Dictation Evaluation Reddit Parser

Garcia, Benjamin Lembke, Logan MacMillan, Kyle
Smith, Christopher Stelter, Andrew

September 21, 2018

Contents

1 Introduction to DERP

Dictation Evaluation Reddit Parser – DERP, is designed to be a Domain Specific Language – DSL for interpreting news, or news-like websites. The system was designed as part of a group project for South Dakota School of Mines & Technology's Compilers course.

1.1 What is DERP: Human Queries For a Digital World

Time is something we are regularly lacking and it would be nice to expedite the intake of online news while doing other things such as preparing and eating breakfast, taking a shower, getting dressed, or driving to work. We should be able to say to our phone, "What's going on in the world today?" and get a response containing relevant information that we care about; not just "I'm sorry, I don't understand 'What's going on in the world today'". If we want news on a specific topic, such as Tesla, it should be as easy as saying, "What's in the news about Tesla?".

The formats of news intake are largely unchanged in recent years; sure, the internet provides a way for news to be shared, but we still have to hunt down the stories we want, whether we get it through a radio or television broadcast, an online or magazine article, or a newspaper. These broadcasts are rigid and we have no control over what news we hear and there are frequently topics that don't concern us. Articles and newspapers are the same way but we are able to skip what we don't like. The issue with newspapers is we are limited to what the editor decided to include. The issue with articles is that it's down to us to track down relevant stories. A common way to track down interesting articles is an aggregation site such as reddit, steemit, band, and voat to name a few.

1.2 DERP Origins

In the fall of 2018 we were tasked with making a DSL for a class in compiler theory. We discussed many possible topics:

- Constrain/Generate
  - Floor Plans
  - Workout Assistant
  - Computer Part Picker
- Perform a Task
  - Robot "AI" Proposal by Dr. Hinker
  - Simple Image Processing
- Configure a Task

- News Reader
- SQL Helper

We decided against a Constrain/Generate model because the ones we came up with all required labeled data which we did not have access to and were not sure we could find or make in a timely manner.

Performing a task seemed like a good idea but the idea of being reliant on getting a robot (in the case of Dr. Hinker's proposal) seemed like a bad idea. We were also unsure of the code that went into it and if we'd be walking into a well documented and fleshed out product or a hack. The image processing was to be something along the lines of simplistic image manipulation.

Lastly we discussed configuration-style tasks. The two proposals were News Reader and an SQL Helper. The basic idea with the news reader is that a user could create custom queries that would obtain news that he/she was interested in while skipping the rest. SQL Helper was an idea to have anyone use SQL. So if a businessman needed some form of report he could ask his secretary to make it and they'd be able to without knowing SQL. The headache of building a system that was robust enough to translate human queries into SQL was deemed too much for the short amount of time we had and so we settled on the News Reader.

We needed a name. Having selected our project we could now determine what to name our team. We settled on Dictation Evaluation Reddit Parser because if we have enough time we'd like it to take spoken words to evaluate. The reddit bit was added because our minimum viable product will be built around reddit.

1.3 Why DERP

DERP allows a developer to create an interface between a person and a feed site such as Reddit, steemit, BAND, and Voat, as described in What is DERP section. This interface allows for a programmer to develop an application that cleanly links an end-user to the feed site(s) of their choice, without the user even being aware that they are using DERP. Without DERP a developer would have to build their own parser to understand what a user wants. DERP allows for a consistent platform for the developer to begin from. Lastly, don't spend time building a solution when one's available for free.

2 DERP Design Goals

2.1 Consolidated Information

The DERP project provides an intuitive way to multiplex multiple news sources into a personalized stream of data. Using DERP, one can specify as many different sources as they want to (provided language plugins are available), and then all of those sources are accessed through the same set of language keywords, regardless of if the source is a subreddit, a news website, or just a file on the user's device. Furthermore, DERP facilitates naming groups of sources and queries to create query macros, allowing users to further personalize what they get out of DERP.

Online news sources regularly expose the same categories of information - date, author, title - DERP acts as an interface for all different article types, allowing users to work only with this high-level information about the articles they are finding. Hiding the details of what exactly is required to find an article that matches a specific criterion allows users to focus more on what they want to read, and less on how they obtain that information.

2.2 Easy to Use Natural English Interface

The second goal of DERP is to design a language that won't feel strange to speak. All of the DERP keywords and syntax are similar to natural English speech patterns. This allows for easy adaptation of DERP into speech-recognition tools such as Google Assistant and Amazon Alexa. Using the natural form of the language, users with one of these devices could speak their program to the interpreter and receive their results immediately.

In addition to being easy to command DERP through natural language, DERP provides output that also feels natural. DERP has a set of phrases available to it for reporting errors or results from user queries. The main output from DERP, articles the user has requested, are outputted as full text so that devices reading them using a text-to-speech system sound as natural as possible.

