Which Scores to Predict in Sentence Regression for Text Summarization?

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz
Research Training Group AIPHES / Knowledge Engineering Group
Department of Computer Science, Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt, Germany
{zopf@aiphes,eneldo@ke,juffi@ke}.tu-darmstadt.de

Abstract

The task of automatic text summarization is to generate a short text that summarizes the most important information in a given set of documents. Sentence regression is an emerging branch in automatic text summarizations. Its key idea is to estimate the importance of information via learned utility scores for individual sentences. These scores are then used for selecting sentences from the source documents, typically according to a greedy selection strategy. Recently proposed state-of-the-art models learn to predict ROUGE recall scores of individual sentences, which seems reasonable since the final summaries are evaluated according to ROUGE recall. In this paper, we show in extensive experiments that following this intuition leads to suboptimal results and that learning to predict ROUGE precision scores leads to better results. The crucial difference is to aim not at covering as much information as possible but at wasting as little space as possible in every greedy step.

1 Introduction

More and more data is generated in textual form in newspapers, social media platforms, and microblogging services and it has become impossible for humans to read, comprehend, and filter all the available data. Automatic summarization aims at mitigating these problems by *"taking an information source, extracting content from it, and presenting the most important content to the user in a condensed form and in a manner sensitive to the users or applications needs"* (Mani, 2001).

Very prominent in automatic text summarization is the idea of extractive summarization. In extractive summarization, summaries are not generated from scratch. Instead, sentences in the source documents, which are supposed to be summarized, are extracted and concatenated to form a summary. To be able to select sentences in a meaningful manner, it is crucial for the extractive systems to be able to estimate the utility of individual sentences.

Supervised extractive methods are usually modeled in a regression framework. Hence, this subfield of automatic summarization is called *sentence regression*. The predicted scores are used to generate a ranking of the sentences, and a greedy strategy is often used in combination with additional redundancy avoidance to select sentences which will be added to the iteratively generated summary (Carbonell and Goldstein, 1998). Another method for the selection is solving an integer linear programming (ILP) problem (Gillick et al., 2008; Hong and Nenkova, 2014) which is, however, an NP-hard problem (Filatova and Hatzivassiloglou, 2004). Even though it can be argued that the complexity is not an issue since there are good solvers for ILPs, it remains a problem when large document collections with many sentences have to be summarized or the system should be used on a large scale for many users. The greedy approach is due its simplicity and efficiency very appealing.

Crucial for building sentence regression models is the choice of the regressands which has to be predicted by the models. Most of the recent works try to predict ROUGE recall scores of individual sentences, which seems to be an obvious choice since the final summaries are also evaluated with ROUGE recall metrics (Lin, 2004; Owczarzak et al., 2012). We show in this paper that following this intuition leads to suboptimal results. In extensive experiments, we investigate sentence regression models with perfect and noisy prediction of different regressand candidates with and without redundancy avoidance. In all experiments, we observe the very same result: learning to predict ROUGE precision scores of sentences leads to better results than learning to predict ROUGE recall scores if the scores are selected

Proceedings of NAACL-HLT 2018, pages 1782–1791
New Orleans, Louisiana, June 1 - 6, 2018.

with a greedy algorithm afterwards. Our findings are in particular important for automatic summarization research since the best models currently available are sentence regression models trained to predict ROUGE recall scores. We expect that simply replacing ROUGE recall scores as regressand with ROUGE precision scores can potentially improve these state-of-the-art models further.

We note in passing that the problem is reminiscent of defining heuristics in inductive rule learning: Individual rules are typically evaluated according to their consistency (minimizing the amount of false positives) and completeness (maximizing the amount of true positives), which loosely correspond to precision and recall (Fürnkranz and Flach, 2005). Heuristics such as weighted relative accuracy, which give equal importance to both dimensions, are successfully used for evaluating single rules in subgroup discovery (Lavrač et al., 2004), but tend to over-generalize when being used for selecting rules for inclusion into a predictive rule set. The reason for this is that a lack of completeness can be repaired by adding more rules, whereas a lack of consistency can not, so that consistency or precision of individual rules should receive a higher weight in the selection task. Transferred to summarization, this means that space wasted by recall-oriented selection cannot be used anymore whereas a low recall in a partial summary can be repaired by adding more sentences.

In the following, we will first formalize the problem of extractive summarization and outline the greedy selection strategy (Section 2). Previously extractive summarization systems, in particularly sentence regression models, are summarized in Section 3. We then present an intuition why predicting ROUGE precision scores can potentially give better results in Section 4. In extensive experiments (Section 5), we actually show the previously stated hypothesis which says that selecting sentence according to ROUGE precision instead of ROUGE recall leads to better results if sentence are selected greedily.

2 Extractive Summarization

In this section, we will first formally define the problem of extractive summarization and then describe the greedy sentence selection strategy which is used by many prior works.

