Considerable Issues to detect Topical-Relevance Upon Free Texts

Arpana Rawal, M K Kowar, Sanjay Sharma
Bhilai Institute of Technology
Durg, Chhattisgarh, India
arpana_rawal@rediffmail.com

H R Sharma
Chhatrapati Shivaji Institute of Technology
Durg, Chhattisgarh, India
hrsharma44@gmail.com

***Abstract*—**Topical Relevance has always been an equivocal concept in IR evaluations upon text documents. Text Miners have been well engaged in unfolding various strategies in finding, how well the core content is expressed in a document and till what extent, the correlated topic gives out the semantic closeness exhibiting subject specificity. These directly and indirectly searched correlated term-associations initiated from the target terms, terminate in finite page vicinities of topical relevance. With this idea into the minds, the page-filtering technique emerges out as a logical approach to outline the degree of content coverage. These results can be further compared with more than one test documents to rank them in order of topical relevance.

Keywords- Topical relevance; Paragraph tagging; Target page; Search vicinity; Semantic page-filtering

I. INTRODUCTION

Most of the challenging task domains in the realm of Information Retrieval is still in on-going research viz. categorization, relevance finding, relevance ranking and summarization tasks. These task domains are inherently dependent of Natural Language Processing (NLP) techniques. In turn, these techniques are found accompanied with ambiguities due to complex grammatical rules of naturally spoken languages. In this direction, many computational linguists have been working upon various approaches to text understanding, considering the sentence syntax and semantics. Many Languages like Traditional Latin, Arabic and Sanskrit have been successfully using dependency grammars over decades for explaining word-to-word agreement, case assignment or any semantic relation in the sentential fragments to form a concept. This has motivated a lot of NLP researchers who in turn, proceeded with systematic study of semantically motivated information such as deep dependency relations or predicate-argument structures.

II. TEXT REPRESENTATION PRELIMINARIES

A. POS tagging the text

It all begins with reading the line streams of character inputs from the free text, that gets recognized by efficient tokenizers, that work on the concept of lexical frequencies and lexical probabilities. But the real annotation of text fragments wholly lies on the underlying very large annotated corpora that have been enterprisingly constructed for American English, here a very frequently used Pen-Tree Bank corpus is taken up, consisting of 4.5 million words (growing till date) and is annotated for obtaining POS-tagged information [1]. This is necessary because according to grammarians, most of the words dominating the subject domain exist as Proper Nouns, Common Nouns preceded with modifiers either qualitative, quantitative or adverbial type. They can be event-describing verbs in different moods, tenses and adverbial forms. Hence, these fragments need to be identified for their lexical sequences and syntactic categories [2] [3].

B. Exploring Text Syntax and Semantics

The next phase of understanding the inputted text includes drafting a set of rules ranging from context-sensitive grammars to context-free grammars, followed by proving the correctness of sentential constructions through a variety of shallow parsers viz. simple-chunk parsers [4], shallow statistical parser and a wide range of treebank based NL dependency parsers that learnt grammar rules using machine-learning techniques [5]. Kalpan, etal. gives a comparative statement to the usage of shallow and deep parsers in the sense, that shallow parsed strings may leave partial bracketings to the parsed strings as parse tree constructions, without giving them meaningful representations. While deep grammars offer predicate-argument structure to the input strings, the parsed outputs still are more normalized representations in the form of dependency relations [6][7]. Hereafter the current work proposes to use dependency relations as developed by NLP community at Stanford University in extracting the generating streams of term-phrases that constitute the semantic concept space for the highlighted topics of any taken-up domain [8].

C. Identifying text-relevance parameters

The current proposal slightly deviates from the opinion of providing free text in order to parse its POS-tagged equivalent into dependency units, with the realization that locality of topical relevance too needs to get identified at the end. This can be computed with help of page-numbers as a topic-accession parameter. Let this process of deriving topical-vicinity be preceded by identifying the paragraph / subsection / section and chapter boundaries. In this way, each paragraph can be assigned unique identification tag, comprising of chapter, section or subsection number, followed by paragraph count on the relevant page-number, as shown in the paragraph tagging-format below, figure1.

