Pasokh: A Standard Corpus for the Evaluation of Persian Text Summarizers

Behdad Behmadi Moghaddas, Mohsen Kahani, Seyyed Ahmad Toosi, Asef Pourmasoumi, Ahmad Estiri
Web Technology Laboratory
Ferdowsi University of Mashhad
Mashhad, Iran
Behdad.behmadi@stu.um.ac.ir, Kahani@um.ac.ir, Ahmad.toosi@alumni.um.ac.ir, Asef.pms@gmail.com
Ahmad.estiri@alumni.um.ac.ir

Abstract—**The increasingly vast amount of information, particularly on the Web, has resulted in a profound need for automatic summarization systems. The systems, in turn, need to be evaluated in terms of how desirably they can retrieve information. The evaluation is done by comparing the machine summaries against a standard reference corpus containing a reasonably large number of text sources and the summaries that human beings have made out of them.**

Due to the lack of such a standard corpus for Persian, the summarizers that were developed used to be evaluated against the small corpora constructed by the developers of the proposed systems. This made the systems non-comparable. Thus, Pasokh was constructed as a standard large enough reference corpus. It took over 2000 man-hours of work.

Keywords-computational processing of Persian;single-document automatic summarization;multi-document automatic summarization; evaluation of automatic summarization;evaluation corpus

I. Introduction

The increasingly vast amount of information, particularly on the Web, has posed the challenge of accessing the gist of huge information in short time. Automatic summarization is the solution to the challenge. In [1], for instance, a method is proposed to automatically output the content of Twitter blogs in summarized form. In technical terms, summarization is defined as reducing the source text to a more concise version, with a computer program, such that the main points of the original document are retained [2].

Automatic summarization systems are classified from various points of view. Two major classifications are based on the following distinctions: between single-document and multi-document types of summarization, between extractive and abstractive types of summarization [3].

It is of prime importance to have an objective measure to evaluate the performance of summarization systems. As the ultimate goal of automatic summarization, of any type, is to produce summaries that are similar to human-generated summaries in quality, it is enormously useful to evaluate the quality of the output of summarization systems against human-generated summaries. The exploitation of such an evaluation technique requires a standard set containing source texts and their corresponding human summaries [4]. Such a corpus serves as the reference of comparison for the evaluation of the performance of different summarization systems.

Pasokh[1] is a dataset consisting of a large number of Persian news documents on various topics. It contains the human-generated summaries of the documents in the forms of single-document, multi-document, extractive, and abstractive summaries. The dataset was produced in such a way to be a sufficiently large representative corpus for the evaluation of Persian text summarizers.[2]

The paper is structured as follows: the next section is a review of the related works. Section III is the detailed description of the method adopted in constructing Pasokh corpus. Section IV concerns the assessment of the produced corpus. Finally, Section V involves the concluding remarks and mention of the future work.

II. A Brief Survey of the Related Products

As noted, a challenge in automatic text summarization is to evaluate the performance of the proposed systems. Precise evaluation requires an appropriate standard dataset to serve as the reference corpus. BBC, CNN, TREC, CAST corpus, DUC corpus, among others, are a number of the major datasets produced thus far. DUC (Document Understanding Conference) datasets [5-7] have been widely used. Below, there is a brief introduction to them.

A. DUC standard datasets for English text

Since 2001, NIST (National Institute of Standards and Technology) started releasing DUC datasets concerning automatic text summarization and seven datasets (DUC2001 to DUC2007) were presented. Each dataset was published with a particular purpose in mind. The main goal of the conference was to help evaluate the techniques of automatic text summarization and examine the methods for the evaluation of summarization techniques. DUC2001 to DUC2004 were produced for the

[1] Pasokh translates as “Answer”. In the original Persian: Peykare-ye (A)estândârd-e Sâmânehâ-ye (O is added for ease of pronunciation) KHolâse-sâz

[2] Accessible from “ijaz.um.ac.ir”, Pasokh is currently in its 2013 version.

978-1-4799-2093-8/13/$31.00 ©2013 IEEE

evaluation of single- and multi-document summarizations. DUC2005 to DUC2007 were produced for multi-document summarization only.

DUC2007 involves 45 topics in total, each incorporating 25 documents. 10 members of NIST were assigned the task of generating the human summaries. The abstractive summaries for each topic were created by four persons who were randomly chosen for doing each part of the task. The 32 summarization systems involved in the project automatically generated summaries for every topic. Therefore, the machine-generated summaries could be compared to the human-generated ones using the ROUGE tool and the systems could be ranked based on the results.

