Summary Evaluation with and without References

Juan-Manuel Torres-Moreno, Horacio Saggion, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-Morales

***Abstract*—We study a new content-based method for the evaluation of text summarization systems without human models which is used to produce system rankings. The research is carried out using a new content-based evaluation framework called FRESA to compute a variety of divergences among probability distributions. We apply our comparison framework to various well-established content-based evaluation measures in text summarization such as COVERAGE, RESPONSIVENESS, PYRAMIDS and ROUGE studying their associations in various text summarization tasks including generic multi-document summarization in English and French, focus-based multi-document summarization in English and generic single-document summarization in French and Spanish.**

***Index Terms*—Text summarization evaluation, content-based evaluation measures, divergences.**

I. INTRODUCTION

TEXT summarization evaluation has always been a complex and controversial issue in computational linguistics. In the last decade, significant advances have been made in this field as well as various evaluation measures have been designed. Two evaluation campaigns have been led by the U.S. agence DARPA. The first one, SUMMAC, ran from 1996 to 1998 under the auspices of the Tipster program [1], and the second one, entitled DUC (Document Understanding Conference) [2], was the main evaluation forum from 2000 until 2007. Nowadays, the Text Analysis Conference (TAC) [3] provides a forum for assessment of different information access technologies including text summarization.

Evaluation in text summarization can be extrinsic or intrinsic [4]. In an extrinsic evaluation, the summaries are assessed in the context of an specific task carried out by a human or a machine. In an intrinsic evaluation, the summaries are evaluated in reference to some ideal model. SUMMAC was mainly extrinsic while DUC and TAC followed an intrinsic evaluation paradigm. In an intrinsic evaluation, an automatically generated summary (*peer*) has to be compared with one or more reference summaries (*models*). DUC used an interface called SEE to allow human judges to compare a *peer* with a *model*. Thus, judges give a COVERAGE score to each *peer* produced by a system and the final system COVERAGE score is the average of the COVERAGE's scores asigned. These system's COVERAGE scores can then be used to rank summarization systems. In the case of query-focused summarization (e.g. when the summary should answer a question or series of questions) a RESPONSIVENESS score is also assigned to each summary, which indicates how responsive the summary is to the question(s).

Because manual comparison of peer summaries with model summaries is an arduous and costly process, a body of research has been produced in the last decade on automatic content-based evaluation procedures. Early studies used text similarity measures such as cosine similarity (with or without weighting schema) to compare peer and model summaries [5]. Various vocabulary overlap measures such as n-grams overlap or longest common subsequence between peer and model have also been proposed [6], [7]. The BLEU machine translation evaluation measure [8] has also been tested in summarization [9]. The DUC conferences adopted the ROUGE package for content-based evaluation [10]. ROUGE implements a series of recall measures based on n-gram co-occurrence between a peer summary and a set of model summaries. These measures are used to produce systems' rank. It has been shown that system rankings, produced by some ROUGE measures (e.g., ROUGE-2, which uses 2-grams), have a correlation with rankings produced using COVERAGE.

In recent years the PYRAMIDS evaluation method [11] has been introduced. It is based on the distribution of "content" of a set of model summaries. Summary Content Units (SCUs) are first identified in the model summaries, then each SCU receives a weight which is the number of models containing or expressing the same unit. Peer SCUs are identified in the peer, matched against model SCUs, and weighted accordingly. The PYRAMIDS score given to a peer is the ratio of the sum of the weights of its units and the sum of the weights of the best possible ideal summary with the same number of SCUs as the peer. The PYRAMIDS scores can be also used for ranking summarization systems. [11] showed that PYRAMIDS scores produced reliable system rankings when multiple (4 or more) models were used and that PYRAMIDS rankings correlate with rankings produced by ROUGE-2 and ROUGE-SU2 (i.e. ROUGE with skip 2-grams). However, this method requires the creation

Manuscript received June 8, 2010. Manuscript accepted for publication July 25, 2010.

Juan-Manuel Torres-Moreno is with LIA/Université d'Avignon, France and École Polytechnique de Montréal, Canada (juan-manuel.torres@univ-avignon.fr).

Eric SanJuan is with LIA/Université d'Avignon, France (eric.sanjuan@univ-avignon.fr).

Horacio Saggion is with DTIC/Universitat Pompeu Fabra, Spain (horacio.saggion@upf.edu).

