Technical Report - Target language independent article generation for popular schools and hospitals

Bhavyajeet Singh, Amogh Bhole, Himani Bhardwaj, Aman Kashyap

November 2020

1 Relevant links

- Github Repository
- Collab link for hospital script
- Youtube video
- Powerpoint

2 Introduction

Today Wikipedia is the go-to place for most people on the internet whenever they want to know anything. While Wikipedia is extremely rich in the content it holds, it is not equally accessible to all. More than 35% of all Wikipedia articles are just composed of 3 languages, English, French and German. This divide shows that there is an immediate need to generate articles in indic languages. While the task of enriching Wikipedia for Indian languages is a vast task the current project focuses on a very specific sub task of the same. The project aims at generating Wikipedia articles for hospitals and schools in Indian languages. Along with that the project also intends to have an architecture that is largely language independent and can be extended to other languages easily.

3 Relevant reading and similar works

3.1 Architecture for multilingual wikipedia

This paper provides an architecture for a multilingual wikipedia. It aims at capturing encyclopedic content in a way that abstracts from natural language, enabling people to read and contribute regardless of language. The approach

proposes two major components for the architecture, Abstract wiki and WikiLambda. Abstract wiki is responsible for creating and maintaining content in an abstract notion, independent of language whereas Wikilambda is a wiki of linguistic based functions which allows people with different languages to work on and maintain the same content
Ref: https://arxiv.org/abs/2004.04733

3.2 LSJBot

The purpose of this bot is very similar to the aim that we are trying to achieve. The Lsj bot primarily uses 2 major resources for generating articles about species- Catalogue of Life (Col database) WikiSpecies (wikiMedia). The code works by concatenating the already generated phrases with language independent scientific data to create articles the code contains a lot of hard-coded segments like the list of resources and domain specific functions
Ref : https://en.wikipedia.org/wiki/Lsjbot

3.3 NLG Template authoring Environment

This paper put forwards a simple to use template authoring environment. You need not be an expert in programming and linguistics to create templates for NLG tasks. A simpler alternative is simpleNLG which is a library in Java. It allows users to define templates using programming.
Ref : Novice-Friendly Natural Language Generation Template Authoring Environment

3.4 Data to text generation

This technique is used for generating text from structured data. The data may be in the form of tables, knowledge graphs, JSON. The problem with current state of the art data to text generation systems is that they are not able to generate text spanning 4-5 lines. End to end neural text generation systems are performing poorly as compared to the template based systems. The sentential coherence does not do justice to the input data provided. These systems are not able to produce correct factual information. This paper proposes content selection and planning within a neural data to text generation system. Earlier work in this field mainly consisted of hand crafted rules and learning weights of various grammatical rules to construct text. We will try to incorporate the content selection and planning from this paper.
Ref: https://arxiv.org/pdf/1809.00582.pdf

4 Understanding of the problem statement

The problem statement required us to come up with an implementation which is language independent and can be scaled up to generate articles in multiple

Indian languages from the same source code. Along the course of the project we realised certain other requirements as well that are desired from the project. Since there was no single reliable source of data available, the code must be able to handle sparse data sets which may not always have all the desired values. Along with that since the script is used to generate thousands of articles, the article should not be entirely identical and should have some human touch onto it. While having domain specific codes is important to capture the details of a domain, the code should still be understandable and portable to other domains if the need arises. The generated articles should not just be ready to deploy and publish onto the Wikipedia and should make good use of the wiki specific features like infoBoxes etc. All these requirements were kept in mind while designing and building the project.

5 Implementation details and phases

5.1 Data collection and Data Cleaning

One of the most important aspects that governs the quality of the content and the articles generated is the data that we use to generate the articles.The first step before implementing anything would involve looking for the right data which can be used for the purpose of article generation.

We started generating articles with a basic dataset which had only the names and address of a few hospitals but After the mid evaluations of the project it was observed hat we needed to expand our dataset in order to generate more detailed articles and thus in order to do that we started looking for better sources.

Since it was difficult to find a single source for all the relevant data, we had to accumulate our data from multiple sources.
Among others the following sources were considered and looked upon for possibilities of scraping.

