Department of Culture and Communication

Master's Program

Language and Culture in Europe

Using the Pyramid Method for generating gold standards for Persian texts

Masoud Jahan Nia

Language and Culture in Europe

Course: Linguistics Specialization

Spring 2013

Supervisor: Arne Jönsson

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 Introduction

We live in a digital world and we are exposed to so much information every day. But we don't have enough time to study long texts and it is very important in the world of technology to get an idea about texts by reading a very short form of them. Sometimes we just want to get the idea of a text and sometimes we don't have much time to read a full text. A general concern is that we have to read a lot of information in a very short time and at the same time we don't want to waste much time on reading long texts. One solution to this is text summarization.

There are different ways of text summarization. Abstract summary vs. extract summary is a common distinction. An extract summary is formed by taking out phrases, complete sentences or paragraphs from texts. Sentence extraction, which is the main focus of this study, is one of the common methods in creating summaries. On the other hand, an abstract summary is created by breaking down the text and rewriting it to a shorter one.

Many researches have already done ways of automatically summarizing texts, but this field of study still requires improvements. These kinds of improvements need to be evaluated, which is done by comparing them with the summaries generated by humans. For instance, Persian is a less resourceful language in Automatic Summarization and further studies need to be done (Shamsfard, 2011).

There is always a demand for standard texts for evaluating automatically generated summaries. "A gold standard is often a compilation of different human created summarizations which is then put together into one." (Carlsson and Jönsson, 2010:1) Gold standards are defined based on summaries created by humans and mostly used in different Computational studies on texts.

In this study we made use of the Pyramid method on extraction based summaries. We asked human summarizers to extract full sentences from the texts. The sentences which were of the highest frequency amongst other sentences were considered to make a gold standard.

Here, human summarizers were asked to extract 30%, 50% and 80% of sentences from the texts. The first purpose was to see how the Pyramid method works for creating gold standards based on different scales of summarization. It was also important for us to know the Pyramid method's pattern for Persian summarized texts of different scales. 30% summaries acts as indicative summaries and they give enough information for readers to understand the theme of the texts and then decide whether to read the texts fully or not. 50% summaries provide the reader with the most important sentences from the texts. 80% summaries are good closeness to the original text, but in shorter form. They give enough information to the readers and can be a very proximate summary to the original text. We defined the length of the extracts with the above mentioned scales, created pyramids according to them and then evaluated the results.

It is impossible to create absolute gold standards that all agree upon, but when we generate a gold standard where most of human summarizers agree upon its content, then it can be used for evaluating automatically generated texts. They can distinct good summaries from bad ones. They are nearer to the majority of people's way of thinking toward text summarization.

This paper aims to analyze the following issues in Persian texts:

- Generating gold standard
- Evaluation of gold standards
- Investigating the applicability of the Pyramid method for creating gold standards
- Analyzing the differences between two types of texts (social and scientific) when generating gold standards

2 Theoretical background

In extraction-based methods, texts are divided into phrases, sentences or paragraphs and then human summarizers or computer software take out a number of them. The phrases, sentences or paragraphs in new texts are ranked according to a scale and then the top highest scale ones are selected. One factor to be considered when creating extraction-based summaries is discourse coherence. It is the degree to which sentences are related to other sentences in a text. It is a very important factor for readers to understand the texts and also in distinguishing good summaries.

Hassel and Mazdak (2004: 35) explained that "the main task in text summarization using extraction method is to find an accurate balance between the coherence and preserving the important information in the text." Considering Figure 1, if the focus of text summarization is on important information, then in an extreme case it extracts lines A, B1 and C2 from the text and if it focuses on cohesion then it would extract lines A, B and C. Neither of cases would result in a good summarization (Hassel and Mazdak, 2004).

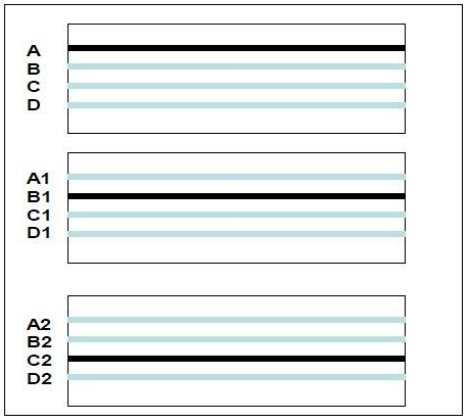

Figure 1: Cohesion and Important Info (Hassel and Mazdak, 2004)

2.1 Evaluation of automatic summaries

It is very important for researchers to find out the degree of efficiency and usefulness of automatic summaries and it is not often an easy task. There are two ways of evaluating summaries, we can either ask humans to read the texts and evaluate them based on their ideas or we can create some gold standards and then compare automatic summaries to those gold standards and find the efficiency of automatic summaries.

There are different metrics for evaluation of automatically generated summaries. ROUGE is the most common metric of content selection quality for researchers. The features that made ROUGE so popular are being *cheap* and *fast*. Many researchers like to apply ROUGE for automatic evaluation of summarization along with a manual evaluation of content. According to Nenkova and McKeown, ROUGE is based on the computation of n-gram overlap between a summary and a set of models (Nenkova and McKeown 2011: 204). Different parameters like word stemming, n-gram size and stop word removal are applied in ROUGE and made it applicable in different settings.

2.2 Creation of gold standards

The gold standards are usually a compilation of some summaries created by humans and based on humans' agreement on phrases, sentences and paragraphs to be included in the extracts. The phrases, sentences or paragraphs, which human summarizers mostly use in their extracts, get higher weights and form the optimal summaries.

The KTH eXtract Corpus (KTHxc) contains a number of original full texts and several man-made extracts (Hassel and Dalianis, 2005). In KTHxc full extract units (most often sentences) are added to the summaries based on how many times a unit has been included in different summaries. One difference between the method of sentence extraction in our project and what Hassel and Dalianis did in KTHxc is that in their case human informants were allowed to extract texts as short as 5 percent and as long as 60 percent of a text, while the mean length of the extracts were between 31 and 34 percent. In this research, we asked informants to extract exactly 30%, 50% and 80% sentences out of the texts.

It is very difficult to create a standard text which everyone accepts. "Our approach acknowledges the fact that no single best model summary exists, and takes this as a foundation rather than an obstacle." (Nenkova and Passonneau: 1) It is not also possible to create gold standards based on just one summary, as there is a bias toward selection specific parts of the original text. According to Nenkova (2006) five human summaries are enough in creation of gold standards. Carlsson (2010) supported this idea by applying it in creating gold standards for Swedish texts.

2.3 The Pyramid method

According to Nenkova, “a pyramid is a representation of a gold standard summary for an input set of documents.” (Nenkova et al., 2007:3) It is a way of generating commonly agreed upon content units. Nenkova et al. (2007) stated that “The pyramid evaluation method has been developed for reliable and diagnostic assessment of content selection quality in summarization and has been used in several large scale evaluations”. Unlike other ways of creating gold standards, the Pyramid method is a quantitative representation of the agreement that human summary writers have on the parts of a text that is important for each of them. The units of meaning are weighted according to the number of appearance in the extracts. The higher weight of units shows their higher frequency in the extracts and also the importance of units compared to other units.

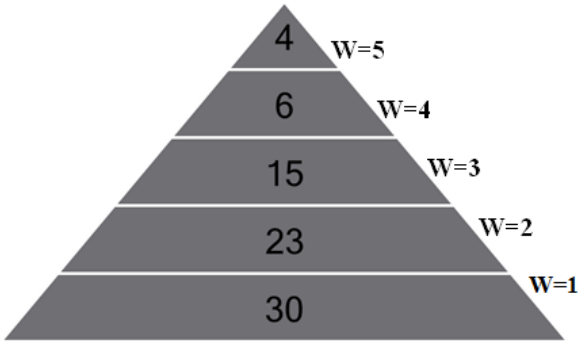

Figure 2: An example of pyramid based on a 30% summary
(the numbers show the total number of sentences at each tier)

Figure 2 shows a pyramid of order 5 as it includes 5 tiers. The sentences in tier 5 are expressed by all five human summarizers; in tier 4 the sentences are expressed by 4 summarizers and so on. The number of sentences in the top tiers is usually fewer than the lower tiers. In order to create an optimal summary out of pyramids, we start from the top tier and add all sentences into an optimal summary. We continue to the next tier and add the sentences of that tier to the summary. If the length of the optimal summary permits, we add sentences from lower tiers. As a general role for maximizing the value of optimal summary, the sentences from tier (n-1) should not be added until all sentences from tier (n) are expressed in the summary.

3 Methodology

As the main goal of this research is to generate gold standards for Persian texts, we gathered a compilation of 10 texts, five of them were scientific and five were social texts. By social we mean every text which relates to the society and what is happening every day to the people. It is a very broad field and can include so many issues in the society. We also considered some criteria for the selection of the texts. The number of words in the texts was supposed to be between 1000 and 1400. The texts were carefully selected from Persian articles published in some popular newspapers and magazines. The texts included different areas of science like astronomy, ecology, oceanography and healthcare. Moreover, in case of social texts they covered social relations, history and common concerns in the society. (See appendix 3 for more information about the texts) Table 1 shows a summary of the texts used in this research.

Table 1: The distribution of texts, the number of sentences and words in each text

Texts	Type of the Text	Number of Sentences in Texts	Number of Words in Texts	Number of 30% summaries	Number of 50% summaries	Number of 80% summaries
T1	Scientific	118	1057	5	5	5
T2	Scientific	117	1179	5	5	5
T3	Scientific	80	1157	5	5	5
T4	Scientific	110	1294	5	5	5
T5	Scientific	114	1228	5	5	5
T6	Social	89	1191	5	5	5
T7	Social	119	1312	5	5	5
T8	Social	111	1017	5	5	5
T9	Social	104	1241	5	5	5
T10	Social	147	1093	5	5	5

The number of words in the articles was between 1017 and 1312 while the number of sentences was between 80 and 147. The average number of words in these texts was 1176; 1183 for scientific texts and 1170 for social texts. The average number of sentences in the ten texts was 110.9. This scale was 107 for scientific texts and it was 114 for social ones. There was not that much difference regarding these factors for these two types of texts. The reason

for that was to have articles with sufficient number of words and sentences, in order to provide more validity to our gold standards.

Nenkova's idea (2006) about the number of human summarizers was applied, i.e., five participants were carefully selected to read the texts and summarize them. They were supposed to summarize the texts into 30, 50 and 80 percent of the full texts. They were free to start with the text and the scale they preferred. There was no supervision in this regard. Generally, when asking people to create a summary of a text, they will include different sentences in the extracts based on their consideration of the most important sentences. In this study, we split up the texts into sentences and then asked human summarizers to create extracts based on selecting the most important and relevant sentences.

