The New Approach for Creating the Knowledge Base Using WikiPedia

Prasad E. Ganesh[1(✉)], H. R. Manjunath[1], V. Deepashree[1], M. G. Kavana[1], and Raviraja[2]

[1] Alva's Institute of Engineering and Technology, Moodbidre, Mijar, Karnataka, India
gprasadl78l43@gmail.com, manjunathdvg@gmail.com
[2] Swiss Re, Bengaluru, India

Abstract. Wikipedia is recognized as one of the largest repositories in the Web. The term knowledge base was in connection with the expert systems as it is the part of Artificial Intelligence. A knowledge base can be created for any entity. The existing system like YAGO, MediaWiki tries to convert Wikipedia into a structured database to provide a vast knowledge base across the domains. It is very difficult to get the information which we want across the domains. So, the solution would be to get a systematic automated approach to build a knowledge base using Wikipedia on entity which we are interested in. The proposed system provides a knowledge base built upon the location as its entity. The system is feeded with seed data, by using these seed data it traverse through the Wikipedia graph and builds knowledge base using similarity measurement between seed data and traversed upcoming pages of wiki graph. Any expert AI systems uses gold standard knowledge base to take any decisions.

Keywords: Natural language processing · Knowledge base · Entity linking

1 Introduction

As digital libraries are regularly increasing in volume, makes it more easy access to the content or information. But it makes it more difficult for a researcher to get a particular information. In that Wikipedia is a large-scale source of network, having all information through the collaboration of contributors. Wikipedia contains information in hierarchical level as articles, link between articles, categories of same kind of articles, multiple language linking etc.

Most records we tend to get from the web in everyday existence is within the style of texts. These texts contain an outsized range of named entities (e.g. person, organization, and place) that are the essential components of texts. However, these entities are extremely ambiguous, thus we want to link them to associate degree existing content in order that folks will apprehend what the entities ask and perceive the texts a lot of properly.

A Knowledge base (KB) could be a special reasonably information for sophisticated structured and unstructured information utilized by automatic processing system. In general, a Knowledge base isn't a static assortment of data, but a dynamic resource.

S. Smys et al. (Eds.): ICCNCT 2019, LNDECT 44, pp. 172–178, 2020.
https://doi.org/10.1007/978-3-030-37051-0_19

The cognitive content construction must extract data like entities and relationships between entities from net texts and add them to the cognitive content. Before filling, the foremost necessary step is to clear up those entities extracted by the system. We have a tendency to decision this method as named entity linking or entity linking. Entity linking could be a task of linking named entities in net texts to their corresponding entities in an exceedingly cognitive content (e.g., DBpedia, and YAGO). The proposed system aims at creating a knowledge base which solves ambiguity of recognizing a place if the place is searched using its alias name and helps in retrieving all possible information about that location.

2 Literature Survey

Beevi and Deivasigamani [1] presented a novel way to deal with the making of information base by extracting knowledge from unstructured web documents. Pre-processing techniques, similarity and redundancy techniques were performed on the extracted documents. The extricated information is sorted out and changed over to XML archives which was then stored in the information base. Though the system is effectively conveyed in genuine condition, refinement is possible in knowledge extraction and representation processes.

Maree et al. [2] Proposed a framework for programmed information base development from heterogeneous information sources including space explicit ontologies, universally useful ontologies, plain messages, picture and video inscriptions which are naturally extricated from site pages. In the proposed framework a few data extraction procedures were integrated to naturally make, enhance and stay up with the latest. Despite the fact that the made information base is utilized to discover arrangements between heterogeneous ontologies in the ecological and agrarian spaces, further expansion al ontologies from online metaphysics vaults on the web can be misused to en-rich and grow the inclusion of the learning base.

Nastase and Strube [3] proposed an approach to derive a large scale and multilingual by misusing a few features of Wikipedia. They described how to expand upon the Wikipedia's current system of classifications and articles to automatically find new relations and their occurrences.

