Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 254–258
Marseille, 11–16 May 2020

A Seed Corpus of Hindu Temples in India

Priya Radhakrishnan*
AI Labs, American Express
Priya.Radhakrishnan@aexp.com

Abstract

Temples are an integral part of culture and heritage of India and are centers of religious practice for practicing Hindus. A scientific study of temples can reveal valuable insights into Indian culture and heritage. However to the best of our knowledge, learning resources that aid such a study are either not publicly available or non-existent. In this endeavour we present our initial efforts to create a corpus of Hindu temples in India. In this paper, we present a simple, re-usable platform that creates temple corpus from web text on temples. Curation is improved using classifiers trained on textual data in Wikipedia articles on Hindu temples. The training data is verified by human volunteers. The temple corpus consists of 4933 high accuracy facts about 573 temples. We make the corpus and the platform freely available. We also test the re-usability of the platform by creating a corpus of museums in India. We believe the temple corpus will aid scientific study of temples and the platform will aid in construction of similar corpora. We believe both these will significantly contribute in promoting research on culture and heritage of a region.

Keywords: Corpus Creation, Information Extraction, Crowd Sourcing

1. Introduction

Motivation : Temples remain an integral part of culture and heritage of India and are centres of religious practice for practicing Hindus (Trouillet, 2017). A scientific study of temples can reveal valuable insights into the Indian culture and heritage. Much of the information about temples is available as text in the open web, which can be utilized to conduct such a study. However this information is not in the form of a learning resource, which can be readily used for such studies.

Two sources of readily available facts on temples is Wikipedia page infoboxes and Google knowledgegraph. But we find these sources to cover only a limited number of popular temples. For instance, as per 2001 census, India has more than 2 million Hindu temples[1] out of which only 518 have wikipedia articles. Wikipedia articles are present only for the popular temples. Even for popular temple like 'Brahmapureeswarar Temple' which has a Wikipedia article, Google knowledge graph does not return any facts.

In this paper we present our initial efforts towards the creation of a corpus of Hindu temples in India, using textual data on temples in the web. Figure 1 depicts temple corpus creation. Consider the temple 'Tirumala Venkateswara Temple' as an example. The input webpage (shown on left side in Figure 1) is textual data on the temple. Our platform converts this into the facts *i.e.*location, deity and mangement, and stores (shown on right side of Figure 1) as temple corpus.

Data Collection : Recent advancements in reading comprehension and question answering literature has produced many systems that perform Natural Language Understanding (NLU). We utilize one such system[2], a Question Answering (QA) system which uses BERT (Devlin et al., 2018) and is among the best performing models on QA and reading comprehension. This system is enhanced to understand the text on temple and generate answers for nine pre-defined questions on a temple. This enhanced system is referred as Temple-QA henceforth. The text on a temple, the nine questions and their answers generated by Temple-QA forms the collected data. This collected data is curated to create the temple corpus.

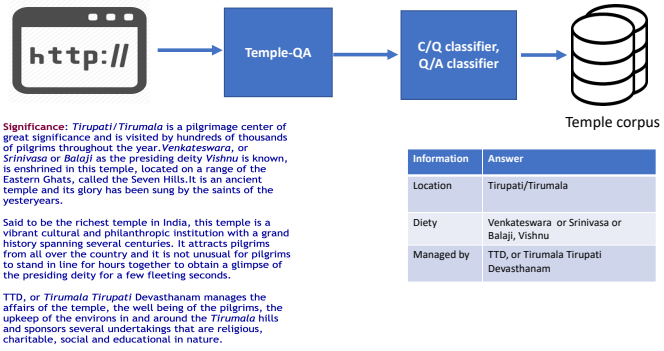

Figure 1: Temple corpus creation process. Temple-QA collects data from web, while C/Q and Q/A classifiers curate the data. Please see section 3. for details.

