The Elementary Scenario Component Metric for Summarization Evaluation

Martin Kirilov[1, 2], **Daan Kolkman**[1, 2, 3], **Bert-Jan Butijn**[2, 4]

[1]Eindhoven University of Technology, Eindhoven, The Netherlands
[2]Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands
[3]University of Applied Sciences Utrecht, Utrecht, The Netherlands
[4]Erasmus School of Accounting and Assurance, Erasmus University, Rotterdam, The Netherlands
marti.kkirilov@gmail.com, d.kolkman@tue.nl, bjbutijn@gmail.com

Abstract

Evaluation is a fundamental step in the development of novel automatic summarization methods. The correlation between commonly used automatic evaluation metrics and golden standard human evaluations is often modest at best. Automatic evaluation metrics have thus not proven an alternative to human evaluation. This presents a problem to the progress of automatic summarization because evaluations conducted by people are time-consuming, inconsistent, and costly. We introduce the Elementary Scenario Component Metric (ESCM), which draws on the creative arts and scenario modelling literature. This metric does not require reference summaries, but uses twelve elementary scenario components, or a sub-selection thereof, to estimate the relevance of summaries instead. We show that the ESCM achieves a correlation of 0.89 with human evaluations and is less time-consuming than the creation of reference summaries.

1 Introduction

Although automatic summarization has a long history (Luhn, 1958), it remains a key challenge within Natural Language Processing (Fabbri et al., 2019). The aim of automatic summarization is to shorten a source text into a condensed version, conserving both the information content and the overall meaning (Kiyani and Tas, 2017). Automatic summarization methods can be classified among two axes: the summarization method and the number of input texts.

Irrespective of the automatic summarization method, an essential step in the development of a summarization system is the evaluation of generated summaries. Evaluation, however, is not without issues. Evaluation protocols differ from one paper to the next (Hardy et al., 2019) and evaluation metrics such as ROUGE are often used well beyond their intended scope (Liu and Liu, 2008). Moreover, (Fabbri et al., 2021) demonstrate that the system-level correlations between fourteen commonly used evaluation metrics and golden standard human evaluations for coherence, consistency, fluency, and relevance are mostly weak to moderate.

A commonality among most evaluation techniques is the need for reference summaries, referred to as gold-standard summaries. Most evaluation techniques calculate a score based on the comparison of the system generated summaries with the reference summaries. A drawback of employing reference summaries is that objectively establishing them is difficult (Steinberger and Ježek, 2009b). These reference summaries are human-written and therefore introduce a considerable level of subjectivity, since there is not a single perfect way of writing a text summary (Saziyabegum and Sajja, 2016). Moreover, writing these reference summaries by humans can be time-consuming and costly for large corpora (Giannakopoulos and Karkaletsis, 2013).

This paper presents the Elementary Scenario Component Metric (ESCM) which is grounded in work on scenarios in the creative arts and scenario modelling literature (De Kock, 2014). The ESCM does not require reference summaries, but utilizes elementary scenario components to estimate the relevance of summaries instead. The contribution of this study to the automatic summarization literature is twofold: ESCM reduces the dependence on people as human-written reference summaries are no longer a requisite for the evaluation of automatic summarization methods. More importantly, the ESCM is grounded in the creative arts literature and has a correlation of 0.89 with human evaluations, suggesting it may be a better proxy than other metrics currently in use. This paper is structured as follows: First we provide a brief overview of the literature on evaluation metrics for automatic summarization. Next we discuss the concept of scenarios as used in the creative arts and scenario planning literature. This informs discussion of the twelve Elementary Scenario Components. We then

introduce the ESCM and apply a sub-selection of five elementary scenario components in an experiment of multi-document crime case summarization. We conclude by offering some reflections and limitations of our work and offer avenues for further development of the ESCM. Our code can be found in the paper's GitHub page[1].

2 Related work

2.1 Evaluation protocols

Although much progress has been made, there is no consensus on how automatic summarization systems should be evaluated (van der Lee et al., 2019). A variety of metrics and procedures exist, Steinberger and Ježek (2009a) put forward a taxonomy of automatic summarization evaluation techniques. They suggest evaluation techniques can be broadly classified in two categories: intrinsic and extrinsic.

