Multilingual Summarization Evaluation without Human Models

Horacio Saggion
TALN - DTIC
Universitat Pompeu Fabra
horacio.saggion@upf.edu

Juan-Manuel Torres-Moreno
LIA/Université d'Avignon
École Polytechnique de Montréal
juan-manuel.torres@univ-avignon.fr

Iria da Cunha
IULA/Universitat Pompeu Fabra
LIA/Université d'Avignon
iria.dacunha@upf.edu

Eric SanJuan
LIA/Université d'Avignon
eric.sanjuan@univ-avignon.fr

Patricia Velázquez-Morales
VM Labs
patricia_vazquez@yahoo.com

Abstract

We study correlation of rankings of text summarization systems using evaluation methods with and without human models. 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 and focus-based multi-document summarization in English and generic single-document summarization in French and Spanish. The research is carried out using a new content-based evaluation framework called FRESA to compute a variety of divergences among probability distributions.

1 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 the summarization evaluation field. Various evaluation frameworks have been established and evaluation measures developed. SUMMAC (Mani et al., 2002), in 1998, provided the first system independent framework for summary evaluation; the Document Understanding Conference (DUC) (Over et al., 2007) was the main evaluation forum from 2000 until 2007; nowadays, the Text Analysis Conference (TAC)[1] provides a forum for assessment of different information access technologies including text summarization.

Evaluation in text summarization can be extrinsic or intrinsic (Spärck-Jones and Galliers, 1996). In an extrinsic evaluation, the summaries are assessed in the context of an specific task a human or machine has to carry out; 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 order to intrinsically evaluate summaries, the automatic summary (*peer*) has to be compared to a *model* summary or summaries. DUC used an interface called SEE to allow human judges compare a peer summary to a model summary. Using SEE, human judges give a *coverage* score to the peer summary representing the degree of overlap with the model summary. Summarization systems obtain a final coverage score which is the average of the coverage's scores associated to their summaries. The system's coverage score can then be used to rank summarization systems. In the case of query-focused summarization (e.g. when the summary has to respond to a question or set 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

[1] http://www.nist.gov/tac

Coling 2010: Poster Volume, pages 1059–1067,
Beijing, August 2010

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 (Donaway et al., 2000), various vocabulary overlap measures such as set of n-grams overlap or longest common subsequence between peer and model have also been proposed (Saggion et al., 2002; Radev et al., 2003). The *Bleu* machine translation evaluation measure (Papineni et al., 2002) has also been tested in summarization (Pastra and Saggion, 2003). The DUC conferences adopted the ROUGE package for content-based evaluation (Lin, 2004). It implements a series of recall measures based on n-gram co-occurrence statistics between a peer summary and a set of model summaries. ROUGE measures can be used to produce systems ranks. It has been shown that system rankings produced by some ROUGE measures (e.g., ROUGE-2 which uses bi-grams) correlate with rankings produced using coverage. In recent years the Pyramids evaluation method (Nenkova and Passonneau, 2004) was introduced. It is based on the distribution of "content" in 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 the 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 used for ranking summarization systems. Nenkova and Passonneau (2004) 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 bi-grams). Still this method requires the creation of models and the identification, matching, and weighting of SCUs in both models and peers.

Donaway et al. (2000) put forward the idea of using directly the full document for comparison purposes, and argued that content-based measures which compare the document to the summary may be acceptable substitutes for those using model summaries. A method for evaluation of summarization systems without models has been recently proposed (Louis and Nenkova, 2009). It is based on the direct content-based comparison between summaries and their corresponding source documents. Louis and Nenkova (2009) evaluated the effectiveness of the Jensen-Shannon (Lin, 1991b) theoretic measure in predicting systems ranks in two summarization tasks query-focused and update summarization. They have shown that ranks produced by Pyramids and ranks produced by the Jensen-Shannon measure correlate. However, they did not investigate the effect of the measure in past 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.

We think that, in order to have a better understanding of document-summary evaluation measures, more research is needed. In this paper we present a series of experiments aimed at a better understanding of the value of the Jensen-Shannon divergence for ranking summarization systems.

