A Principled Framework for Evaluating Summarizers: Comparing Models of Summary Quality against Human Judgments

Maxime Peyrard and **Judith Eckle-Kohler**
Research Training Group AIPHES and UKP Lab
Computer Science Department, Technische Universität Darmstadt
www.aiphes.tu-darmstadt.de, www.ukp.tu-darmstadt.de

Abstract

We present a new framework for evaluating extractive summarizers, which is based on a principled representation as optimization problem. We prove that every extractive summarizer can be decomposed into an objective function and an optimization technique. We perform a comparative analysis and evaluation of several objective functions embedded in well-known summarizers regarding their correlation with human judgments. Our comparison of these correlations across two datasets yields surprising insights into the role and performance of objective functions in the different summarizers.

1 Introduction

The task of extractive summarization (ES) can naturally be cast as a discrete optimization problem where the text source is considered as a set of sentences and the summary is created by selecting an optimal subset of the sentences under a length constraint (McDonald, 2007; Lin and Bilmes, 2011).

In this work, we go one step further and mathematically prove that ES is equivalent to the problem of choosing (i) an objective function θ for scoring system summaries, and (ii) an optimizer O. We use (θ, O) to denote the resulting *decomposition* of any extractive summarizer. Our proposed decomposition enables a principled analysis and evaluation of existing summarizers, and addresses a major issue in the current evaluation of ES.

This issue concerns the traditional "intrinsic" evaluation comparing system summaries against human reference summaries. This kind of evaluation is actually an end-to-end evaluation of summarization systems which is performed *after* θ has been optimized by O. This is highly problematic from an evaluation point of view, because first, θ is typically not optimized exactly, and second, there might be side-effects caused by the particular optimization technique O, e.g., a sentence extracted to maximize θ might be suitable because of other properties not included in θ. Moreover, the commonly used evaluation metric ROUGE yields a noisy surrogate evaluation (despite its good correlation with human judgments) compared to the much more meaningful evaluation based on human judgments. As a result, the current end-to-end evaluation does not provide any insights into the *task of automatic summarization*.

The (θ, O) decomposition we propose addresses this issue: it enables a well-defined and principled evaluation of extractive summarizers on the level of their components θ and O. In this work, we focus on the analysis and evaluation of θ, because θ is a model of the quality indicators of a summary, and thus crucial in order to understand the properties of "good" summaries. Specifically, we compare θ functions of different summarizers by measuring the correlation of their θ functions with human judgments.

Our goal is to provide an evaluation framework which the research community could build upon in future research to identify the best possible θ and use it in optimization-based systems. We believe that the identification of such a θ is the central question of summarization, because this optimal θ would represent an optimal definition of summary quality both from an algorithmic point of view and from the human perspective.

In summary, our contribution is twofold: (i) We present a novel and principled evaluation framework for ES which allows evaluating the objective function and the optimization technique separately and independently. (ii) We compare well-known summarization systems regarding their implicit choices of θ by measuring the correlation

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 26–31
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2005

of their θ functions with human judgments on two datasets from the Text Analysis Conference (TAC). Our comparative evaluation yields surprising results and shows that extractive summarization is not solved yet.

The code used in our experiments, including a general evaluation tool is available at `github.com/UKPLab/acl2017-theta_evaluation_summarization`.

2 Evaluation Framework

2.1 (θ, O) decomposition

Let $D = \{s_i\}$ be a document collection considered as a set of sentences. A summary S is then a subset of D, or we can say that S is an element of $\mathcal{P}(D)$, the power set of D.

Objective function We define an objective function to be a function that takes a summary of the document collection D and outputs a score:

$$\begin{array}{rccl} \theta & : & \mathcal{P}(D) & \rightarrow & \mathbb{R} \\ & & S & \mapsto & \theta_D(S) \end{array} \tag{1}$$

Optimizer Then the task of ES is to select the set of sentences S^* with maximal $\theta(S^*)$ under a length constraint:

$$\begin{aligned} S^* &= \operatorname*{argmax}_S \theta(S) \\ len(S) &= \sum_{s \in S} len(s) \leq c \end{aligned} \tag{2}$$

We use O to denote the technique which solves this optimization problem. O is an operator which takes an objective function θ from the set of all objective functions Θ and a document collection D from the set of all document collections $\mathcal{D}$, and outputs a summary S^*:

$$\begin{array}{rccl} O & : & \Theta \times \mathcal{D} & \rightarrow & \mathcal{S} \\ & & (\theta, D) & \mapsto & S^* \end{array} \tag{3}$$

Decomposition Theorem Now we show that the problem of ES is equivalent to the problem of choosing a decomposition (θ, O).

