

The Impact of Cohesion Errors in Extraction Based Summaries

Evelina Rennes, Arne Jönsson

Linköping University & SICS East Swedish ICT

SE-581 83, Linköping, SWEDEN

evelina.rennes@liu.se, arnjo@ida.liu.se

Abstract

We present results from an eye tracking study of automatic text summarization. Automatic text summarization is a growing field due to the modern world's Internet based society, but to automatically create perfect summaries is challenging. One problem is that extraction based summaries often have cohesion errors. By the usage of an eye tracking camera, we have studied the nature of four different types of cohesion errors occurring in extraction based summaries. A total of 23 participants read and rated four different texts and marked the most difficult areas of each text. Statistical analysis of the data revealed that absent cohesion or context and broken anaphoric reference (pronouns) caused some disturbance in reading, but that the impact is restricted to the effort to read rather than the comprehension of the text. However, erroneous anaphoric references (pronouns) were not always detected by the participants which poses a problem for automatic text summarizers. The study also revealed other potential disturbing factors.

Keywords: Automatic summarization, eye-tracking studies, cohesion

1. Introduction

The task of automatic text summarization consists of reducing the length of a text, while preserving most of its content. It is a growing research field due to the last few decades' development of an Internet based society, characterized by the constant need of easy access to textual information. Except for the obvious benefit of effective information mediation, the ability to summarize texts automatically might be of use to persons with poor reading skills, for example people with dyslexia, cognitive disabilities, aphasia, or the partially sighted. To manually abbreviate and simplify texts is very time consuming, and many documents therefore remain inaccessible for poor readers.

There are various ways in which automatic summarization can be done, for example through *extraction* or *abstraction*. *Abstraction* paraphrases the text content by breaking out the key ideas in order to capture the general idea of the text (Hahn and Mani, 2000), whereas extraction based summarizers extract the most important sentences from a text and use them to produce a summary of the text.

Summaries produced from extraction based summarizers commonly suffer from problems regarding text cohesion, since important relations between words and sentences are lost in the extraction process and the limitation the extraction of complete phrases implies (Hahn and Mani, 2000). A lack of cohesion may result in an erroneous interpretation of a text (Otterbacher et al., 2002), and especially anaphoric references are known to cause problems in automatic text summarization (Hassel, 2000; Mani et al., 1998). The higher the level of summary is, the more errors are found (Kaspersson et al., 2012).

The aim of this study was to investigate how different types of cohesion errors affect the reading of a text summarized by an extraction based automatic text summarizer.

This was explored by tracking scan paths with an eye tracking camera and by letting participants rate and comment on the parts of the texts that were found to be difficult.

Earlier studies have shown that words that are used less frequently demand a longer lexical activation process (Just and Carpenter, 1980; Rayner, 1998). This motivates the choice to, in addition to the previously identified error types, also

look for other factors that might affect the experience of reading automatically summarized texts, for example unusual or difficult words.

2. Error Types

The different kinds of errors that are used in this study are derived from Kaspersson et al. (2012) who categorized three error types and sub-types:

1. Erroneous anaphoric reference
 - (a) Noun-phrases
 - (b) Proper names
 - (c) Pronouns
2. Absent cohesion or context
3. Broken anaphoric reference
 - (a) Noun-phrases
 - (b) Proper names
 - (c) Pronouns

Erroneous anaphoric references, describe errors that occur when an anaphoric reference refers to an incorrect antecedent. This is often the case when the summary has not included the correct antecedent and at the same time there is another antecedent in the text that fits. There are three sub-types of erroneous anaphoric references: noun-phrases, proper names and pronouns.

Absent cohesion or context, describes the case when the extracted sentences lack cohesion or context, which affects the comprehension of the summary.

Broken anaphoric references, are errors that occur when the summarizer does not extract the antecedent that is referred to in an anaphoric reference. There are three sub-types of broken anaphoric references: noun-phrases, proper names and pronouns.