Ma and Zhang [4] proposed the novel approach called economical manifold ranking (EMR) to interchange the normal similarity activity techniques that are unable to replicate the $64000 similarity between completely different modalities of knowledge. Though EMR technique will exploit structure of knowledge to ranking and is additional stable and correct than the normal ways that however as a result of the dimension of the information are going to be lost when the various modalities of knowledge are projected to the common feature area, the performance

Of the rule will be reduced. Thus, the EMR rule is combined with the DNN ways to attain the higher performance for cross-media retrieval.

Saad and Kamarudin [5] concentrated on ascertaining semantic likenesses among sentences and playing out a near examination among recognized comparability estimation systems. They used a large lexical database of English known as WordNet to figure the word-to-word se-mantic likeness. The consequence of the exploration

inferred that the Jaccard and Dice performs better in estimating semantic closeness between sentences.

Trisedya and Inastra [6] explored a few sentence alignment methods which are been used before for another domain and also to check whether the Wikipedia can be used as one of the resources for collecting parallel corpora of some languages. They had used two approaches of sentence alignment by treating Wikipedia as both parallel corpora and comparable corpora which gave positive results. Though the system gave positive results for two particular ethnic languages, it could not give an expected result when tested for other ethnic languages.

Effendi et al. [7] constructed an interpretation corpus dependent on different rudimentary tasks (insertion, deletion, substitution and reordering operations in a publicly supporting stage to produce multi-rework sentences from a source sentence. These rudimentary reword tasks can be used for various applications. Though the system gave good performance for several applications further it can be improved for the utilization by other applications including summarization and machine translation.

Gupta et al. [8] proposed a method for analyzing interlanguage interfaces alongside divert page titles and link text titles and it additionally sifted through off base interpretation applicants utilizing design coordinating. They proposed the utilization of page parameters to give a relativity between the sought string and the interpretation hopefuls. The technique was helpful for particular do-principle explicit terms since exactness and inclusion were superior to anything the bilingual.

Content corpus approach. The framework can be additionally created by including the India dialects which are absent in the online lexicons and by breaking down and discovering some more parameters for better execution.

Whang et al. [9] proposed definite investigation of semantic corpus construction advancements, and proposes another website page deduplication calculation dependent on TF-IDF and word vector separate. The method has established the semantic corpus of common language under cloud administration stage. Though the system proposed an calculation to sift through the non-repetitive corpus put away in HBase and can powerfully store learning in discourse acknowledgment, further it can be improved by optimizing the web joins evacuation calculation and web content duplication expulsion calculation.

Nothman et al. [10] have exhibited a strategy for consequently creating named substance explained message in various dialects from Wikipedia, in light of marking each cordial connection with the element sort of the objective article. The approach was highly effective and efficient for creating NER models in resource-scarce langauges. By using domain-oriented article classification and sentence selection, the method has been utilized for fast development of substance explained corpora specifically spaces however has saddled labeling just the four CoNLL NER types disregards Wikipedia's di-section inclusion of specialized and prevalent areas. Wu et al. [11] introduced the troubles and uses of element linking and fixated on the most approaches to manage those issues. They conjointly listed data bases, datasets and also the analysis criterion and a few challenges on the entity linking. As per their analysis content is employed in substance connecting frameworks are disconnected data or extracted from the net database anyway ailing in programmed update system. To upgrade the exactness of substance connecting, a few frameworks misuse some propelled models to manage this

drawback, as genuine AI models anyway that has higher time quality. In this manner, the framework is frequently improved by executing the approaches to adjust exactness and registering quality.

3 Methodology

The proposed system takes locations name or its alias name as an input and retrieve the Wiki-Pages of the location. Then all the categories under that location is retrieved. Later similarity measurement algorithm is applied to the current page and seed page, if the similarity is greater than 85% then that page is added to the database. Finally the API is created which can be used for the location based applications (Fig 1).

