Efficient Search Mechanism from Large Scale Corpora for Domain-Specific Language Modeling in Speech Recognition

Disha Kaur Phull, G. Bharadwaja Kumar

Abstract**:** ***With the Internet and the World Wide Web revolution, large corpora in variety of forms are germinating ceaselessly that can be manifested as big data. One obligatory area for the usage of such large corpora is language modeling for large vocabulary continuous speech recognition. Language modeling is an indispensable module in speech recognition architecture, which plays a vital role in reducing the search space during the recognition process. Additionally, the language model that is contiguous to the domain of the speech can dwindle the search space and escalate the recognition accuracy. In this paper, an efficient searching mechanism for domain-specific document retrieval from the large corpora has been elucidated using Elasticsearch which is a distributed and an efficient search engine for big data. This assisted us in tuning the language model in accordance with the domain and also by reducing the search time by more than 90% in comparison to conventional search and retrieval mechanism used in our earlier work. A word level and a phrase level retrieval process for creating domain-specific language model has been implemented. The evaluation of the system is performed on the basis of word error rate (WER) and perplexity (PPL) of the speech recognition system. The results shows nearly 10% decrease on WER and a major reduction in the PPL that helped in boosting the performance of the speech recognition process. From the results, it can be consummated that Elasticsearch is an efficient mechanism for domain specific document retrieval from large corpora rather than using topic modeling toolkits.***

Keywords **:** ***Adapted Language Model, BigData, Document Retrieval, ElasticSearch, Indian English, Lecture Speech, Speech recognition.***

I. INTRODUCTION

The increased access to information through World Wide Web and digital revolution has apparently led to the creation, generation and circulation of numerous digital data. A remarkably large number of extremely voluminous data which requires an extensive application of computationally intensive techniques has coerced the introduction of the term 'Big Data'. Generally, data that is too large, complex, and dynamic in a way that it is impractical for any conventional hardware/software tools or systems to manage and process in a timely manner and scalable fashion is defined as BigData [1]. The conventional systems find it challenging to capture, store, analyze, curate, search, share, transfer, visualize, query, update and secure this humongous data. Few technologies, frameworks or architectures like Apache Hadoop, MapReduce, Spark, Apache Lucene and Mahout have emerged for resolving these challenges. Searching and retrieving documents from ones desktop itself is undeniably a difficult and time consuming task if proper file information is incongruous. The task of document retrieval using a user query against a set of free-text records from this humongous data is actually very arduous but yet a vital constraint for many applications including language modeling in speech recognition process. Apart from document retrieval, topic modeling has been rigorously used for retrieving domain-specific information for language modeling in speech recognition.

Speech Recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words [2]. Language model plays an indispensable role in a large vocabulary continuous speech recognition system (LVCSR) by reducing the search space for the recognizer. The performance of the speech recognizer can be improved if the element in the language model is in close proximity with the speech to be recognized. The retrieval of the domain specific documents over a large collection of documents can be time consuming with the traditional searching approach. The inclusion of topic modeling enabled many research works to reduce the search space & time and increase the overall performance of the speech recognition process by yielding the domain-specific documents. Owing to topic modeling, it is imperative to provide the number of topics to be generated in advance but in view of large corpora only an approximation is made. The appropriate number of topics has to be decided with experimentation by changing the number of topics each time for creating topic models. Also, in most of the topic modeling mechanisms, the model is created once all the documents are loaded and if any document has to be included in the later stage, the whole process of model creation and indexing has to be performed to include those documents. In order to overbear these constraints, in this paper,

Revised Manuscript Received on August 30, 2019.

* Correspondence Author
Disha Kaur Phull*, SCSE, VIT University Chennai Campus, Tamil Nadu, India. Email: dkphull@gmail.com
G. Bharadwaja Kumar, SCSE, VIT University Chennai Campus, Tamil Nadu, India. Email: bharadwaja.kumar@vit.ac.in

we have proposed a dynamically adaptive methodology for building the domain-specific language models. The Wikipedia dump and sources from web are used for building the domain- specific language models. Typically, Wikipedia dump comprises of multifarious domains [3] and has been widely used resource in many natural language processing tasks. We have used Elasticsearch for searching and retrieving documents effectively from Wikipedia dump which is having very large collection of documents. Elasticsearch which is a distributed, RESTful analytical search engine with leverages like efficient scalability, predictability, reliability, simplicity and transparency in querying and analyzing the data is an ideal choice for the document retrieval task [4]. In the present work, we have presented a LVCSR system with dynamic language model adaptation for Indian English. Indian English (IE) is referred to as the English spoken by the Indians as their second language [5]. Performing speech recognition on IE is very complex as it has a dominant native language (L1) influence over English (L2 - second language) [6].

