Evaluation of Causal Sentences in Automated Summaries

C. Puente
Advanced Technical Faculty of Engineering ICAI
Pontifical Comillas University
Madrid, Spain
cristina.puente@icai.comillas.edu

A. Villa-Monte and L. Lanzarini
III-LIDI, Faculty of Computer Science
National University of La Plata
La Plata, Argentina
{avillamonte, laural}@lidi.info.unlp.edu.ar

A. Sobrino
Faculty of Philosophy
University of Santiago de Compostela
La Coruña, Spain
alejandro.sobrino@usc.es

J. A. Olivas
Department of Information Technologies and Systems
University of Castilla-La Mancha
Ciudad Real, Spain
joseangel.olivas@uclm.es

***Abstract*—This paper presents an experiment to show the importance of causal sentences in summaries. Presumably, causal sentences hold relevant information and thus summaries should contain them. We perform an experiment to refute or validate this hypothesis. We have selected 28 medical documents to extract and analyze causal and conditional sentences from medical texts. Once retrieved, classic metrics are used to determine the relevance of the causal content among all the sentences in the document and, so, to evaluate if they are important enough to make a better summary. Finally, a comparison table to explore the results is showed and some conclusions are outlined.**

Keywords—Causality; causal sentences; automatic summaries; sentence scoring metrics; Soft Computing.

I. INTRODUCTION

Since the invention of writing, human beings have stored knowledge in texts. The invention of printing in the antique and, recently, of the electronic publishing devices, have caused that the number of texts at our disposal has increased enormously. The up to 2011 Google executive director, Eric Schmidt, said that that the Web stored 5 million terabytes of data (2014). And the forecast is that the growth of information is unstoppable.

Humans interact with what surrounds through senses, coordinated by the brain. In a computer metaphor, the brain processes the information using the stimulus of the senses and memory registers, primarily located at the hippocampus. Although there is disagreement about the memory capacity of the brain [1], we recall many things because we forget at the same time many others. Most of the time, to remember what matters means to isolate the grain from the straw and so, grasp the essential information to keep it in our memory. In the case of written texts, this operation is known as to summarize.

Summarizing is a cognitive characteristic of human intelligence to retain the essentials. Forgetting is bad, but it would be worst to remember everything that we read, because in many cases the brain would collapse. To summarize is to grasp the fundamental for our purposes, to make clear the information that seems relevant to us in order to complete our knowledge and, therefore, worthy of being remembered. Separating relevant information is sometimes a more or less objective process, but in other cases is a context and individual dependent one. Many times a text -especially if it is not a scientific text-, admits different views and thus, alternative ways to discriminate its essentials.

In academic or scientific texts there is a consensus about the role of causal content to establish a mark of relevance. Causal sentences show a link between knowledge that is solidly rooted in agent causes and therefore expresses well-founded intuitions about the world or about ourselves [2]. Therefore, it is reasonable to presume that separating causal sentences in a text should provide some of its essential pieces of information and, so, to contribute to make a good summary of it.

Causality has been traditionally linked to physical laws and Physics –but quantum mechanics- advocates for the use of classical logic showing the coherence of its thesis. For a long time, causality was normalized from Physicists as a crisp relation mimicking natural connections. But science is strongly linked to writing and written texts permit to verify if people, including scientist, express causal judgments in a crisp way or rather using vague language. Puente et al. [3] mined causal sentences in texts from several sources using a semi-automatic procedure and showed that, contrary to the common image of precision, causal sentences, -even from the field of Physics-, used a lot of vague vocabulary, as fuzzy quantifiers, linguistic modifiers or vague predicates.

There are classical definitions slightly different from what is summarization, depending if the focus is placed on the size of the text, its content or both. Thus, according to Hovy, [4], a

978-1-5090-6034-4/17/$31.00

summary is a text produced from one or more texts containing a significant portion of the information of the original text, and no longer than half of it. Following Mani et al., [5], to summarize is the process of distilling the most important information from a text source given a particular user. Those definitions pointed to two different methods getting a summarized text: extractive or abstractive. An extractive summary is achieved choosing appropriate statements from the source text, and later sticking them in a comprehensive message. An abstractive summary results from grasping the main idea or ideas from the text, which will be expressed without using sentences of the source text. This first method is perhaps a first step challenging the second one, closer to what is expected about a quality human summary.