2.1.1 Intrinsic Evaluation

Intrinsic methods are based on how well the summary information content matches the information of a reference summary (Murray et al., 2008). Intrinsic evaluation can be further broken down into text quality evaluation and content evaluation. Evaluating the quality of the text is usually done by people, who rate different aspects of the summary on a predefined scale. These aspects of linguistic quality include grammatically, non-redundancy, reference clarity, and coherence and structure (Steinberger and Ježek, 2009a).

Content evaluation consists of co-selection measures such as precision, recall, and F-score, which ignore the fact that sentences can contain the same information even if written different and content-based measures which do not have that limitation. Content-based measures compare the words in a sentence, rather than the entire sentence, examples include cosine similarity (Louis and Nenkova, 2008), longest common subsequence, n-gram matching, pyramids (Nenkova and Passonneau, 2004), and Latent Semantic Analysis (LSA) based measures (Steinberger and Ježek, 2009a). The disadvantage of such measures is that they do not discriminate very well between summaries that involve differences in meaning (Mani, 2001). In effect, these measures are likely to work with extractive systems better than abstractive ones (Aries et al., 2019).

[1] https://github.com/ESCM-summarization/ESCM-evaluation

The most notable example of n-gram matching is the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric introduced by Lin (2004) which has been the de-facto standard for automatic evaluation of summarization in recent years (Yao et al., 2017). It works by measuring similarity between system generated summaries and reference summaries. Depending on the implementation, it can measure the overlap of uni-grams (ROUGE-1), bi-grams (ROUGE-2), Longest Common Subsequence (ROUGE-L), and others.

Aside from ROUGE, there are other metrics that have been used for summary evaluation like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BLANC (Lita et al., 2005). However, unlike ROUGE, these metrics were originally developed for evaluation of machine translation systems. Consequently, the use of these metrics in the automatic summarization literature is very limited.

A key problem with intrinsic evaluation is that these methods need to match the result summary with an "ideal summary", which is difficult to establish for a number of reasons (Steinberger and Ježek, 2009a). When people have to pick the most relevant sentences from documents in order to produce summaries, they frequently disagree in which sentences best represent the content of a document (Spärck-Jones et al., 2007). There is thus an inherent subjectivity to summarization evaluation (Lloret et al., 2018). Moreover, manual evaluation is expensive and the obtained results may be difficult to reproduce (Giannakopoulos and Karkaletsis, 2013).

More recently, researchers have developed various novel evaluation approaches which do improve upon the well-established intrinsic evaluation methods. For instance BERTScore (Zhang* et al., 2020), originally aimed at other tasks such as machine translation and image captioning. Or QAEval (Deutsch et al., 2021) and QuestEval (Scialom et al., 2021), which are both based on question-answering (QA) approaches. The latter might even be comparable to the ESCM, since it also doesn't require any ground-truth reference. Nonetheless, all three of these metrics require the use of additional models only for the evaluation of summaries. This, undoubtedly, would introduce a lot more unknowns within a summarization pipeline, and make it significantly more complex.

2.1.2 Extrinsic Evaluation

Extrinsic evaluation techniques determine the quality of a summary based on how it affects other tasks - a summary is considered good if it helps with solving these tasks (Gambhir and Gupta, 2017). These techniques are also known as task-based methods. Extrinsic evaluations have the advantage of assessing the utility of summarization in a task, so they can be of tremendous practical value to users of summarization technology (Mani, 2001). On the other hand, they are less helpful in providing insights on how the system actually performs the summarization. According to (Steinberger and Ježek, 2009a), the three most relevant tasks for extrinsic evaluation are document categorization, information retrieval, and question answering.

In the case of document categorization, the evaluation seeks to determine whether the generic summary is effective in capturing whatever information in the document is needed to correctly categorize the document (Steinberger and Ježek, 2009a). A document corpus and the topics of each document are needed in order to apply this method. The results of categorizing summaries are compared to results of categorizing full texts and random sentence extracts (Steinberger and Ježek, 2009a). The main metrics used in this case are the precision and recall of the categorization (or also their F-1 score) (Steinberger and Ježek, 2009a).

