Evolving Web Corpus: Text Powered by Non Text

Zaheer Ahmad[1] Mohammad Abid Khan[2] Rahman Ali[3] Ibrar Ahmad[4] Mohammad Amir[5]

Department of Computer Science, University of Peshawar, Khyber Pakhtunkhwa, Pakistan

[1]ahmad.zaheer@yahoo.com,
[2]abid_khan1961@yahoo.com
[3]rahmanali.scholar@gmail.com
[4]toibrar@yahoo.com
[5]bitox@yahoo.com

Abstract

The World Wide Web has now become a vital turning point for different disciplines. Its publishing capability empowers people to easily contribute to knowledge in their field of interest. In addition, without their awareness, they are summing up text in one genre or another into the largest text and non-text corpus--the Web Corpus. The world's largest corpus in different languages in almost all genres is increasing in its size and nature almost on daily basis. It is thought that, with the proper usage of built-in multimodal (text and non-text) nature of Web Corpus, it will not only help the linguistic studies but will also be of great benefit to other fields of specialization. Researchers from different areas particularly from corpus linguistics, natural language processing, data and text mining and Image and knowledge retrieval are mostly attracted because of its enormous size and multimodality. This research paper provides a comprehensive review of the research carried out in the field of Web Corpus to examine its evolutionary process of Web Corpus as a multimodal and multidisciplinary tool of interest.

Keywords

Web Corpus, Corpora, Natural Language Processing (NLP), Multimodality, Linguistics.

1. Introduction

Billions of webpages have been created by people, unknowingly contributing texts in different languages to the World Wide Web which can be analyzed using NLP techniques, for its linguistic, social and cultural content. This exercise leads the web to be potentially the greatest resource e.g. for lexicographers, linguists, translators, teachers, corpus linguists and researchers in fields such as Natural Language Processing (NLP), Information Retrieval and Text Mining [1]. The Web offers extraordinary accessibility, quantity, variety of genres, languages and cost-effectiveness. Consequently a vast set of tools has been developed to exploit the web resources for different purposes in a number of fields, as well as the publication of numerous scholarly articles and research papers. Web Corpus requires pre-cleaning to convert web text to pure or cleaned text before it could be analyzed, by cleaning the web text we mean to make it useable for analyses. On Web Corpus, all the analysis tools and techniques can be applied with slight change or the way as they work on the classic text corpus. As web text has a vast variety of languages, genres and natures therefore new tools and techniques are emerging to categorize languages and genres for the construction or usage of Web Corpus [1, 9, 13]. Google claimed to cover 3.3 billion web pages in 2004 and has found over a trillion pages in 2008 [2, 6]. In 2005, one billion images (non-text net) were found embedded in web pages [3]. In [4] it is stated that 17% words on a web page are in image form and those words are usually the most semantically important text on the page. A large proportion of 76% of these words in image form does not appear elsewhere in the encoded text and therefore are processed with difficulty using traditional tools and techniques. Furthermore, the textual description or captions of the text images are incomplete, wrong or does not exist in 56% of the cases. Therefore reading non-text contents on a web page becomes essential to

understand the textual information. To make non-text part of the Web Corpus useable, some tools and techniques are developed to caption and annotate images, videos and audios. In the coming sections, detailed discussion and analysis of these tools and techniques will be provided. It will be examined that how new trends and research lead to the changing of the definition of corpus as a whole. Corpus Linguistics in itself is no more limited to text only as image and other non-textual data is required in some cases to fully understand the text. In some cases, portion of text cannot be understood if the relevant image is not read. Similarly to understand the image, text should be comprehended. Therefore text and non- text corpus approach is best suited for different fields of studies. In this research paper, different features of Web as Corpus will be discussed, that makes it a powerful tool. A comparison of classic Corpus with the Web Corpus will be carried out in section 2 in particular and on some other occasions in general in order to understand why Web Corpus is the ultimate answer for many questions. Section 3 replies to some criticism on Web Corpus, section 4 discusses the multimodality nature of Web Corpus, whereas, section 5 and 6 describe about some tools, research and future prospects of Web Corpus respectively. In the last section some conclusions are established from the study.

