Integrating of the Diversity and Swarm Based Methods for Text Summarization

Mohammed Salem Binwahlan
Faculty of Computer Science and Information Systems, University Teknologi Malaysia, 81310, Skudai, Johor, Malaysia, +60177582859
moham2007med@yahoo.com

Naomie Salim
Faculty of Computer Science and Information Systems, University Teknologi Malaysia, 81310, Skudai, Johor, Malaysia, +607 5532208
naomie@utm.my

Ladda Suanmali
Faculty of Science and Technology, Suan Dusit Rajabhat University 295 Rajasrima Rd, Dusit, Bangkok, Thailand 10300
ladda_sua@dusit.ac.th

ABSTRACT
Automatic text summarization is a technique concerning the creation of a compressed form for single document or multi-documents. The summary creation under the condition of the redundancy and the summary length limitation is a challenge problem. The automatic text summarization system which is built based on exploiting of the advantages of different resources in form of an integration model could produce a good summary for the original document. In this paper, we introduced an integration model for automatic text summarization problem; we tried to exploit different resources advantages in building of our model like advantage of diversity based method which can filter the similar sentences and select the most diverse ones and advantage of the differentiation between the most important features and less important using swarm based method. The experimental results showed that our model got the best performance over all methods used in this study.

Keywords
Binary tree, Diversity, MMI, Summarization, Swarm, Summary.

1. INTRODUCTION
Massive amounts of data which are provided either online on the internet or on other different electronic media like CDs forming a serious problem in terms of the exploration and analysis purposes. Automatic text summarization is a technique concerning the creation of compressed form for single document or multi-documents. The benefits of automatic text summarization system’s availability increase the need for existence of such systems; the most important benefits of using a summary is its reduced reading time and providing quick guide to the interesting information. The aim of automatic text summarization techniques is to find the most important text units and present them as summary of the original document. Each technique differs from another in the way of discovering such text units. The automatic text summarization system which is built based on exploiting of the advantages of different resources in form of an integration model could produce a good summary for the original document. Many techniques have been proposed for automatic text summarization problem based on different methodologies, [1; 2; 3] used the shallow features to score the text units and selecting the highest score text units as summary, [4; 5; 6; 7; 8; 9; 10] could add advantage to the mechanism of the text unit scoring which is the exploitation of the data to create a criteria or weights to be used in the scoring coefficient through the applying of the machine learning for automatic text summarization problem. The discourse structure based techniques [11; 12; 13] employed the discourse structure for the sentence scoring.

All techniques mentioned above did not pay attention to the problem of the redundancy which causes the low quality of the created summary. The method which was built for dealing with the problem of redundancy is MMR [14], many methods made use of MMR either directly or after modifying it, [15; 16; 17; 18; 19; 20; 21].

The improvement of the summary quality remains the key research problem and needs much work like incorporate more than one good technique. Aretoulaki [22] proposed a hybrid system, the system was built based on four modules, where each module tries to look for specific features and information in the input text, then the outputs of those modules are passed to Artificial Neural Network (ANN) to score the text units as important and unimportant based on those outputs of the four modules. Alemany and Fort [23] presented a summarizer based on lexical chains, in which, the cohesive properties of the text were combined with coherence relations to produce good summaries. a different hybrid model was introduced by Cunha et al. [24], which combines mainly three systems, each system produces its own extract, then an algorithm creates the final summary by selecting the highest score sentences from the three extracts after scoring of those extract sentences. The summary creation under the condition of the redundancy and the summary length limitation is a challenge problem. Therefore in this paper, we introduce an integration model for automatic text summarization problem; we try to exploit different resources advantages in building of our model like advantage of diversity based method [25] which can filter the similar sentences and select the most diverse ones and advantage of the differentiation between the most important features and low important using swarm based method [26].

The rest of this paper is organized as follows: Section 2 presents MMI diversity based text summarization method, swarm based text summarization method and introduces integrating of the MMI diversity based text summarization and swarm based text summarization. Section 3 discusses experimental design. Section 4 presents the experimental results. Section 5 presents discussion. Section 6 gives conclusion and future work.

2. INTEGRATION MODEL
In this section, we try to investigate the performance of integrating of two methods: MMI diversity based Text
2.1 MMI diversity based Text Summarization

MMI (maximal marginal importance) [25], it is a text summarization diversity based method for summary generation. It depends on the extraction of the highest important sentences from the original text. The text features used in this method are: sentence centrality, title-help sentence (THS), title-help sentence relevance sentence (THSRS), word sentence score (WSS), key word feature and the similarity to first sentence.