2.1 Problem Definition

The task in extractive summarization is to generate a list of sentences S (the summary) from given list of input sentences I (the text to summarize). The size of the generated summary S must not be longer than a predefined length l (usually measured in words or characters).

In order to select sentences, both supervised and unsupervised models are used to predict utility scores of sentences in a first phase. In a second phase, sentences are selected and concatenated to build a summary.

For evaluation, the generated summary is typically compared to human written summaries by automatic means, in many cases by computing so-called ROUGE scores (Lin, 2004).

2.2 Greedy Selection Strategy

A popular strategy to select sentences based on the previously predicted utility scores is the greedy sentence selection strategy which is described in Algorithm 1.

Algorithm 1 Greedy Sentence Selection with Redundancy Avoidance in Extractive Summarization

list of all input sentences $I = s_1, \ldots, s_n$
utility function u
desired summary length l
π = permutation of I s.t. $u(s_{\pi(1)}) \geq \cdots \geq u(s_{\pi(n)})$
$S \leftarrow \emptyset, i \leftarrow 1$
while $|S| < l$ **and** $i < n$ **do**
  if $\text{sim}(s_{\pi(i)}, S) < \theta$ **then**
    $S \leftarrow S + s_{\pi(i)}$
  end if
  $i \leftarrow i + 1$
end while
return S

According to the greedy strategy, the sentence with the highest utility score is selected first. After the best sentence has been selected, it is removed from the input list of available sentences, and the former second best sentence is considered next. Redundancy avoidance strategies are used to ensure that sentences with similar contents are not added multiple times to the summary. A simple strategy computes the similarity of the currently best sentence and all already selected sentences. If the maximum similarity exceeds a predefined threshold θ, the summarizer removes the sentences from the input list without adding it to the summary. The selection process is repeated until the desired summary length is reached. Once a decision is made, it is never revised.

3 Sentence Regression for Extractive Summarization

After the field of automatic summarization has been dominated by unsupervised extractive summarization models for some time (Carbonell and Goldstein, 1998; Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Li et al., 2006), supervised regression models are more commonly used in recent years. The crucial difference is that supervised models learn to predict regressands based on training examples in a training phase whereas unsupervised models do not predict regressands. We focus on supervised extractive regression systems in this paper. Comprehensive overviews of automatic summarization (Nenkova and McKeown, 2011; Gambhir and Gupta, 2016; Yao et al., 2017) also cover unsupervised methods in more detail and include abstractive summarization methods which are out of scope for this paper.

Extractive sentence regression can be described as the task of learning regressands for individual sentences from examples. The general learning problem can be formulated as $y_i = u(x_i) + e_i$ where y_i denotes the regressand (also called dependent variable or target variable) of sentence x_i (the regressor, also called independent variable or features), and e_i denotes the ith residuum (also called error). Sentence regression aims at learning the utility function u from observed sentence-utility pairs in order to minimize the errors for unseen sentences-utility pairs.

Kupiec et al. (1995) proposed one of the first supervised summarization systems, which trains a Bayesian model to predict the probability that a sentence will be included in the summary. They criticized that although a large number of different features had been used in previous unsupervised models, no principled method to select or weight the features had been proposed at this time. Instead of generating summaries, the performance of the model was evaluated based on the classification output of the model for individual sentences. Similarly, Conroy and O'leary (2001) use a Hidden Markov Model to predict the probability that a sentence is included in a reference summary.

The model proposed by Li et al. (2006) already predicts utility scores for individual sentences. The model weights are, however, not learned in a supervised training but assigned by humans. Li et al. (2007) extends this previously proposed unsupervised model and used a support vector regression (SVR) model in the DUC 2007 shared task (Over et al., 2007). Both Li et al. (2006) and Li et al. (2007) use a greedy selection strategy. Instead of learning to predict the probability of appearance of a sentence in a summary (Kupiec et al., 1995; Conroy and O'leary, 2001), Li et al. (2007) use the average and maximum text similarity of candidate sentences and reference summaries as regressands. Ouyang et al. (2011) also applied SVR but used the sum of word probabilities as regressand. Their system therefore also tends to select longer sentences similarly to systems which use ROUGE recall.

PriorSum (Cao et al., 2015b) follows Li et al. (2007) and presents a linear regression framework which uses prior and document dependent features. As regressand, ROUGE-2 recall is used.

Cao et al. (2015a) propose a hierarchical regression process which predicts the importance of sentences based on its constituents. ROUGE-1 recall and ROUGE-2 recall are used as regressand for sentences. For sentence selection, they implement both a greedy selection and a selection based on integer linear programming.

The Redundancy-Aware Sentence Regression (Ren et al., 2016) framework models both importance and redundancy jointly. They train a multi-layer perceptron which then predicts relative importance utilities based on ROUGE-2 recall scores.

