Learning to Score System Summaries for Better Content Selection Evaluation.

Maxime Peyrard and **Teresa Botschen** and **Iryna Gurevych**
Research Training Group AIPHES and UKP Lab
Computer Science Department, Technische Universität Darmstadt
www.aiphes.tu-darmstadt.de, www.ukp.tu-darmstadt.de

Abstract

The evaluation of summaries is a challenging but crucial task of the summarization field. In this work, we propose to learn an automatic scoring metric based on the human judgements available as part of classical summarization datasets like TAC-2008 and TAC-2009. Any existing automatic scoring metrics can be included as features, the model learns the combination exhibiting the best correlation with human judgments. The reliability of the new metric is tested in a further manual evaluation where we ask humans to evaluate summaries covering the whole scoring spectrum of the metric. We release the trained metric as an open-source tool.

1 Introduction

The task of automatic multi-document summarization is to convert source documents into a condensed text containing the most important information. In particular, the question of evaluation is notably difficult due to the inherent lack of gold standard.

The evaluation can be done manually by involving humans in the process of scoring a given system summary. For example, with the *Responsiveness* metric, human annotators score summaries on a LIKERT scale ranging from 1 to 5. Later, the *Pyramid* scheme was introduced to evaluate content selection with high inter-annotator agreement (Nenkova et al., 2007).

Manual evalations are meaningful and reliable but are also expensive and not reproducible. This makes them unfit for systematic comparison.

Due to the necessity of having cheap and reproducible metrics, a significant body of research was dedicated to the study of automatic evaluation metrics. Automatic metrics aim to produce a semantic similarity score between the candidate summary and a pool of reference summaries previously written by human annotators (Lin, 2004; Yang et al., 2016; Ng and Abrecht, 2015). Some variants rely only on the source documents and the candidate summary ignoring the reference summaries (Louis and Nenkova, 2013; Steinberger and Ježek, 2012).

In order to select the best automatic metric, we typically consider manual evalution metrics as our gold standard, then a good automatic metric should reliably predict how well a summarizer would perform if human evaluation was conducted (Owczarzak et al., 2012; Lin, 2004; Rankel et al., 2013).

In practice, we use the human judgment datasets like the ones constructed during the manual evaluation of the Text Analysis Conference (TAC). The system summaries submitted to the shared tasks were manually scored by trained human annotators following the Responsiveness and/or the Pyramid schemes. An automatic metric is considered good if it ranks the system summaries similarly as humans did.

Currently, ROUGE (Lin, 2004) is the accepted standard for automatic evaluation of content selection because of its simplicity and its good correlation with human judgments. However, previous works on evaluation metrics comparison averaged scores of summaries over topics for each system and then computed the correlation with averaged scores given by humans. ROUGE works well in this scenario which compares only systems after aggregating their scores for many summaries. We call this scenario *system-level correlation analysis*.

A more natural analysis, which we use in this work, is to compute the correlation between the candidate metric and human judgments for each topic indivually and then average these correlations over topics. In this scenario, which we call *summary-level correlation analysis*, the performance of ROUGE significantly drops meaning that on average ROUGE does not really identify summary quality, it can only rank systems after aggregation of many topics.

In order to advance the field of summarization we need to have more consistent metrics correlating well with humans on every topic and capable of estimating the quality of individual summaries (not just systems).

We propose to rely on human judgment datasets to learn an automatic scoring metric. The learned metric presents the advantage of being explicitly trained to exhibit high correlation with the "gold-standard" human judgments at the summary level (and not just at the system level). The setup is also convenient because any already existing automatic metric can be incorporated as a feature and the model learns the best combination of features matching human judgments.

We should worry whether the learned metric is reliable. Indeed, typical human judgment datasets (like the ones from TAC-2008 or TAC-2009) contain manual scores only for several system summaries which have a limited range of quality. We conduct a manual evaluation specifically designed to test the metric accross its whole scoring spectrum.

To summarize our contributions: We performed a summary-level correlation analysis to compare a large set of existing evaluation metrics. We learned a new evaluation metric as a combination of existing ones to maximize the summary-level correlation with human judgments. We conducted a manual evaluation to test whether learning from available human judgment datasets yields a reliable metric accross its whole scoring spectrum.

2 Related Work

Automatic evaluation of content has been the subject of a lot of research. Many automatic metrics have been developed and we present here some of the most important ones.