The above concept removed the potential difficulties faced while normalizing document lengths among long and short documents, as too long paragraphs were found written across few page lengths [9] [10]. Now the bridging paragraphs too could be identified with unique identification technique as shown in the extended tagging format, figure 2.

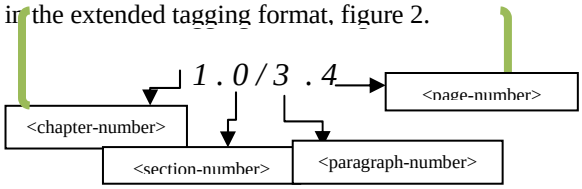

Figure 1. Paragraph-tag for 4th paragraph on page 4

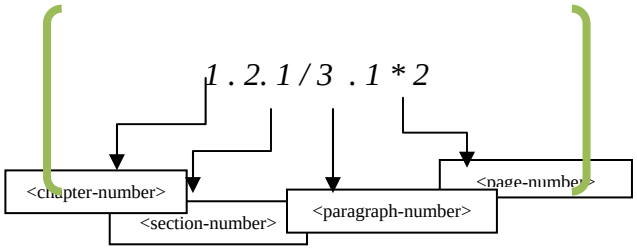

Figure 2. Paragraph-tag for bridging paragraph between pages 1,2

1.0 / 1.6

In addition, as our example network illustrates, an ANS is robust in the sense that it will respond with an output even when presented with inputs that it has never seen before, such as patterns containing noise. If the input noise has not obliterated the image of the character, the network will produce a good guess using those portions of the image that were not obscured and the information that it has stored about how the characters are supposed to look. The inherent ability to deal with noisy or obscured patterns is a significant advantage of an ANS approach over a traditional algorithmic solution. It also illustrates a neural-network maxim: The power of an ANS approach lies not necessarily in the elegance of the particular solution, but rather in the generality of the network to find its own solution to particular problems, given only examples of the desired behavior.

1.0 / 2.6

Once our network is trained adequately, we can show it images of numerals written by people whose writing was not used to train the network. If the training has been adequate, the information propagating through the network will result in a single element at the output having a binary 1 value, and that unit will be the one that corresponds to the numeral that was written. Figure 1.3 illustrates characters that the trained network can recognize, as well as several it cannot.

1.0 / 3.6

In the previous discussion, we alluded to two different types of network operation: training mode and production mode. The distinct nature of these two modes of operation is another useful feature of ANS technology. If we note that the process of training the network is simply a means of encoding information about the problem to be solved, and that the network spends most of its productive time being exercised after the training has completed, we will have uncovered a means of allowing automated systems to evolve without explicit reprogramming.

1.0 / 4.6

As an example of how we might benefit from this separation, consider a system that utilizes a software simulation of a neural network as part of its programming. In this case, the network would be modeled in the host computer system as a set of data structures that represents the current state of the network. The process of training the network is simply a matter of altering the connection weights systematically to encode the desired input-output relationships. If we code the network simulator such that the data structures used by the network are allocated dynamically, and are initialized by reading of connection-weight data from a disk file, we can also create a network simulator with a similar structure in another, off-line computer system. When the on-line system must change to satisfy new operational requirements, we can develop the new connection weights off-line by training the network simulator in the remote system. Later, we can update the operational system by simply changing the connection-weight initialization file from the previous version to the new version produced by the off-line system.

