The Heterogeneity Principle in Evaluation Measures for Automatic Summarization*

Enrique Amigó Julio Gonzalo Felisa Verdejo
UNED, Madrid
{enrique,julio,felisa}@lsi.uned.es

Abstract

The development of summarization systems requires reliable similarity (evaluation) measures that compare system outputs with human references. A reliable measure should have correspondence with human judgements. However, the reliability of measures depends on the test collection in which the measure is meta-evaluated; for this reason, it has not yet been possible to reliably establish which are the best evaluation measures for automatic summarization. In this paper, we propose an unsupervised method called Heterogeneity-Based Ranking (HBR) that combines summarization evaluation measures without requiring human assessments. Our empirical results indicate that HBR achieves a similar correspondence with human assessments than the best single measure for every observed corpus. In addition, HBR results are more robust across topics than single measures.

1 Introduction

In general, automatic evaluation metrics for summarization are similarity measures that compare system outputs with human references. The typical development cycle of a summarization system begins with selecting the most predictive metric. For this, evaluation metrics are compared to each other in terms of correlation with human judgements. The second step consists of tuning the summarization system (typically in several iterations) in order to maximize the scores according to the selected evaluation measure.

There is a wide set of available measures beyond the standard ROUGE: for instance, those comparing basic linguistic elements (Hovy et al., 2005), dependency triples (Owczarzak, 2009) or convolution kernels (Hirao et al., 2005) which reported some reliability improvement with respect to ROUGE in terms of correlation with human judgements. However, in practice ROUGE is still the preferred metric of choice. The main reason is that the superiority of a measure with respect to other is not easy to demonstrate: the variability of results across corpora, reference judgements (Pyramid vs responsiveness) and correlation criteria (system vs. summary level) is substantial. In the absence of a clear quality criterion, the de-facto standard is usually the most reasonable choice.

In this paper we rethink the development cycle of summarization systems. Given that the best measure changes across evaluation scenarios, we propose using multiple automatic evaluation measures, together with an unsupervised method to combine measures called *Heterogeneity Based Ranking* (HBR). This method is grounded on the general Heterogeneity property proposed in (Amigó et al., 2011), which states that the more a measure set is heterogeneous, the more a score increase according to all the measures simultaneously is reliable. In brief, the HBR method consists of computing the heterogeneity of measures for which a

*This work has been partially funded by the Madrid government, grant MA2VICMR (S-2009/TIC- 1542), the Spanish Government, grant Holopedia (TIN2010-21128-C02-01) and the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreement nr. 288024 (LiMoSINe project).

Proceedings of the Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pages 36–43,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

system-produced summary improves each of the rest of summaries in comparison.

Our empirical results indicate that HBR achieves a similar correspondence with human assessments than the best single measure for every observed corpus. In addition, HBR results are more robust across topics than single measures.

2 Definitions

We consider here the definition of similarity measure proposed in (Amigó et al., 2011):

Being Ω the universe of system outputs (summaries) s and gold-standards (human references) g, we assume that a similarity measure is a function $x : \Omega^2 \longrightarrow \Re$ such that there exists a decomposition function $f : \Omega \longrightarrow \{e_1..e_n\}$ (e.g., words or other linguistic units or relationships) satisfying the following constraints; (i) maximum similarity is achieved only when the summary decomposition resembles exactly the gold standard; (ii) adding one element from the gold standard increases the similarity; and (iii) removing one element that does not appear in the gold standard also increases the similarity. Formally:

$$f(s) = f(g) \Longleftrightarrow x(s,g) = 1$$

$$(f(s) = f(s') \cup \{e_g \in f(g) \setminus f(s)\}) \Longrightarrow \\ x(s,g) > x(s',g)$$

$$(f(s) = f(s') - \{e_{\neg g} \in f(s) \setminus f(g)\}) \Longrightarrow \\ x(s,g) > x(s',g)$$

This definition excludes random functions, or the inverse of any similarity function (e.g. $\frac{1}{f(s)}$). It covers, however, any overlapping or precision/recall measure over words, n-grams, syntactic structures or any kind of semantic unit. In the rest of the paper, given that the gold standard g in summary evaluation is usually fixed, we will simplify the notation saying that $x(s,g) \equiv x(s)$.

