# GeSERA: General-domain summary evaluation by relevance analysis

Jessica López Espejel, Gaël de Chalendar, Jorge Garcia Flores, Thierry Charnois, Ivan Vladimir Meza Ruiz

# GeSERA: General-domain Summary Evaluation by Relevance Analysis

**Jessica López Espejel**[*†], **Gaël de Chalendar**[*], **Jorge Garcia Flores**[†], **Thierry Charnois**[†], **and Ivan Vladimir Meza Ruiz**[‡]
[*] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[†] CNRS-LIPN-Université Sorbonne Paris Nord, France
[‡] IIMAS, UNAM, Mexico
jessicanayeli.lopezespejel@univ-paris13.fr, gael.de-chalendar@cea.fr,
jgflores@lipn.fr, thierry.charnois@lipn.univ-paris13.fr, ivanvladimir@turing.iimas.unam.mx



## Abstract

We present GeSERA, an open-source improved version of SERA for evaluating automatic extractive and abstractive summaries from the general domain. SERA is based on a search engine that compares candidate and reference summaries (called queries) against an information retrieval document base (called index). SERA was originally designed for the biomedical domain only, where it showed a better correlation with manual methods than the widely used lexical-based ROUGE method. In this paper, we take out SERA from the biomedical domain to the general one by adapting its content-based method to successfully evaluate summaries from the general domain. First, we improve the query reformulation strategy with POS Tags analysis of general-domain corpora. Second, we replace the biomedical index used in SERA with two article collections from AQUAINT-2 and Wikipedia. We conduct experiments with TAC2008, TAC2009, and CNNDM datasets. Results show that, in most cases, GeSERA achieves higher correlations with manual evaluation methods than SERA, while it reduces its gap with ROUGE for general-domain summary evaluation. GeSERA even surpasses ROUGE in two cases of TAC2009. Finally, we conduct extensive experiments and provide a comprehensive study of the impact of human annotators and the index size on summary evaluation with SERA and GeSERA.

## 1 Introduction

Automatic summary evaluation is a challenging task in Natural Language Processing (NLP). Evaluation is usually done by humans, but manual evaluation is subjective, costly and time expensive (Lin and Hovy, 2002). Automatic evaluation methods (Lin, 2004a; Torres-Moreno et al., 2010; Zhao et al., 2019; Zhang et al., 2020) are an alternative to save time for users who extract the most relevant content from the web using Automatic Text Summarization systems (ATS). There exist two types of evaluation approaches: (1) manual evaluation methods like Pyramid (Nenkova and Passonneau, 2004) and Responsiveness, where human intervention is mandatory, and (2) automatic evaluation methods, where human intervention can be needed as a ground-truth reference (Lin, 2004a; Cohan and Goharian, 2016) or not (Torres-Moreno et al., 2010; Cabrera-Diego and Torres-Moreno, 2018).

*Summary Evaluation by Relevance Analysis* (SERA) (Cohan and Goharian, 2016) is an automatic evaluation method that partially relies on human references to evaluate abstractive summaries from the biomedical domain. It was proposed as an alternative to ROUGE (Lin, 2004a), the widely used automatic metric, that is based on lexical overlaps between candidate and reference summaries. ROUGE is unfair to evaluate abstractive summaries where the ATS paraphrases the text instead of just copying-pasting chunks of it (Cohan and Goharian, 2016).

SERA is based on content relevance and is fairer to evaluate abstractive summaries because it attributes high scores to summaries that are lexically different but semantically related. However, it surpasses ROUGE on the biomedical domain only. In this paper, we modify SERA to make it usable for generic collections. We propose the following contributions:

**1.** Implement an open-source version of SERA from scratch.

**2.** Propose GeSERA (*General-domain SERA*), an improved version of SERA that is domain-independent.

**3.** Conduct extensive experiments with two large indexes (AQUAINT-2 (Graff, 2002) and Wikipedia) and three summarization datasets

(TAC[1] 2008, TAC2009, CNN-Daily Mail (Bhandari et al., 2020)). These datasets are well-suited for general-domain and news abstractive summary evaluation. GeSERA achieves competitive results compared to a range of state-of-the-art evaluation approaches on both abstractive and extractive summaries.

**4.** Make the code and our Wikipedia dataset publicly available to help future researchers.[2]

**5.** Conduct extensive experiments on the impact of the index size and human annotators on summary evaluation with SERA and GeSERA.

## 2 Proposed approach

### 2.1 Baseline: SERA

SERA (Cohan and Goharian, 2016) is based on a content relevance between a candidate summary and its corresponding human-written reference summaries using information retrieval. SERA compares these summaries (called queries) against a set of documents from the same domain (called index), and compares the overlap of retrieved results. SERA refines queries in three different manners (1) *Raw text* - only stop words and numbers are removed, (2) *Noun phrases (NP)* - only noun phrases are kept, and (3) *Keywords (KW)* - only unigrams, bigrams, and trigrams are kept.

SERA is defined in Equation 1.

$$SERA = \frac{1}{M}\sum_{i=1}^{M}\frac{|R_C \cap R_{G_i}|}{|R_C|} \quad (1)$$

where: $R_C$ is the list of retrieved documents for the candidate summary $C$, $R_{G_i}$ is the ranked list of retrieved documents for the reference summary $G_i$, and $M$ is the number of reference summaries.

Another variant of SERA is called SERA-DIS. It takes the order of retrieved documents into consideration (Equation 2).

$$SERA - DIS = \frac{\sum_{i=1}^{M}(\sum_{j=1}^{|R_C|}\sum_{k=1}^{|R_{G_i}|} X_{j,k}}{M * D_{max}}$$
$$X_{j,k} = \begin{cases} \frac{1}{log(|j-k|+2)} & if\ R_C^{(j)} = R_{G_i}^{(k)} \\ 0 & otherwise \end{cases} \quad (2)$$

where: $R_C^{(j)}$ is the $j^{th}$ result in the ranked list $R_C$, and $D_{max}$ is the maximum achievable score used for normalization. In both SERA variants, retrieved results are truncated at 5 and 10 documents (hence the notations SERA-5 and SERA-10 in Section 3). Cohan and Goharian (2016) used articles from PubMed[3] as a index, and summaries from TAC 2014 as queries.

The intuition behind SERA is that a summary context is represented by its most related articles. Thus, two summaries related to the same documents are semantically related, even if they are lexically different. Consequently, SERA is fairer to evaluate abstractive summaries contrarily to the lexical-based ROUGE. However, SERA suffers from a series of limitations: (1) the code is not open-source, (2) no information was provided concerning the subset of PubMed used as an index, and (3) PubMed is specialized in the biomedical domain only. The first two drawbacks make SERA unusable by the community, while the third restricts its usage to the biomedical domain.

### 2.2 GeSERA: General-domain SERA

We build on SERA merits and limitations to propose GeSERA, an open-source version of SERA that evaluates summaries from the general domain. Novelties of GeSERA are the index pool and query reformulation adapted to the evaluation of summaries from the general domain.

**Index documents pool -** Differently from SERA, GeSERA enables general-domain summary evaluation. It is thus necessary to replace the biomedical index used by Cohan and Goharian (2016) with a set of documents related to the general domain. We build many indexes using a variant number of articles from Wikipedia and AQUAINT-2. See Subsection 3.1 for more details.

**Query reformulation (QR) -** It improves retrieval process by removing unnecessary terms from the text. Therefore, we propose a different approach to refine queries in GeSERA that is better suited for general domain summaries.

