ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 5, Issue 4 (Oct. - Dec. 2017)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2017)

PLIS: Proposed Language Independent Stemmer Performance Evaluation

[I]Dr. M. Kasthuri, [II]Dr. S. Britto Ramesh Kumar
[I]Bishop Heber College, Tiruchirappalli, Tamil Nadu, India
[II]St. Joseph's College, Tiruchirappalli, Tamil Nadu, India

Abstract
Information Retrieval (IR) is an emerging discipline that involves methods, models and patterns to find the documents of an unstructured nature in dynamic environment. Search Engines are playing a major role in Information Retrieval Systems (IRS) to identify the morphological variants of the language using Stemming. Stemming is an important pre-processing step in query-based systems such as IRS, Web Search Engine, Natural Language Processing (NLP), Big Data Analysis, etc. The purpose of stemming is to diminish different grammatical or word forms to a common base form. In this digital era, most of the web pages are designed using English and European languages. Similarly, the web pages designed with Indian and other Asian languages are also increasing. The study reveals that the approaches for developing the stemmer involve rule-based, machine learning and hybrid approach. However, each one of them has its own limitations. Additionally, the existing stemmers are severely affected by various problems like over-stemming, under-stemming, etc. Furthermore, there is no effective approach for developing Language Independent Stemmer (LIS) for IR Systems with greater accuracy. Therefore, in this research work, it has been proposed to design the model for Language Independent Stemmer using Dynamic Programming (DP) to retrieve the multi-linguistic web documents with the greater speed and accuracy. However, this research paper presents performance evaluation result.

Keywords
Stemming, Information Retrieval, Dynamic Programming, Language Independent Stemmer, Multi-linguistic

I. Introduction

Information Retrieval (IR) is fast becoming the dominant form of information access. The acquisition of the relevant document is made using query mechanism or by browsing or searching in an information space. The problem of information storage and retrieval has been receiving more attention in the recent years. Information explosion and the need to find relevant information from a huge collection are driving the improvements in Information Retrieval field. In recent years, IR has been established as one of the research disciplines in Computer Science with growing industrial impact. With the growth of the World Wide Web, Information and Communication Technologies, and high-speed Internet connections, the generation and transmission of large volume of data across the world have increased over the last decade. The networks, technologies and information which are being generated require faster and better Information Retrieval Systems. Due to the increase of information day by day the search engines require more efficient techniques for retrieving the data faster and with great accuracy. Especially, there has been a huge increase in the amount of web information available in Indian and other Asian languages **[1]**. Web document in a large number of Indian languages like Hindi, Urdu, Bengali, Oriya, Tamil, Telugu and Marathi is now available in the electronic form **[2]**. Information Retrieval Systems (IRS) play a vital role in providing access to this information **[3]**. In order to achieve this, the Information Retrieval Systems use Search Engine for retrieving the relevant and useful documents with respect to the user input queries.

II. Review of Literature

Stemming is to decrease the inflectional forms and derivationally related forms of a word to a common base form **[4]**. Stemming is used to improve the retrieval effectiveness and to reduce the size of indexing files. Current research on stemming mainly focuses on removing suffixes, prefixes and affixes. In recent times, the focus of the research has shifted to developing language independent stemmer encompassing various dimensions namely performance, strength, accuracy and efficiency. Finding the accurate stemmer for multilingual document is an important requirement in the field of information retrieval. However, it is difficult to find the structure of the words for the specific languages, consider the writing morphology of the language and handle the morphological variants of the language. Hence, in order to develop the language independent stemmer for various languages, an exhaustive study was made on existing stemming algorithms and its approaches.

Adege et al. (2017) have generated a stemmer of Ge'ez language using rule-based approaches **[5]**. There are two techniques were followed such as affix removal and morphological analysis techniques. The experimental result shows that, this research work performed with an accuracy of 82.42%. However, limited rules sets were created, which affects the accuracy of the proposed stemmer. It requires linguistic knowledge to generate rule set. Further, stemming errors like over-stemming and under-stemming problems were observed from affix removal technique.

Zadeh et al. (2017) have proposed a new hybrid stemming method based on a combination of affix stripping and statistical techniques for Persian language **[6]**. The authors conducted a performance test on the proposed stemmer using two different data sets. The experimental result shows that encouraging results were obtained. However, rule-based affix stripping approach applied in this research work requires linguistic inspection and it is a time consuming. Additionally, if small snippets of documents are involved, then the approach will not provide effective results.