In this regard, the study faced some problems. For instance, one of the problems was that there is not a very specific and clear definition for *sentence* in Persian (Mohaghegh, 2010). Finding the number of sentences and separating them was very important in this research. The most problematic part was dividing complex sentences. In this regard, we tried to divide the complex sentences into simple sentences wherever it was possible. "در فاصله‌ای که پیام منتشر شده تا زمانی که ما پیام را دریافت می‌کنیم" and "نیکولا تسلا، یکی از اولین کسانی بود که اواخر قرن ۱۹ میلادی پیشنهاد داد" were two complex sentences from Text 1 and in this research they were divided in two simple sentences. There were other complex sentences which were not divided into their components. As an example "ببینیدآیا آنها تمایلی به برقراری ارتباط دارند یا نه." is a complex sentence that was left intact.

In order to make the texts simpler to the readers, some of the sentences were modified before presenting to human summarizers. As an example in Text 10, the sentence "که شامل ۲بخش مهم حرف زدن و گوش دادن می‌شود یا به عبارتی کم سؤال کردن و خوب گوش دادن." were modified to "که شامل ۲بخش مهم حرف زدن و گوش دادن می‌شود.". Another issue that was considered in this research was the infinitive "توانستن". In the sentences like "پس می‌توان مدعی شد" and "از این رو می‌توان پیش‌بینی کرد" the infinitive "توانستن" is not considered as a sentence and both expressions are regarded as one sentence.

At the end of this process, we gathered 150 texts from 5 human summarizers in different scales of 30%, 50% and 80%. The next step was to extract information and finally get gold standards out of texts. The method we used in this research was a variant of the Pyramid method. The Pyramid method is based on what most of human summarizers agree upon. For each summarization unit, in this case *sentence*, a weight was defined according to the number of the summaries it appeared in. If a unit expressed 3 out of 5 times (equal to the number of summarizers) then the weight of that unit was considered 3. There were different partitions in every pyramid and in each partition the units had the same weight. The number of partitions (tiers) in a pyramid depended upon the number of human summarizers. As we had five human summarizers in our case, then the number of tiers was five, so it is called a pyramid of order 5.

The optimal summary content contains sentences from the top tier and if the number of sentences are not enough, then it includes the sentences from the next tier and then it continues till an optimal summary is achieved. When a sentence weight is higher, it shows the higher importance of it; while when its weight is low, it means that it is not such an important sentence for summarization. At the end of this process maximally informative summaries, gold standards, are formed from top rated (high weighted) sentences for every article. Extracting sentences based upon the number of appearance in human extracts makes a content selection meaningful, stable and informative.

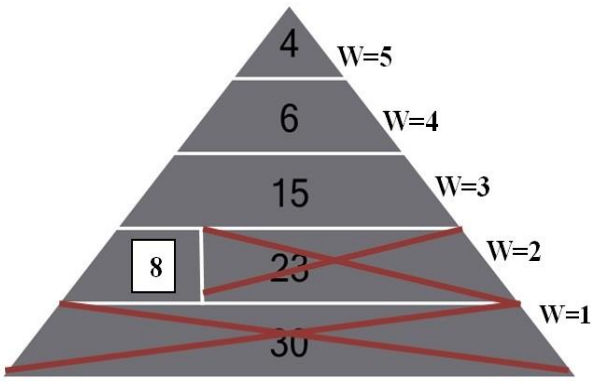

Figure 3: A 30% pyramid for Text 4

Another problem faced during the process of extracting gold standards from pyramids was because of having restrictions for sentences selection. As illustrated in Figure 3, Text 4 had 110 sentences (sentences with weight = 0 are not included in the Pyramid) and for making a 30% summary we needed 33 sentences. In this case, after adding sentences from the top three tiers, we still required 8 sentences for having a 30% scale gold standard for this text. The

problem appeared when we were deciding to include which sentences in the summary. The solution we applied was to add sentences based on their appearance in original texts. The sentences mentioned earlier in original texts had more priority than those appeared later. This idea had two assets. The first asset was that it resulted in more coherent texts as we added sentences from specific part of the texts. The next advantage was that we could easily apply this idea to all texts.

Based on 150 summaries done by human summarizers, 30 pyramids were generated. Then, according to the pyramids, 30 gold standards in 3 scales were created. The next step was to evaluate the gold standards and see how useful the Pyramid method was for evaluation of scientific and social texts in Persian.

Two groups of people were carefully selected for evaluation of the gold standards. The first group was 5 graduates from different fields of study with good understanding of Persian texts, while the second group was 5 Persian language teachers who studied and taught Persian for some years. We actually used a third category of evaluators, students, but we haven't included them in the thesis. They were ignorant about the questions and it seems that they couldn't distinct between the texts well. The other reason was that they couldn't recognize that Question 5 and 6 are negative questions. The standard deviation in the group of students was also high (SD= 1.63) compared to other groups.

They were asked to read the 30% gold standards and answer a questionnaire with 6 questions. The questionnaire was like what Carlsson (2010) used, but with slight changes. There was another question "The summary gives a good understanding of the content of the original document" in Carlsson's research which was deleted because it had meaning overlap with Question 4 "The summary gives a good idea on what is written in the original document." It was difficult to distinct between the two questions in Persian.

Table 2: Questionnaire – English Version

Question 1	The summary has a good length to give an idea on the content in the original text				
Disagree 1	2	3	4	5	6 Fully agree with
Question 2	The summary is experienced to be information rich,				
Disagree 1	2	3	4	5	6 Fully agree with
Question 3	The summary is experienced as strenuous to read.				
Disagree 1	2	3	4	5	6 Fully agree with
Question 4	The summary gives a good idea on what is written in the original document.				
Disagree 1	2	3	4	5	6 Fully agree with
Question 5	The summary is experienced as missing relevant information from the original document				
Disagree 1	2	3	4	5	6 Fully agree with
Question 6	The summary is experienced as a good complement to the original document.				
Disagree 1	2	3	4	5	6 Fully agree with

The questions (Table 2) embodied different ideas like the length of the gold standards, the level of difficulty, relevancy to original documents and richness of the gold standards. (See appendix 2 for the questionnaire in Persian) They were not asked to consider or disregard coherence in texts. They carefully read the texts and answered the questionnaire based on their own judgments about the texts.

4 Results

4.1 Analysis of Summaries

We developed ten gold standards for ten Persian texts. The first criterion to be measured was the degree of sentence overlap in these human summarized texts. The following function shows the way to calculate sentence overlap:

$$Sentence\ Overlap = \frac{Number\ of\ sentences\ included\ in\ any\ summary}{Total\ number\ of\ sentences}$$

Table 3: Sentence overlap distribution in texts

Text	Type of the text	Sentence overlap between 30% summaries
T1	Scientific	66.9%
T2	Scientific	69.2%
T3	Scientific	78.7%
T4	Scientific	70.9%
T5	Scientific	78.9%
AVG	Scientific	72.9%
T6	Social	66.2%
T7	Social	71.4%
T8	Social	65.7%
T9	Social	65.3%
T10	Social	76.1%
AVG	Social	68.94%
AVG	All texts	70.9%

According to Table 3, the average overlap between 30% summaries is 70.9%. It means that 70.9% of sentences were used by human summarizers at least one time in five extracts. There is not a big difference comparing scientific and social texts, but this overlap is marginally more for scientific texts. This agreement is 72.92% for scientific texts, whereas it is 68.94% for social ones. Table 3 depicts the sentence overlap which is between 65.3% for Text 9, that is a social text and 78.9% for Text 5 that is a scientific text.

The next criterion to be considered based on sentence extraction from documents is the number and the pattern of unique sentences in summaries. The first summarizer makes use of some sentences and they are considered as *unique* sentences, for the reason that they are

mentioned for the first time. Firstly, the number of *unique* sentences is equal to the number of sentences in the first summary. When adding the sentences from the second summary, we face some shared sentences and also some unique sentences, which are mentioned for the first time. The normal pattern of uniqueness of sentences can be illustrated by a descending line.

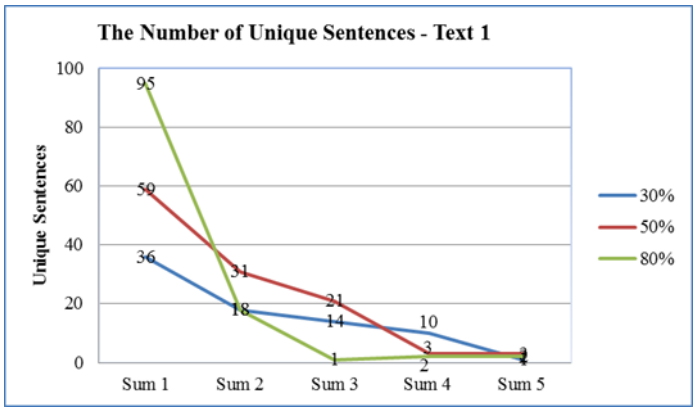

Figure 4: The pattern of unique sentences for Text 1

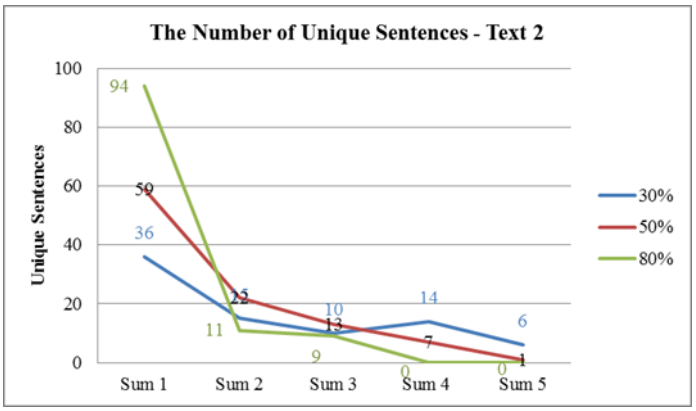

Figure 5: The pattern of unique sentences for Text 2

According to the summaries for Persian texts, this pattern does not always follow the normal one. There are some obstacles that affect this pattern. As illustrated in Figures 4 and 5, the number of unique sentences has a descending pattern. It always starts from a number equals to X% (X= 30%, 50% or 80% of the number of sentences from the ten texts in this research) of the summaries. Then, with a descending slope it shows the number of sentence uniqueness for the second extract and it continues to count for the uniqueness of all five summaries. For example, as illustrated in Figure 4, the first summariser makes his/her 80% extract with 95 sentences and the second extract brings only 18 *unique* sentences. There is also such a decline for the third extract, when it adds just one new unique sentence. But the fourth and fifth extracts add two new unique sentences. We can neglect this deviation. For 50% and 30% summaries, this pattern is more irregular; it is because of the fact that there are more choices for the extraction of sentences. As illustrated in Figure 5, the fourth summarizer adds 14 new unique sentences to his/her 30% extract, which is slightly higher than the third one. Text 2 has 117 sentences and the first three summaries have just extracted 61 sentences, so the fourth one can select up to 56 unique sentences. So it is not peculiar that it results in deviation of the general pattern.