According to Kieuvongngam et al. (2020), nouns in generated summaries represent more accurately the information conveyed by the original abstracts than other POS tags. This study was conducted on Covid-19 medical texts and can explain why SERA achieved a higher correlation than ROUGE for the TAC 2014 biomedical dataset. We led an analysis that consists of analyzing Part-Of-Speech (POS) tags distribution

[1] https://tac.nist.gov/

[2] https://github.com/JessicaLopezEspejel/GeSERA/

[3] http://www.ncbi.nlm.nih.gov/pmc/

for PubMed (biomedical dataset built by Cohan et al. (2018)), AQUAINT-2 (news corpus), and Wikipedia (general domain encyclopedia). Figure 1 shows bar plots for percentages of nouns, verbs, adjectives (Adj.), prepositions (Prep.), and the total percentage of other tags (Others).

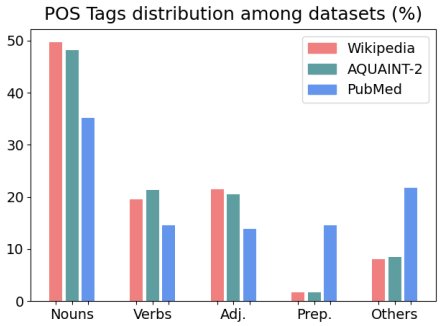



Figure 1: POS Tags distribution percentages for Wikipedia, AQUAINT-2, and PubMed datasets

Our analysis of three datasets which describe different domains confirms the observation of Kieuvongngam et al. (2020) for PubMed. However, it shows that the percentages of verbs and adjectives are higher in AQUAINT-2 and Wikipedia than in PubMed. Equally important, there is a remarkable absence of prepositions in Wikipedia and AQUAINT-2. Based on our analysis, we propose to reformulate queries in GeSERA by only keeping tokens tagged with nouns, verbs, and adjectives, the three most frequent tags in the news and general domain corpora.

**Search engine -** A semantic-based retrieval approach is crucial when handling abstractive summaries. In order to compare the queries against the index, a search engine is needed for information retrieval and scoring. We use the Whoosh[4] search engine with the BM25F (*Best Match 25 Model with Extension to Multiple Weighted Fields*) ranking function (Zaragoza et al., 2004). This model is widely used for semantic search (Pérez-Agüera et al., 2010; Robertson and Zaragoza, 2009). It consists in weighting terms according to their field importance, combining them, and using the resulting pseudo-frequencies for ranking.

## 3 Experiments

SERA was developed in the context of scientific biomedical article summarization with the idea that its semantic specificity is particularly useful for this domain. We hypothesize that if we reformulate queries properly and change the index pool, SERA can assess summaries from other domains for both abstractive and extractive summarization. This hypothesis is based on the fact that SERA considers terms that are not lexically equivalent but are semantically related. We conduct extensive experiments on SERA and GeSERA to test our hypothesis.

### 3.1 Index datasets

The index is a key component of GeSERA approach insofar as it should describe the same domain as the queries. The number of documents in the index is also decisive as we will show in Subsection 4.1. We describe briefly query and index datasets hereafter and provide more information in the supplementary material.

- **AQUAINT-2** (Graff, 2002) is a news corpus built from various sources. We vary the size of the index to include $\mathcal{I} = \{10000, 15000, 30000, 60000, 89760, 179520, 825148\}$ randomly-selected documents.
- **Wikipedia** is a free encyclopedia that contains various information from the general domain. We vary the size of the index to include $\mathcal{I} = \{10000, 15000, 30000, 1778742\}$ randomly-selected documents.

### 3.2 Query datasets

Candidate and reference summaries from the news datasets TAC2008 and TAC2009, and the CNN Daily Mail (CNNDM) version published by Bhandari et al. (2020) are used as queries.

**TAC2008** (/and **TAC2009**) are subsets of AQUAINT-2. They contain 5568 (/4840) candidate summaries proposed by 58 (/55) participants, and 384 (/352) reference summaries, respectively. These datasets are designed for multi-document extractive summarization (one summary is shared by a set of documents).

**CNN Daily Mail** (Bhandari et al., 2020) is a news dataset, and is of great interest to us because it contains candidate summaries obtained from both extractive and abstractive systems. It consists of 100 reference summaries, having each 25 candidate summaries generated by 11 extractive and 14 abstractive systems. It is designed for mono-document abstractive and extractive summarization (one summary for each document).

[4] https://whoosh.readthedocs.io/en/latest/intro.html

### 3.3 Baselines

We compare GeSERA with some of the most influential evaluation metrics from the literature:

- **ROUGE** and **SERA** - two automatic evaluation approaches that rely on human intervention. ROUGE has many variants, but we only report the most popular ones: ROUGE-N ($N = \{1, 2\}$ is the n-gram size) and ROUGE-L (Longest Common Subsequence). For each variant, we report the F-score, Recall and Precision.
- **MoverScore** (Zhao et al., 2019) and **BERTScore** (Zhang et al., 2020) - two automatic evaluation approaches based on BERT.
- **JS-2** (Lin et al., 2006) - Jensen-Shannon divergence between bigram's distribution of the candidate and reference summaries.
- **SummTriver** (ST) (Cabrera-Diego and Torres-Moreno, 2018) - based on trivergence, i.e. a composition ($\mathcal{T}_c$) or multiplication ($\mathcal{T}_m$) of Kullback-Leibler (KL), Jensen-Shannon (JS), and smoothed Jensen-Shannon (sJS) divergences. It does need a human reference.
- **FRESA** (Torres-Moreno et al., 2010) - an automatic evaluation method that does not rely on human intervention. It is based on Jensen-Shannon divergence and has four variants: unigrams (FRESA-1), bigrams (FRESA-2), trigrams (FRESA-3), and skip-grams (FRESA-4).

More information is in Section 5. Implementation details are in the supplementary material.

### 3.4 Evaluation methodology

To compare GeSERA with other state-of-the-art methods, we measure the correlation between the scores provided by each automatic and manual evaluation methods. The manual evaluation approaches used here are:

- **Pyramid** (Nenkova and Passonneau, 2004) exploits the content distribution in human summaries using *Summary Content Units* (SCUs) based on their frequency in the summary corpus.
- **LitePyramid** - (Shapira et al., 2019) is a crowdsource-based lightweight version of Pyramid that relies on statistical sampling instead of exhaustive SCU extraction and testing.
- **Responsiveness** measures the summary's linguistic quality.

Following Cohan and Goharian (2016), we use the correlation metrics: (1) *Pearson* (Benesty et al., 2009), (2) *Spearman* (Kokoska and Zwillinger, 2000), and (3) *Kendall tau-b* (Kendall, 1945).

## 4 Results and discussion

In Subsection 4.1, we vary the index size in SERA and GeSERA and study the variation of their performance on TAC datasets by averaging the score of all four manual annotators $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$, and $\mathcal{A}_4$. Once we determine the best index size, we present in Subsection 4.2 the correlations using the best index size of each method.

### 4.1 Impact of the index size on the performance of SERA and GeSERA

We first present the impact of the indexes sizes built from Wikipedia on SERA and GeSERA, then the ones built from AQUAINT-2.

**Wikipedia Index -** Figures 2-b and 2-d show Pearson correlations of SERA and GeSERA with Pyramid when indexing different values of $\mathcal{I}$ from the Wikipedia dataset, and when using TAC2008 and TAC2009 as query datasets, respectively. Figures show that the best score (0.902) is obtained with SERA-DIS-NP-10 using $\mathcal{I} = 30,000$ for TAC2008 while the best one (0.957) for TAC2009 is obtained with GeSERA-DIS-10 with $\mathcal{I} = 10,000$. We thus use in Subsection 4.2 an index size $\boldsymbol{\mathcal{I} = 30,000}$ and $\boldsymbol{\mathcal{I} = 10,000}$ for TAC2008 and TAC2009, respectively. Surprisingly, the worst scores are obtained with $\mathcal{I} = 1,778,742$, the largest and more diversified index size corresponding to all documents from our Wikipedia corpus.

