



UNIVERSIDAD
POLITÉCNICA
DE MADRID

POLITÉCNICA



Uso de Interlinguas para Búsqueda Documental Multilingüe

Máster Universitario en Inteligencia Artificial

Autor: César García Cabeza
Tutor: Dr. D. Jesús Cardeñosa Lera

Julio 2021

Índice de la presentación

- Introducción
- Estado del Arte
- Planteamiento del problema y modelo
- Experimentación y análisis
- Conclusiones
- Trabajos futuros
- Demo

Introducción



- Recuperación de la Información
- Relevante o no relevante
- Multilingüidad





Estado del Arte

De Gerald Salton al Aprendizaje Profundo

Técnicas de traducción

Basada en el conocimiento

- Diccionarios
- Tesauros

Basada en corpus

- Corpus paralelos
- Corpus comparables

Otras

- Traducción automática
- Interlinguas
- Aprendizaje profundo

▪ Importante la calidad

▪ Clasificación (Oard, 1997)

Técnicas de traducción



Basada en el conocimiento

- Diccionarios
- Tesauros

Basada en corpus

- Corpus paralelos
- Corpus comparables

Otras

- Traducción automática
- Interlinguas
- Aprendizaje profundo

- **Comienzo de diccionarios**
(Ballesteros y Croft, 1996, 1997, 1998)
 - Términos simples
 - Expresiones
 - Técnicas de expansión

Técnicas de traducción

Basada en el conocimiento

- Diccionarios
- Tesauros

Basada en corpus

- Corpus paralelos
- Corpus comparables

Otras

- Traducción automática
- Interlinguas
- Aprendizaje profundo

- **Creación manual** (Salton, 1969, 1972)
- **0 automática** (Brown, 1998)
- **Uso de existentes** (Salvador et al., 2014)

Técnicas de traducción

Basada en el conocimiento

- Diccionarios
- Tesauros



Basada en corpus

- Corpus paralelos
- Corpus comparables

Otras

- Traducción automática
- Interlinguas
- Aprendizaje profundo

▪ Tipo de corpus

▪ Combinación con otra técnica

Técnicas de traducción


Basada en el conocimiento

- Diccionarios
- Tesauros

Basada en corpus

- Corpus paralelos
- Corpus comparables

Otras

- 
- Traducción automática
 - Interlinguas
 - Aprendizaje profundo

▪ Baja calidad en textos cortos

▪ No se ajusta del todo

▪ Caro en comparación

Técnicas de traducción


Basada en el conocimiento

- Diccionarios
- Tesauros

Basada en corpus

- Corpus paralelos
- Corpus comparables

Otras

- 
- Traducción automática
 - Interlinguas
 - Aprendizaje profundo

▪ Características

▪ Multilingüidad masiva

▪ UNL (Cardeñosa et al., 2008, 2009)

Técnicas de traducción

Basada en el conocimiento

- Diccionarios
- Tesauros

Basada en corpus

- Corpus paralelos
- Corpus comparables

Otras

- Traducción automática
- Interlinguas
- Aprendizaje profundo



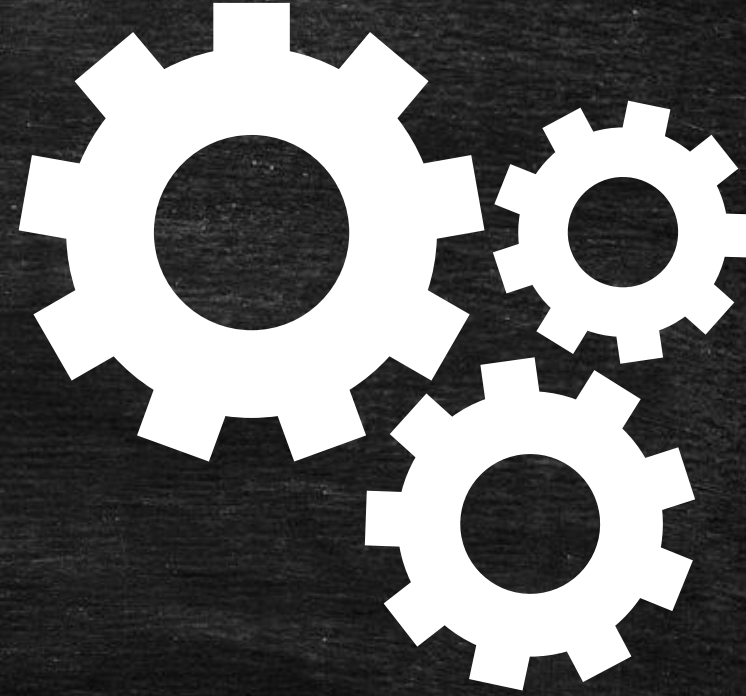
- **Word2Vec** (Mikolov et al., 2013)

- **No supervisado** (Litschko et al., 2018)

- **BERT** (Jiang et al., 2020)