Iria da Cunha is with IULA/Universitat Pompeu Fabra, Spain; LIA/Université d'Avignon, France and Instituto de Ingeniería/UNAM, Mexico (iria.dacunha@upf.edu).

Patricia Velázquez-Morales is with VM Labs, France (patricia_velazquez@yahoo.com).

of models and the identification, matching, and weighting of SCUs in both: models and peers.

[12] evaluated the effectiveness of the Jensen-Shannon ($\mathcal{JS}$) [13] theoretic measure in predicting systems ranks in two summarization tasks: query-focused and update summarization. They have shown that ranks produced by PYRAMIDS and those produced by $\mathcal{JS}$ measure correlate. However, they did not investigate the effect of the measure in summarization tasks such as generic multi-document summarization (DUC 2004 Task 2), biographical summarization (DUC 2004 Task 5), opinion summarization (TAC 2008 OS), and summarization in languages other than English.

In this paper we present a series of experiments aimed at a better understanding of the value of the $\mathcal{JS}$ divergence for ranking summarization systems. We have carried out experimentation with the proposed measure and we have verified that in certain tasks (such as those studied by [12]) there is a strong correlation among PYRAMIDS, RESPONSIVENESS and the $\mathcal{JS}$ divergence, but as we will show in this paper, there are datasets in which the correlation is not so strong. We also present experiments in Spanish and French showing positive correlation between the $\mathcal{JS}$ and ROUGE which is the *de facto* evaluation measure used in evaluation of non-English summarization. To the best of our knowledge this is the more extensive set of experiments interpreting the value of evaluation without human models.

The rest of the paper is organized in the following way: First in Section II we introduce related work in the area of content-based evaluation identifying the departing point for our inquiry; then in Section III we explain the methodology adopted in our work and the tools and resources used for experimentation. In Section IV we present the experiments carried out together with the results. Section V discusses the results and Section VI concludes the paper and identifies future work.

II. Related Work

One of the first works to use content-based measures in text summarization evaluation is due to [5], who presented an evaluation framework to compare rankings of summarization systems produced by recall and cosine-based measures. They showed that there was weak correlation among rankings produced by recall, but that content-based measures produce rankings which were strongly correlated. This put forward the idea of using directly the full document for comparison purposes in text summarization evaluation. [6] presented a set of evaluation measures based on the notion of vocabulary overlap including n-gram overlap, cosine similarity, and longest common subsequence, and they applied them to multi-document summarization in English and Chinese. However, they did not evaluate the performance of the measures in different summarization tasks. [7] also compared various evaluation measures based on vocabulary overlap. Although these measures were able to separate random from non-random systems, no clear conclusion was reached on the value of each of the studied measures.

Nowadays, a widespread summarization evaluation framework is ROUGE [14], which offers a set of statistics that compare peer summaries with models. It counts co-occurrences of n-grams in peer and models to derive a score. There are several statistics depending on the used n-grams and the text processing applied to the input texts (e.g., lemmatization, stop-word removal).

[15] proposed a method of evaluation based on the use of "distances" or divergences between two probability distributions (the distribution of units in the automatic summary and the distribution of units in the model summary). They studied two different Information Theoretic measures of divergence: the Kullback-Leibler ($\mathcal{KL}$) [16] and Jensen-Shannon ($\mathcal{JS}$) [13] divergences. $\mathcal{KL}$ computes the divergence between probability distributions P and Q in the following way:

$$D_{\mathcal{KL}}(P||Q) = \frac{1}{2}\sum_w P_w \log_2 \frac{P_w}{Q_w} \tag{1}$$

While $\mathcal{JS}$ divergence is defined as follows:

$$D_{\mathcal{JS}}(P||Q) = \frac{1}{2}\sum_w P_w \log_2 \frac{2P_w}{P_w+Q_w} + Q_w \log_2 \frac{2Q_w}{P_w+Q_w} \tag{2}$$

These measures can be applied to the distribution of units in system summaries P and reference summaries Q. The value obtained may be used as a score for the system summary. The method has been tested by [15] over the DUC 2002 corpus for single and multi-document summarization tasks showing good correlation among divergence measures and both coverage and ROUGE rankings.

