*International Journal of Advances in Software Engineering & Research Methodology– IJSERM*
*Volume 2: Issue 2* [ISSN : 2374-1619 ]
*Publication Date : 30 October, 2015*

# Engineering Abstract Quality Index using Santos' Move

Syamsiah Mashohor, Tunku Haifaa Tunku Osman, Helen Tan, Chan Swee Heng, and Ain Nadzimah Abdullah
Universiti Putra Malaysia

***Abstract*— Writing informative abstract is challenging especially for novice student writers due to word limit. Until today, there is no tool in helping writers write good abstract. In this paper, a framework was proposed to check the quality of engineering abstract with two important criteria, which are unique keyword and principal move. The execution of the proposed framework on engineering corpus shows the capability of the proposed work in detecting essential moves according to Santos' Move. The obtained Abstract Quality Index (AQI) values show that the quality of abstract could be evaluated in a more detailed manner. The findings from the engineering abstract corpus yield that shorter length of abstract with right keywords produced higher value of AQI.**

***Keywords*— Engineering abstract checker, abstract quality index, Santos' Move.**

## I. Introduction

To portray a well-written document, it is crucial to have an interesting start. Not only interesting, all vital information regarding the write-up need to be included so that a reader can get the idea of the overall content. Most writers, particularly the novice writers, do not take into consideration the importance of an abstract. Perhaps it could be that they lack the skill of producing a good one. Furthermore, it could be that they are not aware of the abstract genre which could act as a guide for effective abstract writing.

Based on the abstract structure, there are important rhetorical moves that have to be included in an abstract. The moves are brief background of the topic, the purpose, method (if appropriate), results and conclusions of an article. An abstract with these distinct moves would definitely have a logical flow of ideas and this in turn will make the piece of writing engaging for the readers.

There are many different formats for abstract writing which have been created since early 1980s. Different kind of abstract has different kind of focus, such as certain format has either four or five moves. One of these abstract formats is Santos' move [1]. Santos' move was created in 1996 and they were said to be able to fulfill certain communicative purpose. Santos' moves have five structure features, which are situating, presenting, describing, summarizing and discussing the research.

## II. Literature Review

Studies on abstract writing abound and such interest shown by writing scholars indicates the importance of abstract writing among academicians. The convention of abstract writing has been studied by several researchers in the field of Applied Linguistics [2], [3] and [1] and they have found that at the macro level, abstracts have a distinctive five rhetorical move structures. They are Move 1- Situating the research, Move 2 - Presenting the research, Move 3 - Describing Methodology, Move 4 - Summarising findings and Move 5 - Discussing the results [3]. These five move structures provide the textual organization of the abstract which allows the writers' ideas to flow coherently and therefore the text becomes more reader friendly.

Besides investigating the rhetorical move structures, researchers in [1] and [3] also examined the linguistic resources of each move. In [1], he revealed that certain move takes on certain linguistic resources such as thematization, tense voice or voice choice to fulfil its communicative intent. For example, Move 1 is typified by the use of the present tense while Move 2 is characterised by the use of the diectic form and reporting verbs. Besides examining the linguistic resources, [3] also looked at the linguistic realization of the authorial stance in the moves. Adapting framework on grammatical subjects in [4] and [5], [3] classified the grammatical subjects into two broad domains namely phenomenal classes and epistemic classes. Under these two classes, there are sub categories; for example, under epistemic class, there are self-reference, other reference and audience. In the study, it was found that most grammatical subjects in Move 1 referred to other reference and the most frequent subject category in Move 3 was objects of research and their attributes [3].

To write an effective abstract, novice writer would need to know how the abstract is structured and how each rhetorical move is realized linguistically. Such input could only be found in the findings of a corpus-based research which addresses the differences in abstracts writings between different disciplines. Although there is a lack of instructional input on abstract writing in writing manuals, an article in [6] showcased how abstract and critique writing could be taught using three simple steps. The author claims that there was positive feedback from her students when abstract writing was taught using the following steps: laying the foundation, communicating expectation and evaluation criteria and scaffolding for success. The attempt is indeed commendable; however, it has been found that the study does not highlight the textual organization of the abstract and the specific linguistic realisations pertaining to each move. This definitely would greatly disadvantage the novice writers.