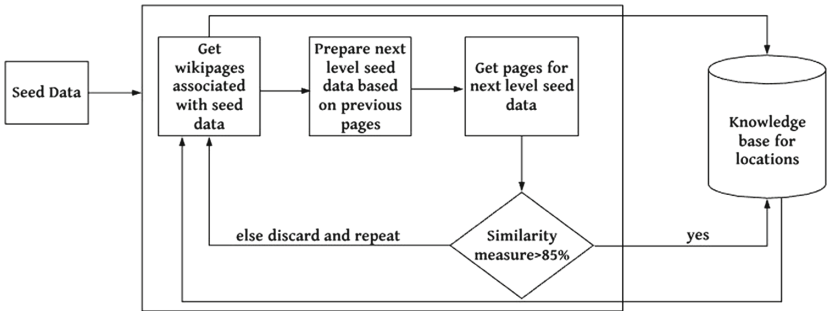

Fig. 1. Block diagram for proposed system

The proposed system process as follows:

- Initially, the seed data is prepared manually by storing the location name, its corresponding Wikipedia URL and seed categories in the database. If the wikipages and categories under the pages are describing about a place or a location, then it is considered as a seed data candidate.
- When user runs the system, creates knowledge base for seed data by fetching the Wiki-page basic details using query API, extract alias names using query API and extract the files and multilingual data using the additional mediawiki queryAPI. The location is covered according to geo coordinates (i.e., latitudes and longitudes).
- Whatever the seed data that have been taken in L1, pages of categories are displayed through API. For L2 data considering sub categories of L1 data comparing with similarity measurements with result more than 85% will be considered as L2 seed data.
- The final product will be the knowledge base for the location.

4 Results and Discussion

4.1 Experimental Setup

- The proposed system uses mongo dB which is an open-source document database and leading NoSQL database. Mongo DB is used to store mainly two collections: Seed categories and Wiki_pages.
- Python 3.7 is the platform to run the proposed system.
- Microsoft Excel Sheets are used to store the seed categories which are collected manually (Fig. 2).

Title	Url	Seed_categories
Karnataka	https://en.wikipedia.org/wiki/Karnataka	South India
Bagalkot	https://en.wikipedia.org/wiki/Bagalkot	Cities and towns in Bagalkot district,Cities in Karnataka
Bengaluru Urban	https://en.wikipedia.org/wiki/Bangalore_Urban_district	Districts of India
Belagavi	https://en.wikipedia.org/wiki/Belgaum	Forts in Karnataka,Tourism in Karnataka,Former capital cities in India,Smart cities in India,Cities and towns in Belgaum district
Bellari	https://en.wikipedia.org/wiki/Bellary	Municipal corporations in Karnataka
Vijayapur	https://en.wikipedia.org/wiki/Bijapur	Tourism in Karnataka
Chamarajanagar	https://en.wikipedia.org/wiki/Chamarajanagar	Cities and towns in Chamarajanagar district
Chikballapur	https://en.wikipedia.org/wiki/Chikkaballapur	Cities and towns in Chikkaballapur district
Chitradurga	https://en.wikipedia.org/wiki/Chitradurga	Cities and towns in Chitradurga district
Dakshina Kannada	https://en.wikipedia.org/wiki/Dakshina_Kannada	Tulu Nadu
Davanagere	https://en.wikipedia.org/wiki/Davanagere	Cities and towns in Davanagere district
Dharwad	https://en.wikipedia.org/wiki/Dharwad	Cities and towns in Dharwad district
Haveri	https://en.wikipedia.org/wiki/Haveri	Western Chalukya Empire,Cities and towns in Haveri district
Kodagu	https://en.wikipedia.org/wiki/Kodagu_district	Proposed states and territories of India
Kolar	https://en.wikipedia.org/wiki/Kolar	Cities and towns in Kolar district
Koppal	https://en.wikipedia.org/wiki/Koppal	Cities and towns in Koppal district
Mandya	https://en.wikipedia.org/wiki/Mandya	Cities and towns in Mandya district
Raichur	https://en.wikipedia.org/wiki/Raichur	Cities and towns in Raichur district
Ramanagara	https://en.wikipedia.org/wiki/Ramanagara	Climbing areas of India
Shivamogga	https://en.wikipedia.org/wiki/Shimoga	Cities and towns in Shimoga district
Tumakuru	https://en.wikipedia.org/wiki/Tumkur	Cities and towns in Tumkur district
Udupi	https://en.wikipedia.org/wiki/Udupi	Cities and towns in Udupi district,Populated coastal places in India,Port cities in India
Yadgiri	https://en.wikipedia.org/wiki/Yadagiri	Cities and towns in Yadgir district
Guntur	https://en.wikipedia.org/wiki/Guntur	Mandal headquarters in Guntur district
Kadapa	https://en.wikipedia.org/wiki/Kadapa	District headquarters of Andhra Pradesh
Krishna	https://en.wikipedia.org/wiki/Krishna_district	Districts of Andhra Pradesh
Kurnool	https://en.wikipedia.org/wiki/Kurnool	Cities in Andhra Pradesh