[12] went even further and, as in [5], they proposed to compare directly the distribution of words in full documents with the distribution of words in automatic summaries to derive a content-based evaluation measure. They found a high correlation between rankings produced using models and rankings produced without models. This last work is the departing point for our inquiry into the value of measures that do not rely on human models.

III. Methodology

The followed methodology in this paper mirrors the one adopted in past work (e.g. [5], [7], [12]). Given a particular summarization task T, p data points to be summarized with input material $\{I_i\}_{i=0}^{p-1}$ (e.g. document(s), question(s), topic(s)), s peer summaries $\{\text{SUM}_{i,k}\}_{k=0}^{s-1}$ for input i, and m model summaries $\{\text{MODEL}_{i,j}\}_{j=0}^{m-1}$ for input i, we will compare rankings of the s peer summaries produced by various evaluation measures. Some measures that we use compare summaries with n of the m models:

$$\text{MEASURE}_M(\text{SUM}_{i,k}, \{\text{MODEL}_{i,j}\}_{j=0}^{n-1}) \tag{3}$$

while other measures compare peers with all or some of the input material:

$$\text{MEASURE}_M(\text{SUM}_{i,k}, I'_i) \quad (4)$$

where I'_i is some subset of input I_i. The values produced by the measures for each summary $\text{SUM}_{i,k}$ are averaged for each system $k = 0, \ldots, s-1$ and these averages are used to produce a ranking. Rankings are then compared using Spearman Rank correlation [17] which is used to measure the degree of association between two variables whose values are used to rank objects. We have chosen to use this correlation to compare directly results to those presented in [12]. Computation of correlations is done using the *Statistics-RankCorrelation-0.12* package[1], which computes the rank correlation between two vectors. We also verified the good conformity of the results with the correlation test of Kendall τ calculated with the statistical software R. The two nonparametric tests of Spearman and Kendall do not really stand out as the treatment of ex-æquo. The good correspondence between the two tests shows that they do not introduce bias in our analysis. Subsequently will mention only the ρ of Sperman more widely used in this field.

A. Tools

We carry out experimentation using a new summarization evaluation framework: FRESA –FRamework for Evaluating Summaries Automatically–, which includes document-based summary evaluation measures based on probabilities distribution[2]. As in the ROUGE package, FRESA supports different n-grams and skip n-grams probability distributions. The FRESA environment can be used in the evaluation of summaries in English, French, Spanish and Catalan, and it integrates filtering and lemmatization in the treatment of summaries and documents. It is developed in Perl and will be made publicly available. We also use the ROUGE package [10] to compute various ROUGE statistics in new datasets.

B. Summarization Tasks and Data Sets

We have conducted our experimentation with the following summarization tasks and data sets:

1) Generic multi-document-summarization in English (production of a short summary of a cluster of related documents) using data from DUC'04[3], task 2: 50 clusters, 10 documents each – 294,636 words.
2) Focused-based summarization in English (production of a short focused multi-document summary focused on the question "who is X?", where X is a person's name) using data from the DUC'04 task 5: 50 clusters, 10 documents each plus a target person name – 284,440 words.
3) Update-summarization task that consists of creating a summary out of a cluster of documents and a topic. Two sub-tasks are considered here: A) an initial summary has to be produced based on an initial set of documents and topic; B) an update summary has to be produced from a different (but related) cluster assuming documents used in A) are known. The English TAC'08 Update Summarization dataset is used, which consists of 48 topics with 20 documents each – 36,911 words.
4) Opinion summarization where systems have to analyze a set of blog articles and summarize the opinions about a target in the articles. The TAC'08 Opinion Summarization in English[4] data set (taken from the Blogs06 Text Collection) is used: 25 clusters and targets (i.e., target entity and questions) were used – 1,167,735 words.
5) Generic single-document summarization in Spanish using the *Medicina Clínica*[5] corpus, which is composed of 50 medical articles in Spanish, each one with its corresponding author abstract – 124,929 words.
6) Generic single document summarization in French using the "Canadien French Sociological Articles" corpus from the journal *Perspectives interdisciplinaires sur le travail et la santé* (PISTES)[6]. It contains 50 sociological articles in French, each one with its corresponding author abstract – 381,039 words.
7) Generic multi-document-summarization in French using data from the RPM2[7] corpus [18], 20 different themes consisting of 10 articles and 4 abstracts by reference thematic – 185,223 words.

