Thought exercise

How to improve search using Natural Language Processing? Can you suggest one approach?

Google search is the most used search engine on the World Wide Web across all platforms, with 92.16% market share as of December 2020 handling more than 5.4 billion searches each day. Google Search also provides many different options for customized searches, using symbols to include, exclude, specify or require certain search behavior, and offers specialized interactive experiences, such as flight status and package tracking, weather forecasts, currency, unit, and time conversions, word definitions, and more.

Text search is easy to use, but there is some vagueness of languages that causes search results to be inaccurate. Thus we need to improve change the approach to provide more accurate search results. The most common approach is to get the specifics of the search query by adding filters: new input fields, checkboxes, radio buttons, etc. This can help us to know what exactly the user is looking for and provide accurate results. However, by introducing such filters we are making it very tedious for the user to input every detail. Therefore, we would like to stick to a single text field and still produce accurate results. This can be done using Natural Language Processing (NLP). We need an end-to-end learning system that understands user queries and provides the appropriate results. We need to provide a structure to the user query and then give this as an input to the system.

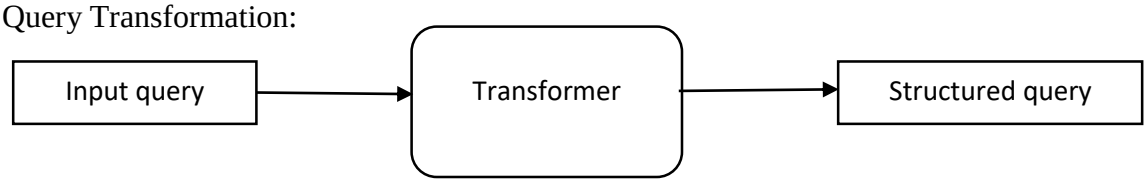
Some of the problems associated with current systems are:

- Token matching: Suppose I am looking for formal shirts, and I query the system and I get results of polo shirts. As a user, I am currently looking for formal shirts but there is a complete mismatch.
- Contextualization: Suppose I am searching for books online and query with just ‘NLP’. As a user, I am interested in learning about Natural Language Processing, whereas the system shows me books on Neuro-linguistic programming.
- Query misunderstanding: Because of the ordering or sequence of words the query might retrieve wrong results.

These problems can be solved by:

- Query Auto-completion: We can help the user auto-complete the search query. By doing so we can eliminate those queries that lead to very fewer results or possibly no results
- Alternate query generation: We can transform the user query into an alternate query and show the user the required results.

Query Transformation:



- Using synonyms: The problem of token matching can be solved by replacing the words with synonym words through a custom dictionary. We find the Part of speech using a library like Spacy and get synonyms for words that have Part of speech as noun, verb, or adjective. We have to keep a cutoff of cosine similarity for selecting similar words to avoid too many or irrelevant synonyms.
- The problem of contextualization can be solved by using User history, User geography, changes information