**AQUAINT-2 Index -** Figures 2-a and 2-c show the Pearson correlation coefficients of SERA and GeSERA with Pyramid when indexing different values of $\mathcal{I}$ from the AQUAINT-2 dataset, and when using TAC2008 and TAC2009 as query datasets. Similarly to Wikipedia, figures show that overall, the best results are obtained with small index sizes. The best score (0.928) was obtained with GeSERA-5 using $\mathcal{I} = 15,000$ for TAC2008, while the best one (0.947) for TAC2009 is obtained with both SERA-DIS-NP-10 and GeSERA-DIS-5 using $\mathcal{I} = 179,520$. Note that results with $\mathcal{I} = 10,000$ are comparable with those obtained with $\mathcal{I} = 179,520$ for TAC2009. We thus use in Subsection 4.2 an index size $\boldsymbol{\mathcal{I} = 15,000}$ and $\boldsymbol{\mathcal{I} = 179,520}$ for TAC2008 and TAC2009, respectively. Once again, the lowest results are obtained with the full AQUAINT-2 corpus corresponding to $\mathcal{I} = 825,148$.

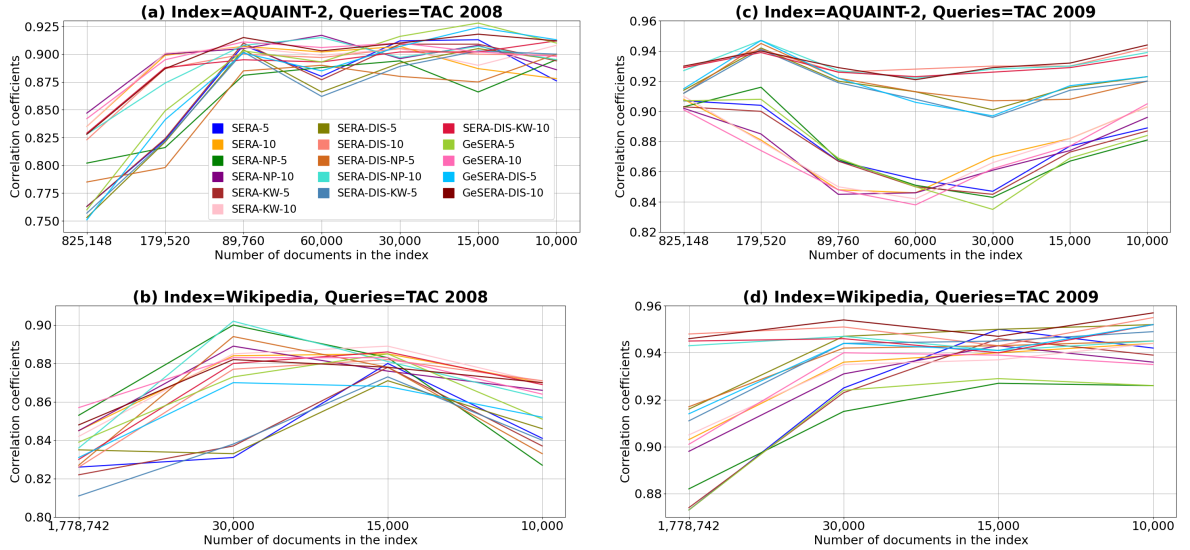



Figure 2: Pearson correlation coefficients using TAC2008 and TAC2009 as queries, and AQUAINT-2 and Wikipedia as indexes. Scores are averaged over all human annotators $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$, and $\mathcal{A}_4$. *Best viewed in color.*

## 4.2 Comparison of GeSERA with the SoTA

We use the best index sizes from Subsection 4.1 to report the detailed results with the best variants of each method from Subsection 3.3. More results are reported in the supplementary material.

### 4.2.1 TAC2008 query dataset

Table 1-*up* shows correlations of the best variants of SERA, GeSERA, ROUGE, SummTriver (ST) and FRESA with two manual evaluation approaches: Pyramid and Responsiveness. Note that for SERA and GeSERA, we fix the query dataset TAC2008, while we vary the index between AQUAINT-2 and Wikipedia.

**AQUAINT-2 Index -** Results show that GeSERA-5 achieves the best scores among all variants of SERA and GeSERA for both Pyramid and Responsiveness. SERA-5 is the best variant of SERA for Pyramid with Pearson and Spearman, while SERA-NP-10 provides better results for Pyramid with Kendall, and Responsiveness with all correlation measures. GeSERA-5 surpasses the two best variants of SERA by 0.015 (Pearson), 0.016 (Spearman), and 0.021 (Kendall) points for Pyramid, and by 0.02, 0.027, and 0.031 points for Responsiveness. The gains are higher for Responsiveness, and for Kendall correlations.

**Wikipedia Index -** Results show that GeSERA-10 is the best variant of GeSERA for Pyramid with all correlation metrics used. GeSERA-DIS-10 gets the best scores for Responsiveness with Pearson and Kendall. Alternatively, the best SERA variant is SERA-DIS-NP-10 for Pyramid with all correlation measures, and with Responsiveness for Spearman and Kendall measures. Interestingly, when using queries from TAC2008 with Wikipedia, GeSERA does not surpass SERA neither for Pyramid, not for Responsiveness.

| **TAC2008** | Pyramid | | | Responsiveness | | |
|---|---|---|---|---|---|---|
| | Pearson | Spearman | Kendall | Pearson | Spearman | Kendall |
| ST-JS-$\mathcal{T}_m$ | -0.889 | -0.827 | -0.643 | -0.820 | -0.801 | -0.608 |
| FRESA-1 | -0.487 | -0.638 | -0.537 | -0.385 | -0.498 | -0.371 |
| FRESA-4 | 0.544 | 0.257 | 0.168 | 0.596 | 0.416 | 0.296 |
| ROUGE-2-R | **0.946** | **0.967** | **0.851** | **0.894** | **0.918** | **0.755** |
| ROUGE-3-F | **0.941** | **0.951** | **0.810** | **0.915** | **0.924** | **0.767** |
| | AQUAINT-2 index ($\mathcal{I} = 15,000$) | | | | | |
| SERA-5 | 0.913 | 0.908 | 0.732 | 0.845 | 0.821 | 0.624 |
| SERA-NP-10 | 0.908 | 0.905 | 0.739 | 0.849 | 0.827 | 0.632 |
| GeSERA-5 | **0.928** | **0.924** | **0.760** | **0.869** | **0.854** | **0.663** |
| | Wikipedia index ($\mathcal{I} = 30,000$) | | | | | |
| SERA-NP-5 | 0.900 | 0.898 | 0.733 | 0.839 | 0.819 | 0.616 |
| SERA-DIS-NP-10 | 0.902 | 0.917 | 0.754 | 0.826 | 0.820 | 0.626 |
| GeSERA-10 | 0.883 | 0.903 | 0.727 | 0.808 | 0.805 | 0.598 |
| GeSERA-DIS-10 | 0.882 | 0.899 | 0.722 | 0.810 | 0.800 | 0.601 |