Fig. 2. Excel sheet for seed data

4.2 Results

The proposed system is efficient than the existing system which fails to identify whether the searched entity is place or a person. The proposed system redirects into the correct page when the place is searched by its alias names. This system has a common API which replaces the different APIs needed for different format of data (text, image, video and so on) and it resolves the language problem by supporting multi languages. Figure 3 represents the incremental knowledge base constructed by the system showing all the seed categories of various locations.

Figure 4 represents database stored in mongo dB which is extracted from the seed categories and provides the information of the location to the user when they search for the location by its alias name.

```
Saved seed_categorie :Forts in Karnataka pagetitle :  Mirjan Fort
Saved seed_categorie :Forts in Karnataka pagetitle :  Mudgal
Saved seed_categorie :Forts in Karnataka pagetitle :  Mundargi
Saved seed_categorie :Forts in Karnataka pagetitle :  Nadivi
Saved seed_categorie :Forts in Karnataka pagetitle :  Nandi Hills, India
Saved seed_categorie :Forts in Karnataka pagetitle :  Pavagada
Saved seed_categorie :Forts in Karnataka pagetitle :  Raichur fort
Saved seed_categorie :Forts in Karnataka pagetitle :  Sadashivgad
Saved seed_categorie :Forts in Karnataka pagetitle :  Savandurga
Saved seed_categorie :Forts in Karnataka pagetitle :  Sira, Karnataka
Saved seed_categorie :Forts in Karnataka pagetitle :  Srirangapatna fort
Saved seed_categorie :Forts in Karnataka pagetitle :  Tekkalakote
Saved seed_categorie :Forts in Karnataka pagetitle :  Uchangidurga
Saved seed_categorie :Forts in Karnataka pagetitle :  Vanadurga Fort
Saved seed_categorie :Forts in Karnataka pagetitle :  Yadagiri
Saved seed_categorie :Tourism in Karnataka pagetitle :  Agumbe
Saved seed_categorie :Tourism in Karnataka pagetitle :  Aihole
Saved seed_categorie :Tourism in Karnataka pagetitle :  Badami
Saved seed_categorie :Tourism in Karnataka pagetitle :  Bagalkot district
Saved seed_categorie :Tourism in Karnataka pagetitle :  Belgaum
Saved seed_categorie :Tourism in Karnataka pagetitle :  Belur, Karnataka
Saved seed_categorie :Tourism in Karnataka pagetitle :  Bijapur
Saved seed_categorie :Tourism in Karnataka pagetitle :  Bidar
Saved seed_categorie :Tourism in Karnataka pagetitle :  Dandeli
Saved seed_categorie :Tourism in Karnataka pagetitle :  The Emerald Route
Saved seed_categorie :Tourism in Karnataka pagetitle :  Golden Chariot
Saved seed_categorie :Tourism in Karnataka pagetitle :  Gommatagiri
Saved seed_categorie :Tourism in Karnataka pagetitle :  Gopinatham
Saved seed_categorie :Tourism in Karnataka pagetitle :  Jamboti
Saved seed_categorie :Tourism in Karnataka pagetitle :  Kanakachalapathi Temple, Kanakagiri
Saved seed_categorie :Tourism in Karnataka pagetitle :  Kanakagiri
Saved seed_categorie :Tourism in Karnataka pagetitle :  Kanakagiri Jain Shri kshetra
Saved seed_categorie :Tourism in Karnataka pagetitle :  Karnataka State Tourism Development Corporation
Saved seed_categorie :Tourism in Karnataka pagetitle :  Kodagu district
Saved seed_categorie :Tourism in Karnataka pagetitle :  Kokrebellur
Saved seed_categorie :Tourism in Karnataka pagetitle :  Lakkundi
Saved seed_categorie :Tourism in Karnataka pagetitle :  Lal Bagh
Saved seed_categorie :Tourism in Karnataka pagetitle :  Madikeri
Saved seed_categorie :Tourism in Karnataka pagetitle :  Mandagadde Bird Sanctuary
Saved seed_categorie :Tourism in Karnataka pagetitle :  Pattadakal (town)
Saved seed_categorie :Tourism in Karnataka pagetitle :  Sakleshpur
Saved seed_categorie :Tourism in Karnataka pagetitle :  Saundatti
Saved seed_categorie :Tourism in Karnataka pagetitle :  Shivaganga
Saved seed_categorie :Tourism in Karnataka pagetitle :  Shravanabelagola
Saved seed_categorie :Tourism in Karnataka pagetitle :  Sirsi, Karnataka
Saved seed_categorie :Tourism in Karnataka pagetitle :  Somanathapura (town)
Saved seed_categorie :Tourism in Karnataka pagetitle :  Somwarpet
```