Patel et al. (2016) have proposed a lightweight stemmer for Gujarati language using supervised learning **[7]**. Initially the authors to make handcrafted rules for prefix and suffixes. Then these rules were checked using linguistic expert for the Gujarati Morphology. They evaluated the proposed algorithm with IRS and improved results were obtained. However, linguistic inspection needs lots of time. Additionally, the primitive knowledge of the language is more important, which not suitable for agglutinative as well as morphologically rich languages.

Ali et al. (2016) have proposed a rule-based stemming method

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

for Urdu Text **[8]**. This approach has evaluated on Urdu headline news datasets. The proposed method provides 90% to 95 % accuracy. However, in order to develop this Urdu stemmer, generic stemming rules and stemming lists have been created in advance. Additionally, various grammar books and Urdu literature were used to generate a list of 60 prefix rules. It is a language dependent stemmer and linguistic inspection should be needed.

Nehar et al. (2015) have proposed a stemmer for Arabic language using Trigram, Transducers and Rational Kernels approach **[9]**. The experimental result shows that stemming improves the quality of classifiers in terms of accuracy. However, the main limitation of this stemmer is that trigram approach requires large storage space and it is not a practical approach. The rational kernels classification is computationally expensive. Additionally, complexity of this stemmer is heavy and it requires long computational time.

Deepamala et al. (2015) have proposed a stemmer for Kannada language using table lookup approach **[10]**. To improve the stemming performance, a stem list is created manually and inserted into the table. However, the major problem in this stemmer is high complexity due to the adoption of various algorithms such as naïve bayes, lookup table approach and maximum entropy for classification and stemming. Large storage space is required to insert each word and its morphological variants for the specific language. Finally, manual inspection is needed to verify valid stem in the table.

Karanikolas, (2015) has proposed a stemmer for English using supervised learning. This approach required two resources such as a list of available suffixes and a set of words in the target language with their translations. The word and its morphological variants have been prepared using the native language experts **[11]**. However, the major limitation of this stemmer is the dependence on the experts' argument which involves time consuming. Additionally, the primitive knowledge of the language is an essential for the IR experts. Finally, this stemmer is not suitable for morphologically rich languages.

Brychcín et al. (2015) have proposed high precision stemmer (HPS) based on unsupervised approach. The primary objective of this stemmer is to find the stem word for multilingual documents using unsupervised **[12]**. High precision stemmer is tested on corpora to find the stem words for Czech, Polish, Slovak, Hungarian, English and Spanish languages. However, the limitation of this stemmer is that the complexity of the algorithm is heavy since this approach uses various algorithms such as clustering, suffix stripping, n-gram probability, maximum entropy classifier and it is not user friendly. Finally, it requires large size of memory.

Gupta, (2014b) has proposed stemmer for Hindi nouns using rule-based approach **[13]**. This stemmer applies suffix stripping approach for performing stemming for nouns. By analyzing the articles from Hindi newspapers, the author has generated 16 noun suffixes and their corresponding stemming rules. It is a light stemmer. However, the major drawback of this stemmer is that it uses only 16 Hindi noun suffixes that lead to over-stemming and under-stemming problems. It requires linguistic knowledge to generate rule set. Additionally, there are not many linguistic resources available for Hindi and it is at infancy stage.

Joshi et al. (2014) have proposed a stemmer for Punjabi language using hybrid approach. This stemmer uses table lookup and rule-based approach **[14]**. This stemmer attempts to provide better results using different algorithms in a hybrid way. However, the main issue identified in this stemmer is limited Punjabi words only are stored in the table, which makes to produce incorrect stem word. Another issue identified in this stemmer is that the table lookup approach needs extensive manual work and requires large storage space. Suffix stripping approach may invoke over-stemming and under-stemming errors. It requires strong linguistic knowledge to create rule sets.

Karanikolas, (2014) has proposed a methodology for building simple but robust language independent stemmer for Serbian, Albanian and Greek languages. The purpose of this stemmer is to identify the stem word without using the knowledge of the target language **[15]**. The advantage of this stemmer is it is a language independent stemmer. However, list of suffixes in the target language is created manually in advance and it requires lots of storage space. This stemmer requires two iterations and it adopts supervised learning which increases the processing time.

