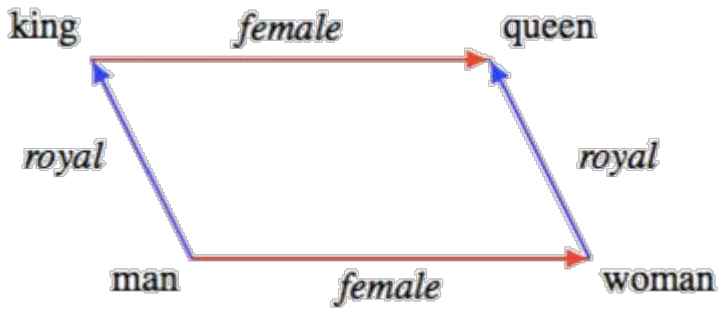
Son representaciones vectoriales de los textos para que las entiendan los modelos.
Algunas simples como Bag of Words (BoW) o Tf-idf.

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Modelos clásicos

Embeddings con redes neuronales

Se aprenden de forma **semi-supervisada** y se usan redes neuronales recurrentes, sobre todo las LSTM. P.ej: GloVe, Word2Vec, FastText, etc.
Ya no son solo cuentas de palabras, sino que se aprende la distribución estadística de las palabras en el idioma, aprende la **semántica.**



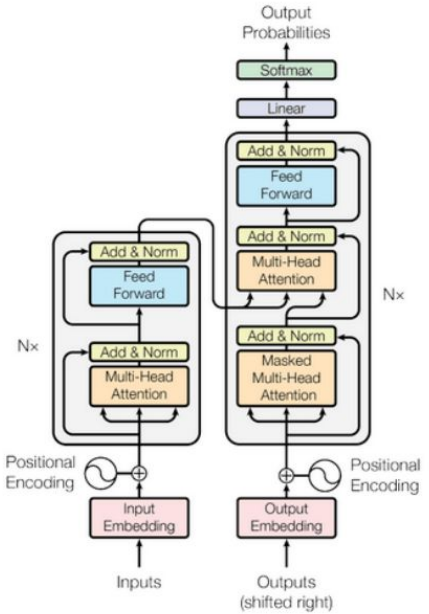
Índice

1. Introducción al NLP
2. Modelos clásicos
3. **Transformers**
4. ¿NLP en español?
5. Ejemplo práctico

Transformers

Attention is all you need

En 2017 aparece el Modelo Transformer. Dicho modelo crea una arquitectura de redes neuronales con **mecanismos de atención** que consigue superar todos los métodos más usados hasta la fecha.
Constaba de parte encoder y parte decoder.



Transformers

BERT

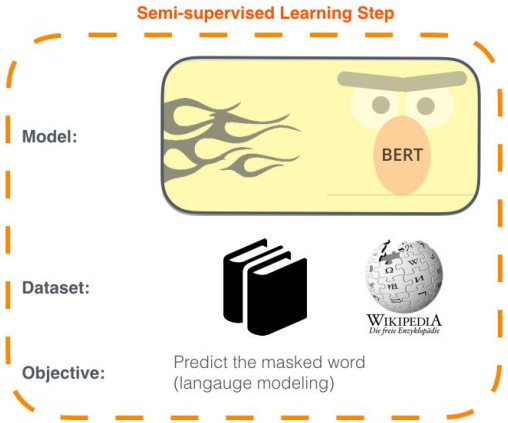
En 2018 se crea BERT. Desde entonces se produce una revolución en el campo del PLN y aparecen cientos de modelos basados en BERT o intentando mejorar la arquitectura transformers.
Entrenado con estrategia **fill-mask,** muy buena para que el modelo vea el contexto.

