GRAD: A Metric for Evaluating Summaries

Liana Ermakova[*] — Anton Firsov[**]

[*] *HCTI – EA 4249, Université de Bretagne Occidentale, Brest, France*
LISIS, CNRS-ESIEE-INRA-UPEM, Champs-sur-Marne, France
Université de Lorraine, Nancy, France
[**] *Perm State National Research University, Perm, Russia*

RÉSUMÉ. *Ce papier vise à proposer une nouvelle métrique pour évaluer les résumés. La plupart de méthodes existantes (e.g. ROUGE) nécessitent une intervention humaine importante car elles comparent le résumé considéré avec un ensemble des résumés de référence (gold standard). De plus, les métriques basées sur le chevauchement de vocabulaires ne sont pas appropriées pour la comparaison avec le texte intégral. La métrique proposée intitulée GRAD vise à dépasser les défauts des mesures existantes et s'appuie sur la représentation graphique du texte. L'hypothèse est qu'un bon résumé doit être composé de sommets qui sont connectés avec un maximum d'autres sommets. En outre, nous introduisons un cadre entièrement automatique pour évaluer les métriques qui n'exige aucune annotation humaine. Les expérimentations conduites sur une collection d'articles scientifiques disponibles sur la plate-forme ISTEX ainsi que sur des articles Wikipédia prouvent que la métrique proposée est meilleure de façon significative comparativement aux mesures existantes pour distinguer les résumés automatiques et ceux créés par les humains.*

ABSTRACT. *Automatic summary evaluation is an important but not solved problem. Manual assessment is expensive and subjective and it is not applicable in real time or on a large corpus. Commonly used metrics for summary evaluation still involve substantial human efforts since they assume comparison with a set of reference summaries. Existing metrics based on vocabulary overlap are not suitable for assessment grounded on comparison with a full text. We tried to overcome their drawbacks by proposing a metric based on graph representation of a text. Moreover, we introduce a completely automatic framework for evaluation of metrics that does not require any human annotation. We conducted experiments on Wikipedia data set and a collection of scientific articles. Our approach significantly outperforms strong baselines on both test collections in distinguishing human abstracts from the extractive summaries.*

MOTS-CLÉS : *Résumés automatiques, mesure, métrique, graphe.*

KEYWORDS: *Automatic summary evaluation, measure, metric, graph.*

1. Introduction

In the last decades, data explosion led people to manage the constantly growing amount of information. That involves a need to compress it. Automatic summary evaluation is an important but not solved problem. However, nowadays there is no common approach to summary evaluation. Summaries can be evaluated in different aspects, however, in this paper we focus on informativeness assessment and discrimination between human abstracts and extractive summaries. Manual assessment is expensive, subjective, and often conflicting (Gholamrezazadeh *et al.*, 2009) and, therefore, not applicable on a large scale collection or in real time (e.g. for algorithm tuning). Traditional automatic methods without any human intervention are not used since these techniques provide low results. Therefore, widely used metrics for automatic summary evaluation involve some human efforts and assume comparison with a set of reference summaries. As a rule, the metrics for informativeness evaluation are based on vocabulary overlap. Probably, the most commonly applied semi-automatic measure of informativeness is ROUGE (Lin, 2004). These kinds of metrics are based on the vocabulary overlap and require a gold standard. Full texts can not serve as a gold standard since in this case the measures are mainly reduced to the ratio of the size of a summary over the one of a full text. They can not be applied for the comparison with the full texts. Thereby, the primary objective of this research is to develop a completely automatic measure for summary evaluation based on the full text.

The key idea of our method is to represent a text as a semantic graph, where vertices correspond to terms used in a text and edges represent semantic closeness between corresponding terms. We hypothesize that a good summary should contain terms that refer to the vertices having the maximal number of neighbors in the semantic graph. And vice versa, if in a text there are many terms remote from summary vertices, then this summary should be scored lower. More formally, we use an inverted sum of distances from every term in a text to its closest term from a summary as a measure of summary quality.

Moreover, we introduce a completely automatic framework for evaluation of metrics that does not require any human annotation.

The rest of the paper is organized as follows. Section 2 describes the state-of-the-art in summary evaluation. In Section 3 we propose an automatic measure for evaluation of summary quality. In Section 4 we present an automatic framework to evaluate assessment metrics. Section 5 provides the obtained results and discusses them. Section 6 concludes the paper.