Having reviewed the literature, it is observed that various approaches are available for stemming in Information Retrieval Systems but each one have several limitations. The existing stemming algorithms used rule-based, statistical, successor variety, table lookup, machine learning and hybrid approach. Rule-based approach needs additional supervision of linguistic consultants to frame the rule set. Lookup approach decreases the stemming errors however it needs lots of manual work **[16]**. Large size of memory is required in lookup approach for storing the stem words and its inflected terms. The limitation with the successor variety stemming is that if very small document snippets are involved, then it will not provide correct result. Supervised learning also requires a very good linguistic knowledge to segment words and get the root or inflectional words. Manual inspection is required to verify the stemmed word every time and also increases the cost. Unsupervised learning is little complex and over-stems the words occasionally. It is motivated to develop Language Independent Stemmer using Dynamic Programming (DP) to retrieve the multi-linguistic web documents with the greater speed and accuracy.

III. Proposed Algorithm for Language Independent Stemmer

The framework of the Proposed Language Independent Stemmer (PLIS) for Information Retrieval Systems (IRS) is useful to analyze the given input query and undergoes various processing steps in the PLIS, which are also explained in the previous research paper **[17]**. The proposed algorithm has tested on various languages like English, French, Tamil and Hindi. Each language has its own morphological structure. It is difficult to manage with the single algorithm for all the language. But the proposed algorithm supports Indian and Non-Indian languages. The PLIS algorithm considers diacritics and accented characters of the language and generates efficient result **[18]**. An innovative attempt is being made to develop a stemming algorithm for a novel conflation method that exploits the quality of words and uses some standard Natural Language Processing tools like Levenshtein Distance and Longest Common Subsequence for Stemming process **[19]**. The proposed algorithm can be used in multi-lingual information retrieval systems. The Proposed Language Independent Stemmer generates accurate result even with agglutinative language like Tamil. Any Information Retrieval Systems can adopt the PLIS as a pre-processing approach and find the stem word for any morphologically rich languages. In this paper we discussed about experimental study and discuss about the result obtained from our research work.

IV. Experimental Study

The experiment study has been conducted to test the performance, Strength and accuracy of the Proposed Language Independent Stemmer (PLIS) using test bed. The test bed has implemented based on PLIS framework, where the functional components are implemented using PHP 5.4, Html5, JavaScript and CSS3.

A) Data Sample Construction

The data sample is constructed from four repositories namely English, French, Tamil and Hindi, arranged into conflation groups. The test data set-I has 1,858 English words, out of which 287 incorrect words are available. The test data set-II contains 1,858 French words, which contains 359 incorrect words. The test data set-III consists of 1,858 Tamil words, there are 356 incorrect words are identified. The test data set-IV contains 1,858 Hindi words, where 347 incorrect words are available. Similarly, ten different data samples are constructed with four data sets. Each data set has its own distinct words with various numbers of words collected from four different repositories. The maximum number of 2,70,674 distinct words are used in ten data samples **[22]**. Table I shows that data sample1 with four data sets, chosen for the test.

Table I: Data Sample 1 with Four Data Sets

Data Set	Total Number of Words	Total Number of Unique Words
Test Data set-I	1858	287
Test Data set-II	1858	359
Test Data set-III	1858	356
Test Data set-IV	1858	347

B) Summerized Performance Analysis

To assess the performance of the stemmer, apply these algorithms to the sample vocabulary downloaded from the repository servers. The proposed stemmer and some of the existing stemmers **[21], [23], [25]** are evaluated with more than 2,70,674 words for languages such as English, French, Tamil and Hindi, which are arranged into conflation groups. Some of them are incorrect words. Table II shows the summarized overall performance analysis of PLIS compared with the existing stemmers.

