Estimating Summary Quality with Pairwise Preferences

Markus Zopf
Research Training Group AIPHES / Knowledge Engineering Group
Department of Computer Science, Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt, Germany
zopf@aiphes.tu-darmstadt.de

Abstract

Automatic evaluation systems in the field of automatic summarization have been relying on the availability of gold standard summaries for over ten years. Gold standard summaries are expensive to obtain and often require the availability of domain experts to achieve high quality. In this paper, we propose an alternative evaluation approach based on pairwise preferences of sentences. In comparison to gold standard summaries, they are simpler and cheaper to obtain. In our experiments, we show that humans are able to provide useful feedback in the form of pairwise preferences. The new framework performs better than the three most popular versions of ROUGE with less expensive human input. We also show that our framework can reuse already available evaluation data and achieve even better results.

1 Introduction

Due to the huge amount of information contained in texts, the task of automatic text summarization (Mani, 2001; Nenkova and McKeown, 2011) is a pressing challenge nowadays and will become even more important in the future. Building summarization systems is, however, not the only challenge in this field. Evaluation of automatically generated summaries is also an active field of research.

Ideally, we would like to ask humans for their opinion about the quality of automatically generated summaries in an extrinsic evaluation (Halteren and Teufel, 2003). Since summaries are generated for humans, they should also be evaluated directly by humans. Unfortunately, manual evaluation cannot be performed at a large scale because of the huge effort which is necessary for evaluation. (Lin, 2004) reported that 3,000 hours of human effort would be required for a simple evaluation of the summaries for the Document Understanding Conference (DUC) 2003, a popular summarization shared task series. This motivates research of automatic evaluation methods for automatic summarization.

ROUGE (Lin, 2004), the current method of choice for evaluating automated text summarization, relies on the availability of gold standard summaries. The gold standard summaries are used to define the optimal output of a summarization system. Writing high-quality summaries, however, requires the availability of expert writers and takes a lot of effort. (Dang, 2005) reported that creating the reference summaries for the DUC 2005 shared task was a difficult endeavor with an effort of five hours to produce each reference summary. Since ROUGE needs at least four reference summaries to become reasonably reliable, the effort sums up to at least 20 hours of annotation effort per topic. For this reason, gold standard summaries are only available for a few, rather small datasets. Also the more accurate (but also even more expensive) Pyramid method (Nenkova and Passonneau, 2004) requires expensive gold standard summaries.

Lack of larger and diverse evaluation corpora limits research in automatic summarization. Furthermore, currently available automatic evaluation methods are viewed with skepticism (Rankel et al., 2013). Proper evaluation is, however, an indispensable ingredient for good research. Computing the similarity between two summaries as in ROUGE is a very difficult task. This seems to be obvious since estimating the similarity between sentences and even words is still an active field of research.

In this work, we present an alternative evaluation framework which does not use gold standard summaries to estimate the quality of summaries. Instead of comparing automatically generated summaries with gold standard summaries,

Proceedings of NAACL-HLT 2018, pages 1687–1696
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

our model is trained with simple and inexpensive pairwise preferences (Thurstone, 1927; Fürnkranz and Hüllermeier, 2010) of sentences. To this end, we provide pairs of sentences from the input document of a summarization task to human annotators and ask which of the two sentences contains more important information. We use here the idea of intrinsic information importance (Hong and Nenkova, 2014; Zopf et al., 2016) which describes that information can be intrinsically important. For example, the information *"Donald Trump won the U.S. presidential election"* is intrinsically important. It is likely that it should also be contained in the generated summary if this information is contained in an input document.

After collecting few preferences, our model uses the preferences to generate a ranking of all sentences according to information importance. Summaries which contain sentences similar the upper part of the ranking are then considered to be better than summaries which contain unimportant sentences from the lower part of the ranking.

Pairwise preferences are an appealing form of annotation, since they are much easier to generate than producing complex gold standard summaries. Not only collecting the annotations is easier, but also using the collected annotations is much simpler. The presented model does not have to solve the difficult task of estimating the similarity between generated and gold standard summaries. Instead, the model uses the ranking to estimate the summary quality.

Figure 1 provides an illustration of the traditional evaluation and our new model. On the left, the input documents are illustrated which should be summarized. In the upper part gold standard summaries are generated by humans and used to estimate the quality of an automatically generated summary. In the lower part, we collect pairwise preferences of sentences and use the preferences for evaluation.