2. Advantages of Web Corpus over the Classic Corpus

Currently, a corpus is expected to bear some properties and features to be considered a solution in the field of linguistics and NLP, some of these features include its availability, size, development time, development labor, representativeness, efficiency and accuracy.

2.1 Multilingualism

Despite the fact that English was the dominant language of internet in the past and little or no presence of other major languages existed, now Web and eventually the Web Corpus is multilingual as other languages are striving hard to attain their place on the web as in November 2007, 45% of the webpages were written in English, 3.8% in Spanish, 4.41% in French, 2.66% in Italian, 1.39% in Portuguese, 0.28% in Romanian and 5.9% in German [5]. These statistics show a great mulilinguality on the Web Corpus.

2.2. A Complete Picture of Language

It is also no denying the fact that Web has the most up to date reflection of a language. Huge work is going on to digitize and publish linguistic archives and other documents having language information on the web. Weblogs, discussions and wikis are only available on the web. These make it possible to find all the new words or phrases of a language on the web. This makes the web as a complete resource for linguists and NLP researchers [1]. Alternatively, the classic corpus takes time to be updated beside the labor involved in the process.

2.3 Size of the Corpus

The Web universe is constantly expanding, so its size is unknowable [1]. It can be more or less considered as monitor corpus. In 2008 Google declared that it had identified (but not actually indexed) over a trillion URLs (Word Wide Web addresses), and that several billions new webpages appear daily [6]. Classic Corpora such as BNC is no match to this size of text corpus that was already getting short to help e.g. lexicographical research [7].

2.4 Representativeness

Today the web represents all the major languages and almost all genres of a language. There might be some genres more represented as compared to others but more represented genres can be controlled using the classification [9, 13] and word counting techniques.

2.5 Development Time and Labor

Development of Web Corpus is not time consuming and a large corpus can be developed in days without much labor that is usually involved in developing classic corpus. Development tools and techniques involved are simple and easily available. Mostly Search Engines (SEs) are used to facilitate the development cycle. Mainly two techniques are used, Key Words Approach using SEs and Web Crawlers for collecting text for Web Corpus [1,8] (Web for Corpus approach) besides using the Web as Corpus (using Web directly as Corpus)

2.5.1 Key Word Approach. In the Keywords approach, the only thing a corpus architect needs to start is a number of key words which the linguist considers relevant to the specialized domain for which a corpus is going to be built. The words chosen to start the process are called “seeds”[8] and are transformed by the system into a set of automated queries submitted to an ordinary search engine. The search engine then retrieves and downloads relevant pages, the architect post-processes them, and finally produces

a corpus from which a new word list is extracted containing new terms to be used as seeds to build a larger corpus through a cyclical process. Common Search Engines and Web Crawlers are used in this technique for the creation of "quick- and-dirty" monolingual and comparable corpora [8]

2.5.2 Manual Creation of Web Corpus. A web corpus can be created by selecting web pages manually or using a semi-automatic technique in which URLs or IPs of the pages are identified in the first phase then all those pages are downloaded using download managers or web archive tools. Some labor is involved in the approach but the text thus collected is more representative and as per the requirements of the researcher.

2.6 Convenience and Availability

The Web Corpus is available equally to the professionals and students [7] and that is also all the time. Web pages can be archived [10] by a third party for future availability, use and reference despite that their owners make it offline.

2.7 Geographic and Social Range

Languages on the Web now truly represent all the geographic locations in the world, from Europe to Africa and America to Asia no major language exist that is not present on the Web. Similarly, all the social genres of these major languages can be found on the web in all of its forms such as webpages, discussions and blogs.

2.8 Private Language

It was near to impossible to accumulate private language in a corpus prior to the birth of Web Corpus. On the Web, social sites, blogs, discussions and chat scripts provide private language in abundance. That is in enormous use in NLP, linguistics and psychology.

3. Criticism on Web Corpus

Some issues discussed in [7], though minor in their nature, were raised which were not in favor of the Web Corpus. But criticism on the Web Corpus is proved incorrect with the development of tools and techniques to analyze and exploit the Web Corpus. Some of these objections with their answers are provided in the subsections below:

3.1. Control

Unlike Web Corpus, offline corpus has control over genre and text collected [7]. As a solution, Web Crawlers can be used to collect categorized text. Besides, sites such as webarchieve [10] provide a well classified collection of web pages e.g. Asian Tsunami Web Archive, a collection of more than 1500 sites related to 2004 Tsunami in Asia. Similarly, Election 2002 Web Achieve on the same site provides 4,000 sites collection. It includes political party, government, advocacy, blogs, public opinion, and miscellaneous websites related to the 2002 United States election.