2. Related Work

Readability, coherence, conciseness, content, grammar, recall, pithiness etc. are usually assessed manually (Lin, 2004 ; Saggion *et al.*, 2002) since often these parameters are not numerically expressed (Gholamrezazadeh *et al.*, 2009). Traditional methods of readability evaluation are based on familiarity of terms (Stenner *et al.*, 1988 ; Chall et Dale, 1995 ; Fry, 1990) or their length (Tavernier et Bellot, 2011) and syntax complexity (Collins-Thompson et Callan, 2004). Another set of methods is based on syntax analysis (Mutton *et al.*, 2007 ; Wan *et al.*, 2005 ; Zwarts et Dras, 2008). Syntactical methods may be combined with statistics (e.g. sentence length, the depth of a parse tree, omission of personal verb, rate of prepositional phrases, noun and verb groups etc.) (Chae et Nenkova, 2009). The latter methods are suitable only for the readability evaluation of a particular sentence and therefore they cannot be used for extracts assessment. Researches also propose to use language models for evaluating summaries (Collins-Thompson et Callan, 2004 ; Si et Callan, 2001). Usually assessors assign score to the readability of text in some range (Barzilay *et al.*, 2002). Syntactical errors, unresolved anaphora, redundant information and coherence influence readability and therefore the score may depend on the number of these mistakes (Bellot *et al.*, 2013). Different non-parametric rank correlation coefficients (e.g. Kendall, Spearman or Pearson coefficients) may be used to find the dependence (Lebanon et Lafferty, 2002). However, as shown in (Lapata, 2003), Kendall coefficient is the most suitable for sentence ordering assessment.

To measure the retained information one assessor team develops a set of questions based on the input texts, while another group answers these questions reading only summaries (Seki, 2005). An assessor may be asked to evaluate the importance of each sentence/passage and the obtained importance annotation allows generating summaries with predefined compression rate (Saggion *et al.*, 2002) or serve as expert extracts which may be used as a gold standard. However, on average, assessment agreement is about 70% due to the fact that judges may have different opinions about summary quality and evaluation metrics (Gholamrezazadeh *et al.*, 2009). Inter-rater agreement is traditionally measured by Cohen's kappa (Carletta, 1996) or Krippendorff's alpha in case of arbitrary number of coders (Krippendorff, 2004).

Summaries may be evaluated according to compression rate, i.e. proportion of summary length over full text length, or retention rate, i.e. proportion of the retained information (Gholamrezazadeh *et al.*, 2009). A good summary should have low compression rate and high retention rate. Compression rate is well defined and can be easily computed while retention rate estimation is more problematic since it involves less formalized concepts. Reference summaries allow to compute the metrics commonly used in information retrieval : recall, precision (Gholamrezazadeh *et al.*, 2009), F-measure (Lin, 2004) over the number of terms/sentences appearing in reference and candidate summaries. The F-measure is widely used in ad-hoc information retrieval but it is less useful in summary evaluation since a search engine result is potentially infinite while a summary is limited. Besides, the sentence-based measures of IR types cannot be applied to abstract assessment since abstracting suppose reformulation of initial sentences. Similarity between reference and candidate summaries may be estimated as cosine, dice or Jaccard coefficient, as well as the number of shared n-grams or longest common subsequence etc. One of the most efficient metrics of summary evaluation is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) used at the Document Understanding Conference (DUC) (Lin, 2004). ROUGE is also based on comparison with a set of reference summaries. Proposed by Chin-Yew Lin in 2004, ROUGE aims at evaluating summaries but it is also used to assess the quality of machine translation. There exist several variants of the ROUGE metrics, e.g. $ROUGE - N$ (n-grams recall), $ROUGE - N_{multi}$ (maximal value of pairwise n-gram recalls), $ROUGE - L$ (longest common substring shared by two sentences). The experiments conducted by the organizers of Automatically Evaluating Summaries of Peers (AESOP) task within Text Analysis Conference (TAC) showed that the metrics proposed by TAC participants, such as ROUGE-BE, DemokritosGR2, catholicasc1, and CLASSY1 significantly outperform ROUGE-2 which is the best from all ROUGE variants (Owczarzak *et al.*, 2012).

Normalized pairwise comparison $LCS_{MEAD}(S_1, S_2)$ (Radev *et al.*, 2002) is similar to ROUGE-L when $\beta = 1$, but LCS_{MEAD} takes the maximal value of LCS (longest common substring), while ROUGE-L deals with the union of LCSs (Hovy et Tratz, 2008). One of the serious shortcomings of LCS is the fact that it does not consider the distance between words. This drawback was tried to be eliminated in weighted LCS which takes into account the length of consecutive matches. LCS based algorithms are a special case of edit distance (Bangalore *et al.*, 2000). The metric ROUGE-S is based on the counting of shared bigrams the elements of which may be separated by arbitrary number of other words. The distance may be limited by d_{skip}. Sometimes unigram smoothing is applied (ROUGE-SU).