In the context of Information Retrieval (IR), summaries and full documents are used as input to an IR system. The similarity between how well the system works with the summaries as opposed to the full documents should serve as an indicator of the quality of summaries (Steinberger and Ježek, 2009a). Steinberger and Ježek suggest several methods to measure this similarity, namely Kendall's tau, Spearman's rank correlation, and linear correlation.

Question-answering is the third relevant task suggested by (Steinberger and Ježek, 2009a). The task is generally about reading comprehension — a human reads original documents or summaries and then answers a multiple-choice test (Mani, 2001). The idea is that if reading a summary allows a human to answer questions as accurately as they would by reading the original document, the summary is highly informative (Mani, 2001).

2.2 Elementary Scenario Components

Intrinsic evaluation techniques have a simple idea in common, they all compare two elements: the subject of evaluation (a text summary) and some kind of reference object (a reference summary). Scholars often use automatic summarization evaluation techniques that need either reference summaries, or some kind of specific task in order to measure the quality of a system generated summary.

With the Elementary Scenario Component Metric (ESCM), we employ a different approach that does not require reference summaries. Instead, our method is based on the idea that all relevant aspects of a some narrative can be described by the means of twelve elementary scenario components (De Kock, 2014). In the creative arts literature, a narrative is generated by a scenario that describes the interactions between characters (ibid.).

There is a rich literature on the nature and role of scenarios throughout history which can be traced back to Artisotle's Poetics. Aristotle is credited for being the first to distinguish between different components of scenarios and many have followed in his footsteps (Janko et al., 1984). Based on an extensive review of the literature on scenarios components in the creative arts, De Kock (2014) identified twelve Elementary Scenario Components (ESC-12) as the building blocks for any scenario. A list of all the components with a description is provided in Table 1. In this paper we suggest that these elementary scenario components can provide the foundation for a new automatic summarization system evaluation metric.

Our work is most closely related to intrinsic evaluation techniques based on "factoids". Such techniques attribute a score to text fragments based on their informativeness. These fragments are considered single coherent semantic units, such as "the Netherlands", "glass of water", and "the car arrived" (Van Halteren and Teufel, 2003). Radev et al. (2004) emphasized that it is necessary to determine not only what factoids should be included in the summary, but also how important they are. The pyramid method introduced by Nenkova and Passonneau (2004) builds on this idea to leverage multiple manually generated summaries. It demonstrates that factoids can be assigned weights based on those references summaries and how highly weighted units can be considered as more essential for a summary than not so highly weighted ones.

Component	Type	Description
Arena	Objective	The location where the story takes place.
Time(frame)	Objective	The time(frame) in which the story takes place.
Context	Objective	The set of circumstances that surround the story.
Protagonist	Objective	The main character of the story around whom the plot evolves.
Antagonist	Objective	The opposition against whom/which the protagonist must contend.
Motivation	Subjective	The psychological features that drive the protagonist.
Primary objective	Subjective	The way by which the protagonist attains his motivation.
Means	Objective	The methods or instruments by which the protagonist achieves his primary objective.
Modus operandi	Objective	The method of operation of the protagonist.
Resistance	Objective	The obstacles the protagonist has to overcome to be able to achieve his objective.
Symbolism	Interpretable	When a component carries a symbolic value for the protagonist, antagonist, or the audience.
Red herring	Interpretable	A misleading occurrence or indicator used to lead someone in the wrong direction of thought.

Table 1: A list of the ESC12 (De Kock, 2014)

In terms of these contributions, the ESCM provides a more general framework for determining relevant factoids, which could then be ranked using the pyramid method.

The ESC-12 are divided into three categories - objective, subjective, and interpretable components: Objective components comprise observable phenomena and are not related to the protagonist's individual feelings and interpretations, Subjective components reflect the protagonist's individual interpretation of experiences and interpretable components do not have a meaning until interpreted by a third party (De Kock, 2014).

3 Elementary Scenario Components Metric

We propose a new evaluation metric based on the ESC-12. Below we introduce the procedure for applying the ESC-12 to an automatic summarization system.

1. Determine relevant ESC and operationalize variable mapping. De Kock (2014) argues that ESC-12 represent general categories that occur in any type of scenario. However, careful tuning of the variables making up the components is necessary to fit a particular domain. For instance, although the *Arena* may be relevant to a narrative on historic geography and a narrative pertaining a criminal case, their operationalization would be different.