We have carried out experimentation with the proposed measure and have verified that in certain tasks (such as those studied by (Louis and Nenkova, 2009)) there is a strong correlation among Pyramids and Responsiveness and the Jensen-Shannon 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 Jensen-Shannon measure and ROUGE.

The rest of the paper is organized in the following way: First in Section 2 we introduce related work in the area of content-based evaluation identifying the departing point for our inquiry; then in Section 3 we explain the methodology adopted in our work and the tools and resources used for experimentation. In Section 4 we present the experiments carried out together with the results. Section 5 discusses the results and Section 6 concludes the paper.

2 Related Work

One of the first works to use content-based measures in text summarization evaluation is due to (Donaway et al., 2000) 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 between rankings produced by recall, but that content-based measures produce rankings which were strongly correlated, thus paving the way for content-based measures in text summarization evaluation.

Radev et al. (2003) 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 measures studied.

Nowadays, a widespread summarization evaluation framework is ROUGE (Lin and Hovy, 2003) which, as we have mentioned before, offers a set of statistics that compare peer summaries with models. Various statistics exist depending on the used n-gram and on the type of text processing applied to the input texts (e.g., lemmatization, stopword removal).

Lin et al. (2006) 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 ($\mathcal{KL}$) (Kullback and Leibler, 1951) and Jensen-Shannon ($\mathcal{JS}$) (Lin, 1991a) divergences. In this work we use the Jensen-Shannon ($\mathcal{JS}$) divergence that is defined as follows:

$$\begin{aligned} D_{\mathcal{JS}}(P||Q) &= \frac{1}{2}\sum_w P_w \log_2 \frac{2P_w}{P_w + Q_w} \\ &+ Q_w \log_2 \frac{2Q_w}{P_w + Q_w} \end{aligned} \quad (1)$$

This measure can be applied to the distribution of units in system summaries P and reference summaries Q and the value obtained used as a score for the system summary. The method has been tested by (Lin et al., 2006) over the DUC 2002 corpus for single and multi document summarization tasks showing good correlation among divergence measures and both coverage and ROUGE rankings.

Louis and Nenkova (2009) went even further and, as in (Donaway et al., 2000), proposed to directly compare the distribution of words in full documents with the distribution of words in automatic summaries to derive a content-based evaluation measure. They found high correlation among rankings produced using models and rankings produced without models. This work is the departing point for our inquiry into the value of measures that do not rely on human models.

3 Methodology

The methodology of this paper mirrors the one adopted in past work (Donaway et al., 2000; Louis and Nenkova, 2009). 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), questions, topics), 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 we use compare summaries with n out of the m models:

$$\text{MEASURE}_M(\text{SUM}_{i,k}, \{\text{MODEL}_{i,j}\}_{j=0}^{n}) \quad (2)$$

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

$$\text{MEASURE}_M(\text{SUM}_{i,k}, I'_i) \quad (3)$$

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 compared using Spearman Rank correlation (Spiegel and Castellan, 1998) used to measure the degree of association between two variables whose values are used to rank objects. We use this correlation to directly compare results to those presented in (Louis and Nenkova, 2009). Computation of correlations is done using the CPAN Statistics-RankCorrelation-0.12 package[2], which computes the rank correlation between two vectors.

3.1 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. 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 to compute various ROUGE statistics in new datasets.

3.2 Summarization Tasks and Data Sets

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

Generic multi-document-summarization in English (i.e. production a short summary of a cluster of related documents) using data from DUC 2004[3] corpus task 2: 50 clusters (10 documents each) – 294,636 words.

Focused-based summarization in English (i.e. production 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 2004 task 5: 50 clusters (10 documents each plus a target person name) – 284,440 words.

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 2008 Update Summarization dataset is used which consists of 48 topics with 20 documents each – 36,911 words.

Opinion summarization where systems have to analyze a set of blog articles and summarize the opinions about a target in the articles. The TAC 2008 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.