REGSUM (Hong and Nenkova, 2014) predicts sentence importance based on word importance and additional features. They use a greedy selection strategy with additional redundancy avoidance which only appends sentences to the summary if the maximum cosine similarity to already selected sentences is lower than a fixed threshold.

We summarize that ROUGE recall is often used in the field of sentence regression in combination with a greedy selection and an additional redundancy avoidance strategy. In the following, we first describe the underlying intuition of using ROUGE recall. Second, we describe why using ROUGE precision instead can be potentially better. Later, we show in the experiments that using ROUGE precision is not only theoretically appealing but also works better in practice than ROUGE recall.

4 ROUGE Recall vs. ROUGE Precision

The ROUGE metric (Lin, 2004) is the method of choice for the evaluation of generated summaries in the field of automatic summarization. Its idea is to compute the similarity between automatically generated summaries and references summaries, which are typically provided by humans.

ROUGE can be viewed as an evaluation measure for an information retrieval task in which precision and recall can be measured. Let E be a set of elements, $R \subset E$ the multiset of desired elements in the reference output, $G \subset E$ is the generated output multiset, and $|.|$ the size of a multiset. Then, the recall is defined as

$$r(G, R) = \frac{|G \cap R|}{|R|} \quad (1)$$

and measures how much of the desired content was returned by the system. On the other hand, precision is defined as

$$p(G, R) = \frac{|G \cap R|}{|G|}, \quad (2)$$

and measures how much of the returned content was actually desirable. We define the intersection $\cap$ of two multisets as the smallest multiset S with $\sigma_S(e) = \min(\sigma_G(e), \sigma_R(e)) \ \forall e \in G, R$, where $\sigma_S(e)$ indicates the number of appearances of element e in set S.

In ROUGE-n, the multiset E is defined as the set of all n-grams, the desired reference multiset R contains all n-grams in a reference summary, and the multiset G contains all n-grams in the system summary. We use multisets and not sets since the same n-gram can be contained multiple times in a text.

When ROUGE was first introduced as the evaluation metric for the DUC 2003 shared task (Over et al., 2007), Lin and Hovy (2003) reported that metrics based on ROUGE recall scores have a good agreement with human judgments. A summary with a high ROUGE recall will contain many n-grams which also appear in the reference summaries. Owczarzak et al. (2012) showed that ROUGE-2 recall is the best variant (highest agreement with human judgments) of ROUGE recall if automatically generated summaries have to be evaluated. ROUGE-2 recall is therefore often used to evaluate automatic summarization systems.

Crucial for the use of ROUGE recall is the length limitation of the generated summaries.

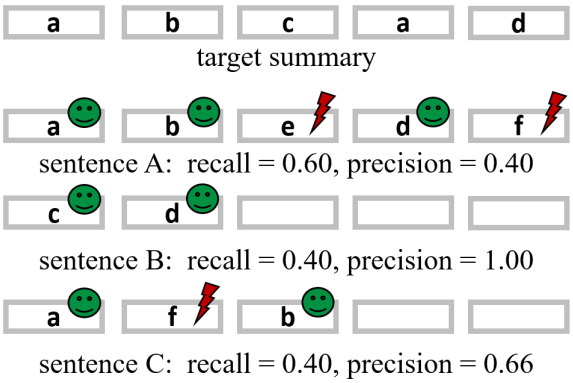

Figure 1: Exemplary illustration of selecting sentences according to precision and recall. The target summary has 5 slots. Sentence A will be selected according to recall since it has a recall scores of 0.6 whereas sentence B and C only have a recall score of 0.4. Sentence A, however, occupies already all available slots in the summary. No more sentence can be selected. Sentence B will be first selected according to precision due to a precision scores of 1.0. After the selection of sentence B, 3 slots are still available in the summary which can be used to fit sentence C to improve the overall summary recall to 0.8.

Usually, the generated summaries are limited to a fixed number of words or characters. Without such a length restriction, systems would be able to generate arbitrary long texts to increase the recall.

Summarization systems aim at maximizing ROUGE recall scores of the generated summaries, since the final summaries are evaluated with ROUGE recall. Greedy extractive summarization approaches try to maximize the overall ROUGE recall of a summary by incrementally adding sentences with a high ROUGE recall to the summary. The idea of this strategy is to pack as much important content as possible into the summary in every step in order to increase the ROUGE recall of the resulting summary. What is usually not considered is the fact that this strategy tends to select longer sentences, since longer sentences tend to have a higher recall. They, however, can contain proportionally more unimportant information, for example in subordinate clauses. As a result, fewer sentences can be selected since the maximum length of the summary is reached earlier.

An alternative strategy, which has not been discussed in the literature so far, is to select sentences according to their ROUGE precision scores. The idea behind this approach is not to cover as much

as information as possible but to waste as little space as possible. Selecting sentences according to precision will not have a bias for longer sentences but for short and dense sentences. Since this strategy tends to selected shorter sentences, more sentence can be included in the summary, which can, in turn, again result in a higher ROUGE recall of the resulting summary.