In [7], the researchers studied on comparison of unstructured and structured abstract by evaluating clarity (measured on a scale of 1 to 10) and completeness (measured with a questionnaire that used 18 items) to 64 participants. The study proposed that Software Engineering journals and conferences must adopt structured abstracts to enhance the readability of abstract.

The most similar concept to abstract assessment is evaluation of readability quality of document and traditionally called as Automated Readability Index at http://read-able.com/. In a survey conducted in [8], there are two categories of computational assessment method of text readability, which are traditional methods and machine learning (ML) based methods. The most well-known traditional readability measures include Flesch - Kincaid Grade Level [9], Degrees of Reading Power [10] and Lexile scores [11]. Many improvements have been done to the traditional method which relies on two main factors: the familiarity of semantic units such as words or phrases, and the complexity of syntax. The main advantage is simplicity of the algorithm and low computing time.

A machine learning based method was initiated in [12] and it is called the artificial intelligence approach to readability. This new approach typically combine a rich representation of the text being evaluated, using a variety of linguistic features, with more sophisticated prediction models based on machine learning. These features are applied on the readability checker and can be used for abstract checker framework.

Similar prototypes to abstract checker are the automated essay grading [13] and automated programming assignment grading [14]. The automated prototypes aim to check the quality of essay or program and provide grade to the submitted documents. The researchers in [13] implemented information retrieval based on stemming and similarity function is used to calculate the distance between submitted essay and the model answer. The final distance will determine the grade given based on a grading scale decided by the teacher. In [14], Latent Semantic Analysis (LSA) is used to analyze the decomposition of structures-documents matrix and similarity is measured in two-dimensional query vector space. The LSA-based assignment grading is based on provided model answers and semantic knowledge is built from those structures.

With the advancement made in computer technology, the learning of abstract has taken a new dimension. No longer confined to writing manuals, researchers have developed software to assist novice writers write abstract efficiently. Researchers in [15] has used RedACTe approach to design a software which is 'oriented to rhetorical and linguistics assistance in research paper abstract writing'. In more recent times, framework in [16] has also developed an abstract checker to assist undergraduate writers to write a more successful abstract. The prototype is developed and evaluated using the researchers' corpus, Learner Corpus of Engineering Abstracts (LCEA, 2014) [17]. The corpus is a collection of the final year project thesis written by Computer and Communications Engineering student in Universiti Putra Malaysia. The main contribution of the framework is the successful detection of move types based on keywords but there is no grade or value given to the checked abstract.

# III. The Proposed Method

## A. *Overall architecture of the framework*

The framework consists of five main stages which are section identification, stemming of keywords, unique keyword rule, principal move rule and Abstract Quality Index (AQI) calculation. The input of the framework is an abstract without the title of thesis and Abstract Quality Index (AQI) which indicates the quality of the uploaded abstract will be calculated. Both unique move rule and principal move rule are optional settings to restrict AQI calculation with the aim to ensure only important information is included in an input abstract. Overall framework relies on listing of keywords that are retrieved by a group of human expert from the sample of training abstracts to represent every Santos' Move. Santos' Move is one of the accepted rule for thesis's abstract and it contains five-move rule which are Background (Move 1), Objectives (Move 2), Methodology (Move 3), Results (Move 4) and Conclusion (Move 5).

Step 1: Section Identification

In this framework, a section could be a sentence or a paragraph depending on the length of abstract. The comparison of keyword-move list is based on individual sentence or paragraph. Therefore, the whole abstract is divided into sentences or paragraphs and numbers of identified sections are counted for further AQI calculation.

Step 2: Stemming of Keywords in Move List

Stemming has been used for effective retrieval of keyword and appropriate for an abstract checker due to the comprehensiveness of an abstract. Suffix stripping is an established method in information retrieval and many researches have been done for English Language [18] and [19]. Keyword in move's list are provided by the human expert and based on the abstracts in LCEA corpus. The keywords are stemmed manually and regular expression matching is used to match words retrieved from the abstract.