For experimentation in the TAC and the DUC datasets we use directly the peer summaries produced by systems participating in the evaluations. For experimentation in Spanish and French (single and multi-document summarization) we have created summaries at a similar ratio to those of reference using the following systems:

– *ENERTEX* [19], a summarizer based on a theory of textual energy;
– *CORTEX* [20], a single-document sentence extraction system for Spanish and French that combines various statistical measures of relevance (angle between sentence and topic, various Hamming weights for sentences, etc.) and applies an optimal decision algorithm for sentence selection;
– *SUMMTERM* [21], a terminology-based summarizer that is used for summarization of medical articles and uses specialized terminology for scoring and ranking sentences;
– *REG* [22], summarization system based on an greedy algorithm;

[1] http://search.cpan.org/~gene/Statistics-RankCorrelation-0.12/

[2] FRESA is available at: http://lia.univavignon.fr/fileadmin/axes/TALNE/Ressources.html

[3] http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html

[4] http://www.nist.gov/tac/data/index.html

[5] http://www.elsevier.es/revistas/ctl_servlet?_f=7032&revistaid=2

[6] http://www.pistes.uqam.ca/

[7] http://www-labs.sinequa.com/rpm2

- $\mathcal{JS}$ summarizer, a summarization system that scores and ranks sentences according to their Jensen-Shannon divergence to the source document;
- a *lead-based* summarization system that selects the lead sentences of the document;
- a *random-based* summarization system that selects sentences at random;
- *Open Text Summarizer* [23], a multi-lingual summarizer based on the frequency and
- commercial systems: *Word*, *SSSummarizer*[8], *Pertinence*[9] and *Copernic*[10].

C. Evaluation Measures

The following measures derived from human assessment of the content of the summaries are used in our experiments:

- COVERAGE is understood as the degree to which one peer summary conveys the same information as a model summary [2]. COVERAGE was used in DUC evaluations. This measure is used as indicated in equation 3 using human references or models.
- RESPONSIVENESS ranks summaries in a 5-point scale indicating how well the summary satisfied a given information need [2]. It is used in focused-based summarization tasks. This measure is used as indicated in equation 4 since a human judges the summary with respect to a given input "user need" (e.g., a question). RESPONSIVENESS was used in DUC and TAC evaluations.
- PYRAMIDS [11] is a content assessment measure which compares content units in a peer summary to weighted content units in a set of model summaries. This measure is used as indicated in equation 3 using human references or models. PYRAMIDS is the adopted metric for content-based evaluation in the TAC evaluations.

For DUC and TAC datasets the values of these measures are available and we used them directly. We used the following automatic evaluation measures in our experiments:

- ROUGE [14], which is a recall metric that takes into account n-grams as units of content for comparing peer and model summaries. The ROUGE formula specified in [10] is as follows:

$$\text{ROUGE-}n(\text{R}, M) = \frac{\sum_m \in M \sum_{n-\text{gram} \in P} \text{count}_{\text{match}}(n - \text{gram})}{\sum_m \in M \sum \text{count(n-gram)}} \quad (5)$$

  where R is the summary to be evaluated, M is the set of model (human) summaries, $\text{count}_{\text{match}}$ is the number of common n-grams in m and P, and count is the number of n-grams in the model summaries. For the experiments presented here we used uni-grams, 2-grams, and the skip 2-grams with maximum skip distance of 4 (ROUGE-1, ROUGE-2 and ROUGE-SU4). ROUGE is used to compare a peer summary to a set of model summaries in our framework (as indicated in equation 3).
- Jensen-Shannon divergence formula given in Equation 2 is implemented in our FRESA package with the following specification (Equation 6) for the probability distribution of words w.

$$P_w = \frac{C_w^T}{N}$$
$$Q_w = \begin{cases} \frac{C_w^S}{N_S} & \text{if } w \in S \\ \frac{C_w^T + \delta}{N + \delta * B} & \text{otherwise} \end{cases} \quad (6)$$