Outline. The remaining paper is organized as follows. Section II gives an account of related work regarding document retrieval, language modeling and bigdata. The overview of the speech recognition process is described in section III. The phase-1 speech recognition framework is mentioned in section IV. The section V depicts the architecture incorporating the bigdata framework and its process flow section VI shows the experimental results for the phase-1 as the base recognition results, word level and the phrase level retrieval involved for language model with comparative results. This would be useful for evaluating the language model's performance. Finally, section VII gives the conclusions for language modeling and bigdata in automatic lecture speech recognition system.

II. RELATED WORKS

Various search and retrieval techniques have been evolving continuously for enhancing the querying process as it is an imperative task for any search engine. Latent Semantic Analysis (LSA) [7] and Latent Dirichlet Allocation (LDA) [8] are two widely used techniques for document retrieval, full-text retrieval [9], information retrieval [10] and topic modeling [11]. Similar techniques have been explored for topic identification and dynamic language model adaptation using vector space model [12], LSA [13], relevance language model [14], semi-supervised language models [15] and topic tracking language model [16]. LDA technique has been widely explored to form unsupervised adapted language model [17] and topic-specific language models for inflectional languages [18]. For many languages the linguistic word level approach [19], syntactico-statistical approach [19] and statistic phrase level approach [20] has been used to build an adapted language model for improving the speech recognition rate. Document retrieval from web content [21, 22. 23, 24] for constructing a language model has also been carried out in-order to improve the accuracy of a speech recognizer. Language models have been formulated using Neural networks and deep neural networks [25] for improving the performance of a speech recognizer. Recently, bigdata technologies have been applied for document retrieval and information mining by incorporating MapReduce framework [26] and Hadoop environment [27]. Hadoop has also been used in speech processing for voice recognition [28] and audio emotion recognition [29]. Substantially, hadoop on MapReduce frame work has also been applied on speech recognition [30, 31] and language modeling with mapreduce [32, 33] has also been explored. Research on lecture recognition for Indian English [34] justifies the need for a modified language modeling strategy for lecture recognition task.

III. OVERVIEW OF SPEECH RECOGNITION PROCESS

A general speech recognition process requires three main models: Acoustic model, language model and pronunciation model. These three models are utilized by the speech recognizer for decoding the speech. The overall speech recognition architecture used in this paper is explained in Fig.1.

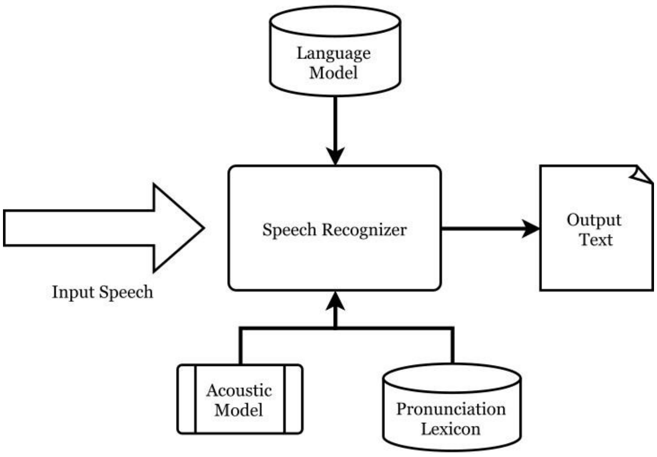

Fig. 1. General speech recognition process.

Acoustic Model: Acoustic modeling of speech typically refers to the process of establishing statistical representations for the feature vector sequences computed from the speech waveform. Hidden Markov Model (HMM) is one of the most common acoustic models used in speech recognition process. In HMM, the states (words or phonemes) are not directly visible, but variables (feature vector sequences) influenced by the states are visible. Each state has a probability distribution over the possible output observation. Each state's output distribution is represented by a multivariate GMM to account for multivariate and multi-modal data due to variations in speaker, accent and gender. Training of context dependent multi-state HMM models comprises of training Context Independent HMM models, Context Dependent HMM models, building decision trees and Gaussian mixture generation. We have used SphinxTrain [1] for acoustic modeling in our LVCSR experiments.

Pronunciation Model: Speech is a continuous audio stream where stable states mix with dynamically changing states. In this sequence of states, one can define more or less similar classes of sounds, or phones. A phonetic dictionary contains a mapping from words to phones. Our phonetic dictionary contains 41 phones that are specific to Indian accent.