3.2. Accessibility

The objection that offline corpus is more accessible than the Web Corpus can be answered with the low cost of computers and bandwidth rates around the world, which increase accessibility to the web by many folds.

3.3. Level of Analysis

Offline corpus is more open to analyze than the Web Corpus: But Almost all sort of analyses which are possible on offline Corpus are now applicable on Web Corpus. Some examples of tools developed and research work are given in [11, 13, 16, 17].Some other objections are Copyright issues, poor non-text quality on the web, usage of foreign language phrases in some languages such as Pak–Indian sub continental languages. Currently all these are been catered for in different research papers.

4. Web Corpus a Phenomena of Multimodality

Web Corpus is a rich resource for linguists as linguistic analysis such as part-of-speech (PoS) tagging and concordancing require machine-readable written text which is available in volumes on the web. The Web Corpus also has enormous usage in other fields of studies such as natural language processing, data and text mining, image processing, knowledge engineering and Information Retrieval. As 17% of the web is composed of images [4], roughly 7% of these images contain text and usually the text image is the most semantically important part of the text on a webpage. Without understanding the image text, the main concept of the text is hard to be consumed. Similarly, video and audio represent a large part of online content. Using some robust tools [12, 13], the built-in

multimodality can be exploited as the main power of the Web Corpus which is earlier considered as its main disadvantage.

5. Tools and Research Evolving the Web Corpus

In this section, a review of the algorithms, techniques and tools is given to infer that the Web Corpus is providing more resources and ways to be exploited for different purposes in various fields than the classic Corpus. This part of the research is organized, starting from developing of Web Corpus tools and techniques, cleaning process, classification, analysis tools, techniques and uses of Web Corpus in different fields. Web crawlers and SEs are used to develop Web Corpus, Web Corpus cleaner systems and tools such as given in [14] are used in different ways to clean the text, remove unnecessary data and boiler- plates or extract the pure text from the web pages, so with the help of these tools, the challenging task to clean up acquired web pages can be achieved. Many new approaches to web page cleaning were encouraged by the CLEANEVAL 2007 contest organized by ACL Web as Corpus interest group. Researchers used heuristic rules as well as different machine learning methods, including Support Vector Machines, decision trees, genetic algorithms and language models. Although methods are fundamentally different, many of them employ similar set language-independent features such as average length of a sentence or ratio of capitalized words in a page segment. Script and genre classification can be carried out before collecting web pages or after purifying the text. Research Studies [15] for automatic script, genre and clusters identification, categorization/classification and duplicate pages [9] removal can be carried out using Neural Networks, Support Vector Machines and other machine learning techniques to develop Web Corpus with no anomalies. Even semi-automated system can be used to classify a single author work as discussed in the paper [13]. Tools such as mentioned in [16, 17] are there to annotate web corpus beside WebCorp[24] to get and analyze linguistic data that is information extraction, retrieval and management from the net. With the help of this tool neologisms and coinages, newly-voguish terms; rare or possibly obsolete terms; rare or possibly obsolete constructions; and phrasal variability can be analyzed. A study [17] for web based term translation proposes domain specific bilingual lexicon. It mines bilingual search-result pages obtained through a search engine using unknown query terms. For facilitation and usage of non-text data such as images, audios and videos to exploit web as multimodal corpus a number of tools and techniques are developed or they are in the process of research and development for different languages. Out of these developments one hot area of research is concept-based image categorization and image search on the Web as discussed in [18]. Optical Character Recognition (OCR) Systems for Latin image-text reading are already in the market but OCR systems for cursive languages such as Urdu and Arabic, are also under research [25, 26] and development. You Tube Video transcription [19] and automatic captioning tools for videos on You Tube [20] have been already developed. Similarly, web based automatic lecture transcription work [21, 22] is under research. Work on Arabic transcription [23] is an example on the research on languages which are not much digitalized. Transcription tools (for example Transcribe! for Mac OS-X 8.00) [23] and many more are freely or with very low price available in the industry. Some of the above tools and techniques might not be developed for Web specifically but the aim of mentioning it here is to describe that these tools can greatly affect the Web Corpus usage in the near future.