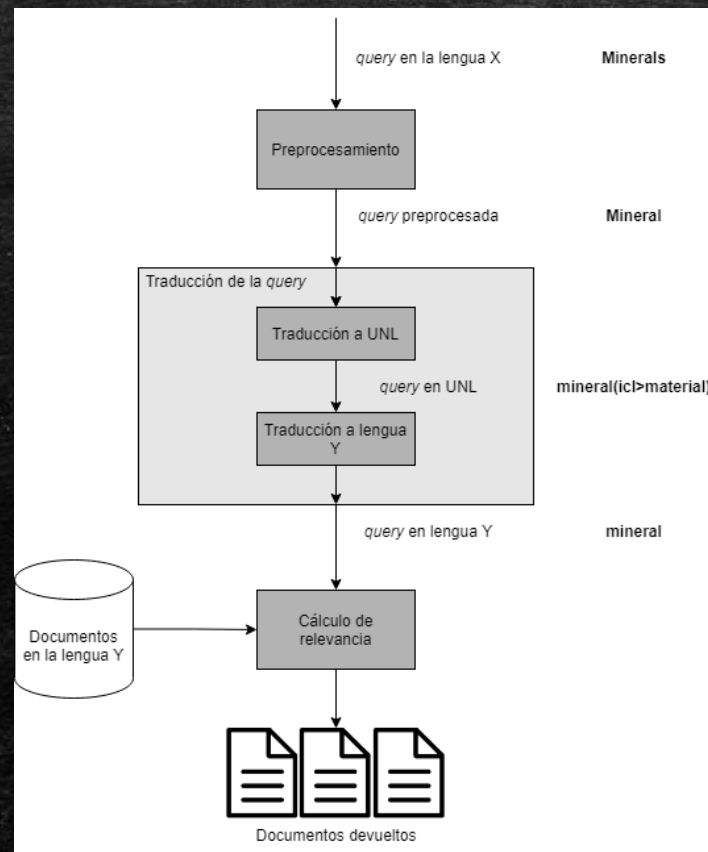
Planteamiento del problema



Modelo

Vector Space Model y UNL

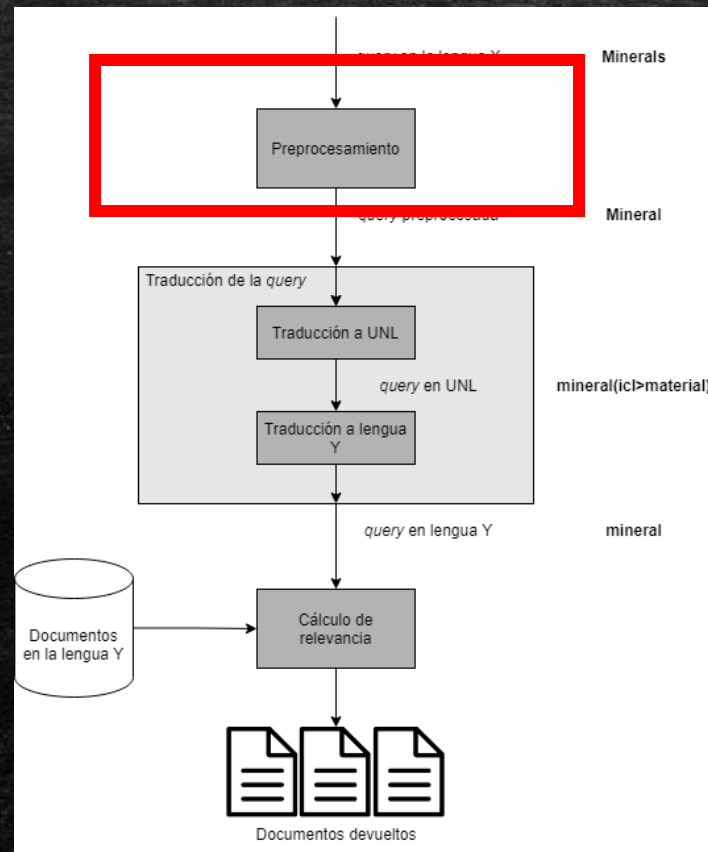
Propuesta de modelo



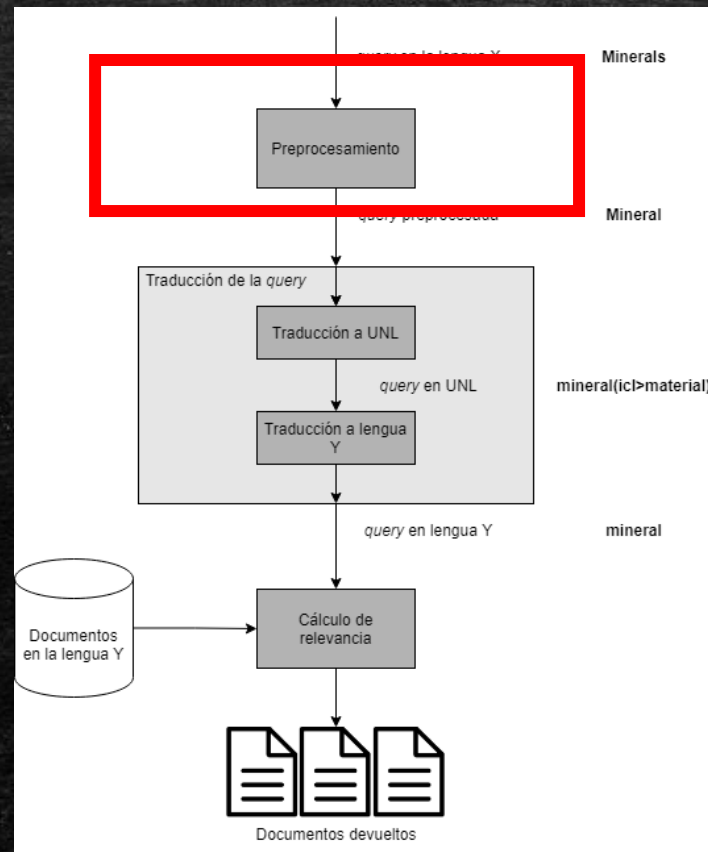
- Preprocesamiento
- Conversión multilingüe
- Buscador