Table II: Summarized Performance Analysis

Analysis of stemmers	Lovins	Iterated Lovins	Porter1	Porter2	Paice/ Husk	Fair-wheather	Porter/ PECL French	Damodharan Tamil	Gupta Hindi	PLIS
Total No.of Words (TW)	1858	1858	1858	1858	1858	1858	1858	1858	1858	1858
Number of Distinct words before stemming (N)	1571	1571	1571	1571	1571	1571	1499	1502	1511	1520
Number of Distinct words after stemming (S)	683	745	756	727	678	677	766	764	758	645
Mean Number of Words (MWC) = N/S	**2.30**	**2.11**	**2.08**	**2.16**	**2.32**	**2.32**	**1.96**	**1.97**	**1.99**	**2.36**
Index Compression Factor (ICF) = ((N-S)/N)*100	**56.52**	**52.58**	**51.88**	**53.72**	**56.84**	**56.91**	**48.90**	**49.13**	**49.83**	**57.57**
Number of words Stemmed (WS)	1363	1248	1259	1237	1319	1309	1267	1324	1161	1366
Words Stemmed Factor (WSF) = (WS/TW)*100	**73.36**	**67.17**	**67.76**	**66.58**	**70.99**	**70.45**	**68.19**	**71.26**	**62.49**	**73.52**
Correctly Stemmed words (CSW)	1210	1112	998	976	1150	1100	1199	1221	1101	1344
Incorrectly stemmed words (ISW)	153	136	261	261	169	209	68	103	60	22
Correctly Stemmed Words Factor (CSWF) =(CSW/WS)*100	**88.77**	**89.10**	**79.27**	**78.90**	**87.19**	**84.03**	**94.63**	**92.22**	**94.83**	**98.39**
Correct Words not stemmed (CW)= (N-WS)	208	323	312	334	252	262	232	178	350	154
No. of Distinct words after conflation (NWC) = (S-CW)	475	422	444	393	426	415	534	586	408	491
Average Words Conflation Factor (AWCF) = ((CSW-NWC)/CSW)	**60.7**	**62.1**	**55.5**	**59.7**	**63.0**	**62.3**	**55.5**	**52.0**	**62.9**	**63.5**
Mean Removal Rate	3.11	3.08	3.01	3.12	3.09	3.13	2.94	2.4	2.23	3.16
Query Latency (ms)	954.40	967.42	886.18	897.18	798.37	999.09	501.79	634.58	988.3	479.4
Query Throughput (NQ)	1.05	1.03	1.13	1.11	1.25	1.00	1.99	1.58	1.01	2.09
Mean Modified Hamming Distance	8.68	9.92	9.94	6.32	7.81	8.03	5.19	4.55	1.55	4.77

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

C) Result and Discussion

Mean number of Words per Conflation class and Index Compression Factor

From the table, Mean number of Words per Conflation class (MWC) obtained by all stemmer algorithms is above 2 characters. Index Compression Factor (ICF) obtained by Porter1 English stemmer is 51.88%, Gupta Hindi Stemmer is 49.83%, Damodharan Tamil Stemmer is 49.13% and Porter French Stemmer is 48.90%, which are comparatively less than Lovins stemmer i.e., 56.52%, Porter2 English stemmer i.e., 53.72% and Iterated Lovins Stemmer i.e., 52.58%. Further, Paice/Husk Stemmer is 56.84%, Fairwheather Language Independent Stemmer is 56.91% and Proposed Language Independent Stemmer (PLIS) is 57.57%, which provides higher Index Compression Factor value than the other stemmers. Index compression reduces the storage overhead of repeated values, so PLIS reduces the storage space.

Word Stemmed Factor

Further the Word Stemmed Factor (WSF) obtained by all the algorithms is greater than 62%, which is above the threshold value i.e., 50 %. This result illustrates that the strength of all the stemmers is strong and all are aggressive in nature. But Lovins Stemmer, Damodharan Tamil Stemmer and PLIS are more aggressive than the above said stemmers.

Correctly Stemmed Word Factor and Average Words Conflation Factor

Correctly Stemmed Word Factor (CSWF) and Average Words Conflation Factor (AWCF) obtained for Lovins Stemmer are 88.77% and 60.7%, for Iterated Lovins Stemmer is 89.10% and 62.1%, for Porter1 Stemmer is 79.27% and 55.5% respectively. These stemmers accuracy of correctly stemmed words and conflating variant words of the same group into correct stem is good, but not satisfactory.

Most of the time Average Word Conflation Factor obtained by Lovins Stemmer, Iterated Lovins Stemmer and Porter1 Stemmer is less than zero (negative values). CSWF and AWCF scores obtained by Porter2, Paice/Husk, Fairwheather, Porter French Stemmer, Damodharan Tamil Stemmer, Gupta Hindi Stemmer and PLIS are good and satisfactory. The accuracy of the existing stemmers is compared to PLIS. However, the accuracy of the PLIS is 98.39%. It is realized that PLIS achieves greater results than the other stemmers.