Generic single-document summarization in Spanish using the "Spanish *Medicina Clínica*"[5] corpus which is composed of 50 biomedical articles in Spanish, each one with its corresponding author abstract – 124,929 words.

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 with their corresponding author abstracts – 381,039 words.

3.3 Summarization Systems

For experimentation in the TAC and the DUC datasets we directly use the peer summaries produced by systems participating in the evaluations. For experimentation in Spanish and French (single-document summarization) we have created summaries at the compression rates of the model summaries using the following summarization systems:

- *CORTEX* (Torres-Moreno et al., 2002), 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;

- *ENERTEX* (Fernandez et al., 2007), a summarizer based on a theory of textual energy;

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

[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/

- *SUMMTERM* (Vivaldi et al., 2010), a terminology-based summarizer that is used for summarization of medical articles and uses specialized terminology for scoring and ranking sentences;

- *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;

- the multilingual word-frequency *Open Text Summarizer* (Yatsko and Vishnyakov, 2007);

- the *AutoSummarize* program of Microsoft Word;

- the commercial SSSummarizer[7];

- the *Pertinence* summarizer[8];

- the *Copernic* summarizer[9].

3.4 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 (Over et al., 2007). Coverage was used in DUC evaluations.

- *Responsiveness* ranks summaries in a 5-point scale indicating how well the summary satisfied a given information need (Over et al., 2007). It is used in focused-based summarization tasks. Responsiveness was used in DUC-TAC evaluations.

- *Pyramids* (briefly introduced in Section 1) (Nenkova and Passonneau, 2004) is a content assessment measure which compares content units in a peer summary to weighted content units in a set of model summaries. 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:

- We use the *Rouge* package (Lin, 2004) to compute various statistics. For the experiments presented here we used uni-grams, bi-grams, and the skip bi-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.

- Jensen-Shannon divergence formula given in Equation 1 is implemented in our FRESA package with the following specification for the probability distribution of words w.

$$P_w = \frac{C_w^T}{N} \tag{4}$$

$$Q_w = \begin{cases} \frac{C_w^S}{N_S} & \text{if } w \in S \\ \frac{C_w^T+\delta}{N+\delta*B} & \text{elsewhere} \end{cases} \tag{5}$$

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 $\delta = 0.005$.

4 Experiments and Results

We first replicated the experiments presented in (Louis and Nenkova, 2009) to verify that our implementation of JS produced correlation results compatible with that work. We used the TAC 2008 Update Summarization data set and computed JS and ROUGE measures for each peer

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

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

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

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 1. These results confirm a high correlation among Pyramids, Responsiveness, and JS. We also verified high correlation between JS and ROUGE-2 (0.83 Spearman correlation, not shown in the table) in this task and dataset.

Measure	Pyr.	p-value	Resp.	p-value
ROUGE-2	0.96	$p < 0.005$	0.92	$p < 0.005$
JS	0.85	$p < 0.005$	0.74	$p < 0.005$

Table 1: Spearman system rank correlation of content-based measures in TAC 2008 Update Summarization task

Then, we experimented with data from DUC 2004, TAC 2008 Opinion Summarization pilot and with single 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 2004, we computed the JS measure for each peer summary in tasks 2 and 5 and we used JS and the official ROUGE, coverage, and Responsiveness scores to produce systems' rankings. The various Spearman's rank correlation values for DUC 2004 are presented in Tables 2 (for task 2) and 3 (for task 5). For task 2, we have verified a strong correlation between JS and coverage. For task 5, the correlation between JS and coverage is weak, and the correlation between JS and Responsiveness weak and negative.

Measure	Cov.	p-value
ROUGE-2	0.79	$p < 0.0050$
JS	0.68	$p < 0.0025$

Table 2: Spearman system rank correlation of content-based measures with coverage in DUC 2004 Task 2

Measure	Cov.	p-value	Resp.	p-value
ROUGE-2	0.78	$p < 0.001$	0.44	$p < 0.05$
JS	0.40	$p < 0.050$	-0.18	$p < 0.25$

Table 3: Spearman system rank correlation of content-based measures in DUC 2004 Task 5

Although the Opinion Summarization task is a new type of summarization task and its evaluation is a complicated issue, we have decided to compare JS rankings with those obtained using Pyramids and Responsiveness in TAC 2008. Spearman's correlation values are listed in Table 4. As can be seen, there is weak and negative correlation of JS with both Pyramids and Responsiveness. Correlation between Pyramids and Responsiveness rankings is high for this task (0.71 Spearman's correlation value).

