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).