We consider also the definition of *heterogeneity* of a measure set proposed in (Amigó et al., 2011):

The heterogeneity $H(\mathcal{X})$ of a set of measures $\mathcal{X}$ is defined as, given two summaries s and s' such that $g \neq s \neq s' \neq g$ (g is the reference text), the probability that there exists two measures that contradict each other.

$$H(\mathcal{X}) \equiv$$

$$P_{s,s' \neq g}(\exists x, x' \in \mathcal{X} / x(s) > x(s') \wedge x'(s) < x'(s'))$$

3 Proposal

The proposal in this paper is grounded on the heterogeneity property of evaluation measures introduced in (Amigó et al., 2011). This property establishes a relationship between heterogeneity and reliability of measures. However, this work does not provide any method to evaluate and rank summaries given a set of available automatic evaluation measures. We now reformulate the heterogeneity property in order to define a method to combine measures and rank systems.

3.1 Heterogeneity Property Reformulation

The heterogeneity property of evaluation measures introduced in (Amigó et al., 2011) states that, assuming that measures are based on similarity to human references, the real quality difference between two texts is lower bounded by the heterogeneity of measures that corroborate the quality increase. We reformulate this property in the following way:

Given a set of automatic evaluation measures based on similarity to human references, the probability of a quality increase in summaries is correlated with the heterogeneity of the set of measures that corroborate this increase:

$$P(Q(s) \geq Q(s')) \sim H(\{x | x(s) \geq x(s')\})$$

where $Q(s)$ is the quality of the summary s according to human assessments. In addition, the probability is maximal if the heterogeneity is maximal:

$$H(\{x | x(s) \geq x(s')\}) = 1 \Rightarrow P(Q(s) \geq Q(s')) = 1$$

The first part is derived from the fact that increasing heterogeneity requires additional diverse measures corroborating the similarity increase ($H(\{x | x(s) \geq x(s')\})$)). The correlation is the result of assuming that a similarity increase according to any aspect is always a positive evidence of true similarity to human references. In other words, a positive match between the automatic summary and the human references, according to any feature, should never be a negative evidence of quality.

As for the second part, if the heterogeneity of a measure set X is maximal, then the condition of the heterogeneity definition ($\exists x, x' \in \mathcal{X}.x(s) > x(s') \wedge x'(s) < x'(s')$) holds for any pair of summaries that are different from the human references. Given that all measures in X corroborate the similarity increase ($X = \{x | x(s) \geq x(s')\}$), the heterogeneity condition does not hold. Then, at least one of the evaluated summaries is not different from the human reference and we can ensure that $P(Q(s) \geq Q(s')) = 1$.

3.2 The Heterogeneity Based Ranking

The main goal in summarization evaluation is ranking systems according to their quality. This can be seen as estimating, for each system-produced summary s, the average probability of being "better" than other summaries:

$$\text{Rank}(s) = \text{Avg}_{s'}(P(Q(s) \geq Q(s')))$$

Applying the reformulated heterogeneity property we can estimate this as:

$$\text{HBR}_{\mathcal{X}}(s) = \text{Avg}_{s'}(H(\{x | x(s) \geq x(s')\}))$$

We refer to this ranking function as the *Heterogeneity Based Ranking* (HBR). It satisfies three crucial properties for a measure combining function. Note that, assuming that any similarity measure over human references represents a positive evidence of quality, the measure combining function must be at least robust with respect to redundant or random measures:

1. HBR is independent from measure scales and it does not require relative weighting schemes between measures. Formally, being f any strict growing function:

$$\text{HBR}_{x_1..x_n}(s) = \text{HBR}_{x_1..f(x_n)}(s)$$

2. HBR is not sensitive to redundant measures:

$$\text{HBR}_{x_1..x_n}(s) = \text{HBR}_{x_1..x_n,x_n}(s)$$

3. Given a large enough set of similarity instances, HBR is not sensitive to non-informative measures. In other words, being x_r a random function such that $P(x_r(s) > x_r(s')) = \frac{1}{2}$, then:

$$\text{HBR}_{x_1..x_n}(s) \sim \text{HBR}_{x_1..x_n,x_r}(s)$$

The first two properties are trivially satisfied: the $\exists$ operator in H and the score comparisons are not affected by redundant measures nor their scale properties. Regarding the third property, the Heterogeneity of a set of measures plus a random function x_r is:

$$H(\mathcal{X} \cup \{x_r\}) \equiv$$

$$P_{s,s'}(\exists x, x' \in \mathcal{X} \cup \{x_r\} | x(s) > x(s') \wedge x'(s) < x'(s')) =$$

$$H(\mathcal{X}) + (1 - H(\mathcal{X})) * \frac{1}{2} = \frac{H(\mathcal{X}) + 1}{2}$$

That is, the Heterogeneity grows proportionally when including a random function. Assuming that the random function corroborates the similarity increase in a half of cases, the result is a proportional relationship between HBR and HBR with the additional measure. Note that we need to assume a large enough amount of data to avoid random effects.

4 Experimental Setting

4.1 Test Bed

We have used the AS test collections used in the DUC 2005 and DUC 2006 evaluation campaigns[1] (Dang, 2005; Dang, 2006). The task was to generate a question focused summary of 250 words from a set of 25-50 documents to a complex question. Summaries were evaluated according to several criteria. Here, we will consider the responsiveness judgements, in which the quality score was an integer between 1 and 5. See Table 1 for a brief numerical description of these test beds.

In order to check the measure combining method, we have employed standard variants of ROUGE (Lin, 2004), including the reversed precision version for each variant [2]. We have considered also the F

[1] http://duc.nist.gov/

[2] Note that the original ROUGE measures are oriented to recall

	DUC 2005	DUC 2006
#human-references	3-4	3-4
#systems	32	35
#system-outputs-assessed	32	35
#system-outputs	50	50
#outputs-assessed per-system	50	50

Table 1: Test collections from 2005 and 2006 DUC evaluation campaigns used in our experiments.

measure between recall and precision oriented measures. Finally, our measure set includes also BE or Basic Elements (Hovy et al., 2006).

4.2 Meta-evaluation criterion

The traditional way of meta-evaluating measures consists of computing the Pearson correlation between measure scores and quality human assessments. But the main goal of automatic evaluation metrics is not exactly to predict the real quality of systems; rather than this, their core mission is detecting system outputs that improve the baseline system in each development cycle. Therefore, the issue is to what extent a quality increase between two system outputs is reflected by the output ranking produced by the measure.

According to this perspective, we propose meta-evaluating measures in terms of an extended version of AUC (Area Under the Curve). AUC can be seen as the probability of observing a score increase when observing a real quality increase between two system outputs (Fawcett, 2006).

$$AUC(x) = P(x(s) > x(s')|Q(s) > Q(s'))$$

In order to customize this measure to our scenario, two special cases must be handled:

(i) For cases in which both summaries obtain the same value, we assume that the measure rewards each instance with equal probability. That is, if $x(s) = x(s')$,$P(x(s) > x(s')|Q(s) > Q(s')) = \frac{1}{2}$.

(ii) Given that in the AS evaluation scenarios there are multiple quality levels, we still apply the same probabilistic AUC definition, considering pairs of summaries in which one of them achieves more quality than the other according to human assessors.

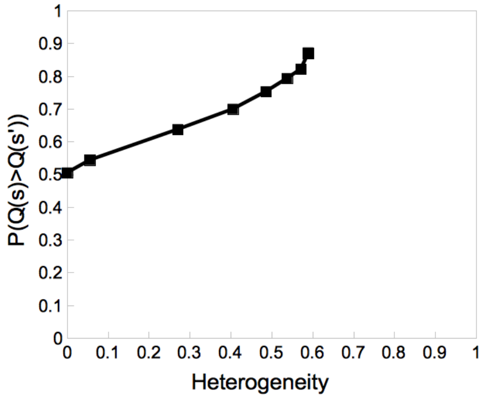

Figure 1: Correlation between probability of quality increase and Heterogeneity of measures that corroborate the increase

5 Experiments

5.1 Measure Heterogeneity vs. Quality Increase

We hypothesize that the probability of a real similarity increase to human references (as stated by human assessments) is directly related to the heterogeneity of the set of measures that confirm such increase. In order to verify whether this principle holds in practice, we need to measure the correlation between both variables. Therefore, we compute, for each pair of summaries in the same topic the heterogeneity of the set of measures that corroborate a score increase between both:

$$H(\{x \in \mathcal{X}|x(s) \geq x(s')\})$$

The Heterogeneity has been estimated by counting cases over 10,000 samples (pairs of summaries) in both corpora.