An evaluation on topics from two standard datasets, looking at predicting the relative ratings of automatically generated summaries, shows that our new evaluation model is as good as or better than existing methods, at a much lower annotation cost.

2 Related Work

In this section, we will recapitulate previous work in automated text summarization evaluation, focusing on three important approaches, namely model-free evaluation, ROUGE, and Pyramid. The evaluation methods are ordered according to their annotation requirements from none (model-free evaluation) to high (Pyramid). In addition to the most prominent methods described below, several evaluation models were developed in the Automatically Evaluating Summaries Of Peers (AESOP) shared tasks. The systems in this shared task also considered reference summaries as additional information to evaluate a reference summary and are therefore as expensive as ROUGE in terms of required human annotation. Similarly, Giannakopoulos and Karkaletsis (2013) use machine learning to learn a linear combination of n-gram methods to evaluate summaries. Mackie et al. (2014), Giannakopoulos (2013), and Cohan and Goharian (2016) investigate evaluation for microblog, multilingual, and scientific summarization, respectively. Our evaluation, on contrary, uses newswire datasets since this is the most prominent application domain for automatic summarization. Furthermore, we focus on evaluating the information content of summaries and do not evaluate linguistic quality. This is, for example, captured by Pitler et al. (2004).

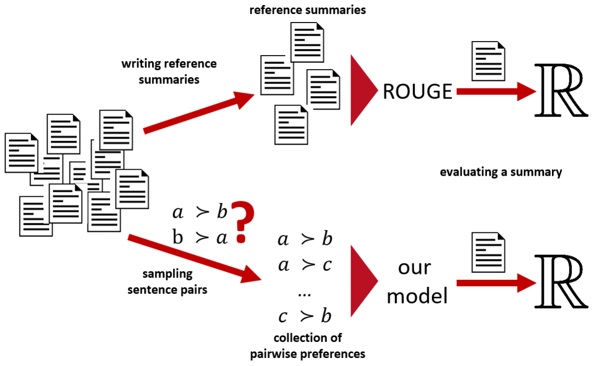

Figure 1: Illustration of traditional evaluation models based on reference summaries (top) and the new model (bottom) which is based on pairwise preferences.

2.1 Model-free Evaluation

Model-free evaluation methods Jensen-Shannon divergence (Louis and Nenkova, 2013) do not require human input such as gold standard summaries and can therefore be applied without additional cost. The quality of model-free evaluation methods is however limited, which is validated in our experiments (see Section 5).

2.2 ROUGE

ROUGE (Lin, 2004) was first used in the Document Understanding Conference (DUC) (Over et al., 2007) and is nowadays the method of choice for automatic evaluation in text summarization. Many popular summarization systems were evaluated with ROUGE (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Gillick et al., 2009; Lin and Bilmes, 2011). It is inspired from the BLEU evaluation method (Papineni et al., 2002) and is based on measuring lexical n-gram overlap of (stemmed) tokens between generated and gold standard summaries. Researchers usually report the n-gram recall of a summary to evaluate the quality of a summary. The quality of ROUGE is often criticized in the research community. Sjöbergh (2007), for example, shows nicely how the ROUGE recall scoring can be fooled easily. A simple greedy language model based on the source documents extracts frequent bi-grams which are likely to occur in the reference summaries. The generated texts are merely lists of bi-grams and not meaningful sentences which cannot be considered to be summaries. However, they achieve superhuman ROUGE recall scores. In the TAC 2008 shared task (Dang and Owczarzak, 2008), both ROUGE-2 and ROUGE-SU4 score automatic systems higher than human summaries, which would lead to the conclusion that these systems are able to produce better summaries than humans. Furthermore, studies show that the correlation between ROUGE scores and human judgments may not be significant in non-newswire genres and other summary types (Liu and Liu, 2008). ROUGE also has many parameters (Graham, 2015), which makes reproduction and comparison of results problematic. Last but not least, ROUGE computes text similarity only based on simple string matching. Expressing the same information with different words is not rewarded by ROUGE. In addition to Graham (2015), Owczarzak et al. (2012) and Rankel et al. (2013) analyze ROUGE in more detail.

Agreement with human judgments (Owczarzak et al., 2012) can be used instead of Pearson's correlation to validate an automatic evaluation model. Measuring agreement allows to obtain a better understanding of the performance of an evaluation model compared to the Pearson correlation. We will also use agreement similarly to Owczarzak et al. (2012) in our experiments.