3. Eye Tracking

Eye tracking is a method with many possible applications. The main concept associated with the method is that the eyes provide a kind of direct link to the cognitive processes and by studying the movement of the eye it is possible to gain insight into the cognitive state of a person executing a certain task. The eye's movement is a result of both goal driven and stimulus driven processes (Duchowski, 2007), and depends strongly on the type of cognitive task that is being performed. In our studies we will measure:

- *Fixations*, the period of time where the eye is relatively still (about 200-300 ms).
- *Fixation duration*. Just and Carpenter (1980) formed a hypothesis that an object or a text is processed exactly as long as a fixation lasts, and therefore implies a relatively easy access to cognitive processing. However, this is not uncontroversial, and the hypothesis has been questioned (Holmqvist, 2011; Rayner, 1998; Reichle et al., 1998).

The fixation duration indicates the effort needed for the cognitive processing, but the average fixation duration varies depending on the task and stimuli. The more complicated a text is, the longer the average fixation durations, and factors like stress might result in shorter fixations (Holmqvist, 2011).

According to Rayner (1998), the average fixation duration is not an adequate measure since it underestimates the duration that the fixations last. The first fixation is often longer than the following fixations on the same word, and the mean duration is therefore in many cases slightly too low. Rayner (1998) claims that the first fixation duration is a better measure of cognitive processing.

In usability research, many short fixations imply that information that was expected to be found is missing (Ehmke and Wilson, 2007).

All words of a text are not fixated during reading. Long words are more likely to be fixated than short ones (Just and Carpenter, 1980), but other aspects such as frequency and predictability from context are also proven to be a reason for shorter fixations or word skipping (Reichle et al., 1998).

- *Pupil size*, which increases during problem solving and correlates to the difficulty of the task which implies that this could be used as a measure of cognitive activity (Hess and Polt, 1964).

The diameter of the pupil can indeed be used to measure cognitive workload, though one has to be aware of the problems this method involves. The pupil size is sensitive to various states of the participant and the environment, factors that should be accounted for in the experimental design.

Except for cognitive workload, pupil size increases as an effect of emotion, anticipation, pain or drug influence, and it might decrease due to factors like fatigue, diabetes or high age. The environmental factors can

be controlled for by ensuring that the presented stimuli are of the same brightness and contrast and that the lighting of the room is kept constant (Holmqvist, 2011).

4. Procedure

The study was conducted on 23 students, 13 men and 10 women. They were all native Swedish speakers without any writing or reading disability and with normal or corrected-to-normal vision. The average age was 23.2 ($SD = 2.76$). The experiment consisted of four parts: answering a questionnaire, text reading, error marking and text rating.

4.1. Equipment

The eye tracking equipment used for this study was SMI iView RED II 50 Hz Pupil/Corneal reflex camera mounted underneath a 19" computer monitor. The softwares used for recording and analyzing the eye tracking data were iView X, Experiment Center 3.0 and BeGaze 2.

4.2. Texts

The texts used in the tests were four texts from the Swedish popular science magazine *Forskning och Framsteg*. The summaries are in Swedish and produced by the automatic text summarizer COGSUM (Smith and Jönsson, 2011).

COGSUM is based on Random Indexing and a modified version of the Weighted Page Rank algorithm, which is used for selecting the sentences that are most relevant in the text (Smith and Jönsson, 2011). The algorithm calculates a rank based on the Random Indexing vectors, which makes sentences that are similar in content support each other, and eventually result in a ranking of the sentences by their importance. The output of the summarizer was not in any way formatted, other than being divided into paragraphs in order to enhance readability. The texts were previously tagged for errors by Kaspersson et al. (2012).

The texts were summarized to a summary level of 33% meaning that 33% of the original text remained chosen in order to get as many errors as possible in a text, while keeping it at a reasonable length that is still readable (Kaspersson et al., 2012).