2. Annotate data. The metric requires an annotated dataset, with labels for (sub)set of ESCs-12 for each article or story (one or more documents can be referring to one story). After a system generates summaries for the input dataset, the evaluation metric is calculated based on the presence of the corresponding ESCs for each story in each summary relative to the presence of these components in the input texts. The formula which is used to calculate this can vary depending on fine-tuning for optimal results.

3. Compute ESCM. In the third step, the ESCM is computed. This requires ESC labeled texts, and summaries generated on these texts by the summarization system.

3.1 Elementary Scenario Component Metric Calculation

The first step is to determine which of the labelled components are available in the pre-processed input texts. This is of importance, because some components might be missing from the input texts after certain levels of pre-processing. For instance, if the inputs are truncated to 500 tokens, some components might not be included. Another possibility - for multi-document summarization- is that a component is present in one source text, which has been discarded in the selection step (e.g. if the Protagonist is mentioned only in document #3, but documents #1 and #2 are selected for truncation).

The second step consists of determining which of the components in the input texts are present in the summaries. This is a fairly complicated process in the context of automatic summarization, therefore we illustrate it with an example. Suppose that the algorithm is trying to match the name of the antagonist obtained from the dataset and present in the input text (e.g. John Doe). In case the character string "John Doe" is also in the summary, there is a 100% match. The problem arises when the summary contains some variant of the antagonist's name, for instance "John D." or "J. D.". Therefore, we include an approximate string matching technique in the evaluation metric algorithm. This allows the algorithm to distinguish between complete and partial string matching.

In the final step, the algorithm calculates a sum of the individual scores of the components of inter-

est. Each individual score can vary between 0 and 100, where 100 means that the two components are identical, and the smaller the score gets, the more different the components are. Equation 1 illustrates how the evaluation score is calculated:

$$\text{ESCM} = \frac{\sum_{(M,R)\in C} \frac{\sum_{\substack{x\in M \\ y\in R}} \frac{FuzzyScore(x,M)}{FuzzyScore(y,R)}}{K}}{N} \quad (1)$$

Where N is the total number of cases, M and R are a pair of model generated summary and reference input text for a single case from the set of all cases C. Furthermore, x and y are pairs of ESCs from the model generated summary M and the reference input text R respectively. K is the number of components that are considered when calculating the metric, which in this case is 5. $FuzzyScore(arg_1, arg_2)$ is a function that returns a number between 0 and 100 depending on how accurately is the ESC label arg_1 represented in the text arg_2.

Finally, a specific threshold is introduced for each of the components. The idea behind these thresholds is to determine if the matched component is referring to the same n-gram as the label, or it is a completely different n-gram. Again, an example may help illustrate how this works: suppose the Protagonist's name is "George S.", but the best match the algorithm is able to find is the string "was reported" with a score of 60. However, if the Protagonist's name was "Brian Nijhof", then the best match would be "Brian N." with a score of 74. The introduction of thresholds helps the algorithm to discard the first example, but keep the second.

The thresholds are used as follows: For each case the algorithm checks if all of the ESCs have a score above the threshold in the input texts for that case. If any of the ESCs have a lower score, the whole case is discarded from the calculation of the evaluation metric. This is necessary to ensure that all cases used in the calculation have all ESCs in the input texts. However, the thresholds are used in a different way when handling the summary texts. For each case, if the summary ESC score is below the threshold it means that it is wrong, and is thus set to 0. This process is illustrated with a few examples in Table 2.

In the first example, the name of the protagonist is mentioned with a score of 70 in the input texts, which is just on the threshold, and consequently, it is also checked in the summary text. However, in the summary, the fuzzy matching returns a score of 46, which is below the threshold. Then the algorithm sets the ESC summary score to 0, and the ESC relative score becomes $0/70 = 0\%$. The second example shows that if the fuzzy score of an ESC is below the threshold for the input texts, the ESC does not receive a relative score, and the whole case is discarded. For the last two examples, both of the input texts and summary text fuzzy scores are above the thresholds, and therefore the relative score can be calculated as a fraction between the fuzzy scores.