ROUGE (Lin, 2004) simply computes the n-gram overlap between a system summary and a pool of reference summaries. It has become a de-facto standard metric because of its simplicity and high correlation with human judgments at the system-level. Afterwards, Ng and Abrecht (2015) extended ROUGE with word embeddings. Instead of hard lexical matching of n-grams, ROUGE-WE uses soft matching based on the cosine similarity of word embedding.

Recently, a line of research aimed at creating strong automatic metrics by automating the Pyramid scoring scheme (Harnly et al., 2005). Yang et al. (2016) proposed PEAK, a metric where the components requiring human input in the original Pyramid annotation scheme are replaced by state-of-the-art NLP tools. It is more semantically motivated than ROUGE and approximates correctly the manual Pyramid scores but it is computationally expensive making it difficult to use in practice.

Some other metrics do not make use of the reference summaries, they compute a score based only on the candidate summary and the source documents (Lin et al., 2006; Louis and Nenkova, 2013). One representative of this class is the Jensen Shannon (JS) divergence, an information-theoretic measure comparing system summaries and source documents with their underlying probability distributions of n-grams. JS divergence is simply the symmetric version of the well-known Kullback-Leibler (KL) divergence (Haghighi and Vanderwende, 2009).

Little work has been done on the topic of learning an evaluation metric. Conroy and Dang (2008) previously investigated the performances of ROUGE metrics in comparison with human judgments and proposed ROSE (ROUGE Optimal Summarization Evaluation) a linear combination of ROUGE metrics to maximize correlation with human responsiveness. We also look for a combination of features which correlates well with human judgements but, in contrast to Conroy and Dang (2008), we include a wider set of metrics: ROUGE scores, other evaluation metrics (like Jensen-Shannon divergence) and features typically used by summarization systems.

Hirao et al. (2007) also proposed a related approach. They used a voting based regression to score summaries with human judgments as gold standard. Our setup is different because we train and evaluate our metric with the summary-level correlation analysis instead of the system-level one. Our experiments are done on multi-document datasets whereas they use single-documents. Finally, we also perform a further manual evaluation to test the metric outside of its training domain.

3 Approach

Let a dataset D contain m topics. A given topic t_i consists of a set of documents $\mathcal{D}_i$, a set of reference summaries θ_i, a set of n system summaries $\mathcal{S}_i$ and the scores given by humans to the n summaries of $\mathcal{S}_i$ noted $\mathcal{R}_i$. We note $s_{i,j}$ the j-th summary of the i-th topic and $r^h_{i,j}$ the score it received from manual evaluation:

$$\begin{aligned} t_i &= (\mathcal{D}_i, \theta_i, \mathcal{S}_i, \mathcal{R}_i) \\ \mathcal{S}_i &= [s_{i,1}, \ldots, s_{i,n}] \\ \mathcal{R}_i &= [r^h_{i,1}, \ldots, r^h_{i,n}] \end{aligned} \tag{1}$$

An automatic evaluation metric is a function taking as input a document set $\mathcal{D}_i$, a set of reference summaries θ_i and a candidate system summary s and outputs a score. For simplicity, we note: $\sigma(\mathcal{D}_i, \theta_i, s) = \sigma_i(s)$ the score of s as a summary of the i-th topic according to some scoring metric σ.

We search an automatic scoring function σ such that $\sigma_i(s_{i,j})$ correlates well with the manual scores $r^h_{i,j}$.

The final score can be computed at the system-level by aggregating scores over topics before and then computing the correlation or at the summary-level by computing the correlation for each topic and then averaging over topics. We briefly present the difference between the two in the following paragraphs.

System-level correlation Let K be any correlation metric operating on two lists of scored elements, then the system-level correlation is computed by the following formula:

$$\begin{aligned} K^{sys}_{avg} = K([\sum_i^m \sigma_i(s_{i,1}), \ldots, \sum_i^m \sigma_i(s_{i,n})], \\ [\sum_i^m r^h_{i,1}, \ldots, \sum_i^m r^h_{i,n}]) \end{aligned} \tag{2}$$

Both terms in K are lists of size n. The scores for the summaries of the l-th summarizer are aggregated to form the l-th element of the lists. The correlation is computed on the two aggregated lists. Therefore, K^{sys}_{avg} only indicates whether the evaluation metrics can rank systems correctly after aggregation of many summary scores but it ignores individual summaries. It has been used before because evaluation metrics were initially tasked to compare systems.