Preprocesamiento: tokenización

- “Esto son dos ejemplos para la presentación”
- [“esto”, “son”, “dos”, “ejemplos”, “para”, “la”, “presentación”]

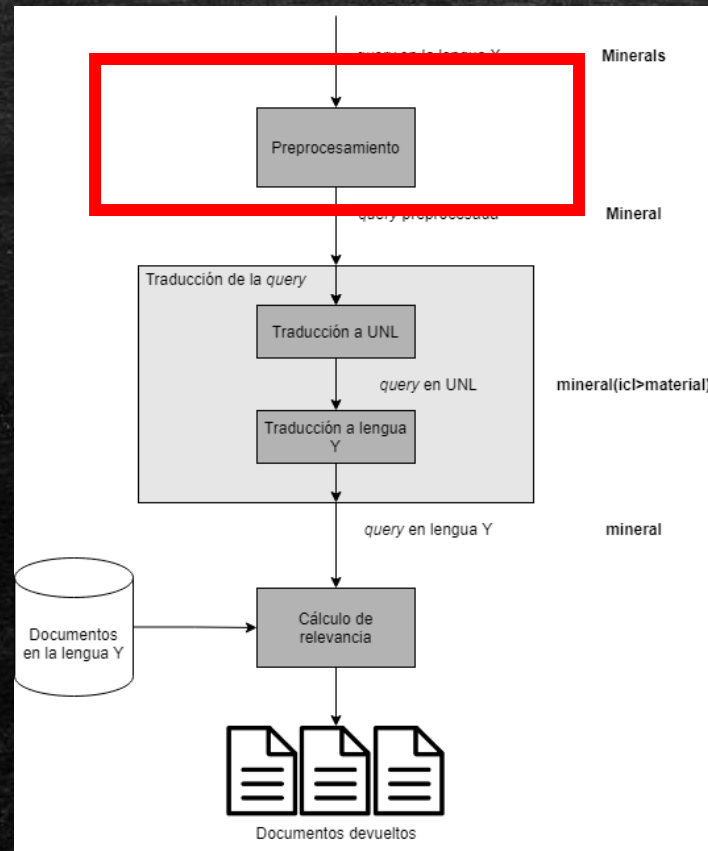


Preprocesamiento: palabras vacías



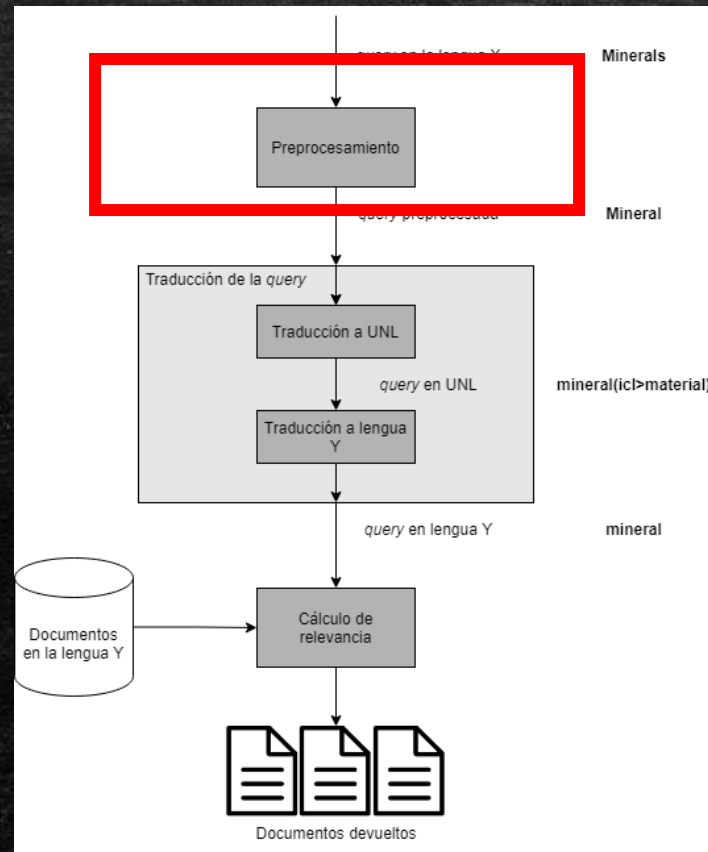
- “Esto son dos ejemplos para la presentación”
- [“esto”, “son”, “dos”, “ejemplos”, “para”, “la”, “presentación”]
- [“ejemplos”, “presentación”]