There is a lot of work on extractive summaries and also many articles about the extraction of causal sentences in texts like the one presented by Kaplan and Berry-Bogge [6]. They approached a knowledge-based inference system to detect causal knowledge in scientific texts using linguistic templates to match causal relations. The main problem that they had with this approach was the scalability in large applications. Rink et al. in [7] dealt with a method for detecting causal relations between events related in a text. The method was able to find if two events from the same sentence present a causal relation by building a graph representation of the sentence, automatically extracting graph patterns from that graph representation and training a binary classifier that decides if an event is causal or not based on the extracted graph patterns.

Many papers on automatic summaries include causal techniques as hooks for extracting sentences with relevant content. We refer to mining relevant sentences by detecting causative verbs [8], causal links [9] or if-then conditionals [3]. Connecting causality and summarization, Endres-Niggemeyer [10] suggests that if events belong to a causal chain, the procedure to read and order the sequence from the beginning to the end of the chain will produce a good quality summary. Particular events or isolated ones are more difficult to connect, as they would be meaningless, or have to be set up into a context; on the other hand, if these events are ordered in a causal chain, the context is already given, and the quality of the resultant summary will be higher. But so far there have been no studies evaluating the extent to which mining causal sentences help to improve an extractive summary. This paper seeks to shed light on this subject and, to that end, is structured as follows: In point 2 we will describe a process to extract and classify causal and conditional sentences from text to create a causal knowledge base. In point 3 we will describe the metrics that we have used to measure the relevance of the sentences obtained in the whole document so to evaluate the quality of them. In point 4, we will describe an experiment with 28 documents to check how good are causal sentences to form summaries. In point 5 we will discuss the obtained results which will lead us to conclusions and future works.

II. EXTRACTION AND ANALYSIS OF CAUSAL SENTENCES

Taking the presented works of the introduction into account, in [3], we presented an algorithm to extract, classify and represent causal and conditional sentences though a causal graph. The first stage of this algorithm was to select and classify causal sentences from text documents. So that we used the morphological analyzer Flex plus C code to create a program able to detect 20 syntactical patterns frequently used in the English language to express causality, as seen in Fig. 1.

Fig. 1. Patterns selected to be extracted in a document.

Structure 1: if + present simple + future simple.
Structure 2 : if + present simple + may/might.
Structure 3 : if + present simple + must/should.
Structure 4 : if + past simple + would + infinitive.
Structure 5 : if + past simple + might/could.
Structure 6 : if + past continuous +would + infinitive.
Structure 7 : if + past perfect +would + infinitive.
Structure 8 : if + past perfect + would have+ past participle.
Structure 9 : if + past perfect + might/could have + past participle.
Structure 10 : if + past perfect + perfect conditional continuous.
Structure 11 : if + past perfect continuous + perfect conditional
Structure 12 : if + past perfect + would + be + gerund
Structure 13 : for this reason, as a result.
Structure 14 : due to, owing to.
Structure 15 : provided that.
Structure 16 : have something to do, a lot to do.
Structure 17 : so that, in order that.
Structure 18 : although.
Structure 19 : in case that.
Structure 20 : on condition that, supposing that.

We performed several experiments with text belonging to different scopes like legal texts, scientific texts, news, gospel, etc., obtaining better and more accurate results in scientific and medical texts. To check it, we performed a Gold standard test analysing 50 pages of texts from different areas such as news or medicine and the following results were obtained:

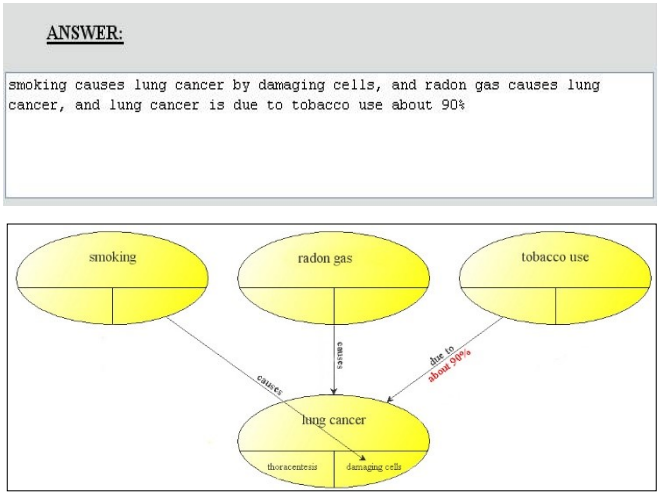
TABLE I. GOLD STANDARD TEST WITH DIFFERENT TEXT GENRES

Type of Text	Detected	Classified	Classified (Manual)	Recall	Precisión	F-Measure
Scientific	62	52	80	0.65	0.839	0.73239
Medical	11	10	13	0.7692	0.909	0.83333
Best Sellers	22	12	37	0.3243	0.545	0.40678
News	14	11	19	0.5789	0.786	0.66667
Gospel	30	21	42	0.5	0.7	0.58333

This data shows better performance with medical texts (recall factor, 77% and precision, 90% -the highest value-) and scientific texts than with general-purpose texts (novels), Gospel texts or news, where the language used is not as direct and concise. That is why from now on we decided to perform summaries with these type of texts.

Using medical texts as source, we provided an algorithm to draw the mined causal sentences into a causal graph. By reading the nodes of this graph another program automatically provides a comprehensive story of the causal links between several factors and their effects as seen in Fig. 2 [11].

Fig. 2. Example of answer by reading a causal graph.

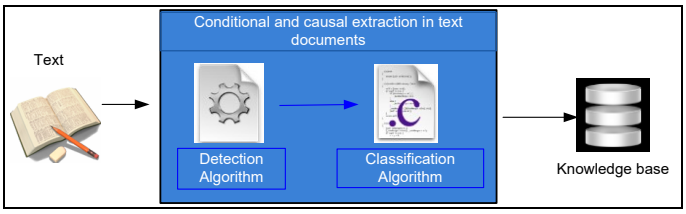


But this approach had two main problems:

- All nodes in the graph exhibit the same importance, as we had no way to evaluate the relevance of the sentences used compared with other causal sentences.
- Redundancy of nodes should be solved, as we had many implicit ways to define the same concept, eg: "Tobacco use", "smoking".

So we needed a criterion to select 'the best' causal sentences of the paper, or the most interesting to be included in our summary. With the metrics and the algorithm that we present in this paper we are able to evaluate causal sentences among a document, and so in the future, establish a ranking of relevance among them that could lead in a more accurate causal graph, and so in a more suitable summary. So, the first step of the algorithm to do this, is to create a causal knowledge base with the sentences extracted to apply the metrics defined in the next section.

Fig. 3. Steps of the algorithm to create a causal knowledge base.



III. SENTENCE SCORING METHODS FOR TEXT SUMMARIZATION

There are different metrics used to get an extractive summary, and which allow to apply different criteria when sorting the statements in a document. Each metric analyzes a specific characteristic of the sentences and, based on that, assigns a score to each sentence. Then, these scores are used to sort statements from highest to lowest. After this, a threshold is applied to get the most relevant statements in relation to the characteristic being considered, which allows controlling the size of the resulting summary.

The methods used range from identifying certain expressions within the text (such as "most importantly," "finally," "in summary," "this article describes," etc.) to more complex calculations such as how central a sentence is (calculating the number of co-occurrences of the words in it with the rest of the document).

In this article, the summaries obtained applying six known metrics are analyzed and they are also compared with the summary formed by causal sentences. The metrics selected are calculated from statement position, length and word frequency. Below, we describe briefly each of these metrics based on S_i (i-th sentence in document D).

A. Sentence Position

This metric, defined by Baxendale in 1958 [12], measures how close the sentence is to the end, the beginning, and the ends of the document (both the beginning and the end), as the equations 1, 2 and 3, respectively show:

$$POS_L(S_i) = i \tag{1}$$

$$POS_F(S_i) = \frac{1}{i} \tag{2}$$

$$POS_B(S_i) = \max\left(\frac{1}{i}, \frac{1}{n-i+1}\right) \tag{3}$$

being n the total number of sentences in D and being i a number between 1 and n that is assigned sequentially to each sentence based on their occurrence within the document, from the beginning to the end. Its calculation may vary depending on whether sentence position within a section, paragraph, etc. is considered or not.

B. Sentence Length

This metric is used to apply a penalization to sentences that are too short, since it is expected that these are excluded from the summary. Defined by Nobata et al. in 2001 [13], can be calculated using either the number of words in the statement or the number of characters in it, as shown by equations 4 and 5, respectively.

$$LEN_W(S_i) = |words(S_i)| \tag{4}$$

$$LEN_{CH}(S_i) = |characterers(S_i)| \tag{5}$$

where $|\ .\ |$ indicates set cardinality.