Curation by Crowd : The collected data is curated using two classifiers. These classifiers are trained using temple domain data verified by volunteers. The training data is created from the text of wikipedia articles on temples. From this text, Temple-QA extracts answers for nine predefined questions about a temple. The text, questions and answers as shown in Figure 2, are presented to human-volunteers. The volunteer are required to review each Question:Answer pair and confirm the correctness of the answers by refering to the text provided. Volunteer's decision is shown in the rightmost column in the figure. As the volunteer is provided with text to refer, he/she is not required to have any prior background knowledge about the temple. By using the volunteer inputs as feedback, the classifiers learn to

*The author was a student at IIIT, Hyderabad. This work was done at IIIT,Hyderabad. The author can also be reached at priyaradhakrishnan0@gmail.com

[1]places of worship by 2001 cencus

[2]Using the BertForQuestionAnswering implementation in https://huggingface.co/transformers/index.html

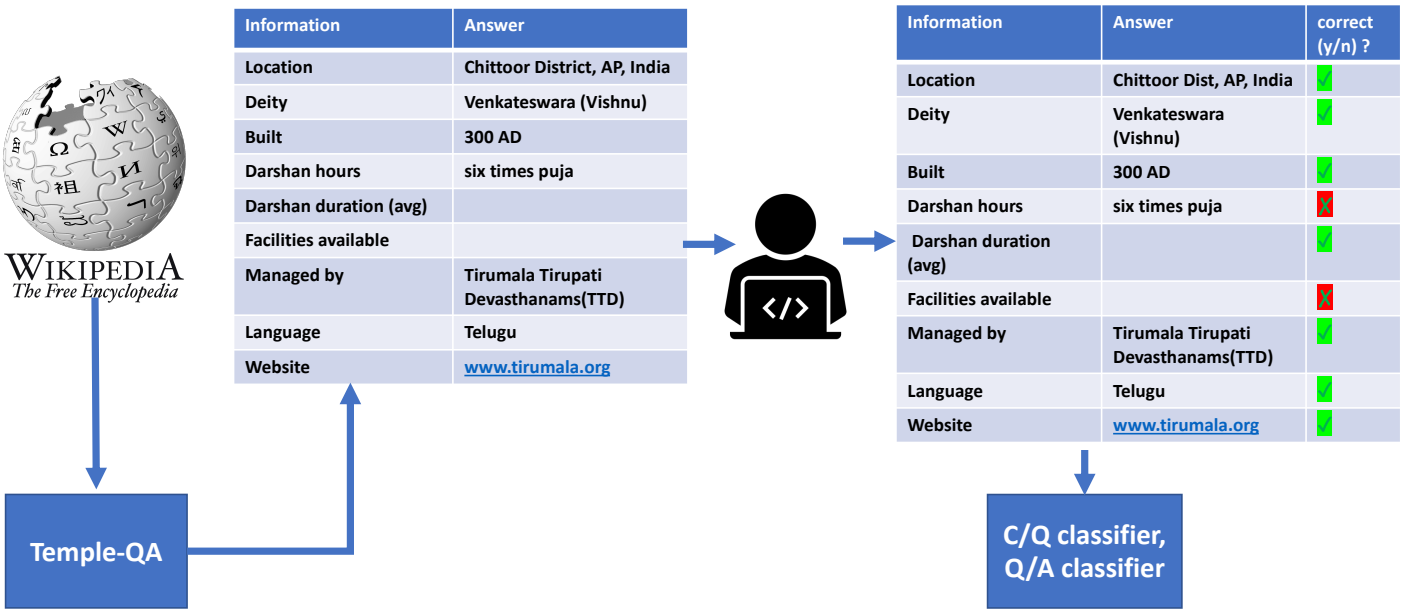

Figure 2: Temple corpus curation platform. From a list of Wikipedia articles on temples, the training data is created by Temple-QA, which is verified by voluteers. Volunteers mark the information as correct or wrong. The human verified training data is used to train the C/Q and Q/A classifiers for better curation. Please see section 4. for details.

better extract answers of these nine questions from temple-related-text.