  Where P is the probability distribution of words w in text T and Q is the probability distribution of words w in summary S; N is the number of words in text and summary $N = N_T + N_S$, $B = 1.5|V|$, C_w^T is the number of words in the text and C_w^S is the number of words in the summary. For smoothing the summary's probabilities we have used $\delta = 0.005$. We have also implemented other smoothing approaches (e.g. Good-Turing [24], that uses the CPAN Perl's Statistics-Smoothing-SGT-2.1.2 package[11]) in FRESA, but we do not use them in the experiments reported here. Following the ROUGE approach, in addition to word uni-grams we use 2-grams and skip n-grams computing divergences such as $\mathcal{JS}$ (using uni-grams) $\mathcal{JS}_2$ (using 2-grams), $\mathcal{JS}_4$ (using the skip n-grams of ROUGE-SU4), and $\mathcal{JS}_M$ which is an average of the $\mathcal{JS}_i$. $\mathcal{JS}$s measures are used to compare a peer summary to its source document(s) in our framework (as indicated in equation 4). In the case of summarization of multiple documents, these are concatenated (in the given input order) to form a single input from which probabilities are computed.

IV. EXPERIMENTS AND RESULTS

We first replicated the experiments presented in [12] to verify that our implementation of $\mathcal{JS}$ produced correlation results compatible with that work. We used the TAC'08 Update Summarization data set and computed $\mathcal{JS}$ and ROUGE measures for each peer summary. We produced two system rankings (one for each measure), which were compared to rankings produced using the manual PYRAMIDS and RESPONSIVENESS scores. Spearman correlations were computed among the different rankings. The results are presented in Table I. These results confirm a high correlation among PYRAMIDS, RESPONSIVENESS and $\mathcal{JS}$. We also verified high correlation between $\mathcal{JS}$ and ROUGE-2 (0.83 Spearman correlation, not shown in the table) in this task and dataset.

Then, we experimented with data from DUC'04, TAC'08 Opinion Summarization pilot task as well as single and

[8] http://www.kryltech.com/summarizer.htm

[9] http://www.pertinence.net

[10] http://www.copernic.com/en/products/summarizer

[11] http://search.cpan.org/~bjoernw/Statistics-Smoothing-SGT-2.1.2/

TABLE I
SPEARMAN CORRELATION OF CONTENT-BASED MEASURES IN TAC'08 UPDATE SUMMARIZATION TASK

Mesure	PYRAMIDS	p-value	RESPONSIVENESS	p-value
ROUGE-2	0.96	$p < 0.005$	0.92	$p < 0.005$
$\mathcal{JS}$	0.85	$p < 0.005$	0.74	$p < 0.005$

multi-document summarization in Spanish and French. In spite of the fact that the experiments for French and Spanish corpora use less data points (i.e., less summarizers per task) than for English, results are still quite significant. For DUC'04, we computed the $\mathcal{JS}$ measure for each peer summary in tasks 2 and 5 and we used $\mathcal{JS}$, ROUGE, COVERAGE and RESPONSIVENESS scores to produce systems' rankings. The various Spearman's rank correlation values for DUC'04 are presented in Tables II (for task 2) and III (for task 5). For task 2, we have verified a strong correlation between $\mathcal{JS}$ and COVERAGE. For task 5, the correlation between $\mathcal{JS}$ and COVERAGE is weak, and that between $\mathcal{JS}$ and RESPONSIVENESS is weak and negative.

Although the Opinion Summarization (OS) task is a new type of summarization task and its evaluation is a complicated issue, we have decided to compare $\mathcal{JS}$ rankings with those obtained using PYRAMIDS and RESPONSIVENESS in TAC'08. Spearman's correlation values are listed in Table IV. As it can be seen, there is weak and negative correlation of $\mathcal{JS}$ with both PYRAMIDS and RESPONSIVENESS. Correlation between PYRAMIDS and RESPONSIVENESS rankings is high for this task (0.71 Spearman's correlation value).

For experimentation in mono-document summarization in Spanish and French, we have run 11 multi-lingual summarization systems; for experimentation in French, we have run 12 systems. In both cases, we have produced summaries at a compression rate close to the compression rate of the authors' provided abstracts. We have then computed $\mathcal{JS}$ and ROUGE measures for each summary and we have averaged the measure's values for each system. These averages were used to produce rankings per each measure. We computed Spearman's correlations for all pairs of rankings.

Results are presented in Tables V, VI and VII. All results show medium to strong correlation between the $\mathcal{JS}$ measures and ROUGE measures. However the $\mathcal{JS}$ measure based on uni-grams has lower correlation than $\mathcal{JS}$s which use n-grams of higher order. Note that table VII presents results for generic multi-document summarization in French, in this case correlation scores are lower than correlation scores for single-document summarization in French, a result which may be expected given the diversity of input in multi-document summarization.