[1] http://cmusphinx.sourceforge.net/wiki/tutorialam

Language Model: Language models help any speech recognizer to figure out how likely a word sequence is independent of the acoustics. Also, language models play a vital role in resolving acoustic confusions arising due to co-articulation, assimilation and homophones during the decoding process. In addition, continuous speech recognition suffers from difficulties such as variation due to sentence structure (prosodies), interaction between adjacent words (crossword co-articulation) and absence of clear acoustic markers to delineate word boundaries.

Decoding: The HMM decoders find the most probable word sequences by searching triphone sequences present in acoustic model for a particular word. The decoder makes the search decisions in speech recognition using viterbi algorithm which follows the dynamic programming approach. The decoder requires the acoustic model, pronunciation lexicon and the language model for efficient speech recognition.

IV. PHASE-1 SPEECH RECOGNITION

In general, LVCSR systems use language models which are representatives of a broad corpus. But, in actual practice, however, recognition is usually on a coherent text covering a single topic, suggesting that knowledge of the topic at hand can be used to advantage. In addition, the perplexity and out of vocabulary can lead to poor decoding if one uses generic language models. Hence, a base model can be augmented with information from a small sample of domain-specific language data to significantly improve recognition performance [35].

Initially, when any speech is subjected to a recognizer, prior knowledge about the domain of speech will be unknown and to dynamically channel the recognition process towards the possible topic under recognition, we require baseline/phase-1 recognition along with an appropriately adapted language model. Hence, we explore the adaptation of language model for channeling domain-specific speech recognition by retaining the conventional recognition system as phase-1 and have carried out our experimentation consequent to phase-1. Phase-1 speech recognition is explained in this section and the consequent domain-specific LVCSR is explained in section V.

The language model for phase-1 recognition is assembled such that it forms the baseline language model which can be utilized for any domain in speech recognition. Phase-1 language model constitutes of two corpuses, first is the base corpus, which comprises of the transcriptions from the training data to complement the speech in the lecture. The second corpus used is the Wikipedia dump[2] which is included to increase the comprehensive variety to inculcate various domains. Preprocessing is performed on the base corpus and Wikipedia dump to make it congruent for creating language model and it decreases the ambiguities prevalent in the text corpus. Two language models are constructed: a tri-gram language model created from the baseline corpora and a bi-gram language model from Wikipedia dump. These baseline corpora and the Wikipedia dump are combined for creating the language model for the Phase-1 recognition process. The speech recognizer engine uses the acoustic model, pronunciation lexicon and the Phase-1 language model to generate the output transcription for the input speech.

V. INGRESSION OF BIG DATA IN SPEECH RECOGNITION

Generally, data sets that are very large or complex where the traditional data processing application software is inadequate to perform data manipulation are defined as big data. A technology which can help us to deal with big data is really on demand, as there is a massive increase in the volume and variety of data. Hence, the roots of big data are slowly penetrating into speech recognition also an initial experimentation to engulf the benefits of big data have been triggered. However the various potentials of bigdata are yet to be explored and utilized. For our present work, we require a powerful searching, indexing and querying engine for performing document retrieval from Wikipedia dump which comprises of nearly 1.9 million files and nearly 8GB in size after selectively choosing the files more than 1KB and some amount of pre-processing. The humongous nature of this data makes it much more challenging to retrieve documents even after the ingression of big data technologies and framework such as Hadoop. After contemplating the benefits and drawbacks of several possible bigdata technologies, we capitulate this humongous nature of data by incorporating Elasticsearch engine for document retrieval in this paper.

Elasticsearch is a highly scalable open-source full-text search and analytics engine [36]. It allows you to store, search, and analyze huge volume of data quickly and in a real time environment. It is generally used as the underlying engine/technology which powers applications that have complex search features and requirements. Some of the beneficial components in Elasticsearch which enables the smooth functioning of the search and analysis are Near Realtime (NRT), cluster, node, index, type, document and shards. Conventionally, an index can potentially store a large amount of data that can exceed the hardware limits of a single node which causes retrieval overheads. To solve this problem, Elasticsearch provides the ability to subdivide your index into multiple pieces called shards which help in efficient indexing and faster retrieval. An index is stored in a set of shards, which themselves are Lucene indices. Due to this fact, if the documents have to be included into Elasticsearch at any later stage, the indexing of those documents are performed while loading of those documents, as the indexing on the whole document again is needless. When comparing to the topic modeling mechanism, in Elasticsearch, it is inessential to provide the number of topics to be generated from the corpora.

