CS6200 Information Retrieval:
Implementing an English-Spanish Cross-Lingual Information Retrieval System

Amie Kong
kong.am@northeastern.edu
Khoury College of Computer and Information Science
Northeastern University

April 27, 2021
Github Source Code: https://github.com/amiekong/cross-lingual-retrieval

Introduction

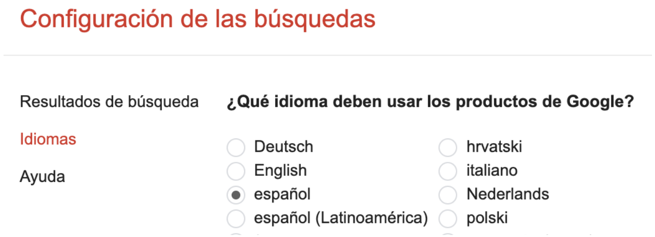
With online media being a huge information source, diverse information has created language barriers and hindrances in communication across different cultures. Google data shows that roughly 30% of media on the internet in the U.S. is consumed by browsers using English and Spanish interchangeably, whether it be in messages, searches, or viewing pages. Unfortunately, the U.S. is behind the rest of the world in going multilingual online due to the linguistic culture of prioritizing English and not other languages [3]. This serves as a disadvantage for bilingual citizens living in the U.S. who either want to retrieve documents in another foreign language to improve their fluency in a language or for children and families of first-generation immigrants who want to obtain relevant information and rely on the internet as their source. Furthermore, it also hinders second-language acquisition of learning a second language.

Cross-Language Information Retrieval deals with retrieving information written in a different language from the user's query language. CLIR systems use different techniques including: dictionary-based CLIR, parallel corpora based CLIR, comparable corpora based CLIR, and machine translator based CLIR. Much improvement has been done on CLIR since the first workshop during the SIGR96 conference that the "most accurate systems of cross lingual adhoc retrieval today are nearly as effective as monolingual systems" [5]. However, tasks such as translational ambiguity and phrase identification still pose challenges to CLIR systems.

Problem

One of the issues of using Twitter and Google search when filtering the results to a specific language is that it highly reflects on the language of the typed query. There is no option to translate the typed query to a specific language - only the results. So if a user wants to search for "popular food" and sets the results filter option to "Spanish," the top results that include the terms "popular food" are returned and users have the option to translate the article but there are no original Spanish articles, thus making the results biased towards the source language. In order for the user to have original Spanish articles returned for popular food, the query must be typed in the target language (in this case "comida popular").

(Changed search setting to Spanish.)



Configuración de las búsquedas

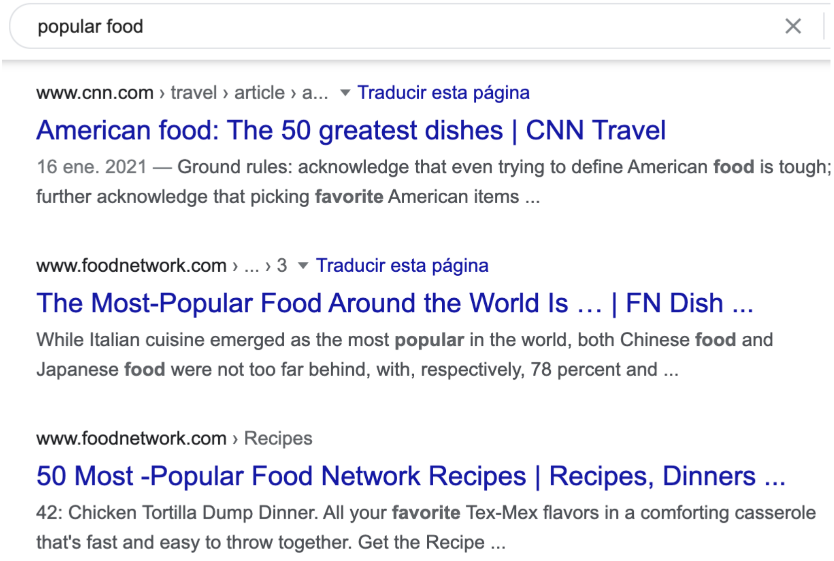
Resultados de búsqueda

Idiomas

Ayuda

¿Qué idioma deben usar los productos de Google?