V. DISCUSSION

The departing point for our inquiry into text summarization evaluation has been recent work on the use of content-based evaluation metrics that do not rely on human models but that compare summary content to input content directly [12]. We have some positive and some negative results regarding the direct use of the full document in content-based evaluation.

We have verified that in both generic muti-document summarization and in topic-based multi-document summarization in English correlation among measures that use human models (PYRAMIDS, RESPONSIVENESS and ROUGE) and a measure that does not use models ($\mathcal{JS}$ divergence) is strong. We have found that correlation among the same measures is weak for summarization of biographical information and summarization of opinions in blogs. We believe that in these cases content-based measures should be considered, in addition to the input document, the summarization task (i.e. text-based representation, description) to better assess the content of the peers [25], the task being a determinant factor in the selection of content for the summary.

Our multi-lingual experiments in generic single-document summarization confirm a strong correlation among the $\mathcal{JS}$ divergence and ROUGE measures. It is worth noting that ROUGE is in general the chosen framework for presenting content-based evaluation results in non-English summarization.

For the experiments in Spanish, we are conscious that we only have one model summary to compare with the peers. Nevertheless, these models are the corresponding abstracts written by the authors. As the experiments in [26] show, the professionals of a specialized domain (as, for example, the medical domain) adopt similar strategies to summarize their texts and they tend to choose roughly the same content chunks for their summaries. Previous studies have shown that author abstracts are able to reformulate content with fidelity [27] and these abstracts are ideal candidates for comparison purposes. Because of this, the summary of the author of a medical article can be taken as reference for summaries evaluation. It is worth noting that there is still debate on the number of models to be used in summarization evaluation [28]. In the French corpus PISTES, we suspect the situation is similar to the Spanish case.

VI. CONCLUSIONS AND FUTURE WORK

This paper has presented a series of experiments in content-based measures that do not rely on the use of model summaries for comparison purposes. We have carried out extensive experimentation with different summarization tasks drawing a clearer picture of tasks where the measures could be applied. This paper makes the following contributions:

– We have shown that if we are only interested in ranking summarization systems according to the content of their automatic summaries, there are tasks were models could be subtituted by the full document in the computation of the $\mathcal{JS}$ measure obtaining reliable rankings. However, we have also found that the substitution of models by full-documents is not always advisable. We have

TABLE II
Spearman ρ of content-based measures with Coverage in DUC'04 Task 2

Mesure	Coverage	p-value
Rouge-2	0.79	$p < 0.0050$
$\mathcal{JS}$	0.68	$p < 0.0025$

TABLE III
Spearman ρ of content-based measures in DUC'04 Task 5

Mesure	Coverage	p-value	Responsiveness	p-value
Rouge-2	0.78	$p < 0.001$	0.44	$p < 0.05$
$\mathcal{JS}$	0.40	$p < 0.050$	-0.18	$p < 0.25$

TABLE IV
Spearman ρ of content-based measures in TAC'08 OS task

Mesure	Pyramids	p-value	Responsiveness	p-value
$\mathcal{JS}$	-0.13	$p < 0.25$	-0.14	$p < 0.25$

TABLE V
Spearman ρ of content-based measures with Rouge in the *Medicina Clínica* Corpus (Spanish)

Mesure	Rouge-1	p-value	Rouge-2	p-value	Rouge-SU4	p-value
$\mathcal{JS}$	0.56	$p < 0.100$	0.46	$p < 0.100$	0.45	$p < 0.200$
$\mathcal{JS}_2$	0.88	$p < 0.001$	0.80	$p < 0.002$	0.81	$p < 0.005$
$\mathcal{JS}_4$	0.88	$p < 0.001$	0.80	$p < 0.002$	0.81	$p < 0.005$
$\mathcal{JS}_M$	0.82	$p < 0.005$	0.71	$p < 0.020$	0.71	$p < 0.010$

found weak correlation among different rankings in complex summarization tasks such as the summarization of biographical information and the summarization of opinions.

– We have also carried out large-scale experiments in Spanish and French which show positive medium to strong correlation among system's ranks produced by Rouge and divergence measures that do not use the model summaries.