To summarize a document using this method, it is required first to cluster the document sentences into clusters (using k-means clustering algorithm) where each cluster contains the most similar sentences. The clusters number is determined automatically by the summary length (number of sentences in the final summary). Each sentences cluster is represented as one binary tree or more. The first sentence which is presented in the binary tree is that sentence with higher number of friends (higher number of similar sentences), then the sentences which are most similar to already presented sentence are selected and presented in the same binary tree. The sentences in the binary tree are ordered based on their scores. The score of the sentence in the binary tree building process is calculated based on the importance of the sentence and the number of its friends using eq. 1.

\[
\text{Score}_{BT}(s_i) = \text{impr}(s_i) + (1 - (1 - \text{impr}(s_i) \times \text{friendsNo}(s_i)))
\]  

Where: Score_{BT}(s_i) is the score of the sentence s_i in the binary tree building process, impr(s_i) is importance of the sentence s_i, eq. 2 and friendNo(s_i) is the number of sentence friends.

\[
\text{IMPR}(s_i) = \text{avg}(WSS(s_i) + \text{SC}(s_i) + \text{SSNG}(s_i) + \text{sim}_f(s_i) + \text{kwrd}(s_i))
\]  

Each level in the binary tree contains \(2^\ln\) of the higher score sentences, where \(\ln\) is the level number, \(\ln = 0, 1, 2, \ldots, n\), the top level contains one sentence which is a sentence with highest score. In case, there are sentences remaining in the same cluster, a new binary tree is built for them by the same procedure.

MMI is used to select one sentence from the binary tree of each sentence cluster to be included in the final summary. In the binary tree, a level penalty is imposed on each level of sentences which have high scores. Where \(n\) equal the predefined summary length and the sentences are scored using the same features presented in section 2.1. The score of each feature is adjusted by a weight. The features weights are generated by training of the particle swarm optimization (PSO) [27]. 100 documents were selected from Document Understanding Conference (DUC) [28] data collection, DUC 2002 and used as training and testing data. The swarm method is defined as combination of adjusted features scores as in (5)

\[
\text{Score}(s) = \sum_{i=1}^{5} w_i \times \text{score}_f_i(s)
\]  

Where Score(s) is the score of the sentence s, \(w_i\) is the weighted of the feature i produced by PSO, \(i = 1\) to \(5\) and score\(_f_i(s)\) is the score of the feature i.

In MMI method, the score of the sentence in the binary tree is calculated using:

\[
\text{Score}_{BT}(s_i) = \text{impr}(s_i) + (1 - (1 - \text{impr}(s_i) \times \text{friendsNo}(s_i)))
\]

2.2 Swarm Based Text Summarization

The swarm based text summarization method [26] generates a summary of the original document by picking up the top \(n\) sentences which have high scores. Where \(n\) equal the predefined summary length and the sentences are scored using the same features presented in section 2.1. The score of each feature is adjusted by a weight. The features weights are generated by training of the particle swarm optimization (PSO) [27]. 100 documents were selected from Document Understanding Conference (DUC) [28] data collection, DUC 2002 and used as training and testing data. The swarm method is defined as combination of adjusted features scores as in (5)

\[
\text{Score}(s) = \sum_{i=1}^{5} w_i \times \text{score}_f_i(s)
\]  

Where Score(s) is the score of the sentence s, \(w_i\) is the weighted of the feature i produced by PSO, \(i = 1\) to \(5\) and score\(_f_i(s)\) is the score of the feature i.

In MMI method, the score of the sentence in the binary tree is calculated using:

\[
\text{Score}_{BT}(s_i) = \text{impr}(s_i) + (1 - (1 - \text{impr}(s_i) \times \text{friendsNo}(s_i)))
\]

In this equation the importance of the sentence appears in two positions, both two importances of the sentence are calculated using a simple combination of the text features score. We replaced the sentence importance in the second position by the sentence importance which is calculated using swarm weights based adjusted features scores. The new formula of scoring of the sentence in the binary tree became as the following:

\[
\text{Score}_{BT}(s_i) = \text{impr}(s_i) + (1 - (1 - \text{swarm}_{impr}(s_i) \times \text{friendsNo}(s_i)))
\]

The reason of making the features are the centre of the integrating of the two methods is because the features are the cornerstone in the generation process of the text summary. The summary quality is sensitive for those features in terms of how the sentences are scored based on the used features. Therefore the exploiting of the advantages of different resources can be a good way for evaluating the sentences.

3. EXPERIMENTAL DESIGN

The DUC 2002 document sets (D061j, D062j, D063j, D064j, D065j, D066j, D067f, D068f, D069f, D070f, D071f, D072f, D073b and D077b) comprising 100 documents were used for testing [28]. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit [29] is used for evaluation, where ROUGE compares a system generated summary against a human generated summary to measure the quality. ROUGE is the main metric in the DUC text summarization evaluations. It has different variants, in our experiment, we use ROUGE-N (N=1 and 2) and ROUGE-L. In DUC 2002 document sets, each document set contains two model or human generated summaries for each document. We gave the names H1 and H2 for those two model summaries. The human summary H2 is used as benchmark to measure the quality of our proposed model summary, while the human summary H1 is used as reference summary. Beside the human with human benchmark (H2-H1) (H2 against H1); we also use another benchmark which is MS word summarizer (Msword).
4. EXPERIMENTAL RESULTS
We ran three experiments using the same data set, first is MMI diversity based method, second is swarm based method, and third is the integrating of the two previous methods (MMI diversity based method, swarm based method). We present the three evaluation measures ROUGE-1, ROUGE-2 and ROUGE-L with the metrics recall, precision and f-measure.