The four texts varied in length from 11 to 14 sentences and the number of tagged errors varied from 6 to 12 per text. In total there were 34 errors. The error types and number of errors for each type that were present in the texts were:

- 1(c) Erroneous anaphoric reference - Pronouns, a total of 4 errors
- 2. Absent cohesion or context, a total of 16 errors
- 3(a) Broken anaphoric reference - Noun-phrases, a total of 4 errors
- 3(c) Broken anaphoric reference - Pronouns, a total of 10 errors

The remaining error types were not present in the texts. Table 1 shows the amount of tagged cohesion errors for each text and the number of sentences for each text. The row labeled *Percentage* represents the ratio of the number

of errors and the number of sentences. Text 2 was the shortest text, with the least errors which resulted in a relatively low percentage of errors per sentence.

Text 3 and text 4 were of the same length (14 sentences) but text 3 had a higher percentage of errors per sentence, in fact it had the overall highest score of errors per sentence (85.71%).

Table 1: Descriptives of the texts used in the test.

	Text 1	Text 2	Text 3	Text 4
No. of errors	7	6	12	9
No. of sentences	12	11	14	14
Percentage	58.33%	54.55%	85.71%	64.29%

The order in which the texts was presented was not the same for all participants.

4.3. Questionnaire

The questionnaire was created with the intention of capturing the participants' reading strategies and prior attitudes to reading. The questionnaire items were answered using a unipolar Likert scale varying from 1 to 5, where 1 corresponded to *do not agree* and 5 represented *agree completely*. The participants also filled in age, gender, profession or current education, and whether glasses or contact lenses were used during the experiment.

4.4. Experimental Procedure

The participants were informed that the participation was completely voluntary, that they were going to be anonymized and that they were allowed to terminate the experiment if they did not want to continue.

Before positioning in front of the eye tracking equipment, the participants filled in the questionnaire described above, answering questions about reading strategies, and attitudes towards reading. The participants were positioned in front of a computer screen with the RED eye tracking camera positioned under the screen. Before the actual test, a calibration of the eye tracking camera was performed. The participants were asked to find a comfortable position before starting the calibration, since it is important to keep the same position during the test. The calibration was repeated until a satisfying calibration value was achieved.

The reading part of the test consisted of the four texts presented one by one. The participants were not aware that the texts were summarized. They were instructed to read the texts for as long as they wanted until they felt they understood it, and then continue to the next text. They were told that they were going to perform a task after having finished reading, but they were naive to what the task consisted of.

After reading, the participants were asked to mark the parts of each text that they considered most problematic to read, using a highlighter pen on a printed copy of the texts. They were allowed to mark as many as they wanted, and were then asked to rank the marked areas on a scale 1-3 where 1 was the least difficult area and 3 the most difficult area. They were then allowed to comment on their markings, and the comments were recorded.

The participants were also asked to rate the texts regarding difficulty, how boring they were, how interesting they were and how exhausting they were. A Likert scale from 1 to 5 was used, where 1 represented *do not agree* and 5 represented *agree completely*.

After the test, the participants were asked whether they felt that the presence of the eye movement camera had any impact of their performance, and if their attitude towards the texts would be different if they knew in advance that the summaries were automatically produced.

4.5. Areas of Interest

To analyze the data recorded by the eye tracking equipment, areas of interest (AOIs) were defined. There were four different AOIs corresponding to the four error types. In the case of error type 2, absent cohesion or context, it is often difficult to detect the specific place in a sentence where the error occurs, which motivated to mark the whole sentence as an AOI of type 2. In some sentences there was more than one error, and all AOIs were placed so that they did not overlap, with the result that some errors lack data from an area corresponding to the area of the other error type in the same sentence.

The rating of each error was used in order to motivate definitions of further AOIs, that did not correspond to any of the pre-defined error types. If more than half of the participants marked an area as difficult, that same area was defined as an AOI. However, this was not the case and no other area was considered in the analysis except for the already defined error types.