– We have also presented a new framework, Fresa, for the computation of measures based on $\mathcal{JS}$ divergence. Following the Rouge approach, Fresa package use word uni-grams, 2-grams and skip n-grams computing divergences. This framework will be available to the community for research purposes.

Although we have made a number of contributions, this paper leaves many open questions than need to be addressed. In order to verify correlation between Rouge and $\mathcal{JS}$, in the short term we intend to extend our investigation to other languages such as Portuguese and Chinesse for which we have access to data and summarization technology. We also plan to apply Fresa to the rest of the DUC and TAC summarization tasks, by using several smoothing techniques. As a novel idea, we contemplate the possibility of adapting the evaluation framework for the phrase compression task [29], which, to our knowledge, does not have an efficient evaluation measure. The main idea is to calculate $\mathcal{JS}$ from an automatically-compressed sentence taking the complete sentence by reference. In the long term, we plan to incorporate a representation of the task/topic in the calculation of measures. To carry out these comparisons, however, we are dependent on the existence of references.

Fresa will also be used in the new question-answer task campaign INEX'2010 (http://www.inex.otago.ac.nz/tracks/qa/qa.asp) for the evaluation of long answers. This task aims to answer a question by extraction and agglomeration of sentences in Wikipedia. This kind of task corresponds to those for which we have found a high correlation among the measures $\mathcal{JS}$ and evaluation methods with human intervention. Moreover, the $\mathcal{JS}$ calculation will be among the summaries produced and a representative set of relevant passages from Wikipedia. Fresa will be used to compare three types of systems, although different tasks: the multi-document summarizer guided by a query, the search systems targeted information (focused IR) and the question answering systems.

Acknowledgment

We are grateful to the Programa Ramón y Cajal from Ministerio de Ciencia e Innovación, Spain. This work is partially supported by: a postdoctoral grant from the National Program for Mobility of Research Human Resources (National Plan of Scientific Research, Development and Innovation 2008-2011, Ministerio de Ciencia e Innovación, Spain); the research project CONACyT, number 82050, and the research project PAPIIT-DGAPA (Universidad Nacional Autónoma de México), number IN403108.

TABLE VI
SPEARMAN ρ OF CONTENT-BASED MEASURES WITH ROUGE IN THE PISTES CORPUS (FRENCH)

Mesure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-SU4	p-value
$\mathcal{JS}$	0.70	$p < 0.050$	0.73	$p < 0.05$	0.73	$p < 0.500$
$\mathcal{JS}_2$	0.93	$p < 0.002$	0.86	$p < 0.01$	0.86	$p < 0.005$
$\mathcal{JS}_4$	0.83	$p < 0.020$	0.76	$p < 0.05$	0.76	$p < 0.050$
$\mathcal{JS}_M$	0.88	$p < 0.010$	0.83	$p < 0.02$	0.83	$p < 0.010$

TABLE VII
SPEARMAN ρ OF CONTENT-BASED MEASURES WITH ROUGE IN THE RPM2 CORPUS (FRENCH)

Measure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-SU4	p-value
$\mathcal{JS}$	0.830	$p < 0.002$	0.660	$p < 0.05$	0.741	$p < 0.01$
$\mathcal{JS}_2$	0.800	$p < 0.005$	0.590	$p < 0.05$	0.680	$p < 0.02$
$\mathcal{JS}_4$	0.750	$p < 0.010$	0.520	$p < 0.10$	0.620	$p < 0.05$
$\mathcal{JS}_M$	0.850	$p < 0.002$	0.640	$p < 0.05$	0.740	$p < 0.01$

REFERENCES

[1] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, "Summac: a text summarization evaluation," *Natural Language Engineering*, vol. 8, no. 1, pp. 43–68, 2002.

[2] P. Over, H. Dang, and D. Harman, "DUC in context," *IPM*, vol. 43, no. 6, pp. 1506–1520, 2007.

[3] *Proceedings of the Text Analysis Conference*. Gaithesburg, Maryland, USA: NIST, November 17-19 2008.

[4] K. Spärck Jones and J. Galliers, *Evaluating Natural Language Processing Systems, An Analysis and Review*, ser. Lecture Notes in Computer Science. Springer, 1996, vol. 1083.