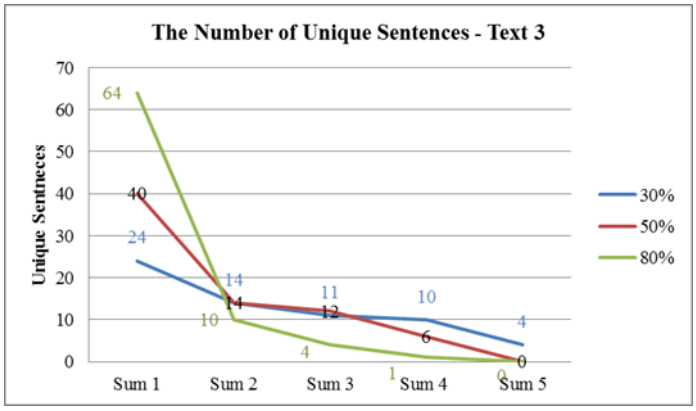

Figure 6: The pattern of unique sentences for Text 3

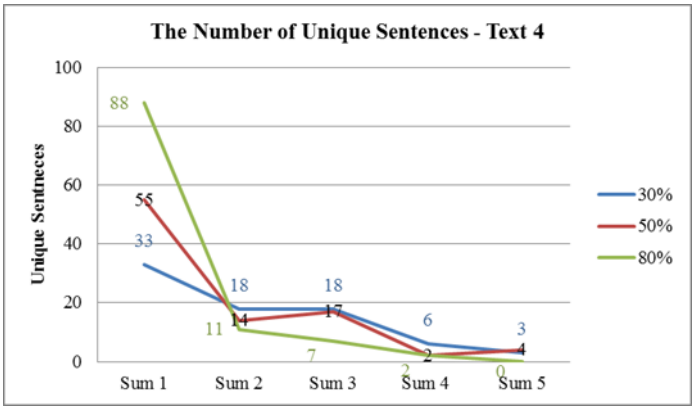

Figure 7: The pattern of unique sentences for Text 4

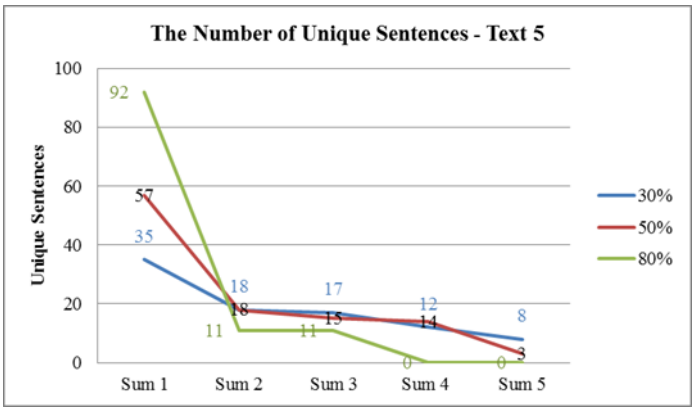

Figure 8: The pattern of unique sentences for Text 5

Moreover, there is also a change in our general pattern in Text 4 (Figure 7). The third summarizer can select unique sentences out of 41 sentences, which first and second summarizers have not included in their summaries. So, in this case we can see a little increase from 14 to 17. The increment between fourth and fifth summarisers can also be neglected. Figure 6 and 8 depict the pattern for Text 3 and Text 5 and both are in accordance with the general pattern.

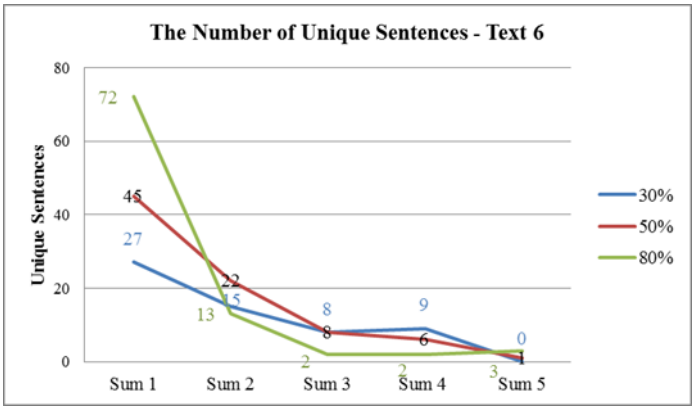

Figure 9: The pattern of unique sentences for Text 6

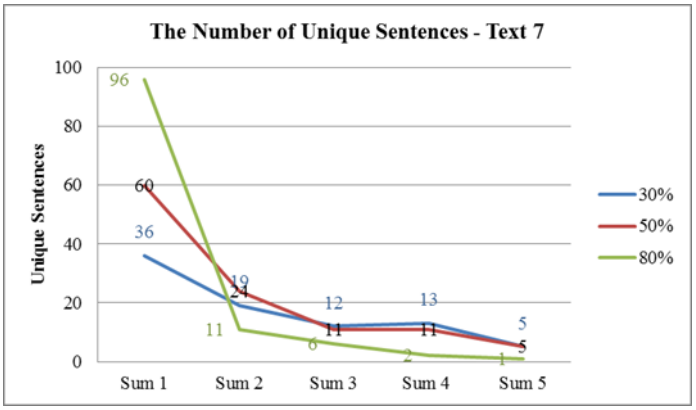

Figure 10: The pattern of unique sentences for Text 7

Furthermore, the pattern for social texts is similar to scientific texts (see Figure 9 - 13); therefore, there are not many differences between the types of the texts. Additionally, the figures show some deviations from the general pattern with social texts like scientific ones.

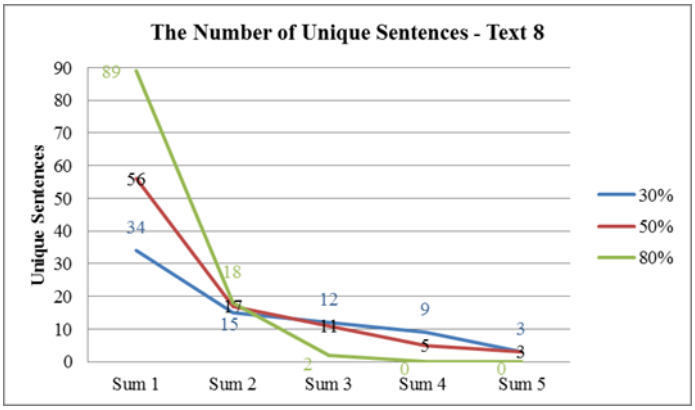

Figure 11: The pattern of unique sentences for Text 8

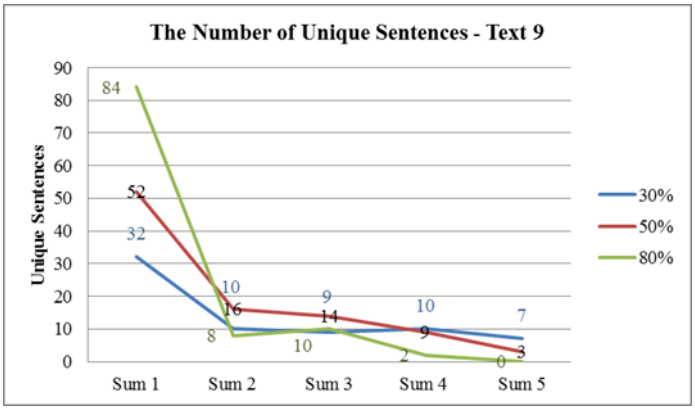

Figure 12: The pattern of unique sentences for Text 9

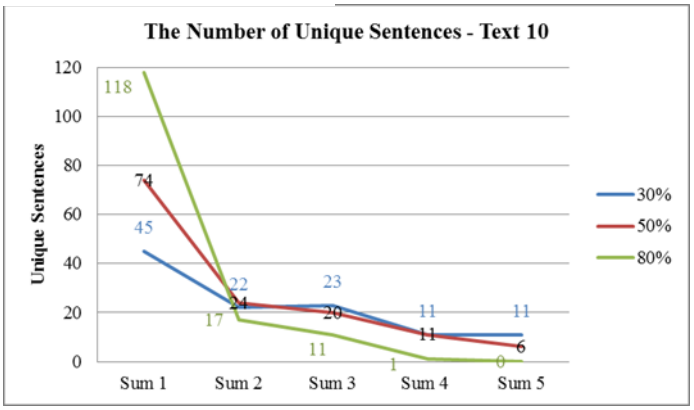

Figure 13: The pattern of unique sentences for Text 10

A general pattern that can be understood from the above figures is that the number of unique sentences is the highest in the first summary among all the texts and it is the lowest in the last one, except for a little deviation on 80% extraction of Text 4 which can be neglected. In general, social and scientific texts are not so different in this regard.

As for the Pyramids, the information is extracted and the first data to be analyzed is the pattern of sentence distribution among different tiers in every pyramid. The following tables show the distribution of sentences in 30%, 50% and 80% summaries.

Table 4: Distribution of sentences at each weight for 30% summaries (T1 to T5 are scientific texts, while T6 to T10 are social ones)

Texts	Sentence Weight 30% Pyramids						Number of Sentences included in the Pyramid	Weighted Sum	Mean Sentence Weight
	0	1	2	3	4	5			
T1	39	24	22	22	9	2	79	180	2.27
T2	36	27	25	20	2	7	81	180	2.22
T3	17	32	12	10	7	1	63	119	1.88
T4	32	30	23	15	6	4	78	165	2.11
T5	24	33	38	11	7	1	90	175	1.94
T6	30	13	26	11	8	1	59	131	2.20
T7	34	33	22	18	11	1	85	180	2.11
T8	38	25	16	19	9	4	73	170	2.32
T9	36	24	13	20	5	6	68	160	2.35
T10	35	46	35	18	10	3	112	225	2.00
AVG	32.1	28.7	23.2	16.4	7.4	3.0	78.8	168.5	2.14

In Table 4, the first column shows the texts, columns 2- 7 show the distribution of sentence weight in ten pyramids and columns 8-10 illustrate the number of sentences included in every pyramid, sum of the weight of sentences and the mean sentence weight in each pyramid. In the last three columns, the sentences with (weight = 0) are ignored. The following function shows how Weighted Sum is calculated:

$$\text{Weighted Sum} = \sum_{weight\,=\,1}^{5} \text{the number of sentences in each weight} \times \text{sentence } weight$$

Considering the last row, the general pattern of 30% summaries shows that the frequency of sentences is the highest in tier 1 and it is the lowest in tier 5. Meanwhile, there are some parts in the pyramids which do not follow this pattern. As an example, in Text 2 the number of sentences in tier 5 is equal to 7; while it is 2 in tier 4.

The last column of the above table shows the mean sentence weight for every summary. It shows the mean weight of all sentences in the summary with (weight>0). The following function shows how it is calculated:

$$Mean\ Sentence\ Weight = \frac{Weighted\ Sum}{Number\ of\ Sentences\ included\ in\ the\ pyramid}$$

The average mean sentence weight is 2.14 for all the texts. Text 9 has the highest mean sentence weight, while Text 3 has the lowest one. The average mean sentence weight is 2.08 for scientific texts and it is 2.19 for social texts. The mean sentence weight is higher in social texts than scientific ones.