| **TAC2009** | Pyramid | | | Responsiveness | | |
|---|---|---|---|---|---|---|
| | Pearson | Spearman | Kendall | Pearson | Spearman | Kendall |
| ST-JS-$\mathcal{T}_m$ | -0.526 | -0.755 | -0.623 | -0.650 | -0.744 | -0.587 |
| FRESA-1 | -0.610 | -0.650 | -0.491 | -0.594 | -0.565 | -0.410 |
| FRESA-2 | -0.630 | 0.046 | -0.026 | -0.385 | -0.074 | -0.063 |
| ROUGE-1-F | **0.951** | **0.915** | **0.788** | **0.835** | 0.793 | 0.622 |
| ROUGE-3-F | 0.842 | **0.964** | **0.841** | 0.622 | **0.852** | **0.675** |
| ROUGE-3-R | 0.848 | **0.964** | **0.845** | 0.627 | **0.845** | **0.673** |
| | AQUAINT-2 index ($\mathcal{I} = 179,520$) | | | | | |
| SERA-NP-5 | 0.916 | 0.831 | 0.670 | 0.816 | 0.692 | 0.530 |
| SERA-NP-10 | 0.885 | 0.828 | 0.662 | 0.806 | 0.702 | 0.529 |
| SERA-DIS-10 | 0.941 | 0.836 | 0.671 | 0.806 | 0.673 | 0.521 |
| SERA-DIS-NP-10 | 0.947 | 0.825 | 0.665 | 0.806 | 0.683 | 0.518 |
| GeSERA-5 | 0.908 | 0.835 | 0.678 | 0.834 | 0.697 | 0.530 |
| GeSERA-DIS-5 | 0.947 | 0.836 | 0.688 | 0.831 | 0.684 | 0.525 |
| | Wikipedia index ($\mathcal{I} = 10,000$) | | | | | |
| SERA-10 | 0.944 | 0.892 | 0.741 | **0.845** | **0.784** | **0.607** |
| SERA-KW-10 | 0.944 | 0.894 | 0.738 | 0.839 | 0.771 | 0.588 |
| SERA-DIS-10 | **0.955** | 0.896 | 0.751 | 0.791 | 0.781 | 0.603 |
| SERA-DIS-KW-10 | 0.952 | **0.899** | 0.753 | 0.785 | 0.782 | **0.607** |
| GeSERA-10 | 0.935 | 0.870 | 0.710 | **0.839** | 0.737 | 0.571 |
| GeSERA-DIS-5 | 0.952 | 0.867 | 0.717 | 0.819 | 0.768 | 0.592 |
| GeSERA-DIS-10 | **0.957** | 0.882 | 0.710 | 0.800 | 0.748 | 0.577 |

Table 1: Correlations on TAC2008 and TAC2009 datasets, in terms of Pearson, Spearman and Kendall, of automatic evaluation methods with Pyramid and Responsiveness. The best first (red), second (blue) and third (black) scores of each column are in bold

While ROUGE-2-R and ROUGE-3-F provide the best results for all correlation measures on TAC2008, GeSERA and SERA largely surpass the scores of SummTriver and FRESA with both Pyramid and Responsiveness. In the case of GeSERA-5, it achieves higher correlations than ST-JS-$\mathcal{T}_m$, the best variant of SummTriver, by 0.039, 0.097, and 0.117 for Pyramid, and 0.049, 0.053, and 0.055 for Responsiveness. Finally, FRESA baseline achieves the lowest correlation scores in all configurations. The performance of SummTriver and FRESA is not surprising insofar as they do not rely on any human reference.

#### 4.2.2 TAC2009 query dataset

Table 1-*bottom* shows correlation coefficients of SERA, GeSERA, ROUGE, SummTriver and FRESA with two manual evaluation approaches: Pyramid and Responsiveness. Once again, we fix the query dataset TAC2009, while we vary the index between AQUAINT-2 and Wikipedia.

**AQUAINT-2 Index -** ROUGE provides the highest scores against SERA and GeSERA when we index documents from AQUAINT-2. Importantly, GeSERA-DIS-5 and GeSERA-5 achieve higher correlations than SERA with Pyramid and Responsiveness, respectively. Note that the scores of SERA vary more between its variants, while results of GeSERA are more stable and the best ones are obtained with only two of its variants. This finding highlights the robustness of our approach against variations of configurations.

**Wikipedia Index -** SERA and GeSERA surpass ROUGE against the Pyramid manual method in terms of Pearson correlation when indexing documents from Wikipedia. The best correlations are obtained by GeSERA-DIS-10 and SERA-10 against Pyramid and Responsiveness, respectively.

Interestingly, for TAC2009, GeSERA-DIS-10 achieves better Pearson correlation than ROUGE with Pyramid, and GeSERA-10 with Responsiveness. This finding proves the effectiveness of GeSERA to evaluate summaries from the general domain. Equally, GeSERA reduces the gap between SERA and ROUGE in most of other cases.

SummTriver achieves reasonably good results in Table 1 even without the use of any human reference. This baseline is useful when human summaries are costly or hard to find. However, when such references are available, SummTriver does not take advantage of them, leading its correlation to be low compared to human-based evaluation approaches such as ROUGE and SERA.

FRESA shows the lowest scores among evaluation approaches tested here. It drops approximately from 0.1 to 0.3 point compared to the lowest results obtained by SERA. This is mainly because FRESA is based only on the divergence between the evaluated summary and its source documents, without including any comparison with summaries generated by other participants, as Summtriver does. Thus, FRESA is barely correlated with manual evaluation in many cases where the correlation gets so close to zero (for instance, FRESA-2 with TAC2009 using Kendall correlation). Note that SummTriver and FRESA have mostly negative correlations because they are based on a divergence measure which increases when the summary's quality is low and decreases when its quality is high.

#### 4.2.3 CNNDM query dataset

Based on results obtained in Subsection 4.1 regarding the effectiveness of SERA and GeSERA with small index sizes, we decided to index $\mathcal{I} = 10,000$ documents from Wikipedia to run the two methods on CNNDM. Results are in Table 2.

| **CNNDM** | Pearson | Spearman | Kendall |
|---|---|---|---|
| ROUGE-1-R | **0.914** | **0.922** | **0.773** |
| ROUGE-2-R | **0.962** | **0.958** | **0.860** |
| ROUGE-L-F | 0.526 | 0.368 | 0.278 |
| BERTScore-1-P | -0.021 | 0.093 | 0.064 |
| BERTScore-1-R | 0.768 | 0.738 | 0.552 |
| MoverScore | 0.443 | 0.367 | 0.284 |
| JS-2 | 0.780 | 0.665 | 0.512 |
| SERA-10 | 0.858 | 0.789 | 0.616 |
| SERA-KW-10 | 0.864 | 0.782 | 0.621 |
| SERA-DIS-10 | 0.827 | 0.781 | 0.605 |
| SERA-DIS-NP-5 | 0.599 | 0.554 | 0.391 |
| GeSERA-5 | 0.623 | 0.527 | 0.387 |
| GeSERA-10 | **0.880** | **0.872** | **0.719** |
| GeSERA-DIS-10 | 0.817 | 0.788 | 0.605 |

Table 2: Correlation coefficients on CNNDM, in terms of Pearson, Spearman and Kendall, of multiple automatic evaluation methods with LitePyramid.

Table 2 shows that the highest correlations of ROUGE are obtained with ROUGE-2-R, followed by ROUGE-1-R. Globally, the highest correlations in ROUGE are obtained with the recall metric (ROUGE-R), followed by ROUGE-F, and finally by ROUGE-P. The following highest correlations are obtained with GeSERA-10. Once again, GeSERA surpasses all the SERA variants, and the state-of-the-art methods presented in this table.

Although SERA-KW-10 has the best score in terms of Pearson and Kendall, all SERA variants present very similar scores. Behind the SERA method, BERTScore and JS-2 measures present very similar scores. Meanwhile, MoverScore shows the lowest correlations. Results show the effectiveness of GeSERA to evaluate extractive and abstractive summaries since CNNDM contains both approaches.