Figure 1 shows an example in which selecting sentences according to ROUGE precision leads to a higher ROUGE recall score of the resulting summary than selecting sentences according to ROUGE recall. In the following section, we will show that the intuition described in this section is not only appealing in theory, but can also be substantiated in empirical experiments.

We summarize that selecting sentences according to ROUGE precision scores can, intuitively, be better than selecting sentences according to ROUGE recall scores even though the final summaries are always evaluated with ROUGE recall metrics.

5 Experimental Setup

We now present the experimental setups in which we test different regressand candidates for sentence regression in three different, well-known multi-document summarization (MDS) corpora. We used the MDS corpora from the DUC 2004[1], TAC 2008, and TAC 2009[2] summarization shared tasks. All corpora contain 10 input documents and 4 reference summaries for each topic. The number of topics are 50, 46, and 44, respectively. We simulate in the experiments the outcomes of regression models which use different regressands. This will provide us with theoretical insights on which regressand candidates should be considered in regression models and will answer the main question of this paper: *Which scores to predict in sentence regression for text summarization?* For our experiments, we produce summaries containing 665 characters for DUC2004 and summaries containing 100 words for TAC2008 and TAC2009.

5.1 Regressand Candidates

The key ingredient of greedy extractive summarization is the utility function $u(.)$, which is used for sorting the sentences in the first step of Algorithm 1. In this paper, we examine 7 different regressand candidates (**in boldface**) which can be used as regressands when the utility function u is learned via supervised regression.

ROUGE-1 recall (**R1 Rec**) and ROUGE-2 recall (**R2 Rec**) are computed according to Equation 1 for all sentences in the input documents. ROUGE-n recall counts the n-gram overlap of the input sentence and the reference summaries. The more n-grams in the reference documents are covered by a sentence, the higher the score is. These regressands are usually used by prior sentence regression works.

We also compute the ROUGE-1 precision (**R1 Prec**) and ROUGE-2 precision (**R2 Prec**) for all sentences according to Equation 2. A sentence has a high ROUGE-n precision if a high rate of n-grams in the sentence match with n-grams in the reference documents. Sentences with a high density of matching n-grams are therefore preferred by ROUGE precision. The main claim of this paper is that ROGUE precision scores should be primarily considered in sentence regression works instead of ROUGE recall scores. We therefore expect that R1 Prec and R2 Prec will perform better than R1 Rec and R2 Rec.

As a reference point, we compute for each sentence the maximum similarity (**maxADW**) for and the average similarity (**avgADW**) with all sentences in the reference summaries (denoted by list S) according to a state-of-the-art ADW similarity measure (Pilehvar et al., 2013). ADW computes the semantic similarity of two sentences by finding an optimal alignment of word senses contained in the two sentences.

$$\text{maxADW}(s) = \max_{t \in \text{S}} \text{ADWsim}(s, t) \quad (3)$$

$$\text{avgADW}(s) = \frac{1}{|S|} \cdot \sum_{t \in \text{S}} \text{ADWsim}(s, t) \quad (4)$$

Computing the maximum similarity aligns with the idea that a good sentence in the input documents matches well with one sentence in the reference summary. A sentence is representative for the whole summary if it has a high average similarity with all the reference summary sentences. For each sentence, we also randomly generated (**random**) sentence scores which are used as regressand.

[1] http://duc.nist.gov
[2] https://tac.nist.gov

6 Results

6.1 Optimal Prediction without Redundancy Avoidance

In the first experiment, we investigate how helpful the predicted scores are under the assumption that the regressand candidates can be predicted perfectly. The experiment therefore shows how a systems will perform in the optimal case. We do not consider redundancy avoidance strategies in this experiment so that observed performance differences are solely due to differences in the used regressand candidates.

	DUC2004		TAC2008		TAC2009	
	R-1	R-2	R-1	R-2	R-1	R-2
R1 Rec	38.63	08.99	39.28	11.08	34.31	08.37
R2 Rec	39.23	12.07	42.39	16.20	37.42	13.03
R1 Prec	**41.29**	11.18	**43.56**	14.65	**39.45**	12.17
R2 Prec	39.18	**12.73**	43.46	**18.19**	37.81	**13.64**
maxADW	37.60	10.13	42.55	15.46	34.56	11.05
avgADW	38.50	09.62	40.97	12.43	35.48	09.34
random	31.76	04.66	29.58	04.60	29.88	04.63

Table 1: Summarization results in three different multi-document summarization corpora without redundancy avoidance. Columns R-1 and R-2 display the summary quality according to ROUGE-1 recall and ROUGE-2 recall scores, respectively.