Curated Corpus : We generated a corpus of 4933 facts on temples from the web text (from websites in lists of Appendix 1). We collected web text of 573 temples, the details of the temples and its webtext source (URL) is provided along with the corpus. We used 518 Wikipedia articles on temples[3] to generate the temple domain training data for the curation classifiers. We present the temple corpus as a 'seed' corpus as it is an initial effort and will be used as seed to our future information extraction efforts from blogs, community forums and social media

Example Application : In order to create a corpus of any domain using our platform, one can start with a list of Wikipedia pages in that domain and our platform can be trained on pre-defined questions for that domain. We tested this by creating a corpus of museums in India starting with a list of Wikipedia pages on indian museums.

Contributions : Our contributions are

1. The first publicly available[4] corpus of temples in India.
2. A simple, re-usable and intelligent platform for collection and curation of information on geo-local entities.

2. Related Work

Textual Corpus of temples in India Past research on textual corpus on places of worship shows text dataset on places of worship across United States[5] and image dataset on temples(Ghorai et al., 2018). Studies have also explored creating domain specific knowledge from Wikipedia (Zhao et al., 2017). However there is a derth of publicly available dataset on temples in India. Creation of dataset of Hindu temples in India is difficult as sources of information are often diverse and scarce and the information is either poorly structured or unstructured (Maheshwari et al., 2018). There are multiple websites which provide lists of temples in India. We compiled a list of these websites and provided it in Appendix 1. These have been created by different people over a period of time for different purposes. The information contained in these websites are disperate in nature.

Human-in-loop AI systems for curation of corpus (Wang et al., 2012) propose a hybrid human-machine approach for curation of corpus. They used machines to do an initial, coarse pass over all the data, and people verify only the most likely matching pairs. This approach is endorsed as a human-only approach was infeasible with increasing data set sizes. For the human evaluator we consider general crowd over domain experts. Korben. et al.(Kobren et al., 2014) find evaluation of domain specific KB facts by crowd workers better than that by experts.

Temple cultures and the Hindu diaspora have been researched from different angles of investigation viz. economic, socio-political, ritual, iconographic and architectural (Larios and Voix, 2018). Our efforts are towards creation of a corpus or learning resource.

3. Data Collection and Challenges

Information on temples is avaiable in open web. Using this resource, we try to create a corpus of indian temples storing the following nine facts about a temple.

1. Where is temple located?
2. The temple is dedicated to which deity?
3. When was the temple built?
4. What are the *darshan*[6] hours?

[3] Wikipedia temple list

[4] https://github.com/priyaradhakrishnan0/templeKB/corpus

[5] https://data.world/awram/us-places-of-worship

[6] Word meaning 'an opportunity to see or an occasion of seeing the image of a deity'. *orig.* Hindi

5. What is the average *darshan* duration?
6. What are the facilities available?
7. Who manages the temple?
8. What is the local language?
9. email / phone / website?

Information extraction from open web is known to be a very challenging task (Sarawagi, 2008). We find that the task is relatively less challenging when the temple has a Wikipedia page. But Wikipedia page is present only for a small fraction of all temples in India. This makes collecting information from open web for temples without a wikipedia page, a very challenging task.

The key challenges in collecting information from a web page are in figuring out *(i)Does the web page contain the answer?* and *(ii)Is the answer correct?* Addressing the first challenge increases recall while addressing the second increases precision of our temple corpus. The first challenge was posed by temples that do not have a wikipedia article, as for the temples with wikipedia article, the wikipedia article invariably contains the answers. The second challenge is due to the low accuracy of extracting answers where the-platform selects a wrong answer for the question.

4. Data Curation Approach

To address the two challlenges listed above, we build two classifiers. The first classifier is a Context-Question (C/Q) classifier which predicts whether the question can be answered from the given context (web text). The second classifier is Question:Answer (Q/A) classifier which predicts if the answer is correct for the given question.

These two classifiers are created by training initially on the SQuAD dataset (Rajpurkar et al., 2018). The SQuAD dataset consists of Contexts (C), Questions (Q) and Answers (A). Contexts are free form text from passages in Wikipedia articles. Questions are posed by crowdworkers on the context. Answer to every question is a segment of text, or span, from the context. Further the question might be unanswerable. The C/Q classifier was built with context-question pairs with answered ones as positive samples and unanswered as negative samples. The Q/A classifier was built with question-answer pairs as positive samples. Negative samples were created by pairing given question with answer of a random question . C/Q and Q/A classifiers are used to curate the Temple-QA output.