C. Average Word Frequency

This metric, defined by Vanderwende in 2007 [14], calculates the average frequency of the words in sentence S_i, as shown in equation (6).

$$TF(S_i) = \frac{\sum_{w \in words(S_i)} \mathrm{tf_w}}{|words(S_i)|} \quad (6)$$

IV. EXPERIMENT AND RESULTS

In this paper, we assess the quality of the summaries formed by the causal sentences of a document and those obtained by applying each of the sentence scoring methods to that same document. To determine the quality of a summary generated automatically, it is compared individually with the summary created by a human being, which is considered to be the ideal, expected summary.

For the experiment, we used a set of free access articles published in the medical journal PLOS Medicine [15] on biomedical, environmental, social and political health issues. As mentioned in Section II, medicine area was chosen because causal sentences are best detected.

TABLE II. CHARACTERISTICS OF THE DOCUMENTS

FILE ID	NUMBER OF			
	detected causal sentences	sentences in the document	words per document sentence	words in author summary
0040119	5	85	15	137
0040209	12	125	12	64
1000401	10	158	16	108
1000398	10	215	14	112
1000396	18	173	16	119
1000404	5	90	16	116
1000400	8	145	17	143
1001129	8	137	16	133
1001131	5	98	18	109
1001130	4	81	18	122
1001250	38	155	17	117
1001247	11	124	19	125
1001263	8	79	19	159
1001260	12	109	17	119
1001258	15	180	18	100
1001298	11	131	15	113
1001404	16	136	16	158
1001416	5	132	15	157
1001384	17	112	16	129
1001421	18	102	17	134
1001391	12	123	15	145
1001390	8	139	13	123
1001385	20	120	15	139
1001415	20	144	15	162
1001641	10	130	16	179
1001693	5	130	14	145
1001730	12	202	14	104
1001701	15	158	13	92

From all available documents, those whose summaries had more than 6 sentences and were not subdivided into several sections were selected. Table II details each of the documents used, indicating the number of causal sentences detected, number of sentences, average number of words per sentence, and number of words in the summary produced by the authors.

The documents were downloaded in XML format through the Internet and prepared as applicable for the experiment to be carried out. First, the summary created by the authors was put aside and the title of the article was removed, as well as the titles of any sections in the article (having previously discarded entire, non-relevant sections such as References and Acknowledgments, as well as all figures). Then, the rest of the document was segmented by dividing the text into smaller portions using full stops as delimiters, except when it was used as separator and to form abbreviations.

From the set of sentences in each document, exactly as they appear, causal sentences were then identified using the morphological parser described in Section II.

Also, each of the metrics described in Section III was calculated for each sentence. To do this, the words in each sentence had to be separated first using white spaces and punctuation marks. For simplicity, in this pre-processing stage a "word" was considered to be formed solely by alphabetic characters. Then, stopwords were removed, and finally, words were reduced to their stems.

The Python programming language was used both for document download and pre-processing, as well as for calculating the corresponding metrics and comparing them to the causal variations. The stemming algorithm used was Porter with the implementation provided in package NLTK [16], including stopword list for the English language.

To assess summary quality, ROUGE [17] was used. This is a software package developed by Chin-Yew Lin that allows the automatic assessment of summaries. Among the measurements provided in this package, ROUGE-N [18] was selected because it is one of those frequently used in literature. This evaluation metric is based on n-gram co-occurrence, whose equation is shown below:

$$\frac{\sum_{S \in \{author\ summary\}} \sum_{n-gram \in S} count\_match(n-gram)}{\sum_{S \in \{author\ summary\}} \sum_{n-gram \in S} count\ (n-gram)} \quad (7)$$

where the denominator is the sum of all occurrences of all n-grams in the summary created by the author, and the numerator is the sum of all co-occurrences of the n-grams in the automatic summary and the summary created by the author. An n-gram is a contiguous sequence of n words from a given text. In this article we calculated ROUGE-1, which uses unigrams (n-grams of size 1), because we are interested in the number of simple words that coincide with the author's abstract. For this same reason, the TF metric was calculated by word and not by bigrams or trigrams.