The results of the experiment are shown in Table 1. It can be seen that in all corpora the use of ROUGE-1 precision regressands of the sentences leads to better results than using ROUGE-1 recall regressands if ROUGE-1 recall is used as evaluation metric for the final summary. Analogous results can be observed for ROUGE-2 scores. This indicates that using ROUGE recall as regressand in a sentences regression framework is not very promising. Thus, the results are a first confirmation of the previously described intuition that predicting precision scores can be better than predicting recall scores.

Table 2 provides details about the lengths of the produced summaries according to number of stems and number of sentences. The hypothesis that an algorithm that selects sentences according to recall tends to select longer sentences (stated in Section 4) is confirmed. The results therefore also confirm that longer sentences tend to have a higher recall.

In addition to the standard DUC and TAC corpora, we also report results for 2 German datasets, namely the DBS corpus (Benikova et al., 2016) and a subset of the German part of the auto-hMDS corpus (Zopf et al., 2016; Zopf, 2018). The DBS corpus contains topics from the educational domain. auto-hMDS contains heterogeneous topics retrieved from Wikipedia and automatically collected source documents retrieved from web sites. The results are displayed in Table3 and show that the results can be transferred to German. We additionally observe that ROUGE-1 precision seems to be a bit stronger in DBS compared to ROUGE-2 precision even if the resulting summaries are evaluated with ROUGE-2 recall.

	avg. stems			avg. sentences		
	D04	T08	T09	D04	T08	T09
R1 Rec	166	132	141	3.42	2.67	2.70
R2 Rec	160	129	132	4.26	3.46	3.55
R1 Prec	157	125	127	7.76	6.75	6.07
R2 Prec	157	129	126	7.10	6.13	6.09
maxADW	158	127	129	6.56	5.06	5.11
avgADW	158	126	126	5.12	4.13	4.02
random	164	131	131	6.66	5.21	4.89

Table 2: Averaged lengths of resulting summaries measured in number of stems (avg. stems) and number of sentences (avg. sentences). D04 refers to DUC2004 and T08 and T09 refer to TAC2008 and TAC2009, respectively. We count also partially contained sentences which have been cut by the ROUGE length limitation.

	DBS		hMDS	
	R-1	R-2	R-1	R-2
R1 Rec	33.48	13.89	31.94	13.38
R2 Rec	38.67	21.77	40.67	24.39
R1 Prec	**42.20**	**25.55**	**43.25**	23.01
R2 Prec	37.01	23.12	41.65	**24.96**
random	23.27	04.23	20.63	02.36

Table 3: Results as in Table 1, but for 2 datasets (DBS and auto-hMDS) containing German documents.

6.2 Optimal Prediction of F-Scores

The previous experiment clearly showed that selecting sentences according to ROUGE precision outperforms a selection according to ROUGE recall. In this experiment, we will evaluate if a trade-off between recall and precision can lead to even better results. It is, e.g., known that in inductive rule learning, parametrized measures such as the m-estimate, which may be viewed as a trade-off between precision and weighted relative accuracy, can be tuned to outperform its constituent heuristics (Janssen and Fürnkranz, 2010). In retrieval tasks, the F-measure provides a more commonly

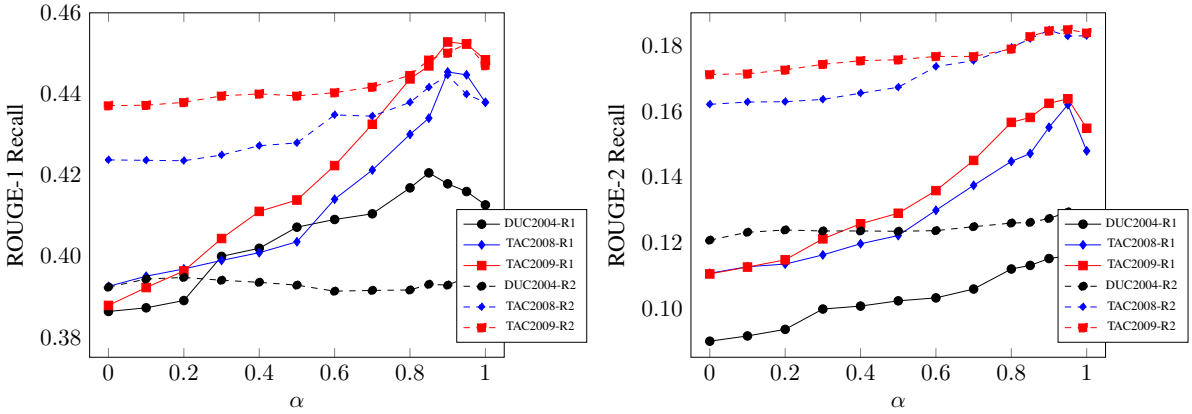

Figure 2: Results of mixing ROUGE-1/2 precision and ROUGE-1/2 recall using $F_\alpha(p, r)$-Measure in different datasets evaluated with ROUGE-1 recall (left) and ROUGE-2 recall (right). For example, the curve labeled DUC2004-R1 shows the results of mixing ROUGE-1 precision and ROUGE-1 recall in the DUC 2004 corpus.

used trade-off between precision and recall, so we chose to use this measure for our experiments. We compute for all sentences the F-measure with $0 \leq \alpha \leq 1$ as

$$F_\alpha(p, r) = \frac{1}{\frac{\alpha}{p} + \frac{1-\alpha}{r}} \quad (5)$$

where a $\alpha = 0$ is equivalent to recall and $\alpha = 1$ equals precision.