Table 5: Distribution of sentences at each weight for 50% summaries (T1 to T5 are scientific texts, while T6 to T10 are social ones)

Texts	Sentence Weight 50% Pyramids						Number of Sentences included in the Pyramid	Weighted Sum	Mean Sentence Weight
	0	1	2	3	4	5			
T1	1	16	43	41	15	2	117	295	2.52
T2	15	24	19	22	18	19	102	295	2.89
T3	8	13	17	22	13	7	72	200	2.77
T4	18	14	19	29	14	16	92	275	2.98
T5	7	32	20	17	28	10	107	285	2.66
T6	7	15	23	22	12	10	82	225	2.74
T7	8	25	24	32	19	11	111	300	2.70
T8	19	20	14	23	12	23	92	280	3.04
T9	10	11	37	20	15	11	94	260	2.76
T10	12	25	31	36	40	3	135	370	2.74
AVG	10.5	19.5	24.7	26.4	18.6	11.2	100.4	278.5	2.78

There is not any specific pattern for 50% summaries. Table 5 shows that in most of the texts there are more sentences in tier 3 than other tiers in pyramids. The mean sentence weight is 2.78 for 50% summaries. The highest mean sentence weight is for Text 8, which is a social text and the lowest is for Text 1 that is a scientific text. The mean sentence weight is 2.76 for scientific texts, while it is equal to 2.79 in the case of social texts. Social texts are marginally better in this case. Considering the last row, we can infer that 50% pyramids generally have a non-linear pattern and the peak is 26.4 for sentences of weight 3. It means that in average, the number of sentences with (frequency=3) is more than other sentences in a pyramid.

Applying the Pyramid method for 50% summaries results in distortion of the pyramid. So, it can be concluded that the Pyramid method does not create a pyramid for 50% summaries.

Table 6: Distribution of sentences at each weight for 80% summaries (T1 to T5 are scientific texts, while T6 to T10 are social ones)

Texts	Sentence Weight 80% Pyramids						Number of Sentences included in the Pyramid	Weighted Sum	Mean Sentence Weight
	0	1	2	3	4	5			
T1	0	2	7	24	38	47	118	475	4.02
T2	3	3	10	19	20	62	114	470	4.12
T3	1	2	9	12	16	40	79	320	4.05
T4	2	5	6	15	32	50	108	440	4.07
T5	0	2	12	16	34	50	114	460	4.03
T6	0	6	4	14	24	41	89	357	4.01
T7	3	4	6	16	34	56	116	480	4.13
T8	2	0	8	19	38	44	109	445	4.08
T9	0	2	13	13	27	49	104	420	4.03
T10	0	1	9	27	60	50	147	590	4.01
AVG	1.1	2.7	8.4	17.5	32.3	48.9	109.8	445.7	4.05

Table 6 illustrates the preceding pattern for 80% summaries. This pattern is the other way around compared to 30% pyramids. According to the table, there are more sentences in tier 5 than any other tiers in pyramids. The mean sentence weight is 4.05 for all extracts. The highest mean sentence weight is 4.13 for Text 7, which is a social text and the lowest ratio is 4.01 for two texts of 6 and 10 which are again social texts. This scale is 4.05 for scientific texts and 4.07 for social texts. Similar to 30% and 50% summaries, the mean sentence weight of social texts of 80% summaries is better than scientific ones.

In 80% summaries, there are fewer sentences with lower weights and more sentences with higher weights. It is because of the fact that summarisers extract 80% of all sentences for making a summary and this leads to more coverage of sentences and we can find a distortion of the Pyramid method for 80% summaries.

Regarding all three tables, we can say that in general 30% summaries have more sentences in lower tiers and less number of sentences belongs to the higher tiers of the pyramid. 50% summaries do not have a general pattern and it is found that they have more sentences in central tiers. In 80% summaries, the frequency of sentences is higher in top tiers and less in low tiers. The mean sentence weight of all the texts raises by increasing the scale of

summarization. It grows from 2.14 for 30% summaries to 2.78 for 50% summarization and then to 4.05 for 80% summaries.

Considering the mean sentence weight for different summarization scales results in what Figure 14 shows. The average for 30% summaries is 2.14, while for 50% summaries it is 2.78 and finally 4.05 for 80% summaries. This figure depicts that the longer the summaries, the more consistent they are. But, it does not have a linear pattern. Despite, more consistency is expected for 50% summaries.

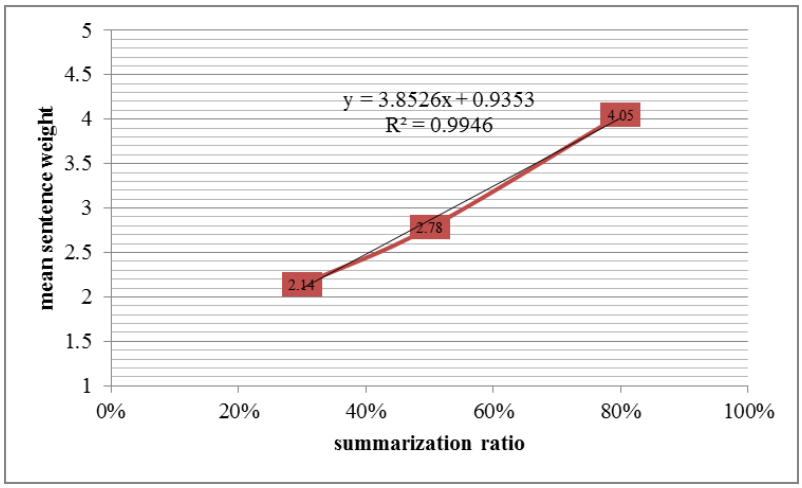

Figure 14: Mean sentence weight of different scales for all texts

Figure 15 - 24 illustrate the cardinality of sentences at each weight in pyramids and it is a plot from the information in columns 2-7 of the preceding tables.

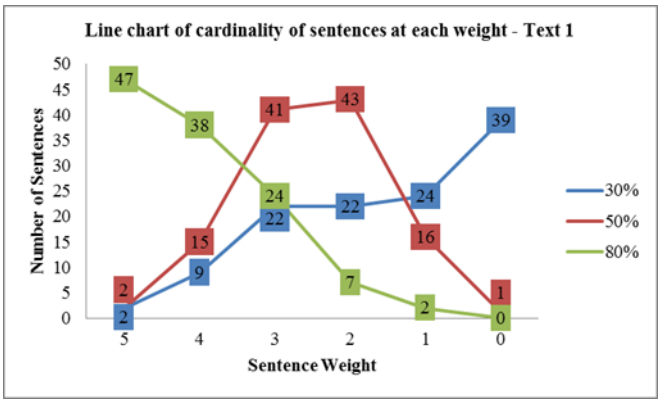

Figure 15: Line chart of cardinality of sentences for Text 1

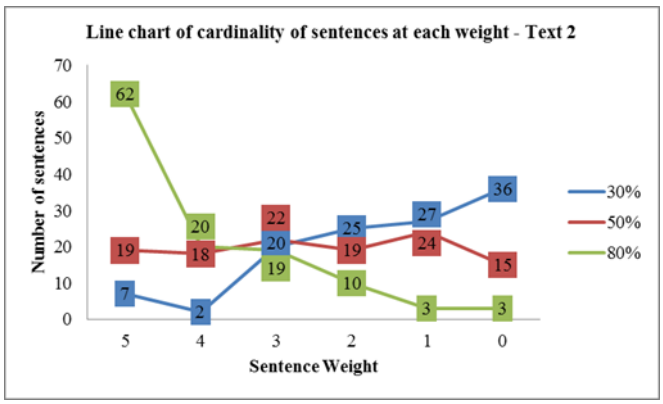

Figure 16: Line chart of cardinality of sentences for Text 2

The X-axis depicts each sentence weight from the highest weight (5), to the lowest weight (0). Weight 0 means that the sentences have not been mentioned in any of five summaries. Then, the Y-axis illustrates the cardinality of sentences at each weight.

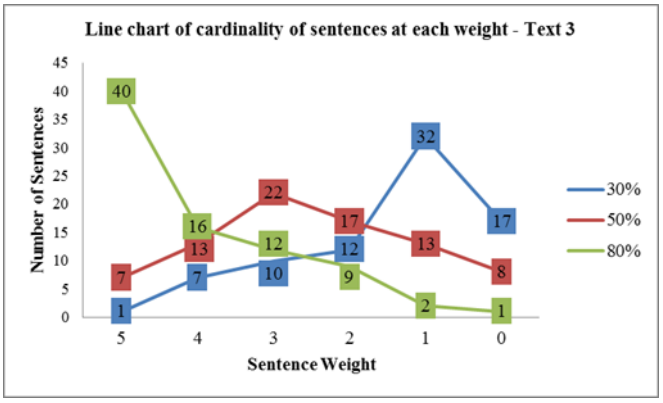

Figure 17: Line chart of cardinality of sentences for Text 3

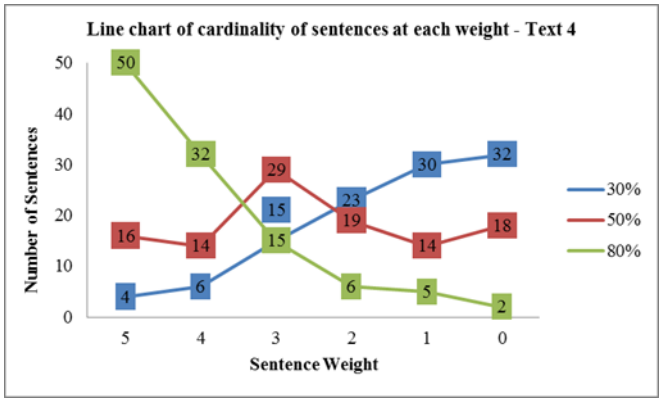

Figure 18: Line chart of cardinality of sentences for Text 4

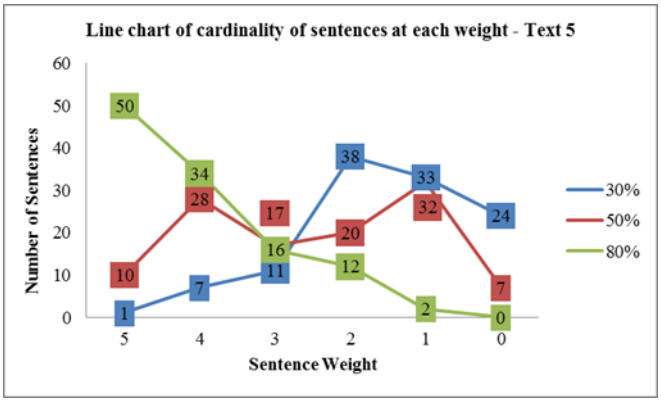

Figure 19: Line chart of cardinality of sentences for Text 5

As illustrated in these figures, 80% summaries have more consistent pattern than the other two scales. It is due to the fact that the summarizers have more common sentences in 80% extracts and this results in having more sentences at top tiers and fewer sentences at lower tiers.