### 4.3 Impact of human annotators on the performance of SERA and GeSERA

For the sake of comparability with state-of-the-art approaches, we presented in the previous section correlations computed with the four manual annotators. According to Lin and Hovy (2002), human evaluation is subjective. We confirm experimentally this finding and highlight that human annotators affect the performance of automatic evaluation approaches. To know how much each annotator can affect the correlation against human evaluations, and which annotator gets the lowest and highest correlations, we compute scores using different combination of human annotators. We compute the correlation of each human annotator individually, using three human annotators: ($\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$), ($\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_4$), ($\mathcal{A}_2$, $\mathcal{A}_3$, $\mathcal{A}_4$), and finally using the four human annotators ($\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$, $\mathcal{A}_4$).

Note that we only report here the results on TAC2009 dataset insofar as the results on TAC2008 are not conclusive, where the best score vary considerably from one annotator to another depending on the configuration. We present results on TAC2008 in the supplementary material.

Figure 3 provides SERA and GeSERA correlations with Pyramid using TAC2009 as a query dataset and AQUAINT-2 and Wikipedia as indexes. Results show that the best human annotator is always $\mathcal{A}_1$ as he provides summaries with the best correlations of SERA and GeSERA in terms of Pearson, Spearman, and Kendall. Inversely, the worst human annotator is always $\mathcal{A}_3$ for SERA as he achieves the worst scores in terms of all correlation metrics used here. For GeSERA, the worst human annotator is $\mathcal{A}_3$ for Wikipedia in terms of all correlation metrics, while it is $\mathcal{A}_4$ for AQUAINT-2 in terms of Spearman and Kendall and $\mathcal{A}_3$ in terms of Pearson.

In Table 3, we compare results obtained with the four manual annotators versus those obtained with the best three annotators ($\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_4$) for TAC2009. Results show that there is a clear gain when discarding the most unreliable annotator. We conclude that human annotators partially participate in the quality of automatic summary evaluation. This bias is caused by the quality of their manually written summaries.

| | AQUAINT-2 | | | | Wikipedia | | | |
|---|---|---|---|---|---|---|---|---|
| | SERA-DIS-NP-10 | | GeSERA-DIS-5 | | SERA-DIS-10 | | GeSERA-DIS-10 | |
| Annotators | 4 | 3 | 4 | 3 | 4 | 3 | 4 | 3 |
| Pearson | 0.947 | **0.949** | 0.947 | **0.951** | 0.955 | **0.959** | 0.957 | **0.959** |
| Spearman | 0.825 | **0.835** | 0.836 | **0.841** | 0.896 | **0.909** | **0.882** | **0.882** |
| Kendall | 0.665 | **0.681** | 0.668 | **0.691** | 0.751 | **0.776** | 0.710 | **0.743** |

Table 3: Impact of human annotators on the evaluation with SERA and GeSERA using TAC2009.

## 5 Related work

Evaluation methods are fundamental techniques to assess if summaries generated by an automatic system capture the original document's idea. Different evaluation methods have been developed in the last decade for the evaluation of automatically-generated summaries. There exist two types of evaluation methods: (1) manual evaluation methods, and (2) automatic evaluation methods. The first group of methods requires human intervention as ground-truth references. Pyramid (Nenkova and Passonneau, 2004) and Responsiveness are the most popular such methods. The second group of methods is divided itself into two subsets: (1) methods that need human intervention like ROUGE (Lin, 2004a) and SERA (Cohan and Goharian, 2016), and (2) methods that do not need any human reference like SummTriver (Cabrera-Diego and Torres-Moreno, 2018) and FRESA (Torres-Moreno et al., 2010).

The most popular automatic metric used by the community is ROUGE (Lin, 2004a). It needs reference summaries, and is based on their lexical overlaps with candidate summaries. That is why it is more useful to evaluate extractive summaries where chunks of the text are copied and pasted to form the summary. However, in the case of abstractive summaries where the ATS paraphrases the text with possibly new vocabulary, the ROUGE metric becomes unfair. To overcome this issue, researchers have been proposing in the last few years other automatic metrics to fairly evaluate both extractive and abstractive summaries.

The first type of automatic evaluation methods relies partially on human judgment as ROUGE does. The simplest method is based on a *Jensen-Shannon* (JS-2) (Lin et al., 2006) divergence between bi-gram's distribution of candidate and ref-

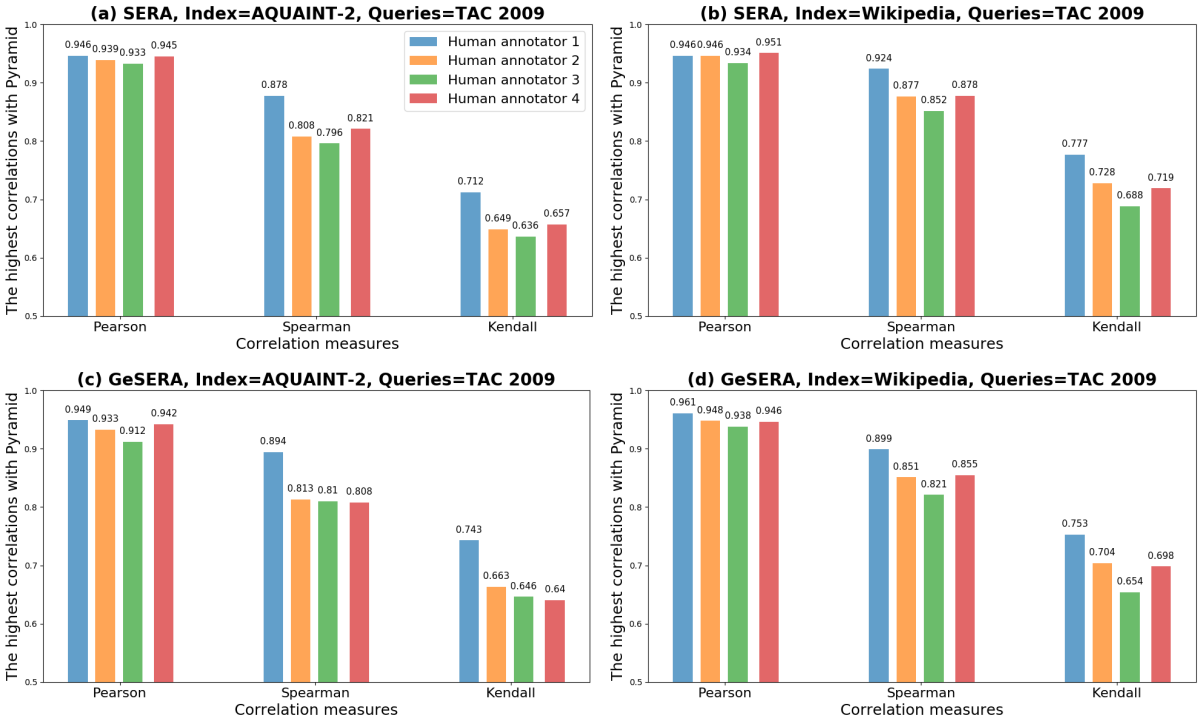



Figure 3: Correlation coefficients in terms of Pearson, Spearman and Kendall obtained by each annotator $\mathcal{A}_i$ using TAC2009 for queries, and AQUAINT-2 and Wikipedia as indexes. *Best viewed in color.*

erence summaries. More sophisticated systems include MoverScore (Zhao et al., 2019) that is based on fine-tuning the BERT model and combining contextualized representations with Earth Mover Distance (EMD) from Rubner et al. (2000). BERTScore (Zhang et al., 2020) also is based on BERT model. Unlike ROUGE (Lin, 2004b), BERTScore makes use of contextual embeddings that are effective for paraphrase detection. Similarly to BERTScore, *Semantic Similarity for Abstractive Summarization* (SSAS) (Vadapalli et al., 2017) is based on semantic matching between candidate and reference summaries. The second type of automatic evaluation methods does not need any human intervention. For instance, *FRamework for Evaluating Summaries Automatically* (FRESA) (Torres-Moreno et al., 2010) is based on divergences among probability distributions between the summary to evaluate and its source document. Another well-known metric is SummTriver (Cabrera-Diego and Torres-Moreno, 2018). It is based on Trivergences between the summary to evaluate, its source document(s), and a set of summaries related to the same source document(s) but generated with other ATS systems.