Summary-level correlation Instead, we advocate for the summary-level correlation which is computed by the following formula:

$$\begin{aligned} K^{summ}_{avg} = \frac{1}{m} \cdot \sum_{t_i \in D} K([\sigma_i(s_{i,1}), \ldots, \sigma_i(s_{i,n})], \\ [r^h_{i,1}, \ldots, r^h_{i,n}]) \end{aligned} \tag{3}$$

Here, we compute the correlation between human judgments and automatic scores for each topic and then average the correlation scores over topics. This measures how well evaluation metrics correlate with human judgments for summaries and not only for systems which is important in order to have finer grain of understanding.

From now on, when we refer to correlation with human judgments we will refer to the summary-level correlation.

Correlation metrics There exist many possible choices for K. As different correlation metrics measure different properties, we use three complementary 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 the automatic metric and the human judgments.

Spearman's ρ is a rank correlation metric which compares the ordering of systems induced by the automatic metric 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 the automatic metric can recognize the best summaries.

3.1 Features

The choice of features is a crucial part of every learning setup. Here, we can benefit from the large amount of previous works studying signals of summary quality. We can classify these signals in three categories.

First, any existing automatic scoring metric can be a feature. These metrics use the candidate summary and the reference summary to output a score.

The second category contains the previous summarization systems having an explicit formulation of summary quality. These systems can implicitly score any summary, then they extract the summary with maximal score via optimization techniques

(Gillick and Favre, 2009; Haghighi and Vanderwende, 2009). Optimization-based systems have recently become popular (McDonald, 2007). Such features score the candidate summary based only on the document sources and the summary itself.

The last category contains the metrics producing a score based only on the summary. Examples of such metrics include readability or redundancy.

Clearly, features using reference summaries (existing automatic metrics) are expected to be more useful for our task. However, it has been shown that some metrics of the second category (like JS divergence) also contain useful signal to approximate human judgments (Louis and Nenkova, 2013). Therefore, we use features coming from all three categories expecting that they are sensitive to different properties of a good summary.

We considered only features cheap to compute in order to deliver a simple and efficient tool. We now briefly present the selected features.

Features using reference summaries **ROUGE-N** (Lin, 2004) computes the n-gram overlap between the candidate summary and the pool of reference summaries. We include as features the variants identified by Owczarzak et al. (2012) as strongly correlating with humans: ROUGE-2 recall with stemming and stopwords not removed (giving the best agreement with human evaluation), and ROUGE-1 recall (the measure with the highest ability to identify the better summary in a pair of system summaries).

ROUGE-L (Lin, 2004) considers each sentence of the candidate and reference summaries as sequences of words (after stemming). It interprets the longest common subsequence between sentences as a similarity measure. An overall score for the candidate summary is given by combining the scores of individual sentences. One advantage of using ROUGE-L is that it does not require consecutive matches but in-sequence matches reflecting sentence-level word order.

JS divergence measures the dissimilarity between two probability distributions. In summarization, it was also used to compare the n-gram probability distribution of a summary and souce documents (Louis and Nenkova, 2013), but here we employ it for comparing the n-gram probability distribution of the candidate summary with the reference summaries. Thus, it yields an information-theoretic measure of the dissimilarity between the candidate summary and the reference summaries.

If θ_i is the set of reference summaries for the i-th topic, then we compute the following score:

$$JS_{ref}(s, \theta_i) = \frac{1}{|\theta_i|} \sum_{ref \in \theta_i} JS(s, ref) \quad (4)$$

ROUGE-WE (Ng and Abrecht, 2015) is the variant of ROUGE-N replacing the hard lexical matching by a soft matching based on the cosine similarity of word embeddings. We use ROUGE-WE-1 and ROUGE-WE-2 as part of our features.

FrameNet-based metrics ROUGE-WE proposes a statistical approach (word embeddings) to alleviate the hard lexical matching of ROUGE. We also include a linguistically motivated one. We replace all nouns and verbs of the reference and candidate summaries with their FrameNet (Baker et al., 1998) frames. This frame annotation is done with the best-performing system configuration from Hartmann et al. (2017) pre-trained on all FrameNet data. It assigns a frame to a word based on the word itself and the surrounding context in the sentence.

Frames are more abstract than words, thus different but related words might be associated with the same frames depending on the meaning of the words in the respective context. ROUGE-N can now match related words through their frames. We also use the unigram and bigram variants (Frame-N).

Semantic Vector Space Similarities In general, automatic evaluation metrics comparing system summaries with reference summaries propose a kind of semantic similarity between summaries. Finding good automatic evaluation metric is hard because the task of textual semantic similarity is challenging. With the development of word embeddings (Mikolov et al., 2013), several semantic similarities have arisen exploiting the inherent similarities built in vector space models. We include one such metric: AVG_{SIM}, the cosine similarity between the average word embeddings of the system summary and the reference summaries. To reduce noise, we exclude stopwords.