**Definition 1**: $s_l$ is a representation of section in an abstract, whereby $l$= {1,2,3,...,$maxl$}.

**Definition 2**: $i$=1,2,3,4,5 represents five-move rule which are Background ($i$=1), Objectives ($i$= 2), Methodology ($i$= 3), Results ($i$= 4) and Conclusion ($i$= 5).

**Definition 3**: $m$ is a move detected and it will vary from $m_1$ until $m_5$.

TABLE I. ANALYSIS OF AN INPUT ABSTRACT FOR IDEAL CASE WHEN UNIQUE=0 AND PRINCIPAL=0

| | $f_{m1}$ | $f_{m2}$ | $f_{m3}$ | $f_{m4}$ | $f_{m5}$ |
|---|---|---|---|---|---|
| $s_1$ | 1 | 0 | 0 | 0 | 0 |
| $s_2$ | 0 | 1 | 0 | 0 | 0 |
| $s_3$ | 0 | 1 | 0 | 0 | 0 |
| $s_4$ | 0 | 0 | 1 | 0 | 0 |
| $s5$ | 0 | 0 | 1 | 0 | 0 |
| $s_6$ | 0 | 0 | 0 | 1 | 0 |
| $s_7$ | 0 | 0 | 0 | 1 | 0 |
| $s_{maxl}=8$ | 0 | 0 | 0 | 0 | 1 |
| $f_{mtotal}$ | 1 | 2 | 2 | 2 | 1 |

TABLE II. ANALYSIS OF AN INPUT ABSTRACT WHEN PRINCIPAL IS INACTIVE (SET TO 0)

| | Principal=0 | | | | |
|---|---|---|---|---|---|
| | $f_{m1}$ | $f_{m2}$ | $f_{m3}$ | $f_{m4}$ | $f_{m5}$ |
| $s_1$ | 2 | 1 | 0 | 0 | 0 |
| $s_2$ | 1 | 2 | 0 | 0 | 0 |
| $s_3$ | 0 | 1 | 2 | 0 | 0 |
| $s_4$ | 0 | 1 | 2 | 0 | 0 |
| $s_5$ | 0 | 1 | 3 | 1 | 0 |
| $s_6$ | 0 | 0 | 2 | 3 | 0 |
| $s_7$ | 0 | 0 | 1 | 4 | 0 |
| $s_{maxl}=8$ | 0 | 0 | 0 | 1 | 3 |
| $f_{mtotal}$ | 3 | 6 | 10 | 9 | 3 |

TABLE III. ANALYSIS OF AN INPUT ABSTRACT WHEN PRINCIPAL IS ACTIVE (SET TO 1)

| | Principal=1 | | | | | |
|---|---|---|---|---|---|---|
| | $f_{m1}$ | $f_{m2}$ | $f_{m3}$ | $f_{m4}$ | $f_{m5}$ | $P_m$ |
| $s_1$ | 2 | 1 | 0 | 0 | 0 | $m_1$ |
| $s_2$ | 1 | 2 | 0 | 0 | 0 | $m_2$ |
| $s_3$ | 0 | 1 | 2 | 0 | 0 | $m_3$ |
| $s_4$ | 0 | 1 | 2 | 0 | 0 | $m_3$ |
| $s_5$ | 0 | 1 | 3 | 1 | 0 | $m_3$ |
| $s_6$ | 0 | 0 | 2 | 3 | 0 | $m_4$ |
| $s_7$ | 0 | 0 | 1 | 4 | 0 | $m_4$ |
| $s_{maxl}=8$ | 0 | 0 | 0 | 1 | 3 | $m_5$ |
| $f_{mtotal}$ | 2 | 2 | 7 | 7 | 3 | |

**Definition 4**: $f_m$ is a frequency of move detected in a sentence, $s_l$. Therefore, every move type has its own frequency in a sentence. $f_{m\text{total}}$ is a frequency of particular move's occurrence in the whole abstract.