A. Incorporation of Elasticsearch for Language Model Adaptation

The inclusion of Elasticsearch framework provides an overall change in the construction of a domain specific language model for speech recognition and is depicted in Fig.2.

[2] http://en.wikipedia.org/wiki/Wikipedia:Database

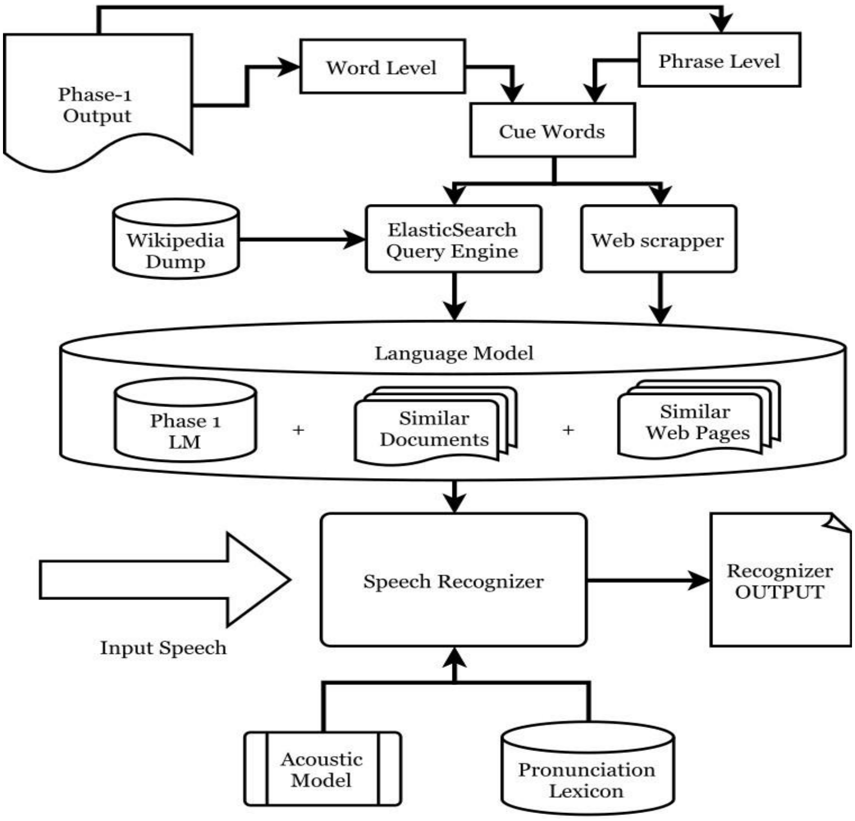

Fig. 2. Incorporating Elasticsearch for adaptation of language model in speech recognition process.

Table- I: Example of finding the bag-of-words as cue words from the ASR output.

Words	PoS Tags
the	DT
database	**NN**
requirements	**NNS**
are	VBP
essentially	RB
method	**JJ**
what	WP
are	VBP
the	DT
data	**NN**
elements	**NNS**

Table- II: Example of finding the multi-words as cue words from the ASR output.

Example	the database requirements are essentially method what are the data elements
Multi-Word	database requirements A-N

Cue Words:

Typically, a cue word is obtained from a user query as an indicator for carrying out the subsequent search process in any conventional search and retrieval system. In LVCSR, the cue words for carrying out the domain specific document search can be captured from the Phase-1 output. A word or a multi-word can be considered as a potential cue word from Phase-1 output by using Parts of Speech (POS) tags, multi-word expressions and Named Entity Recognition features. The word level and multi-word level cue word extraction and retrieval process is described as follows:

Word Level Process: In our experimentation, POS tagging has been used for retrieving the cue words relevant to a topic using Natural Language Toolkit (NLTK) with python [37]. After analysis, the significant POS tags considered are NNS (Noun Plural), NN (Noun Singular), JJ (Adjective), VBN (verb past participle), JJS (Adjective superlative) and VB (Verb Baseform). These POS tags are found to be most effective to get the theme of the domain. We have constructed a bag of words which contains words repeated more than 8 times along with the POS tags which is further used for word level retrieval in speech recognition process. A small text, for example, with its POS tags is shown in Table I, where the highlighted words are considered to be the cue words. This example depicts that the topic to which the text belongs could be “database management systems” when we consider the context of the highlighted words.

From this, it is evident that an efficient set of cue words can be used for document retrieval which is further utilized for word level retrieval process.