TABLE III. SIZE OF EACH TYPE OF SUMMARY AND PERCENTAGE OF ABSTRACTS NOT USED

	CAUSAL	POS_L	POS_F	POS_B	LEN_CH	LEN_W	TF
SUMMARY SIZE	0.11	0.10	0.10	0.10	0.16	0.16	0.12
UNUSABLE SUMMARY	0.77	0.71	0.68	0.69	0.75	0.75	0.73

Table III, in the first row, shows the size of each summary as average word percentage of the size of the documents. In the second row, for each metric, the average proportion of words in

the automatic summary that do not match the expected summary is showed.

Table IV shows the value of ROUGE-1 for each type of summary and for each document. In the case of the causal summary, it is built from all detected causal sentences. For the remaining metrics, values were ordered from highest to lowest, and summaries were built with the first n best ranking sentences, where n is the number of causal sentences detected for the document being summarized.

TABLE IV. OBTAINED ROUGE-1 VALUES FOR THE DOCUMENT. THE VALUES OF EACH ROW WERE COLORED USING A GRADIENT BETWEEN GREEN AND RED DEPENDING ON THE HIGHEST VALUE OBTAINED AND THE LOWEST RESPECTIVELY.

FILE ID	ROUGE-1 VALUE FOR EACH METRIC						
	CAUSAL	POS_L	POS_F	POS_B	LEN_CH	LEN_W	TF
0040119	0.24	0.16	0.23	0.20	0.28	0.31	0.19
0040209	0.35	0.52	0.52	0.52	0.52	0.52	0.52
1000401	0.44	0.21	0.45	0.39	0.60	0.68	0.43
1000398	0.36	0.51	0.51	0.51	0.51	0.51	0.51
1000396	0.60	0.58	0.60	0.65	0.65	0.64	0.56
1000404	0.19	0.20	0.20	0.24	0.31	0.31	0.31
1000400	0.27	0.33	0.51	0.52	0.56	0.58	0.40
1001129	0.32	0.34	0.38	0.41	0.63	0.57	0.58
1001131	0.29	0.18	0.31	0.23	0.35	0.33	0.23
1001130	0.15	0.26	0.30	0.22	0.32	0.34	0.24
1001250	0.67	0.77	0.65	0.90	0.85	0.82	0.72
1001247	0.40	0.42	0.50	0.39	0.51	0.53	0.34
1001263	0.33	0.66	0.40	0.47	0.61	0.56	0.46
1001260	0.44	0.51	0.26	0.53	0.44	0.42	0.37
1001258	0.45	0.45	0.57	0.48	0.60	0.68	0.48
1001298	0.47	0.41	0.46	0.41	0.55	0.58	0.49
1001404	0.52	0.39	0.36	0.39	0.53	0.53	0.36
1001416	0.20	0.31	0.29	0.35	0.39	0.39	0.20
1001384	0.60	0.52	0.73	0.52	0.84	0.83	0.61
1001421	0.64	0.62	0.68	0.64	0.79	0.82	0.61
1001391	0.41	0.41	0.47	0.57	0.52	0.51	0.38
1001390	0.35	0.37	0.32	0.35	0.42	0.42	0.25
1001385	0.57	0.49	0.51	0.49	0.66	0.63	0.56
1001415	0.46	0.65	0.65	0.65	0.65	0.65	0.65
1001641	0.34	0.32	0.41	0.49	0.56	0.53	0.27
1001693	0.17	0.29	0.36	0.40	0.40	0.46	0.21
1001730	0.29	0.40	0.47	0.40	0.47	0.49	0.33
1001701	0.16	0.44	0.34	0.38	0.41	0.46	0.39

As can be seen in Table IV, the summary formed by the causal sentences obtains the lowest ROUGE-1 value in approximately 50% of cases. That is an unexpected result. In our view, may be due to the mismatch between the words included in the causal sentences and those that form the abstract, showing that the causal setences are not intended to contain the words that make up the abstract.

On the other hand, the LEN metrics obtain the best ROUGE-1 value since, with the same number of sentences as the other metrics, when considering the longest in terms of words and characters, they are more likely to contain the words of the summary.