Fig. 3. Incremental knowledge base

	_id ObjectId	title String	url String	summary String	content String	categories Array
1	5c9a4d685f38a62f84884077	"South India"	"https://en.wikipedia.org/wiki/	"South India is the area includ	"South India is the area includ	[] 25 elements
2	5c9a4d6c5f38a62f84884078	"Andhra Pradesh"	"https://en.wikipedia.org/wiki/	"Andhra Pradesh () (pronunciati	"Andhra Pradesh () (pronunciati	[] 27 elements
3	5c9a4d6f5f38a62f84884079	"Karnataka"	"https://en.wikipedia.org/wiki/	"Karnataka (Karnāṭaka) is a sta	"Karnataka (Karnāṭaka) is a sta	[] 23 elements
4	5c9a4d715f38a62f8488407a	"Kerala"	"https://en.wikipedia.org/wiki/	"Kerala () is a state on the so	"Kerala () is a state on the so	[] 28 elements
5	5c9a4d745f38a62f8488407b	"Lakshadweep"	"https://en.wikipedia.org/wiki/	"Lakshadweep (; ISO: Lakṣadvīp	"Lakshadweep (; ISO: Lakṣadvīp	[] 23 elements
6	5c9a4d775f38a62f8488407c	"Puducherry"	"https://en.wikipedia.org/wiki/	"Puducherry (, literally New To	"Puducherry (, literally New To	[] 21 elements
7	5c9a4d7a5f38a62f8488407d	"Bharatanatyam"	"https://en.wikipedia.org/wiki/	"Bharatanatyam (Tamil: பரதநாட்ய	"Bharatanatyam (Tamil: பரதநாட்ய	[] 14 elements
8	5c9a4d7c5f38a62f8488407e	"Carnatic music"	"https://en.wikipedia.org/wiki/	"Carnatic music, Karnāṭaka saṃg	"Carnatic music, Karnāṭaka saṃg	[] 14 elements
9	5c9a4d7f5f38a62f8488407f	"Cinema of South India"	"https://en.wikipedia.org/wiki/	"The Cinema of South India is u	"The Cinema of South India is u	[] 6 elements
10	5c9a4d825f38a62f84884080	"Coastline of Andhra Pradesh"	"https://en.wikipedia.org/wiki/	"The Coastline of Andhra Prades	"The Coastline of Andhra Prades	[] 6 elements
11	5c9a4d845f38a62f84884081	"Coastline of Tamil Nadu"	"https://en.wikipedia.org/wiki/	"The Coastline of Tamil Nadu is	"The Coastline of Tamil Nadu is	[] 7 elements
12	5c9a4d875f38a62f84884082	"Coromandel Coast"	"https://en.wikipedia.org/wiki/	"The Coromandel Coast is the so	"The Coromandel Coast is the so	[] 28 elements
13	5c9a4d8a5f38a62f84884083	"South Indian cuisine"	"https://en.wikipedia.org/wiki/	"South Indian cuisine includes	"South Indian cuisine includes	[] 11 elements
14	5c9a4d8c5f38a62f84884084	"Dravidian architecture"	"https://en.wikipedia.org/wiki/	"Dravidian architecture is an a	"Dravidian architecture is an a	[] 10 elements
15	5c9a4d8f5f38a62f84884085	"Earliest color films in South	"https://en.wikipedia.org/wiki/	"The South Indian film industry	"The South Indian film industry	[] 6 elements
16	5c9a4d925f38a62f84884086	"Economy of South India"	"https://en.wikipedia.org/wiki/	"Economy of South India after i	"Economy of South India after i	[] 5 elements
17	5c9a4d955f38a62f84884087	"Geography of South India"	"https://en.wikipedia.org/wiki/	"The Geography of South India c	"The Geography of South India c	[] 5 elements

