Approximate unsupervised summary optimisation for selections of ROUGE

Natalie Schluter[1, 2] Héctor Martínez Alonso[3]
(1) MobilePay, Danske Bank, Copenhagen, Denmark
(2) IT University of Copenhagen, Denmark
(3) Alpage, INRIA, Univ. Paris Diderot, Sorbonne Paris Cité (France)
nasc@danskebank.dk, hector.martinez-alonso@inria.fr

RÉSUMÉ

Il est d'usage de mesurer les performances d'un système de résumé automatique en utilisant la métrique ROUGE. Malheureusement, cette métrique n'est pas appropriée pour des approches non supervisées. Nous montrons qu'il est cependant possible d'effectuer une optimisation pour une solution approchée de ROUGE-n en utilisant une fonction objective fondée sur une version pondérée par document de ROUGE : "document-weighted ROUGE". Cette méthode permet d'obtenir des performances au niveau de l'état de l'art pour des systèmes de résumé automatique pour le français et l'anglai. Et ceci malgré le fait qu'il n'y ait pas de corrélation entre la métrique ROUGE pondérée au niveau des documents et les jugements humains, contrairement à la métrique ROUGE originale. Ces résultats suggèrent l'existence en théorie d'une relation d'approximation entre ces deux métriques.

ABSTRACT

Approximate summary optimisation for selections of ROUGE

It is standard to measure automatic summariser performance using the ROUGE metric. Unfortunately, ROUGE is not appropriate for unsupervised summarisation approaches. On the other hand, we show that it is possible to optimise approximately for ROUGE-n by using a document-weighted ROUGE objective. Doing so results in state-of-the-art summariser performance for single and multiple document summaries for both English and French. This is despite a non-correlation of the document-weighted ROUGE metric with human judgments, unlike the original ROUGE metric. These findings suggest a theoretical approximation link between the two metrics.

MOTS-CLÉS : résumé automatique, couverture maximale, évaluation.

KEYWORDS: automatic summarisation, maximum coverage, evaluation.

1 Introduction

A standard metric in automatic multi- and single-document summarisation is ROUGE. Certain variants of the metric have been shown to correlate well with human judgments (Lin, 2004). In particular, ROUGE-2 is the ROUGE variant shown to correlate best with human judgments.

While it is impossible to directly optimise for these variants in unsupervised summarisation systems, a state-of-the-art unsupervised summariser optimises for a sort of document-weighted coverage version of ROUGE-2 (Gillick & Favre, 2009). However, optimisation of this coverage version does

not seem to imply optimisation of the original ROUGE metric. Moreover, we show in this paper that the document-weighted metric scores in themselves do not exhibit any correlation with human judgments as certain selections of ROUGE do. Despite this fact, state-of-the-art ROUGE scores for systems optimising for the document-weighted metric are achieved by (Gillick & Favre, 2009). We therefore take up study of this seemingly contradictory situation from an *empirical* point of view : we want to obtain more comprehensive empirical evidence for or against the correlation between document-weighted optimisation and corresponding ROUGE scores.

In this paper, we consider the direct consequences of optimising for separate document-weighted coverage versions of variants of ROUGE that have been shown to correlate most with human judgments. Our results follow that of (Gillick & Favre, 2009), showing that optimising for the document-weighted coverage version of the metric generally yields improvements in the corresponding ROUGE variant. The results are surprising given the non-correlation with human judgments.

We carry out our experiments over both French and English data, for both single- and multi-document summarisation. In doing so, we introduce a new single-document summarisation dataset for French, within the legal text domain, and provide the first results for this dataset with a state-of-the-art summariser.

2 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 elements shared between A and B. The ROUGE-n score of a system summary S with respect to R is defined as follows, where sums are over n-gram types (and not tokens) (Lin, 2004). [1]

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

In extractive summarisation, it is impossible to optimise for ROUGE-n using just the input document as in unsupervised summarisation. Indeed, replacing the reference summary by the input document makes all sentences of the the same length equally beneficial to the summary :

$$\text{unsup-ROUGE-}n(S) := \frac{\sum_{g \in S} |\{g|g \in S\} \cap \{g|g \in D\}|}{\sum_{g \in D} |\{g|g \in D\}|} = \frac{\sum_{g \in S} |\{g|g \in S\}|}{\sum_{g \in D} |\{g|g \in D\}}$$

$$= \frac{(\text{summary budget}) - n}{\sum_{g \in D} |\{g|g \in D\}|} = C_D$$

where C_D is some constant with respect to the fixed input document—in particular, it is independent of any particular output summary S.