Since the AOIs varied in size, the number of fixations was corrected by dividing by the size of the AOIs, in order to get comparable scores.

5. Results

This section presents the results from all parts of the experiment conducted in this study. First, the data collected by the questionnaire and text rating are presented, followed by the results of the error marking and subjective rating. Finally, the results of the eye tracking data from the reading sessions are presented. The variables used for the statistical analysis of the eye tracking data were the number of fixations, fixation duration and pupil size.

5.1. Attitude to reading

The questionnaire that evaluated the participants' prior attitudes to reading gave the results presented in Table 2. The participants generally considered themselves to be good readers.

5.2. Text rating

The texts used in this study were evaluated regarding three different criteria: whether they were easy to understand, boring or exhausting to read. The results are shown in Table 3.

The texts differed slightly. According to the means, text 2 was considered the easiest, least boring and least exhausting text while text 1 was the most boring text and text 3 was the most exhausting text to read.

Table 2: Mean and standard deviation of participants' self rated reading abilities and attitudes towards reading.

Assertion	Mean	Std.Dev.
I usually understand what I read	4.61	.58
I am a slow reader	2.52	.90
I find it easy to read	4.70	.70
I find it exhausting to read	1.65	.88
I am often pleased to get a rough idea of a text's content	3.70	1.02

Table 3: Mean and standard deviation (within parentheses) of the text ratings.

Assertion	Text 1	Text 2	Text 3	Text 4
Easy	3.43 (.90)	3.96 (1.33)	2.91 (1.20)	3.78 (1.17)
Boring	3.52 (1.08)	1.78 (.85)	3.00 (1.08)	2.17 (.98)
Exhausting	2.96 (1.14)	2.00 (1.09)	3.48 (.99)	2.39 (1.1)

A statistical analysis (repeated measures ANOVA) revealed that for the criterium *easy*, texts differed significantly $F(3, 66) = 4.02, p < .05$. A Bonferroni post-hoc test showed that there is a significant difference between text 2 and text 3 ($p < .05$), implying that text 2 was significantly easier than text 3.

For the criterium *boring*, significant differences were found $F(3, 66) = 15.28, p < .001$. Bonferroni post-hoc test showed that there are significant differences between text 1 and text 2 ($p < .001$), text 1 and text 4 ($p < .05$), and between text 2 and text 3 ($p < .05$). The results show that text 1 and text 3 were significantly more boring than text 2. For the criterium *exhausting* significant differences were found $F(3, 66) = 9.37, p < .001$. Bonferroni post-hoc test showed that there are significant differences between text 1 and text 2 ($p < .05$), text 2 and text 3 ($p < .001$), and between text 3 and text 4 ($p < .05$). Text 1 and text 3 were significantly more exhausting to read than text 2, and text 3 were significantly more exhausting to read than text 4.

5.3. Error marking and subjective rating

From the analysis of the error markings made by the participants, other areas than the previously tagged errors were marked, see Table 4. The errors that had been identified in advance made up 38.3% of the total amount of markings. The second most frequent reason of marking was different types of language related problems, for example long sentences or complicated word order (17.55%). Difficult words accounted for 11.7% of the total amount of markings. General problems to understand the context were represented in 9.04% of the markings and summarizer errors and numbers made up in 7.45% and 4.79% respectively of the total amount of marked areas. The category Other collects comments where the participants were not able to explain why they marked a certain area as problematic. It

accounts for 11.17% of the markings.

Table 4: Distribution of cohesion errors and other categories that were marked by the participants.

Category	Percentage
Cohesion error	38.3%
Language	17.55%
Difficult words	11.7%
Context	9.04%
Summarizer errors	7.45%
Numbers	4.79%
Other	11.17%

The mean of the subjective rating, between 1 (least difficult) and 3 (most difficult), of each error that the participants gave the marked errors is presented in Table 5.