1. data.gov.in
2. pin-code.org
3. medindia.net

Since no single source was able to provide a 'ready-to-use' dataset, we can to build our own knowledgebase by combining data from multiple sources. A script was written to scrape data from some of these sites (like pin-code.org). shows the output of the scraping script for a single hospital.

For the purpose of data cleaning, auxiliary scripts were written to first analyse how sparse the collected data is and how many values are present in each field. After thoroughly observing the collected data it was seen that the data had a lot of garbage and blank values which had to be removed in order to generate meaningful articles. This was done manually with the help of Excel filters or through the help of certain scripts which can be found in the repository.

The final data set used for Hospital articles was still sparse and had the following fields :

- Location Coordinates
- Location
- Hosptial name
- Hospital Category
- Hospital Care Type
- Discipline Systems of Medicine
- State
- District
- Pincode
- Telephone
- Hospital Primary Email Id
- Website
- Specialties
- Facilities

The final data set used for School articles had the following fields :

- village
- city
- name
- gender category
- student count
- teacher count
- school type (pvt, public etc)
- management type
- language medium

5.2 Architecture and specifications

The project primarily depends on a template based approach to generate articles. We have pre defined templates in the required languages and these templates are filled with the relevant data from the database in order to form the articles. The bot would also generate info boxes in the same manner.
The database is language independent and the attribute values are fetched from the database and transliterated (or translated, if required) in the target language before filling the templates.

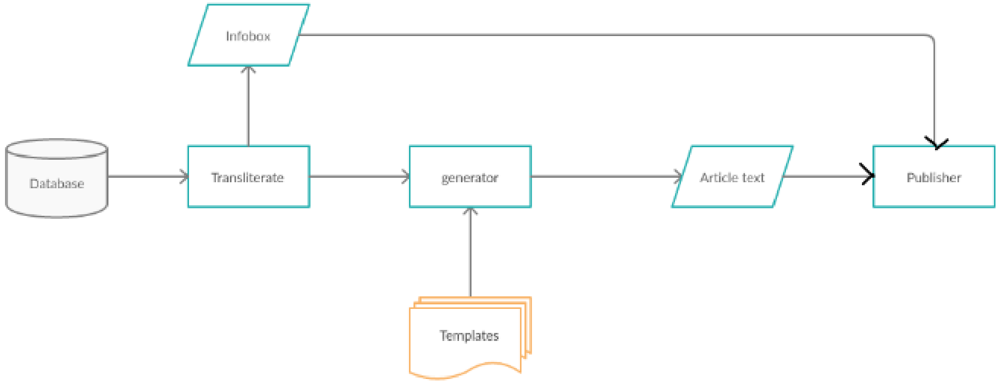


Figure 1: Basic flow

5.2.1 Templates

In order to introduce some variety in the type of articles that are generated, we plan on using modular templates where each attribute of the database has the possibility of being written in multiple ways in the template and any one out of these multiple options is chosen at random while generating the article text.
This approach of storing phrases separately also has an advantage that in case the data is sparse where not all attributes are available for all entries, we can choose to omit the sentences accordingly.
A user-friendly format for the dataset dependent templates was designed which provides easy scalability and modifications .Templates and the language specific part are entirely separated from the code. Multiple options for writing the same sentence can be encoded in the template in order to add variance to the article and no programming knowledge is required to expand templates .

5.3 Generator

The generator scripts read the templates from the file and convert it into a dictionary which is to be used throughout The values from the dataset are translated or transliterated to the target language The script also Generates infoboxes and article content by replacing the available placeholders in the template by the values from the dataset and Prints the article source code in a publishing ready format It Can take the city or IDs as input to selectively generate articles

5.4 Language meta-data

Users can add the language specific meta data in additional files. This contains lists of english words which appear often and are not satisfactorily translated or transliterated by the existing modules. User is free to add multiple or even no words to the meta data file

5.5 Publishing the articles

PyWikiBot is a Python library and collection of scripts that automate work on MediaWiki sites. PyWikibot was set up to publish the generated articles. Though this can only publish on test wiki for now due to licensing and permission complications, but once the necessary permissions are attained, the same pipeline can be used to publish articles on the hindi, marathi or telugu wikipedia as well.