[5] R. L. Donaway, K. W. Drummey, and L. A. Mather, "A comparison of rankings produced by summarization evaluation measures," in *NAACL Workshop on Automatic Summarization*, 2000, pp. 69–78.

[6] H. Saggion, D. Radev, S. Teufel, and W. Lam, "Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics," in *COLING 2002*, Taipei, Taiwan, August 2002, pp. 849–855.

[7] D. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Çelebi, D. Liu, and E. Drábek, "Evaluation challenges in large-scale document summarization," in *ACL'03*, 2003, pp. 375–382.

[8] K. Papineni, S. Roukos, T. Ward, , and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *ACL'02*, 2002, pp. 311–318.

[9] K. Pastra and H. Saggion, "Colouring summaries BLEU," in *Evaluation Initiatives in Natural Language Processing*. Budapest, Hungary: EACL, 14 April 2003.

[10] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: ACL-04 Workshop*, M.-F. Moens and S. Szpakowicz, Eds., Barcelona, July 2004, pp. 74–81.

[11] A. Nenkova and R. J. Passonneau, "Evaluating Content Selection in Summarization: The Pyramid Method," in *HLT-NAACL*, 2004, pp. 145–152.

[12] A. Louis and A. Nenkova, "Automatically Evaluating Content Selection in Summarization without Human Models," in *Empirical Methods in Natural Language Processing*, Singapore, August 2009, pp. 306–314. [Online]. Available: http://www.aclweb.org/anthology/D/D09/D09-1032

[13] J. Lin, "Divergence Measures based on the Shannon Entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 145-151, 1991.

[14] C.-Y. Lin and E. Hovy, "Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics," in *HLT-NAACL*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 71–78.

[15] C.-Y. Lin, G. Cao, J. Gao, and J.-Y. Nie, "An information-theoretic approach to automatic evaluation of summaries," in *HLT-NAACL*, Morristown, USA, 2006, pp. 463–470.

[16] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. of Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[17] S. Siegel and N. Castellan, *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1998.

[18] C. de Loupy, M. Guégan, C. Ayache, S. Seng, and J.-M. Torres-Moreno, "A French Human Reference Corpus for multi-documents summarization and sentence compression," in *LREC'10*, vol. 2, Malta, 2010, p. In press.

[19] S. Fernandez, E. SanJuan, and J.-M. Torres-Moreno, "Textual Energy of Associative Memories: performants applications of Enertex algorithm in text summarization and topic segmentation," in *MICAI'07*, 2007, pp. 861–871.

[20] J.-M. Torres-Moreno, P. Velázquez-Morales, and J.-G. Meunier, "Condensés de textes par des méthodes numériques," in *JADT'02*, vol. 2, St Malo, France, 2002, pp. 723–734.

[21] J. Vivaldi, I. da Cunha, J.-M. Torres-Moreno, and P. Velázquez-Morales, "Automatic summarization using terminological and semantic resources," in *LREC'10*, vol. 2, Malta, 2010, p. In press.

[22] J.-M. Torres-Moreno and J. Ramirez, "REG : un algorithme glouton appliqué au résumé automatique de texte," in *JADT'10*. Rome, 2010, p. In press.

[23] V. Yatsko and T. Vishnyakov, "A method for evaluating modern systems of automatic text summarization," *Automatic Documentation and Mathematical Linguistics*, vol. 41, no. 3, pp. 93–103, 2007.

[24] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, 1999.

[25] K. Spärck Jones, "Automatic summarising: The state of the art," *IPM*, vol. 43, no. 6, pp. 1449–1481, 2007.

[26] I. da Cunha, L. Wanner, and M. T. Cabré, "Summarization of specialized discourse: The case of medical articles in spanish," *Terminology*, vol. 13, no. 2, pp. 249–286, 2007.

[27] C.-K. Chuah, "Types of lexical substitution in abstracting," in *ACL Student Research Workshop*. Toulouse, France: Association for Computational Linguistics, 9-11 July 2001 2001, pp. 49–54.

[28] K. Owkzarzak and H. T. Dang, "Evaluation of automatic summaries: Metrics under varying data conditions," in *UCNLG+Sum'09*, Suntec, Singapore, August 2009, pp. 23–30.

[29] K. Knight and D. Marcu, "Statistics-based summarization-step one: Sentence compression," in *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2000, pp. 703–710.