2.3 Pyramid

The Pyramid method (Nenkova et al., 2007) (similar to (Teufel and Van Halteren, 2004)) was used in the Text Analysis Conference (TAC) (Dang and Owczarzak, 2008) and goes beyond lexical comparisons. It is based on *Summarization Content Units* (SCUs, later also called Summary Content Units). An SCU is a set of lexical expressions with same meaning (e.g. {"2 people passed away", "two persons died"}). After generating the gold standard summaries, SCUs are extracted from these summaries and are weighted by their occurrence frequency in an additional annotation step. Furthermore, every generated summary has to be annotated individually with SCUs before the Pyramid method can be applied. (Nenkova and Passonneau, 2004) have already reported that a large-scale application of the Pyramid method is infeasible. (Over et al., 2007) report a huge effort for the annotation process in the DUC challenges. This additional annotation effort is unattractive for researchers, who prefer automatic methods such as ROUGE. This is validated by the few applications of the Pyramid method until today. The need for more human annotation also introduces an additional source for annotation mistakes. Inspecting the annotations in the TAC 2008 dataset in detail reveals that this is not only a theoretical issue but has practical implications.[1] PEAK (Yang et al., 2016) is an attempt to automate the Pyramid evaluation (similar to (Passonneau et al., 2013)). PEAK also requires reference summaries and is therefore as expensive as ROUGE.

2.4 Qualitative Feedback in Other NLP Tasks

Simple and inexpensive qualitative human feedback has already been used in the field of machine translation (Callison-Burch, 2009; Callison-Burch et al., 2012; Guzmán et al., 2015). (Snow et al., 2008) showed that in a wide variate of NLP tasks, cheap non-expert labels can replace expensive expert annotations. In comparison to our work, we are not asking non-experts to perform the same task as expert annotators (namely writing references summaries) but replace the complex task with a simpler tasks (providing qualitative feedback in form of pairwise preferences).

[1] We found several issues such as not annotating parent SCUs, missing SCUs in sentences, and different annotations for equal sentences.

3 Problem Definition

First, we define T to be the set of all possible texts which can be considered to be summaries. For a given set of source documents D, we define a binary relation $>_D \subset T \times T$ with the intuition that $\mathbf{a} >_D \mathbf{b}$ holds for two texts $\mathbf{a}, \mathbf{b} \in T$ if and only if $\mathbf{a}$ is considered to be the better summary of document collection D than $\mathbf{b}$. Whenever the context is clear, we will omit D and write $\mathbf{a} > \mathbf{b}$ for short. The relation $>$ induces a strict total order (given that ties are not allowed) over T. A text which ranks high according to $>_D$ is a good summary of the document collection D.

How good summaries are is annotated in summarization corpora only for a very small subset of assessed texts $T^+ \subset T$ by human annotators. We use the relation $>^*$ to express in which order summaries are ranked by humans in summarization corpora for each document set D (also called topic or cluster). The relation $>^*$ therefore models the human judgments. For not assessed texts $T^- = T \setminus T^+$ the human judgment is unknown.

The quality of an evaluation method E can be assessed by measuring the agreement with the human judgments. Evaluation models define (implicitly) a ranking $>^E$ by assigning scores to summaries or predicting the ranking directly. Calculating the agreement of the ranking $>^E$ with the human ranking defined by $>^*$ provides a scores which can be used to assess the performance of evaluation models. Measuring the agreement between two relations (which are sets) can be easily done by computing the intersection of both sets:[2]

$$\text{Agreement}(>^*, >^E) = \frac{|>^* \cap >^E|}{|>^*|} \qquad (1)$$

This evaluation of evaluation models is similar to the definition of *Agreement* and *Contradiction* in Owczarzak et al. (2012): *"Agreements occur when the two evaluation metrics make the same distinction between System A and System B (...). Contradictions occur when both metrics find a (...) difference between A and B, but in opposite directions."* A perfect evaluation model, which predicts the preference for all pairs of summaries correctly, will have an agreement of 1 whereas a random evaluation model, which always predicts the preference randomly, will have an expected value of 0.5 according to this measure.

Donald Trump won the election and became president.	$\succ$	*The U.S. Congress certified the results on January 6.*

Figure 2: Example of a pairwise preference annotation of two sentences. The first sentence is preferred over the second sentence because the first sentence contains more important information given that the information is not already known.