**Definition 5**: $w_s$ is a word retrieved from a sentence and $k_{mj}$ is a keyword contained by the move's list. *mj* represents the move type, *m* and *j* is the index for a keyword in the list. Any matching word $w_s$ and keyword $k_{mj}$ will increase the $f_m$ for that particular sentence, $s_l$. Therefore, ideally, the summation of $f_m$ for all sentences will be equal to the value of *maxl*, as shown in Table I.

Step 3: Unique Move Rule

Unique move rule is necessary to overcome occurrence of a keyword representing multi moves in single section or more. Uniqueness of keyword is lost when a keyword appears in more than one keyword-move list. This process will analyse the occurrences of multiple moves, $f_m$ in any section, $s_l$ which contributed by an identified keyword. Therefore, the *j* number of keyword for each move type, *m* will be different in different input abstract. The rule for removal of the identified keyword, $k_{mj}$ will be performed if the unique move rule is activated (set to 1).

Step 4: Principal Move Rule

Principal move analysis overcomes the occurrences of multiple moves, $f_m$ in a sentence, $s_l$ which determine a single move is representing a sentence. This is appropriate when a sentence is meant to deliver specific information for one move, as in an ideal case. The differences in analysis when principal is active or inactive can be seen in Table II and Table III.

**Definition 6**: $Pm_l$ indicates the principal move for a sentence, $s_l$ and maximum value of $f_m$ for respective move type *m* is considered as a principal move. The value of $f_{mtotal}$ dependent to which move that has been identified as $Pm_l$, whenever the principal setting is activated (=1).

Step 5: AQI Calculation

$$AQI = \frac{\sum_{s=1}^{maxl} \sum_{m=1}^{5} w_m f_m}{maxl} \quad (1)$$

Whereby $w_m$ is a weight given to every move, *m* and this is depend on the importance of information for a specific type of abstract. For engineering field, informative abstracts are commonly used [20] and require detail information on the aim (m2), methodology (m3) and results (m4).

# IV. Results and Discussions

### *B. Analysis on Engineering Abstract Samples (LCEA, 2014)*

Learner Corpus of Engineering Abstracts (LCEA, 2014) [17] is a copyrighted product aimed for linguistic researchers to help engineering student writers write successful abstract writing. In this corpus, the average number of sentence is 10 sentences per abstract and there are 998 engineering abstracts compiled in the corpus. The corpus is available and can be requested from researchers in [17].

From Figure 1, it is observed that all abstracts contained Move 2 (aim (m2)), Move 3 (methodology (m3)) and Move 4 (results (m4)) as required in informative abstract. It is very rare in these abstracts that Move 5 (conclusion (m5)) is mentioned. Move 1 (background of work (m1)) is mentioned in 61% of abstracts but the selection of keywords

for this move is hard due to the vast area of background in engineering field. The performance of detected moves based on selected keywords implemented in this framework is matched to manual assessment by a group of experts as reported in [16]. The human experts are experienced writers in engineering field and they are trained by the English Language lecturers from the Faculty of Modern Language and Communication, Universiti Putra Malaysia in performing the assessment. From the observation, the quality is relatively poor due to missing essential structure in the abstracts.

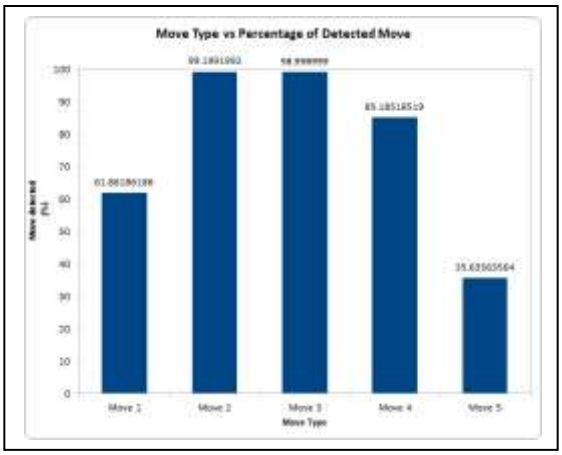


Figure 1. Percentage of move detection for every move type for all abstracts.