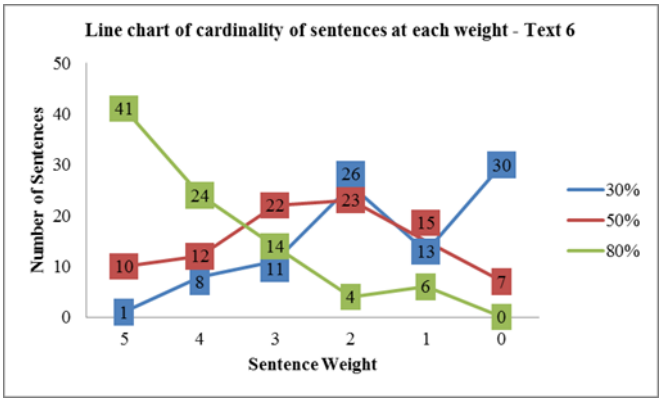

Figure 20: Line chart of cardinality of sentences for Text 6

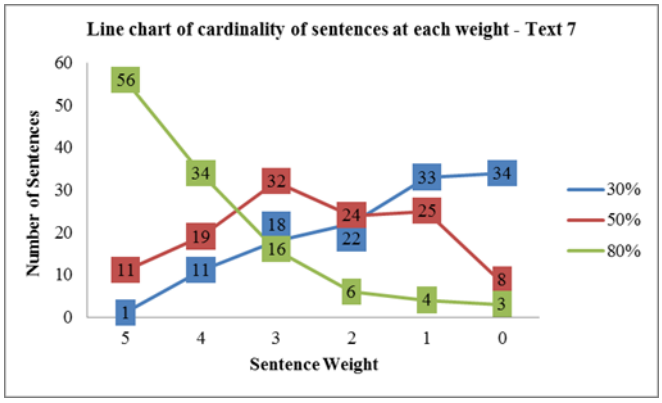

Figure 21: Line chart of cardinality of sentences for Text 7

Comparing scientific and social texts, it can be interpreted that there is more deviation from the general pattern in social texts than scientific texts. In 80% summaries of scientific texts, there is just one deviation in Text 2, but in 3 out of 5 in social texts we can find such kind of deviation. In 30% summaries, scientific texts have more ordered pattern than social texts.

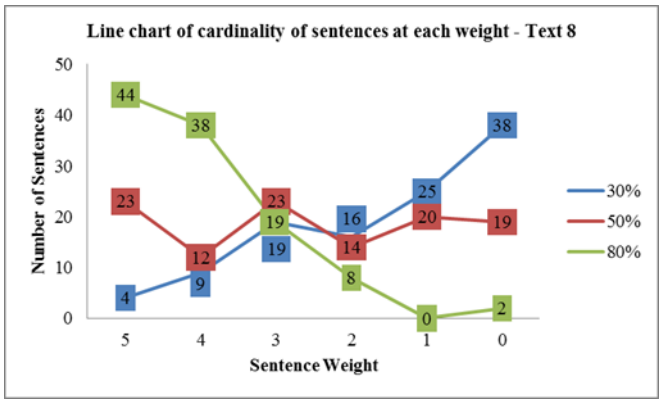

Figure 22: Line chart of cardinality of sentences for Text 8

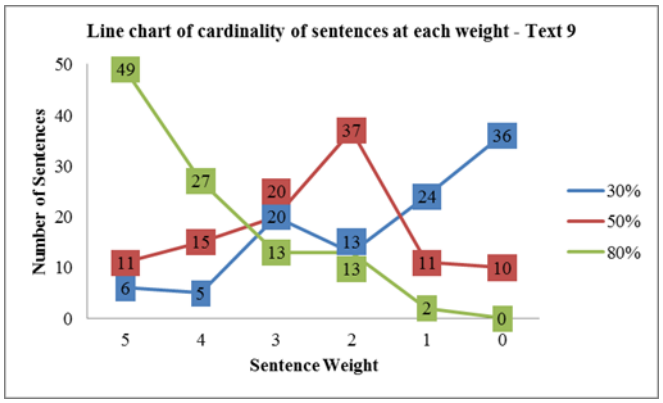

Figure 23: Line chart of cardinality of sentences for Text 9

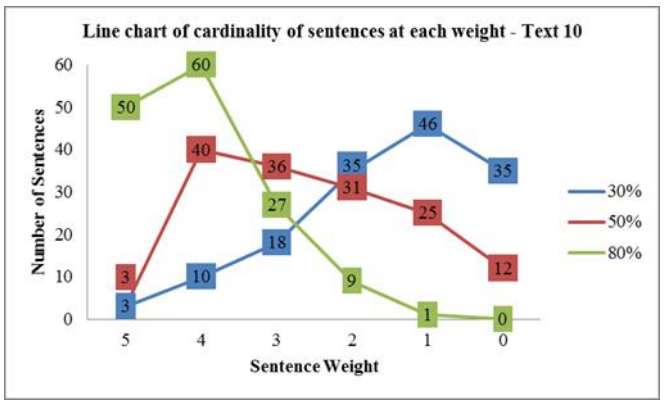

Figure 24: Line chart of cardinality of sentences for Text 10

Although, there is not that much difference between 50% pyramids of scientific and social texts, based on the above figures, it is understood that social texts are slightly better in this case.

Another pattern that can be extracted from the summaries is the distribution of sentences in different parts of them. Figure 25 and 26 for the first two summaries show which sentences are extracted by human summarizers. Appendix 1 shows the figures for the rest of the texts. X-axis in these figures represents the sentence numbers, while Y-axis shows the frequency of the sentences in 5 extracts. The pattern is illustrated for 30%, 50% and 80% summaries.

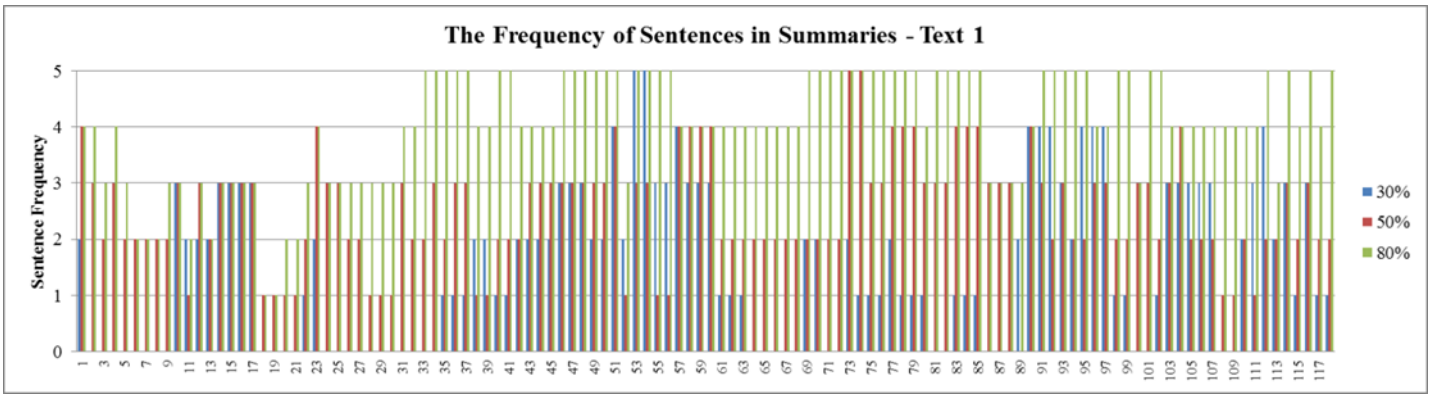

Figure 25: Column chart of the frequency of sentences in summaries for Text 1

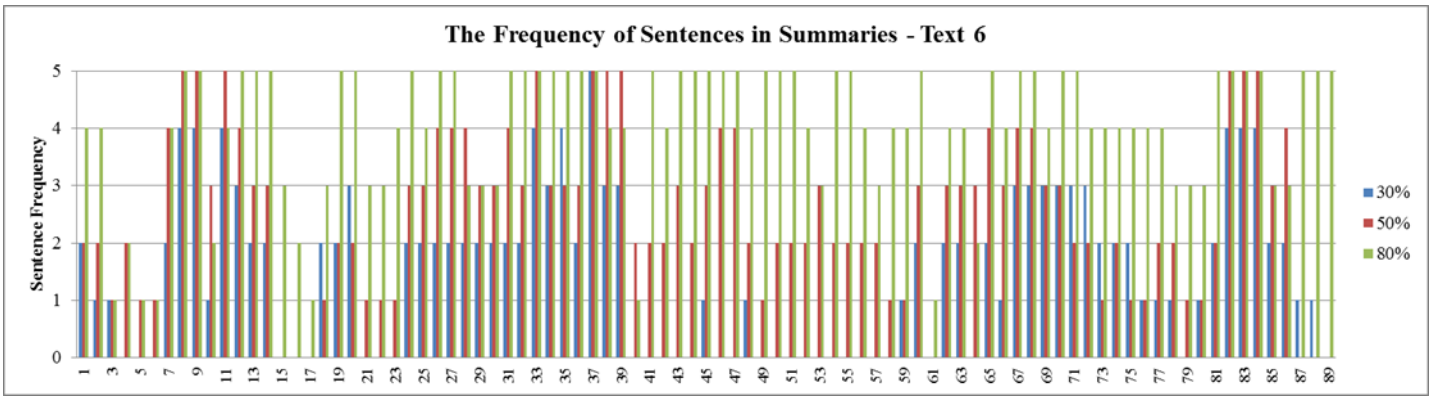

Figure 26: Column chart of the frequency of sentences in summaries for Text 6

Taking a careful look at the figures for all summaries, it is hard to get an exact idea on the pattern of sentences extraction from the texts. According to these figures, human summarizers select sentences from different parts of the texts.

4.2 Analysis of Gold Standards

Table 7 shows the questionnaire which was used in this research. It is repeated for convenience to the readers. As Question 3 “The summary is experienced as strenuous to read” and Question 5 “The summary is experienced as missing relevant information from the original document” were negative questions, compared to the other 3 questions; their values were inverted.

Table 7: Questionnaire – English Version

Question 1	The summary has a good length to give an idea on the content in the original text				
Disagree 1	2	3	4	5	6 Fully agree with
Question 2	The summary is experienced to be information rich,				
Disagree 1	2	3	4	5	6 Fully agree with
Question 3	The summary is experienced as strenuous to read.				
Disagree 1	2	3	4	5	6 Fully agree with
Question 4	The summary gives a good idea on what is written in the original document.				
Disagree 1	2	3	4	5	6 Fully agree with
Question 5	The summary is experienced as missing relevant information from the original document				
Disagree 1	2	3	4	5	6 Fully agree with
Question 6	The summary is experienced as a good complement to the original document.				
Disagree 1	2	3	4	5	6 Fully agree with

Table 8 and 9 depicts the answers to the questionnaire from both groups. Table 8 illustrates the data from graduates' questionnaires, while Table 9 is the representation of teachers' questionnaires. As it was interesting to know the spread and diversion of responses in every group, the standard deviation was calculated for every question in every text.