This could be improved if metrics are used to rank causal sentences after they have been identified. In addition, it would be relevant to analyze other evaluation mechanisms that allow to weight the importance of each word within the summary since ROUGE-1 compares only by quantity.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a method to extract causal sentences from text documents and due to certain metrics evaluate how relevant they are to compose an extractive summary. We have checked that despite causal sentences contain a great deal of information linking concepts, it is quite ambitious to create an extractive summary just using these type of sentences. To do so, we have compared a causal summary (only created with causal sentences) with what could be a regular summary of the document, and we have measured how close they are. Despite in some cases they are not that far, it is honest to say that the combination of causal sentences with other sentences better ranked in the document could produce a better summary. This observation will serve us in the future to attempt to create better extractive summaries by doing this, and to solve a very important problem with no solution in the past.

In previous works [11], as we said in Section II, we created a causal graph to create the summary. The problem there was that we did not know how to rank the sentences to create the graph, and so the summary. In that work, we chose randomly 15 sentences related to the topic to compose the graph. With the metrics presented in this paper, another algorithm to rank these causal sentences according to their importance in the document can be designed and so create a more accurate graph allowing, to weight the nodes, and so to cast the best causal path between antecedents and consequents. This way a more relevant summary should be generated.

ACKNOWLEDGMENT

This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant MERINET:TIN2016-76843-C4-2-R (AEI/FEDER, UE) and under grant TIN2014-56633-C3-1-R.

Augusto Villa Monte thanks the National University of La Plata for funding his PhD in Computer Science in this university through a type B postgraduate fellowship.

REFERENCES

[1] Reber, P. 'What is the memory capacity of human brain?, Scientific American, May, 2010.

[2] Mackie, John L. (1988), *The Cement of the Universe: A study in Causation*. Clarendon Press, Oxford, England.

[3] Puente C., Sobrino A., Olivas J. A., Merlo R., 'Extraction, Analysis and Representation of Imperfect Conditional and Causal sentences by means of a Semi-Automatic Process'. Proceedings IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010). Barcelona, Spain, pp. 1423-1430, 2010.

[4] Hovy, E. H. (2005), 'Automated Text Summarization'. In R. Mitkov (ed), *The Oxford Handbook of Computational Linguistics*, chapter 32, 583–598. Oxford University Press, 2005.

[5] Mani, I. (2001), 'Summarization evaluation: An overview'. In Proceedings of the North American chapter of the association for computational linguistics (NAACL) Workshop on automatic summarization, 2001.

[6] Kaplan R.M., G. Berry-Bogghe, Knowledge-based acquisition of causal relationships in text, Knowledge Acquisition, 1991, 3, 317-337.

[7] Rink B., Bejan C. A., Harabagiu S., Learning textual graph patterns to detect event relations, Proceedings of the Twenty-Third Int. Florida

Artificial Intelligence Research Society Conference (FLAIRS 2010), 265-270.

[8] Thomson, J. J., "Verbs of action," Synthese, vol. 72, no. 1, pp. 103–122, 1987.

[9] Altenberg, B. (1984), "Causal linking in spoken and written english," Studia linguistica, vol. 38, no. 1, pp. 20–69.

[10] Hilton, D. J., "Conversational processes and causal explanation", Psychological Bulletin, vol, 107(1), Jan 1990, 65-81.

[11] Sobrino A., Puente C. and Olivas J.A. "Extracting Answers from causal mechanisms in a medical document". Neurocomputing vol. 135 (2014) pp.53–60. ISSN: 0925-2312.

[12] P. B. Baxendale, "Machine-made index for technical literature—an experiment," IBM Journal of Research and Development, vol. 2, no. 4, pp. 354–361, Oct 1958.

[13] C. Nobata, S. Sekine, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, "Sentence extraction system assembling multiple evidence," in Proceedings of the Second NTCIR Workshop Meeting, 2001.

[14] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," Inf. Process. Manage., vol. 43, no. 6, pp. 1606–1618, Nov. 2007.

[15] PLOS Medicine: A Peer-Reviewed Open Access Journal [Online]. Available: http://journals.plos.org/plosmedicine/

[16] Natural Language Toolkit [Online]. Available: http://www.nltk.org/

[17] Recall-Oriented Understudy of Gisting Evaluation. A software package for automated evaluation of summaries [Online]. Available: http://www.berouge.com/Pages/default.aspx

[18] C. Lin, "Rouge: a package for automatic evaluation of summaries," in Proceedings of the ACL-04 Workshop, pp. 74-81, Jul. 2004.