Transformers

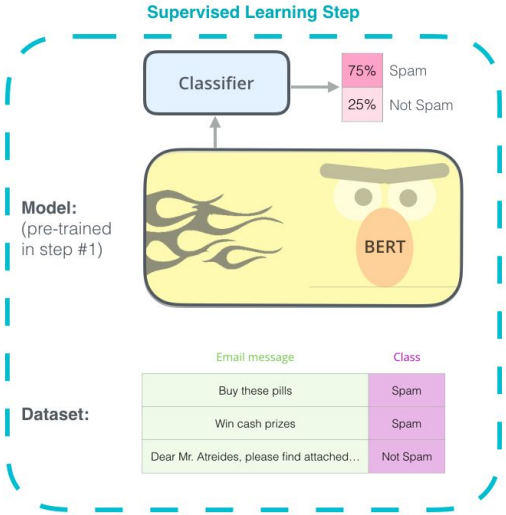
Fine Tuning

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.



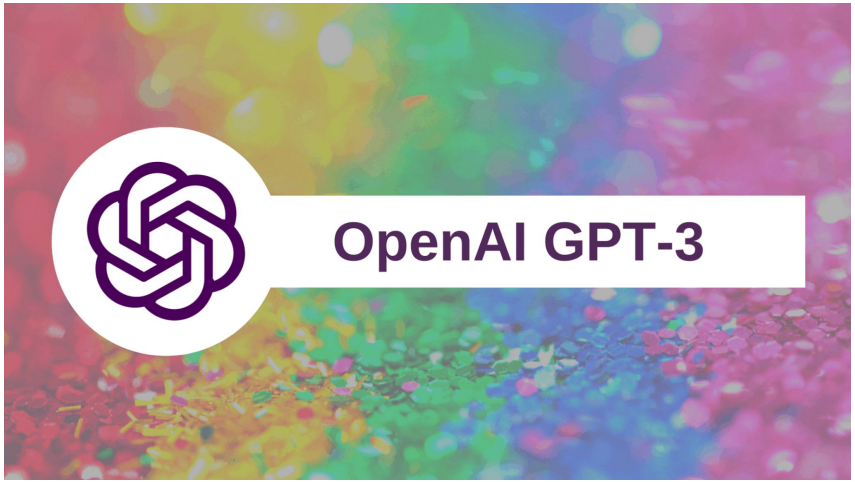
2 - Supervised training on a specific task with a labeled dataset.



Transformers

Familia GPT

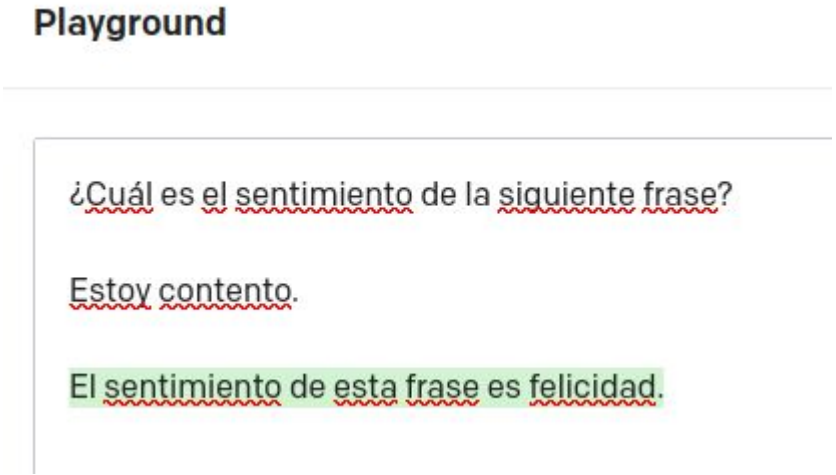
Simplificando, son muchas capas de decoder juntas, de hecho cuantas más, mejor.
Muy buenos resultados en tareas de tipo Winograd, comprueban relaciones gramáticas complejas, no solo semánticas.



Transformers

Zero/Few Shot Learning

Esta técnica de “inferencia” se basa en usar un modelo base lo suficientemente grande como para no tener que hacer fine tuning sobre él.
Esto se acerca más a lo que se podría entender por una inteligencia más “general” (aunque realmente es que el modelo ha memorizado internet).



Transformers

Zero/Few Shot Learning

Playground

¿Cuál es el sentimiento de la siguiente frase? Opciones: Negativo, Positivo, Neutro

Estoy contento.