Table 5: Mean and standard deviation of the subjective rating for each error type.

Error type	Mean	Std.Dev.
Error 1c	1.82	0.77
Error 2	1.85	0.81
Error 3a	1.70	0.86
Error 3c	1.88	0.78

All error types had similar scores ranging from 1.70 to 1.88. No statistical significance was found between the subjective ratings of each error type.

All participants reported that their attitude would be more lenient if they knew in advance that the texts used in the test were summaries.

5.4. Eye tracking results

This section presents the eye tracking results from the reading session. None of the participants reported that the presence of the eye movement camera had any significant impact on their reading performance.

Table 6 presents the eye tracking data. The row labeled *None* corresponds to the area that has not been marked as an error type (i.e. AOI), and is thus seen as the rest of the text.

A repeated measures ANOVA was used to test for differences between the four error types and the rest of the text. For fixation duration and pupil size no difference were found ($p > .05$). For the corrected number of fixations there was a significant difference $F(2, 160, 47.522) = 251.86, p < .001$, Greenhouse-Geisser corrected.

The graph in Figure 1 presents the number of fixations (means), corrected for the size of the AOI, distributed over the error types. Error type 2 and error type 3c have the highest number of fixations.

Bonferroni post-hoc tests revealed significant differences, as presented in Table 7. As before, *None* represents the rest of the text. Statistically significant differences are marked in bold.

Table 6: Mean and standard deviation of the number of fixations, fixation duration and pupil size for each error type. The values corrected for the size of the AOI are within parentheses.

Error	Number of fixations	
	Mean	Std.Dev.
1c	13.61 (2.28)	6.22 (1.04)
2	210.30 (8.69)	51.91 (2.15)
3a	12.70 (1.75)	4.30 (.59)
3c	22.61 (6.14)	5.08 (1.38)
None	841.44 (1.25)	193.77 (.29)
	Fixation Duration	
	Mean	Std.Dev.
1c	291.88	76.14
2	280.25	41.63
3a	269.17	52.83
3c	279.20	59.19
None	273.33	41.13
	Pupil Size	
	Mean	Std.Dev.
1c	10.82	1.20
2	10.76	1.21
3a	10.73	1.20
3c	10.79	1.22
None	10.82	1.21

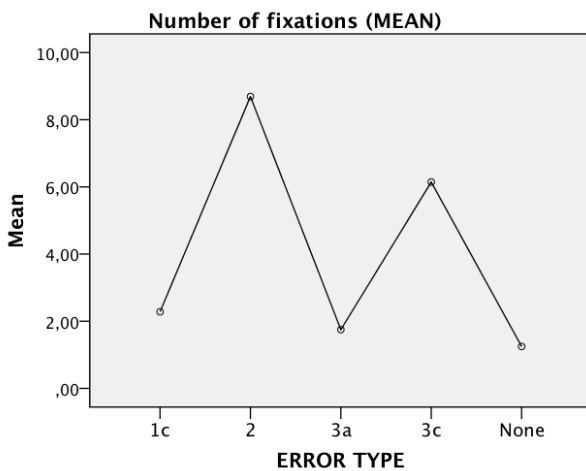


Figure 1: The number of fixations (mean) distributed over the different error types.

All error types were fixated significantly more than the rest of the text ($p < .05$).

Error type 2, absent cohesion or context, had significantly more fixations than all other error types ($p < .001$). Error type 3c, broken anaphoric reference (pronouns), had significantly more fixations than error type 1c and 3a ($p < .001$). Significant differences were found between all error types except for 1c, erroneous anaphoric reference (pronouns), and 3a, broken anaphoric reference (noun-phrases) ($p = .065$). The marginal significance level suggests a ten-

Table 7: Pairwise comparisons from the Bonferroni post-hoc test. Significant differences are marked in bold.