Figure 3. Tagged Paragraphs 1-4 of section 1.0 on page-number 6

processing element-44,input space*4,information flow*4,example inputs*5,ANS approach*6,ANS approach*5,ANS approach*6,character set*4,processing unit*4,training mode*6,production mode*6,network structure*4,ANS technology*6,network operation*6,output units*4,ANS models*7,character input*4,output layer*4,output units*4,output structure*4,input noise*6,building blocks*7,network structure*4,data structures*6,disk file*6,row vectors*5,row vector*5,network simulator*6,output units*4,strategy simplifies*4,network simulator*6,computer system*6,initialization file*6,network simulator*6,input pattern*5,software simulation*6,data structures*6,

Figure 4. Noun-phrase chain for syllabus string, ‘processing element’

III. Determining Semantic Vicinities

The ongoing research orients itself upon a live case study to find the most promising teaching content for a provided syllabus fragment of technical domain of ‘Neural Networks’. For this, an exemplary text material is selected as a book on the above domain, entitled “Neural Networks : Algorithms, Applications and Programming techniques” authored by James Freeman and David M. Strapetus. In the beginning, the page numbers that are detected to embed these search keywords, serve as a set of target pages, pointed to by page-links of metadata list Domain Vocabulary, behaving as the underlying Ontology [11].

A. The Target Pages

How ever, it would not be inappropriate to mention that the targeted topics may or may not lie embedded as provided in inputted patterns in the above domain of search. For instance, ‘neocognitron character recognition’ is described in chapter 1 of the book, but narrated in a sentential fragment like, ‘recognition of hand-drawn alphanumeric characters’. Thus, assuming the paragraphs are undergone a step of preprocessing with paragraph-tagging, if the page-numbers are not traced according to the above criteria, the search gets deviated to find search strings within paragraph-tagged content of all the chapters. Also, it is always better to customize the searches with desired word-permutations of search key phrases. In this way, if a 4-gram search for upon the key-phrase as ‘neocognitron hand-written digit recognition’ does not give any

***(proposed) **(computed)**

results, the end-user can trace the target pages with other permuted combination of sub-string patterns, like 'hand-written recognition', 'hand-written digits', 'digit recognition', 'neocognitron recognition', etc. These are called as hit-word patterns in the proposed logic [12]. This helped in finding coherence of precise topical relevance densities to the most useful page limits.

B. The Search Vicinities

Now, there arises an issue, as to what extent, the term-co-occurrences should be searched for the search key-phrase, beginning from the set of target pages. The search vicinity can be considered to be all the paragraphs lying between the target pages till the end of the relevant sections or sub-sections. The idea behind this was taken up with the usual observations that any topic is discussed by any author till the section ends in its extreme length.

C. The triggered search

Speaking in general, the main parameter for exploring the semantics of any NL sentence goes round about the most significant portions namely noun phrases in Subject and Object roles. This affirms the authors to trigger an exhaustive search in the drafted search vicinities, as computed in the previous section. This time, grammatical relations extracted in form of Stanford typed dependencies are chosen as the search content. The salient feature of Stanford typed dependencies is that there is a distinctly *defined* relation for every word lexicon of the same or adjacently meaningfully related sentences.

D. Computing the semantic vicinities

Thus, there crops up a strategy to find term-to-term co-occurrences from typed dependency equivalents of the

Unit No.	Syllabus Strings (Seed words)	Hit-word patterns	Target search pages*	Target searchp ages**	Term-search vicinity	Semanti cally filtered page range
1	Processing Element	Processing element, Element	4	4	4-7	4-7
2	Neocognitr on character recognition	Neocognitro n character recognition, character recognition, recognition	5	2,3,7	5-7	5-7
2	Neocognitr on handwritte n digital recognition	Neocognitro n handwritten digital recognition,	7	2,3,7	7-7	6-7
3	Neural Network Models	Neural Network Models, Network Models, Models	3	3	3-7	3-7

identified or paragraph-tagged text.

TABLE I. PAGE-FILTERED VICINITIES FOR SEED WORDS

As stated above, the search is initiated to find only Noun-Phrase patterns co-occurring directly or indirectly with hit-patterns, with the underlying fact that these components may collectively formulate the subject matter for the queried topic in domain of context. The search algorithm also takes care of aligning components of multi-word noun phrases together, so that the noun-phrases can take up n-gram representations. Figure 4 extracts such a chain of noun-phrases for the search key-phrase, 'processing element' from pages 4 to 7. A few selected noun-phrases that get logically extracted in the chain from a set of exemplary paragraphs '1.0/1.6, 1.0/2.6, 1.0/3.6 and 1.0/4.6' are illustrated in figure 3 from the book domain of case study.