We run Temple-QA on web text of the temple, for the nine predefined questions (listed in section 3.) and get the answers. Here context is the web text of the temple, questions are nine predefined questions and answers are Temple-QA generated answers. Questions for which Temple-QA does not generate answer are considered unanswerable. In our running example of Tirumala Venkateswara Temple, questions on location, deity and management are answered while questions on built (period), *darshan*, facilities, language and contact are unanswered. The Question:Answer pairs on curation by C/Q and Q/A classifiers are called 'facts' and stored (in json format) as Temple Corpus.

Classifier	Test Set	Accuracy (SQuAD)	Accuracy ($SQuAD_{++}$)
C/Q	web	0.62	0.72
	wiki	0.43	0.74
	wiki+web	0.60	0.72
Q/A	web	0.58	0.62
	wiki	0.67	0.88
	wiki+web	0.57	0.94

Table 1: Curation accuracy on a held-out test set of 50 temples. Classifiers trained on $SQuAD_{++}$dataset have higher accuracy than classifiers trained on SQuAD dataset.

Fact	Decision	Interpretation	training sample
Question: Answer	correct	correct answer	positive Q/A
Question: Answer	wrong	wrong answer	negative Q/A
Question: No-answer	correct	unanswerable	positive C/Q
Question: No-answer	wrong	answerable	negative C/Q

Table 2: Volunteer decision on collected data. Based on volunteer decision the training samples of C_T,Q_T and A_T are added to SQuAD dataset to create $SQuAD_{++}$ dataset. Please refer section 4. for details.

We assess the accuracy of the temple corpus by evaluating the context-question-answer data using C/Q and Q/A classifier. Specifically we evaluate on a set of 50 temples, randomly selecting 26 temples (264 Question:Answer pairs) with wikipedia article and 24 temples (359 Question:Answer pairs) without wikipedia article. Results are presented in table 1. Temples with and without wikipedia articles are denoted 'wiki' and 'web' respectively, while results on the the entire 50 temples is denoted 'wiki+web'.

To improve the accuracy of curation classifiers, we do the following. We run Temple-QA for temples with wikipedia articles. Here contexts, questions and answers are Wikipedia text of the temple, the nine predefined questions and their Temple-QA generated answers respectively. This is denoted by C_T, Q_T and A_T. This data is curated by humans using Temple corpus curation platform (shown in figure 2). Volunteers annotate the information (mark the Question:Answer and Question:No-answer pairs) as correct or wrong. The interpretation of the four volunteer decision is presented in Table 2.

The corrected samples of C_T, Q_T and A_T is added to the SQuAD dataset. Resulting dataset is denoted $SQuAD_{++}$ henceforth. Following the interpretation (of table 2) the Question:No-answer pairs marked correct and wrong are positive and negative samples respectively for the C/Q classifier. Similarly, the Question:Answer pairs marked correct and wrong are positive and negative samples respectively for the Q/A classifier. The C/Q and Q/A classifiers are

SI	Number of temples	Average number of facts extracted
wiki	518	13.6
web	573	**8.6**

Table 3: Temple corpus has 4933 facts extracted from 573 temples, extracting on an average 8.6 facts for temple. Please see section 5. for details.

trained on $SQuAD_{++}$ dataset and we repeat the evaluation on the same held-out test set of 50 temples as earlier. Results are presented in the fourth column of table 1. Here we see the curation accuracy has improved for classifier trained on $SQuAD_{++}$ dataset. Improvement of 0.10 points on the C/Q classifier is the improvement in recall while improvement of 0.04 in Q/A classifier is the improvement in precision of the platform.By adding the samples of C_T, Q_T and A_T to the SQuAD dataset, we essentially imparted temple domain specific training to the C/Q and Q/A classifiers which gave improved results.