### C. *Overall AQI Analysis*

For this implementation, informative abstract structure is used whereby Move 2 (aim), Move 3 (methodology) and Move 4 (results) are obligatory components. Hence, weight given to every move is $w_1$=0.1, $w_2$=0.2, $w_3$=0.3, $w_4$=0.3 and $w_5$=0.1. Figure 2 shows the minimum, median and maximum values of AQI obtained whenever unique move rule and principal move rule is activated (set to 1) and deactivated (set to 0). There is a significant reduction of maximum AQI obtained, which is 30% of reduction when principal is activated and unique is inactive. The reduction happened caused by elimination of multi move type that represent a sentence. Only single move with highest frequency of occurrence, $f_m$ in a sentence is counted in the calculation of AQI.

When unique is active and principal is inactive, the reduction of maximum AQI is 45%. The reduction happened due to removal of keyword, $k_{mj}$ in the keyword-move list. Lesser keyword listed will reduce the frequency of move, $f_m$ and this contributes directly to AQI value.

The massive reductions in maximum values confirmed that the quality assessment is restricted. The occurrences of multiple moves in single sentence or a keyword representing different moves will increase the AQI and it gives wrong impression on quality of written abstract. If no rules are applied, these phenomena may produce unstructured abstract and it will be difficult for writer to improve the content of abstract.

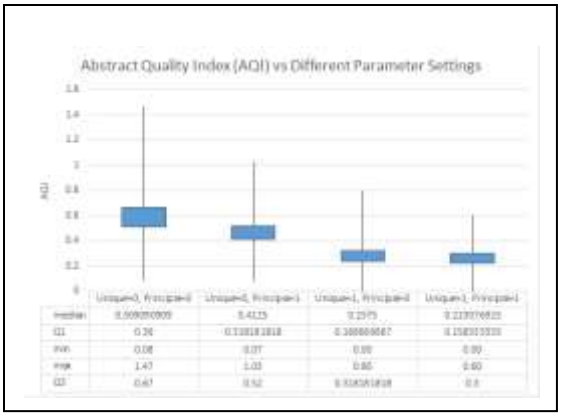


Figure 2. AQI with different unique and principal settings.

Median and minimum values of AQI do not differ much when the rules are applied. It is also observed that poor abstracts are very short and most of the moves do not exist. Table IV show that AQI is more dependable to represent the quality of an abstract compared to the number of sentences. Shorter length of abstract could obtain higher AQI value if right keywords are used to describe the five Santos' Move. Referring to Table V, the highest AQI values are produced by short length abstracts (less than 10 sentences) which normally have all the moves. However, the number of word per sentence is not considered in this study and perhaps it should be included in future studies. Additionally, it is also observed that Move 5 (conclusion) is often missing even in abstracts with high AQI values.

TABLE IV. AVERAGE AQI FOR DIFFERENT LENGTH OF ABSTRACT

| Number of sentences per abstract | Number abstract with sentence length | Average AQI |
|---|---|---|
| 1-9 sentences | 467 | 0.60 |
| 10-20 sentences | 516 | 0.50 |
| >20 sentences | 18 | 0.35 |

TABLE V. TOP 5 HIGHEST RANKED ABSTRACT ACCORDING TO AQI VALUES (BOTH RULES ARE ACTIVE).

| Filename | *maxl* | $fm_1$ | $fm_2$ | $fm_3$ | $fm_4$ | $fm_5$ | AQI |
|---|---|---|---|---|---|---|---|
| 577.txt | 7 | 0 | 9 | 7 | 1 | 0 | 0.60 |
| 825.txt | 6 | 2 | 2 | 10 | 0 | 0 | 0.60 |
| 818.txt | 6 | 3 | 2 | 5 | 4 | 2 | 0.60 |
| 034.txt | 6 | 0 | 0 | 6 | 5 | 2 | 0.58 |
| 491.txt | 6 | 3 | 5 | 4 | 3 | 0 | 0.57 |

# v. Conclusions and Future Works

To conclude, the framework is successful in checking the quality of an abstract by producing AQI value. Embedding two additional rules have produced better abstract checking. The restriction rules will help the writers to write their message clearly and writing a sentence to represent a single move is a good practice. This early version, however, needs to be further validated for more effective use. The prototype is currently constrained to a limited level of keyword comparison to identify move patterns that describe abstract writing.

