The limits of automatic summarisation according to ROUGE

Natalie Schluter

Department of Computer Science
IT University of Copenhagen
Copenhagen, Denmark
natschluter@itu.dk

Abstract

This paper discusses some central caveats of summarisation, incurred in the use of the ROUGE metric for evaluation, with respect to optimal solutions. The task is NP-hard, of which we give the first proof. Still, as we show empirically for three central benchmark datasets for the task, greedy algorithms empirically seem to perform optimally according to the metric. Additionally, overall quality assurance is problematic: there is no natural upper bound on the quality of summarisation systems, and even humans are excluded from performing optimal summarisation.

1 Introduction

Research in automatic summarisation today has reached a stalemate. Despite continuing innovation of promising algorithms for carrying out automatic summarisation, recent research over conventional benchmark datasets has suggested the following: according to the most widely accepted automatic evaluation metric, ROUGE, there has been no substantial improvement in performance on central datasets in the field in the last decade (Hong et al., 2014). Additionally, according to ROUGE, there seems to be little significant benefit to supervised over unsupervised learning, or to exact over greedy approximate algorithmic solutions. Moreover, there is little understanding as to what a perfect score is according to ROUGE, or how naturally this describes a human's idea of an *optimal* summary.

In this paper we substantiate these issues with evidence, observing that by ROUGE numbers:

(1) **Perfect scores for extractive summarisation are theoretically computationally hard to achieve.** We provide the first proof of NP-hardness for optimisation of extractive summarisation with respect to ROUGE. Yet empirically the metric shows that greedy and exact global decoding method performances are similar.

(2) **100% perfect scores are impossible for higher quality datasets.** The metric returns an average of ROUGE scores over multiple reference summaries in order to avoid bias (Nenkova and Passonneau, 2004). This means that it is impossible to obtain 100% ROUGE-n scores unless the reference summaries contain precisely the same n-grams.

(3) **Relative perfect scores are highly diverse and unattainable by humans.** ROUGE scores are generally rather low for short summaries and seem to get higher for datasets with longer summary length budgets, even when document length also substantially increases. We know that 100% perfect scores are impossible, so what is a perfect score according to ROUGE? How do we know when no improvement is possible? Previous research on evaluation metrics for automatic summarisation has tried to empirically show a correlation between human judgments and system output quality (Lin, 2004; Lin and Hovy, 2003; Liu and Liu, 2008; Graham, 2015). But this does not address the upper bound issue. Indeed, we demonstrate there is no possible relative perfect score, even if one has access to the sentences of the reference summaries. So, for example, even humans are doomed to perform sub-optimally (Cf. Marujo et al. (2016)).

(4) **State-of-the-art automatic summarisation is unsupervised.** There have been recent advances in supervised summarisation mainly

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 41–45,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

with respect to supervised learning using neural networks (for example (Rush et al., 2015; Chopra et al., 2016)). However, due to data size requirements, these systems are constrained to title generation systems and therefore not in the scope of this work. Hong et al. (2014) survey the state-of-the-art using the central DUC 2004 dataset. Of these, `ICSISum` (Gillick and Favre, 2009) is the only global summariser using an *exact* algorithm; it obtains the best ROUGE-2 score without supervision. All the other approaches use greedy strategies/approximations, even if they intend to model global optimisation. This raises the following important question: If one shifts from a greedy strategy to an exact global one, does supervision give substantial system performance improvement?

In this paper, we do not consider or compare evaluation metrics. This work is all under the assumption that ROUGE (under its currently used parameters) provides an accurate account of summarisation quality.[1]

Throughout, we refer to as **reference summaries** the gold standard that accompanies the summarisation dataset. Reference summaries are probably abstractive. On the other hand, by **gold summaries**, we refer to optimal summaries consisting of sentences from the input document.

2 Preliminaries

ROUGE. Let g be an n-gram and R and S be multiset representations of reference and system summaries, respectively. We define the intersection $A \cap B$ of two multisets A, B as a multiset containing all multiples of their shared elements.

$$\text{ROUGE-}n(S) := \frac{\sum_{g \in S} |\{g|g \in S\} \cap \{g|g \in R\}|}{\sum_{g \in R} |\{g|g \in R\}|} \tag{1}$$

When there is more than one reference summary, then the individual ROUGE scores are calculated per reference and the average is returned.