5.6 User Interaction

The user can provide the program the name of the hospital that the user wishes to get an article for, or run the program to generate all possible articles using the database. The user can also provide the name of a district to generate articles for all hospitals belonging to that district.

5.7 Languages

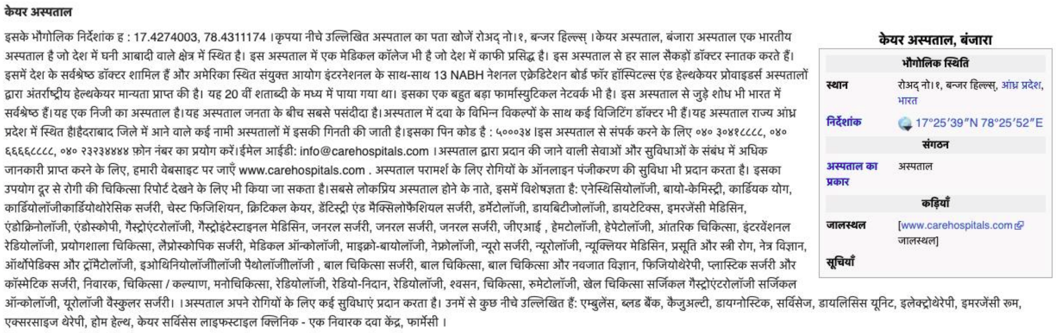
The project intends to have a language independent architecture where any target language will have a template and the code would translate or transliterate the necessary content and publish articles in that language. The current repository includes templates for Hindi, Marathi and Telugu, but this can be very easily extended to any of the other languages

6 Results and Demonstration

In order to generate articles the user needs to follow the following

1. Selection of relevant fields from the available dataset
2. Creating templates which contains sentences for the chosen fields in the desired language
3. Choose which keys should be translated or transliterated for better finetuning (optional)