3.2 Fuzzy String Matching

For multi-document summarization there is higher chance of discrepancies between the data labels and the actual strings these labels refer to in the articles. Information in some news articles might be missing or different compared to other news articles. For example, let us say "John Doe" is the antagonist label for a case, while some of the articles only contain the name "John D.". Consequently, it would be required to measure how similar these two strings are. Directly matching strings is not a viable option, because there are many examples where the difference is only a few characters, and it is evident that they are referring to the same thing. Thus, there is a need for a method which would allow for the implementation of fuzzy string matching. Such a method is Levenshtein distance (Levenshtein, 1966), of which we used the Token Sort Ratio method for our FuzzyScore.

4 Experiment

To demonstrate the effectiveness of the ESCM, we conduct an experiment with automatic summarization of crime cases. There are plenty examples in literature for application of NLP techniques for crime data analysis (Ku and Leroy, 2014; van Banerveld et al., 2014; Ku et al., 2008; Iriberri and Leroy, 2007; Wang et al., 2007). The vast majority of such studies focus on information extraction (e.g. Named-entity recognition), crime classification, and crime analysis (Ku and Leroy, 2014). However, there is a lack of research about automatic summarization in the crime domain. The only example of a summarization system for crime texts that was found at the time of writing is the SALOMON project (Moens et al., 1997; Moens, 2000). For our experiment we implement three models: Hi-Map (Fabbri et al., 2019), Transformer

ESC label	Input texts		Summary text		ESC Relative score
	ESC match	Fuzzy score	ESC match	Fuzzy score	
Rudolf Käsenbier	Rudolf K.	70	of Enschede	46	0%
Anouar B.	book and	50	-	-	-
Henk Haalboom	Henk Haalboom	100	Haalboom	76	76%
Michael E.	Michel E.	94	Michel E.	94	100%

Table 2: Examples of different combinations of fuzzy scores and their corresponding relative score for the Protagonist ESC in the Homicide dataset.

(Vaswani et al., 2017), and TextRank (Mihalcea, 2004). For the first two models we tried a truncation of 500 and 1000 tokens. For TextRank we only implemented a 500 token model. These truncation lengths are chosen to be inline with the setting of relevant studies such as Fabbri et al. (2019)

4.1 Homicide dataset

The original version of the Homicide dataset has been created by Pandora Intelligence. It consists of 100 manually chosen homicide cases that occurred in the Netherlands. For each case there are relevant data about some of the ESCs (usually not all), as well as multiple source articles about the case. These articles were web-scraped from manually selected URLs. The methodology of selecting these URLs is mainly based on the results of Google Search queries for the most popular Dutch homicide cases from the last few decades. The dataset comes in a Dutch and English version. The Dutch version is created from web-scraping these URLs of mostly Dutch news websites. The English version is made by domain experts, who automatically translated the Dutch version via Google Translate API[2], and manually reviewed all translations.

On average, the number of articles per case is 13.86 and there are no cases with less than five sources. This high number of source articles makes the Homicide dataset very suitable for multi-document summarization. A more detailed information about the distribution of the source articles can be found in Table 4.

4.2 Application of the ESCM procedure

The ECSM procedure offers an effective and durable set of components to describe, characterise and model a criminal incident (De Kock, 2014). We follow the procedure outlined in section three. For means of illustration, and to limit the degree of subjectivity in the labelling, we exclude both subjective and interpretable components. In *Step 1* of the procedure, we thus select the following objective components: *Arena*, *Timeframe*, *Protagonist*, *Antagonist*, and *Modus operandi*.

Next, we provide the operationalization of these concepts in the context of our experiment. More specifically, we specify the variable or set of variables for each ESC that we have selected. Table 3 provides an overview of this mapping. A subset of the Homicide dataset was then made, only consisting of cases, which contain annotations about all 5 pre-selected objective components mentioned in *Step 2*. As a result, this yielded 26 cases which were suitable for the application of our ESCM evaluation. In *Step 3* we compute the ESCM using the following thresholds for fuzzy string matching: Arena 65, Timeframe 75, Protagonist 70, Antagonist 75, Modus operandi 75. The thresholds were manually fine-tuned upon basic data exploration and characteristics of the ESC annotations.