The data. Empirical results of this paper are calculated over datasets from three separate domains.
<u>duc04</u>: 30 newswire article set-summary set pairs first used in the DUC 2004 summarisation task 2.[2] We use both the original 665 bytes summary budget as well as the 100 word summary budget used by (Hong et al., 2014).
<u>echr</u>: judgment-summary pairs scraped from the European Court of Human Rights case-law website, HUDOC.[3] The test set consists of 138 pairs. We adopt the same summary budget length: 805 words used by Schluter and Søgaard (2015).
<u>wiki</u>: Wikipedia leading paragraphs-article pairs (all labeled “good article”) from a comprehensive dump of English language Wikipedia articles.[4] The test set consists of 111 pairs. We use the same summary budget of 335 used by Schluter and Søgaard (2015).

3 ROUGE optimisation for extraction

We now provide a proof of NP-hardness of exact oracle extractive summarisation with respect to ROUGE. We first prove the result for ROUGE-1 and later extend the result to ROUGE-n.

Theorem 1. *Given a document, its manually written non-extractive summary, and the ROUGE-1 metric for $N \in \mathbb{Z}_+$, building an extractive summary that maximises the ROUGE-1 metric is NP-hard.*

Proof. The objective is to optimise ROUGE-1 by maximising the number of word tokens paired up between system and reference summaries. That is, one is trying to choose the sentences, within budget, that cumulatively maximise the number of unigram tokens that can be paired with those of reference summaries. We can reduce the NP-hard *max k-weighted dominating set problem* to the oracle extractive summarisation problem with ROUGE-1 as the metric.

Given a graph $G = (V, E)$, the max k-*dominating set problem* requires a solution of k vertices that are adjacent to the maximum number of vertices in G. The max k-dominating set problem is NP-hard, even for cubic graphs (graphs in which the degree of all vertices is equal to 3) (Garey and Johnson, 1979).

Suppose further that each vertex $s \in V$ is associated with a weight w_v. The max k'-weighted dominating set problem consists in determining a subset of vertices of total weight k' that are adjacent with the maximum number of vertices in G.

[1] We use the current version ROUGE-1.5.5 `http://www.berouge.com`, with the following parameters unless otherwise stated: `-n 2 -m -x -f A -t 0 {-b|-l} [length] -a -r 1000 -c 95`.

[2] `http://duc.nist.gov/duc2004/`

[3] `http://hudoc.echr.coe.int/`

[4] `https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles-multistream.xml.bz2`

In particular, if we set $w_v = 1$ for each vertex, then the two problems are identical, showing the corresponding NP-hardness of this weighted version of the problem.

Let $G = (V, E)$ be a cubic graph. Let $N(v)$ be the neighbourhood of vertex v. Now let the weight of each vertex w_v be $|N(v) \cup \{v\}| = 4$. With $k' = 4k$ it is easy to see that the max k'-weighted dominating set problem is equivalent to the max k-dominating set problem for cubic graphs. A solution is a dominating set S' such that $|\{u \mid u \in (N(v) \cup \{v\}), v \in S'\}|$ is maximised for $\sum_{v \in S'} w(v) = 4k$.

We reduce the $4k$-weighted dominating set problem to the problem of exact summarisation with respect to ROUGE-1 as follows.

We create an input document $D = \{s_v \mid v \in V\}$, where $s_v := N(v) \cup \{v\}$ is a sentence (its components written in any order). Evaluation is carried out against a single reference summary V (the set of vertices of our original graph written out in any order). Let S be an output extractive summary from D within our budget of size $4k$. We want to maximise

ROUGE-1$(S) =$

$$= \frac{\sum_w |\{w|w \in \bigcup_{s_v \in S} s_v\} \cap \{w|w \in V\}|}{\sum_w \{w|w \in V\}}$$

$$= \frac{|(\bigcup_{s_v \in S} s_v) \cap V|}{|V|} = \frac{|(\bigcup_{s_v \in S} s_v)|}{|V|}$$

$$= \frac{|\{u \mid u \in (N(v) \cup \{v\}), s_v \in S\}|}{|V|} \tag{2}$$

where the second equality follows from the fact that no vertex occurs more than once in the reference summary V.