B. *Persian datasets*

No standard dataset has been presented for the evaluation of Persian automatic text summarization thus far. The reason is to be sought in the time and funds that the process of constructing such a corpus consumes, besides requiring a well-trained team for producing the summaries. The lack of an appropriate corpus is an important reason behind the scant research on Persian automatic summarization. Even the existing pieces of research cannot be evaluated and compared to one another. Pasokh reference dataset has been constructed following the latest international standards to resolve this problem.

III. 3. The Proposed Method

The process of building the Pasokh corpus is divided into the two main phases of constructing the single-document summarization corpus and the multi-document one. These are explained separately below. The specifications for the datasets produced for single- and multi-document summarization are shown in Tables I and II, respectively.

TABLE I. THE SPECIFICATIONS FOR THE DATASET PRODUCED FOR SINGLE-DOCUMENT SUMMARIZATION

Item	Item Count
Documents in the dataset	100
News genres	6
News agencies	7
Extractive summaries per document	5
Abstractive summaries per document	5

TABLE II. THE SPECIFICATIONS FOR THE DATASET PRODUCED FOR MULTI-DOCUMENT SUMMARIZATION

Item	Item Count
Topics	50
Documents per topic	20
News agencies	7
Extractive summaries per topic	5
Abstractive summaries per topic	5
Compression rate	30%

A. *The creation of single-document reference summaries*

The input of a single-document summarization system is only one document [3]. This kind of summarization is far less complex than multi-document summarization; the reason is that there is only one document to be summarized in this mode. The document probably discusses one single story in a cohesive manner and lacks contradictory information [8-9]. Multi-document summarization, on the other hand, involves a considerable number of documents and covering the key information in all of the documents is a complicated task.

The source documents used in constructing the single-document dataset were chosen from 7 famed news agencies that cover news from different perspectives. 100 documents of varying length were selected from the news agencies (mentioned in Table III). Table IV shows the categorization of news genres.

TABLE III. THE NEWS AGENCIES USED IN CHOOSING SOURCE DOCUMENTS FOR THE SINGLE-DOCUMENT CORPUS

No.	News	Web Address	The
1	Tabnak	http://www.tabnak.ir	TAB
2	Press TV	http://www.presstv.ir	PTV
3	Fars	http://www.farsnews.com	FAR
4	Irna	http://irna.ir	IRN
5	Hamshahri	http://www.hamshahrionline.ir	HAM
6	Alef	http://www.alef.ir	ALF
7	Jam-e-Jam	http://www.jamejamonline.ir	JAM

TABLE IV. THE GENRES OF THE SOURCE DOCUMENTS USED FOR THE SINGLE-DOCUMENT CORPUS

No.	Genre	The Abbreviated Name
1	Economic	EC
2	Cultural	CU
3	Social	SO
4	Political	PO
5	Sports	SP
6	Scientific	SC

B. *The organizing of the source documents used in single-document summarization*

The below convention for naming was used to facilitate access to the source documents:

<The abbreviated name of the news agency>.<The abbreviated name of the genre>.<Publication date>.<Document number>

Figure 1. The convention for naming single-document source texts

Thus, the document named PTV.PO.13901228.068.xml has been downloaded from Press TV News Agency, is categorized under political genre, has been published on the 28th of Esfand (the 12th month), the year 1390 (in Solar Hijri Calendar) and is given the number 068. The documents were released in xml format as shown in Figure 2.

978-1-4799-2093-8/13/$31.00 ©2013 IEEE

```
<?xml version="1.0" encoding="utf-8" ?>
<DOC>
<DOCNO>document number</DOCNO>
<DATE-TIME>publication date</DATE-TIME>
<CATEGORY>genre</CATEGORY>
<TITLE>document title</TITLE>
<SUMMARY>news summary (given in the news website</SUMMARY>
<TEXT>the complete news text</TEXT>
</DOC>
```

Figure 2. XMLstructure of the source documents in single-document summarization

C. The process of building the single-document reference corpora

10 male and female undergraduate students were employed for the corpus creation. The task of creating summaries for each source document was assigned to 5 individuals to lower the effect of the personal tendencies on the output. Machine summaries, therefore, will be compared to the average of five human summaries.

For each summarization task, an extractive and an abstractive summary, 3 to 7 sentences long, were selected/created. The summaries had to represent the central content of the source text. The summarization team was trained to follow the below guidelines:

1) Avoid repetition and redundancy while generating summaries

2) The summary should be in concordance with the key points of the original text

3) The summary should hold appropriate readability, particularly in abstractive form

4) The summary should have cohesive content

5) The summary length should not exceed the set compression rate

In total, 1000 reference summaries were produced for 100 news articles, 500 summaries being abstractive and 500 ones extractive. The large size of the dataset was a source of difficulty in the management of the source documents and the produced summaries. In order to overcome the problem, a program called "Kholâse-yâr" (roughly translating as "summarization aid") was developed; it can be used to produce any kind of human summary.