Over-Stemming and Under-Stemming

Additionally, the following performance results are also observed. The Word Stemmed Factor (WSF) acquired by Lovins Stemmer is 73.36%, Iterated Lovins Stemmer is 67.17%, Porter1 Stemmer is 67.76% and Porter2 Stemmer is 66.58%. These results are considered good. But CSWF is comparatively low and most of the times, these stemmer produces negative value for AWCF. It is so because the word has inflectional and derivational suffixes. This also alters the root words to incorrect stem. While considering over-stemming and under-stemming errors it is possible that those occur in the above specified stemmers. However, occurrence of over-stemming errors is more than under-stemming errors. The occurrence of under-stemming errors is high in Damodharan Tamil Stemmer as compared to the other stemmers and NWC is high also in the same stemmer. This happens because the variant words of same conflation class are transformed to different stems.

When calculating AWCF, Porter1 stemmer obtains negative value compared to porter2 stemmer, due to the occurrence of more over-stemming errors. The performance of Proposed Language Independent Stemmer is slightly better than the other stemmers on MWC, ICF, WSF and CSWF. The AWCF obtained by PLIS is 63.5%, which is comparatively higher than the other stemmers. The reason behind this is that PLIS stemmer is more aggressive than the other stemmers as WSF obtained is 73.52%, which is higher than the other stemmers.

Number of Unique Words after Conflation

The Number of unique Words after Conflation (NWC) is greater in Damodharan Tamil Stemmer than Lovins Stemmer, Porter French Stemmer and PLIS. But NWC obtained by Lovins Stemmer, Porter French Stemmer, and PLIS are greater than the remaining stemmers. Hence the number of incidence of over-stemming and under-stemming errors on above stemmers is less compared to the other stemmers. Mean Modified Hamming Distance of Porter1 Stemmer provides higher score i.e., 9.94 and Gupta Hindi Stemmer gives lower score i.e., 1.55 compared to the other stemmers.

Query Throughput and Query Latency

Query throughput is the number of queries processed per second. Query latency is the time between issuing a query and receiving a response, measured in millisecond. Query Latency score of PLIS is lower than the others stemmers and thus it has been concluded that processing time of PLIS takes less time than the other stemmers. Query Throughput is comparatively higher in PLIS than the other stemmers. But most of the stemmers give good result in Query throughput and Latency. Therefore, from the above result it is observed that the PLIS provides better results compared to the existing stemmers based on its performance, strength and accuracy. The PLIS can be very well adopted in the applications that employ the stemming process. Thus PLIS achieves national and international scope of the research.

V. Conclusion

Stemming approaches in Information Retrieval Systems focus now on increasing the retrieval performance, consuming less time but providing greater accuracy, strength and supporting multi-linguistic documents need more attention. Adoption of rule-based, lookup and hybrid approaches are useful and have advantages, there are several limitations and problems. In view of the above aspects, this research work proposes the Language Independent Stemmer **[17], [18]** for Information Retrieval Systems using Dynamic Programming concepts.

The key features of the proposed stemmer are the following:

- Design the Language Independent Stemmer for Information Retrieval Systems with the support of accessing multi-linguistic web documents.
- Assured to generate accurate stem word without prior knowledge about the language.
- Effective performance of Character Analyzer using Dynamic Programming.
- Enhance the accuracy using Rule-Based Filtration.
- Improves the filtration using Words Filter.
- Handling the different morphological variants and generate accurate stem word.
- Support accented characters and diacritics efficiently.
- Time and memory consumption.
- Support left-to-right and right-to-left writing styles of the language.

The noteworthy advantages are the reduction of computational cost, manpower, time and providing greater accuracy. The dynamic programming concept incorporated in the PLIS generates high query throughput and the low query latency. The scores obtained for Mean number of Words per Conflation class of the proposed stemmer are positive. Therefore, the performance of the PLIS will be high. The Proposed Language Independent Stemmer gives higher Index Compression Factor, which reduces the storage overhead of repeated values. Additionally, average Words Stemmed Factor for PLIS gives high value, which is greater than minimum threshold value. Hence, the strength of the PLIS is strong and aggressive in nature. Correctly Stemmed Words Factor of the PLIS is high and PLIS reaches 98.397% of accuracy. Furthermore, Average Words Conflation Factor of the PLIS achieves higher percentage. Hence, the PLIS generates accurate result.