The metric BLEU commonly used for machine translation evaluation is also suitable for assessment of any generated text (Lin, 2004). As ROUGE, BLEU is also estimated as the number of shared n-grams. BLEU and edit distance may be applied for relevance judgment as well as for readability evaluation. In contrast to BLEU, the Meteor metric is able to treat spelling variants, WordNet synsets and paraphrase tables and distinguishes function and content words (Denkowski et Lavie, 2011). However, this metric is designed for machine translation evaluation and fails to deal with texts of different length.

The organizers of INEX Tweet Contextualization Track 2011-2014 evaluated extractive summaries by comparing them with the pool of passages judged as relevant. As the distance they used the Kullback-Leibler divergence or simple log difference (Bellot *et al.*, 2013). They state that the Kullback-Leibler divergence is very sensitive to smoothing in case of small number of relevant passages in contrast to the absolute log-diff between frequencies (Bellot *et al.*, 2013). (Cabrera-Diego *et al.*, 2016) introduced a trivergent model that outperformed the divergence score.

Tratz and Hovy proposed to use Basic Elements (BE) which can be considered as paraphrases (Hovy et Tratz, 2008). A BE is a syntactic unit up to 3 words with associated tags such as named entities and parts of speech. BE can deal lemmas, synonyms, hyponyms and hyperonyms, identical prepositional phrases, spelling variants, nominalization and denominalization (derivation in WordNet), transformations like prenominal noun - prepositional phrase, noun swapping for IS-A type rules, pronoun transformations, pertainym adjective transformation.

In practice resampling methods are often used, e.g. jackknifing (using subsets of available data) or bootstrapping (random replacement of points in the data set). In this case assessment is the mean of all computed values (Hovy et Tratz, 2008).

In (Campr et Ježek, 2015), the authors proposed to use the similarity within semantic representation such as LSA, LDA, Word2Vec and Doc2Vec. However, ROUGE-1 outperformed all these metrics. In (Ng et Abrecht, 2015), ROUGE metric was modified by word embeddings but this variant showed lower results than the standard one.

A Pyramid score is based on the number of repetitions of information in the gold-standards (Nenkova *et al.*, 2007). In (Owczarzak *et al.*, 2012), a responsiveness metric is proposed. This metric shows how well a summary satisfies the user's information need expressed by a given query.

In (Louis et Nenkova, 2013), the authors suggest an automatic approach for summary evaluation without a gold standard. Instead of a set of reference summaries, a full text is used. They estimate summary score by Kullback-Leibler divergence, Jensen Shannon divergence, and cosine similarity measure. Although these metrics have a some correlation with ROUGE score, ROUGE-1 demonstrated better results.

Traditionally, the quality of assessment metrics is evaluated as a correlation between expert results and candidate metrics (e.g. Kendall, Spearman or Pearson coefficients) (Lin, 2004). A good metric should give low score to summaries which have low score according to human judgment and high score otherwise.

3. GRAD Measure

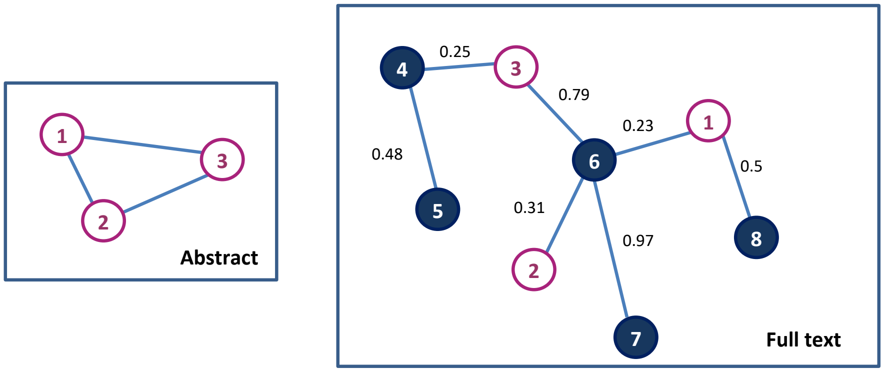
In this paper we propose a completely automatic metric for assessing summary quality which we called GRAD (GRAph Distance). The main idea of our method is to estimate how well summary terms are connected to full text terms.

In the areas of Natural Language Processing and Information Retrieval, on a par of the bag-of-words model, a text is often represented by a graph where vertices denote terms and edges correspond to relations between them (Blanco et Lioma, 2012). We adapted this representation for our research.