D. The creation of multi-document reference summaries

In multi-document summarization, multiple documents are fed into the system. This kind of summarization is closely related to question-answering systems and question-biased summarization [10]. This kind of summarization involves multiple separate source documents on a shared theme but from various viewpoints. Consider the theme of "The Global Challenge of Water Scarcity", for instance. "Water Shortage in Iran" and "Water Shortage in Pakistan" are the titles of two potential news articles, among others, that can be categorized under this theme.

Multi-document summarization poses greater challenges than the single-document mode. The major ones are as follows [11]:

1) The original documents deal with the same topic from different, occasionally contradictory, viewpoints; therefore, it is difficult to create a summary of high readability.

2) As multiple independent documents are involved and all discuss the same theme, redundant or overlapping information from multiple sources are likely to appear in the summary.

3) It requires enormous care to extract all the different ideas existing in the source documents and include in the summary as many of the most important ones as possible.

The selection of a number of themes is an initial stage in the process of building the multi-document database. 50 topics in various areas (sporting events, political news, etc.) were chosen some of which are listed in Table V. All the news had been published in the years 1389 through 1391 (in the SH Calendar).

TABLE V. A NUMBER OF THE THEMES SELECTED FOR THE MULTI-DOCUMENT CORPUS

Theme No.	Theme
D91A01	Behdad Salimi
D91A02	Earthquake in the east of Turkey
D91A03	Political unrest in Bahrain
D91A04	Political unrest in Syria
D91A05	The regulations of military service

For each topic, 20 documents were gathered from famous Iranian news agencies of differing views. In total, 1000 documents were collected for multi-document summarization corpus construction. Then they were distributed among the members of the summarization team. For every topic, 5 extractive and 5 abstractive multi-document summaries were created. In total, 500 multi-document summaries were produced that can be used to evaluate systems for multi-document Persian text summarization.

E. Kholâse-yâr software system

As stated above, the great number of documents and the associated summaries makes it difficult to handle the processes of content management and organization of the tasks done by the human summarizers. Thus, Kholâse-yâr system was designed and developed, in C# and under .Net Framework 3.5, as a facilitating tool for the human summarizers and to increase the speed and precision of the corpus creation process.

The source news articles, in xml format, are the input to the system. The program user interface consists of the two sections of single- and multi-document summarization. Entering each section, each human summarizer, uniquely identified by a user name, can see the source text(s) and produce either extractive or abstractive summary for the text(s). The section related to single-document summarization includes three separate fields, i.e. Reasons, Abstractive, and Extractive, to be filled in with the evidence supporting the selections, the abstractive and extractive summaries, respectively. The summaries produced by every human summarizer are gathered in a database as the output. Thus, the collection of the summaries produced by all the users for each given source text is readily accessible from the database. The collection of the source texts together with their associated summaries constitutes the single-document corpus. The naming convention of the files makes it unlikely for the different reference summaries created for a single file to be mistaken for one another. The naming convention is illustrated in Figure 3.

<user name>.<single-document or multi-document>.<abstractive or extractive>.<the full name of the source document>

Figure 3. The naming convention for reference summaries

If the summary type is abstractive, the character A and if it is extractive, the character E is put into the file name. The character S abbreviates single-document and M abbreviates multi-document summarization. Therefore, JAM.SO.13901203.081.E.S.F.F.C states that the JAM.SO.13901203.081 document, extracted from Jam-e-Jam news agency on the 28th of Esfand (the 12th month), the year 1390 (in SH Calendar), was summarized by the user named F.F.C in extractive type and single-document mode.

The structure is similar for multi-document summarization. The difference is that, in naming the output files, the document name is replaced with the topic name. The reason is that, in multi-document summarization, only one summary is generated for each topic consisting of multiple files.

IV. Assessment

A vital step after the corpora have been created is to ensure the quality of the created summaries. In general, at least two characteristics should be examined while assessing the reference summaries [12]: the compression rate (how short the summary is in comparison to the original text) and the preservation rate (how much information is preserved).

The assessment techniques are divided into intrinsic and extrinsic ones [13]. The former assesses the summary in itself regardless of its purpose. The latter, on the other hand, focuses on the end-user. Extrinsic assessment techniques tend to be employed in applications such as information retrieval and question-answering where the relevance of the summary to the original text is tested [12].

From the intrinsic techniques, the ones concerning text cohesion and informativeness, and from the extrinsic ones, the question games were exploited to measure the quality of the generated summaries in Pasokh corpus. The following is a description of the methods used and the results obtained.

Cohesion: the summarized texts generated by extraction-based methods (the copy and paste operations on expressions, sentences, and paragraphs) occasionally suffer from semantic irrelevance between sentences. A method to measure the summary cohesion is rating sentences in terms of the degree of cohesion they exhibit. The total degree of cohesion can then be compared to the score of the reference summaries, the score of the source sentences, or the score of other summarization systems.