Fig. 4. Mongo dB database

5 Conclusion and Future Work

Any AI system which is location-based application can use the proposed system's API to access the information of that location along with its categories. The proposed system works only for the Indian places. The ambiguity of redirecting to the exact page is resolved when the location is searched by its alias names.

In future, the system can be developed for locations all over the world and Knowledge graph can be constructed by using this knowledge base. Currently the system is using similarity measurement algorithm which can be replaced with other

Machine Learning process to get more accuracy. Further the knowledge base can be created for multi entity which closely related or for multilingual so that this knowledge base can be used for translation also.

References

1. Beevi, J. H., Deivasigamani, N.: A new approach to the design of knowledge base using XCLS clustering. In: Proceedings of IEEE International Conference on Pattern Recognition, Informatics and Medical Engineering, pp. 14–19 (2012)
2. Maree, M., Alhashmi, S.M., Belkhatir, M., Hawit, A.: Automatic construction of a domain-independent knowledge base from heterogeneous data sources. In: Proceedings of IEEE International Conference on Fuzzy Systems and Knowledge Discovery, pp. 1483–1488 (2012)
3. Nastase, V., Strube, M.: Transforming wikipedia into a large scale multilingual concept network, pp. 62–85. Elsevier (2012)
4. Ma, S.Q., Zhang, H.: Efficient manifold ranking for cross-media retrieval. In: 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 335340 (2018)
5. Saad, S.M., Kamarudin, S.S.: Comparative analysis of similarity measures for sentence level semantic measurements of text. In: Proceedings of IEEE International Conference on Control System, Computing and Engineering, pp. 90–94 (2013)
6. Trisedya, B.D., Inastra, D.: Creating Indonesian-javanese parallel corpora using wikipedia articles. In: Proceedings of IEEE, pp. 239–245 (2014)
7. Effendi, J., Sakti, S., Nakamura, S.: Creation of a multi paraphrase corpus based on various elementary operations. In Proceedings of IEEE Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA), pp. 177–182 (2017)
8. Gupta, A., Goyal, A., Bindal, A., Gupta, A.: Meliorated approach for extracting bilingual terminology from wikipedia, world academy of science, engineering and technology. In: Proceedings of 11th IEEE International Conference on Computer and Information Technology, pp. 78–85 (2008)
9. Wang, S.Z., Zhang, Q.C., Zhang, L.: Natural language semantic corpus construction based on cloud service platform. In: Proceedings of IEEE International Conference on Machine Learning and Cybernetics, pp. 670–674 (2017)
10. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from Wikipedia, Elsevier, pp. 151175 (2012)
11. Wu, G., He, Y., Hu, X.: Entity linking: an issue to extract corresponding entity with knowledge base. In: Proceedings of IEEE, vol. 6, pp. 6220–6231 (2018)
12. Tehseen, M., Javed, H., Shah, I.H., Ahmed, S.: A light- weight key negotiation and authentication scheme for large scale WSNs. In: Recent Trends and Advances in Wireless and IoT-enabled Networks, pp. 225–235. Springer (2019)
13. Sun, S., Au, K.S., Li, Y., Barber, P.: Systems and Methods for Authentication. US Patent Application 16/159,235, filed 14 Feb (2019)