Features using document sources are inspired by existing summarization systems:

TF⋆IDF comes from the seminal work from Luhn (1958). Each sentence in the summary is scored according to the TF*IDF of its term. The score of the summary is the sum of the scores of its sentences. We computed the version based on unigrams and bigrams (TF∗IDF-N).

N-gram Coverage is inspired by the strong summarizer ICSI (Gillick and Favre, 2009). Each n-gram in the summary is scored with the frequency it has in the source documents. The final score of the system summary is the sum of the scores of its n-grams. We also use the variants based on unigrams and bigrams (Cov-N).

KL and JS measures the KL or JS divergence between the word distributions in the summary and source documents. We use as features both KL and JS based on unigram and bigram distributions (KL-N and JS-N).

Features using the candidate summary only Finally, we also include a redundancy metric based on n-gram repetition in the summary. It is the number of unique n-grams divided by the total number of n-grams in the summary. We also use unigrams and bigrams (Red-N).

3.2 Model

For a given topic t_i, let ϕ be the function taking as input a document set $\mathcal{D}_i$, a set of reference summaries θ_i and a system summary s and outputting the set of features described earlier. We note $\phi(\mathcal{D}_i, \theta_i, s) = \phi_i(s)$, the feature set representing s as a summary of the topic i.

We aim to learn a function σ_ω with parameters ω scoring summaries similarly as humans would. If $\sigma_\omega(\phi_i(s))$ is the score given by the learned metric to the summary s, we look for the set of parameters ω which maximizes the summary-level correlation defined by equation 3. It means we are trying to solve the following problem:

$$\underset{\omega}{\text{argmax}} \sum_{t_i \in D} K([\sigma_\omega(\phi_i(s_{i,1})), \ldots, \sigma_\omega(\phi_i(s_{i,n}))], [r^h_{i,1}, \ldots, r^h_{i,n}]) \quad (5)$$

We can approach this problem either with a *learning-to-rank* or with a *regression* framework. Learning-to-rank seems well suited because it captures the fact that we are interested in ranking summaries, however we selected the regression approach in order to keep the model simple. It solves a different but closely related problem:

$$\underset{\omega}{\text{argmax}} \sum_{t_i \in D} \sum_{j}^{n} \frac{\|\sigma_\omega(\phi_i(s_{i,j})) - r^h_{i,j}\|^2}{2} \quad (6)$$

The regression finds the parameters predicting the scores closest to the ones given by humans. We use an off-the-shelf implementation of Support Vector Regression (SVR) from scikit-learn (Pedregosa et al., 2011).

4 Experiments

We conducted both automatic and manual testing of the learned metric. We present here the datasets and results of the experiments.

4.1 Datasets

We use two multi-document summarization datasets from the Text Analysis Conference (TAC) shared tasks: 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. 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.

With our notations, for example with TAC-2009, we have $n = 55$ scored system summaries, $m = 44$ topics, $\mathcal{D}_i$ contains 10 documents and θ_i contains 4 reference summaries.

We also use the recently created German dataset DBS-corpus (Benikova et al., 2016). It contains 10 topics consisting of 4 to 14 documents each. The summaries have variable sizes and are about 500 words long. For each topic, 5 summaries were evaluated by trained human annotators but only for content selection with Pyramid.

We experiment with this dataset because it contains heterogeneous sources (different text types) in German about the educational domain. This contrasts with the English homogeneous news documents from TAC-2008 and TAC-2009. Thus, we can test our technique in a different summarization setup.

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

4.2 Correlation Analysis

Baselines Each feature presented earlier is evaluated individually. [2] Indeed, they all produce scores for summaries meaning we can measure their correlation with human judgments. Classical evaluation metrics, like ROUGE-N variants, are therefore also included in this analysis and serve as baselines. Identifying which metrics have high correlation with human judgments constitutes an initial feature analysis.

Most of the features do not need language dependent information, except those requiring word embeddings or frame identification based on a frame inventory. We do not include the frame identification features when experimenting with the German DBS-corpus. However, for the other language dependent features, we used the German word embeddings developed by Reimers et al. (2014). For the English datasets, we use dependency-based word embeddings (Levy and Goldberg, 2014).