○ Deutsch
○ English
● español
○ español (Latinoamérica)
○ hrvatski
○ italiano
○ Nederlands
○ polski



popular food

www.cnn.com › travel › article › a... ▾ Traducir esta página

American food: The 50 greatest dishes | CNN Travel

16 ene. 2021 — Ground rules: acknowledge that even trying to define American **food** is tough; further acknowledge that picking **favorite** American items ...

www.foodnetwork.com › ... › 3 ▾ Traducir esta página

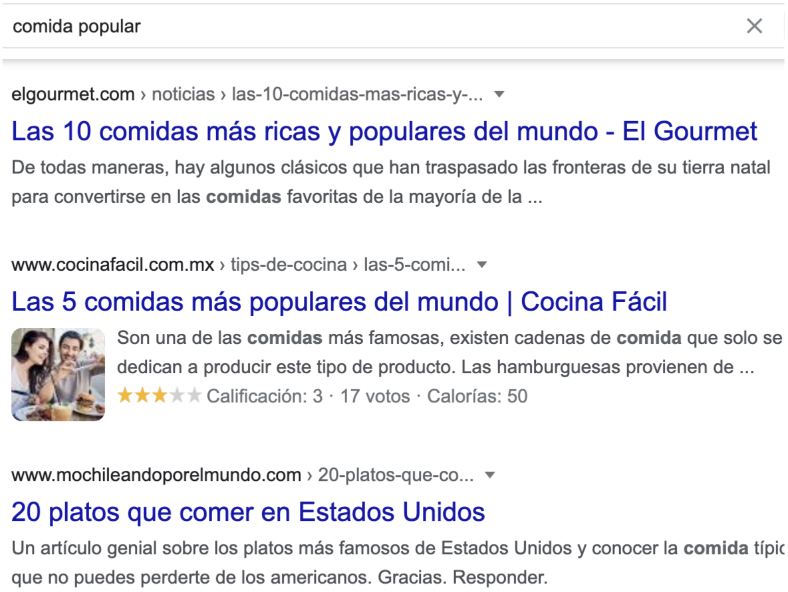
The Most-Popular Food Around the World Is ... | FN Dish ...

While Italian cuisine emerged as the most **popular** in the world, both Chinese **food** and Japanese **food** were not too far behind, with, respectively, 78 percent and ...

www.foodnetwork.com › Recipes

50 Most -Popular Food Network Recipes | Recipes, Dinners ...

42: Chicken Tortilla Dump Dinner. All your **favorite** Tex-Mex flavors in a comforting casserole that's fast and easy to throw together. Get the Recipe ...



comida popular

elgourmet.com › noticias › las-10-comidas-mas-ricas-y-... ▾

Las 10 comidas más ricas y populares del mundo - El Gourmet

De todas maneras, hay algunos clásicos que han traspasado las fronteras de su tierra natal para convertirse en las **comidas** favoritas de la mayoría de la ...

www.cocinafacil.com.mx › tips-de-cocina › las-5-comi... ▾

Las 5 comidas más populares del mundo | Cocina Fácil

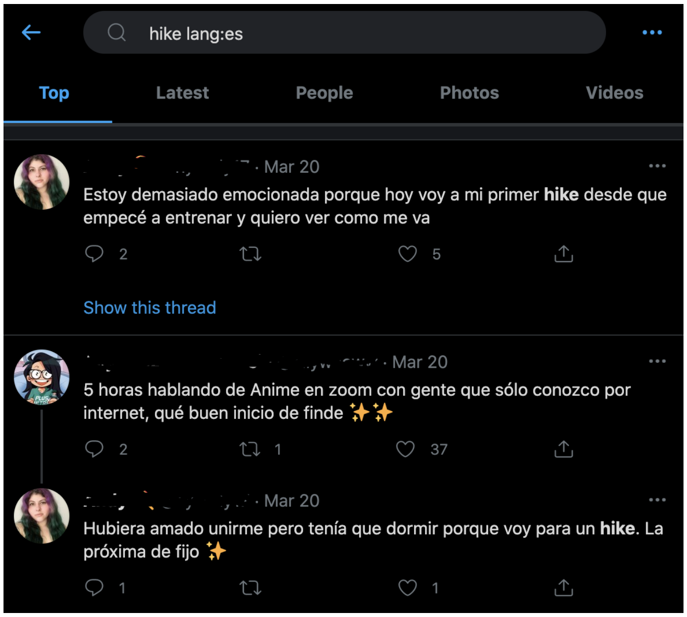
Son una de las **comidas** más famosas, existen cadenas de **comida** que solo se dedican a producir este tipo de producto. Las hamburguesas provienen de ...
★★★☆☆ Calificación: 3 · 17 votos · Calorías: 50

www.mochileandoporelmundo.com › 20-platos-que-co... ▾

20 platos que comer en Estados Unidos

Un artículo genial sobre los platos más famosos de Estados Unidos y conocer la **comida** típic que no puedes perderte de los americanos. Gracias. Responder.

