Re-using high-quality resources for continued evaluation of automated summarization systems

Laura Alonso*, Maria Fuentes†, Marc Massot‡, Horacio Rodríguez†

* GRIAL
Dept. de Lingüística General
Universitat de Barcelona
lalonso@lingua.fil.ub.es

† TALP Research Centre
Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
{mfuentes,horacio}@lsi.upc.es

‡ Dept. d'Informàtica
i Matemàtica Aplicada
Universitat de Girona
marc.massot@udg.es

Abstract

In this paper we present a method for re-using the human judgements on summary quality provided by the DUC contest. The score to be awarded to automatic summaries is calculated as a function of the scores assigned manually to the most similar summaries for the same document. This approach enhances the standard n-gram based evaluation of automatic summarization systems by establishing similarities between *extractive* (vs. *abstractive*) summaries and by taking advantage of the big quantity of evaluated summaries available from the DUC contest. The utility of this method is exemplified by the improvements achieved on a headline production system.

1. Motivation

Automatic Summarization has become in last years an active line of research, first promoted by TIPSTER's SUMMAC (SUMMAC, 1998) and more recently by the DUC competition (DUC, 2004). Initially reduced to a textual, monolingual, single-document condensation task, the field has evolved for covering currently a wide spectrum of summarisation tasks. However, the spectacular progress in summarization systems is diminished by the fact that there is no satisfactory methodology for evaluating summaries. As a consequence, there is no clear way to direct research efforts, because there is no clear way to assess whether a line of work improves the resulting summaries.

Knowing all this, it is clear that the NIST sponsored DUC contest represents a highly valuable opportunity for the automatic summarization community. All systems that participate in DUC have their automatic summaries manually evaluated by NIST assessors, and the performance of the various systems is compared. Thus, DUC establishes a common comparison ground for a wide range of systems, but, most importantly, it provides them with high-quality, human evaluation of their resulting summaries.

However, this yearly evaluation is not enough to assist dayly research. Summarization systems can be very complex, and many important decisions have to be taken for a system to work. It would be desirable that each of the important decisions could be evaluated separately as they are incorporated in the system. This would allow to direct the efforts in development time.

In this paper we present a method for re-using the high quality judgements of DUC for continued evaluation of summarization systems. This method has been applied to evaluate the further improvements of a system that participated in DUC 2003. The results of these continued evaluations are being of much help to direct development efforts.

The rest of the paper is structured as follows. In the following Section we describe the methodology for evaluation by re-use, which is illustrated by evaluating a summarization system. The basic summarization system is presented in Section 3., and the improvements achieved by continued evaluation are presented in Section 4..

2. Methodology for evaluation by re-use

2.1. The goodness of word-based similarity measures for summary comparison

It has often been claimed that word-based similarity measures, like unigram-overlap, fail to account for the goodness of automatic summaries in comparison with human produced *gold standards*, because they cannot capture similarities in meaning without a correspondence to similarities in form.

However, in a recent study (Lin and Hovy, 2003) it is shown that "*automatic evaluation using unigram co-occurrences between summary pairs correlates surprisingly well with human evaluations*". As a consequence of this finding, n-gram based evaluation measures have been established as the main method for evaluating the goodness of DUC'04 summarization systems, by means of ROUGE (Lin, 2004). Scores are assigned to automatic summaries by comparison with summaries created by humans from the same source documents, by n-gram overlap.

The methodology we present here goes beyond the evaluation proposed by ROUGE, taking advantage of two facts:

- n-gram overlap is more adequate to account for similarities between *extractive* summaries than between *abstractive* summaries
- a big quantity of human-made judgements on summary quality is available for a big number of extractive summaries, produced by NIST assessors for DUC

In effect, ROUGE establishes comparisons between automatic, mostly *extractive* summaries, and human, *abstractive* summaries. It can be expected that similarities between pairs of *extractive* summaries are even better represented by n-gram overlap, because the variability in linguistic realization is lower in extractive summaries than in human-generated summaries, since words used to produce the summaries all come from the same original text.