Let S be a document represented by a set of sentences $A = {a_k}_{k=1}^{|A|}$ generating a vocabulary V. Let then introduce a graph $G = (V, E)$, where V is a set of vertices corresponding to terms from the document S and E is a set of edges representing semantic links between corresponding terms. In contrast to other representations, relations in our graph are not named but are assigned positive real number proportional to the inverse distance between terms. Thus, the closer terms are semantically, the less distance is between them. We consider that terms are semantically related if they co-occur in the same sentence :

$$E = \{(v_i, v_j) : (\exists a_k \in A | v_i \in a_k, v_j \in a_k)\} \qquad [1]$$

Figure 1. *Example of graph representation of an abstract and a full text*



Analyzing co-occurrence graphs allow discovering lexical semantic affinities beyond grammatical restrictions (Blanco et Lioma, 2012) similar to free association graphs (Steyvers et Tenenbaum, 2005). On the one hand, co-occurrence graphs are widely used in information retrieval, e.g. LDA (Blei *et al.*, 2003), however, state-of-the art models usually ignore indirect connections between terms. On the other hand, knowledge inference in RDF-like representations is a common task in the semantic web, but building such semantic representation manually is extremely labor-intensive and time consuming while automatic construction is restricted by syntactic relations (Blanco et Lioma, 2012). In our approach we combine the simplicity and the power of co-occurrence graphs with the idea of indirect connections in knowledge representations.

Let consider the following sentences :

Example 1 *ROUGE is a metric for evaluating automatic summarization.*
In contrast to ROUGE, GRAD is not based on vocabulary overlap.

It is not explicitly said that GRAD is a metric for evaluating automatic summarization. However, from the second sentence we can conclude that GRAD is connected to ROUGE and in turn ROUGE is connected to evaluation and summarization.

Figure 1 illustrates the suggested representation of a text. Magenta vertices denote terms from a summary. Blue vertices represent other terms in a full text. Edges link terms appearing in the same sentence. For example, the term 4 is a neighbor of term 3 (from the abstract) and 5 (not appearing in the abstract), because they co-occur in the same sentence. Edge weights are the inverse number of times when the terms are neighbors :

$$w(v_i, v_j) = \frac{1}{\sum_{a_k \in A}[v_i \in a_k \wedge v_j \in a_k]} \qquad [2]$$

where [•] is the Iverson bracket. Thus, in the Example 1 $w(ROUGE, GRAD) = \frac{1}{1}$ since *ROUGE* and *GRAD* co-occur in only one sentence, while $w(ROUGE, is) = \frac{1}{2}$.

The intuition behind this is that the more times terms are used in the same sentence, the more semantically close they are and the less distance is between them in the graph. If the edge between the pair of terms is absent then the terms are not semantically connected and the distance is infinite. However, indirect connection still can exist between those terms.

We hypothesize that a good summary is made of the terms that refer to the central vertices in the semantic graph, i.e. the terms that are connected to the maximal number of other terms in a full text. According to our metric, the score of a summary is estimated as a normalized inverted sum of distances from every term in the text to its closest term appearing in the summary S :

$$score(S) = \frac{1}{|S| \sum_{v_i} \min_{v_j \in V \cap S} d(v_j, v_i)} \qquad [3]$$

where $d(v_j, v_i)$ is the shortest path between v_i and v_j. Normalization is done by dividing the score by the number of terms in the summary. The normalization is necessary because without it the metric would have unwarranted preference for longer summaries.

Potentially this approach can work with different methods used for semantic graph building. However here we investigate only one of possible methods for creation of such graph.

To calculate minimal distances from every term in the text to its closest term from the summary we used a modified Dijkstra's algorithm (Dijkstra, 1959). The algorithm has many variants. Dijkstra's original variant aimed at searching the shortest path between two vertices, but a more common variant produces a shortest-path tree from a fixed single source node to all other vertices in the graph. Our modification is an assignment of zero distance to every vertex representing some term from the summary :

$$d^*(v_i \in V \cap S) = 0 \qquad [4]$$

The values of our metric vary in the range $[0, 1]$.

Let consider another example. Let Example 2 represent the full text and let the abstract be just *GRAD*. The corresponding graph is given in Figure 3. Then, our algorithm computes the minimal distances from each word to the word *GRAD* in this graph.

Example 2 *ROUGE is a metric.*
As ROUGE, GRAD is word based.

The asymptotic complexity of the algorithm used to find distance from every term to closest summary term is the same as complexity of original Dijkstra algorithm. The complexity of algorithm used for building semantic graph is $O(W^2 \times |A|)$ where W is an average number of words in a sentence and $|A|$ is a number of sentences in the article.