6. Current Trends and Future Prospects of Web Corpus

Currently, researchers from different fields such as search engines and text mining are regularly publishing papers which enormously benefit corpus linguists and NLP professionals. Work is going on to annotate images, to transcript audios and videos and assign them captions automatically for their intelligent retrieval and indexing purposes. Concordance is the field which attracted most of the classic corpus linguists towards the web corpus because web is richer in its contents to advance the research to a new level. Organization e.g. WaCky and WaCUK are striving hard to collect and develop the largest Web Corpus which was unimaginable before the Web. Automatic language, text and genre classification provide a base to collect text in a specific domain. OCRs are currently the hot area of research which can be used to convert image text from the web into readable text to develop or use web as corpus.

6.1 Urdu Web and Future of Urdu Web Corpus

Urdu webpages appeared on the internet in 1994 [27]. Initially Urdu script was not supported in the ASCII character set, therefore, being one of the largest languages of the world Urdu has nominal representation than it was supposed to enjoy on the web. Urdu web pages was developed in the form of

images or in the form of roman script and still these make a good share of Urdu script on the web due to the lack of Urdu support in ASCII character set and Urdu standardization[27], besides, unavailability of robust Urdu text editors in the past. This, on the first instance not only made it difficult for simple SE to search Urdu text, secondly it made it impossible to develop a viable Urdu web corpus using the already available tools and techniques designed for the development of English web corpus. Currently, with the provision of Urdu support in the form of Unicode and availability of effective Urdu text editors, web developers have started dynamic web pages in Urdu. Now, web provides a rich set of Urdu genres to be exploited for the development of Web Corpus. Some of the main Urdu text genres available on the web are news, religious scripts, literature and poetry, fashion and cooking, medical text, forums and blogs besides Urdu audio and videos in these genres [28]. By now Google is already providing search facilities for Urdu, therefore, with the enhancement of Urdu Optical Character Recognitions (OCRs) systems such as [25, 26] and Urdu text categorization techniques, a viable Urdu web corpus can be dreamed of.

7. Conclusions

Web Corpus has made many fairy tales of NLP and linguistics true and still it is evolving. The Web Corpus is converting the notion of a corpus into multimedia/multi modal Corpus that will indeed help and facilitate to study languages and machine translation as the images, audio and video in the Web Corpus will further enhance the capabilities of researchers and linguists to study languages. However, it will not be just used for linguistics studies. In the near future, Web Corpus will be an important tool for NLP research, Image, Video and Audio, transcriptions, knowledge, image and Audio-Video retrieval, classification, genre identification, and speech recognition.

References

[1]. W.H. Fletcher, "*Corpus Analysis of the World Wide Web*", Date Retrieved (06/14/2010), Available: webascorpus.org /Corpus_Analysis_of_the_World_Wide_Web .pdf

[2]. *Size of the web*, Date Retrieved (14-06-2010), Available: http://www.caslon. com.au/metricsguide2.htm#pages

[3]. *The non-text net*, , retrieved on (14-06-2010), Available : http://www.caslon. com.au/metricsguide2.htm

[4]. S J Perantonis B Gatos and V Maragos, "*A novel Web image processing algorithm for text area identification that helps Commercial OCR engines to improve their Web image recognition efficiency*", in proc, of Second International Workshop on Web Document Analysis, United Kingdom, Edinburgh, pp. 61-64,2003.

[5]. *Global Internet Usage*, Date Retrieved (14,6,2010), available: http://en.wikipedia. org/wiki/Global_Internet_usage

[6]. *World Wide Web*, Date Retrieved (14, 05, 2010), Available: http://en.wikip edia. org/wiki/World_Wide_Web.

[7]. Marianne Hundt, Nadja Nesselhauf, Carolin Biewer, *,"Corpus linguistics and the web"*, RODOPI, New York, USA, 2004.

[8]. M.Gatto,"*From language to culture and beyond: building and exploring comparable web corpora",* Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, pages 72–78 Malta, 22 May 2010.