4.3 Evaluation

We evaluate our procedure and the ESCM using a questionnaire administered to several expert respondents. We followed the guidelines for expert evaluation proposed by van der Lee et al. (2019). Although evaluation by a more general audience is sometimes preferred, we opted for expert evaluation in an effort to collect the highest quality data.

The survey for the human evaluation experiment was distributed among police offcers from the Dienst Regionale Informatieorganisatie (DRIO) department of Police Oost-Brabant. In total, 21 participants filled in the survey. This can be regarded as a high number of participants for an expert-focused study, which typically use up to four experts (van der Lee et al., 2019). The cases presented in each variant were randomly selected from the 26 available.

4.3.1 Text quality criteria

It is very common for automatic summarization studies which perform human evaluation to report text quality. However, text quality criteria differ across tasks, and there is a significant variety in

[2] https://cloud.google.com/translate

Component	Operationalization
Arena	The city where the homicide took place.
Timeframe	The date on which the crime took place, which is provided in a DD-MM-YYYY format.
Protagonist	The name of the murderer or murderers.
Antagonist	The name of the victim or victims.
Means	The murder weapon.
Modus operandi	The type of murder (e.g. manslaughter or first degree murder).

Table 3: The operationalization of the ESCM for the Homicide dataset

Number of sources	Frequency
Up to 5	2
From 6 to 10	23
From 11 to 15	44
From 16 to 20	23
More than 20	8

Table 4: Distribution of source articles in the Homicide dataset.

naming conventions for measures of text quality (van der Lee et al., 2019). There is also absence of common evaluation guidelines for NLG tasks (Belz and Hastie, 2014), which means that the measured criteria should be explicitly defined when implementing a human evaluation experiment (van der Lee et al., 2019). Although the ESCM is primarily intended to measure accuracy, following van der Lee et al. (2019) also measure relevance, and fluency as text quality criteria.

4.3.2 Questionnaire design

The survey consisted of three parts. Participants were presented with an introduction to the research topic, the ESC framework, and the goals of the survey. The second part was comprised of five different text summaries, each followed by a set of ESC-related questions and two text quality questions. The third and final part included general demographic questions and concluded the survey.

Based on the goals of this evaluation experiment, it was decided to include three types of questions – text quality questions, questions evaluating the accuracy of ESCs in various text summaries, and general demographic questions. For the first two types, we use a 5-point Likert scale, the full survey can be found on GitHub.

4.4 Results

4.4.1 Human evaluation

The average scores for each case-summary combination are reported in Table 5. Let X be the variable containing ESCM scores of multiple case summary variants. More specifically, X contains the 15 scores labelled as *ESCM* in Table 5. Let Y be the variable containing the average scores per case and summary variant obtained from the survey results. Y contains the 15 scores labelled with *Survey* in Table 5. Upon calculating Pearson's Correlation Coefficient, the variables X and Y were found to be strongly positively correlated $(\boldsymbol{r(13) = .89, p < .001})$. A scatter plot with a trend line is illustrated in Figure 1.

The results of the text quality, averaged per criterion are presented in Table 6. As with the first part of the questionnaire, we used a 5-point Likert scale. Expectedly, the Transformer models score best in terms of subjective fluency and relevance, even though the results for all model combinations are quite close to the average on the scale.

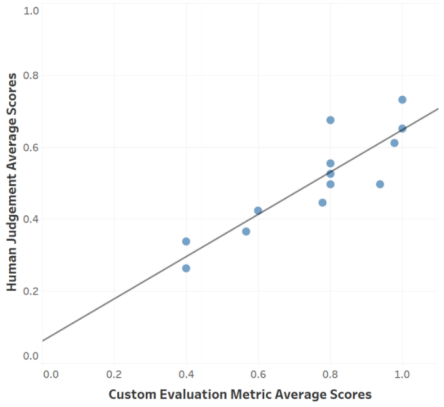

Figure 1: Scatter plot of the average survey and ESCM scores.

van der Lee et al. do recommend to always report inter-rater reliability (IRR). IRR measures degree of consensus among ratings provided by various human evaluators. Krippendorff's alpha Krippendorff (1970) is a frequently used IRR measure and can be used regardless of the number of observers, levels of measurement, sample sizes, and presence or absence of missing data Krippendorff (2004). Three distinctive groups of summaries were presented to three different groups of participants, we report Krippendorff's alpha for each in Table 7. The coefficients show positive agreement among the participants in the human evaluation.