2.3 Search More, Search Faster

By amalgamating news sources into a personalized feed, DERP allows users to find the information they want with fewer operations. Rather than check each of their favorite subreddits, news sites, and RSS feeds, a user can simply load those sources into DERP and make a general query about them. DERP will do the heavy lifting of finding things the user might be interested, which means the user will be able to spend more of their time actually consuming the information provided and deciding on new topics to get information over.

Because it is an extensible language, DERP users are limited only by the language plugins they use. While some keywords are understood specially by the language, such as those pertaining to dates and times, any language plugin

can provide additional fields and keywords, making the language flexible and powerful while keeping the complexity out of the core of the language.

3 The DERP Execution Model

3.1 Interactive

3.2 Hierachical Data (website → boards (subreddits) → posts

3.3 Constraints

3.4 Savable Criteria (data structures)

3.5 Composed Constraints

4 The DERP Query Language

4.1 Main Features

1. Syntax is similar to natural language
2. Syntax is compatible with integration with voice control
3. Ability to save selections, criteria, and sources
4. Ability to combine saved selections, criteria, and sources
5. Can value copy saved selections into the selection currently being built
6. Modal interpreter
7. Extensible source design, defaults support Reddit, can create modules for other sources
8. Syntax is tolerant of articles and some 'natural' variants of keywords
9. Capable of creating a selection on a variety of criteria (fields?)
10. tolerant of missing fields on merged selections (i.e. posts are missing tags)
11. ordering of results can be controlled, with source specific defaults

4.2 Primitives

There are N primitives in the language

1. Date - field type used for queries based on date fields
2. Number - field type used for queries on numeric fields
3. Boolean - field type used for queries on true/false fields
4. String - field type used for queries on text fields

4.3 Higher Order Types...

1. Selection - represents a complete query
2. Filter - represents a predicate for removing elements from a query's results
3. Source - represents a data source, expected fields, a default ordering, and other details specific to the source

4.4 Memory Management

Memory management is automatic, and handled by the Python runtime. Selections, Sources, and Filters are persisted to files, and loaded into their representative objects at runtime.

4.5 Derp Operators

4.5.1 String Operators

- with the exact - String comparison equality operator. Includes results where the specified field is an exact match for the provided string.
- like - String comparison fuzzy equality operator. Includes results where the specified field is an approximate match for the provided string.
- in the - Sub-String comparison equality operator. Includes results where a sub-string of the specified field is an exact match for the provided string.

4.5.2 Date Operators

- Date on - Date comparison equality operator. Includes results where the specified field is exactly the same date as the provided date.
- date after - Date comparison greater-than operator. Includes results where the specified field is strictly after the provided date.
- date before - Date comparison less-than operator. Includes results where the specified field is strictly before the provided date.

4.5.3 Boolean Operators

- which are | are - Boolean comparison equality operator. Includes results where the specified field contains the boolean value of 'true'.
- which are not | are not - Boolean comparison non-equality operator. Includes results where the specified field contains the boolean value of 'false'.

4.5.4 Numeric Operators

- with exactly | exactly - Number comparison equality operator. Includes results where the specified field contains the same numeric value as the provided numeric value.
- with over | over - Number comparison greater-than operator. Includes results where the specified field contains a value strictly greater than the provided numeric value.
- with under | under - Number comparison less-than operator. Includes results where the specified field contains a value strictly less than the provided numeric value.
- with roughly | roughly - Number comparison epsilon equality operator. Includes results where the specified field contains a value within an epsilon value greater or less than the provided numeric value.

4.5.5 Misc Operators

- matching - Criteria composition operator. The criteria corresponding to the provided name will be textually included into the current criteria or selection.
- from - Selection designation operator. The selection corresponding to the provided name will be textually included into the current selection.
- and | or - Combine the results from the statements on either side of the operator.

4.5.6 Keywords

- exit - Exits the program, and may only be used in the mode selection mode.
- stop - end mode keyword. Ends selection or criteria creation mode and clears the active state.
- clear - reset mode state keyword. Without leaving the current creation mode, clears the active state.
- recall - (in create mode) read back the current state. Will read back all statements that have been entered.
  (in mode selection) read back specified selection or criteria as if it were the active interpreter state.
- save as - store the current selection or criteria with a specified name. A saved selection or criteria may be used in the creation of other selections and criteria.
- read - (in create mode) execute the selection in its current state and present the results.
  (in mode selection) execute the specified selection and present the results.
- add - add a (set of) statements that include or exclude results from a source to a selection or criteria.
- remove - add a (set of) statements that exclude results from one or more sources to a selection or criteria.s