[9]. J.Gibson,B.Wellner,S.Lubar,"*Identification of Duplicate News Stories in Web Pages*" , in proc. of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google?, Marrakech, Morocco, 2008.

[10]. *Asian Tsunami Web Archive/ Election 2002*, Date Retrieved (07,11,2010), available:http://www.archive.org/web/web.ph

[11]. P.Rayson, J. Walkerdine ,"*Annotated Web as Corpus*", in proc. ACL Workshops archive Proceedings of the 2nd International Workshop on Web as Corpus, Association for Computational Linguistics Morristown, NJ, USA,2006

[12]. S.Evert, "*A lightweight and efficient tool for cleaning Web pages*", in proc. of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco,2008

[13]. R. Guzmán-Cabrera, M.Montes-y-Gómez "*A Web-Based Self-training Approach for Authorship Attribution*",, Book Advances in Natural Language Processing, Springer Berlin / Heidelberg, 2008.

[14]. M.Spousta, M. Marek, Pavel ,"*Victor: the Web-Page Cleaning Tool*", in proc. Of the 4th Web as Corpus Workshop Web as Corpus Workshop (WAC-4) Can we beat Google?, Marrakech, Morocco,2008.

[15]. B.H.ChandraShekar, Dr.G.Shoba, "*Classification Of Documents Using Kohonen's , Self-Organizing Map and*

Classifying Webcorpora into domain and genre", in proc. International Journal of Computer Theory and Engineering, Vol. 1, No. 5, 2009.

[16]. A.Kornai "*Google for the Linguist on a Budget*", in proc. Of the 4th Web as Corpus Workshop Web as Corpus Workshop (WAC-4) Can we beat Google?, Marrakech, Morocco,2008.

[17]. Jenq-Haur Wang1, Jei-Wen Teng1 ,"*Exploiting the Web as the Multilingual Corpus , for Unknown Query Translation*", Journal of the American Society for Information Science and Technology, John Wiley & Sons, Inc. New York, NY, USA,2006.

[18]. K.Yanai , "*Finding Visual Concepts by Web Image Mining*" , in proc. of the 15th international conference on World Wide Web, Edinburgh, Scotland 2006

[19]. *YouTube Audio Transcription*, Date Retrieved (07,10,2010) available: http://googlesystem.blogspot.com/2009/11/youtube-audio-transcription.html

[20]. *Automatic Caption in You Tube*, Date Retrieved (07,10,2010) available: http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html

[21]. C.Munteanu1 , G.Penn1,"*Web-Based Language Modelling for Automatic Lecture Transcription*", in proc. In Proceedings of the Tenth European Conference on Speech Communication and Technology - EuroSpeech / Eighth INTERSPEECH, Antwerp, Belgium, August 2007.

[22]. Chu, S.M. Hong-kwang Kuo Yi Y Liu Yong Qin Qin Shi Zweig, G, "*The IBM Mandarin Broadcast Speech Transcription System*", in proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.

[23]. Soltau, H., Saon, G., Povey, D., Mangu, L., Kingsbury, B., KJ., Oma, M., and Zweig, G., "*The IBM 2006 GALE ArabASR system*", in Proc. ICASSP, vol. 4, pp. 349–352, 2007.

[24]. A. Renouf, A. Kehoe, and J. Banerjee. 2007. WebCorp: an integrated system for web text search. Corpus Linguistics and the Web.

[25]. Z.Ahmad, J. K. Orakzai, "*Urdu compound Character Recognition using feed forward neural networks*" , in proc. 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, China, 2009.

[26]. Z.Ahmad, J. K. Orakzai, "*Urdu Nastaleeq OCR (Optical Character Recognition*", Proceedings of World Academy of Science, Engineering and Technology, Bangkok, Thailand, 2007.

[27]. *Urdu Encoding and Collation Sequence for Localization*, Date Retrieved (14, 09, 2010), Available http://www.panl10n.net/english/outputs/Pakistan/Urdu-Encoding-Collation.pdf

[28]. *An overview of Urdu on the Web*, Date Retrieved (14, 09, 2010), Available http://www.urdustudies.com/pdf/20/25ResourcesHoda.pdf