Summary informativeness: A way to measure the informativeness of a generated summary is comparing it to the source text or a reference summary to see how much of source information is retained in the summary.

Question game: this aims to test the reader comprehension of the summary and the power of the summary to reproduce the key ideas of the source text. This is done through following stages; first, the experimenter reads the source texts and highlights the key parts. Then questions regarding the key parts are asked. The assessor answers the questions three times: once without seeing any text (baseline 1), another time after they have seen a generated summary, and finally after they have seen the source text (baseline 2). A summary that accurately reproduces the central points of the text should enable the assessor to answer most questions (by being closer to baseline 2 than baseline 1) [14]. Based on how correctly the questions have been answered, the summary under assessment will be given a score.

Ten students of linguistics scored the generated summaries by the above-mentioned methods. In the case of assessing the cohesion and informativeness of the summaries, every assessor rated each summary using the scale below:

1:very weak, 2:weak, 3:acceptable, 4:good, 5:very good

Dividing the resultant scores by 5, we were left with some number between 0 and 1. Averaging the ten scores given by all the assessors yielded a more precise score. The average scores, in percentage terms, for the whole corpus are shown in table VI.

TABLE VI. Results of the assessment of Pasokh corpus

	Cohesion	**Informativeness**	**Question game**
Single-document	83.56%	88.19%	70.08%
Multi-document	60.24%	63.51%	73.33%

In average, single-document reference summaries generally score higher than multi-document ones. A reason for this difference is the varying writing styles of multiple source documents that contribute to the content of a single reference summary.

Those summaries that scored lower than the acceptable limit will be identified for the modification of the corpus in the future revisions.

V. Conclusion

Considering the vast amount of existing written information and the shortage of time, optimal summarization of the myriad of books, articles, news reports, etc. on the Web is a major concern of researchers. Of no less importance is the evaluation of summarization systems. Huge-size datasets consisting of reference human-generated summaries are a key part of evaluation systems. The quality of automatic summarization systems is measured in comparison to the reference summaries.

The lack of a standard corpus for the evaluation of summarization systems for the Persian language was profound. Following internationally recognized standard procedures, Pasokh corpus was constructed to fulfill the mentioned need of Persian NLP. It took over 2000 man-hours to produce the dataset. It contains a large number of reference summaries for the evaluation of single- and multi-document summarization systems.

ACKNOWLEDGMENT

The authors are grateful for the aid they received from Dr. Nader Jahangiri and the students of the Department of Linguistics at Ferdowsi University of Mashhad. They also acknowledge the members of Web Technology Laboratory at Ferdowsi University of Mashhad.

REFERENCES

[1] A. Sood, "Towards summarization of written text conversations," International Institute of Information Technology, India, 2013.

[2] I. Mani and M. T. Maybury, Advances in automatic text summarization: the MIT Press, 1999.

[3] C.Y. Lin and E. Hovy, "From single to multi-document summarization: A prototype system and its evaluation," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 457-464.

[4] H. Saggion, et al., "Multilingual summarization evaluation without human models," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 1059-1067.

[5] S. Ye, et al., "NUS at DUC 2005: Understanding documents via concept links," in Proceedings of Document Understanding Conferences, 2005.

[6] B. Schiffman, et al., "Experiments in multidocument summarization," in Proceedings of the second international conference on Human Language Technology Research, 2002, pp. 52-58.

[7] B. Favre, et al., "The LIA-Thales summarization system at DUC-2006," in Proceedings of the 2006 DUC Workshop, 2006.

[8] R. Mihalcea and P. Tarau, "A language independent algorithm for single and multiple document summarization," in Proceedings of IJCNLP, 2005.

[9] K. M. Svore, et al., "Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources," in EMNLP-CoNLL, 2007, pp. 448-457.

[10] T. Hirao, et al., "An extrinsic evaluation for question-biased text summarization on QA tasks," in Proc. of the NAACL 2001 Workshop on Automatic Summarization, 2001, pp. 61-68.

[11] X. Wan, et al., "Manifold-Ranking Based Topic-Focused Multi-Document Summarization," in IJCAI, 2007, pp. 2903-2908.

[12] E. Hovy and C.-Y. Lin, "Automated text summarization and the SUMMARIST system," in Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998, 1998, pp. 197-214.

[13] M. Hassel, "Evaluation of automatic text summarizaiton: a practical implementation," Karlstad University, 2004.

[14] I. Mani, et al., "The TIPSTER SUMMAC text summarization evaluation," in Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, 1999, pp. 77-85.

978-1-4799-2093-8/13/$31.00 ©2013 IEEE