We formalize an extractive summarizer σ as a set function which takes a document collection $D \in \mathcal{D}$ and outputs a summary $S_{D,\sigma} \in \mathcal{P}(D)$. With this formalism, it is clear that every (θ, O) tuple forms a summarizer because $O(\theta, \cdot)$ produces a summary from a document collection.

But the other direction is also true: for every extractive summarizer there exists at least one tuple (θ, O) which perfectly describes the summarizer:

Theorem 1 $\forall \sigma, \exists (\theta, O)$ *such that:*
$\forall D \in \mathcal{D}, \sigma(D) = O(\theta, D)$

This theorem is quite intuitive, especially since it is common to use a similar decomposition in optimization-based summarization systems. In the next section we illustrate the theorem by way of several examples, and provide a rigorous proof of the existence in the supplemental material.

2.2 Examples of θ

We analyze a range of different summarizers regarding their (mostly implicit) θ.

ICSI (Gillick and Favre, 2009) is a global linear optimization that extracts a summary by solving a maximum coverage problem considering the most frequent bigrams in the source documents. ICSI has been among the best systems in a classical ROUGE evaluation (Hong et al., 2014). For ICSI, the identification of θ is trivial because it was formulated as an optimization task. If c_i is the i-th bigram selected in the summary and w_i its weight computed from D, then:

$$\theta_{ICSI}(S) = \sum_{c_i \in S} c_i * w_i \tag{4}$$

LexRank (Erkan and Radev, 2004) is a well-known graph-based approach. A similarity graph $G(V, E)$ is constructed where V is the set of sentences and an edge e_{ij} is drawn between sentences v_i and v_j if and only if the cosine similarity between them is above a given threshold. Sentences are scored according to their PageRank score in G. We observe that $\theta_{LexRank}$ is given by:

$$\theta_{LexRank}(S) = \sum_{s \in S} PR_G(s) \tag{5}$$

where PR is the PageRank score of sentence s.

KL-Greedy (Haghighi and Vanderwende, 2009) minimizes the Kullback Leibler (KL) divergence between the word distributions in the summary and D (i.e $\theta_{KL} = -KL$). Recently, Peyrard and Eckle-Kohler (2016) optimized KL and Jensen Shannon (JS) divergence with a genetic algorithm. In this work, we use KL and JS for both unigram and bigram distributions.

LSA (Steinberger and Jezek, 2004) is an approach involving a dimensionality reduction of the term-document matrix via Singular Value Decomposition (SVD). The sentences extracted should cover the most important latent topics:

$$\theta_{LSA} = \sum_{t \in S} \lambda_t \tag{6}$$

where t is a latent topic identified by SVD on the term-document matrix and λ_t the associated singular value.

Edmundson (Edmundson, 1969) is an older heuristic method which scores sentences according to cue-phrases, overlap with title, term frequency and sentence position. $\theta_{Edmundson}$ is simply a weighted sum of these heuristics.

TF⋆IDF (Luhn, 1958) scores sentences with the TF*IDF of their terms. The best sentences are then greedily extracted. We use both the unigram and bigram versions in our experiments.

3 Experiments

Now we compare the summarizers analyzed above by measuring the correlation of their θ functions with human judgments.

Datasets We use two multi-document summarization datasets from the Text Analysis Conference (TAC) shared task: TAC-2008 and TAC-2009.[1] TAC-2008 and TAC-2009 contain 48 and 44 topics, respectively. Each topic consists of 10 news articles to be summarized in a maximum of 100 words. We use only the so-called initial summaries (A summaries), but not the update part.

For each topic, there are 4 human reference summaries along with a manually created Pyramid set. In both editions, all system summaries and the 4 reference summaries were manually evaluated by NIST assessors for readability, content selection (with Pyramid) and overall responsiveness. At the time of the shared tasks, 57 systems were submitted to TAC-2008 and 55 to TAC-2009. For our experiments, we use the Pyramid and the responsiveness annotations.

System Comparison For each θ, we compute the scores of all system and all manual summaries for any given topic. These scores are compared with the human scores. We include the manual summaries in our computation because this yields a more diverse set of summaries with a wider range of scores. Since an ideal summarizer would create summaries as well as humans, an ideal θ would also be able to correctly score human summaries with high scores.

For comparison, we also report the correlation between pyramid and responsiveness.