Since replacing references by documents for unsupervised ROUGE optimisation is worthless when conducting actual summarization, we must look for alternatives. One such alternative is *document-weighted coverage*.

1. We use and extend the current version ROUGE-1.5.5 `http://www.berouge.com`, with the following generally used parameters unless otherwise stated : `-n 2 -m -x -f A -t 0 -b 665 -a -r 1000 -c 95`.

Document-weighted coverage. A document-weighted coverage version of ROUGE follows the same definition as in Equation 1, except now each match is weighted by the n-gram's frequency in the input document rather than by the multiplicity of the match.

$$\text{doc-weight-ROUGE-}n(S) := \frac{\sum_{g \in S} \max(|\{g|g \in S\}|, |\{g|g \in D\}|)}{\sum_{g \in D} |\{g|g \in D\}|} = \frac{\sum_{g \in S} |\{g|g \in D\}|}{\sum_{g \in D} |\{g|g \in D\}|} \quad (2)$$

Maximum coverage exact summarisation. Since the quantity described by Equation 2 does not depend on any reference summary, it can serve as an unsupervised objective function. With $n = 2$, this is the objective function used by (Gillick & Favre, 2009)'s state-of-the-art unsupervised extractive summariser.

Note that optimising for Equation 2 is the same as optimising without the denominator :

$$\text{doc-weight-ROUGE-}n(S) \propto \sum_{g \in S} |\{g|g \in D\}|.$$

ROUGE, doc-weight-ROUGE and human judgments. Because the maximum coverage objective optimisation has lead to a state-of-the-art summariser, it would be of interest to know whether the metric actually correlates with human judgments on summary quality as (Lin, 2004) has shown ROUGE does. Table 1 provides the correlations for the DUC2004 dataset, where we observe that this is not the case. [2]

	W/ STOPWORDS	W/O STOPWORDS
R1	0.45	0.5
doc-R1	-0.14	-0.009
R2	0.453	0.472
doc-R2	-0.17	-0.063

TABLE 1 – Spearman (Sp-r) correlations with median human coverage judgments.

We will show that despite the poor correlation of document-weighted ROUGE-n, using this metric as an objective function seems to help optimise for the original ROUGE-n.

3 Data

We carry out experiments over both English and French data for single- and multi-document summarisation.

2. Note that the ROUGE correlations presented here are somewhat lower than is found in (Lin, 2004) over a similar dataset. We carried out extensive parameter search to check whether we could obtain more similar scores to the original. These parameter settings indeed provide the best correlations. The human judgment SEE data is available with the dataset distribution.

Single-document summarisation datasets (echr). For single-document summarisation, our datasets consist of judgment-summary pairs scraped from the European Court of Human Rights case-law website, HUDOC, [3] the English judgments (**echr_en**) of which were first used by (Schluter & Søgaard, 2015). The English test set consists of 138 pairs. We adopt the same summary budget length as in their work, namely 805 words.

We introduce the French counterpart of this dataset, produced in precisely the same manner as for English. The French data (**echr_fr**) was divided into training, test, and development sets respectively comprising 1295, 154 and 146 judgment-summary pairs each. The average summary size of the training set is 840 words. As in (Schluter & Søgaard, 2015) we adopt this average for our system summary budget for our test set input.

Multi-document summarisation datasets. The multi-document sets come from different domains depending on the language. For English we use the canonical **duc04** dataset, composed of 30 newswire set-summary set pairs, first used in the DUC 2004 summarisation task 2. [4] We use both the original 665 bytes summary budget.

For French we use the **rpm2** summarization corpus (de Loupy *et al.*, 2010), a collection of French newswire texts encomprising 20 topics that allows multi-document and multi-reference summarization. This corpus has been used in summarization experiments such as Boudin & Torres-Moreno (2009) or Bossard & Guimier De Neef (2011). We use the 100 word budget used in these previous studies.

4 Experiments

Gillick & Favre (2009) showed that in optimising for document-weighted ROUGE-2, improvements in ROUGE-2 scores were achieved. We have shown how document-weighted ROUGE-n seems to hold no correlation with human judgments, so the question remains : is this a coincidence ?