Similarly, the same issue holds for Twitter when we filter the results to return for the Spanish language. However, in Twitter we see that when an English-typed query is entered bilingual posts are returned that include the specified English terms. For example, when “hike” is entered to return Spanish tweets, only bilingual Spanish posts that include “hike” are returned.



hike lang:es

[redacted] · Mar 20
Estoy demasiado emocionada porque hoy voy a mi primer **hike** desde que empecé a entrenar y quiero ver como me va

2 5

Show this thread

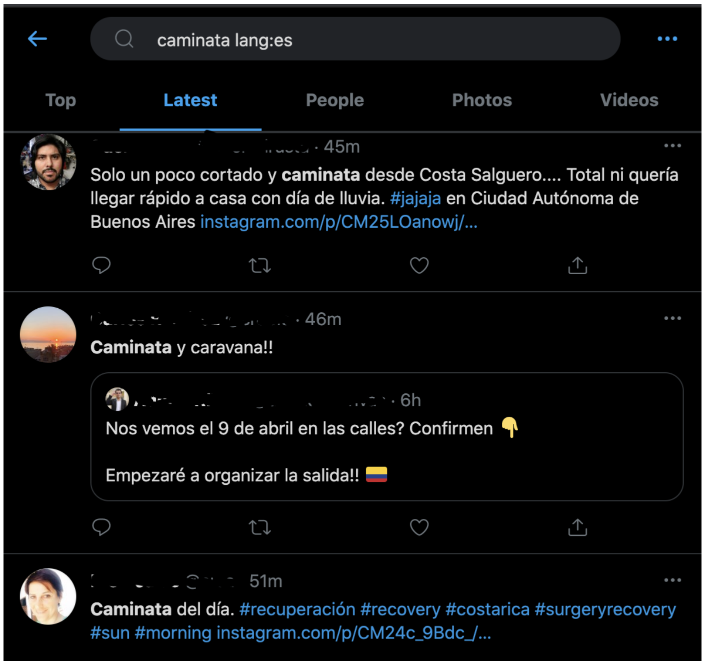
[redacted] · Mar 20
5 horas hablando de Anime en zoom con gente que sólo conozco por internet, qué buen inicio de finde ✨✨

2 1 37

[redacted] · Mar 20
Hubiera amado unirme pero tenía que dormir porque voy para un **hike**. La próxima de fijo ✨

1 1

When “caminata” or “senderismo” (derivations of “hike”) are entered to return Spanish tweets, we get the relevant posts that we want. It will be annoying for the user to have to use a translator just to copy & paste their query into a search box.



caminata lang:es

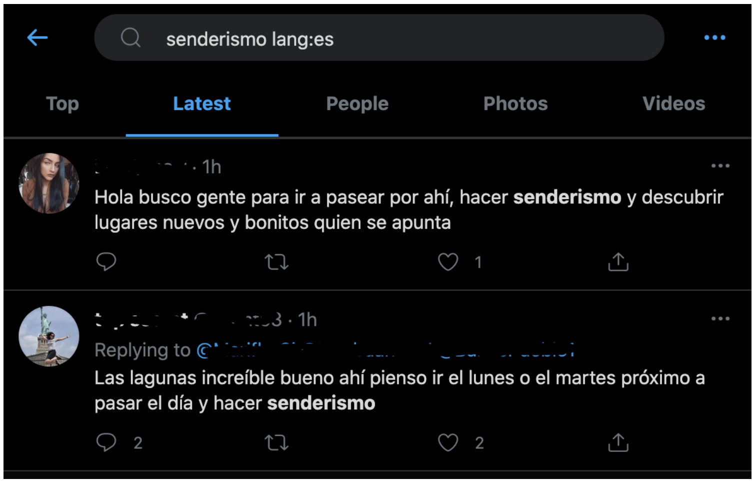
[redacted] · 45m
Solo un poco cortado y **caminata** desde Costa Salguero.... Total ni quería llegar rápido a casa con día de lluvia. #jajaja en Ciudad Autónoma de Buenos Aires instagram.com/p/CM25LOanowj/...