Preprocesamiento: normalización



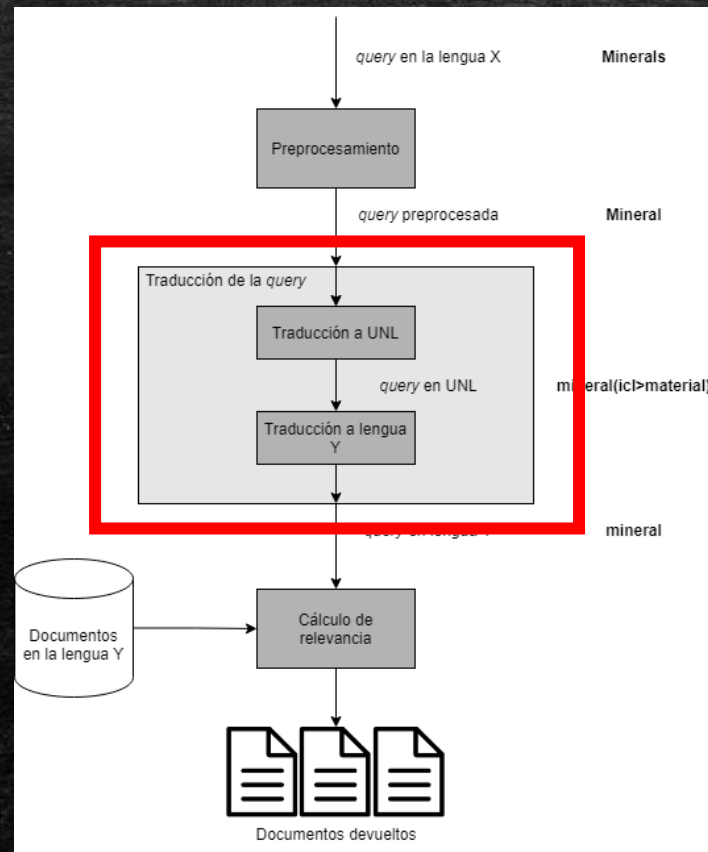
- "Esto son dos ejemplos para la presentación"
- ["esto", "son", "dos", "ejemplos", "para", "la", "presentación"]
- ["ejemplos", "presentación"]
- ["ejemplo", "presentación"]

Preprocesamiento: etiquetación morfológica



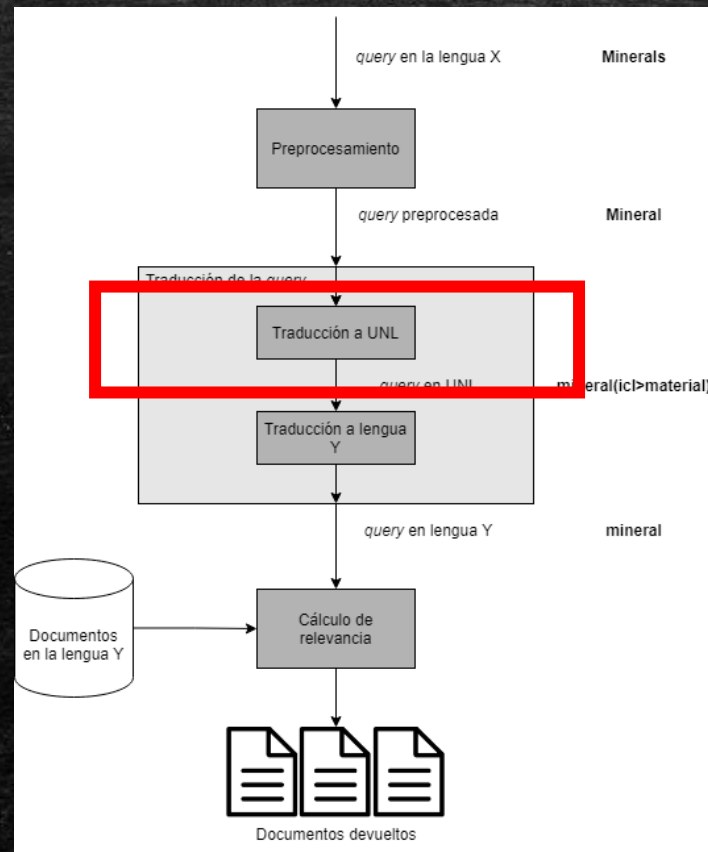
- "Esto son dos ejemplos para la presentación"
- ["esto", "son", "dos", "ejemplos", "para", "la", "presentación"]
- ["ejemplos", "presentación"]
- ["ejemplo", "presentación"]
- ["ejemplo":sustantivo, "presentación":sustantivo]