4. Run the script



केयर अस्पताल

इसके भौगोलिक निर्देशांक ह : 17.4274003, 78.4311174 ।कृपया नीचे उल्लिखित अस्पताल का पता खोजें रोअद् नो।१, बन्जर हिल्ल्स् ।केयर अस्पताल, बंजारा अस्पताल एक भारतीय अस्पताल है जो देश में घनी आबादी वाले क्षेत्र में स्थित है। इस अस्पताल में एक मेडिकल कॉलेज भी है जो देश में काफी प्रसिद्ध है। इस अस्पताल से हर साल सैकड़ों डॉक्टर स्नातक करते हैं। इसमें देश के सर्वश्रेष्ठ डॉक्टर शामिल हैं और अमेरिका स्थित संयुक्त आयोग इंटरनेशनल के साथ-साथ 13 NABH नेशनल एक्रेडिटेशन बोर्ड फॉर हॉस्पिटल्स एंड हेल्थकेयर प्रोवाइडर्स अस्पतालों द्वारा अंतर्राष्ट्रीय हेल्थकेयर मान्यता प्राप्त की है। यह 20 वीं शताब्दी के मध्य में पाया गया था। इसका एक बहुत बड़ा फार्मास्युटिकल नेटवर्क भी है। इस अस्पताल से जुड़े शोध भी भारत में सर्वश्रेष्ठ हैं।यह एक निजी का अस्पताल है।यह अस्पताल जनता के बीच सबसे पसंदीदा है।अस्पताल में दवा के विभिन्न विकल्पों के साथ कई विजिटिंग डॉक्टर भी हैं।यह अस्पताल राज्य आंध्र प्रदेश में स्थित है।हैदराबाद जिले में आने वाले कई नामी अस्पतालों में इसकी गिनती की जाती है।इसका पिन कोड है : ५०००३४ ।इस अस्पताल से संपर्क करने के लिए ०४० ३०४१८८८८, ०४० ६६६६८८८८, ०४० २३२३४४४४ फ़ोन नंबर का प्रयोग करें।ईमेल आईडी: info@carehospitals.com ।अस्पताल द्वारा प्रदान की जाने वाली सेवाओं और सुविधाओं के संबंध में अधिक जानकारी प्राप्त करने के लिए, हमारी वेबसाइट पर जाएँ www.carehospitals.com . अस्पताल परामर्श के लिए रोगियों के ऑनलाइन पंजीकरण की सुविधा भी प्रदान करता है। इसका उपयोग दूर से रोगी की चिकित्सा रिपोर्ट देखने के लिए भी किया जा सकता है।सबसे लोकप्रिय अस्पताल होने के नाते, इसमें विशेषज्ञता है: एनेस्थिसियोलॉजी, बायो-केमिस्ट्री, कार्डियक योग, कार्डियोलॉजीकार्डियोथोरेसिक सर्जरी, चेस्ट फिजिशियन, क्रिटिकल केयर, डेंटिस्ट्री एंड मैक्सिलोफैशियल सर्जरी, डर्मेटोलॉजी, डायबिटीजोलॉजी, डायटेटिक्स, इमरजेंसी मेडिसिन, एंडोक्रिनोलॉजी, एंडोस्कोपी, गैस्ट्रोएंटरोलॉजी, गैस्ट्रोइंटेस्टाइनल मेडिसिन, जनरल सर्जरी, जनरल सर्जरी, जनरल सर्जरी, जीएआई , हेमटोलॉजी, हेपेटोलॉजी, आंतरिक चिकित्सा, इंटरवेंशनल रेडियोलॉजी, प्रयोगशाला चिकित्सा, लैप्रोस्कोपिक सर्जरी, मेडिकल ऑन्कोलॉजी, माइक्रो-बायोलॉजी, नेफ्रोलॉजी, न्यूरो सर्जरी, न्यूरोलॉजी, न्यूक्लियर मेडिसिन, प्रसूति और स्त्री रोग, नेत्र विज्ञान, ऑर्थोपेडिक्स और ट्रॉमैटोलॉजी, इओथिनियोलॉजीलॉजी पैथोलॉजीलॉजी , बाल चिकित्सा सर्जरी, बाल चिकित्सा, बाल चिकित्सा और नवजात विज्ञान, फिजियोथेरेपी, प्लास्टिक सर्जरी और कॉस्मेटिक सर्जरी, निवारक, चिकित्सा / कल्याण, मनोचिकित्सा, रेडियोलॉजी, रेडियो-निदान, रेडियोलॉजी, श्वसन, चिकित्सा, रुमेटोलॉजी, खेल चिकित्सा सर्जिकल गैस्ट्रोएंटरोलॉजी सर्जिकल ऑन्कोलॉजी, यूरोलॉजी वैस्कुलर सर्जरी। ।अस्पताल अपने रोगियों के लिए कई सुविधाएं प्रदान करता है। उनमें से कुछ नीचे उल्लिखित हैं: एम्बुलेंस, ब्लड बैंक, कैजुअल्टी, डायग्नोस्टिक, सर्विसेज, डायलिसिस यूनिट, इलेक्ट्रोथेरेपी, इमरजेंसी रूम, एक्सरसाइज थेरेपी, होम हेल्थ, केयर सर्विसेस लाइफस्टाइल क्लिनिक - एक निवारक दवा केंद्र, फार्मेसी ।

केयर अस्पताल, बंजारा

भौगोलिक स्थिति	
स्थान	रोअद् नो।१, बन्जर हिल्ल्स्, आंध्र प्रदेश, भारत
निर्देशांक	17°25′39″N 78°25′52″E
संगठन	
अस्पताल का प्रकार	अस्पताल
कड़ियाँ	
जालस्थल	[www.carehospitals.com जालस्थल]
सूचियाँ	

Figure 2: Hindi article generated for hospital

7 Future Scope

Future works include Expandint to other similar domains like religious places, colleges etc and Increasing the data available to create more elaborate articles We can also Add templates for other languages to increase multilingual support and explore Natural language generation approaches to reduce the dependency on templates