IV. RESULTS AND DISCUSSIONS

The proposed page-filtering algorithm put forth in one of the previous works by A.Rawal, etal. was manually traced upon the set of thirteen syllabus strings against their respective target pages and search page ranges [12]. In order to mechanize the same, the series of on-going experiments have already explored and preprocessed the free text in chapter 1 of book domain. So, the current scenario was able to focus on few of the syllabus strings that are found matched in target pages of this chapter. Table 1 summarizes such results obtained by searching hit-word patterns related to inputted syllabus strings that themselves are represented as seed patterns in column 2. For instance, search upon trigram 'Neocognitron character recognition' also initiated searches for bigram search upon 'character recognition' and unigram search upon 'recognition'. As traced by domain experts, the target page for the hit-pattern, 'character recognition' was found on page 5 that lay as a part of text in figure caption of that page. But, when all the paragraphs of the chapter were tagged, the figure captions were removed and a bridging paragraph between pages 3 and 4 got shifted as a last paragraph on page 3, which contained one of the hit-patterns, 'recognition'. These findings resulted in a shift in target pages from page 5 to pages 2, 3 and 7, as tabulated in column 5. The enumerations upon page-number tags augmented to each co-occurring noun-phrase in the extracted noun-phrase chains were traced as a set of page-numbers that indicated the semantically filtered page-vicinities as shown in column 7 of the experimental results.

V. CONCLUSION

The current approach leaves a question behind, as why not to consider free-text like figure caption fragments also in the text corpus of the domain. This may increase the chances of successfully searching the hit patterns in total, without going for subsequent searches of their sub-pattern combinations. Taking another scenario, if there would have been few more target pages, locating the sub-pattern, 'recognition', then the designed system is destined to compute co-occurring terms in all the contexts identified, out of which finding page vicinities of correct context coverage may increase further run-time complexities. Henceforth, logic behind page-filtering process needs thorough testing phases that lie as further scope in the research work.

ACKNOWLEDGMENT

The current work is planned to be reported as a partial fulfillment of the semester-wise progress of the on-going research. The authors would like to express deep gratitude to the Research & Development Cell of Bhilai Institute of Technology for providing necessary infrastructure towards the project. Ani Thomas and Sarang Pitale deserve special thanks for the implementation of the designed algorithm. Finally, the authors would like to thank Sujata Kataria for providing the fully downloaded version of Stanford parser and tagger in time for the necessary implementation.

REFERENCES

[1] Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz "Building a large annotated corpus of English: the Penn Treebank." *Computational Linguistics*, Vol.19, 1993. Reprinted in *Using Large Corpora*, S. Armstrong (ed.), MIT Press, 1994.

[2] K. Toutanova, and C.D. Manning, "Enriching the Knowledge Sources Used in Maximum Entropy Part-Of-Speech Tagger", Department of Computer Science, Stanford University, California, USA, Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Hong Kong.

[3] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer, "Feature Rich Part-Of-Speech Tagging with a Cyclic Dependency Network", Department of Computer Science, Stanford University, California, USA, *In Proceedings of HLT-NAACL 2003* pages 252-259..

[4] Philip Brooks, "SCP: Simple Chunk Parser", Artificial Intelligence Center, University of Georgia, Athens, Georgia, USA, May 2003, www.ai.uga.edu/mc/ProNTo/Brooks.pdf.

[5] Marie-Catherine de Marne_e and Christopher D. Manning, "Stanford typed dependencies manual", September 2008.

[6] Marie-Catherine de Marne_e, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. in 5th International Conference on Language Resources and Evaluation (LREC 2006), 2006.

[7] G. Attardi, and M. Ciaramita, "Tree Revision Learning for Dependency Parsing", Proceedings of NAACL HLT 2007, Association of Computational Linguistics, Rochester, NY, pp. 388-395, April 2007.

[8] A. Cahill, R. O' Donovan, J.V. Genabith, M. Burke, S. Riezler and Andy Way, "Wide Coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation", Dublin City University, IBM Center for Advanced Studies, © Association of Computational Linguistics, pp-Vol. 34, No. 1, pages 81-124, 2008.