[redacted] · 46m
Caminata y caravana!!

[redacted] · 6h
Nos vemos el 9 de abril en las calles? Confirmen 👇

Empezaré a organizar la salida!! 🇨🇴

[redacted] · 51m
Caminata del día. #recuperación #recovery #costarica #surgeryrecovery #sun #morning instagram.com/p/CM24c_9Bdc_/...



senderismo lang:es

[redacted] · 1h

Hola busco gente para ir a pasear por ahí, hacer **senderismo** y descubrir lugares nuevos y bonitos quien se apunta

1

[redacted] · 1h

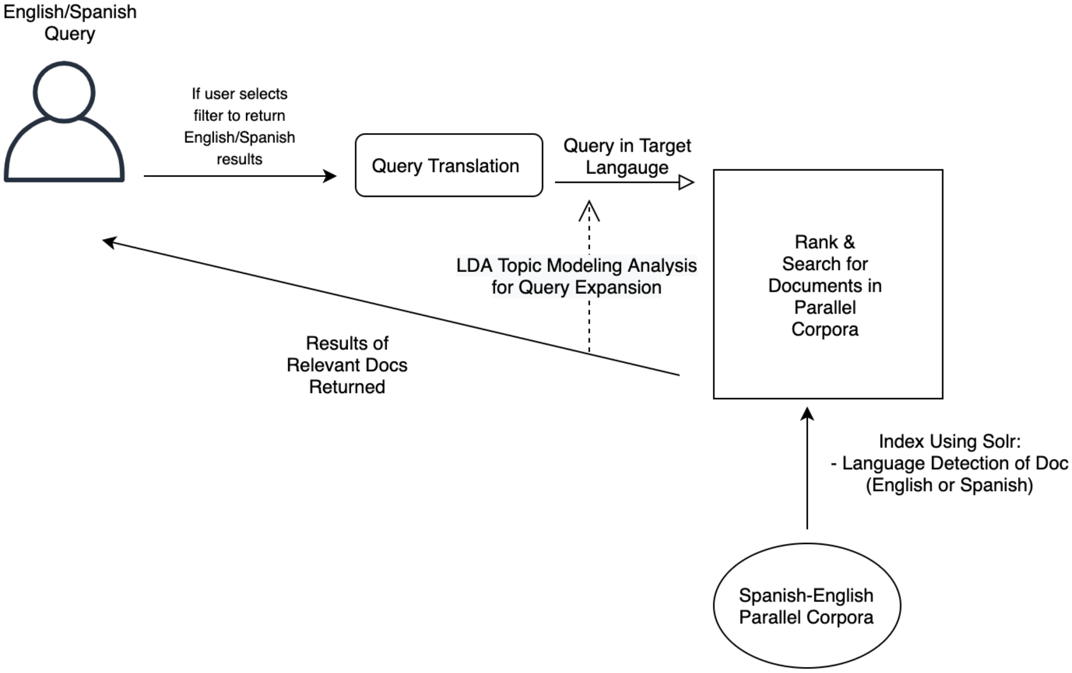
Replying to [redacted]

Las lagunas increíble bueno ahí pienso ir el lunes o el martes próximo a pasar el día y hacer **senderismo**

2 2

Solution and Methodology

Implementing a **Cross-Lingual Information Retrieval System** for **English-Spanish** where the returned results are in the specified target language regardless of the query's source language will be a solution to this problem. When a user enters an English query and wants to filter it to Spanish results it will only return Spanish results. The user can type an English or Spanish query and have the option to filter it to "return Spanish results" or "return Spanish results". The results returned will return a list of Spanish or English documents with the document title followed by the document text when the query is translated and indexed against the dataset. By extracting latent topic information from the returned target documents, an automatic **query expansion feature** will be enabled to reformulate a new query by adding an extra term or two in an attempt to improve the relevance ranking of the relevant documents retrieved if the precision of the results were low.