The results of the experiment, which are displayed in Figure 2, show that precision ($\alpha = 1.0$) is already close to the optimum but that incorporating also a small fraction of recall ($\alpha \approx 0.9$) leads to the best results which indicates that a slight bias towards longer sentences can improve the result even further. A possible explanation is that there are short sentences in the input documents which are considerably redundant to other high precision sentences. However, overall the trend in the results (increasing evaluation scores with increasing α, which means increasing impact of ROUGE precision) substantiate the general hypothesis of this paper, namely that sentence selection measures should target precision instead of recall.

6.3 Optimal Prediction with Redundancy Avoidance

Summarization systems usually apply a redundancy avoidance strategy in order to avoid including the same information multiple times in the summary. In this experiment, we investigate whether incorporating a simple redundancy avoidance strategy will lead to different results.

During the greedy selection process, we compute the similarity of the currently highest scoring sentence and all already selected sentences (see Algorithm 1, line 4). The highest scoring sentence will be skipped if the maximum similarity of the sentence and the already selected sentences is higher than a predefined threshold θ. We use the state-of-the-art ADW similarity measure to compute the similarities and test the quality of the generated summaries as in the previous experiments with ROUGE-1 and ROUGE-2 recall. The results of the experiment for the thresholds $\theta = 0.4, 0.5, \ldots, 1.0$ are displayed in Figure 3.

We see that sentence selection using ROUGE-1/2 precision scores (red and blue solid lines) consistently leads to better results than with ROUGE-1/2 recall scores (red and blue dashed lines) for all chosen redundancy thresholds. Selecting according to maximum ADW similarity leads to consistently better results than selecting according to the average ADW similarity. This indicates that it is better to search for sentences which align well with a part of the summary than selecting sentences which align relatively well with the whole summary. The best results are achieved with thresholds of $\theta = 0.5$ and $\theta = 0.6$ which worked well for both ROUGE-1 and ROUGE-2 recall in both datasets.

6.4 Noisy Predictions

In the previous experiments, we showed the results of a greedy summarizer which selects sentences according to perfectly predicted scores. Summarization systems are, however, not capable of pre-

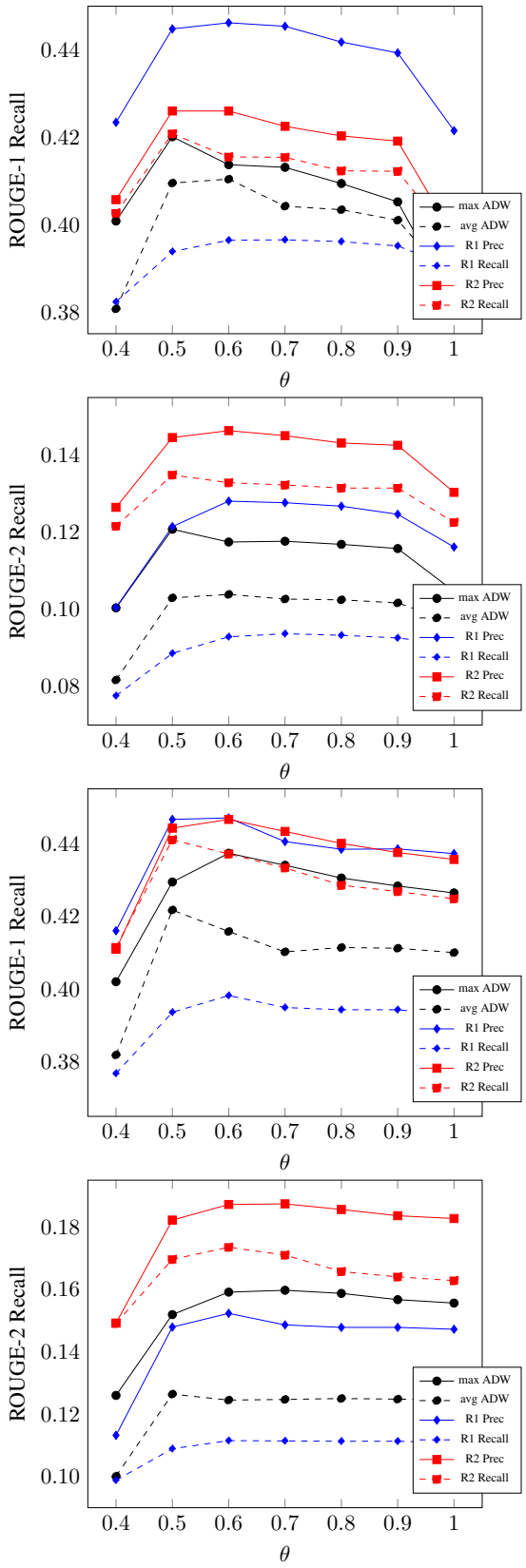

Figure 3: Summary quality assessed with ROUGE-1 recall and ROUGE-2 recall with different redundancy avoidance thresholds θ in the DUC 2004 (top half) and TAC 2008 (bottom half) datasets.