Then, we have sorted each pair $\langle s, s' \rangle$ according to its related heterogeneity. We have divided the resulting rank in 100 intervals of the same size. For

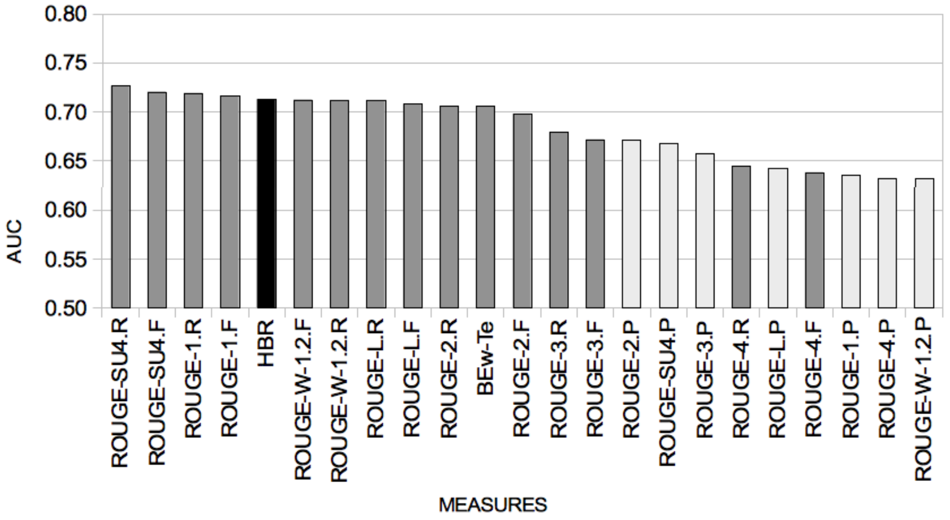

Figure 2: AUC comparison between HBR and single measures in DUC 2005 and DUC 2006 corpora.

each interval, we have computed the average heterogeneity of the set and the probability of real quality increase ($P(Q(s) \geq Q(s'))$).

Figure 1 displays the results. Note that the direct relation between both variables is clear: a key for predicting a real quality increase is how heterogeneous is the set of measures corroborating it.

5.2 HBR vs. Single Measures

In the following experiment, we compute HBR and we compare the resulting AUC with that of single measures. The heterogeneity of measures is estimated over samples in both corpora (DUC 2005 and DUC 2006), and HBR ranking is computed to rank summaries for each topic. For the meta-evaluation, the AUC probability is computed over summary pairs from the same topic.

Figure 2 shows the resulting AUC values of single measures and HBR. The black bar represents the HBR approach. The light grey bars are ROUGE measures oriented to precision. The dark grey bars include ROUGE variants oriented to recall and F, and the measure BE. As the Figure shows, recall-based measures achieve in general higher AUC values than precision-oriented measures. The HBR measure combination appears near the top. It is improved by some measures such as ROUGE_SU4_R, although the difference is not statistically significant ($p = 0.36$ for a t-test between ROUGE_SU4_R and HBR, for instance). HBR improves the 10 worst single measures with statistical significance ($p < 0.025$).

5.3 Robustness

The next question is why using HBR instead of the "best" measure (ROUGE-SU4-R in this case). As we mentioned, the reliability of measures can vary across scenarios. For instance, in DUC scenarios most systems are extractive, and exploit the maximum size allowed in the evaluation campaign guidelines. Therefore, the precision over long n-grams is not crucial, given that the grammaticality of summaries is ensured. In this scenario the recall over words or short n-grams over human references is a clear signal of quality. But we can not ensure that these characteristics will be kept in other corpora, or even when evaluating new kind of summarizers with the same corpora.

Our hypothesis is that, given that HBR resembles the best measure without using human assessments, it should have a more stable performance in situations where the best measure changes.