The performances of the baselines on TAC-2008 and TAC-2009 are displayed in Table 1, and Table 2 depicts scores for the DBS-corpus. In order to have an insightful view, we report the scores for the three correlation metrics presented in the previous section: Pearson's r, Spearman's ρ and Ndcg.

Feature Analysis There are fewer scored summaries per topic in the DBS-corpus (5 compared to 55 in TAC-2008). Shorter ranked lists generally have higher scores which explains the overall higher correlation scores in the DBS-corpus. It also contains longer summaries (500 words compared to 100 words for TAC) which provides a reason behind the better performances of JS features. Indeed, word frequency distributions are more representative for longer texts.

First, we see that classical evaluation metrics like ROUGE-N have lower correlation when computed at the summary-level. Here the correlations are around 0.60 spearman's ρ while they often surpass 0.90 in the system-level scenario (Lin, 2004).

However, the experiments confirm that ROUGE-N, especially ROUGE-2, are strong when compared to other available metrics. Even the more semantically motivated metrics like ROUGE-N-WE or Frame-N (ROUGE-N enriched with frame annotations) can not outperform the simple ROUGE-N. The added semantic information might be too noisy to really give improvements. Simple lexical comparison still seems to be better for evaluation of summaries.

Interestingly, it is the other simple evaluation metric $JS_{ref} - N$ which competes with ROUGE-N. This metric only compares the distribution of n-grams in the reference summaries with the distribution of n-grams in the candidate summary and it outperforms ROUGE-N for pearson's r. However, ROUGE-N still outperforms $JS_{ref} - N$ for Ndcg. It indicates that this metric can be complementary with ROUGE-N even though it was rarely used for evaluation before.

Finally, we observe that the features not using the reference summaries have poor performances. It is troubling because these are the strategies used by classical summarization systems in order to decide which summary to extract. Overall, they have Ndcg scores higher than 0.5 meaning they can decently identify some of the best summaries explaining why these systems can produce good summaries.

Our Models For each dataset, we trained two models. The first model (S^3_{full} for *Supervised Summarization Scorer*) uses all the available features for training. However, the previous feature analysis revealed that some features are poor. We hypothesized that they might harm the learning process. Therefore we trained a second model S^3_{best} using only 6 of the best features. [3] We normalize human scores so that they every topic has the same mean.

Both models are trained and tested in a leave-one-out cross-validation scenario ensuring proper testing of the approach. The results for TAC-2008 and TAC-2009 are presented in Table 1 while the results for the DBS-corpus are in Table 2. For comparison we also added the correlation between pyramid and responsiveness when both annotations are available.

Model analysis As expected we observe that using the restricted set of non-noisy features gives stronger results. S^3_{best} is the best metric and outperforms the classical ROUGE-N. Thanks to the combination of ROUGE-N and $JS_{ref} - N$, it gets the best of both worlds and has consistent performances accross datasets and correlation measures.

[2] We do not include Red-N in the result table because it does not aim to measure content selection