dicting the scores perfectly. We will therefore investigate whether imperfect predictions have an influence on our results in the next experiment. This will also give insights about the robustness of a greedy summarizer in the presence of imprecise predictions.

In order to get model-independent results, we simulate imperfect precisions by adding two different kinds of noise to simulate imperfect predictions, namely additive uniformly distributed continuous noise $\mathcal{U}(a, b)$ and additive Gaussian noise $\mathcal{N}(\mu, \sigma^2)$. For the uniform noise $\mathcal{U}(a, b)$, we test boundaries from $a = -0.2$, $b = 0.2$ to $a = -0.4$, $b = 0.4$. For Gaussian noise, we use mean $\mu = 0$ and variance $\sigma^2 \in \{0.05, 0.1, 0.2\}$. Based on the results in the previous section, we fix the redundancy threshold to 0.6 in this experiment. Due to the random noise, the experiments are no longer deterministic. We therefore run each experiment 10 times and report averaged results.

The results of these experiments (see Table 4) confirm that predicting ROUGE precision is always better than predicting ROUGE recall, in the presence of different kinds of noises and different noise intensities. In case strong Gaussian noise is applied (Table 4, last block), the quality of the

	score	DUC2004		TAC2008		TAC2009	
		R-1	R-2	R-1	R-2	R-1	R-2
$\mathcal{U}(-0.2, 0.2)$	R1 Rec	37.22	07.71	36.73	08.79	37.06	08.99
	R2 Rec	36.93	08.74	36.45	09.91	37.83	11.06
	R1 Prec	**42.53**	10.87	**42.19**	12.57	**43.65**	13.58
	R2 Prec	40.37	**12.04**	40.63	**14.23**	42.25	**15.49**
$\mathcal{U}(-0.3, 0.3)$	R1 Rec	36.78	07.43	35.70	08.00	36.04	08.27
	R2 Rec	35.45	07.54	34.62	08.58	36.08	09.43
	R1 Prec	**42.02**	10.45	**41.42**	11.75	**42.75**	12.83
	R2 Prec	39.56	**11.16**	38.94	**12.64**	40.91	**14.29**
$\mathcal{U}(-0.4, 0.4)$	R1 Rec	36.10	06.92	34.91	07.48	35.85	07.93
	R2 Rec	34.92	07.32	34.08	07.85	3.545	08.70
	R1 Prec	**41.27**	09.98	**40.44**	11.04	**41.63**	11.92
	R2 Prec	39.02	**10.63**	38.22	**11.74**	39.51	**12.97**
$\mathcal{N}(0, 0.05)$	R1 Rec	37.53	07.93	36.99	09.31	37.40	09.36
	R2 Rec	35.46	07.60	35.50	09.41	36.07	09.96
	R1 Prec	**43.55**	11.99	**43.59**	13.98	**45.58**	15.56
	R2 Prec	41.06	**12.92**	42.80	**16.46**	43.97	**17.48**
$\mathcal{N}(0, 0.1)$	R1 Rec	35.63	06.83	34.45	07.31	35.06	07.57
	R2 Rec	33.39	06.04	32.76	06.93	32.88	07.98
	R1 Prec	**41.70**	10.19	**41.41**	12.09	**43.06**	13.23
	R2 Prec	38.41	**10.33**	38.27	**12.43**	40.15	**13.94**
$\mathcal{N}(0, 0.2)$	R1 Rec	33.59	05.72	32.00	05.78	32.36	05.99
	R2 Rec	32.64	05.28	30.76	05.48	31.47	06.01
	R1 Prec	**38.19**	**08.01**	**37.34**	**09.00**	**38.75**	**10.06**
	R2 Prec	35.07	07.45	34.08	08.45	34.71	09.08

Table 4: Summarization results in three different multi-document summarization corpora with noisy score prediction with uniform noise (top) and Gaussian noise (bottom).

summaries decreases more strongly if ROUGE-2 precision scores are predicted, which means that predicting ROUGE-1 precision might be better than predicting ROUGE-2 precision in the case of low prediction quality.