We prefer to use the agreement as defined in Equation 1 for evaluation since it can be much better interpreted (Owczarzak et al., 2012). Furthermore, Pearson's correlation is known to be sensitive to outliers, is only able to measure linear correlations, and requires normally distributed, interval scaled residuals (Anscombe, 1973). These properties cannot be assumed to be given when comparing human scores and automated evaluation measures. We therefore prefer to use the agreement according to human judgments as defined in Equation 1 instead of calculating Pearson's correlation.

4 Preference-based Evaluation of Summaries

In this section, we present a novel framework which does not infer a ranking of automatically generated summaries based on gold standard summaries but based on pairwise preferences. The fundamental idea is not to rely on expensive gold standard summaries as previous work does, but to ask annotators for their preferences about sentences. Annotates label pairs of sentences with a preference label which indicates which sentence contains more important information. Figure 2 illustrates such a pairwise preference annotation. A human would likely prefer the first sentence to be included in a summary instead of the second sentence because the first sentence contains, compared to the second sentence, relatively important information. Based on the preferences, our model generates a ranking which reflects the importance of information which is contained in the sentences. Sentences with important information will be ranked high whereas sentences containing only less important information will be ranked low.

[2] We require that an evaluation metric has to make a decision for two summaries if the two summaries are different according to the human judgment. Formally: $\mathbf{a} >^* \mathbf{b} \rightarrow \mathbf{a} >^E \mathbf{b}$ or $\mathbf{b} >^E \mathbf{a}$.

4.1 Sampling Preference Annotations

The easiest strategy to select pairs of sentences for which preferences should be annotated is to sample pairs of sentences randomly and to ask annotators to provide a preference label for each sampled pair (i.e. annotating whether $a \succ b$ or $b \succ a$ for two randomly sampled $a, b \in S^*$). The sentences are sampled from all source document of a topic and are therefore independent from the automatically generated summaries. Our model will therefore not only be able to evaluate already generated summaries but also summaries which will be generated in the future.

All preferences are stored in a matrix M. An entry of n at position M_{i_j} indicates that sentence with index i was preferred n-times over sentence with index j. To reduce the number of annotations, we apply a *smooth propagation of knowledge*. The idea is that we do not only obtain information about the sampled sentence pair but also about pairs which are similar to the sampled pair.

To estimate how much information can be transfered from one to another pair, we calculate the similarities between all sentences. As similarity measure we use the average of the well-known and simple Cosine similarity of TF-IDF vectors and Jaccard similarities. The combination allows to both rely the similarity computation on lexical similarity (Jaccard) and on important content words (Cosine). We define the set of all sentences in the source documents of a topic as S^*. Let (a, b), $a, b \in S^*$ be one annotated sentence pair and $\dot{a}, \dot{b}$ the vector of similarities between a and b and all sentences (i.e. $\dot{a}_i$ denotes the similarity between a and the i-th sentence in S^*).

We define the similarity of the pair (a_1, b_1) and the pair (a_2, b_2) as $\text{sim}(a_1, a_2) * \text{sim}(b_1, b_2)$. If a_1 is the exact same sentence as a_2 and b_1 is similar with a degree of 0.7 to b_2, we will transfer 0.7 of the information from (a_1, b_1) to the pair (a_2, b_2). Transferring information means that we generate additional preferences based on human preferences. If a_1 was preferred over b_1 by a human annotator, we will additionally generate a weighted preferences of with a weight of 0.7 between a_2 and b_2. This can be modeled by the outer product $\dot{a}_1 \otimes \dot{b}_1$ of a_1 and b_1. For each annotated pair (a, b), in which a was preferred by a human over b, we update matrix M by $M \leftarrow M + \dot{a} \otimes \dot{b}$.

4.2 Sentence Score Prediction

The proposed usage of pairwise preferences between sentences is close to the idea of generating a ranking of sports teams by playing individual matches. Instead of competitions between teams, we observe competitions between sentences. The outcome of a match between teams equals to the annotation of a pair of sentences by a human annotator. Since different people can have different opinions about the importance of information (Gambhir and Gupta, 2016), we expect that one sentence will not always be preferred by humans similarly to the situation that the better sports team does not always win against a weaker opponent. This is expressed by the winning probability between teams (or sentences).