Table 8: Mean from graduates' responses on 30% gold standards (T1 to T5 are scientific texts, while T6 to T10 are social texts.) The numbers in parentheses show Standard Deviation (SD) between five human summarizers for every question in every text.

	Question 1	**Question 2**	**Question 3**	**Question 4**	**Question 5**	**Question 6**	**Mean**
T1	3 (1.41)	3 (1.22)	3.8 (1.30)	2 (1.41)	2.8 (1.48)	2.4 (1.95)	2.8 (1.46)
T2	4.4 (1.14)	3.6 (1.14)	4.8 (0.45)	3.6 (1.67)	4.4 (1.34)	4.2 (1.64)	4.1 (1.23)
T3	3.6 (1.14)	3.8 (0.84)	3.2 (1.79)	4.8 (1.30)	4.4 (0.89)	4.4 (0.89)	4.03 (1.14)
T4	3.2 (2.04)	3 (1.22)	1.2 (0.83)	2.6 (1.67)	3.4 (1.52)	4 (0.55)	2.9 (1.33)
T5	3.6 (1.81)	4 (1)	4.4 (0.55)	4 (1.58)	3.8 (1.30)	4.4 (1.14)	4.03 (1.23)
T6	5 (1.73)	3.6 (1.14)	5 (0)	3.6 (2.07)	4.2 (1.79)	3.8 (1.64)	4.2 (1.4)
T7	4.6 (0.89)	4.6 (1.14)	4.8 (0.45)	4.2 (1.92)	3.8 (1.64)	4.2 (1.30)	4.3 (1.22)
T8	3 (1.34)	3 (1.58)	5 (0)	3.4 (2.50)	4 (1)	3 (2.34)	3.5 (1.63)
T9	3.4 (0.89)	3.8 (0.84)	3.4 (1.82)	3.2 (1.30)	3.6 (1.67)	4.4 (0.55)	3.6 (1.18)
T10	3.8 (2.17)	3.4 (1.82)	5 (0)	3.6 (1.67)	4 (1.73)	3.4 (2.07)	3.8 (1.57)
	3.7 (1.56)	3.5 (1.19)	4.06 (0.71)	3.5 (1.71)	3.8 (1.44)	3.8 (1.42)	**3.7 (1.34)**

The group of graduates finds the gold standards different in quality. In their opinion Text 7 (mean = 4.3), which is a social text, has the best gold standard, while Text 1 (mean = 2.8) has the lowest quality among other texts. They agree more on Question 4 (mean = 4.06) and it means that they consider that the texts gave good ideas on what is written in the original document. The dispersion of the answers is the highest for Text 8(SD = 1.63), while it is the lowest for Text 3 (SD = 1.14). This scale is the highest for Question 4 (SD=1.71) and the lowest for Question 3 (SD=0.71). Mean standard deviation of all texts and all questions is 1.34 for graduates.

Table 9: Mean from teachers' responses on 30% gold standards (T1 to T5 are scientific texts, while T6 to T10 are social texts.) The numbers in parentheses show Standard Deviation (SD) between five human summarizers for every question in every text.

	Question 1	**Question 2**	**Question 3**	**Question 4**	**Question 5**	**Question 6**	**Mean**
T1	4.6 (1.34)	4.2 (0.84)	4.4 (0.55)	4.6 (0.89)	4.4 (0.55)	4.6 (0.89)	4.4 (0.84)
T2	4.6 (0.55)	3.2 (1.30)	4.6 (0.55)	4.2 (0.45)	4.8 (0.45)	4.4 (1.52)	4.3 (0.80)
T3	4.4 (0.89)	4 (1.22)	4 (1.22)	4.8 (1.30)	4.8 (0.45)	5 (0.70)	4.5 (0.97)
T4	3.8 (1.1)	3.4 (1.52)	2.8 (1.64)	4.2 (0.84)	3.6 (1.52)	3.8 (1.79)	3.6 (1.4)
T5	4 (1.58)	3.8 (1.64)	4.4 (0.55)	4.6 (0.89)	4.4 (0.55)	4 (1.22)	4.2 (1.07)
T6	5.2 (0.83)	3.6 (1.52)	4.8 (0.45)	4.4 (1.67)	4 (1)	4 (1.58)	4.3 (1.17)
T7	5 (1)	4.2 (0.84)	4.4 (0.89)	5.4 (0.55)	4.6 (0.55)	5.4 (0.89)	4.8 (0.79)
T8	4.6 (1.95)	4.2 (0.84)	4.2 (0.55)	5.2 (0.84)	4.6(0.55)	4.6 (1.14)	4.5 (1.02)
T9	4.8 (0.84)	4 (1.73)	4.4 (0.89)	5.2 (0.84)	4.8 (0.45)	4.8 (0.84)	4.6 (0.93)
T10	4.8 (1.64)	4.2 (1.64)	4.2 (1.09)	4.6 (1.52)	4.6 (0.55)	4.8 (1.64)	4.5 (1.35)
	4.5 (1.17)	3.8 (1.31)	4.2 (0.87)	4.7 (0.98)	4.4 (0.66)	4.5 (1.22)	**4.3 (1.03)**

Table 9 is the representation of the answers from teachers' point of view. They agree more on good quality of the gold standard for Text 7(Mean = 4.8), and also agree less on good quality of Text 4 (Mean = 3.6). The degree of variability (SD) is the lowest for Text 7 (SD=%0.79) and highest for Text 4. Question 5 have the lowest standard deviation (SD= 0.66), while Question 2 had the highest. (SD=1.31) The mean standard deviation in teachers' group is 1.03.

Both groups agree more upon the better quality of gold standards for social texts. The graduates prefer social texts more with (mean = 3.88) compared to 3.57 for scientific texts. In this regard, the mean from teachers' shows that they agree upon better quality of social texts with (mean = 5.4), while the mean for scientific texts is 4.2. The graduates agree more upon the simplicity of the gold standards and less upon their richness. They also think that the texts do not give a good idea from the original texts. The teachers think that the texts are not information rich, while they agree that the texts give a good idea on what is written in the original document.

5 Conclusions and Discussion

The results in this research correlated with previous researches like Nenkova (2006) and Carlsson (2010), which stated that five human summarizers were enough for creating gold standards. Based on five human summarizers 150 extracts were generated. Sentence overlap was on the first scale to be considered in this research. Sentence overlap between summaries extracted from the same texts in this research was between 65.3% and 78.9%. Carlsson (2010) created 5 gold standards from Swedish texts. The agreement was between 57.5% and 76.6% in his study. Likewise, Hassel and Dalianis (2005) created gold standards for Swedish texts; the sentence overlap in their study was between 61% and 73%. We already know that the lengths of summaries in their study are different. Comparing these two studies on Swedish texts with Persian texts we can infer that Persian summarisers have slightly more agreement on the sentences to include in their extracts.

Another criterion described in this research was the distribution of sentences in different parts of the summarized texts. Finding a specific pattern for sentences distribution in these summaries was very difficult. This is either because of the fact that there is no exact pattern for Persian texts or maybe it is not just about the Persian language and it is more about the difference between scientific or social texts and other kinds of texts. For news texts the first paragraph gives the whole idea of a text. (Smith et al.: 2012) But we can't see this pattern in these figures.

From the ideas discussed before, we can infer that the Pyramid method has different patterns for different summary scales of 30%, 50% and 80%. Figure 27 shows the mean pattern of the pyramid for all the texts.

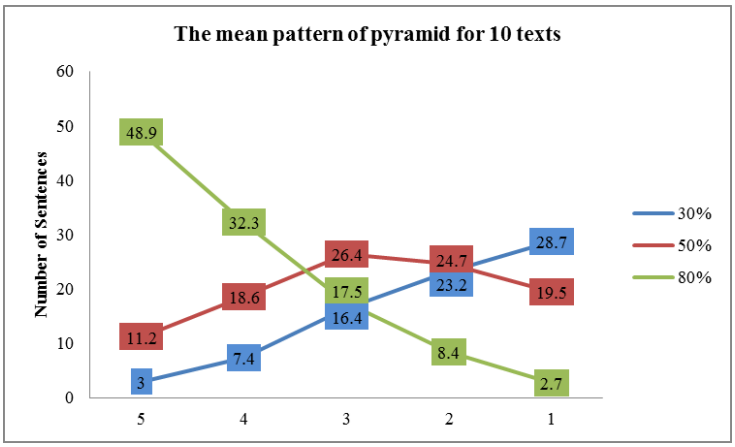

Figure 27: Line chart of the mean pattern of pyramid for 10 texts

Figure 28 is an interpretation of the preceding figure and it shows that the Pyramid method works precisely for 30% summaries but when including more sentences to reach 50%, the pyramid loses its form and for 80% we confront an upside-down pyramid. So, it seems that there is a limit for when the Pyramid method creates a pyramid.

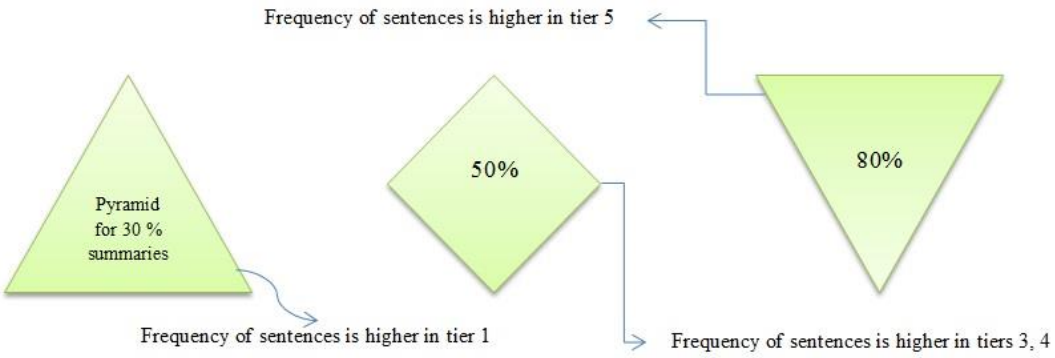

Figure 28: The applicability of the Pyramid method for 30%, 50% and 80% summaries

5.1 Limitations

Some of the problems we confronted during this research were related to Persian language structure. Persian is a less resourceful language and it has not yet been computationally developed well like other languages. Moreover, there are few good academic works on Persian text Summarization. It seems that there is a big gap in this field and it still demands more works to be done. Some gold standards have been already defined for Persian documents. We tried to access them, but we failed. We hope that the gold standards defined in this research help other researchers in their studies.

