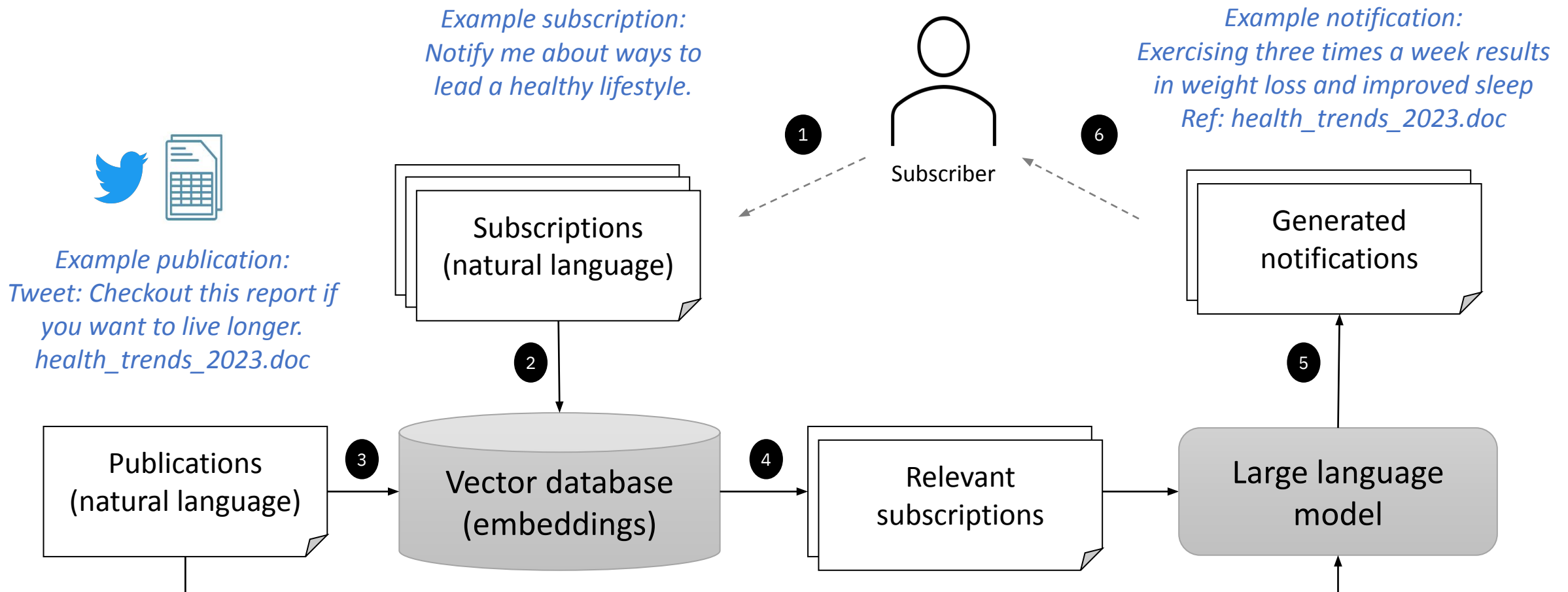


Publish-subscribe with large language models (LLMs): Improving expressiveness with natural language and generated custom

Vatche Isahagian, Vinod Muthusamy, Aleksander Slominski

IBM Research

Proposed LLM-based pub/sub architecture



Overview

- Match publications and subscriptions by comparing embeddings
 - Index subscriptions in a vector database
- Publications can be NL documents □ expand the set of events that can be processed
- Subscriptions can be expressed in NL □ more intuitive, lower barrier to entry, ability to express complex interests
 - E.g., no need to specify event source; system identifies relevant events across all sources
- Notifications are generated and customized □ more personalized and relevant notifications

Examples from the nq_open dataset. The first set motivates the need for generated NL notifications. The second set motivates the need for stateful subscriptions.

Subscription / Query	Matching publications / passages (snippets)	Generated notification
Who has the most world series wins in mlb history	As of 2017, the World Series has been contested 113 times ... In the American League, the New York Yankees have played in 40 World Series and won 27 ... In the National League ... the Brooklyn / Los Angeles Dodgers have appeared in 19 and won 6 ...	New York Yankees
Who has won the most baseball world series		New York Yankees
Has the la dogers ever won a world series		Yes
What are the 2 parks at euro disney	Walt Disney Studios Park, opened on 16 March 2002 celebrating showbusiness, films, and behind - the - scenes Shopping, dining, and entertainment ...	Disneyland Park, opened with the resort on 12 April 1992 ... Walt Disney Studios Park, opened on 16 March 2002
	... encompasses two theme parks, many resort hotels, a shopping, dining, and entertainment complex, and a golf course ... Disneyland Park is the original theme park of the complex, opening with the resort on 12 April 1992	

Conclusions

- Using LLMs for pub/sub matching and notifications is a promising approach, particularly attractive to non-technical users as it hides the technical details of writing a query and formatting the output
- There is a plethora of unstructured data sources, some of which have streams of data
 - new papers published arXiv
 - press articles
 - support tickets
- Users want to be notified when data of interest is generated
- Users want to answers to their questions, not just a list of matching publications
 - E.g., “What are the potential users of my customer database”
 - Examine github issues to determine interest in accessing this database

Challenges

- **Stateful subscriptions:** matching is done based on one publication and it would be useful to remember which publications have already matched a subscription to generate better notifications
- **New covering techniques** are needed when subscriptions are expressed in NL to scale pub/sub systems is to apply subscription covering in order to generate a small number of synthetic subscriptions that aggregate the interests of multiple subscriptions.
- **Creating dataset with NL queries and results based on documents**, such as from Wikipedia or arXiv, would benefit the community and serve as a benchmark to guide future research.
- **Multiple modalities:** foundation models are evolving to support multiple modalities such as video, image, audio, and text. These models should help further increase the expressiveness, improve the ease of use, and extend the applicability of pub/sub systems.