[3] ROUGE-1, ROUGE-2, ROUGE-WE-1, ROUGE-WE-2, $JS_{ref} - 1$ and $JS_{ref} - 2$

	TAC-2008						TAC-2009					
	responsiveness			Pyramid			responsiveness			Pyramid		
	r	ρ	Ndcg	r	ρ	Ndcg	r	ρ	Ndcg	r	ρ	Ndcg
TF$*$IDF-1	.1760	.2248	.5040	.1833	.2376	.3594	.1874	.2226	.3912	.2423	.2845	.2349
TF$*$IDF-2	.0478	.1540	.5962	.0496	.1827	.4833	.0476	.1674	.5079	.0972	.2337	.3949
Cov-1	.2552	.2635	.6137	.2812	.3035	.5140	.2267	.2212	.5627	.2765	.2871	.4776
Cov-2	.1056	.1878	.6154	.1136	.2287	.5228	.1382	.0787	.5602	.1170	.1336	.4936
KL-1	.1774	.2240	.4922	.1996	.2682	.3470	.1696	.2220	.4139	.2328	.2939	.2568
KL-2	.0042	.1654	.6188	.0038	.1921	.5160	.0602	.1373	.6311	.0355	.2011	.5641
JS-1	.2517	.2771	.4411	.2811	.3214	.2839	.2160	.2352	.3896	.2742	.3119	.2273
JS-2	.0409	.1708	.5874	.0447	.2058	.4804	.0013	.1548	.5646	.0310	.2166	.4734
ROUGE-1	.7035	.5786	.9304	.7479	.6329	**.9125**	.7043	.5657	.8901	.8085	.6922	.9323
ROUGE-2	.6955	.5725	.9333	.7184	.6358	.9064	.7271	.5837	.9039	.8031	.6949	.9272
ROUGE-1-WE	.5714	.4503	.9042	.5798	.4587	.8434	.5865	.4377	.8724	.6534	.5163	.8792
ROUGE-2-WE	.5665	.3971	.8972	.5563	.3888	.8258	.6072	.4130	.8749	.6712	.4811	.8709
ROUGE-L	.6815	.5207	.9300	.7028	.5688	.8937	.7305	.5631	**.9083**	.7799	.6529	.9159
AVG_{SIM}	.1351	.0904	.6890	.0747	.0543	.5521	.2389	.1557	.6861	.2306	.1597	.5956
Frame-1	.6587	.5083	.9174	.6861	.5294	.8867	.6786	.5270	.8827	.7626	.6280	.9158
Frame-2	.6769	.5190	.9194	.6917	.5560	.8885	.7152	.5555	.9000	.7814	.6486	.9191
$JS_{ref}-1$	.6907	.5642	.3786	.7527	.6481	.1862	.7125	.5834	.3091	.8328	.7286	.1214
$JS_{ref}-2$	.6943	.5579	.3961	.7187	.6253	.2101	.7291	.5862	.3195	.8105	.7007	.1342
S^3_{full}	.6960	.5582	.9256	.7537	.6520	.9073	.7310	.5522	.9002	.8384	.7240	.9373
S^3_{best}	**.7154**	**.5954**	**.9330**	**.7545**	**.6527**	.9077	**.7386**	**.5952**	.9015	**.8429**	**.7315**	**.9354**
Pyramid	.7030	.6604	.8528	—	—	—	.7152	.6386	.8520	—	—	—

Table 1: Correlation of automatic metrics with human judgments for TAC-2008 and TAC-2009.

Thanks to the combination of metrics, our model has more consistent performances accross different correlation metrics. It especially benefits from the complementarity of ROUGE and JS_{ref}.

While the improvements are sometimes good, they are not dramatic. A bigger and more diverse training data should give further improvements. With a better training set, it might even not be necessary to manually remove the noisy features as the model will learn when to ignore which features.

4.3 Percentage of failure

By analysing the average correlation between the different metrics and human judgments over all topics, we only get an average overview. It would be useful to estimate the number of topics on which a metric *fails* or *works*. One could plot cumulative distribution graphs where the x-axis is the correlation range (from 0 to 1 in absolute values) and the y-axis indicates the number of topics on which the metric's correlation with humans was above the given x point. However, this would require 460 plots (3 datasets * 20 metrics * 6 correlations measures) which would not be readable.

Instead, we define a threshold for each correlation measure and count the percentage of topics for which the metric's correlation with humans was below the threshold. The threshold value is

	Pyramid		
	r	ρ	Ndcg
TF$*$IDF-1	.2902	.2016	.8077
TF$*$IDF-2	.2903	.2396	.8181
Cov-1	.0997	.0544	.8891
Cov-2	.0991	.0638	.8965
KL-1	.7299	.6992	.7348
KL-2	.3089	.1967	.8316
JS-1	.2909	.1680	.8324
JS-2	.1531	.1385	.8496
ROUGE-1	.7016	.7412	.9841
ROUGE-2	.8272	**.8892**	.9985
ROUGE-1-WE	.6842	.7140	.9782
ROUGE-2-WE	.7643	.7937	.9914
ROUGE-L	.7908	.8268	.9957
AVG_{SIM}	.7844	.8309	.9924
$JS_{ref}-1$	**.9712**	.8732	.6881
$JS_{ref}-2$	.9689	.8793	.6879
S^3_{full}	.9077	.8781	.9988
S^3_{best}	.9483	.8755	**.9988**

Table 2: Correlation of automatic metrics with human judgments for the DBS-corpus.