C/Q and Q/A classifiers are implemented using a four layer neural network. The first (input) layer is the embedded layer. Second layer is a bidectional LSTM layer with 100 memory units. Third layer is a dense layer with ReLU activation. Finally, as this is a classifier we use a dense output layer with a single neuron and a sigmoid activation function to make 0 or 1 predictions for the two classes in the problem.

5. Temple Corpus

India has more than 2 million Hindu temples recorded during the 2001 census[7]. Out of these only 518 have wikipedia articles[8]. To collect information on the vast majority of temples which do not have a wikipedia article, one can use the approach proposed here. Our approach uses state-of-the-art NLU system to collect the information and curates it using classifiers. The classifiers are domain adapted using information from wikipedia articles on temples which is further improved by volunteer help. The temple corpus and the collection and curation platform is publicly available.

Appendix 1 provides websites which enlist temples in India. We collect the (web) textual data on these templesand create temple corpus as shown in Figure 1. From temple-related-text on 573 temples we created a corpus 4933 facts. Average number of facts extracted from the web-pages is presented in table 3, where temples with and without wikipedia articles are denoted 'wiki' and 'web' respectively. On an average we extracted 8.6 facts from web pages of temples without wikipedia article. The accuracy of the corpus increased on using C/Q and Q/A classifiers trained on $SQuAD_{++}$ dataset.

6. Example Application - Museum Corpus

In order to create a corpus of any domain using our platform, one can start with list of Wikipedia pages in that domain and our platform can be trained on pre-defined questions for that domain. We demonstrate the same for creating a corpus of museums in India starting with a list of Wikipedia pages on museums in India. We train our platform on these museum specific questions.

Question	Avg. facts	Question	Avg. facts
location	2.5	deity	1.8
built (period)	0.4	*darshan*	1.2
duration	1.0	facilities	0.3
management	0.5	language	0.4
contact	0.5		

Table 4: Average number of questions answered in each of the nine predefined questions in temple corpus. Please refer 7.2. for details.

1. When was the museum established?
2. What are the opening days of the museum?
3. What are the visiting hours?
4. What is the entry fee?
5. What is the average tour duration?
6. What are the facilities available?
7. Who manages the museum?
8. Is docent guide available?
9. What is the language?
10. email / phone / website?

The curated corpus is evaluated with list of museums[9] *i.e.*. We create the corpus by running our platform on web-pages on museums and evaluate the facts across the facts from museums-list. We found the accuracy to be 78%. The accuray was good for questions on docent guide and contact details(email/phone/website), while poor on establishment(date).

7. Discussion

7.1. How does temple corpus fare against search engine retrieved facts for temples ?

One of the easiest and common ways to gather information about a temple is using a search engine. Google, for example, retrieves facts about a temple using Knowledge Graph which is presented as infobox facts. This method works for very popular temples like 'Tirumala Venkateswara Temple'. However for a less popular temple like 'Brahmapureeswarar Temple' which has a Wikipedia article (and hence popular by our measures), Google knowledge graph does not return any infobox facts. We also note that a lot of web text is also available on this temple, which can be used by our platform to create temple corpus of this temple. Thus we see temple corpus is a source of facts on lesser known temples also.

[7] places of worship by 2001 cencus

[8] Wikipedia article is available only for the popular temples.

[9] https://www.sahapedia.org/museums-india

7.2. Which facts are best extracted in temple corpus ?

Table 4 presents the average number of facts derived for each of the nine pre-defined questions for a temple in temple corpus. We find that for questions on location, deity, *darshan* timings and *darshan* duration, the average number of facts derived are above 1.

8. Conclusion and Future work

In this paper we have attempted to create the first publicly available corpus of Hindu temples in India. We describe the temple corpus and the platform for creating and curating the corpus. We explain how we improved the curation performance using training data verified by volunteers. We make the temple corpus and the platform freely available. We believe both the corpus and the platform will aid in construction of similar corpora which contribute significantly in promoting research on culture and heritage of a region. Our study is only an initial effort as it has only considered information in webpages. We plan to enhance it with information in blogs, community forums and social media by using this corpus as seed set. This paper only covers the language resource creation. Enhancements to the language resource and an application utilizing this language resource is our future work.