Another issue for summarization is that there is not a precise definition for a good summary and different people have different criteria for categorizing a summary. So, it is more subjective and it is based on the readers' ideas, opinions and their personal thoughts. Consequently, the gold standards defined in this research can be just one variation of the standard texts for Persian and we don't claim that these texts are the best. But we believe that they can be used for evaluation of automatic summarizers for Persian.

6 Summary and future work

Generating gold standards for Persian texts was one of the main purposes of this research. To this end, five human summarizers created 150 extracts out of 5 social and 5 scientific texts in different scales of 30%, 50% and 80%. Creating the pyramids for every extract was the next step. Gold standards were defined based on every pyramid. The lengths of gold standards were 30, 50 and 80 percent of the original texts. Finally, the gold standards were evaluated by two groups of *graduates* and *teachers* by means of a questionnaire. Based on the questionnaire both groups found the gold standards appropriate, but the teachers liked the summaries more than the graduates.

The Pyramid method created different patterns for different summarization ratios. For 30% summarization, it had a form of pyramid for most of the texts. For the majority of 50% summaries, it resulted in a diamond shape pattern, while for many of 80% summaries it created an upside-down pyramid. So, it can be claimed that when utilizing the Pyramid method for creating gold standards, 30% summarization ratio is probably the best scale that results in a pyramid.

Future work includes developing an automatic summarizer for Persian and use the gold standards developed in this research to automatically evaluate the summarized texts.

7 References

Carlsson, B. (2009). *Guldstandarder – dess skapande och utvärdering*. *Kandidatuppsats Linköpings*Universitet.

Carlsson, B. & Jönsson, A. (2010). *Using the pyramid method to create gold standards for evaluation of extraction based text summarization techniques*. Proceedings of *SLTC 2010*.

Hassel, M., 2004, *Evaluation of Automatic Text Summarization - A Practical Implementation*, Licentiate Thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden.

Hassel, M. & Mazdak, N. (2004): FarsiSum - A *Persian text summarizer*. *In the Proceedings of Computational Approaches to Arabic Script-based Languages,* Workshop at Coling 2004, the 20th International Conference on Computational Linguistics, August 28 2004, Geneva, Switzerland

Hassel, M. & Dalianis, H. (2005). Generation of Refternce Summaries. In the proceedings of the 2nd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, April 21-23, 2005, Poznan, Poland.

Mohaghegh, M. & Sarrafzadeh, A. (2010). *Performance Evaluation of Statistical English-Persian Machine Translation*. Proceedings of JADT2010 10th International Conference on Statistical Analysis of Textual Data. Sapienza, University of Rome, 9-11 June.

Nenkova, A., Passonneau, R. & McKeown, K. (2007). *The pyramid method: Incorporating human content selection variation in summarization evaluation*. *ACM Transactions on Speech and language processing*,4(2) pp.1–23.

Nenkova, A. & McKeown, Kathleen. (2011). Automatic Summarization. *Foundations and Trends® in Information Retrieval*, 5(2), pp.103–233.

Nenkova, A. & Passonneau, R. (2004). *Evaluating content selection in summarization: The pyramid method*. *Proceedings of HLT-NAACL.*

Passonneau, RJ. & Nenkova, A. (2005). *Applying the pyramid method in DUC 2005,* In *Proceedings of the Document Understanding Conference (DUC'05).*

Shamsfard, M. (2011). *Challenges and open problems in persian text processing*. *Proceedings of the 5th Language and Technology* Conference (LTC), pp.65–69, 2011.

Smith, Ch., Danielsson H.& Jönsson, A., (2012) *A More Cohesive Summarizer* Proceedings of the 24th International Conference on Computational Linguistics, Mumbai, India, 2012.