In sports, the term *power ranking* is used to describe a ranking which does not only rank the individual teams but also assigns a score to each team, the *skill*. A well-known method to generate a power ranking is the Bradly-Terry (BT) model (Bradley and Terry, 1952). It estimates the utilities $v(a), v(b)$ of two teams (or two sentences) a and b so that the winning probability of a against b equals the score of a divided by the sum of the scores of a and b:

$$p(a \text{ is prefered over } b) = \frac{v(a)}{v(a) + v(b)} \qquad (2)$$

An algorithm to find a maximum-likelihood estimator (MLE) has already been proposed in (Zermelo, 1929). To find the MLE, we iteratively perform Equation 3 for all sentences s_i until the difference between two iterations is sufficiently small.[3] $\text{wins}(s_i)$ denotes the total number of wins of s_i and $\text{duels}(s_i, s_j)$ the number of duels played between sentences s_i and s_j. This information was collected in the previous step and is stored in matrix M.

$$v(s_i) \leftarrow \text{wins}(s_i) \sum_{i \neq j} \frac{\text{duels}(s_i, s_j)}{v(s_i) + v(s_j)} \qquad (3)$$

We normalize the resulting skill vector after each iteration since every multiple of the solution is also a correct solution and therefore restrict the model to converge to one particular solution.

[3] We initialize all scores $v(s_i)$ equally with $v(s_i) \leftarrow \frac{1}{|\mathbf{s}|}$.

4.2.1 Summary Score Prediction

We estimate the score of summary s with function $u : T \rightarrow \mathbb{R}$ as follows:

$$u(\mathbf{s}) = \sum_{i=1}^{|\mathbf{s}|} w_{s_i} \cdot v(\arg\max_{s \in S^*} \text{sim}(s, s_i)) \qquad (4)$$

The utility of a summary is therefore defined as the weighted sum of the sentence utilities v. Since we do not want to restrict our model to purely extractive summaries (which would mean that all sentences contained in the automatic summary have to be exactly contained in the source documents), we estimate the score of a sentence s_i in the summary by searching for the most similar sentence s in the source documents with similarity function $\text{sim} : S \times S \rightarrow [0, 1]$. As weight of s_i, we use $\frac{|s_i|}{|\mathbf{s}|}$ where $|.|$ denotes the length of the summary and sentence measured in number of characters. The intuition of the weight is that a sentence contributes more to the overall score of a summary if it is longer. The score of a summary will decrease if a large fraction of the summary is occupied with a poor sentence. By using a similarity function instead of a hard matching, our method is able to generalize to unseen sentences.

The definition of u in Equation 4 does not consider redundancy. Including a sentence s twice would result in adding the score of s twice to the summary score. This behavior of the evaluation measure is not desirable. We therefore include a redundancy penalization which does not reward redundant information. For a summary $\mathbf{s}$, we reduce the score of sentence s by

$$v_{\text{red}}(s) = v(s) \frac{1}{|s|} \sum_{g \in s} \frac{\text{num}(g, s)}{\text{num}(g, \mathbf{s})} \qquad (5)$$

where $\text{num}(g, s)$ and $\text{num}(g, \mathbf{s})$ denote the number of occurrences of the bi-gram g in s and $\mathbf{s}$, respectively. $|s|$ denotes the number of bi-grams in s.

4.3 Reusing Available Annotation Information

For already existing summarization corpora, reference summaries and/or Pyramid annotations have already been created. Instead of generating new preference annotations by asking human annotators, we can also reuse the available data to simulate annotations. To this end, we define functions w_{r} and w_{p} which estimate the score of a single sentence based on reference summaries (r) and Pyramid scores (p), respectively. We will use the scores generated by w_{r} and w_{p} to simulate annotations of sentence pairs. For two sentence a, b we can simulate a human preference annotation of $a \succ b$ if $w_{\text{r}}(a) > w_{\text{r}}(b)$ and a win of b over a otherwise (equivalent for w_{p}).

For a set of gold standard summaries R, we define $w_{\text{r}} : S \rightarrow \mathbb{R}$ simply to be the maximum similarity to the sentences in the gold standard summaries:

$$w_r(s) = \max_{t \in \mathbf{r}, \mathbf{r} \in R} (\text{sim}(s, t)) \qquad (6)$$

If a very similar sentence appears in a gold standard summary, s will receive a high score. If no similar sentences are in the gold standard summaries the sentence will receive a low score.