As for the big quantity of judgements, all summaries submitted by every system to DUC are available for every participant, together with the score assigned to them. It can be expected, then, that word-based measures do account for similarity between automatic summaries that have received comparable scores in DUC, because most of the participating systems took an extraction-based approach.

2.2. Assigning scores by transitivity

As follows from the previous section, the human scoring of a new summary can be approximated by weighting of the scores assigned by NIST assessors to similar summaries submitted to DUC. More precisely, our scoring simply computes the weigthted average of the scores assigned by human judges of the N most similar summaries to the summary to be evaluated. Similarity between the new summary and the evaluated summaries is calculated by unigram overlap.

For testing our proposal we applied this methodology (setting N to 3) to the systems participating in Task 1 of DUC 2003. We used as scores DUC *coverage* and *length-adjusted coverage* (LAC), because they account for the informativity of the summaries. We compared the actual scores obtained by the different systems with the average score obtained with our system, as follows:

$$score = \frac{\sum_{i=1}^{3} v_i s_i^2}{\sum_{i=1}^{3} s_i^2} \quad (1)$$

where
s is the similarity between a summary and the summary to be scored, by unigram overlap[1]
v is the score for coverage or LAC assigned to that summary.

As can be seen in Figure 2, the scores assigned automatically to a given summary present correspond very well with the scores assigned to the three other most similar summaries by unigram overlap: the correlation coefficient amounts to 0.99 between approximated and manual scores both for coverage and LAC.

3. The summarization system to be evaluated

We tested the evaluation methodology presented in the previous Section in the continuous evaluation of a summarization system specialized in headline production (Fuentes et al., 2003). A headline is a highly concise representation of the most relevant points contained in a document. It can consist of a sentence, either extracted from the document or automatically generated, or, sometimes, of a list of relevant terms. This subarea of text summarization has experienced an important growth in the last years, mostly because the DUC contest has proposed one such task.

Our headline extraction system combines a Machine Learning approach with manual rules to obtain informative, readable headlines at parametrisable length. Headline extraction is carried out in four steps (see Figure 1):

1. **Enrichment**: the document is segmented in Textual Units (TUs) and enriched with features relevant to the task in three phases:
   (a) **Pre-processing**: general NLP tasks that provide information necessary for further processes, namely: Sentence Segmentation, Tokenization and Morphological Analysis, Named Entity Recognition, POS Tagging and Semantic Tagging (by attaching WordNet synsets, with no attempt to Word Sense Disambiguation). The DUC 2002 segmenter has been used for segmentation, details of the other tasks can be found in (Fuentes and Rodríguez, 2002).
   (b) **Lexical Chainer**: computes lexical and NE chains, following the work of (Morris and Hirst, 1991) and (Barzilay, 1997). We have adapted to English the lexical chainer for Spanish described in (Fuentes and Rodríguez, 2002).
   (c) **Feature Extraction**: extracts the features needed for classification of each TU. Currently the system uses the features described in Table 1. Numeric features are discretized, the number and limits of the intervals have been empirically set with the training data.
2. **Classification**: each TU is classified as belonging to the summary or not, according to its features and a set of classification rules induced from a training corpus. A Decision Tree has been learned using the Sipina shell (SIPINA, 2000). The training corpus was a set of 147 documents with human built extracts, obtained from the DUC 2001 data (Conroy et al., 2001). A confidence score is assigned to each decision, based on the confidence associated to the rule applied, and the set of summary TUs is ranked accordingly.
3. **Summary Content**: from the set of ranked TUs, the one ranked hightest by the ML algorithm is selected for compression.
4. **Simplification**: the selected TU is parsed by MINIPAR (MINIPAR, 1998) and compression rules are applied on the parse to achieve the targeted length and maintain informativity, as follows:
   - find the main verb(s)
   - take syntactically required arguments of main verb(s): subject and objects
   - take complements of main verb(s) that were necessary from the point of view of truth value, for example negative particles
   - take complements of verbal arguments that may specify their truth value, like lexical modifiers
   - take discursively salient sentence constituents, namely, adjuncts marked by a discursive part icle signalling relevance
   - fulfill well-formedness requirements

4. Improvements by continued evaluation

4.1. Starting evaluation

The system has been manually evaluated at the DUC 2003 contest in Task 1, aimed at producing 10 word single

[1]Unigram overlap is normalized by the size of the strings to be compared.