Pharse Level Process: A multi-word is also considered as cue word for retrieving documents relevant to a topic. The multi-word extraction process is carried out by selecting the combination of words with specific POS tags. The combinations considered are noun - adjective, adjective - noun, noun - noun and noun - preposition - noun.

These combinations are found to be effective for filtering out the theme of the topic. Multi-words which follow any of these combinations and also have a relatively high score throughout the document are taken as cue words for further phrase level speech recognition process. For example, consider a sample output as shown in Table II, the multi-word extracted for this output is “database requirements” which follows the combination of adjective - noun with a score of 0.9. Linguakit [38] has been used for multi-word extraction and these extracted multi-words are used for document retrieval.

Retrieval Process

The word level or multi-word level cue words are now given as a query for retrieval process to the web scrapper and Elasticsearch engine. This list is designed to follow the bag of words approach and can be specified as word level and phrase level (multi-word) retrieval.

- **Web scrapper:** It is used to extract pages from the World Wide Web. The algorithm for the web scrapper is given in algorithm 1. It requires the cuewords ‘k’ and an output file ‘f’ will be created which has the extracted web specific pages according to the cuewords passed. The cue words are sent as a query to web which retrieves the documents relevant to the domain under recognition. The search query retrieves URLs with a threshold of first 20 domain specific web pages. A tri-gram Language model is formed with these extracted specific web pages for each domain.

```
Web Scrapper (k,f):
for each url in search(k) do
    file=“test.pdf”;
    if pdfDownloadFile(url, file) then
        x=convertPdfToText( file);
    else
        x=extract(url);
    end
    f.append(x);
end
```

Algorithm 1: The algorithm for web scrapper.

- **Elasticsearch engine:** It is used to search required documents from the Wikipedia Dump. The algorithm for the Elasticsearch engine is given in algorithm 2. It requires the cuewords ‘k’, location ‘W’ of Wikipedia Dump and an output file ‘Es’ will be created which has the extracted similar documents according to the cuewords passed. The cue words are also sent as a query to Elasticsearch query engine which has indexed the Wikipedia dump documents. For a document to be retrieved, we have confined that the document should suffice two criteria: the first, a threshold of ‘n’ word/multi-word should be present in the document with respect to the search query and second, a minimum document score of ‘s’ should be maintained (value of n and s varies for word level and phrase level). This in turn helps us in retrieving a set of similar documents for a particular domain. The main advantage of Elasticsearch is that it takes less than 30 seconds to retrieve the relevant documents from any large corpora. A tri-gram language model is formed using contextually similar documents retrieved for each domain by the Elasticsearch engine.

```
ElasticSearchEngine (W,k):
Result: Es: The extracted Wikipedia files appended
for each file in W do
    for each word in  file do
        if ‘n’ keywords match in the file then match=True;
        else
            match=False;
        end
    end
    if match==True then
        if documentScore of file greater than 's' then
            Es.append( file);
        end
    end
end
```

Algorithm 2: The algorithm for ElasticSearch Engine.

The language model from phase-1 is retained as a generic model; the retrieved documents and the web pages are further preprocessed and augmented into this language model to form an ‘adapted language model’. This adapted language model can be used for decoding in a speech recognizer. Language Model coalescing is an effective technique for incorporating domain-specific information into a language model and abridges vocabulary necessary for a domain-specific LVCSR system. The interpolation of the language models from phase-1, web similar pages and Wikipedia similar documents has been carried out using SRILM toolkit [39] for respective domain. LVCSR system uses the acoustic model, pronunciation lexicon as discussed in section III and utilizes this adapted language model for recognizing the input speech domain to generate the output text.

VI. EXPERIMENTAL EVALUATIONS

A. Data and Setup

NPTEL[3] lecture videos have been used for building the acoustic models. These lectures are a perfect mix of variation in terms of topic, domain and speakers (various regions of India). We have considered lecture videos of 75 speakers and these lectures have been transcribed to train the acoustic model. The lectures have been video recorded in 44 kHz sampling frequency which has consequently down-sampled to 16 kHz and 16 bit mono file format. We have considered a minimum of 15 minutes of speech for each speaker for training. The total speech data for training comprises of 19 hours, the test data consists of 20 minutes audio each for 10 different domains/topics of NPTEL video lectures. The 10 different domains considered for testing are Product Life cycle (PL), Population Studies (PS), Computer organization (CO),

[3] http://nptel.ac.in

Database (DB), Computer Architecture (CA), Computational Techniques (CT), Scalar Random variables (SR), Axioms Probability (AP), Enzymes (EZ) and Amino Acids (AA). The total test set comprises of 3 hours and 20minutes (200 mins) of speech for the evaluation and is not a part of our training set.