Correlations are measured with 3 metrics: Pearson's r, Spearman's ρ and Normalized Discounted Cumulative Gain (Ndcg). Pearson's r is a value correlation metric which depicts linear relationships between the scores produced by θ and the human judgments. Spearman's ρ is a rank correlation metric which compares the ordering of systems induced by θ and the ordering of systems induced by human judgments. Ndcg is a metric that compares ranked lists and puts more emphasis on the top elements by logarithmic decay weighting. Intuitively, it captures how well θ can recognize the best summaries. The optimization scenario benefits from high Ndcg scores because only summaries with high θ scores are extracted.

Previous work on correlation analysis averaged scores over topics for each system and then computed the correlation between averaged scores (Louis and Nenkova, 2013; Nenkova et al., 2007). An alternative and more natural option which we use here is to compute the correlation for each topic and average these correlations over topics (CORRELATION-AVERAGE). Since we want to estimate how well θ functions measure the quality of summaries, we find the summary level averaging more meaningful.

Analysis The results of our correlation analysis are presented in Table 1.

In our (θ, O) formulation, the end-to-end approach maps a set of documents to exactly one summary selected by the system. We call the (classical and well known) evaluation of this single summary end-to-end evaluation because it measures the end product of the system. This is in contrast to our proposed evaluation of the assumption made by individual summarizers shown in Table 1. A system summary was extracted by a given system because it was high scoring using its θ, but we ask the question whether optimizing this θ made sense in the first place.

We first observe that scores are relatively low. Summarization is not a solved problem and the systems we investigated can not identify correctly what makes a good summary. This is in contrast to the picture in the classical end-to-end evaluation with ROUGE where state-of-the-art systems score relatively high. Some Ndcg scores are higher (for TAC-2008) which explains why these systems can extract relatively good summaries in the end-to-end evaluation. In this classical evaluation, only the single best summary is evaluated, which means that a system does not need to be able to rank all

[1] http://tac.nist.gov/2009/Summarization/, http://tac.nist.gov/2008/Summarization/

	TAC-2008						TAC-2009					
	responsiveness			Pyramid			responsiveness			Pyramid		
θ	r	ρ	Ndcg	r	ρ	Ndcg	r	ρ	Ndcg	r	ρ	Ndcg
TF$*$IDF-1	.1777	.2257	.5031	.1850	.2386	.3575	.1996	.2282	.3826	.2514	.2890	.2280
TF$*$IDF-2	.0489	.1548	.5952	.0507	.1833	.4811	.0061	.1736	.4984	.1073	.2383	.3844
ICSI	.1069	.1885	.6153	.1147	.2294	.5228	.1050	.1821	.5707	.1379	.2466	.5016
JS-1	**.2504**	**.2762**	.4411	**.2798**	**.3205**	.2804	.2021	.2282	.3896	.2616	.3042	.2272
JS-2	.0383	.1698	.5873	.0410	.2038	.4804	.0284	.1475	.5646	.0021	.2084	.4734
LexRank	.1995	.1821	.6618	.2498	.2168	.5935	.2831	.2585	.6028	**.3714**	.3421	.5764
LSA	.0437	.1137	**.6772**	.1144	.1131	**.5997**	**.2965**	.2127	**.6641**	.3677	.2935	**.6467**
Edmunds.	.2223	.2686	.6372	.2665	.3164	.5521	.2598	**.2604**	.5852	.3647	**.3720**	.5594
KL-1	.1796	.2249	.4899	.2016	.2690	.3439	.1827	.2275	.4047	.2423	.2981	.2466
KL-2	.0023	.1661	.6165	.0023	.1928	.5135	.0437	.1435	.6171	.0211	.2060	.5462
Pyramid	.7031	.6606	.8528	—	—	—	.7174	.6414	.8520	—	—	—

Table 1: Correlation of θ functions with human judgments across various systems.

possible summaries correctly.

We see that systems with high end-to-end ROUGE scores (according to Hong et al. (2014)) do not necessarily have a good model of summary quality. Indeed, the best performing θ functions are not part of the systems performing best with ROUGE. For example, ICSI is the best system according to ROUGE, but it is not clear that it has the best model of summary quality. In TAC-2009, LexRank, LSA and the heuristic Edmundson have better correlations with human judgments. The difference with end-to-end evaluation might stem from the fact that ICSI solves the optimization problem exactly, while LexRank and Edmundson use greedy optimizers. There might also be some side-effects from which ICSI profits: extracting sentences to improve θ might lead to accidentally selecting suitable sentences, because θ can merely correlate well with properties of good summaries, while not modeling these properties itself.

It is worth noting that systems perform differently on TAC2009 and TAC2008. There are several differences between TAC2008 and TAC2009 like redundancy level or guidelines for annotations; for example, responsiveness is scored out of 5 in 2008 and out of 10 in 2009. The LSA summarizer ranks among the best systems in TAC2009 with pearson's r but is closer to the worst systems in TAC2008. While this is difficult to explain we hypothesize that the model of summary quality from LSA is sensitive to the slight variations and therefore not robust. In general, any system which claims to have a better θ than previous works should indeed report results on several datasets to ensure robustness and generality.