Conversión multilingüe



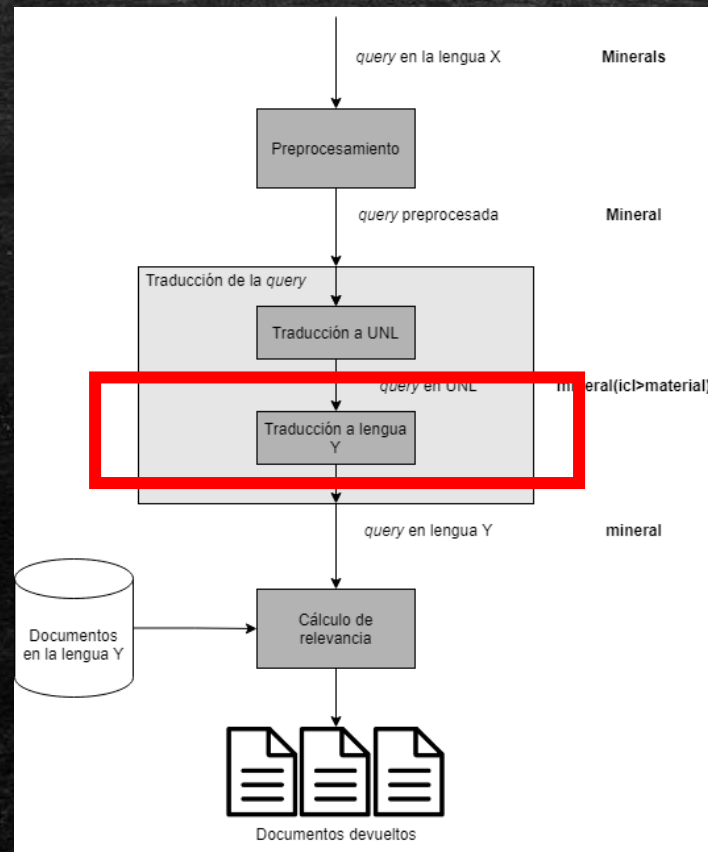
- Universal Networking Language
- En dos fases:
 - ✓ Inglés a UNL
 - ✓ UNL a Español

Conversión multilingüe: Inglés a UNL



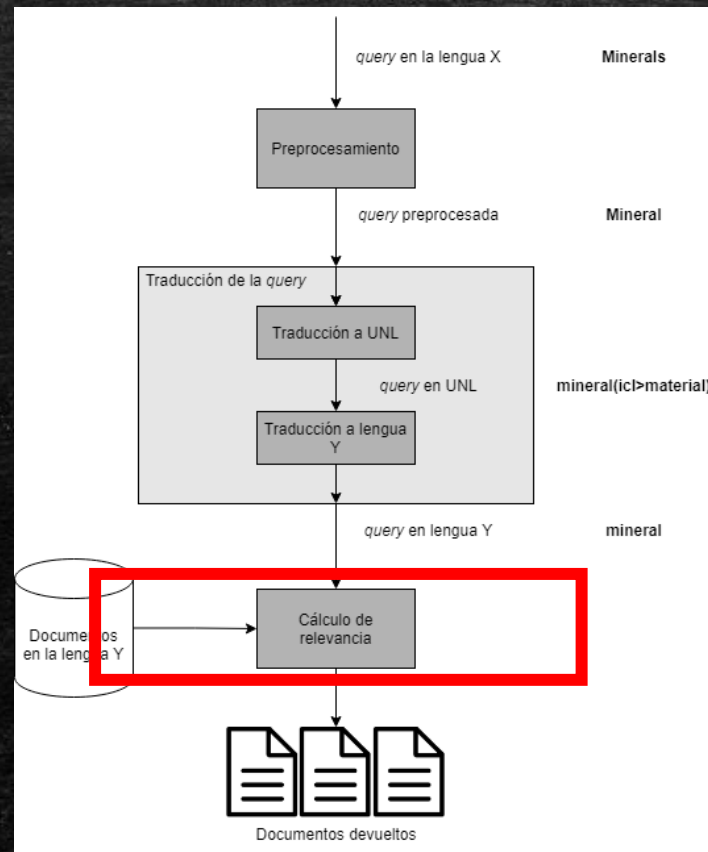
- ["example:sustantivo",
"presentation:sustantivo"]
- ["example(icl>information>
thing)",
"presentation(icl>represent
ation>thing,equ>display)"]

Conversión multilingüe: UNL a Español

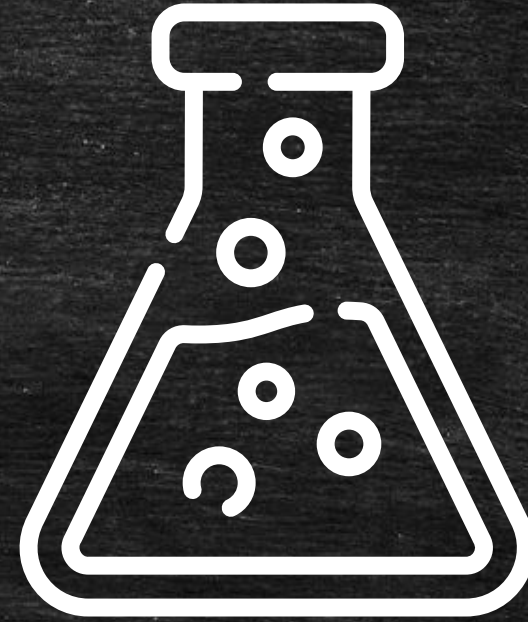


- ["example(icl>information>thing)", "presentation(icl>representation>thing, equ>display)"]
- ["ejemplo", "presentación"]