Pairwise Comparisons		M. Diff.	Sig.
Error 1c	Error 2	-6.41	.000
	Error 3a	.53	.065
	Error 3c	-3.86	.000
	None	1.03	.000
Error 2	Error 1c	6.41	.000
	Error 3a	6.94	.000
	Error 3c	2.55	.000
	None	7.44	.000
Error 3a	Error 1c	-.53	.065
	Error 2	-6.94	.000
	Error 3c	-4.39	.000
	None	.50	.002
Error 3c	Error 1c	3.86	.000
	Error 2	-2.55	.000
	Error 3a	4.39	.000
	None	4.90	.000
None	Error 1c	-1.03	.000
	Error 2	-7.44	.000
	Error 3a	-.50	.002
	Error 3c	-4.90	.000

dency of slightly more fixations on error type 1c than on error type 3a.

6. Discussion

This section presents a discussion about both the results and the method used in this study. First, the results of the different parts of the tests are discussed, followed by a discussion regarding the experimental procedure.

6.1. Text ratings

There were several differences between the texts. In the text rating, text 2 stood out, being considered the most easy, least boring and least exhausting text. This can be explained by the fact that it was the shortest of all texts, and that it had the lowest percentage of errors per sentence. Text 3 was considered the most exhausting text. It was the text that had the highest number of errors per sentence and it was also one of the longest texts. Text 4, which had the same length as text 3 but less errors, was considered less exhausting than text 3. No difference was found regarding the difficulties of these two texts. This suggests that the experience of the text is influenced by the number of cohesion errors, rather than the text length.

Text 1 was considered the most boring text. Seen to the length and number of errors, it was similar to text 2, but differed in rating. The reason to why text 1 was considered more boring might be that the topic was considered boring. Text 1 treated the Nobel Prize while text 2 treated polar bears, and it is possible that the second topic seemed more attractive to the reader.

No participant was aware of that the texts were summarized, which resulted in a critical attitude towards the texts. After finishing the test, the participants were asked whether

their attitude towards the texts would be different if they knew in advance that the texts they had read were summaries. All participants claimed that they would be more lenient with the texts if they knew that they were automatically summarized, and this is probably an important factor when evaluating the automatic text summarizer. When used in real situations and when the user is aware of this fact, it is likely that the different errors are not seen as severe as in this study. It would be interesting to investigate whether the summaries are preferred over the original texts, despite of their errors.

6.2. Error markings

As expected, the majority of areas marked by the participants (38.3%) were marked due to the previously identified cohesion errors. However, other areas than those previously identified as errors were found. A number of problematic aspects are not necessarily due to the text being automatically summarized. For instance, 11.7% of the markings represented difficult words in the text and 17.55% of the markings were due to problems that arose from linguistic factors, such as long sentences, or phrases with a difficult word order.

A number of errors emerge as the automatic text summarizer is extraction based. Information disappear from the original text, causing difficulties to understand the general context, or leaving sentences with a strange word order behind, even if this does not result in cohesion errors. For instance, 9.04% of the markings were areas that the participants claimed were out of context, although not tagged as absent cohesion or context error. The reason to this is probably that the error types affect other parts of the text as well, and that the error type is vague and hard to narrow down to apply for only one sentence. 11.17% of the markings belonged to the *Other* category which may be due to missing context but other factors may also influence, such as that the sentence conveys information that is difficult to comprehend.

The subjective rating of the errors, rated 1, 2 or 3, depending on how difficult the participants experienced the areas to be, scored similarly and showed no significant difference, suggesting that no error type was considered more problematic.

The subjective rating showed that the reader complains about missing context in other parts of the text than the sentences actually tagged by an error type. This suggests one of two things: either that error type 2 does not cover all cases satisfactory and should be expanded, or that the error type is vague and cannot possibly include all cases of absent context.