English/Spanish Query
If user selects filter to return English/Spanish results
Query Translation
Query in Target Langauge
Rank & Search for Documents in Parallel Corpora
LDA Topic Modeling Analysis for Query Expansion
Results of Relevant Docs Returned
Index Using Solr:
- Language Detection of Doc (English or Spanish)
Spanish-English Parallel Corpora

Parallel Corpora Dataset

A subset of the Medical Spanish-English Corpora (MeSpEn) that was presented at the *LREC 2018 Workshop MultilingualBIO: Multilingual Biomedical Text Processing* was used which contains aggregations of datasets from multiple sources such as IBECS, SciELO, Pubmed and MedlinePlus. Consisting of health related documents in Spanish and English, MeSpEn is useful for building parallel corpora for training and evaluating Spanish-English medical machine translation systems, as well as generating multilingual automatic term extraction tools. The data set includes Spanish and Latin American biomedical and clinical literature along with content with information about diseases, conditions, and wellness issues for patients [1].

Specifically, the data set used for this project was MedlinePlus in TEI format, consisting of clean raw text and XML files of each article, structured by sections and paragraphs on topics limited to diseases, illnesses, symptoms, injuries, surgeries, health conditions, wellness issues, drugs herbs and supplements. The raw text of 11,157 articles in English and Spanish were collected and imported into the Solr instance using a Python program (*combiner.py*) that serves to combine the text files into two separate XML files (English documents and Spanish documents) that will be used to add the documents to the Solr instance [2].

Implementing a Cross-Language Information Retrieval System (CLIR)

Accepting questions in one language (in this case English) and retrieving information in a different language (e.g. Spanish) defines CLIR. There are two different approaches to handle CLIR: translate the source language query into the target language and then retrieve the documents (query translation) or translate the entire corpora in the source language and then perform the retrieval (document translation); however, the second option requires a lot of resources and time so the first approach of **query translation** will be used. Translational ambiguity is expected in query translation, especially for short query text due to the limited context. After translation is done, the task is then reduced into a monolingual IR task [4].

Handling Multiple Languages in a Single Index

To handle multiple languages in the Solr core instance, two separate fields for Spanish and English text (“text_en” and “text_es”) were included in the managed_schema file, along with the field types and appropriate Stemmers for Spanish and English [6].

Query Translation

A Python tool (*deep-translator*) that uses multiple translators was installed to translate the detected source language of the query to the target language. The translated query was then used to search against the Solr instance.

Pseudo-Relevance Feedback: Query Expansion Based On Topic Distributions of Retrieved Documents

Using pseudo-relevance feedback (PRF), the user’s new formulated query will be based on the top-ranked retrieved documents in the first retrieval round. Terms will be extracted to enhance the user’s requirement from the top-ranked documents in the first retrieval round and then expand a query used in the next retrieval round. PRF has shown an increase in retrieval performance by several studies [7]. A Latent Dirichlet Allocation Topic Model from scikit, along with preprocessing and tokenization of the retrieved documents was used to create a bag-of-words model and perform topic distribution of the top retrieved documents and a new query was reevaluated in the final round of ranking (*lda.py*).

Ranking Using Solr-Lucene

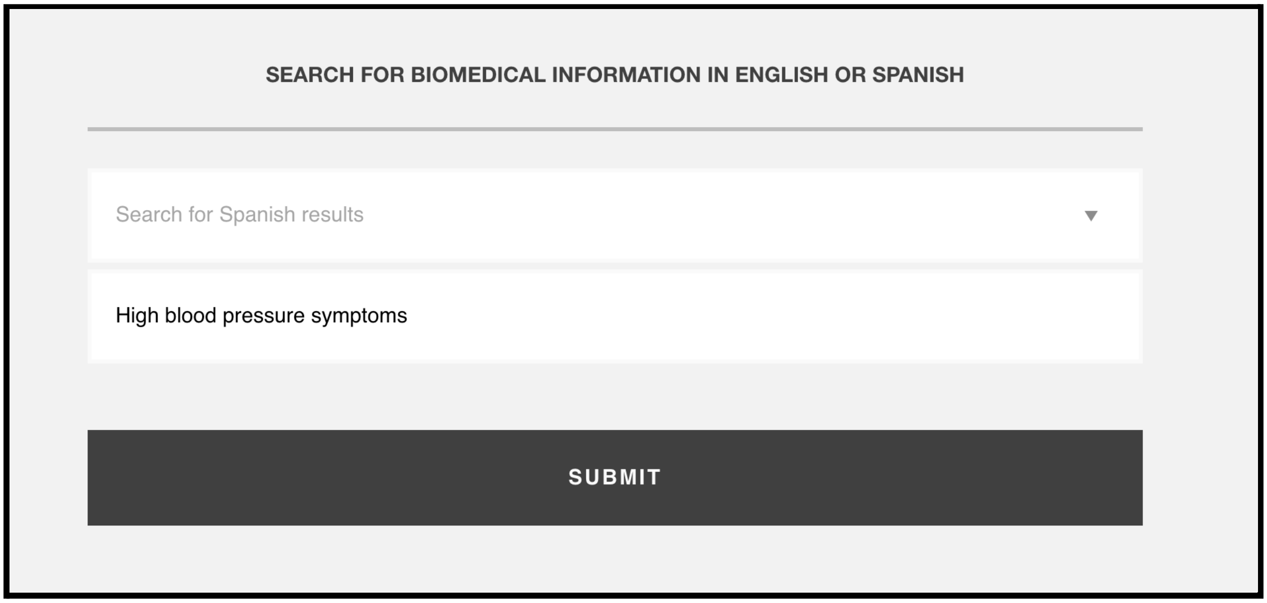
A similarity model using the built-in TF-IDF scoring algorithm in Lucene was used to give relevance scores to each document in the search result and rank the documents.