Therefore, the proposed approach for Language Independent Stemmer in this research work is well formulated, universal for morphologically different languages and widespread with efficiency using Dynamic Programming. It overcomes the limitations encountered in Language Dependent and Independent Stemmers proposed earlier.

References

[1]. [Husain, 2012] *Mohd. Shahid Husain, "An Unsupervised Approach to Develop Stemmer", In: International Journal on Natural Language Computing (IJNLC), ISSN: 2278-1307, Volume.1, Issue.2, India, 2012.*

[2]. [Khan et al. 2012] *Sajjad Ahmad Khan, Waqas Anwarl, Usama Ijaz Bajwa1, and Xuan Wang, "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language", In: Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pp.69–78, India, 2012.*

[3]. [Sethi, 2013] *Dhabal Prasad Sethi, "Design of Lightweight Stemmer for Odia Derivational Suffixes", In: International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Print): 2319-5940, ISSN (Online): 2278-1021, Volume.2, Issue.12, India, 2013.*

[4]. [Parlak, 2012] *Sıddıka Parlak, "Performance Analysis and Improvement of Turkish Broadcast News Retrieval", In: IEEE Transactions on Audio, Speech and Language Processing, ISSN: 1558-7916, Volume.20, Issue.3, pp.731-741, New Jersey, 2012.*

*[5]. [**Adege et al. 2017]** Abebe Belay Adege, Yibeltal Chanie Manie, "Designing a Stemmer for Ge'ez Text Using Rule Based Approach", In: International Journal of Scientific & Engineering Research, ISSN: 2229-5518 Volume.8, Issue.1, pp.1574-1578, Ethiopia, 2017.*

[6]. [Zadeh et a. 2017] *Hossein Taghi-Zadeh , Mohammad Hadi Sadreddini, Mohammad Hasan Diyanati and Amir Hossein Rasekh, "A new hybrid stemming method for persian language", In: Digital Scholarship Humanities, Volume.32, Issue.1, pp.209-221, Pakistan2017.*

[7]. [Patel et al. 2016] *Chandrakant D. Patel and Jayeshkumar M. Patel, "Improving a Lightweight Stemmer Language for Gujarati", International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016.*

[8]. [Ali et al. 2016] *Mubashir Ali, Shehzad Khalid, Haneef Saleemi, Waheed Iqbal, Armughan Ali and Ghayur Naqvi, "A Rule based Stemming Method for Multilingual Urdu Text", In: International Journal of Computer Applications , ISSN: 0975 – 8887, Volume.134, No.8, Palistan 2016.*

[9]. [Nehar et al. 2015] *Attia Nehar, Djelloul Ziadi, and Hadda Cherroun, "Rational Kernels for Arabic Stemming and Text Classification", In: arXiv:1502.07504v1, Algerie, 2015.*

[10]. [Deepamala et al. 2015] *Deepamala, N., and Ramakanth Kumar, P., "Kannada Stemmer and Its Effect on Kannada Documents Classification", In: Computational Intelligence in Data Mining, Volume.33, pp.75-85, Springer, India, 2015.*

[11]. [Karanikolas, 2015] *Nikitas Karanikolas, N., "Supervised learning for building stemmers", In: Journal of Information Science, Volume.41, Issue.3, pp.315-328, Greece, 2015.*

[12]. [Brychcín et al. 2015] *Tomaš Brychcín, and Miloslav Konopik, "HPS: High Precision Stemmer", In: Information Processing and Management, Volume.51, Issue.1, pp.68–91, ELSEVIER, Czech Republic, 2015.*

[13]. [Gupta, 2014a] *Vishal Gupta, "Hindi Rule Based Stemmer for Nouns", In: International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277-128X, Volume.4, Issue.1, pp.62-65, India, 2014.*

[14]. [Joshi et al. 2014] *Garima Joshi, and Kamal Deep Garg, "Enhanced Version of Punjabi Stemmer Using Synset", In: International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277-128X, Volume.4, Issue.5, India, 2014.*