6.3. Eye tracking results

The results from the eye tracking constitute the main part of the analysis. In the previous chapter it was shown that there are significantly more fixations in the areas marked as 2, absent cohesion or context, and 3c, broken anaphoric reference (pronouns).

The results of the statistical analysis of the eye tracking data suggests that error type 2 and error type 3c are the areas that cause the most reading disturbances. However, no dif-

ference could be observed for the duration of the fixations or the pupil size, which indicates that these areas are not more cognitively engaging than the rest of the text. These two claims are somewhat contradictory, and therefore interesting. According to the general hypothesis of fixation duration, long fixations means deep cognitive processing, which would imply that the errors did not cause any substantial effort. However, according to Ehmke and Wilson (2007), many short fixations might indicate confusion when expected information is missing. Although this claim is made within the field of usability research and is applied on a web stimulus, it could be seen as a possible interpretation to the pattern of many but short fixations within the areas of error type 2 and 3c.

Error type 1c, erroneous anaphoric reference, had significantly more fixations than the rest of the text, but less fixations compared to the other error types (except for error 3a where no statistical significance was found). The reason to why this error type is fixated less might be because it is difficult to identify, since the anaphoric expression refers to an existing (erroneous) antecedent. For the practical use of an automatic text summarizer, it is preferable that the errors are found. If the reader does not discover that the antecedent is erroneous, the comprehension of the text may be inaccurate.

Although the average fixation duration is not always considered to be an adequate measure, since it tends to underestimate the duration that the fixations last, this variable was chosen for data analysis. This is motivated by the fact that the cohesion errors may force the reader to make regressions and return to previously read passages, which makes the first fixation duration an insufficient measure of the time spend on a certain word or sentence. Another reason to choose the average fixation duration over the first fixation duration was that it is then possible to compare the AOIs with the rest of the text, since the first fixation duration would give an erroneous value on the rest of the text.

There was no significant change in pupil size which suggests that the participants did not find the cohesion errors more cognitively demanding. The given instructions were *read for as long as you want until you feel that you understand*, and the participants were not informed about the task until after reading all texts. Since there was not a specific task to perform, the cognitive workload might have been lowered, and the problematic areas where only shown in the number of fixations.

Since the size of the pupil might increase or decrease due to other factors, it is a metric that demands an experimental design that controls for all other potential factors. The experiment conducted in this study did not control for factors like fatigue or light variation, which might be another possible explanation to the little change in pupil size.

7. Conclusion

The results of the experiment led to the following conclusions:

- It is clear that cohesion errors affect the experience of reading a summary negatively. The number of fixations was significantly higher in areas belonging to error type 2 (absent cohesion or context) and error type

3c (broken anaphoric reference, sub-type pronouns) which could suggest that the participants experienced difficulties when trying to read these error types in particular.

- It may be that the more cohesion errors there are in a text, the more exhausting it is to read. Yet a text with a high amount of cohesion errors per sentence is not significantly more difficult, which suggests that the errors indeed cause problems during reading, but that the impact is restricted to the effort to read rather than to comprehend the text.
- There are other factors except for cohesion errors that constitute a source of disturbance. The majority of areas marked by the participants (38.3%) were marked due to the previously identified cohesion errors, but there were other aspects of the texts that seemed to cause problems to the reader, for example linguistic factors (17.55%) and difficult words (11.7%). This suggests that there are other factors affecting the experience of reading automatically produced summaries, factors that are not specific to summaries, such as difficult words.
- The number of fixations suggests that cohesion errors affect reading, but the disturbances need not be severe. The non-significant difference in average fixation duration and pupil size supports the claim that the participants did not find cohesion errors more cognitively involving than the rest of the text.

This study has investigated cohesion errors in texts summarized by the extraction based summarizer COGSUM, but we believe that the results are valid for any summarizer that does not consider cohesion, e.g. (DUC, 2002; Chatterjee and Mohan, 2007; Hassel and Sjöbergh, 2007; Gong, 2001; Mihalcea and Tarau, 2004).