[9] Jaana Kekäläinen. Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. International Journal of Information Processing & Management, 41, pp 1019-1033, 2005.

[10] Luis Tari, Phan Huy Tu, Barry Lumpkin, Robert Leaman, Graciela Gonzalez and Chitta Baral. Passage relevancy through semantic similarity, Proceedings of the Text Retrieval Conference (TREC-2007) Genomics track, NIST, 2007.

[11] M.K. Kowar, S. Sharma, A. Thomas, and A. Rawal "Learning Ontologies and Semantic Concept Spaces for Automatic Document Relevance Ranking", Proceedings of 2nd International Conference on Resource Utilization and Intelligent Systems, INCRUIS 2008, INDIA, pp. 591-595, 2008.

[12] A. Rawal, M.K. Kowar, H.R.. Sharma and S.Sharma "Role of Exact Page Accessions for Determining Text Relevance Measures" – Proceedings of Third International Conference on Information Processing, Visveswaraya College of Engineering, Bangalore University, Bangalore, INDIA, pp. 26-34, August 2009.

AUTHORS PROFILE

Mrs. Arpana Rawal is Associate Professor in Department of Computer Science & Engineering at Bhilai Institute of Technology, Durg (C.G.) , India. She obtained her Bachelor of ENgineering from Nagpur University and Masters degree from NIT,Raipur. She is heading the affairs of Department of Information Technology, Bhilai Institute of Technology,Durg since 2006.Her research mainly focuses on text mining techniques for machine learning on text relevance and ranking as application domains. She has five research papers to her credit, in the mentioned area of research, been published at various peer reviewed journals. She has got some of her technical papers published in eight National level, four International level and one State level conference platforms too.

Prof. H. R. Sharma is a Professor in Department of Computer Science & Engineering and Principal at Chhattrapati Shivaji Institute of Technology, Durg, Chhattisgarh, India. He is a postgraduate in Applied Mathematics from Jabalpur University in 1966, completed his another post graduation course in Computer Science from Delhi University in 1991. Further, his research proceeded in the field of Mathematics in the area of Numerical Analysis & high Speed Computation" that was awarded to him from IIT Delhi in the year 1970. Having a total teaching experience for more than 38 years, he has supervised 03 Ph.D. scholars at Maharishi Dayanand University, Rohtak, Haryana , and 02 Ph.D. scholars at Pt. Ravishankar Shukla University, Raipur, Chhattisgarh in the discipline of Engineering and Technology (Computer Sciences). Professor H R Sharma is an Editor to the reputed International Journals of Computer Science and Information Technology, Electronics and Electrical Engineering and also Emerging Journal of Engineering Science and Technology.

Dr. Manoj Kumar Kowar received his B.E., M.Tech. and Ph.D. degree from University of Calcutta. He started his teaching career from Birla Institute of Technology, Mesra, Ranchi and is presently the Principal in Bhilai Institute of Technology at Durg (C.G.) India. He has total teaching experience of 23 years. Dr. Kowar's primary research interests focus on Modelling & Simulation of Fabrication Processes, Bio-Medical instrumentation, Fuzzy logic in Medical Expert systems and Secure Communications. He has a total of 129 Research papers published in refereed International Journal and Conference Proceedings.

Dr. Sanjay Sharma is a Professor in the Department of Mathematics in Bhilai Institute of Technology at Durg (C.G.) , India. He received his Ph.D. degree from Pt. Ravi Shankar Shukla University Raipur, Chhattisgarh. Dr. Sharma's primary research interests focus on analyzing and studying the effects of Solar Radiation Pressure and other nature's forces on Interconnected Satellite system orbits. During the nineteen Years of Teaching Experience, he has published many technical papers in reputed Journals like NASA, Bulletin of the Astronomical Society of India, Indian Journal of Pure and Applied Mathematics, Bulletin of the Calcutta Mathematical Society , The Mathematics Education : Vol. XXXVI etc.