We want to test whether there is any empirical evidence showing that optimising for document-weighted ROUGE-n leads to improvements in original ROUGE-n scores. To do this we extend Gillick & Favre (2009)'s summariser `icsisumm` [5] to optimise over the required n-grams for n-gram coverage corresponding to document-weighted ROUGE-n.

French stemming. The summarization system of (Gillick & Favre, 2009) makes use of stemming to build its concept inventory. Likewise, ROUGE offers the possibility of using stemming for evaluation. However, both tools use English stemmers. In order to run comparable systems for both English and French, we have incorporated French stemming into our summarization and evaluation by replacing the English stemmer of both tools with the French Snowball stemmer provided in NLTK. [6]

Moreover, both systems make also use of English stop lists, which we have replaced with the French stoplist from NLTK. (Note that the ROUGE scores presented here are obtained without stopword filtering.)

3. `http://hudoc.echr.coe.int/`
4. `http://duc.nist.gov/duc2004/`
5. `https://github.com/benob/icsisumm`
6. `http://www.nltk.org/api/nltk.stem.html`

4.1 Results

Tables 2 and 3 show the ROUGE-1 through ROUGE-4 scores (columns) when optimising for each document-weighted ROUGE metric.

For English, the ROUGE-1 and ROUGE-2 scores are close to state-of-the-art, as expected, when optimising for document-weighted ROUGE-2. Interestingly, we achieve the best ROUGE-2 score for both domains and tasks when optimising for document-weighted ROUGE-3. However, the important trend to notice here is that except for ROUGE-1, scores in the diagonal for each domain are in the top 2 of their column. This provides support evidence for the relation between optimising for document-weighted ROUGE and the original ROUGE, despite zero correlations with human judgments.

	duc04				echr_en			
objective	R1	R2	R3	R4	R1	R2	R3	R4
doc-weight-R1	37.76	5.78	1.5	0.52	65.68	23.77	9.81	5.64
doc-weight-R2	39.82	9.83	3.61	2.00	68.38	29.95	13.50	7.86
doc-weight-R3	39.05	10.21	4.27	2.37	67.62	30.70	15.45	9.24
doc-weight-R4	37.14	8.66	3.66	2.06	66.94	30.18	15.25	9.55

TABLE 2 – English ROUGE scores.

	rpm2				echr_fr			
objective	R1	R2	R3	R4	R1	R2	R3	R4
doc-weight-R1	35.91	8.82	3.44	1.39	59.42	20.71	9.48	6.11
doc-weight-R2	36.99	12.23	6.17	3.52	59.81	24.29	12.37	8.31
doc-weight-R3	33.93	10.80	5.31	3.01	60.1	25.67	13.71	9.35
doc-weight-R4	35.17	11.72	6.38	3.93	60.97	25.99	14.32	10.00

TABLE 3 – French ROUGE scores.

The trend of high diagonal scores continues for the French data. For French, we note that, the ROUGE scores are similar to those reported in Boudin & Torres-Moreno (2009). However, our variant of summarization and evaluation for French includes French stemming, making the figures not strictly comparable.

5 Conclusions

We cannot optimise for ROUGE in an unsupervised fashion. However, we have shown in this paper that it is possible to optimise approximately for ROUGE by using a document-weighted ROUGE objective. Doing so results in state-of-the-art summariser performance. This is despite a non-correlation of the document-weighted ROUGE metric with human judgments, unlike the original ROUGE metric. These empirical results should further verified on complementary datasets. Future study will attempt

to make explicit the theoretical ties between the two metrics from which the empirical results shown here should stem.

Références

BOSSARD A. & GUIMIER DE NEEF E. (2011). Etude de l'impact du regroupement automatique de phrases sur un système de résumé multi-documents. In *huitième Conférence en Recherche d'Information et Applications*, p.8̃, Avignon, France.

BOUDIN F. & TORRES-MORENO J.-M. (2009). Résumé automatique multi-document et indépendance de la langue : une première évaluation en français. *Proceedings of Traitement Automatique de la Langue Naturelle (TALN'09), Senlis*.

DE LOUPY C., GUÉGAN M., AYACHE C., SENG S. & TORRES-MORENO J.-M. (2010). A french human reference corpus for multi-document summarization and sentence compression. In *LREC*.

GILLICK D. & FAVRE B. (2009). A scalable global model for summarization. In *Proc of ILP*, p. 10–18.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Proc of WAS*, Barcelona, Spain.

SCHLUTER N. & SØGAARD A. (2015). Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In *Proc of ACL*, Beijing, China.