## 6 Conclusion and perspectives

We introduced GeSERA, an open-source system for general-domain summary evaluation. We redefine query reformulation of SERA based on POS Tags analysis of datasets from different domains, and replace the biomedical index with documents from AQUAINT-2 and Wikipedia. GeSERA achieves competitive results compared to state-of-the-art approaches. Overall, GeSERA surpasses SERA and reduces its gap with ROUGE, and in two cases, it even surpasses ROUGE, the lexical-based method. Unsurprisingly, the comparison with evaluation methods that do not rely on human references reveals a large gap in favor of GeSERA since it relies on human references while the others do not. Extensive experiments show that the index size has a considerable effect on the performance of SERA and GeSERA that tend to perform better with small-size indexes. Finally, the study of human annotators shows their impact on the performance of automatic evaluation methods that rely on human intervention. Our code is publicly available to facilitate reproducibility. We will: (1) explore other variants of the search engine to know its impact on GeSERA, (2) propose a new version of GeSERA that does not rely on human intervention by exploiting information from the source text, (3) apply prepossessing on the index and search for other solutions to improve query reformulation, and (4) explore larger query datasets such as Multi-News (Fabbri et al., 2019).

## References


Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Luis Adrián Cabrera-Diego and Juan-Manuel Torres-Moreno. 2018. Summtriver: A new trivergent model to evaluate summaries automatically without human references. *Data Knowl. Eng.*, 113:184–197.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. *CoRR*, abs/1604.00400.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.

David Graff. 2002. *The AQUAINT corpus of English news text:[content copyright] Portions© 1998-2000 New York Times, Inc.,© 1998-2000 Associated Press, Inc.,© 1996-2000 Xinhua News Service*. Linguistic Data Consortium.

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, pages 239–251.

Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. 2020. Automatic text summarization of COVID-19 medical research articles using BERT and GPT-2. *CoRR*, abs/2006.01997.

Stephen Kokoska and Daniel Zwillinger. 2000. *CRC standard probability and statistics tables and formulae*. Crc Press.

Chin-Yew Lin. 2004a. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.

Chin-Yew Lin. 2004b. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152.

José R Pérez-Agüera, Javier Arroyo, Jane Greenberg, Joaquin Perez Iglesias, and Victor Fresno. 2010. Using bm25f for semantic search. In *Proceedings of the 3rd international semantic search workshop*, pages 1–8.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40(2):99–121.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Juan-Manuel Torres-Moreno, Horacio Saggion, Iria da Cunha, Eric Sanjuan, and Patricia Velázquez-Morales. 2010. Summary evaluation with and without references. *Polibits*, 42:13–20.

Raghuram Vadapalli, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2017. SSAS: semantic similarity for abstractive summarization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 198–203. Asian Federation of Natural Language Processing.

Hugo Zaragoza, Nick Craswell, Michael J. Taylor, Suchi Saria, and Stephen E. Robertson. 2004. Microsoft cambridge at TREC 13: Web and hard tracks. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 563–578. Association for Computational Linguistics.

## Supplementary material

In this supplementary material, we provide: (1) more details about evaluation datasets (Section A), (2) implementation details (Section B), (3) detailed results of all tested approaches (Section C), and (4) the impact of human annotators on TAC2008 dataset (Section D)

## A Datasets details

- **AQUAINT-2** is a news corpus built from New York Times, Associated Press, and Xinhua News Agency. Indexes built from this corpus are balanced, except for the largest one ($\mathcal{I} = 825,148$), that contains all documents. Note that AQUAINT-2 is not open-source, and we cannot distribute it. However, obtained results can be helpful in academic research.
- **Wikipedia** is a free online encyclopedia from the general domain. The largest index ($\mathcal{I} = 1,778,742$) contains all available documents.
- **TAC2008 (/TAC2009)** contains two sets. Each set contains 48 (/44) topics. Each topic includes 10 documents and 4 reference summaries. Candidate summaries are proposed by 58 (/55) participants, where each one provides a candidate summary per topic. In total, there are 960 (/880) documents, 5568 (/4840) candidate summaries, and 384 (/352) reference summaries.

## B Implementation details

SERA and GeSERA were implemented in Python. For information retrieval, we used the Okapi BM25F ranking function from Whoosh, a flexible and pure python search engine framework.

We used the authors' public implementations to run ROUGE, SummTriver, and FRESA. The latter was basically designed for mono-document evaluation. Thus, we concatenated all the articles of the same topic to be able to run it on TAC.

To compute the correlations with LitePyramid of ROUGE, BERTScore, MoverScore, and JS-2 on the CNN Daily Mail dataset, we used the scores provided by Bhandari et al., 2020 in their GitHub repository[5]. Based on experiments on the TAC datasets in Subsection 4.3, we use an index size of $\mathcal{I} = 10,000$ in SERA and GeSERA.

For the sake of comparability, scores are averaged for each participant before computing the correlations with manual methods.

[5] https://github.com/neulab/REALSumm/

## C More results

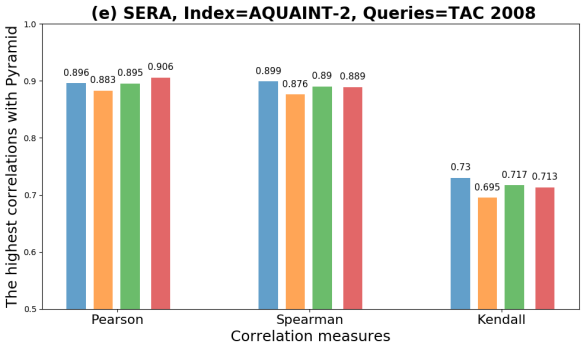
Table 4 and Table 5 provide more results on CNNDM, TAC2008 and TAC2009 datasets. Both tables present variants of the evaluation metrics that we did not report in the main paper.