	TAC-2008						TAC-2009					
	responsiveness			Pyramid			responsiveness			Pyramid		
	r	ρ	Ndcg	r	ρ	Ndcg	r	ρ	Ndcg	r	ρ	Ndcg
ROUGE-1	**.2500**	.3958	**.0208**	.1250	.3125	**.1250**	.2727	.4318	**.2272**	.0455	.1364	.0223
ROUGE-2	.3125	.4167	.0208	.2708	**.2292**	.1667	.2500	.3864	**.2272**	.0682	.1591	**.0000**
ROUGE-1-WE	.7083	.7708	.1042	.6875	.6875	.4583	.5455	.7500	.2500	.4318	.5682	.2955
ROUGE-2-WE	.6667	.8333	.1667	.6667	.8333	.6458	.5455	.7727	.2500	.3409	.6364	.3636
$JS_{ref} - 1$	.2917	.4375	1.000	**.1042**	.2917	1.000	**.2045**	.4091	1.000	**.0227**	.1136	1.000
$JS_{ref} - 2$	.3542	.4375	1.000	.2708	.3125	1.000	.2500	.3864	1.000	**.0227**	**.0909**	1.000
S^3_{best}	**.2500**	**.2917**	**.0208**	.1458	.2708	.1458	.2272	**.3409**	**.2272**	**.0227**	.1136	.0227

Table 3: Percentage of topics for which the correlation between the metric and human judgments is below the chosen thresholds for TAC-2008 and TAC-2009.

an indicator of when the metrics fails to correctly model human judgments on a given topic. We chose: 0.65 for pearson's r, 0.55 for spearman's ρ and 0.85 for Ndcg. The values are chosen arbitrarily but in order to get a meaningful picture, if we choose a threshold too low then all metrics are always above, if the threshold is too high all metrics are always below. We report the scores for the set of best features and our best metric S^3_{best} on TAC datasets in Table 3.

We observe that our metric performs well and has low percentage of *failure*. It exhibits again its robustness accross different correlation measures. We also observe the strong performances of the JS_{ref} especially the unigram version, however it fails completely for the Ndcg metrics which indicates that it always has problems to identify the top best summaries even though its overall correlation is good. Again this confirms that our metric benefits from the complementarity of JS_{ref} and ROUGE because ROUGE has performs well with Ndcg.

4.4 Manual annotation

Our models are trained with human judgment datasets constructed during the shared tasks, meaning that only some system summaries and the 4 references summaries have been evaluated by humans. Systems have a limited range of quality as they rarely propose excellent summaries, and bad summaries are usually due to unrelated errors (like empty summaries). This is a concern because our learned metric will certainly perform well in this quality range, but it should also perform well outside of this range. It has to be capable to correctly recognize the new and better summaries that will be proposed by future systems.

As the learning is constrained to a specific quality range, we need to check that the whole scoring spectrum of the metric correlates well with humans. We check that what is considered upper-bound (resp. random) by the metric is also considered as excellent (resp. bad) by humans.

Annotation setup We collect summaries by employing a meta-heuristic solver introduced recently for extractive MDS by Peyrard and Eckle-Kohler (2016). Specifically, we use the tool published with their paper.[4]

Their meta-heuristic solver implements a *Genetic Algorithm* to create and iteratively optimize summaries over time. In this implementation, the individuals of the population are the candidate solutions which are valid extractive summaries. Each summary is represented by a binary vector indicating for each sentence in the source document whether it is included in the summary or not. The size of the population is a hyper-parameter that we set to 100. Two evolutionary operators are applied: the mutation and the reproduction. Mutations happen to several randomly chosen summaries by randomly removing one of its sentences and adding a new one that does not violate the length constraint. The reproduction is performed by randomly extracting a valid summary from the union of sentences of randomly selected parent summaries. Both operators are controlled by hyper-parameters which we set to their default values.

We use our metric S^3_{best} as the fitness function and, after the algorithm converges, the final population is a set of summaries ranging from almost random to almost upper-bound. For 15 topics of TAC-2009, we automatically selected 10 summaries of various quality from the final population and asked two humans to score them following the

[4] https://github.com/UKPLab/coling2016-genetic-swarm-MDS

	Responsiveness		
	r	ρ	Ndcg
Best baseline	.6945	.6701	.9210
S^3_{full}	.7198	.6818	.9323
S^3_{best}	**.7318**	**.6936**	**.9355**

Table 4: Correlation of automatic metrics with human accross the whole scoring spectrum of S^3_{best}.

guidelines used during DUC and TAC for assessing responsiveness. To select the summaries, we ranked them according to their S^3_{best} scores and for a population of 100 we picked 10 evenly spaced summaries (the first, the tenth and so on). We observe an inter-annotator agreement of 0.74 Cohen's κ. The results are displayed in Table 4 where S^3_{best} is compared to the best baseline (ROUGE-2) and S^3_{full}.

The S^3_{best} metric gets consistent correlation scores with human judgments as it had with responsiveness in the previous experiements (on TAC-2009, for responsiveness, S^3_{best} has 0.7386 pearson's r, 0.5952 spearman's ρ and 0.9015 Ndcg) . It is a strong indicator that the metric is reliable even outside of its training domain. It also outperforms ROUGE-2 in this experiment.