The evaluation metrics used are word error rate (WER) and perplexity (PPL). Word error rate (WER) is computed as $\mathrm{WER} = \frac{S+D+I}{N}$. Where, S denotes the number of substitutions, D the number of deletions, I the number of insertions and N the total number of words present.

Perplexity can be computed as a discrete probability distribution p, defined as $2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)}$. Where, $H(p)$ is the entropy distribution and depends on the words.

B. Result Analysis and Discussion

The performance of the phase-1 recognition is evaluated using Word Error Rate (WER) and perplexity (PPL). The least WER has found to be 19.2% for AA and a highest of 54.6% for CA. The average WER for phase-1 recognition has been found to be around 40.08%. It can be noticed from Table III that AA has the lowest WER owing to its low perplexity measure. CA with highest WER has been obtained because of its higher perplexity. From this section, the overall analysis for the phase-1 process for lecture speech recognition is elucidated and the phase-1 is considered as a baseline result for further comparison.

Further experimentation has been carried out by incorporating the Elasticsearch on a bigdata framework as already mentioned in section V. The LVCSR process after the inclusion of word level and phrase level document retrieval was performed and the recognition results show a notable difference from the Phase-1 process. The overall performance of the speech recognition system for Phase-1 with word level and phrase level retrieval is evaluated using PPL and WER and is depicted in Fig.3. The WER and PPL of CA in the test set has been reduced drastically after the inclusion of domain-specific retrieval at both word level and phrase level. Although, the equipotential performance of word level and phrase level retrieval is clear from the WER and PPL values, both induced a huge improvement in the recognition rate of the LVCSR.

From Fig.3, it can be clearly noted that the language models perplexity has a positive impact on the WER of the speech recognition output. There is a concurrent decrease in the WER and PPL which improves the performance of the LVCSR system. The indifference between word level and phrase level search performance can be due to the fact that the ASR output from the phase-1 was not able obtain any distinct retrievals. Also, multiword expression identification for the phrase level search from a continuous lecture speech itself was quite challenging due to the presence of ungrammatical sentences.

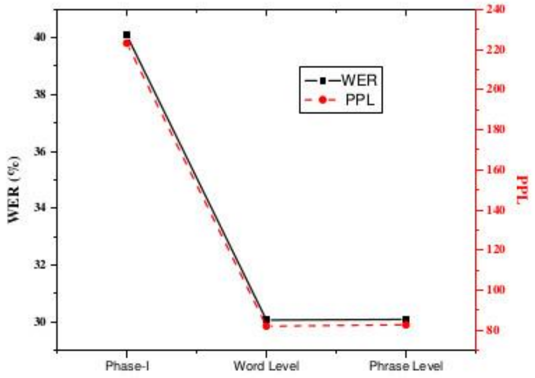

Fig. 3. The impact of perplexity of the language model on the WER achieved.

From the analysis of results, it can be clearly stated that language model plays a vital role in enhancing the performance of the speech recognition. It is also noticed that the inclusion of Elasticsearch for indexing and searching from a large corpora has helped in efficient retrieval of relevant documents, which made the adapted language model incline towards the specific domain of speech and has resulted in improved recognition. The result depicted by our approach is comparable to the results in the literature [40] and even better. This work explicates the need and advantage of the Elasticsearch framework to process large scale corpora in language modeling for speech recognition.

VII. CONCLUSION

The objective of this paper is to efficiently utilize large corpora to acclimate an adaptive language model and experiment its impact on the efficiency of the speech recognition process. The present paper takes cue words from the output of the baseline speech recognition process (Phase-1) to fetch the domain-specific documents for language model towards the domain of the input speech. Document retrieval is done in two parts: one from Wikipedia dump and another from the web simultaneously. It is a well-known fact that Wikipedia dump is a large multifarious domain corpus; we have used Elasticsearch for searching and retrieving domain specific documents from this large corpus. The cue words were formed at word level and phrase (multi-word) level for the retrieval of the documents from the large corpora to validate its impact on the speech recognition process. The cue words were sent as queries for the retrieval task which helps us to reach the close proximity to the topic of input speech. In the phase-1 process the WER was 40.08% and perplexity measure has been nearly 222. The domain-specific document retrieval which is further classified into word level and phrase level processing, outperformed from the phase-1 recognition process. The results shows that there was approximately 10% decrease in WER and the perplexity measure reduced to around 80. The obscurity in the cue words led to fuzziness in the document retrieval which is depicted in the word level and phrase level results.