In order to check empirically this assertion, we have investigated the lower bound performance of measures in our test collections. First, we have ranked measures for each topic according to their AUC values; Then, we have computed, for every measure, its rank regarding the rest of measures (scaled from 0 to 1). Finally, we average each measure across the 10% of topics in which the measure

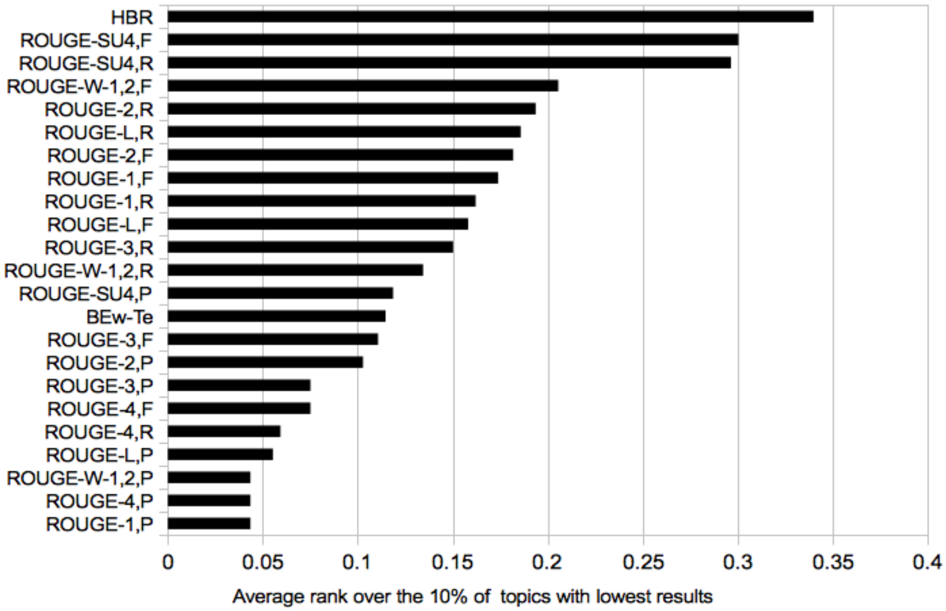

Figure 3: Average rank of measures over the 10% of topics with lowest results for the measure.

gets the worst ranks. Figure 3 shows the results: the worst performance of HBR across topics is better than the worst performance of any single measure. This confirms that the combination of measures using HBR is indeed more robust than any measure in isolation.

5.4 Consistent vs. Inconsistent Topics

The Heterogeneity property is grounded on the assumption that any similarity criteria represents a positive evidence of similarity to human references. In general, we can assert that this assumption holds over a large enough random set of texts. However, depending on the distribution of summaries in the corpus, this assumption may not always hold. For instance, we can assume that, given all possible summaries, improving the word precision with respect to the gold standard can never be a negative evidence of quality. However, for a certain topic, it could happen that the worst summaries are also the shortest, and have high precision and low recall. In this case, precision-based similarity could be correlated with negative quality. Let us refer to these as *inconsistent* topics vs. *consistent* topics. In terms of AUC, a measure represents a negative evidence of quality when AUC is lower than 0.5. Our test collections contain 100 topics, out of which 25 are inconsistent (i.e., at least one measure achieves AUC values lower than 0.5) and 75 are consistent with respect to our measure set (all measures achieve AUC values higher than 0.5).

Figure **??** illustrates the AUC achieved by measures when inconsistent topics are excluded. As with the full set of topics, recall-based measures achieve higher AUC values than precision-based measures; but, in this case, HBR appears at the top of the ranking. This result illustrates that (i) HBR behaves particularly well when our assumptions on similarity measures hold in the corpus; and that (ii) in practice, there may be topics for which our assumptions do not hold.

6 Conclusions

In this paper, we have confirmed that the heterogeneity of a set of summary evaluation measures is correlated with the probability of finding a real quality improvement when all measures corroborate it. The HBR measure combination method is based on this principle, which is grounded on the assumption that any similarity increase with respect to human references is a positive signal of quality.