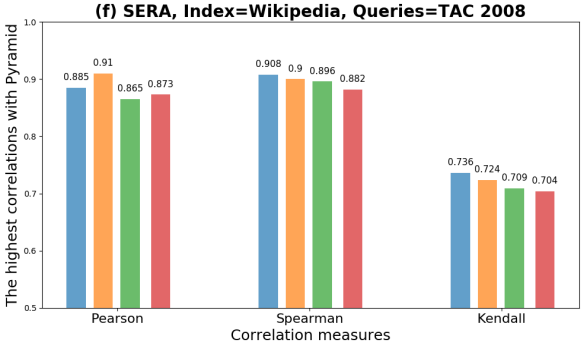
| | Pearson | Spearman | Kendall |
|---|---|---|---|
| ROUGE-1-F | 0.600 | 0.468 | 0.358 |
| ROUGE-1-P | -0.175 | -0.212 | -0.117 |
| ROUGE-2-F | 0.648 | 0.452 | 0.311 |
| ROUGE-2-P | 0.099 | 0.050 | 0.023 |
| ROUGE-L-P | -0.045 | -0.148 | -0.070 |
| ROUGE-L-R | 0.871 | -0.914 | 0.759 |
| BERTScore-1-F | 0.385 | 0.374 | 0.258 |
| MoverScore | 0.443 | 0.367 | 0.284 |
| JS-2 | 0.780 | 0.665 | 0.512 |
| SERA-5 | 0.773 | 0.710 | 0.508 |
| SERA-NP-5 | 0.690 | 0.639 | 0.452 |
| SERA-NP-10 | 0.743 | 0.705 | 0.502 |
| SERA-KW-5 | 0.784 | 0.711 | 0.508 |
| SERA-DIS-5 | 0.748 | 0.668 | 0.472 |
| SERA-DIS-NP-10 | 0.671 | 0.568 | 0.393 |
| SERA-DIS-KW-5 | 0.758 | 0.657 | 0.465 |
| SERA-DIS-KW-10 | 0.828 | 0.752 | 0.565 |
| GeSERA-DIS-5 | 0.566 | 0.469 | 0.315 |

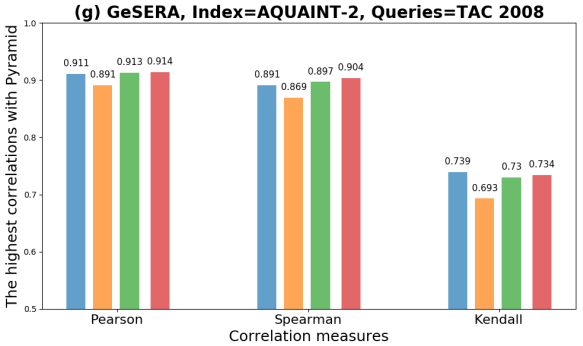
Table 4: Correlations of CNNDM dataset, in terms of Pearson, Spearman and Kendall, of multiple automatic evaluation methods with LitePyramid.

## D Impact of human annotators

Figure 4 provides SERA and GeSERA correlations with Pyramid using TAC2008 as a query dataset and AQUAINT-2 and Wikipedia as indexes. Contrarily to TAC2009, it is hard to define for TAC2008 the impact of human annotators on the evaluation with SERA and GeSERA, as the best scores change from one case to another. For instance, $\mathcal{A}_1$ is the best human annotator for SERA with both AQUAINT-2 and Wikipedia corpora in terms of Spearman and Kendall. However, in terms of Pearson correlation, the best annotator is $\mathcal{A}_4$ for AQUAINT-2 and $\mathcal{A}_2$ for Wikipedia. Alternatively, the best annotator for GeSERA is always $\mathcal{A}_2$ for Wikipedia while the same annotator provides the worst results with AQUAINT-



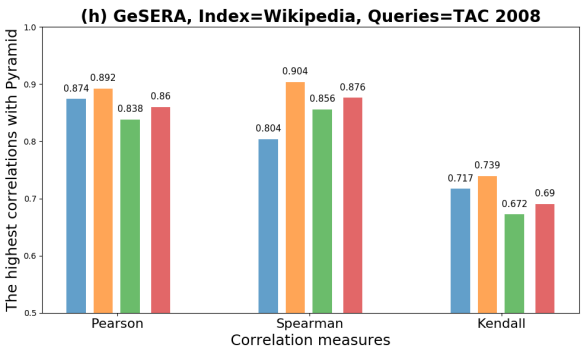




Figure 4: Correlations coefficients obtained by each annotator $\mathcal{A}_i$ using TAC2008 dataset for queries, and AQUAINT-2 and Wikipedia as indexes.