Positivo

Transformers

Zero/Few Shot Learning

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

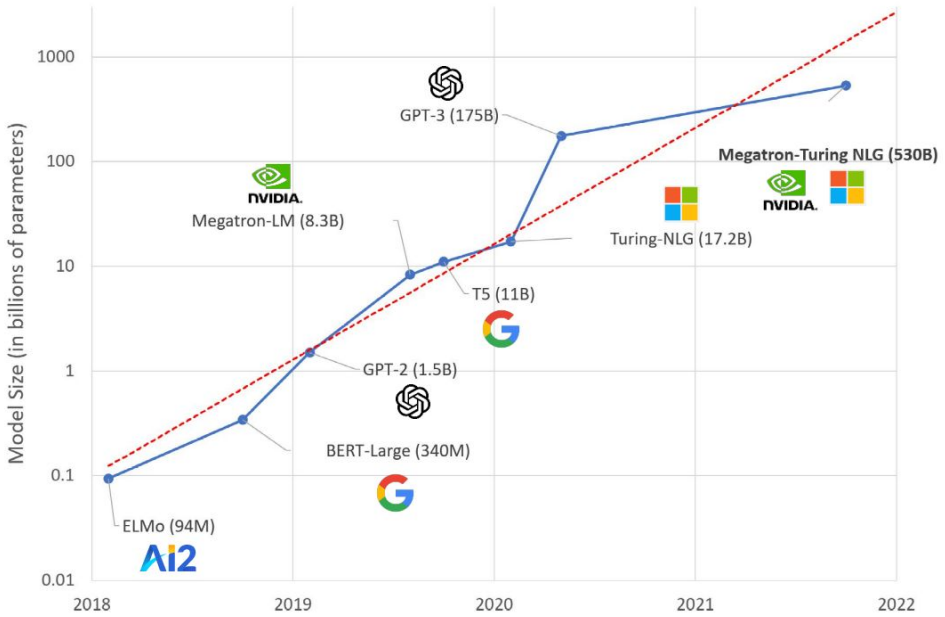
To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Transformers

¿El tamaño importa?

Parece que cuanto más grandes son los modelos, mejor funcionan. ¿Esto puede llegar hasta el infinito?

¿Quién se puede permitir entrenar tamaños modelos?



Índice

1. Introducción al NLP
2. Modelos clásicos
3. Transformers
4. **¿NLP en español?**
5. Ejemplo práctico

BERT

RoBERTa

GTP-2

GTP-3

BART

DeBERTa

¡Oye Blas! ¿Qué tal va tu modelo de lenguaje **español**?

¿NLP en español?

BERT Multi-idioma

Modelo BERT único para 100 idiomas.

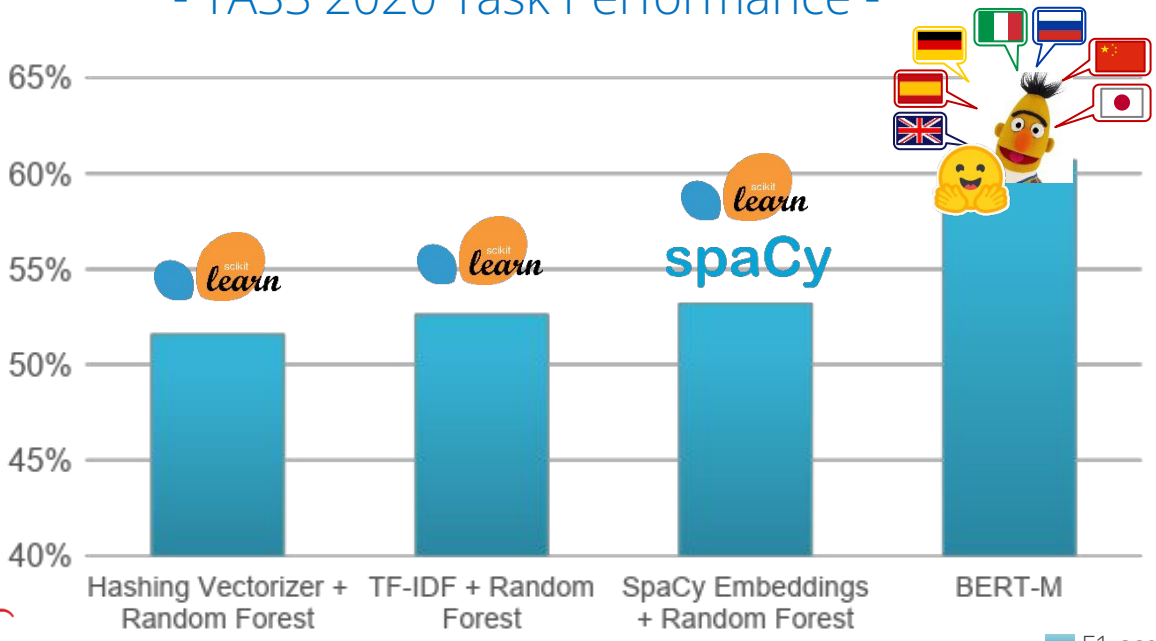
Entrenado con 100 wikipedias.