Here is the pseudocode of the proposed algorithm :

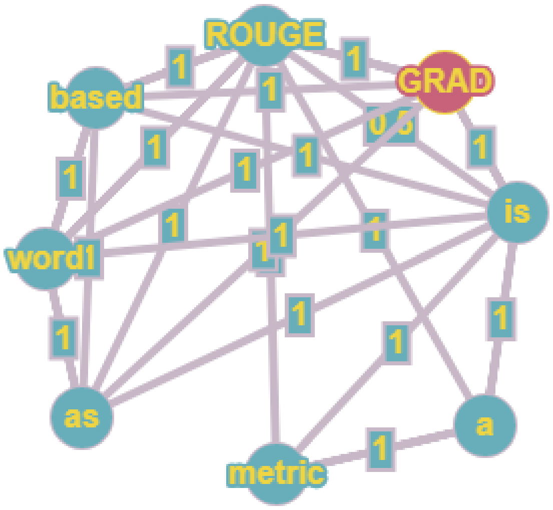


Figure 2. *Example of GRAD graph*

4. Evaluation Framework

The intuition underlying the proposed evaluation framework is that a good assessment metric should assign a high score to a good summary and a low score to a bad one. In contrast to (Lin, 2004), rather than calculating the correlation between the scores assigned to summaries by assessors and metrics, we propose to compare the percentage of times when a good summary of a text is scored lower than a worse one of the same text. This approach requires at least two summaries for every article. The data sets we used contain one human provided summary. Another two summaries we generate automatically using methods described below. We assume that human provided summaries are better than the generated ones and a good metric should reflect that. Therefore the framework does not require explicit human assigned scores and may be performed on very large collections. However, rank correlation coefficients are not applicable for these data since there is no ground truth beyond a pair of summaries and, thus, it is impossible to compare summaries of different full texts.

The second experimental setup aimed at evaluating the capacity of our method to identify the appropriate abstract among human provided summaries. We performed pair-wise comparison between each of 20 documents and abstracts coming from different texts.

Algorithm 1 GRAD(A,S)

```
for all v$\in$V do
    for all u$\in$V do
        w[v,u] = w[u,v] $\leftarrow$ 1.0 / neighbors(v,u)
    end for
end for
notVisited $\leftarrow$ V
for all v$\in$V do
    d[v] $\leftarrow \infty$
end for
for all v$\in$S do
    d[v] $\leftarrow$ 0
end for
while notVisited$\neq \emptyset$ do
    v $\leftarrow$ getClosestNotVisited(V)
    notVisited.remove(v)
    for all u : notVisited do
        if d[u] $>$ d[v] + w[v, u] then
            d[u]$\leftarrow$ d[v] + w[v, u]
        end if
    end for
end while
totalDist $\leftarrow$ 0
for all dd : d do
    totalDist $\leftarrow$ totalDist+dd
end for
return $\frac{1}{|S| \times totalDist}$
```

4.1. *Data Sets*

We conducted experiments on two data sets with existing abstracts : scientific and Wikipedia articles.

4.1.1. *Scientific Articles.*

For our experiments we used the data from the ISTEX (Excellence Initiative of Scientific and Technical Information) platform[1] that contains collections of scientific literature in all disciplines covering journal archives, digital books, databases, texts corpora etc. from the following publishers : Elsevier, Wiley, Springer, Oxford University Press, British Medical Journal, IOP Publishing, Nature, Royal Society of Chemistry, De Gruyter, Ecco Press, Emerald, Brill, Early English Books Online. We collected articles that contain authors abstracts, editorial or/and web summaries. We

[1] http ://www.istex.fr/

selected 4,234 full texts in TXT format with corresponding summaries provided as XML descriptions.

4.1.2. *Wikipedia.*

The second data set we used was a cleaned recent English Wikipedia XML dump created by the INEX organizers for Tweet Contextualization Track (Bellot *et al.*, 2013). All notes, history and bibliographic references were removed. Thus, a page was composed of a title (*title*), an abstract (*a*) and sections (*s*). A section had a header (*h*). Abstract and sections contained paragraphs (*p*) and entities (*t*) referring to other pages. We selected 100,000 longest articles. However, only 43,611 articles had both abstracts and not-empty sections.

4.2. *Methods for Summary Generation*

4.2.1. *Random Summary.*

The first method we used is extractive summarization that randomly selects the sentences from the full texts while the total number of words does not exceed a predefined threshold. To avoid bias because of the different summary sizes, the size of the generated summaries was set to the average size of human-provided ones (200). We call this method **RandSum**.