5 Discussion

The experiments showed that even semantically motivated metrics struggle to outperform ROUGE-N. However, the simple JS_{ref} and ROUGE-N using only n-gram are the best baselines. Reporting these two metrics together might be more insightful than simply reporting ROUGE-N because they are complementary. Our learned metric is benefiting from this complementarity to achieve its scores.

However, finding a good evaluation metric for summarization is a challenging task which is still not solved. We proposed to tackle this problem by learning the metric to approximate human judgments with a regression framework. A learning-to-rank approach could give stronger results because it might be easier to rank summaries. Even after normalization human scores are noisy and topic-dependent. We expect ranking to be more transferable from one topic to another. Here, we constrained ourselves to a simple approach in order to provide a user-friendly tool and the regression offered a simple and effective solution.

Our experiments revealed that the available human judgment datasets are somehow limited. While it is possible to learn a reliable combination of existing metrics, one would need better and bigger human judgment datasets to really get strong improvements. In particular, it is important to extend the coverage of these datasets because we rely on them to compare evaluation metrics. These annotations are the key to understand what humans consider to be good summaries. Statistical analysis on such datasets will likely be beneficial to develop both evaluation metrics and summarization systems (Peyrard and Eckle-Kohler, 2017).

The metric was evaluated on English news datasets and on a German dataset of heterogeneous sources but a wider study might be needed in order to measure the generalization of the learned metric to other datasets and domains. Such generalization capabilities would be interesting because one would not need to re-train a new metric for every domain.

We believe it is important to develop evaluation metrics correlating well with human judgments at the summary-level. This gives a more insightful and reliable metric. If the metric is reliable enough, one can use it as a target to train supervised summarization systems (Takamura and Okumura, 2010; Sipos et al., 2012) and approach summarization as a principled machine learning task.

6 Conclusion

We presented an approach to learn an automatic evaluation metrics correlating well with human judgments at the summary-level. The metric is a combination of existing automatic scoring strategies learned via regression. We release the metric as an open-source tool. [5] We hope this study will encourage more work on learning evaluation metrics and improving the human judgement datasets. Better human judgment datasets will be greatly beneficial for improving both evaluation metrics and summarization systems.

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).

[5] https://github.com/UKPLab/emnlp-ws-2017-s3

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90. Association for Computational Linguistics.

Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. 2016. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1039 – 1050.

John M. Conroy and Hoa Trang Dang. 2008. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. In *Proceedings of the 22Nd International Conference on Computational Linguistics (COLING)*, volume 1, pages 145–152.

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*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

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*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

Aaron Harnly, Rebecca Passonneau, and Owen Rambow. 2005. Automation of Summary Evaluation by the Pyramid Method. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 226–232, Borovets, Bulgaria.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet Semantic Role Labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 471–482. Association for Computational Linguistics.

Tsutomu Hirao, Manabu Okumura, Norihito Yasuda, and Hideki Isozaki. 2007. Supervised Automatic Evaluation for Summarization with Voted Regression Model. *Information Processing and Management*, 43(6):1521–1535.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 302–308.

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. Association for Computational Linguistics.

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 at NAACL*, pages 463–470, New York City, USA.

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*, pages 557–564, Rome, Italy. Springer-Verlag.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

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).

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montreal, Canada. Association for Computational Linguistics.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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)*, pages 247 – 257, Osaka, Japan. The COLING 2016 Organizing Committee.

Maxime Peyrard and Judith Eckle-Kohler. 2017. A principled framework for evaluating summarizers: Comparing models of summary quality against human judgments. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, volume Volume 2: Short Papers. Association for Computational Linguistics.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. GermEval-2014: Nested Named Entity Recognition with Neural Networks. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 117–120.

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin Learning of Submodular Summarization Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 224–233, Avignon, France. Association for Computational Linguistics.

Josef Steinberger and Karel Ježek. 2012. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Hiroya Takamura and Manabu Okumura. 2010. Learning to Generate Summary as Structured Output. In *Proceedings of the 19th ACM international Conference on Information and Knowledge Management*, pages 1437–1440, Toronto , ON, Canada. Association for Computing Machinery.

Qian Yang, Rebecca Passonneau, and Gerard de Melo. 2016. PEAK: Pyramid Evaluation via Automated Knowledge Extraction. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*, Phoenix, AZ, USA. AAAI Press.