Given that Pyramid annotations are available (as in the TAC 2009 corpus, for example), we can define the score of a sentence as the sum of the weights of the matched unique SCUs (similar to the Pyramid method). Annotations are, unfortunately, only available for all sentences in the documents in T^+ and not for sentences in S^*. We therefore search for sentence s in S^* for the most similar sentence $\hat{s}$ in the documents in T^+

$$\hat{s} = \arg\max_{t \in \mathbf{t}, \mathbf{t} \in T^+} \text{sim}(s, t) \qquad (7)$$

and set the score of s to

$$w_p(s) = \sum_{scu \in \hat{s}} \text{weight}(scu) \qquad (8)$$

where $scu \in t$ are all unique SCUs contained in t and $\text{weight}(scu)$ denotes the weight of an SCU as defined in (Nenkova and Passonneau, 2004). As described above, we will observe wins and losses between pairs based on the estimated scores.

5 Experiments

We provide in this section a detailed analysis of our proposed evaluation method. For the experiment, we use eight topics from two popular multi-document summarization datasets, the DUC 2004 (DUC04) and TAC 2009 (TAC09) corpora, which are freely available upon request.[4] Each topic in the datasets contains ten source documents. Each topic contains automatically generated summaries which were generated in the DUC 2004 and TAC 2009 shared tasks. All automatically

[4] http://duc.nist.gov and https://tac.nist.gov

	JS	R1	R2	R3	R4	SU4	man
DUC04	0.480	0.651	0.639	0.649	0.606	0.558	**0.673**
TAC09	0.565	0.638	0.668	0.660	0.674	0.663	**0.688**

Table 1: Agreement of preference based evaluation as defined in Equation 1 of different versions of Jensen-Shannon, ROUGE and our novel model based on manually labeled pairwise preferences.

generated summaries were evaluated by humans. Each summary was labeled with a score from 1 to 5 (DUC04) or 1 to 10 (TAC09) indicating the information content of the summary. Evaluation of grammatically, writing style, etc. is not included in the scores. An evaluation model predicts the preference for two selected summaries correctly if the model predicts the same preference according to the annotated reference scores and incorrectly otherwise. We do not consider ties in the experiments. In the following, we report the agreement as described in Equation 1 for various experiments. We use the abbreviations **JS** (Jensen-Shannon), **R1** - R4 (ROUGE-1 - ROUGE-4), **SU4** (ROUGE-SU4), and **PY** (Pyramid (Nenkova and Passonneau, 2004)) to denote the reference systems.

5.1 Manual Annotations

In the first experiment, we investigate whether humans are able to provide useful feedback in the form of pairwise preferences.

To evaluate our model, we annotated 200 randomly sampled sentence pairs for the first four topics in the DUC04 and the first four topics in the TAC09 corpus with pairwise preferences. The preferences were used (including the previously described smoothed sampling) as input for the proposed model. The results are shown in Table 1. Column **man** denotes the performance of our now model and column **Time (min)** indicates how much time was required to generate the annotations. This information is in particular important for this paper since our main aim is to develop a cheap evaluation framework. In average, our model achieves an agreement of 0.673 in DUC04 and 0.688 in TAC09. This means, that 67.3/68.8 percent of all pairs of manually rated summaries were predicted correctly. This outperforms the best versions of ROUGE in the respective corpora (SU4 with 65.1 percent in DUC04 and R2 with 66.0 percent in TAC09).

With an average annotation time per topic of 53 and 54 minutes our model needs much less annotation effort than ROUGE.

	R1	R2	R4	PY	man	man	man
					+ref	+py	+ref
							+py
DUC04	0.651	0.639	0.606	n/a	**0.722**	n/a	n/a
TAC09	0.638	0.668	0.674	0.715	0.682	0.707	**0.717**

Table 2: Agreement of different versions of ROUGE and Pyramid (PY) and our novel models based on human and automatically generated pairwise preferences in addition to manually labeled preferences.