This study used participants that were unaware of the fact that the texts were summarized, and it is possible that the result would be different if prior knowledge was different. All participants claimed that they would be more lenient with the texts if they knew that they were automatically summarized, and an approach for a future investigation could be whether summaries are preferred over original texts, despite of their weaknesses.

Acknowledgments

We would like to thank Stiftelsen Marcus och Amalia Wallenbergs Minnesfond and SICS East Swedish ICT AB for funding this research.

8. References

Nilhadri Chatterjee and Shiwali Mohan. 2007. Extraction-Based Single-Document Summarization Using Random Indexing. In *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence – (ICTAI 2007)*, pages 448–455.

DUC. 2002. Document Understanding Conference. <http://duc.nist.gov/pubs.html#2002>.

Andrew T. Duchowski. 2007. *Eye tracking methodology: Theory and practice*. Springer-Verlag, London.

HP Edmundson. 1969. New Methods in Automatic Extracting. *Journal of the ACM*.

Claudia Ehmke and Stephanie Wilson. 2007. Identifying Web Usability Problems from Eye-Tracking Data. In *Proceedings of HCI 2007*, volume 1.

Yihong Gong. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Udo Hahn and Inderjeet Mani. 2000. The Challenges of Automatic Summarization. *Computer*, 33(11):29–36.

Martin Hassel and Jonas Sjöbergh. 2007. Widening the HolSum Search Scope. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (Nodalida)*, Tartu, Estonia, May.

Martin Hassel. 2000. Pronominal Resolution in Automatic Text Summarisation. Master's thesis, Master thesis in Computer Science, Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden.

Eckhard H. Hess and James M. Polt. 1964. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science*, 132(3611):1190–1192.

Kenneth Holmqvist. 2011. *Eye Tracking - A Comprehensive Guide to Methods and Measures*. Oxford University Press.

M. A. Just and P. A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Pentti Kanerva. 1988. *Sparse distributed memory*. Cambridge MA: The MIT Press.

Thomas Kaspersson, Christian Smith, Henrik Danielsson, and Arne Jönsson. 2012. This also affects the context - Errors in extraction based summaries. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Robin Keskisärkkä and Arne Jönsson. 2012. Automatic Text Simplification via Synonym Replacement. In *Proceedings of the Fourth Swedish Language Technology Conference, Lund, Sweden*.

Robin Keskisärkkä. 2012. Automatic Text Simplification via Synonym Replacement. Master's thesis, Linköping University, Department of Computer and Information Science.

Hans P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(159-165).

Inderjeet Mani, Eric Bloedorn, and Barbara Gates. 1998. Using Cohesion and Coherence Models for Text Summarization. In *AAAI Technical Report SS-98-06*.

Ethel Martin. 1974. Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81(12):899–917.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Jahna C. Otterbacher, Dragomir R. Radev, and Airong Luo. 2002. Revisions that Improve Cohesion in Multi-

- document Summaries: A Preliminary Study. In *Proceedings of the Workshop on Automatic Summarization (including DUC 2002)*, Philadelphia, pages 27–36.
- Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.
- Erik D. Reichle, Alexander Pollatsek, Donald L. Fisher, and Keith Rayner. 1998. Toward a Model of Eye Movement Control in Reading. *Psychological Review*, 105(1):125–157.
- Magnus Sahlgren and Jussi Karlgren. 2005. Counting Lumps in Word Space: Density as a Measure of Corpus Homogeneity. *Analysis*, pages 151–154.
- Magnus Sahlgren. 2005. An Introduction to Random Indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.
- Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information Status Distinctions and Referring Expressions: An Empirical Study of References to People in News Summaries. *Computational Linguistics*, 37(4):811–842.
- Christian Smith and Arne Jönsson. 2011. Enhancing extraction based summarization with outside word space. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand*.