9. Acknowledgements

We extend our thanks to all the volunteers for their time and effort in reviewing the training data for curation classifiers.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ghorai, M., Santra, S., Samanta, S., Purkait, P., and Chanda, B., (2018). *An Image Dataset of Bishnupur Terracotta Temples for Digital Heritage Research*, pages 269–291. Springer Singapore, Singapore.

Kobren, A., Logan, T., Sampangi, S., and McCallum, A. (2014). Domain specific knowledge base construction via crowdsourcing. December.

Larios, B. and Voix, R. (2018). Introduction. wayside shrines in india: An everyday defiant religiosity.

Maheshwari, A., Kumar, V., Ramakrishnan, G., and Nath, J. S. (2018). Entity resolution and location disambiguation in the ancient Hindu temples domain using web data. In *NAACL-HLT 2018*, June.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Sarawagi, S. (2008). Information extraction. *Found. Trends databases*, 1(3):261–377, March.

Trouillet, P.-Y. (2017). Hindu temples and development of localities in tamil nadu (south india). pages 311–334.

Wang, J., Kraska, T., Franklin, M. J., and Feng, J. (2012). Crowder: Crowdsourcing entity resolution. *Proc. VLDB Endow.*, 5(11):1483–1494, July.

Zhao, X., Xing, Z., Kabir, M. A., Sawada, N., Li, J., and Lin, S.-W. (2017). HDSKG: Harvesting domain specific knowledge graph from content of webpages. In *SANER'17*, pages 56–67. IEEE, February.

Appendix 1: List of web resources for temple data

http://www.vaikhari.org/index.html
http://www.keralatemples.net/index.html
http://www.templenet.com/index.html
https://www.tourmyindia.com/blog/top-30-famous-temples-in-india/
https://www.tourmyindia.com/blog/best-destinations-for-spiritual-tour-south-india/
https://en.wikipedia.org/wiki/List_of_Hindu_temples
http://www.indianmirror.com/temples/temples-by-state.html
http://directory.krishna.com/temples
www.indiaeasytrip.com
http://www.mapsofindia.com/my-india/india/famous-temples-in-delhi
https://www.whatsuplife.in/kolkata/blog/popular-religious-places-famous-temples-in-kolkata/
http://indiatoday.intoday.in/education/story/indian-temples-outside-india/1/459399.html
http://www.tamilselvi.com/List-of-Temples-in-Tamil-Nadu.html
http://www.neatorama.com/2007/09/19/10-most-amazing-temples-in-the-world/
http://www.puneonline.in/city-guide/temples-in-pune
http://singaporehindutemples.com/templelist.html
http://www.explorebihar.in/
http://www.kumbakonam.info/
http://www.tnhrce.org/#
http://eodisha.org/shiva-temples-in-odisha/
https://www.karnataka.com/bangalore/temples/
http://www.asi.nic.in/asi_monu_alphalist.asp
http://www.malabardevaswom.kerala.gov.in/index.php?option=com_cont
http://manatemples.net/Pages/TE_dist_Temples.htm
http://www.inkakinada.com/category/temples
http://www.karnatakavision.com/karnataka-temples.php
https://www.ixigo.com/temples-in-maharashtra-lp-1273349
http://www.apendowments.gov.in/templeListing
http://www.ernakulamonline.in/city-guide/temples-in-ernakulam
http://www.tamilspider.com/resources/3581-List-Temples-Trichy-District.aspx
http://www.bankexamstoday.com/2015/07/famous-temples-india.html
http://travel.vibrant4.com/#
http://www.kamakoti.org/kamakoti/details/branches.html
http://www.godchecker.com/pantheon/indian-mythology.php?list-gods-names
https://shaivam.org/temples-lord-shiva-temples-of-india/temples-of-India

Appendix 2: Sample web resources for Museum data

http://www.arnajharna.org/
http://www.heritagetransportmuseum.org/
http://nrmindia.com/ https://www.irfca.org/articles/patiala-monorail.html