Buscador



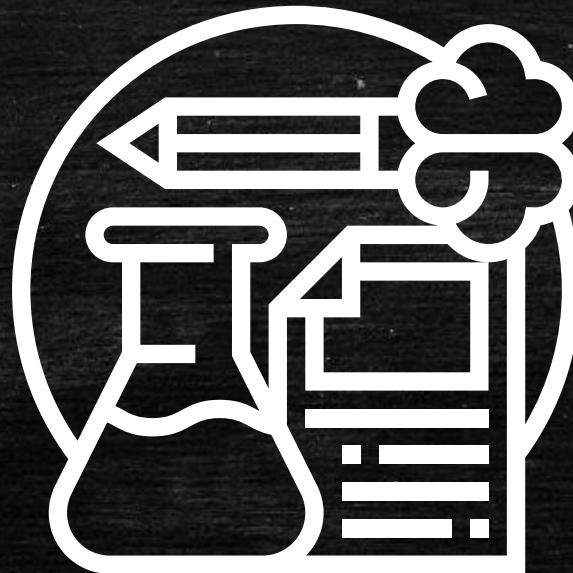
- TF-IDF
- Similitud del coseno



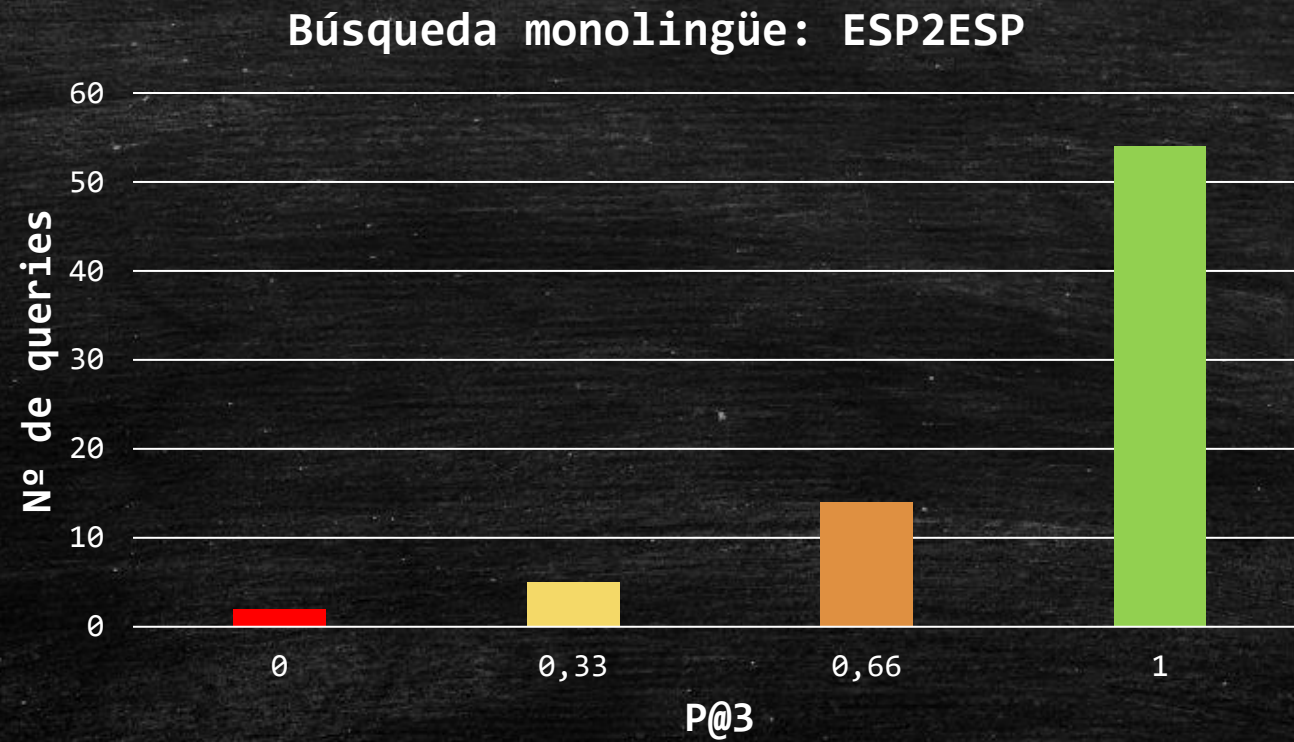
Experimentación y análisis

Experimentos a realizar

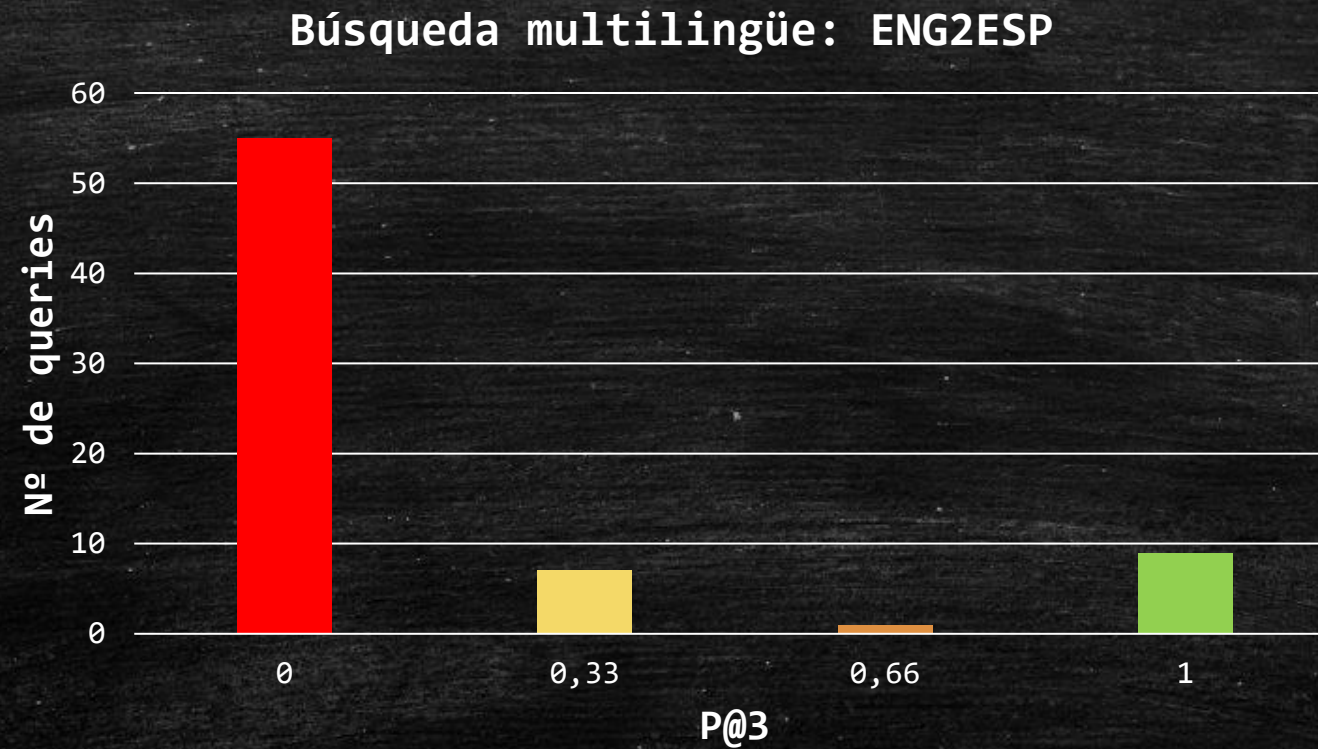
- Experimento N°1 : Búsqueda monolingüe vs búsqueda multilingüe
- Experimento N°2 : Calidad de la traducción