User Interface

Flask and **HTML/CSS** were used to create the user interface. There are two web pages for the user to view. Depending on the target language of the returned results, one webpage will include headers and text in Spanish stored in the *resultos.html* template and the English results will include headers and text in English stored in the *results.html* template. The query from the text field is retrieved from the HTML, along with the filter values of the scroll-down form to denote the target language of the results. There is a button on the results page that allows the user to go back and enter in a new query.

Results

Input of English query of *high blood pressure symptoms* in search text field and set filter to return Spanish results are prompted:

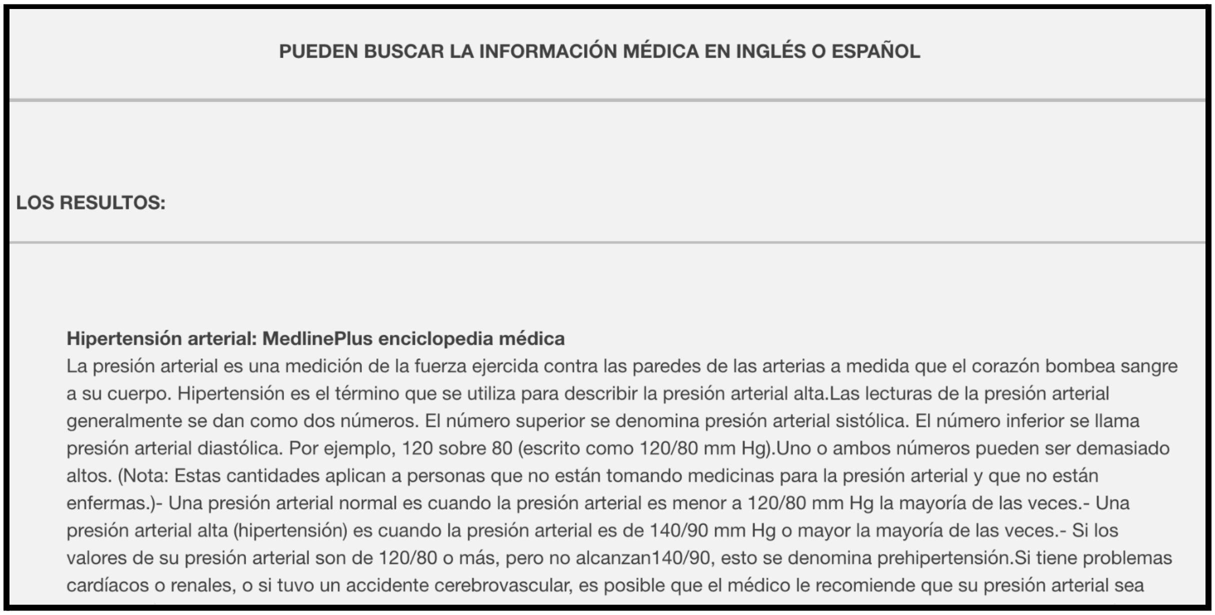


SEARCH FOR BIOMEDICAL INFORMATION IN ENGLISH OR SPANISH

Search for Spanish results ▼

High blood pressure symptoms

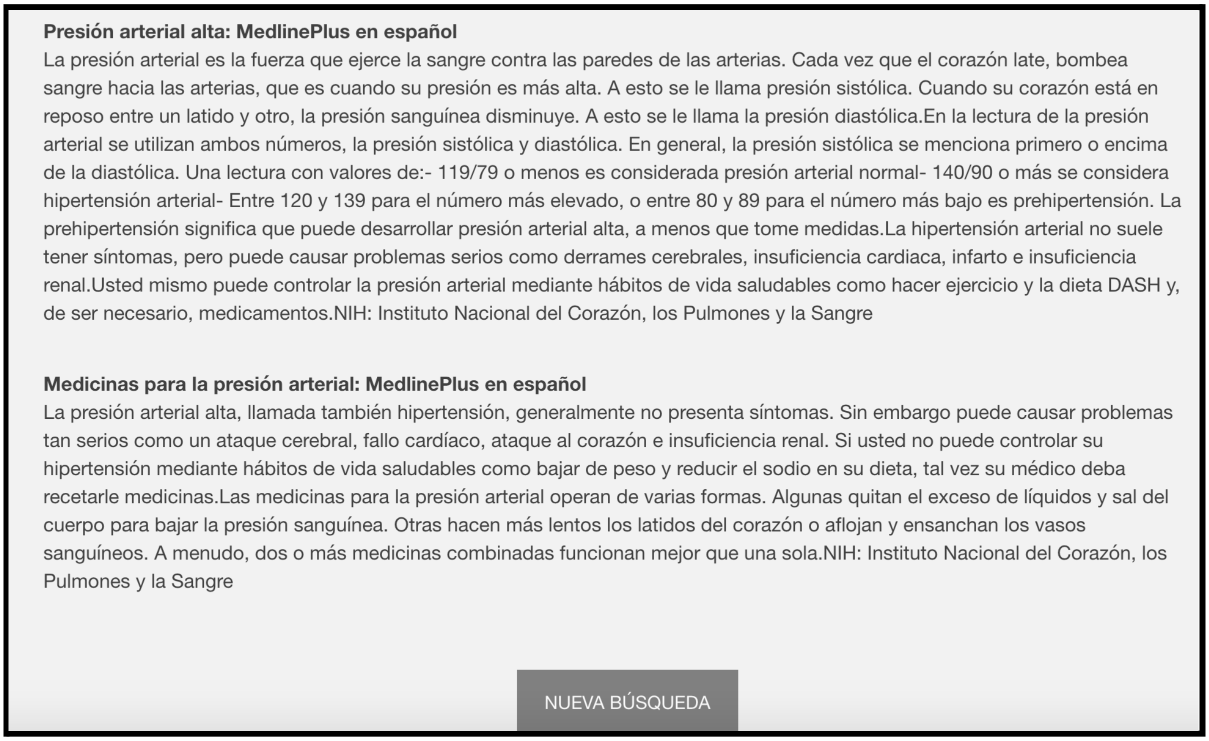
Returned relevant results of the query in the Spanish language are returned. The results are relevant for this input query since the documents shown are articles that include *hipiertensión arterial* and *presión arterial alta* which translates to *hypertension* and *high blood pressure* (10 relevant results were returned). We also see that only Spanish results are returned as specified by the user, which indicates that we have resolved the problem of search engine filters not returning the correct target language documents.