Capa de embeddings de 110K trozos de palabras, compartido entre todos los juegos de caracteres: latino, cirílico, kanji, hiragana, hangul, ...

Versión comprimida (DistilBERT) también disponible.

BERT multi-idioma
¿Qué tal funciona?

- TASS 2020 Task Performance -



TASS
Taller de Análisis Semántico en la SEPLN

TASS detección de emociones: positiva, negativa o neutral.

Mezcla de diferentes variedades del español

Training data:
TASS 2012 + 2018 + 2019 + 2020

Test data:
TASS 2020 dev dataset

Modelos específicos según idioma

Siguiendo las estrategias de pre-entrenamiento de BERT y RoBERTa, se han desarrollado modelos de lenguaje específicos para diferentes idiomas. Algunos ejemplos pintorescos:

CamemBERT (francés)

https://camembert-model.fr/

RobBERT (holandés)

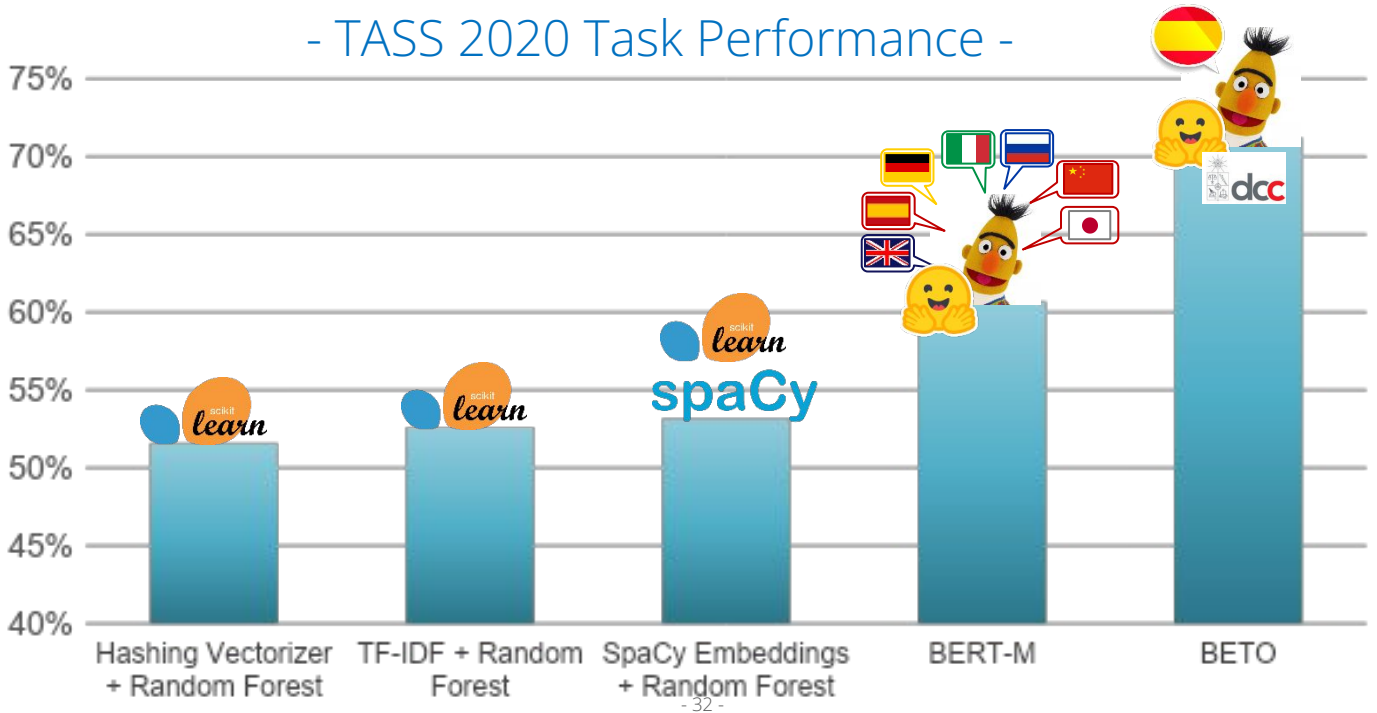
https://github.com/iPieter/RobBERT

GreekBERT (griego)

https://github.com/nlpaueb/greek-bert

BETO- Modelo de lenguaje **español**

- TASS 2020 Task Performance -



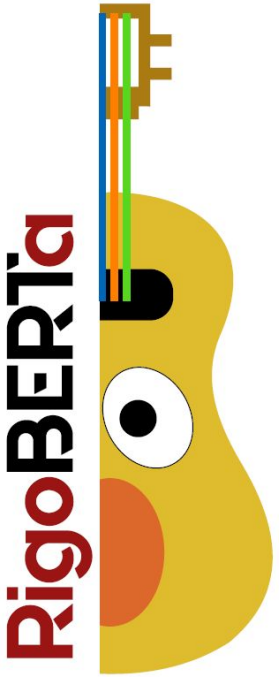
RigoBERTa

El mejor modelo en español

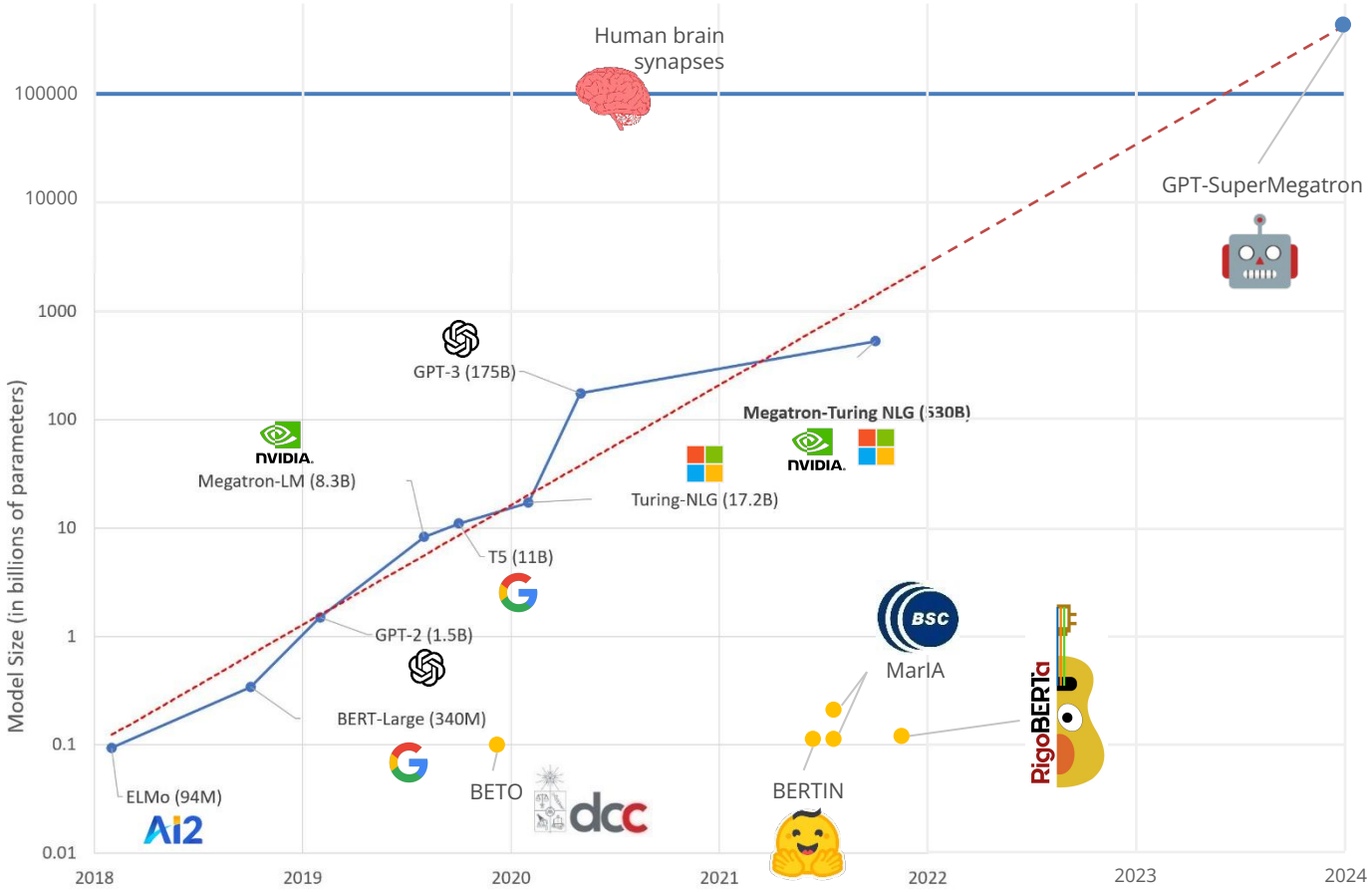
RigoBERTa usa un gran corpus curado por un equipo de lingüistas para descartar datos de mala calidad.