7 Conclusions

Current state-of-the-art sentence regression systems for automatic summarization learn to predict ROUGE recall scores of individual sentences and apply a greedy sentence selection strategy in order to generate summaries. We show in a wide range of experiments that this design choice leads to suboptimal results. In all experiments, we observed the same pattern. The resulting summaries will have a lower quality if ROUGE recall scores for sentences are used instead of ROUGE precision – no matter whether or not redundancy avoidance is considered and whether or not the scores can be predicted perfectly.

In an experiment where we combined both ROUGE recall and ROUGE precision with an F-score computation, we confirmed the previously described observation that the quality of summaries tends to improve with a growing ratio of ROUGE precision vs. ROUGE recall, with a maximum performance for a ratio of $\alpha \approx 0.9$. Biasing the sentence selection slightly to longer sentences is therefore promising. This goes in line with an often applied pre-processing step in which very short sentences are discarded without further analysis (Erkan and Radev, 2004; Cao et al., 2015b).

We also presented an intuition why a selection according to ROUGE precision leads to better results. A system which selects according to ROUGE recall will tend to select longer sentences, since longer sentences tend to have a higher recall. We conclude that systems should instead of fitting iteratively as much as possible into a summary rather aim at wasting as little space as possible in every step.

For future works, it is very simple to incorporate the findings presented in this paper. Instead of learning to predict ROUGE recall scores, the regressand can simply be exchanged and the ROUGE precision can be used instead. Based on the findings in this paper, we expect that the models will benefit from this modification. We furthermore conclude that comparisons between ILP and greedy methods (Cao et al., 2015a) are biased in favor of ILP. A better comparison is possible if precision scores are used as input for greedy systems instead of recall scores.

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.

References

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 22nd International Conference on Computational Linguistics: Technical Papers*. pages 1039–1050.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015a. Ranking with recursive neural networks and its application to multi-document summarization. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* pages 2153–2159. https://doi.org/10.1162/153244303322533223.

Ziqiang Cao, Furu Wei, Sujian Li, Wenjie Li, Ming Zhou, and Houfeng Wang. 2015b. Learning Summary Prior Representation for Extractive Summarization. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* pages 829–833.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR 1998)*. ACM, pages 335–336.

John M. Conroy and Dianne P. O'leary. 2001. Text summarization via hidden Markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. pages 406–407. https://doi.org/10.1145/383952.384042.

Günes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22:457–479.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. A formal model for information selection in multi-sentence extraction. *Proc. International Conference on Computational Linguistics (COLING)* pages 397–403.

Johannes Fürnkranz and Peter A. Flach. 2005. ROC 'n' Rule Learning - Towards a Better Understanding of Covering Algorithms. *Machine Learning* 58(1):39–77. https://doi.org/10.1007/s10994-005-5011-x.

Mahak Gambhir and Vishal Gupta. 2016. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47(1):1–66. https://doi.org/10.1007/s10462-016-9475-9.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tür. 2008. The ICSI Summarization System at TAC 2008. *In Proceedings of the Text Analysis Conf. Workshop* .

Kai Hong and Ani Nenkova. 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pages 712–721.

Frederik Janssen and Johannes Fürnkranz. 2010. On the quest for optimal rule learning heuristics. *Machine Learning* 78(3):343–379. https://doi.org/10.1007/s10994-009-5162-2.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 68–73.

Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. 2004. Subgroup Discovery with CN2-SD. *The Journal of Machine Learning Research* 5:153–188.

Sujian Li, You Ouyang, and Bin Sun. 2006. Peking University at DUC 2006. *Proceedings of Document Understanding Conference 2006* .

Sujian Li, You Ouyang, Wei Wang, and Bin Sun. 2007. Multi-document summarization using support vector regression. *Proceedings of Document Understanding Conference 2007* .

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. pages 25–26.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 71–78.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Publishing Co.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. volume 85, pages 404–411. https://doi.org/10.3115/1219044.1219064.

Ani Nenkova and Kathleen McKeown. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval* 5(3):103–233. https://doi.org/10.1561/1500000015.

You Ouyang, Wenjie Li, Sujian Li, and Qin Lu. 2011. Applying regression models to query-focused multi-document summarization. *Information Processing and Management* 47(2):227–237. https://doi.org/10.1016/j.ipm.2010.03.005.

Paul Over, Hoa Dang, and Donna Harman. 2007. DUC in context. *Information Processing and Management* 43(6):1506–1520. https://doi.org/10.1016/j.ipm.2007.01.019.

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.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* pages 1341–1351.

Pengjie Ren, Furu Wei, and Zhumin Chen. 2016. A Redundancy-Aware Sentence Regression Framework for Extractive Summarization. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 33–43.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems* 53(2):297–336. https://doi.org/10.1007/s10115-017-1042-4.

Markus Zopf. 2018. auto-hMDS: Automatic Construction of a Large Heterogeneous Multi-Document Summarization Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*. page (to appear).

Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler. 2016. The Next Step for Multi-Document Summarization : A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 1535–1545.