## *Acknowledgment*

This project is sponsored by Research University Grant Scheme (RUGS) Universiti Putra Malaysia and the continuous work has been sponsored by Ministry of Education Malaysia under Fundamental Research Grants Scheme (FRGS).

## *References*

[1] Santos, M.B.D.: The textual organization of research paper abstracts in applied linguistics. Text, 16 (4), 481-499.(1996)

[2] Ren, H.V. and Li, Y.Y.: A comparison study on the rhetorical moves of abstract in published research articles and master's foreign language theses. English Language Teaching, 4(1), 162-166. (2011)

[3] Pho: P.D.Research article abstracts in applied linguistics and educational technology: a study of linguistic realizations of rhetorical structure and authorial stance. Discourse Studies, 10(2), 231-250. (2008)

[4] MacDonald, S.P.: A method for analysing sentence-level differences in disciplinary knowledge making.Written Communication, 9(4), 533-569.(1992)

[5] MacDonald, S. P.: Professional Academic Writing in the Humanities and Social Sciences. Carbondale and Edwardsville:Southern Illinois University Press.(1994)

[6] Harris, M.J.: Three steps to teaching abstract and critique writing. International Journal of Teaching and Learning in Higher Education, 17(2), 136-146. (2006)

[7] Budgen, D. et al.: Presenting software engineering results using structured abstracts: a randomised experiment. Empirical Software Engineering, 13, 435–468. (2008)

[8] Collins-Thompson, K. Computational assessment of text readability: A survey of past, present, and future Research. Retrieved at http://www-personal.umich.edu/~kevynct/ (Last access on 17 July 2014). (2014)

[9] Klare, G. R.: Assessing readability.Reading Research Quarterly, 10, 62-102. (1974-1975).

[10] Koslin, B.L., Zeno, S., and Koslin, S. The DRP: An effective measure in reading. New York: College Entrance Examination Board. (1987)

[11] Stenner, A. J.: How Accurate Are Lexile Measures?. Journal of Applied Measurement, 7(3), 3017-322. (2006)

[12] Thomas Francois and Eleni Miltsakaki: Do NLP and machine learning improve traditional readability formulas?.Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations (PITR '12), pp 49-57. (2012)

[13] S.M.F.D Syed Mustapha, Norisma Idris and Rukaini Abdullah: Embedding Information Retrieval and Nearest-Neighbour Algorithm into Automated Essay Grading System, Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05), pp 169-172.(2005)

[14] Kartinah Zen, D.N.F Awang Iskandar and Ongkir Linang: Using Latent Semantic Analysis for Automated Grading Programming Assignments, International Conference on Semantic Technology and Information Retrieval, pp 82-88. ( 2011)

[15] Castel V.M.: XML technology assisted research paper abstract writing. Paper read at FAAPI conference 2006. Multiple Literacies-Beyond four skills, Rosario, Argentina. (2006)

[16] Tunku Haifaa Tunku Osman: Engineering Abstract Checker. Bachelor Thesis. Faculty of Engineering, Universiti Putra Malaysia. (2014)

[17] Tan, H., Chan, S.H., Ain Nadzimah and Syamsiah, M: Learner Corpus of Engineering Abstracts, Universiti Putra Malaysia (copyright: 7 March, 2014).

[18] M.F. Porter: An algorithm for suffix stripping. Program, Vol. 14 Iss 3 pp. 130 - 137. (1980)

[19] Peter Willett: The Porter stemming algorithm: then and now, Program, Vol. 40 Iss 3 pp. 219 - 223. (2006)

[20] Velany Rodrigues: How to write an effective title and abstract and choose appropriate keywords.http://www.editage.com/insights/how-to-write-an-effective-title-and-abstract-and-choose-appropriate-keywords. (Last access: 13th February 2015)