Además usa la arquitectura DeBERTa, que mejora las arquitecturas usadas por otros modelos del lenguaje en español.

En los benchmarks se ha conseguido superar el estado del arte en la mayoría de pruebas.



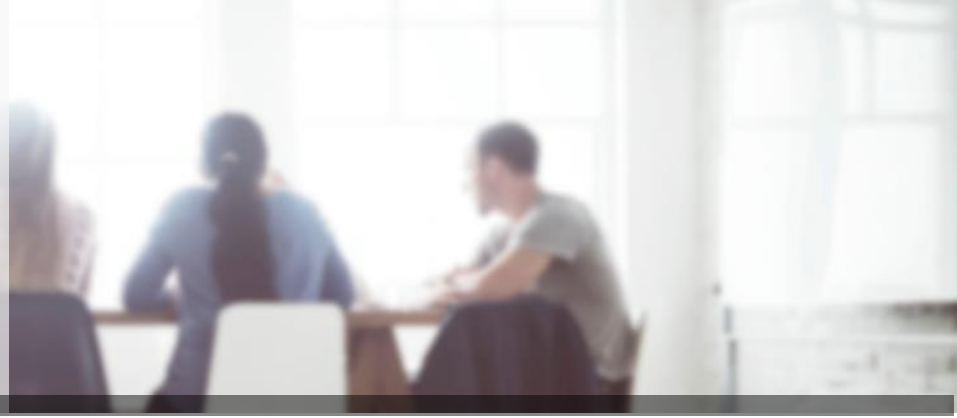
¿El futuro?



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/

Índice

Ejemplo práctico



jupyter

Guillem García Subies

https://github.com/guillemgsubies

Guillem García Subies
Data Scientist en el Instituto de Ingeniería del Conocimiento

You can check more articles of innovation on our Blog:
www.iic.uam.es/blog/

C/ Francisco Tomás y Valiente, nº 11,
EPS, Edificio B, 5ª planta
UAM Cantoblanco. 28049 Madrid
Tel.: (+34) 91 497 2323

Graphic elements of support obtained in :

Elementos gráficos de apoyo obtenidos en:
designed by freepik.com
pixabay