4.2.2. *Frequency Based Summary.*

The second approach we applied for summary generation is based on the cosine similarity measure between bag-of-words representations of a candidate sentence C and a full texts A :

$$cos(C,A) = \frac{\sum_{i=1}^{n} C_i A_i}{\sqrt{\sum_{i=1}^{n} C_i^2}\sqrt{\sum_{i=1}^{n} A_i^2}} \qquad [5]$$

A_i and C_i refer to term TF-IDF from the corresponding texts. The sentences with the highest scores are selected until the total number of words in a summary is less than a predefined threshold. This method is further denoted by **CosSum**.

Generated this way summaries have bad readability because of abrupt topic changes and unresolved anaphoras. However from informativeness standpoint we assume they will be competitive with human provided summaries. We believe that simple methods for summary generation are sufficient since (1) even in this case vocabulary-based metrics fail and (2) there is no need to check manually if they are less good than the human ones.

4.3. *Competing Metrics*

4.3.1. *Cosine Similarity Measure.*

The similarity between a summary and a full text is estimated by the cosine measure hereinafter referred to as **COS** :

$$cos(S,A) = \frac{\sum_{i=1}^{n} S_i A_i}{\sqrt{\sum_{i=1}^{n} S_i^2}\sqrt{\sum_{i=1}^{n} A_i^2}} \qquad [6]$$

$$cos(S,A) \in [0,1] \qquad [7]$$

where S is the summary under evaluation and A is the corresponding full text, S_i and A_i are TF-IDF of the $i-$th term in the vector representation S and A respectively. IDFs were learned on the entire Wikipedia collection.

4.3.2. *Rouge-N.*

$ROUGE - N$ metric shows the n-gram recall (Lin, 2004) :

$$ROUGE - N(T,S) = \frac{\sum_{T_i \in T} \sum_{g \in T_i} c\_g_m(T_i,S)}{\sum_{T_i \in T} \sum_{g \in T_i} c\_g(T_i)} \qquad [8]$$

$$ROUGE - N(T,S) \in [0,1] \qquad [9]$$

where $T = \{T_i\}_{i=1}^{|T|}$ is a set of reference summaries, $c\_g_m(T_i,S)$ is the maximum number of shared n-grams in a reference summary T_i and in the candidate one S, and $c\_g(T_i)$ is the number of n-grams in a reference summary. $ROUGE - N$ implies that a summary get higher score as it contains more n-grams co-occurring with reference summaries. In our case there is only one reference, i.e. a full text. Thus, the equation (8) can be rewritten as :

$$ROUGE - N(A,S) = \frac{\sum_{g \in A} c\_g_m(A,S)}{\sum_{g \in A} c\_g(A)} \qquad [10]$$

where A is a full text, S is a summary under evaluation. This metric is called further as **ROUGE**.

4.3.3. *INEX Metric.*

The organizers of INEX/CLEF Tweet Contextualization track introduced the following measure of the dissimilarity between a summary and a reference (Bellot *et al.*, 2013) :

$$Dis(S,T) = \sum_{t \in T} \frac{f_{T(t)}}{f_T} \times \left(1 - \frac{\min(\log P, \log Q)}{\max(\log P, \log Q)}\right) \qquad [11]$$

where T is the set of terms in the pool of relevant passages, $f_{T(t)}$ is the frequency of a term t in the pool, f_T is the total number of terms in the pool, $f_{S(t)}$ is the frequency of a term t in a summary, f_S is the total number of terms in a summary. P and Q are computed as :

$$P = \frac{f_{T(t)}}{f_T} + 1 \qquad [12]$$

$$Q = \frac{f_{S(t)}}{f_S} + 1 \qquad [13]$$

The lower values of $Div(S, T)$ corresponds to higher matching of tokens in a pool and a summary. The complement of this dissimilarity measure $1 - Dis(S, T) \in [0, 1]$ has similar properties to usual IR Interpolate Precision measures. Thus, we used this complement as a competing metric hereafter referred to as **INEX**.

4.3.4. *Random Score.*

Since vocabulary overlap based metrics have very low performance when a full text is used as a gold standard, we also compared our results with a metric that assigns a random score to a summary in the range $[0, 1]$ (hereafter called **RAND**).

4.4. *Implementation Details*

We parsed full texts, human provided abstracts and generated summaries by the Stanford CoreNLP parser which allows sentence chunking and tokenization (Manning *et al.*, 2014).

Since our metric tends to the small values, we applied the logarithm :

$$score(S) = \frac{1}{\log(|S| \sum_{v_i} \min_{v_j \in V \cap S} d(v_j, v_i))} \qquad [14]$$

5. Results

Table 1 provides the evaluation results of the proposed metric (GRAD) and the competitors measured as the percentage of times when a human-provided abstract is scored higher/lower/equally than/to a generated summary. The table testifies that our measure significantly outperforms the state-of-the-art metrics as well as the baseline assigning score randomly.