PUEDEN BUSCAR LA INFORMACIÓN MÉDICA EN INGLÉS O ESPAÑOL

LOS RESULTOS:

Hipertensión arterial: MedlinePlus enciclopedia médica
La presión arterial es una medición de la fuerza ejercida contra las paredes de las arterias a medida que el corazón bombea sangre a su cuerpo. Hipertensión es el término que se utiliza para describir la presión arterial alta.Las lecturas de la presión arterial generalmente se dan como dos números. El número superior se denomina presión arterial sistólica. El número inferior se llama presión arterial diastólica. Por ejemplo, 120 sobre 80 (escrito como 120/80 mm Hg).Uno o ambos números pueden ser demasiado altos. (Nota: Estas cantidades aplican a personas que no están tomando medicinas para la presión arterial y que no están enfermas.)- Una presión arterial normal es cuando la presión arterial es menor a 120/80 mm Hg la mayoría de las veces.- Una presión arterial alta (hipertensión) es cuando la presión arterial es de 140/90 mm Hg o mayor la mayoría de las veces.- Si los valores de su presión arterial son de 120/80 o más, pero no alcanzan140/90, esto se denomina prehipertensión.Si tiene problemas cardíacos o renales, o si tuvo un accidente cerebrovascular, es posible que el médico le recomiende que su presión arterial sea