Our empirical results indicate that the Heterogeneity Based Ranking achieves a reliability similar to the best single measure in the set. In addi-

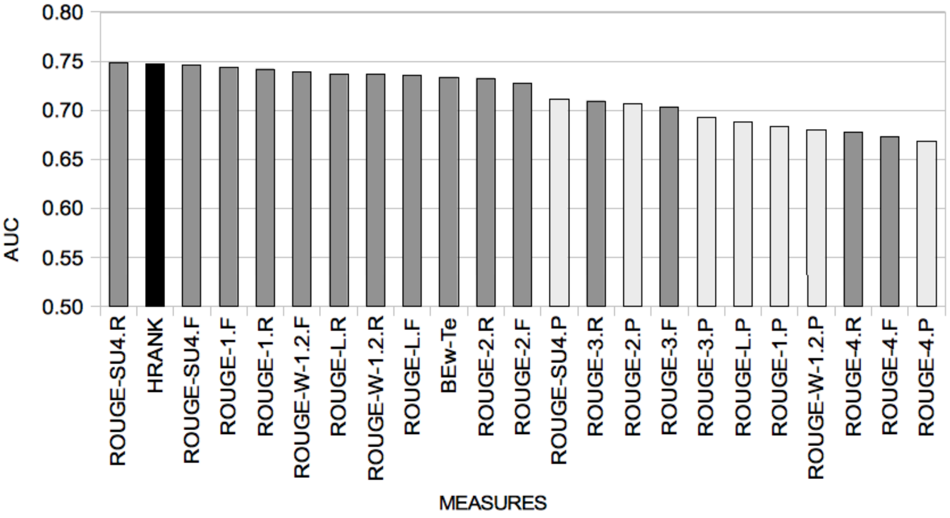

Figure 4: AUC comparison between HBR and single measures in corpora DUC2005 and DUC 2006 over topics in which all measures achieve AUC bigger than 0.5.

tion, HBR results are more robust across topics than single measures. Our experiments also suggest that HBR behaves particularly well when the assumptions of the heterogeneity property holds in the corpus. These assumptions are conditioned by the distribution of summaries in the corpus (in particular, on the amount and variability of the summaries that are compared with human references), and in practice 25% of the topics in our test collections do not satisfy them for our set of measures.

The HBR (Heterogeneity Based Ranking) method proposed in this paper does not represent the "best automatic evaluation measure". Rather than this, it promotes the development of new measures. What HBR does is solving –or at least palliating– the problem of reliability variance of measures across test beds. According to our analysis, our practical recommendations for system refinement are:

1. Compile an heterogenous set of measures, covering multiple linguistic aspects (such as n-gram precision, recall, basic linguistic structures, etc.).

2. Considering the summarization scenario, discard measures that might not always represent a positive evidence of quality. For instance, if very short summaries are allowed (e.g. one word) and they are very frequent in the set of system outputs to be compared to each other, precision oriented measures may violate HBR assumptions.

3. Evaluate automatically your new summarization approach within this corpus according to the HBR method.

Our priority for future work is now developing a reference benchmark containing an heterogenous set of summaries, human references and measures satisfying the heterogeneity assumptions and covering multiple summarization scenarios where different measures play different roles.

The HBR software is available at http://nlp.uned.es/∼enrique/

References

Enrique Amigó, Julio Gonzalo, Jesus Gimenez, and Felisa Verdejo. 2011. Corroborating text evaluation results with heterogeneous measures. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 455–466, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the 2005 Document Understanding Workshop*.

Hoa Trang Dang. 2006. Overview of DUC 2006. In *Proceedings of the 2006 Document Understanding Workshop*.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27:861–874, June.

Tsutomu Hirao, Manabu Okumura, and Hideki Isozaki. 2005. Kernel-based approach for automatic evaluation of natural language generation technologies: Application to automatic summarization. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 145–152, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating DUC 2005 using Basic Elements. Proceedings of Document Understanding Conference (DUC). Vancouver, B.C., Canada.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated Summarization Evaluation with Basic Elements. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 899–902.

Chin-Yew Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Karolina Owczarzak. 2009. Depeval(summ): dependency-based evaluation for automatic summaries. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 190–198, Morristown, NJ, USA. Association for Computational Linguistics.