Maximising the last term (2) is the same as maximising without its denominator. Take $S' := \{v \mid s_v \in S\}$ for the solution of the original $4k$-weighted dominating set problem. Suppose S' was not a maximum solution. Then there is a better solution $\hat{S}$ of weight $4k$. But then $\{s_v \mid v \in \hat{S}\}$ is a better solution for summarisation. This gives the result. □

We can extend the reduction in the proof of Theorem 1 from $4k$-weighted dominating set to extractive summarisation with respect to ROUGE-n with budget $2 \cdot (4k)$ by introducing a dummy symbol d into our documents and summaries for padding sentences. We first introduce some notation for the new sentences of documents and reference summaries.

We will now write sentences s_v from the proof of Theorem 1 with the superscript 1, s^1_v, corresponding to the type of gram (1-gram) measured in ROUGE-1. We set an ordering on V, numbering the vertices so that $V := \{v_1, \ldots, v_{|V|}\}$ (though this ordering is purely for ease in description). Instead of simply choosing any order to write the nodes from $N(v_{i_1}) \cup \{v_{i_1}\} = \{v_{i_1}, v_{i_2}, v_{i_3}, v_{i_4}\}$, we write $s^1_{v_{i_1}}$ according to the ordering of the node indices. So, if $i_1 < i_2 < i_3 < i_4$, then $s^1_{v_{i_1}} = v_{i_1} v_{i_2} v_{i_3} v_{i_4}$.

We generalise this to order-n sentences. The order-n sentence s^n_v is just s^1_v (first order sentence) with each vertex padded to the right by the string $d^{(n-1)}$, and prefixed with $d^{(n-1)}$ to the resulting string, where d is a dummy symbol not in V. For example, $s^2_{v_{i_1}} = dv_{i_1} dv_{i_2} dv_{i_3} dv_{i_4} d$, and in general, $s^n_{v_{i_1}} = d^{(n-1)} v_{i_1} d^{(n-1)} v_{i_2} d^{(n-1)} v_{i_3} d^{(n-1)} v_{i_4} d^{(n-1)}$. So order-$n$ sentences have length $4 + 5(n - 1)$. Order-n sentences will be used for creating documents D_n and reference summaries V_n for the NP-hardness proof of exact oracle summarisation with respect to ROUGE-n, with a budget of $k(4 + 5(n - 1))$.

Note how if v occurs in a first order sentence s^1, then there are exactly 2 bigrams containing v in the corresponding second order sentence s^2: dv and vd. Similarly, there are exactly n n-grams containing v in the corresponding sentence s^n: $d^{(n-1)}v, d^{(n-2)}vd, \ldots, dvd^{(n-2)}, vd^{(n-1)}$. This is the set-up for the document D_n in the reduction of $(4k)$-weighted dominating set to exact extractive summarisation with respect to ROUGE-n.

We set up the reference summary in a similar way. For $V = V_1$, we write the vertices in order. For V_n we pad the right of each symbol in V_1 with the string $d^{(n-1)}$ and attach the same string as a prefix. So, once again, a 1-gram in V_1 corresponds to exactly n n-grams in V_n. ROUGE-n is maximised when the number of matched n-grams of V_n is maximised, which is precisely when the number of 1-grams of V_1 is maximised. The reduction from $(4k)$-weighted dominating set to exact extractive summarisation with respect to ROUGE-n and with budget $(4+5(n-1))k$ follows, yielding the following generalisation of Theorem 1.

Theorem 2. *Given a document, its manually writ-*

ten non-extractive summary, and the ROUGE-n metric for $n \in \mathbb{Z}_+$, building an extractive summary that maximises the ROUGE-n metric is NP*-hard.*

Because the ROUGE optimisation problem is NP-hard, one may suspect that exchanging a greedy strategy out for an exact global approach would lead to substantial improvements in system performance. Therefore, for our three datasets, we generate gold extractive summaries using both exact and greedy global oracle approaches. If our suspicions are true, then we expect these approaches to generate poor quality gold extractive summaries with the greedy algorithm in comparison to exact one.

	opt	Greedy		Exact	
	w.r.t.	R1	R2	R1	R2
duc04	R1	50.5	13.87	49.91	13.98
	R2	48.27	19.61	46.92	16.79
wiki	R1	64.14	22.49	63.41	21.81
	R2	59.68	27.81	59.43	27.11
echr	R1	83.57	51.01	84.17	50.34
	R2	81.38	57.31	82.04	56.67

Table 1: Exact and greedy oracle summarisation ROUGE-n scores in percentages, for $n \in [2]$.