Presión arterial alta: MedlinePlus en español
La presión arterial es la fuerza que ejerce la sangre contra las paredes de las arterias. Cada vez que el corazón late, bombea sangre hacia las arterias, que es cuando su presión es más alta. A esto se le llama presión sistólica. Cuando su corazón está en reposo entre un latido y otro, la presión sanguínea disminuye. A esto se le llama la presión diastólica.En la lectura de la presión arterial se utilizan ambos números, la presión sistólica y diastólica. En general, la presión sistólica se menciona primero o encima de la diastólica. Una lectura con valores de:- 119/79 o menos es considerada presión arterial normal- 140/90 o más se considera hipertensión arterial- Entre 120 y 139 para el número más elevado, o entre 80 y 89 para el número más bajo es prehipertensión. La prehipertensión significa que puede desarrollar presión arterial alta, a menos que tome medidas.La hipertensión arterial no suele tener síntomas, pero puede causar problemas serios como derrames cerebrales, insuficiencia cardiaca, infarto e insuficiencia renal.Usted mismo puede controlar la presión arterial mediante hábitos de vida saludables como hacer ejercicio y la dieta DASH y, de ser necesario, medicamentos.NIH: Instituto Nacional del Corazón, los Pulmones y la Sangre

Medicinas para la presión arterial: MedlinePlus en español
La presión arterial alta, llamada también hipertensión, generalmente no presenta síntomas. Sin embargo puede causar problemas tan serios como un ataque cerebral, fallo cardíaco, ataque al corazón e insuficiencia renal. Si usted no puede controlar su hipertensión mediante hábitos de vida saludables como bajar de peso y reducir el sodio en su dieta, tal vez su médico deba recetarle medicinas.Las medicinas para la presión arterial operan de varias formas. Algunas quitan el exceso de líquidos y sal del cuerpo para bajar la presión sanguínea. Otras hacen más lentos los latidos del corazón o aflojan y ensanchan los vasos sanguíneos. A menudo, dos o más medicinas combinadas funcionan mejor que una sola.NIH: Instituto Nacional del Corazón, los Pulmones y la Sangre

Sample queries including: high blood pressure, heart disease, jaundice, Mad Cow Disease and cystic fibrosis performed well on the cross-lingual search engine, while other queries (iron deficiency, labor) did not perform well due to translational ambiguity.

Conclusion

This English-Spanish Cross-Language Information Retrieval System for medical journals and articles was built as an application to resolve the problem with popular search engines in handling information with multiple languages. The goal of implementing a CLIR system where the returned results are in the specified target language regardless of the query's source language was achieved. However, PRF using query expansion of the topic distribution of retrieved documents did not affect the results of this retrieval system due to the low quality feedback of the returned documents since the performance relies heavily on the number and feedback quality of documents. There are improvements that can be made to this system to deal with translational ambiguity such as performing a global topic model analysis on the entire corpora and using that to improve the query by suggesting common medical topics in the documents.

References

[1] Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, & Martin Krallinger. (2019). MeSpEn_Parallel-Corpora (Version 2019-12-04) [Data set]. Presented at the LREC 2018 Workshop MultilingualBIO: Multilingual Biomedical Text Processing (MultilingualBIO), Miyazaki, Japan: Zenodo. http://doi.org/10.5281/zenodo.3562536

[2] https://github.com/amiekong/cross-lingual-retrieval

[3] Leddy, M. (2019, October 1). *How to target the US Spanish-English bilingual ecommerce market*. Weglot blog. https://blog.weglot.com/targeting-us-bilingual-spanish-english-market-ecommerce-retailers/

[4] Croft, W., & Wei, X. (2007). Topic models in information retrieval.

[5] Zhang, R. (2019, March 7). *A Brief Introduction to Cross-Lingual Information Retrieval*. Medium. https://medium.com/lily-lab/a-brief-introduction-to-cross-lingual-information-retrieval-eba767fa9af6.

[6] https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781849512183/1/ch01lvl1sec12/handling-multiple-languages-in-a-single-index

[7] Serizawa M., Kobayashi I. (2013) A Study on Query Expansion Based on Topic Distributions of Retrieved Documents. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science, vol 7817. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37256-8_31