- Experimento N°3 : Búsqueda multilingüe vs multilingüe refinada



Experimento Nº1: Monolingüe vs Multilingüe

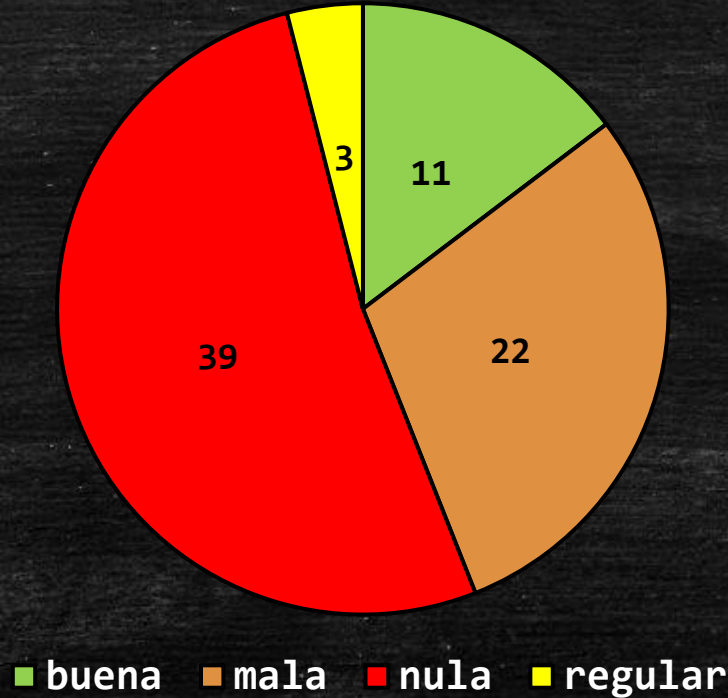


Experimento Nº1: Monolingüe vs Multilingüe



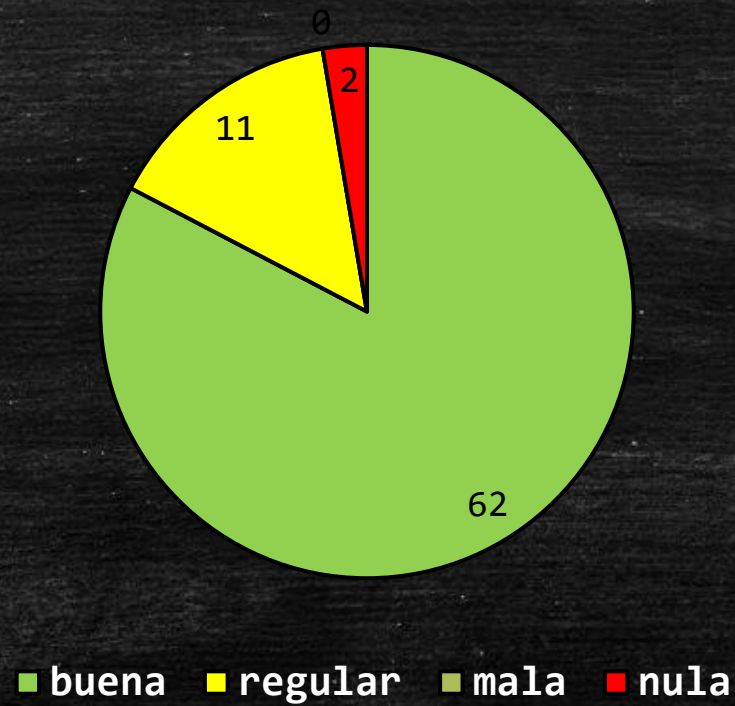
Experimento Nº2: Calidad de las traducciones