The results are shown in the tables 1, 2 and 3 using ROUGE-1, ROUGE-2 and ROUGE-L respectively. We can see in table 1 for example there is a big difference between recall and precision; therefore, we will compare all methods based on the average f-measure evaluation which is a balance between recall and precision.

Based on the average f-measure evaluation for ROUGE-1, table 1, the swarm based method is better than MMI diversity based method but the advantage of MMI diversity based method is that it takes into account the redundancy problem while the swarm based method does not consider that problem. The integrating of the MMI diversity and swarm based methods outperforms them individually. The benchmark Msword got less performance than the individual methods and integrated method. The same thing can be said on the performance of the methods for ROUGE-2 in table 2 and for ROUGE-L in table 3.

5. DISCUSSION
In this experiment, we have tested the hypothesis of exploiting of the advantages of different resources in form of an integration model could produce a good summary, it was not reject. The integration point of the two methods (MMI diversity based method and swarm based method) has been made in the scoring phase of the sentences because the sentences scoring is the unique way for determining the most important ideas in the text. The swarm based method played an important role in the differentiation between the most important features and low important using the weights produced by PSO. MMI diversity based method could filter the similar sentences and select the most diverse ones. The experimental results of the proposed model testing have shown good performance when comparing with the two methods (MMI diversity based method and where swarm based method) and the benchmark methods used in this study.

6. CONCLUSION AND FUTURE WORK
In this study, we introduced an integration model for automatic text summarization problem. We built our model based on the integrating of two techniques (the first technique is diversity based and the other technique is non diversity based). The proposed model has advantage of exploiting of the abilities of different resources. The advantage of the diversity based methods is their ability to filter the most similar sentences and select the most diverse ones and advantage of the non diversity method used in this study is the differentiation between the most important features and less important. The experimental results have shown that the proposed model performs well. For future work, we plan to introduce a different hybrid model of the two techniques combined in the current model based on a different methodology for creating the summaries.

7. ACKNOWLEDGMENTS
This project is sponsored partly by the Ministry of Science, Technology and Innovation under E-Science grant 01-01-06—SF0502, Malaysia.

Table 1: MMI, swarm, integrated swarm_MMI, msword summarizer and H2-H1 comparison: average recall, average precision and average f-measure using rouge-1 at the 95%-confidence interval.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>AVG-R</th>
<th>AVG-P</th>
<th>AVG-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMI</td>
<td>0.42288</td>
<td>0.48837</td>
<td>0.4442</td>
</tr>
<tr>
<td>SWARM</td>
<td>0.43028</td>
<td>0.47741</td>
<td>0.44669</td>
</tr>
<tr>
<td>Integrated_SWARM_MMI</td>
<td>0.42678</td>
<td>0.49368</td>
<td>0.44848</td>
</tr>
<tr>
<td>Msword</td>
<td>0.39306</td>
<td>0.48487</td>
<td>0.42477</td>
</tr>
<tr>
<td>H2-H1</td>
<td>0.49657</td>
<td>0.49613</td>
<td>0.49605</td>
</tr>
</tbody>
</table>

Table 2: MMI, swarm, integrated swarm_MMI, msword summarizer and H2-H1 comparison: average recall, average precision and average f-measure using rouge-2 at the 95%-confidence interval.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>AVG-R</th>
<th>AVG-P</th>
<th>AVG-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMI</td>
<td>0.18213</td>
<td>0.22242</td>
<td>0.19504</td>
</tr>
<tr>
<td>SWARM</td>
<td>0.18828</td>
<td>0.21622</td>
<td>0.19776</td>
</tr>
<tr>
<td>Integrated_SWARM_MMI</td>
<td>0.18629</td>
<td>0.22817</td>
<td>0.19971</td>
</tr>
<tr>
<td>Msword</td>
<td>0.16325</td>
<td>0.21066</td>
<td>0.17947</td>
</tr>
<tr>
<td>H2-H1</td>
<td>0.20957</td>
<td>0.2094</td>
<td>0.20938</td>
</tr>
</tbody>
</table>

Table 3: MMI, swarm, integrated swarm_MMI, msword summarizer and H2-H1 comparison: average recall, average precision and average f-measure using rouge-4 at the 95%-confidence interval.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>AVG-R</th>
<th>AVG-P</th>
<th>AVG-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMI</td>
<td>0.18213</td>
<td>0.22242</td>
<td>0.19504</td>
</tr>
<tr>
<td>SWARM</td>
<td>0.18828</td>
<td>0.21622</td>
<td>0.19776</td>
</tr>
<tr>
<td>Integrated_SWARM_MMI</td>
<td>0.18629</td>
<td>0.22817</td>
<td>0.19971</td>
</tr>
<tr>
<td>Msword</td>
<td>0.16325</td>
<td>0.21066</td>
<td>0.17947</td>
</tr>
<tr>
<td>H2-H1</td>
<td>0.20957</td>
<td>0.2094</td>
<td>0.20938</td>
</tr>
</tbody>
</table>
8. REFERENCES