8 Appendices

Appendix 1- The Frequency of Sentences in Summaries

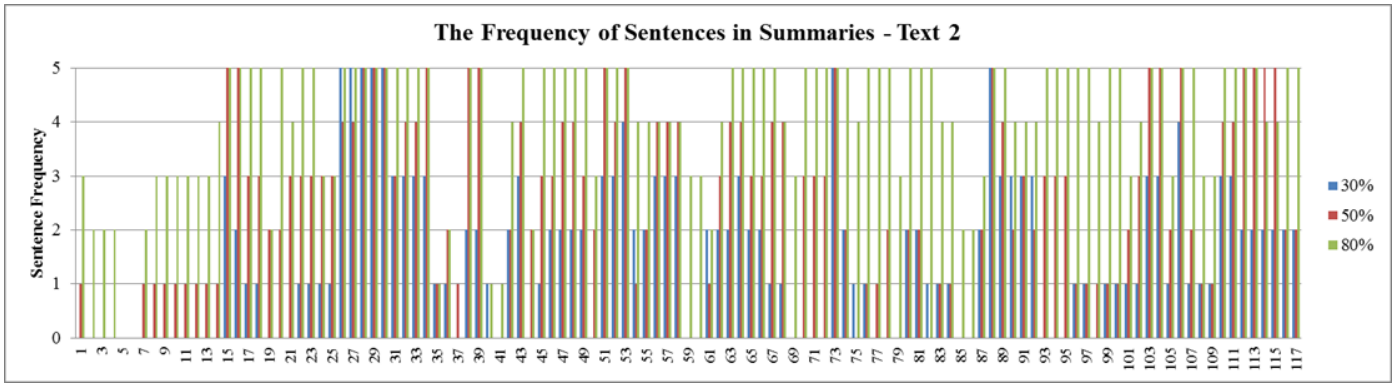

Figure 29: Column chart of the frequency of sentences in summaries for Text 2

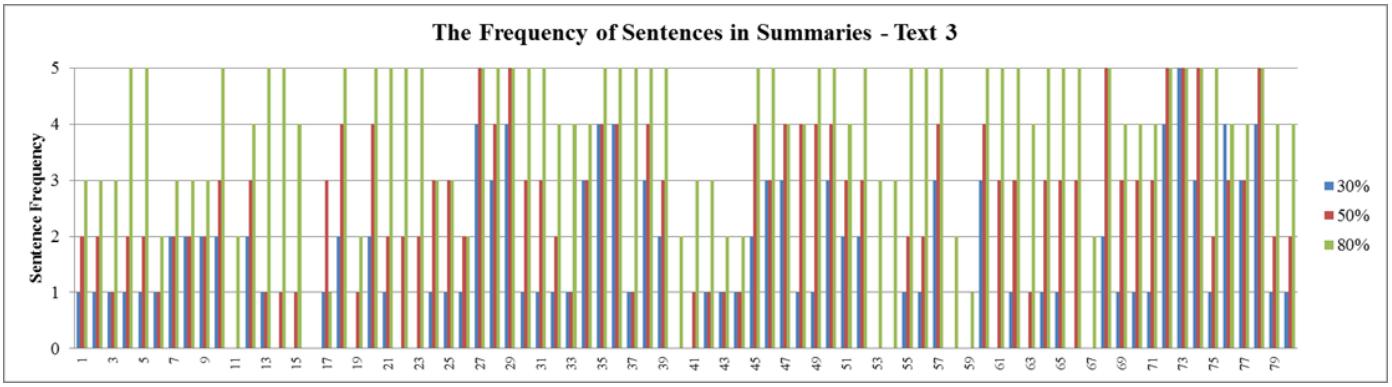

Figure 30: Column chart of the frequency of sentences in summaries for Text 3

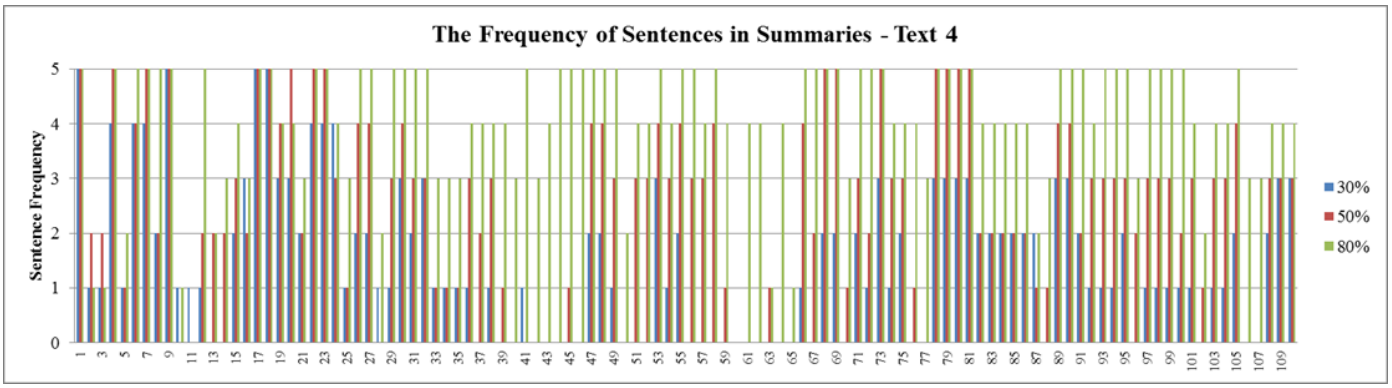

Figure 31: Column chart of the frequency of sentences in summaries for Text 4

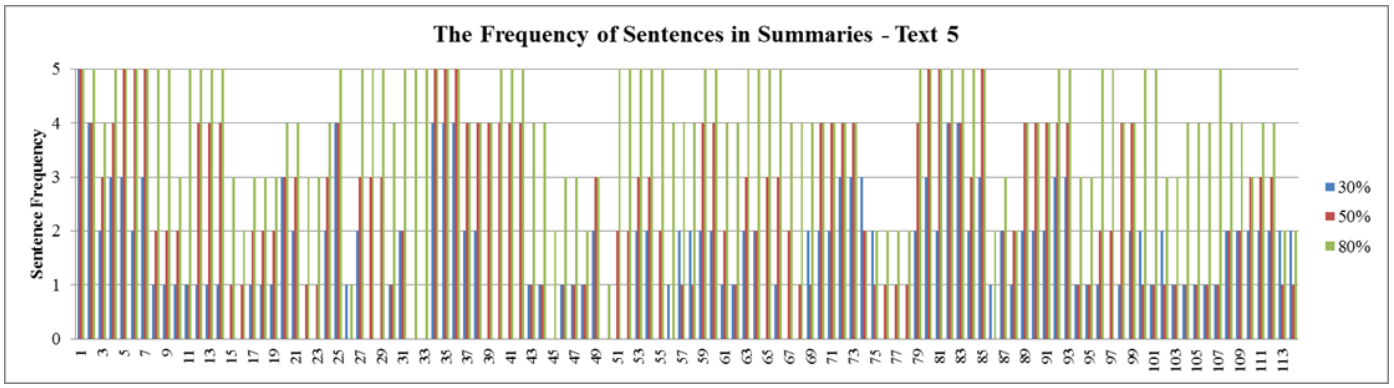

Figure 32: Column chart of the frequency of sentences in summaries for Text 5

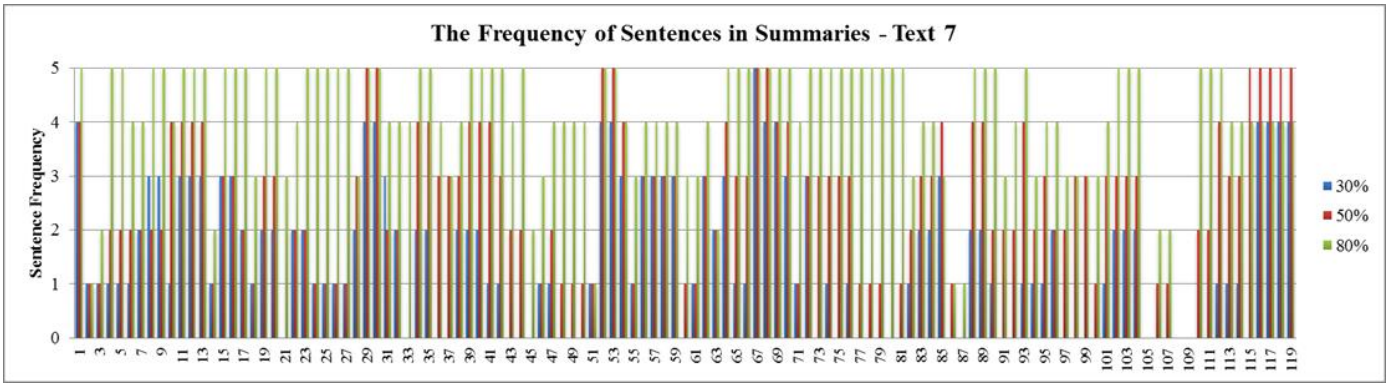

Figure 33: Column chart of the frequency of sentences in summaries for Text 7

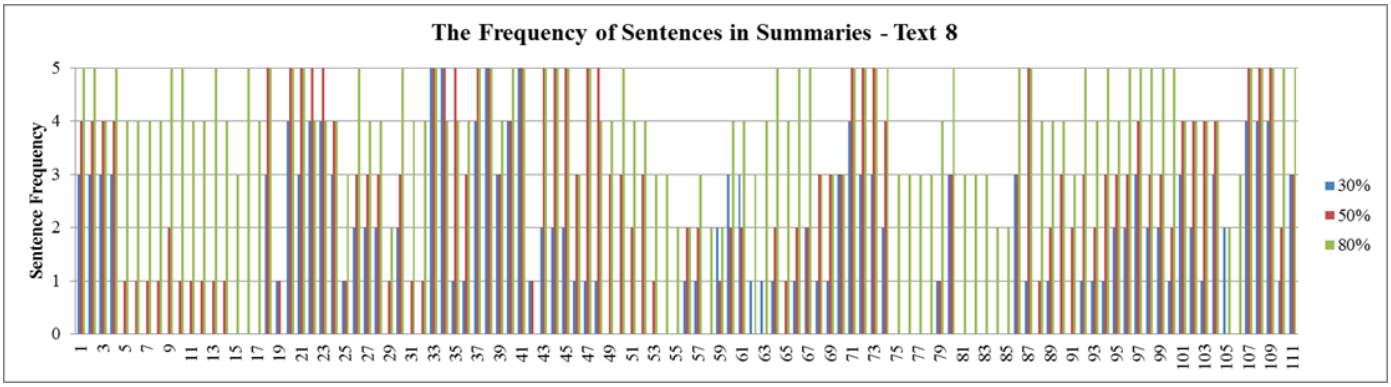

Figure 34: Column chart of the frequency of sentences in summaries for Text 8

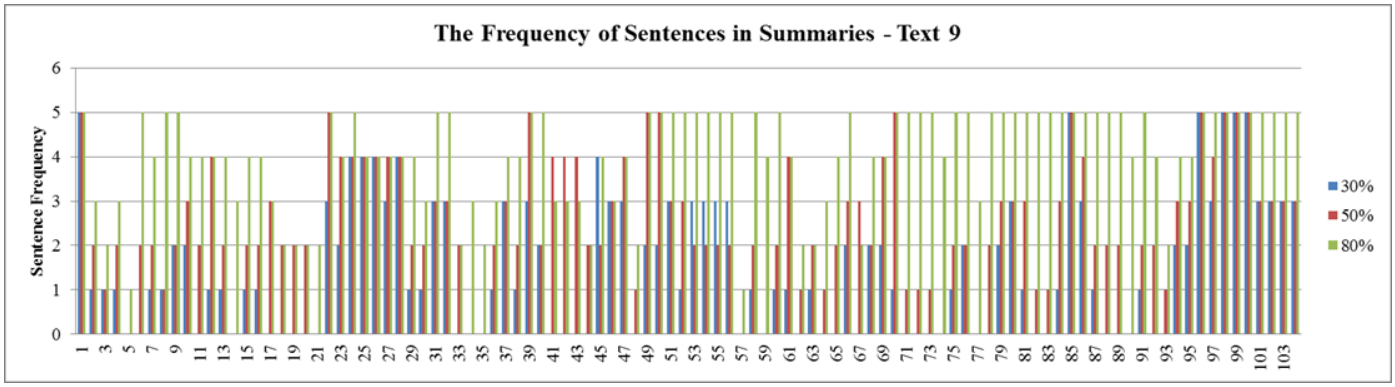

Figure 35: Column chart of the frequency of sentences in summaries for Text 9

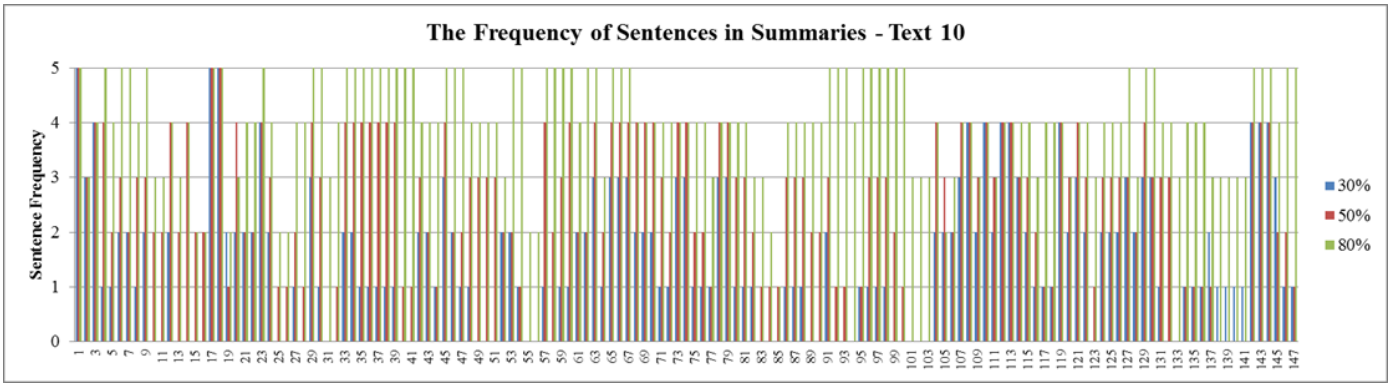

Figure 36: Column chart of the frequency of sentences in summaries for Text 10

Appendix 2 Questionnaire – Persian Version

نام متن:

با توجه به متنی که خوانده اید به سوالات زیر پاسخ بدهید.

1- این متن حجم (طول) مناسبی دارد برای اینکه یک ایده از متن اصلی را بدهد.

مخالفم ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 کاملا موافقم

2- این متن حاوی(مملو از) اطلاعات ارزشمند هست.

مخالفم ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 کاملا موافقم

3- این متن برای خواندن (خواننده) دشوار است.

مخالفم ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 کاملا موافقم

4 – این متن ایده خوبی از آنچه در متن اصلی نوشته شده می دهد.

مخالفم ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 کاملا موافقم

5- این متن فاقد اطلاعات مرتبط با متن اصلی هست.

مخالفم ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 کاملا موافقم

6- این متن مکمل خوبی برای متن اصلی هست (به نظرمی رسد).

مخالفم ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 کاملا موافقم

Appendix 3 – The Texts

۱. در جستجوی همتای آسمانی

1. In search of a celestial counterpart

Source: www.jamejamonline.ir

Link: http://www.jamejamonline.ir/NewsPreview/942579090596329438

پنج شنبه ۰۳ اسفند ۱۳۹۱

Date of Publication: February 21, 2013

۲. زیرو ـ اِس؛ سبکی نو برای آینده

2. Zero S, A new style for future

Source: www.jamejamonline.ir

Link: http://www.jamejamonline.ir/NewsPreview/939910787419538139

سه شنبه ۰۱ اسفند ۱۳۹۱

Date of Publication: February 19, 2013

۳. بازیافت قرن بیست و یکم: گرما

3. Recycling in 21st century: heat

Source: Daneshmand, an Iranian magazine of science and technology since 1963, No. 567

Link: http://www.daneshmandonline.ir/users/index.aspx#

۴. ژن قصه گو

4. An Extremeophile Gene

Source: www.aftabir.com

Link:

http://www.aftabir.com/articles/view/science_education/biology/c3_1347775473p1.php/%DA%98%D9%86-%D9%82%D8%B5%D9%87-%DA%AF%D9%88

۲۶ شهریور ۱۳۹۱

Date of Publication: September 16, 2012

۵. ستاره های دنباله دار

5. Comets

Source: aseman-nojom.blogfa.com

Link: http://aseman-nojom.blogfa.com/post-18.aspx

دوشنبه ششم دی ۱۳۸۹

Date of Publication: December 27, 2010

۶. امیدهای زیبا به کشور «عاری» از ایدز

6. Beautiful hopes for our country free of AIDS

Source: Etemad, Iranian daily newspaper, no. 2554

۸ آذر ۱۳۹۱

Date of Publication: November 28, 2012

۷. توقف رشد جمعیت در سال ۱۴۱۵ با تک‌فرزندی

7. Halt in population growth in 1415 with One-child Policy

Source: jamejamonline.ir

Link: http://www.jamejamonline.ir/NewsPreview/927471557381818010

شنبه ۲۱ بهمن ۱۳۹۱

Date of Publication: February 09, 2013

۸. قاضی مدرسه برای کاهش جرم می‌آید

8. The school's judge comes for crime reduction

Source: www.jamejamonline.ir

Link: http://www.jamejamonline.ir/NewsPreview/914685677954966406

چهارشنبه ۱۱ بهمن ۱۳۹۱

Date of Publication: February 02, 2013

۹. گذشتگان ساختند، ما خراب می‌کنیم

9. The departed built, we destroy

Source: Bahar, Iranian daily newspaper

Link: http://www.baharnewspaper.com/Page/Paper/91/10/11/10

دوشنبه،۱۱ دی ۱۳۹۱

Date of Publication: December 31, 2012

۱۰. مهارت‌های برقراری ارتباط

10. Communication Skills

Source: Hamshahri, Iranian daily newspaper

Link: http://hamshahrionline.ir/details/143647

شنبه ۲۹ مرداد ۱۳۹۰

Date of Publication: August 20, 2011