ENG2ESP: Calidad de las traducciones



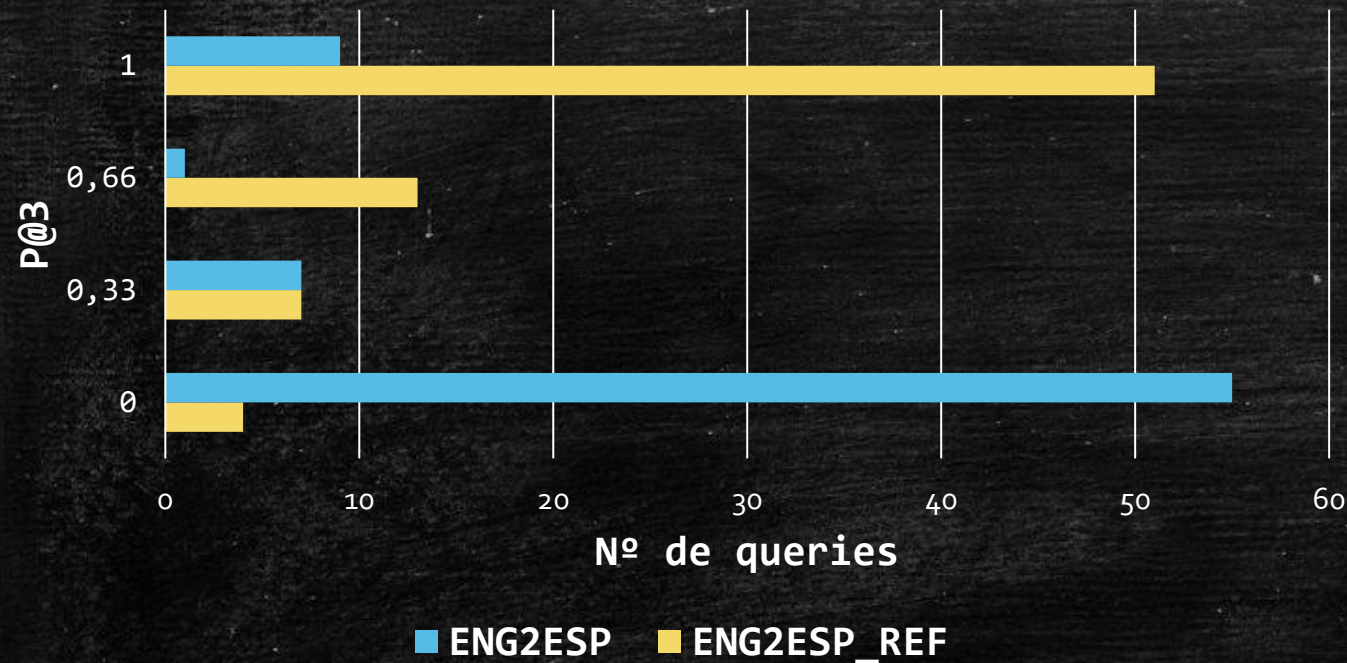
Experimento Nº3: Multilingüe refinada

ENG2ESP_REF: Calidad de las traducciones



Experimento Nº3: Multilingüe refinada

Búsqueda multilingüe vs multilingüe refinada



| | P@3 | 0 | 0.33 | 0.66 | 1 |
|-------------|-----|----|------|------|----|
| ESP2ESP | | 2 | 5 | 14 | 54 |
| ENG2ESP_REF | | 4 | 7 | 13 | 51 |
| Variación | | +2 | +2 | -1 | -3 |

Conclusiones

- Uso de una **interlingua** es un **enfoque válido**
- **Calidad** de los recursos léxicos **importa**
 - Cuello de botella
 - Facilidad de uso
- **Escalabilidad**
- No siempre los enfoques con **aprendizaje profundo** son los mejores

Trabajos futuros

- Incorporación de otros **recursos léxicos** --> expansión
- Técnica más elaborada de **desambiguación**
- **Detección automática** de la lengua de la búsqueda
- Ampliar el **tamaño de la colección** de los experimentos
- **Indexar** en UNL
- Modelo de buscador capaz de capturar la **semántica**

Demo



Muchas gracias por su atención



UNIVERSIDAD
POLITÉCNICA
DE MADRID

POLITÉCNICA



Uso de Interlinguas para Búsqueda Documental Multilingüe

Máster Universitario en Inteligencia Artificial

Autor: César García Cabeza
Tutor: Dr. D. Jesús Cardeñosa Lera

Julio 2021