Interestingly, we observe that the correlation between Pyramid and responsiveness is better than in any system, but still not particularly high. Responsiveness is an overall annotation while Pyramid is a manual measure of content only. These results confirm the intuition that humans take into account much more aspects when evaluating summaries.

4 Related Work and Discussion

While correlation analyses on human judgment data have been performed in the context of validating automatic summary evaluation metrics (Louis and Nenkova, 2013; Nenkova et al., 2007; Lin, 2004), there is no prior work which uses these data for a principled comparison of summarizers.

Much previous work focused on efficient optimizers O, such as ILP, which impose constraints on the θ function. Linear (Gillick and Favre, 2009) and submodular (Lin and Bilmes, 2011) θ functions are widespread in the summarization community because they can be optimized efficiently and effectively via ILP (Schrijver, 1986) and the greedy algorithm for submodularity (Fujishige, 2005). A greedy approach is often used when θ does not have convenient properties that can be leveraged by a classical optimizer (Haghighi and Vanderwende, 2009).

Such interdependencies of O and θ limit the expressiveness of θ. However, realistic θ functions are unlikely to be linear or submodular, and in the well-studied field of optimization there exist a range of different techniques developed to tackle difficult combinatorial problems (Schrijver, 2003; Blum and Roli, 2003).

A recent example of such a technique adapted to extractive summarization are meta-heuristics used to optimize non-linear, non-submodular objective functions (Peyrard and Eckle-Kohler, 2016).

Other methods like Markov Chain Monte Carlo (Metropolis et al., 1953) or Monte-Carlo Tree Search (Suttner and Ertel, 1991; Silver et al., 2016) could also be adapted to summarization and thus become realistic choices for O. General purpose optimization techniques are especially appealing, because they offer a decoupling of θ and O and allow investigating complex θ functions without making any assumption on their mathematical properties. In particular, this supports future work on identifying an "optimal" θ as a model of relevant quality aspects of a summary.

5 Conclusion

We presented a novel evaluation framework for ES which is based on the proof that ES is equivalent to the problem of choosing an objective function θ and an optimizer O. This principled and well-defined framework allows evaluating θ and O of any extractive summarizer – separately and independently. We believe that our framework can serve as a basis for future work on identifying an "optimal" θ function, which would provide an answer to the central question of what are the properties of a "good" summary.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1, and via the German-Israeli Project Cooperation (DIP, grant No. GU 798/17-1).

References

Christian Blum and Andrea Roli. 2003. Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. *ACM Computing Surveys* 35(3):268–308.

H. P. Edmundson. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery* 16(2):264–285.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal of Artificial Intelligence Research* pages 457–479.

Satoru Fujishige. 2005. *Submodular functions and optimization*. Annals of discrete mathematics. Elsevier, Amsterdam, Boston, Paris.

Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*. Association for Computational Linguistics, Boulder, Colorado, pages 10–18.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, pages 362–370.

Kai Hong, John Conroy, benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pages 1608–1616.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, Barcelona, Spain, pages 74–81.

Hui Lin and Jeff A. Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 510–520.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics* 39(2):267–300.

Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development* 2:159–165.

Ryan McDonald. 2007. A Study of Global Inference Algorithms in Multi-document Summarization. In *Proceedings of the 29th European Conference on IR Research*. Springer-Verlag, Rome, Italy, pages 557–564.

Nicholas Metropolis, Arianna Rosenbluth, Marshall Rosenbluth, Augusta Teller, and Edward Teller. 1953. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21:1087 – 1092.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing (TSLP)* 4(2).

Maxime Peyrard and Judith Eckle-Kohler. 2016. A General Optimization Framework for Multi-Document Summarization Using Genetic Algorithms and Swarm Intelligence. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 247 – 257.

Alexander Schrijver. 1986. *Theory of Linear and Integer Programming*. John Wiley & Sons, Inc., New York, NY, USA.

Alexander Schrijver. 2003. *Combinatorial Optimization - Polyhedra and Efficiency*. Springer, New York.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of the 7th International Conference on Information Systems Implementation and Modelling (ISIM '04)*. Rožnov pod Radhoštěm, Czech Republic, pages 93–100.

Christian Suttner and Wolfgang Ertel. 1991. Using Back-Propagation Networks for Guiding the Search of a Theorem Prover. *International Journal of Neural Networks Research & Applications* 2(1):3–16.