Since the used methods for summary generation are the extractive ones, the metrics based on the overlap between terms should have a low performance. ROUGE and COS are in average twice as worse as the random baseline on both test collections. Although, the INEX measure showed quite competitive results on the ISTEX data set, on the Wikipedia articles its results are much lower than the results obtained by RAND baseline. Thus, we can conclude that full texts can not serve as a gold standard especially in case of extractive summaries since in this case the measures are mainly reduced to the ratio of the size of a summary over the one of a full text.

Tableau 1. *% of times when a human-provided abstract (H) is scored higher/lower/equally than/to a generated summary (S). The best values of $H > S$ are in bold*

Data	Method	RandSum			CosSum		
		$H > S$	$H < S$	$H = S$	$H > S$	$H < S$	$H = S$
ISTEX	COS	16.60%	83.40%	0.00%	16.33%	83.66%	0.01%
	ROUGE	14.49%	85.31%	0.19%	17.61%	82.13%	0.27%
	INEX	61.39%	38.61%	0.00%	62.04%	37.88%	0.08%
	RAND	49.67%	50.33%	0.00%	50.01%	49.99%	0.00%
	GRAD	**62.97%**	35.54%	1.48%	**79.63%**	19.24%	1.13%
Wikipedia	COS	23.13%	76.85%	0.02%	17.17%	82.80%	0.02%
	ROUGE	14.48%	85.31%	0.21%	22.21%	77.37%	0.43%
	INEX	14.08%	85.75%	0.17%	12.66%	87.17%	0.17%
	RAND	49.49%	50.51%	0.00%	49.00%	51.00%	0.00%
	GRAD	**71.60%**	21.80%	6.60%	**92.91%**	1.84%	5.25%

One of the possible explanation why our method significantly outperforms the vocabulary overlap based metrics is that other measures are dealing with pure term frequencies regardless of their context. Under the circumstances, a term may be frequent in a particular context or a specific document part but not very informative for the entire text. In contrast, the GRAD metric takes into account how strongly a term is connected to all other terms in the text.

However, our method fails to distinguish native abstracts from other human summaries. Only in 15% of cases the score assigned to the native abstract was higher than the scores assigned to foreign abstracts.

6. Conclusions

The main contribution of this paper is an automatic metric for assessment of summary quality based on graph representation of a full text. Besides, we proposed a completely automatic framework for evaluation of metrics that does not require any human annotation. We conducted experiments on Wikipedia data set and a collection of scientific articles. Our approach significantly outperforms strong baselines on both test collections. The obtained results also certify that metrics based on vocabulary overlap are not suitable for measuring the quality of a summary with regard to a full text. Although, our method fails to distinguish native abstracts from other human summaries, it can be applied to identify whether a summary looks like a human created one. One of the promising directions is investigating the impact of the IDF of terms and POS tags as well as multi-word terms in the performance of GRAD measure. We also intend to research other semantic graph representations of texts, e.g. based on the topic-comment structure.

Remerciements

The authors would like to thanks Marianne Noël, Frédérique Bordignon, Nicolas Turenne, and Université Paris Descartes, Frontières du vivant.

7. Bibliographie

Bangalore S., Rambow O., Whittaker S., « Evaluation metrics for generation », *Proceedings of the first international conference on*p. 1-8, 2000.

Barzilay R., Elhadad N., McKeown K. R., « Inferring Strategies for Sentence Ordering in Multidocument News Summarization », *Journal of Artificial Intelligence Research*p. 35-55, 2002. 17.

Bellot P., Doucet A., Geva S., Gurajada S., Kamps J., Kazai G., Koolen M., Mishra A., Moriceau V., Mothe J., Preminger M., SanJuan E., Schenkel R., Tannier X., Theobald M., Trappett M., Wang Q., « Overview of INEX 2013 », *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*, p. 269-281, 2013.

Blanco R., Lioma C., « Graph-based term weighting for information retrieval », *Information Retrieval*, vol. 15, n^{o} 1, p. 54-92, February, 2012.

Blei D. M., Ng A. Y., Jordan M. I., « Latent Dirichlet Allocation », *Journal of Machine Learning Research*p. 993-1022, 2003. 3.

Cabrera-Diego L. A., Torres-Moreno J.-M., Durette B., *Evaluating Multiple Summaries Without Human Models : A First Experiment with a Trivergent Model*, Springer International Publishing, Cham, p. 91-101, 2016.

Campr M., Ježek K., *Comparing Semantic Models for Evaluating Automatic Document Summarization*, Springer International Publishing, Cham, p. 252-260, 2015.