| | TAC2008 | | | | | |
|---|---|---|---|---|---|---|
| | Pyramid | | | Responsiveness | | |
| | Pearson | Spearman | Kendall | Pearson | Spearman | Kendall |
| ST-sJS-$\mathcal{T}_m$ | -0.885 | -0.822 | -0.637 | -0.822 | -0.797 | -0.605 |
| ST-KL-$\mathcal{T}_m$ | -0.694 | -0.700 | -0.510 | -0.706 | -0.695 | -0.504 |
| ST-JS-$\mathcal{T}_c$ | -0.858 | -0.805 | -0.613 | -0.771 | -0.777 | -0.578 |
| ST-sJS-$\mathcal{T}_c$ | -0.857 | -0.805 | -0.612 | -0.771 | -0.777 | -0.577 |
| ST-KL-$\mathcal{T}_c$ | -0.216 | -0.168 | -0.123 | 0.025 | 0.134 | 0.091 |
| FRESA-2 | 0.474 | -0.062 | -0.064 | 0.523 | 0.076 | 0.034 |
| FRESA-3 | 0.539 | 0.241 | 0.162 | 0.593 | 0.362 | 0.250 |
| ROUGE-1-F | 0.908 | 0.941 | 0.787 | 0.853 | 0.883 | 0.702 |
| ROUGE-1-P | 0.730 | 0.841 | 0.643 | 0.698 | 0.803 | 0.626 |
| ROUGE-1-R | 0.911 | 0.935 | 0.774 | 0.851 | 0.858 | 0.665 |
| ROUGE-2-F | 0.940 | 0.965 | 0.843 | 0.892 | 0.915 | 0.746 |
| ROUGE-2-P | 0.911 | 0.942 | 0.788 | 0.873 | 0.901 | 0.730 |
| ROUGE-3-P | 0.926 | 0.934 | 0.783 | 0.909 | 0.918 | 0.766 |
| ROUGE-3-R | 0.945 | 0.951 | 0.811 | 0.914 | 0.922 | 0.763 |
| ROUGE-L-F | 0.878 | 0.925 | 0.756 | 0.823 | 0.868 | 0.689 |
| ROUGE-L-P | 0.711 | 0.823 | 0.632 | 0.679 | 0.794 | 0.611 |
| ROUGE-L-R | 0.882 | 0.927 | 0.762 | 0.823 | 0.856 | 0.661 |
| ROUGE-W-1.2-F | 0.901 | 0.940 | 0.782 | 0.848 | 0.878 | 0.701 |
| ROUGE-W-1.2-P | 0.712 | 0.822 | 0.631 | 0.688 | 0.794 | 0.620 |
| ROUGE-W-1.2-R | 0.897 | 0.940 | 0.785 | 0.841 | 0.871 | 0.684 |
| ROUGE-SU4-F | 0.917 | 0.949 | 0.805 | 0.870 | 0.904 | 0.728 |
| ROUGE-SU4-P | 0.839 | 0.910 | 0.728 | 0.805 | 0.869 | 0.689 |
| ROUGE-SU4-R | 0.927 | 0.950 | 0.800 | 0.874 | 0.908 | 0.736 |
| | AQUAINT-2 index ($\mathcal{I} = 15,000$) | | | | | |
| SERA-10 | 0.887 | 0.871 | 0.693 | 0.817 | 0.766 | 0.572 |
| SERA-NP-5 | 0.866 | 0.863 | 0.681 | 0.792 | 0.770 | 0.569 |
| SERA-KW-5 | 0.909 | 0.901 | 0.721 | 0.841 | 0.809 | 0.611 |
| SERA-KW-10 | 0.890 | 0.880 | 0.705 | 0.823 | 0.779 | 0.579 |
| SERA-DIS-5 | 0.905 | 0.885 | 0.713 | 0.840 | 0.800 | 0.593 |
| SERA-DIS-10 | 0.900 | 0.888 | 0.711 | 0.829 | 0.797 | 0.592 |
| SERA-DIS-NP-5 | 0.875 | 0.864 | 0.679 | 0.805 | 0.774 | 0.573 |
| SERA-DIS-NP-10 | 0.907 | 0.905 | 0.735 | 0.843 | 0.819 | 0.616 |
| SERA-DIS-KW-5 | 0.903 | 0.885 | 0.712 | 0.837 | 0.801 | 0.597 |
| SERA-DIS-KW-10 | 0.902 | 0.888 | 0.709 | 0.832 | 0.804 | 0.601 |
| GeSERA-10 | 0.902 | 0.890 | 0.708 | 0.843 | 0.800 | 0.604 |
| GeSERA-DIS-5 | 0.924 | 0.910 | 0.746 | 0.861 | 0.836 | 0.641 |
| GeSERA-DIS-10 | 0.918 | 0.896 | 0.724 | 0.852 | 0.806 | 0.610 |
| | Wikipedia index ($\mathcal{I} = 30,000$) | | | | | |
| SERA-5 | 0.831 | 0.839 | 0.673 | 0.763 | 0.751 | 0.560 |
| SERA-10 | 0.884 | 0.900 | 0.724 | 0.812 | 0.798 | 0.594 |
| SERA-NP-10 | 0.890 | 0.912 | 0.738 | 0.806 | 0.812 | 0.618 |
| SERA-KW-5 | 0.837 | 0.838 | 0.667 | 0.767 | 0.749 | 0.552 |
| SERA-KW-10 | 0.885 | 0.906 | 0.727 | 0.812 | 0.806 | 0.603 |
| SERA-DIS-5 | 0.833 | 0.825 | 0.655 | 0.781 | 0.757 | 0.568 |
| SERA-DIS-10 | 0.877 | 0.887 | 0.707 | 0.815 | 0.790 | 0.588 |
| SERA-DIS-NP-5 | 0.894 | 0.884 | 0.718 | 0.838 | 0.809 | 0.604 |
| SERA-DIS-KW-5 | 0.838 | 0.837 | 0.667 | 0.783 | 0.761 | 0.567 |
| SERA-DIS-KW-10 | 0.881 | 0.894 | 0.719 | 0.817 | 0.797 | 0.598 |
| GeSERA-5 | 0.873 | 0.865 | 0.698 | 0.803 | 0.774 | 0.581 |
| GeSERA-DIS-5 | 0.870 | 0.865 | 0.701 | 0.802 | 0.773 | 0.580 |
| | TAC2009 | | | | | |
| ST-sJS-$\mathcal{T}_m$ | -0.511 | -0.751 | -0.620 | -0.636 | -0.739 | -0.585 |
| ST-KL-$\mathcal{T}_m$ | -0.371 | -0.681 | -0.558 | -0.518 | -0.683 | -0.550 |
| ST-JS-$\mathcal{T}_c$ | -0.477 | -0.718 | -0.582 | -0.619 | -0.710 | -0.563 |
| ST-sJS-$\mathcal{T}_c$ | -0.475 | -0.717 | -0.581 | -0.618 | -0.709 | -0.562 |
| ST-KL-$\mathcal{T}_c$ | -0.138 | -0.062 | -0.040 | -0.014 | -0.007 | -0.005 |
| FRESA-3 | -0.556 | 0.055 | 0.056 | -0.298 | 0.180 | -0.147 |
| FRESA-4 | -0.516 | 0.189 | 0.142 | -0.217 | 0.363 | -0.278 |
| ROUGE-1-P | 0.923 | 0.845 | 0.678 | 0.791 | 0.789 | 0.630 |
| ROUGE-1-R | 0.926 | 0.892 | 0.748 | 0.814 | 0.764 | 0.591 |
| ROUGE-2-F | 0.930 | 0.955 | 0.839 | 0.740 | 0.831 | 0.664 |
| ROUGE-2-P | 0.906 | 0.937 | 0.796 | 0.716 | 0.829 | 0.658 |
| ROUGE-2-R | 0.937 | 0.952 | 0.841 | 0.746 | 0.820 | 0.654 |
| ROUGE-3-P | 0.828 | 0.940 | 0.800 | 0.610 | 0.839 | 0.656 |
| ROUGE-L-F | 0.865 | 0.604 | 0.461 | 0.649 | 0.414 | 0.294 |
| ROUGE-L-P | 0.801 | 0.546 | 0.406 | 0.573 | 0.360 | 0.255 |
| ROUGE-L-R | 0.875 | 0.622 | 0.474 | 0.663 | 0.414 | 0.298 |
| ROUGE-W-1.2-F | 0.882 | 0.654 | 0.512 | 0.651 | 0.462 | 0.341 |
| ROUGE-W-1.2-P | 0.798 | 0.514 | 0.393 | 0.558 | 0.337 | 0.237 |
| ROUGE-W-1.2-R | 0.889 | 0.671 | 0.529 | 0.659 | 0.469 | 0.340 |
| ROUGE-SU4-F | 0.934 | 0.940 | 0.818 | 0.747 | 0.808 | 0.639 |
| ROUGE-SU4-P | 0.893 | 0.910 | 0.761 | 0.702 | 0.804 | 0.638 |
| ROUGE-SU4-R | 0.942 | 0.924 | 0.787 | 0.756 | 0.789 | 0.619 |
| | AQUAINT-2 index ($\mathcal{I} = 179,520$) | | | | | |
| SERA-5 | 0.904 | 0.818 | 0.656 | 0.814 | 0.664 | 0.502 |
| SERA-10 | 0.881 | 0.817 | 0.651 | 0.813 | 0.675 | 0.513 |
| SERA-KW-5 | 0.900 | 0.816 | 0.654 | 0.807 | 0.665 | 0.503 |
| SERA-KW-10 | 0.880 | 0.810 | 0.646 | 0.807 | 0.670 | 0.511 |
| SERA-DIS-5 | 0.942 | 0.829 | 0.666 | 0.811 | 0.660 | 0.501 |
| SERA-DIS-NP-5 | 0.945 | 0.831 | 0.670 | 0.809 | 0.687 | 0.529 |
| SERA-DIS-KW-5 | 0.941 | 0.822 | 0.659 | 0.808 | 0.653 | 0.496 |
| SERA-DIS-KW-10 | 0.939 | 0.826 | 0.658 | 0.802 | 0.669 | 0.515 |
| GeSERA-10 | 0.874 | 0.813 | 0.652 | 0.814 | 0.686 | 0.517 |
| GeSERA-DIS-10 | 0.940 | 0.818 | 0.657 | 0.818 | 0.673 | 0.514 |
| | Wikipedia index ($\mathcal{I} = 10,000$) | | | | | |
| SERA-5 | 0.942 | 0.870 | 0.717 | 0.843 | 0.768 | 0.592 |
| SERA-NP-5 | 0.926 | 0.863 | 0.704 | 0.831 | 0.749 | 0.573 |
| SERA-NP-10 | 0.936 | 0.863 | 0.709 | 0.835 | 0.759 | 0.592 |
| SERA-KW-5 | 0.939 | 0.863 | 0.701 | 0.834 | 0.761 | 0.588 |
| SERA-DIS-5 | 0.952 | 0.877 | 0.729 | 0.809 | 0.778 | 0.602 |
| SERA-DIS-NP-5 | 0.945 | 0.842 | 0.684 | 0.811 | 0.733 | 0.563 |
| SERA-DIS-NP-10 | 0.945 | 0.845 | 0.688 | 0.785 | 0.746 | 0.579 |
| SERA-DIS-KW-5 | 0.949 | 0.868 | 0.713 | 0.801 | 0.773 | 0.596 |
| GeSERA-5 | 0.926 | 0.854 | 0.701 | 0.838 | 0.737 | 0.570 |

Table 5: Correlations on TAC2008 and TAC2009.