[15]. [Karanikolas, 2014] *Nikitas Karanikolas, N., "A Methodology for Building Simple but Robust Stemmers without Language Knowledge: Stemmer Configuration", In: Social and Behavioral Sciences Volume.147, pp.370 – 375, ELSEVIER, Greece, 2014.*

[16]. [Kasthuri et al. 2014a] *Kasthuri, M., and Dr. Britto Ramesh Kumar, S., "A Comprehensive Analyze of Stemming Algorithms for Indian and Non-Indian Languages", In: International Journal of Computer Engineering and Applications (IJCEA), ISSN: 2321-3469, Volume.7, Issue.3, pp.1-8, India, 2014.*

[17]. [Kasthuri et al. 2015] *M. Kasthuri, Dr. S. Britto Ramesh Kumar, "A Framework for Language Independent Stemmer Using Dynamic Programming", In: International Journal of Applied Engineering Research (IJAER), Print ISSN 0973-4562, Volume.10, Number.18, pp.39000-39004 Online ISSN 1087-1090, India, 2015.*

[18]. [Kasthuri et al. 2016] *M. Kasthuri, Dr. S. Britto Ramesh Kumar, "PLIS: Proposed Language Independent Stemmer for Information Retrieval Systems Using Dynamic Programming", In: 2016 World Congress on Computing and Communication Technologies, ISBN: 978-1-5090-5573-9, pp.132-135, IEEE, India, 2016.*

[19]. [Kasthuri et al. 2014b] *M. Kasthuri and Dr. S. Britto Ramesh Kumar, "Multilingual Phonetic Based Stem Generation", In: Proceedings of the Second International Conference on Emerging Research in Computing, Information Communication and Applications (ERCICA-2014), **ELSEVIER** Science::Technology India, ISBN: 9789351072607, Volume.1, pp.437-442, 01-02 August, Bangalore, India, 2014.*

[20]. [Porter, 2001] *Porter, M.F., "Snowball: A Language for Stemming Algorithms", Cambridge, 2001.*

[21]. [Fairwheather, 2011] *John Fairwheather, "Language Independent Stemming", In: Patent no: US 8015175 B2,*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Publication number: US8015175B2, California, 2011.

[22]. [EMILLE, 2014] *Emille Corpus, "Tamil and Hindi corpus", France, 2014. [http://catalog.elra.info/ product_info.php]*

[23]. [Rajalingam, 2013] *Damodharan Rajalingam, "A Rule Based Iterative Affix Stripping Stemming Algorithm for Tamil", In: 12th International Tamil Internet Conference, pp.28-33, Malaysia, 2013.*

[24]. [Gupta, 2014b] *Vishal Gupta, "Hindi Rule Based Stemmer for Nouns", In: International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277-128X, Volume.4, Issue.1, pp.62-65, India, 2014.*

[25]. [Chintala et al. 2014] *Dileep Reddy Chintala and Madhusudhana Reddy, E., "An Approach to Enhance the CPI Using Porter Stemmer Algorithm", In: International Journal of Advanced Computer Science, Volume.4, Issue.12, pp.556-564, India, 2014.*

Acknowledgement

I wholeheartedly thank University Grants Commission (UGC), Hyderabad for sanctioning the grant to complete the Minor Research Project entitled Language Independent Stemmer. I would like to express my sincere thanks to the Corpus Provider (EMILLE Corpus) who generously allowed me to download data from their website with free of cost.

Author's Profile

Dr. M.Kasthuri is working as an Assistant Professor in the Department of Computer Applications, Bishop Heber College, Tiruchirappalli, Tamil Nadu, India. She had completed her Doctorate of Philosophy in Computer Science in June 2017 at Bharathidasan University, Tiruchirappalli. She has published a number of National and International level research papers related to Web Mining and Stemming concepts. She has completed UGC sponsored Minor Research Project entitled as Language Independent Stemmer.

Dr. S.Britto Ramesh Kumar is working as an assistant Professor in the Department of Computer Science, St.Joseph's College, Tiruchirappalli, Tamil Nadu, India. He had completed his Doctorate of Philosophy in Computer Science in June 2011 at Bharathidasan University, Tiruchirappalli. His Interested Research Areas are Software Architecture, Mobile Technologies, Web Services and Information Security. He has published many papers in National and International level Journal and in Conference Proceedings. He also conferred with Best Researcher Award for 2008 by BHC management.