This paper was able to show the benefaction of integrating the Elasticsearch engine efficiently in document retrieval on domain-specific language modeling for Speech Recognition.

Table- III: Speech Recognition results Using Word Level and Phrase level retrieval.

Topic	Phase I		WordLevel		PhraseLevel	
	WER (%)	PPL	WER (%)	PPL	WER (%)	PPL
PL	34.6	291.381	34.2	233.21	34.2	233.25
PS	36.2	378.729	12.4	23.47	12.4	23.47
CA	54.6	267.86	29.3	12.84	29.3	12.84
CO	27.8	52.808	32.5	66.7	32.5	73.57
CT	34.3	188.071	17.1	13.54	17.1	13.54
SR	44.1	232.205	40.2	133.64	40.3	133.66
AP	51.2	208.29	44.5	106.97	44.6	106.97
EZ	50.3	294.32	25.8	24.72	25.8	24.72
AA	19.5	38.96	20.6	46.66	20.6	46.66

REFERENCES

1. G. B. Kumar, "An encyclopedic overview of big data analytics", *International Journal of Applied Engineering Research*, vol. 10, no. 3, 2015, pp. 5681–5705.
2. R. Lawrence, Fundamentals of speech recognition. Pearson Education India, 2008.
3. Y. Zhang, J. Niehues, and A. Waibel, "Integrating encyclopedic knowledge into neural language models," *in IWSLT workshop*, 2016.
4. A. Paro, ElasticSearch cookbook. Packt Publishing Ltd, 2015.
5. J. C. Wells, Accents of English, vol. 1. Cambridge University Press, 1982.
6. R. Gargesh, "Indian English: Phonology", Varieties of English: Africa, South and Southeast Asia. New York: Mouton de Gruyter, 2008, pp. 231–243.
7. T. K. Landauer, Latent semantic analysis. Wiley Online Library, 2006.
8. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", *Journal of machine Learning research*, vol. 3, no. Jan, 2003, pp. 993–1022.
9. D. C. Blair and M. E. Maron, "An evaluation of retrieval effectiveness for a full-text document retrieval system", *Communications of the ACM*, vol. 28, no. 3, 1985, pp. 289–299.
10. C. D. Manning, P. Raghavan, H. Sch¨utze, et al., Introduction to information retrieval, vol. 1. Cambridge University press Cambridge, 2008.
11. H. M. Wallach, "Topic modeling: beyond bag-of-words", *in Proceedings of the 23rd international conference on Machine learning*, ACM, 2006 pp. 977–984.
12. H. Yamazaki, K. Iwano, K. Shinoda, S. Furui, and H. Yokota, "Dynamic language model adaptation using presentation slides for lecture speech recognition", *Proc. INTERSPEECH* 2007, 2007, pp. 2349–2352.
13. J. D. Echeverry-Correa, J. Ferreiros-L´opez, A. Coucheiro-Limeres, R. C´ordoba, and J. M. Montero, "Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition", *Expert Systems with Applications*, vol. 42, no. 1, 2015, pp. 101–112.
14. B. Chen and K.-Y. Chen, "Leveraging relevance cues for language modeling in speech recognition", *Information Processing & Management*, vol. 49, no. 4, 2013, pp. 807–816.
15. S. Novotney, R. Schwartz, and S. Khudanpur, "Getting more from automatic transcripts for semisupervised language modeling", *Computer Speech & Language*, vol. 36, 2016, pp. 93–109.
16. S. Watanabe, T. Iwata, T. Hori, A. Sako, and Y. Ariki, "Topic tracking language model for speech recognition", *Computer Speech & Language*, vol. 25, no. 2, 2011, pp. 440–461.
17. M. A. Haidar and D. O'Shaughnessy, "Unsupervised language model adaptation using lda-based mixture models and latent semantic marginals", *Computer Speech & Language*, vol. 29, no. 1, 2015, pp. 20–31.
18. T. Brychc´ın and M. Konop´ık, "Latent semantics in language models", *Computer Speech & Language*, vol. 33, no. 1, 2015, pp. 88–108.
19. A. Toral, P. Pecina, L.Wang, and J. van Genabith, "Linguistically-augmented perplexity-based data selection for language models", *Computer Speech & Language*, vol. 32, no. 1, 2015, pp. 11–26.
20. X. Liu, M. J. Gales, and P. C. Woodland, "Paraphrastic language models", *Computer Speech & Language*, vol. 28, no. 6, 2014, pp. 1298–1316.
21. S. Oger and G. Linar`es, "Web-based possibilistic language models for automatic speech recognition", *Computer Speech & Language*, vol. 28, no. 4, 2014, pp. 923–939.
