

Dialogue Systems in Hindi: Weather

Shelly Jain, 20171008

Dialogue systems

— — —

- Intent definition - narrow down the scenarios addressed; automatically identify intents in queries
- Entity extraction - key aspect in defining user needs; entities must be defined/described for later identification
- Dialogue resolution - address the user request; system must gather all the required information
- Response generation - solve the query and offer response
- Conversational flow - flow designed to direct user to some objective

This system

For rule-based system, queries answered based on rules.
Frame-based systems are mixed initiative - use the frame structure to guide dialogue. User can ask multiple queries at once; systems also questions user to fill required slots. Once frame is full, database query is done.

Data structures store relevant data - one (detailed, with more fields) for current day data, one (with fewer fields) for weekly forecast data. These are accessed/referenced as required by query and corresponding frame.

Implementation

— — —

- Prompt from system for query
- Query accepted by system
- Any keywords present in query are matched to their corresponding frames
- Response is retrieved depending on frame
- Repeat procedure starting from prompt, till the user indicates that no more questions will be asked
- Sign off and end dialogue session

Keywords

Defined and used

- days - indicates which day is the subject of the query, from current ('zero') to the following six; 'seven' indicates the week as a whole
- describe - the defining keywords of the query, which indicates the topic of interest (sun, wind, precipitation, weather, temperature)
- prompts - list of sentences which are provided to user, asking about further queries

Keywords

Defined and used

- negatives - list of accepted negatives in response to prompts; indicates termination of session
- sign offs - list of sentences used to conclude the session, after negative is detected
- invalids - list of responses to queries in unrecognised formats; deals with erroneous questions
- loc codes - dictionary of recognised cities (locations) with their associated location codes as per IMD

Keywords

Defined but not used

- tense - indicates the tense of the question; used to determine where query is about past or future weather, especially in the case of '*kala*' or '*parasom*' which are relative dates
- question - contains the keywords which indicates different types of questions (what, how, amount)

Frames

Detects the day which is subject of the query.

Searched for keyword and then matches it (if present) to dictionary key. Output is provided accordingly. For current day, four formats - features, precipitation, weather and temperature. For forecasts of following days, single format with general response.

If no keywords detected, provides automated response to indicate invalid input format. Prompts for further queries provided after response, negatives recognised as terminals.

Dataset

The dataset for the system was not obtained but generated. It consists of possible scenarios faced by the system. Hence it is a collection of questions that may be posed by the user to the system.

Since the system is rule-based, such questions are somewhat specific to it (thus, not very varied). Also, the list of questions provided in the dataset is not an exhaustive list of what the system can handle - it only provides a few general examples of each type of query.

Challenges

Since the system is rule-based, its capabilities are limited to what it can recognise or extrapolate from the data it possesses. All unrecognised formats are treated as errors.

To provide more accurate and varied responses, a larger list of more complicated rules is required. Also, data would need to be more detailed which uses huge databases, depending on the specificity required.

Limitations

The input and output is provided in WX format, due to difficulty of parsing and matching with Devanagari script.

List of accepted locations has been restricted to the capital cities, based on data provided by IMD. It indicates the procedure followed for a larger-scale model. Also, single location per session.

Names of cities (old/new) accepted in english spelling, not WX. This is due to pronunciation differences all over India.

Limitations

— — —

Due to the limitation of data or formatted information, queries may be of two types. Queries about current day are diverse, but queries about the weekly forecast are restricted to general formats only.

Results

The system performed as expected, given the limitations. It is able to adequately handle a variety of queries within the domain, as well as invalid queries.

Thank You