Measure	Pyr.	p-value	Resp.	p-value
JS	-0.13	$p < 0.25$	-0.14	$p < 0.25$

Table 4: Spearman system rank correlation of content-based measures in TAC 2008 Opinion Summarization task

For experimentation in Spanish and French, we have run 11 multi-lingual summarization systems over each of the documents in the two corpora, producing summaries at a compression rate close to the compression rate of the provided authors' abstracts. We have computed 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 5-6. All results show medium to strong correlation between JS and ROUGE measures. However the JS measure based on uni-grams has lower correlation than JSs which use n-grams of higher order.

5 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 (Louis and Nenkova, 2009). 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 sum-

Measure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-SU4	p-value
JS	0.56	$p < 0.100$	0.46	$p < 0.100$	0.45	$p < 0.200$
JS_2	0.88	$p < 0.001$	0.80	$p < 0.002$	0.81	$p < 0.005$
JS_4	0.88	$p < 0.001$	0.80	$p < 0.002$	0.81	$p < 0.005$
JS_M	0.82	$p < 0.005$	0.71	$p < 0.020$	0.71	$p < 0.010$

Table 5: Spearman system rank correlation of content-based measures with ROUGE in the *Medicina Clinica* Corpus (Spanish)

Measure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-2	p-value
JS	0.70	$p < 0.050$	0.73	$p < 0.05$	0.73	$p < 0.500$
JS_2	0.93	$p < 0.002$	0.86	$p < 0.01$	0.86	$p < 0.005$
JS_4	0.83	$p < 0.020$	0.76	$p < 0.05$	0.76	$p < 0.050$
JS_M	0.88	$p < 0.010$	0.83	$p < 0.02$	0.83	$p < 0.010$

Table 6: Spearman system rank correlation of content-based measures with ROUGE in the PISTES Sociological Articles Corpus (French)

marization 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 (the Jensen Shannon 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 consider in addition to the input document, the summarization task (i.e. its text-based representation) to better assess the content of the peers, 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 Jensen-Shannon 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 of the articles and this is in fact the reason for choosing this corpus. As the experiments in (da Cunha et al., 2007) 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. 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 (Owkzarzak and Dang, 2009). In the French corpus PISTES, we suspect the situation is similar to the Spanish case.

6 Conclusions and Future Work

This paper has presented a series of experiments in content evaluation in text summarization to assess the value of 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 where models could be substituted by the full document in the computation of the Jensen-Shannon divergence measure obtaining reliable rankings. However, we have also found that the substitution of models by full-documents is not always advisable. We have found weak correlation among different rankings in complex summarization tasks such as the summarization of biographical information and the summa-

Measure	ROUGE-1	p-value	ROUGE-2	p-value	ROUGE-2	p-value
JS	0.83	$p < 0.002$	0.66	$p < 0.05$	0.741	$p < 0.01$
JS_2	0.80	$p < 0.005$	0.59	$p < 0.05$	0.68	$p < 0.02$
JS_4	0.75	$p < 0.010$	0.52	$p < 0.10$	0.62	$p < 0.05$
JS_M	0.85	$p < 0.002$	0.64	$p < 0.05$	0.74	$p < 0.01$

Table 7: Spearman system rank correlation of content-based measures with ROUGE in the RPM2 Corpus (French)

rization of opinions about an "entity".

- 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 Jensen-Shannon divergence. Following the ROUGE approach, FRESA implements word uni-grams, bi-grams and skip n-grams for the computation of divergences. The framework is being made available to the community for research purposes.