Case ID	Score type	Summary variant (model, truncation, selection)				
		Hi-MAP 500; 3	Transformer 500; 3	TextRank 500; 3	Hi-MAP 1000; 2	Transformer 1000; 2
Case 44	ESCM	77.80	100.00	80.00	40.00	97.80
	Survey	64.57	85.14	87.43	53.71	81.14
Case 12	ESCM	80.00	60.00	100.00	93.80	56.60
	Survey	69.71	62.29	93.14	69.71	56.57
Case 31	ESCM	40.00	80.00	80.00	80.00	40.00
	Survey	53.71	75.43	72.57	75.43	46.29

Table 5: Average scores of the ESCM and the survey results.

Criterion	Summary variant (model, truncation, selection)				
	Hi-MAP 500; 3	Transformer 500; 3	TextRank 500; 3	Hi-MAP 1000; 2	Transformer 1000; 2
Fluency	2.52	**3.14**	2.67	2.67	3.05
Relevance	2.57	**3.14**	3.10	2.86	2.29

Table 6: Text quality criteria average scores for each of the summary variants presented in the human evaluation experiment.

Case ID	Number of annotators	Number of questions	Krippendorff's alpha
Case 44	7	25	.38
Case 12	7	25	.60
Case 31	7	25	.36

Table 7: Inter-Rater Reliability coefficients for the human evaluation experiment.

4.4.2 Robustness check

To further explore the robustness of our findings we asked independent annotators to write golden standard summaries for the same homicide cases used in our previous experiments, namely 12, 31, and 44. We calculated four different ROUGE measures per case and averaged all ROUGE F-1 scores for all cases per ROUGE type. We calculated the Pearson's Correlation Coefficient of these values in relation to the average ESCM and Human evaluation scores. The results are presented in Table 8. The correlation of the best match-up (R-SU4 and human evaluators) is considerably lower than that of the ESCM and the human evaluation scores. R-SU4 performs best out of the ROUGE metrics which in line with early findings (Lin, 2004) and thus is further proof of the robustness of our experiment.

	R-1	R-2	R-L	R-SU4
ESCM	0.60	0.49	0.53	0.67
Human	0.64	0.58	0.63	0.72

Table 8: Correlation coefficients for the ESCM, human evaluation scores, and various ROUGE scores.

5 Conclusion

We presented a novel automatic evaluation metric for automatic summarization based on the ESC framework: the ESCM. The ESCM is a recall-based metric and we recommend it be used alongside precision-based metrics. With our experiments, we demonstrate the capabilities of the ESCM. However, our metric is subject to some limitations. With only 26 cases, the dataset we used is relatively small. Furthermore, in relation to the calculation of the ESCM, we used a threshold to determine if an ESC is contained in a text or not. These thresholds were determined based on the results from the fuzzy string matching. It is unclear whether these thresholds would be generalizable to other (non)homicide-related datasets.

Our findings show the potential of the ESCM, but more research is necessary to explore its usefulness. Future work could experiment with the ESCs by including more components from the subjective or interpretable type. Furthermore, the selected components could be utilized better by increasing the level of detail (e.g. using the full date instead of just the year). A next step in the validation of the ESCM could be done using the dataset from Text Analysis Conference (TAC) 2010 summarization track. Part of the TAC 2010 dataset consists of texts and summaries of criminal attacks, that are labelled in a similar manner to the ESCM.

Although the primary focus of the ESCM is relevance, it achieved a strong correlation with the average of human evaluations for relevance, accuracy, and fluency. Moreover, those that labelled the texts and wrote summaries reported that labelling for the ESCs was considerably less time consuming. Although this is merely anecdotal evidence, it echoes findings by Lloret et al. (2013) who suggest that producing reference summaries takes 8 – 10 times longer than answering a series of questions about a text. Based on this and the high correlation with human evaluation, we believe the ESCM may present a useful alternative to existing metrics, especially in applications domains that are under-resourced or where writing and evaluating summaries requires domain expertise.