5.2 Weak Supervision with Additionally Simulated Annotations

In the next experiment, we are interested whether we can simulate additional annotations based on already available reference summaries and Pyramid annotations. The automatically annotated pairs can be considered to be an additional weak supervision for the model. We simulated 200 additional annotations based on reference summaries and/or Pyramid annotations in addition to the 200 manual annotations per topic. To this end, we randomly sampled 200 additional pairs and annotated the pairs with a preference label based on reference summaries and/or Pyramid annotations. Table 2, column **man+ref** contains the results for 200 manual + 200 simulated reference summary-based annotations; column **man+py** contains the results for 200 manual + 200 simulated Pyramid score-based annotations; and column **man+ref+py** contains results for 200 manual + 200 reference summary-based + 200 Pyramid score-based annotations. The results show that we can improve the agreement with additional simulated annotations based on reference summaries in DUC04 by 5 percent points. Additional annotations increased Agreement in TAC09 by 3 percent points. This leads to the conclusion that we can use already available reference summaries in order to substitute more human preference annotations, which makes the trade-off between performance and annotations effort of our model even better.

5.3 Solely Using Simulated Annotations

Now, we investigate if simulated preferences are already sufficient to produce reasonable good results. Table 3, columns **ref** and **py** contain the results of an experiment where we sampled 1,000 simulated pairwise annotations. Without any additional annotation effort, the new model is able to

	R1	R2	R4	PY	ref	py
DUC04	0.651	0.639	0.606	n/a	**0.716**	n/a
TAC09	0.638	0.668	0.674	**0.715**	0.644	0.709

Table 3: Agreement with human judgments for reference systems and our model fed with only automatically generated preferences labels.

perform much better than ROUGE at DUC 2004. In TAC 2009, our model achieves similar performance as the best performing evaluation based on Pyramid annotations. We conclude that automatically generating pairwise preferences based on already available reference summaries is already sufficient to outperform ROUGE. Pairwise preferences generated based on the more expensive Pyramid annotations do not improve the performance.

5.4 Convergence

In the next experiment, we investigate how agreement changes with an increasing amount of annotations. Figure 3 shows how agreement improves with more annotations. We sampled n annotations (horizontal axis) randomly from the human annotations and averaged the resulting agreement scores (vertical axis) of 100 runs to obtain reliable results. We observe a continuous improvement of agreement in all four topics in the TAC 2009 dataset which indicates that sampling more annotations can further improve the performance of our system.

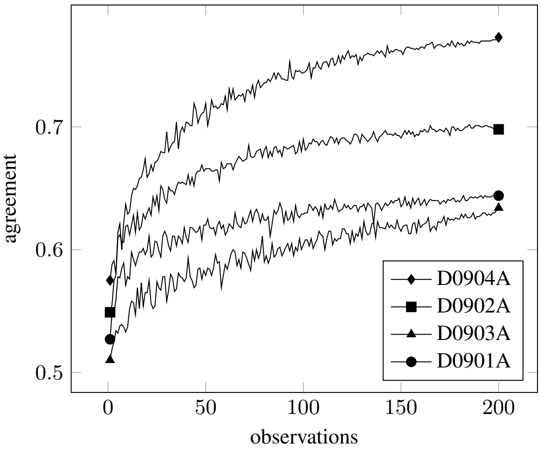

Figure 3: Agreement trajectories averaged over 100 runs per topic in the TAC 2009 corpus.

5.5 Ranking Evaluation

We now investigate the ranking generated by our model directly. Since individual sentences are annotated in the TAC 2009 corpus with SCUs, we can generate a ranking of the sentences and directly compare this ranking with the ranking generated by our model. Table 4 shows the percentage of correctly ordered sentence pairs (similar to Kendall's τ) for our model without and with smoothed sampling.

non-smoothed			smoothed		
man	pyr	ref	man	pyr	ref
0.683	0.987	0.661	0.727	0.941	0.698

Table 4: Percentage of correctly ordered sentence pairs in the TAC 2009 corpus for both a non-smoothed and a smoothed sampling.

Smoothed sampling improves the raking of the model if we use 200 manual or 200 reference summary-based preferences in the TAC 2009 corpus. Given that we can sample pairs based on Pyramid scores, the model is able to reconstruct the ranking almost perfectly if we do not use smoothed sampling. With smoothed sampling, the performance decreases in this case. The result confirms the previously observed performance at summary scoring where preferences based on Pyramid annotations performed best followed by manually generated preference annotations.

6 Conclusions & Outook

Evaluating automatically generated summaries is a challenging task and creating annotations which are required by applications such as ROUGE or Pyramid is laborious and expensive. We presented an alternative model which does not rely on reference summaries or Pyramid annotations but only on simple pairwise preferences of sentences.