We use an open source solver to find exact optimal solutions.[5] Note that in the exact set-up sentences cannot be clipped to meet the boundary budget constraint, which is a more natural setting for automatic summarisation. To build an extractive summary greedily, we iteratively add the sentence with highest ROUGE score to the summary, normalising by sentence length. The measure automatically chops sentences that otherwise bring summary lengths over the limit. Table 1 gives the results for greedy and exact oracle gold extractive summaries across our three domains.

Greedy is good. We observe that across the board, the greedy strategy performs comparably to the exact strategy for global optimisation. With the shorter summaries required by the duc04 dataset, the greedy strategy yields higher ROUGE scores, possibly by chopping the last sentence of summaries. This chopping reward lessens, it seems, as summary budgets increase, but the two methods stay competitive with each other.

No data necessary. This also provides good evidence that is no substantial benefit in switching from unsupervised exact global state-of-the-art approaches to supervised exact global approaches for extractive summarisation on conventional datasets.

Far from perfection. For extractive summarisation, the perfect scores (in Table 1) are far from 100% as well as diverse, according to dataset.

Evaluation against multiple, rather than single reference summaries is generally recognised as leading to fairer, better quality, evaluation: different human summaries appear to be good even though they do not have identical content (Nenkova and Passonneau, 2004). However, averaging ROUGE scores across multiple summaries, as is standard practice, makes a perfect 100% score unattainable, even for abstractive systems. This is because the word frequencies required by ROUGE suddenly become unattainable.

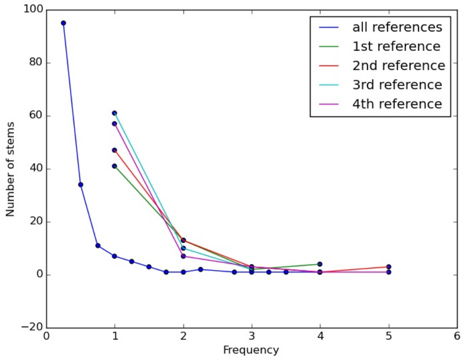

Figure 1: Stemmed word frequencies for reference summary set d30001t from duc04: averaged across all reference summaries and for single reference summaries.

As illustration, consider the frequencies required by the reference summaries for a duc2004 document set in Figure 1. The number of 1-grams to match has increased: this was the original intent—to allow for equally important but different content. We have gone from around 60 stemmed words to 160 stemmed words. However, for example, in the case of our example summary set, 136/160 matches are really only part matches (with weight < 1).

This leads to the contradictory situation where, according to the ROUGE metric, humans cannot summarise well (though they are thought to be

[5] 3gnu.org/software/glpk

able to judge summary quality accurately). Indeed, evaluating one reference summary against the other three for the duc04 dataset achieves 39.92 ROUGE-1 and 9.39 ROUGE-2—far below optimal performance. Since humans are generally abstractive summarisers this provides a sort of upper bound on abstractive summarisation performance according to ROUGE.

4 Concluding remarks

Previous work on summarisation evaluation has mainly considered the positive aspects of ROUGE; namely correlation to human judgments. In this paper we hope to have raised some concerns with respect to ROUGE and our expectations for *optimal* summarisers. We have also given the first NP-hardness proof for global optimisation with respect to ROUGE.

References

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California.

M.R. Garey and D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman and Company.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of ILP*, pages 10–18.

Yvette Graham. 2015. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal.

Kai Hong, John M. 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 *Proc of LREC*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics. In *Proceedings of HLT-NAACL*, Edmonton, AB, Canada.

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

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.

Luís Marujo, Ling Wang, Ricardo Ribeiro, Anatole Gershman, Jaime Carbonell, David Marins de Matos, and João Paulo da Silva Neto, 2016. *Exploring Events and Distributed Representations of Text in Multi-Document Summarization*, volume 94, pages 34–42. Elsevier.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152, Boston, Massachusetts, USA.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal.

Natalie Schluter and Anders Søgaard. 2015. Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 840–844, Beijing, China.