Carletta J., « Assessing agreement on classification tasks : The kappa statistic », *Computational Linguistics*, vol. 22, n^{o} 2, p. 249-254, 1996.

Chae J., Nenkova A., « Predicting the fluency of text with shallow structural features : case studies of machine translation and human-written text », *Proceedings of the 12th Conference of the European Chapter of the ACL*p. 139-147, 2009.

Chall J. S., Dale E., *Readability revisited : The new Dale-Chall readability*, MA : Brookline Books, Cambridge, 1995.

Collins-Thompson K., Callan J., « A Language Modeling Approach to Predicting Reading Difficulty », *Proceedings of HLT/NAACL*, 2004.

Denkowski M., Lavie A., « Meteor 1.3 : Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems », *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*p. 85-91, 2011.

Dijkstra E. W., « A note on two problems in connexion with graphs », *Numerische Mathematik*, vol. 1, n^{o} 1, p. 269-271, 1959.

Fry E., « A readability formula for short passages », *Journal of Reading*, vol. 8, p. 594-597, 1990. 33.

Gholamrezazadeh S., Salehi M. A., Gholamzadeh B., « A Comprehensive Survey on Text Summarization Systems », *Computer Science and its Applications*p. 1-6, 2009.

Hovy E., Tratz S., « Summarization Evaluation Using Transformed Basic Elements », *Proceedings TAC 2008*, 2008.

Krippendorff K., *Content Analysis : An Introduction to Its Methodology*, Sage, 2004.

Lapata M., « Probabilistic Text Structuring : Experiments with Sentence Ordering », *Proceedings of ACL*p. 542-552, 2003.

Lebanon G., Lafferty J., « Cranking : Combining rankings using conditional probability models on permutations », *Machine Learning : Proceedings of the Nineteenth International Conference*p. 363-370, 2002.

Lin C.-Y., « ROUGE : A Package for Automatic Evaluation of Summaries », *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*p. 74-81, 2004.

Louis A., Nenkova A., « Automatically Assessing Machine Summary Content Without a Gold Standard », *Comput. Linguist.*, vol. 39, n^o 2, p. 267-300, June, 2013.

Manning C. D., Surdeanu M., Bauer J., Finkel J., Bethard S. J., McClosky D., « The Stanford CoreNLP Natural Language Processing Toolkit », *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 55-60, 2014.

Mutton A., Dras M., Wan S., Dale R., « Gleu : Automatic evaluation of sentence-level fluency », *ACL-07*p. 344-351, 2007.

Nenkova A., Passonneau R., McKeown K., « The Pyramid Method : Incorporating Human Content Selection Variation in Summarization Evaluation », *ACM Trans. Speech Lang. Process.*, May, 2007.

Ng J.-P., Abrecht V., « Better Summarization Evaluation with Word Embeddings for ROUGE », *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, p. 1925-1930, September, 2015.

Owczarzak K., Conroy J. M., Dang H. T., Nenkova A., « An Assessment of the Accuracy of Automatic Evaluation in Summarization », *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1-9, 2012.

Radev D., Teufel S., Saggion H., Lam W., Blitzer J., Elebi A., Qi H., Drabek E., Liu D., Evaluation of text summarization in a cross-lingual information retrieval framework, Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2002.

Saggion H., Radev D., Teufel S., Lam W., Strassel S. M., « Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment », *LREC*p. 747-754, 2002.

Seki Y., « Automatic Summarization Focusing on Document Genre and Text Structure », *ACM SIGIR Forum*, vol. 39, n^o 1, p. 65-67, 2005.

Si L., Callan J., « A statistical model for scientific readability », *Proceedings of the tenth international conference on Information and knowledge management*p. 574-576, 2001.

Stenner A. J., Horablin I., Smith D. R., Smith M., « The Lexile Framework. Durham, NC : Metametrics », 1988.

Steyvers M., Tenenbaum J. B., « The Large-scale structure of semantic networks : Statistical analyses and a model of semantic growth », *Cognitive science*, vol. 29, n^{o} 1, p. 41-78, 2005.

Tavernier J., Bellot P., « Combining relevance and readability for INEX 2011 Question-Answering track », p. 185-195, 2011.

Wan S., Dale R., Dras M., « Searching for grammaticality : Propagating dependencies in the viterbi algorithm », *Proceedings of the Tenth European Workshop on Natural Language Generation*, 2005.

Zwarts S., Dras M., « Choosing the right translation : A syntactically informed classification approach », *Proceedings of the 22nd International Conference on Computational Linguistics*p. 1153-1160, 2008.