References

Abdelkrime Aries, Walid Khaled Hidouci, et al. 2019. Automatic text summarization: What has been done and what has to be done. *arXiv preprint arXiv:1904.00688*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anja Belz and Helen Hastie. 2014. *Comparative evaluation and shared tasks for NLG in interactive systems*, page 302–350. Cambridge University Press.

P. De Kock. 2014. *Anticipating criminal behaviour: Using the narrative in crime-related data*. Ph.D. thesis, Tilburg University.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66.

George Giannakopoulos and Vangelis Karkaletsis. 2013. Summary evaluation: Together we stand npower-ed. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 436–450. Springer.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

A. Iriberri and G. Leroy. 2007. Natural language processing and e-government: Extracting reusable crime report information. In *2007 IEEE International Conference on Information Reuse and Integration*, pages 221–226.

Richard Janko et al. 1984. *Aristotle on comedy: towards a reconstruction of Poetics II*, volume 2. Univ of California Press.

Farzad Kiyani and Oguzhan Tas. 2017. A survey automatic text summarization. *Pressacademia*, 5(1):205–213.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Chih Hao Ku, Alicia Iriberri, and Gondy Leroy. 2008. Natural language processing and e-government: crime information extraction from heterogeneous data sources. In *Proceedings of the 9th Annual International Conference on Digital Government Research, Partnerships for Public Innovation, DG.O 2008, Montreal, Canada, May 18-21, 2008*, volume 289 of *ACM International Conference Proceeding Series*, pages 162–170. Digital Government Research Center.

Chih-Hao Ku and Gondy Leroy. 2014. A decision support system: Automated crime report analysis and classification for e-government. *Government Information Quarterly*, 31(4):534 – 544.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. 2005. Blanc: Learning evaluation metrics for mt. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 740–747, USA. Association for Computational Linguistics.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2013. Analyzing the capabilities of crowdsourcing services for text summarization. *Language resources and evaluation*, 47(2):337–369.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52(1):101–148.

Annie Louis and Ani Nenkova. 2008. Automatic summary evaluation without human models. In *TAC*.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Inderjeet Mani. 2001. Summarization evaluation: An overview. In *Proceedings of the 2nd NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization*, Tokyo: National Institute of Informatics.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 170–173, Barcelona, Spain. Association for Computational Linguistics.

Marie-Francine Moens. 2000. *Automatic Indexing and Abstracting of Document Texts*, volume 6. Springer Science & Business Media.

Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. 1997. Abstracting of legal cases: The salomon experience. In *Proceedings of the 6th International Conference on Artificial Intelligence and Law*, ICAIL '97, page 114–122, New York, NY, USA. Association for Computing Machinery.

Gabriel Murray, Thomas Kleinbauer, Peter Poller, Steve Renals, Jonathan Kilgour, and Tilman Becker. 2008. Extrinsic summarization evaluation: A decision audit task. In *Machine Learning for Multimodal Interaction*, pages 349–361, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hltnaacl 2004*, pages 145–152.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Saiyed Saziyabegum and Priti S. Sajja. 2016. Literature review on extractive text summarization approaches. *International Journal of Computer Applications*, 156(12):28–36.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karen Spärck-Jones, Stephen E Robertson, and Mark Sanderson. 2007. Ambiguous requests: implications for retrieval tests, systems and theories. In *ACM SIGIR Forum*, volume 41, pages 8–17. ACM New York, NY, USA.

Josef Steinberger and Karel Ježek. 2009a. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Josef Steinberger and Karel Ježek. 2009b. Text summarization: An old challenge and new approaches. In *Foundations of Computational, IntelligenceVolume 6*, pages 127–149. Springer.

Maarten van Banerveld, Nhien An Le-Khac, and M. Tahar Kechadi. 2014. Performance evaluation of a natural language processing approach applied in white collar crime investigation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8860:29–43.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Hans Van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 57–64.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

X. Wang, D. E. Brown, and J. H. Conklin. 2007. Crime incident association with consideration of narrative information. In *2007 IEEE Systems and Information Engineering Design Symposium*, pages 1–4.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.