We showed in our experiments that the proposed model is able to perform better than the current state-of-the-art ROUGE method with less expensive annotations and that humans are able to provide useful feedback in the form of pairwise preferences. In combination with already available references summaries and Pyramid annotations, we were able to simulate more annotations, which improved performance further.

We conclude that gold standard summaries are not the only usable human feedback which can be used for summary evaluation. Investigating other kinds of feedback such as pairwise preferences might be a promising future research direction.

In future work, we would like to investigate whether we can use crowd-sourcing platforms to collect pairwise preferences on a large scale. Furthermore, we want to investigate whether we can reduce the number of required preferences with smarter sampling methods. Active learning methods can be used to replace the simple random sampling strategy. Additionally, the investigation of more sophisticated similarity functions can potentially improve the model's performance.

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

F. J. Anscombe. 1973. Graphs in Statistical Analysis. *The American Statistician* 27(1):17–21.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs. *Biometrika* 39(3):324–345.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* 1(August):286–295. https://doi.org/10.3115/1699510.1699548.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation* pages 10–51. https://doi.org/10.3115/1626431.1626433.

Arman Cohan and Nazli Goharian. 2016. Revisiting Summarization Evaluation for Scientific Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. pages 806–813.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*. https://doi.org/10.3115/1654679.1654689.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of the First Text Analysis Conference*.

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.

Johannes Fürnkranz and Eyke Hüllermeier, editors. 2010. *Preference Learning*. Springer.

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.

G. Giannakopoulos and V. Karkaletsis. 2013. Summary Evaluation: Together We Stand NPowER-ed. *14th International Conference on Computational Linguistics and Intelligent Text Processing* pages 436–450.

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing Workshop. *MultiLing 2013* page 20.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tür, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The ICSI/UTD Summarization System at TAC 2009. In *Proceedings of the Second Text Analysis Conference*.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 128–137.

Francisco Guzmán, Shafiq Joty, Llu\'is Màrquez, and Preslav Nakov. 2015. Pairwise Neural Machine Translation Evaluation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* pages 805–814. http://www.aclweb.org/anthology/P15-1078.

Hans Van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. *Proceedings of the HLT-NAACL2003 Workshop on Text Summarization* pages 57–64. https://doi.org/10.3115/1119467.1119475.

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.

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

Hui Lin and Jeff Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. pages 510–520.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries. In *Proceedings of the 46th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies. pages 201–204.

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

Stuart Mackie, Richard Mccreadie, Craig Macdonald, and Iadh Ounis. 2014. On Choosing an Effective Automatic Evaluation Metric for Microblog Summarisation. *Proceedings of the 5th Information Interaction in Context Symposium* pages 115–124. https://doi.org/10.1145/2637002.2637017.

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.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. pages 145–152.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid Method. In *ACM Transactions on Speech and Language Processing*. volume 4, pages 1–23. https://doi.org/10.1145/1233912.1233913.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. July, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Rebecca J. Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated Pyramid Scoring of Summaries using Distributional Semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pages 143–147.

Emily Pitler, Annie Louis, and Ani Nenkova. 2004. Automatic Evaluation of Linguistic Quality in Multi-Document Summarization. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* pages 544–554. http://www.aclweb.org/anthology/P10-1056.

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.

Jonas Sjöbergh. 2007. Older versions of the ROUGEeval summarization evaluation system were easier to fool. *Information Processing and Management* 43(6):1500–1505. https://doi.org/10.1016/j.ipm.2007.01.014.

Rion Snow, Brendan O Connor, Daniel Jurafsky, Andrew Y Ng, Dolores Labs, and Capp St. 2008. Cheap and fast - but is it good? Evaluation non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* pages 254–263. https://doi.org/10.1.1.142.8286.

Simone Teufel and Hans Van Halteren. 2004. Evaluating Information Content by Factoid Analysis: Human annotation and stability. In *Proceedings of the 2004 Conference on Empirical Methods on Natural Language Processing*. pages 419–426.

Louis Leon Thurstone. 1927. A law of comparative judgement. *Psychological Review* 34:278–286.

Qian Yang, Rebecca J. Passonneau, and Gerard de Melo. 2016. PEAK: Pyramid Evaluation via Automated Knowledge Extraction. In *Proceedings of the 30th Conference on Artificial Intelligence*. pages 2673–2679.

Ernst Zermelo. 1929. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 29:436–460.

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Beyond Centrality and Structural Features: Learning Information Importance for Text Summarization. In *Proceedings of the 20th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 84–94.