Although we have made a number of contributions, this paper leaves many questions open that need to be addressed. In order to verify correlation between ROUGE and JS, in the short term we intend to extend our investigation to other languages and datasets such as Portuguese and Chinese for which we have access to data and summarization technology. We also plan to apply our evaluation framework to the rest of the DUC and TAC summarization tasks to have a full picture of the correlations among measures with and without human models. In the long term we plan to incorporate a representation of the task/topic in the computation of the measures.

Acknowledgements

We thank three anonymous reviewers for their valuable and enthusiastic comments. Horacio Saggion is grateful to the Programa Ramón y Cajal from the Ministerio de Ciencia e Innovación, Spain and to a Comença grant from Universitat Pompeu Fabra (COMENÇA10.004). This work is partially supported by a postdoctoral grant (National Program for Mobility of Research Human Resources; National Plan of Scientific Research, Development and Innovation 2008-2011) given to Iria da Cunha by the Ministerio de Ciencia e Innovación, Spain.

References

da Cunha, Iria, Leo Wanner, and M. Teresa Cabré. 2007. Summarization of specialized discourse: The case of medical articles in spanish. *Terminology*, 13(2):249–286.

Donaway, Robert L., Kevin W. Drummey, and Laura A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *NAACL-ANLP 2000 Workshop on Automatic Summarization*, pages 69–78, Morristown, NJ, USA. ACL.

Fernandez, Silvia, Eric SanJuan, and Juan-Manuel Torres-Moreno. 2007. Textual Energy of Associative Memories: performants applications of Enertex algorithm in text summarization and topic segmentation. In *MICAI'07*, pages 861–871.

Kullback, S. and R.A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Lin, C.-Y. and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 71–78, Morristown, NJ, USA. ACL.

Lin, Chin-Yew, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 463–470, Morristown, NJ, USA. ACL.

Lin, J. 1991a. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(145-151).

Lin, Jianhua. 1991b. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37:145–151.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens, Stan Szpakowicz, editor, *Text Summarization Branches Out: ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July.

Louis, Annie and Ani Nenkova. 2009. Automatically Evaluating Content Selection in Summarization without Human Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 306–314, Singapore, August. ACL.

Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. 2002. Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of NAACL-HLT 2004*.

Over, Paul, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Owkzarzak, Karolina and Hoa Trang Dang. 2009. Evaluation of automatic summaries: Metrics under varying data conditions. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 23–30, Suntec, Singapore, August. ACL.

Papineni, K., S. Roukos, T. Ward, , and W. J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL'02: 40th Annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pastra, K. and H. Saggion. 2003. Colouring summaries Bleu. In *Proceedings of Evaluation Initiatives in Natural Language Processing*, Budapest, Hungary, 14 April. EACL.

Radev, Dragomir R., Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drábek. 2003. Evaluation challenges in large-scale document summarization. In *ACL*, pages 375–382.

Saggion, H., D. Radev, S. Teufel, and W. Lam. 2002. Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics. In *Proceedings of COLING 2002*, pages 849–855, Taipei, Taiwan, August 24-September 1.

Spärck-Jones, Karen and Julia Rose Galliers, editors. 1996. *Evaluating Natural Language Processing Systems, An Analysis and Review*, volume 1083 of *Lecture Notes in Computer Science*. Springer.

Spiegel, S. and N.J. Castellan, Jr. 1998. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill International.

Torres-Moreno, Juan-Manuel, Patricia Velźquez-Morales, and Jean-Guy Meunier. 2002. Condensś de textes par des méthodes numŕiques. In *JADT'02*, volume 2, pages 723–734, St Malo, France.

Vivaldi, Jorge, Iria da Cunha, Juan-Manuel Torres-Moreno, and Patricia Velázquez-Morales. 2010. Automatic summarization using terminological and semantic resources. In *LREC'10*, volume 2, page 10, Malta.

Yatsko, V.A. and T.N. Vishnyakov. 2007. A method for evaluating modern systems of automatic text summarization. *Automatic Documentation and Mathematical Linguistics*, 41(3):93–103.