22. C. Eickho_ and A. P. de Vries, "Robust statistical methods in web retrieval", *ACM SIGWEB Newsletter, no. Winter*, 2016, p. 4.
23. C. Munteanu, G. Penn, and R. Baecker, "Web-based language modelling for automatic lecture transcription.", *in INTERSPEECH*, 2007, pp. 2353–2356,.
24. A. Sethy, P. G. Georgiou, and S. Narayanan, "Building topic specific language models from webdata using competitive models.", *in INTERSPEECH*, 2005, pp. 1293–1296.
25. X. Liu, M. J. Gales, and P. C. Woodland, "Language model cross adaptation for lvcsr system combination", *Computer Speech & Language*, vol. 27, no. 4, 2013, pp. 928–942.
26. W. Zhao, H. Ma, and Q. He, "Parallel k-means clustering based on mapreduce", *in IEEE International Conference on Cloud Computing Springer*, 2009, pp. 674–679.
27. E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello, "Sindice. com: a document-oriented lookup index for open linked data", *International Journal of Metadata, Semantics and Ontologies*, vol. 3, no. 1, 2008, pp. 37–52.
28. Y.-S. Chang, S.-H. Hung, N. J. Wang, and B.-S. Lin, "Csr: A cloud-assisted speech recognition service for personal mobile device," *in International Conference Parallel Processing (ICPP), 2011 IEEE*, 2011, pp. 305–314.
29. M. S. Hossain, G. Muhammad, M. F. Alhamid, B. Song, and K. Al-Mutib, "Audio-visual emotion recognition using big data towards 5g", Mobile Networks and Applications, vol. 21, no. 5, 2016, pp. 753–763.
30. Y. S. Tan, J. Tan, E. S. Chng, B.-S. Lee, J. Li, H. P. Chak, X. Xiao, A. Narishige, et al., "Hadoop framework: impact of data organization on performance", Software: Practice and Experience, vol. 43, no. 11, 2013, pp. 1241–1260.
31. N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "An application of pretrained deep neural networks to large vocabulary conversational speech recognition", *in Proc. Interspeech*, 2012.
32. K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified kneser-ney language model estimation.", *in ACL (2)*, 2013, pp. 690–696.
33. R. L¨ammel, "Googles mapreduce programming model revisited", *Science of computer programming*, vol. 70, no. 1, 2008, pp. 1–30.
34. D. K. Phull and G. B. Kumar, "Investigation of Indian English speech recognition using cmusphinx", *International Journal of Applied Engineering Research*, vol. 11, no. 6, 2016, pp. 4167–4174.
35. A. I. Rudnicky, "Language modeling with limited domain data", *proc. ARPA Spoken Language Technology Workshop*, vol 1, 1995.
36. C. Gormley and Z. Tong, Elasticsearch: The Definitive Guide. ", O'Reilly Media, Inc.", 2015.
37. S. Bird, "Nltk: the natural language toolkit", *in Proceedings of the COLING/ACL on Interactive presentation sessions Association for Computational Linguistics*, 2006, pp. 69–72.
38. F. M. B. Rodrıguez, E. D. Noya, P. G. Otero, M. L. Martınez, E. M. M. Mato, G. Rojo, M. P. S. del Rıo, and S. S. Docıo, "A corpus and lexical resources for multi-word terminology extraction in the field of economy in a minority language*", in Proc. of 3rd Language & Technology Conference*, 2007.
39. A. Stolcke and et.al, "Srilm-an extensible language modeling toolkit," *INTERSPEECH*, 2002.

40. S. Marquard, Improving searchability of automatically transcribed lectures through dynamic language modelling. PhD thesis, University of Cape Town, 2012.

AUTHORS PROFILE

Disha Kaur Phull holds bachelor's degree and master's degree in computer applications. She is currently pursuing her PhD from school of computing science and engineering from VIT University Chennai campus. Her research interest includes Speech Recognition and Natural Language Processing.

G. Bharadwaja Kumar holds a PhD degree in computer science and his research interest include machine learning, data analytics, Internet of things, speech and natural language processing. He is very passionate about developing resources and applications for Indian Languages in the areas of Natural Language Processing and Speech.