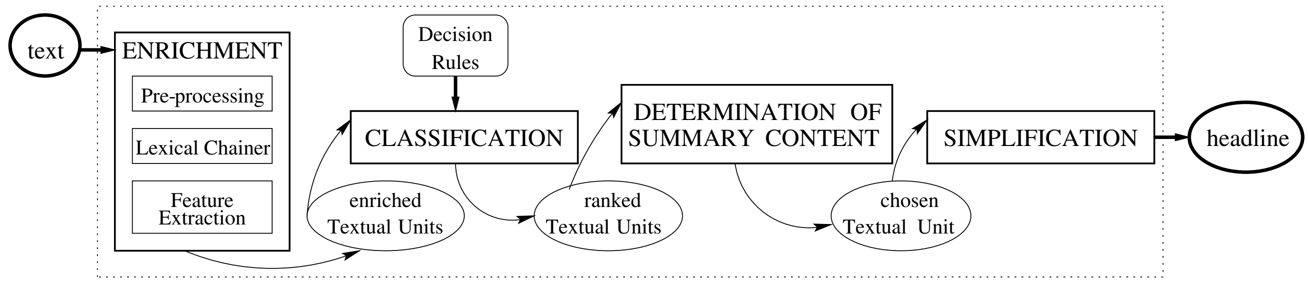

Figure 1: Architecture of the System for Headline Extraction.

Feature Types	Feature Names	Value Type
Length	words, characters, relative_length	one of 5 possible intervals
Position	pos_d	one of 6 possible intervals depending on position of TU in document
Unigram Overlap	uni_1, uni_2, uni_3, uni_4, uni_5	number of TUs in document with unigr_overlap with current TU within interval
Bigram Overlap	bi_0, bi_1	number of TUs in document with not null bigram overlap with current TU
Simple Cosine	scos_1, scos_2, scos_3, scos_4, scos_5	number of TUs in document with cosine with current TU within interval
Weighted Cosine	cos_1, cos_2, cos_3, cos_4, cos_5	number of TUs in document with weighted cosine with current TU within interval
Lexical Chains	strong_lex_chains	number of strong lexical chains crossing current TU (numeric value)

Table 1: Features describing TUs for classification as belonging to summary.

document summaries for pieces of news. Figure 2 displays the results of this evaluation, together with the results for the rest of the systems participating in DUC 2003. Two kinds of scores are displayed: those assigned manually by NIST assessors for *coverage* and *length-adjusted coverage* (LAC), and those approximating the manual scores by the methodology presented in Section 2.2..

Systems are named with the identification used in DUC. The baseline provided by DUC consists in returning the original headlines of the documents, if available. Additionally, we also display approximated results for our improved system and a baseline, created by concatenation of the ten most relevant words in the document (strong LC members and frequent words, leaving stopwords out).

The results of our system were somewhat dissapointing, not reaching .2 coverage or LAC. In order to improve this, a careful analysis of the results was carried out, and some causes of misperformance were identified.

In some cases, the textual unit chosen as most likely to be included in a summary was not actually so. Restrictions in summary length usually supposed a sacrifice in informativity. Additionally, some of the summaries were ungrammatical, either because of parsing errors or by inadequate compression rules, mostly because length restrictions overrode grammaticality constraints. However, ungrammaticality does not affect similarity measures based on unigram overlap, nor the approximated scores based upon them.

4.2. Improvements on the system

An improved version of our system provides solutions for some of the errors in the results submitted to DUC.

Heuristics for choosing the textual unit to be simplified have been refined. First, units with no content other than authorship, location of issue, etc., are identified and discarded by means of pattern matching. This allows progressively decreasing the minimal required length when the combination of heuristics results too restrictive. Moreover, in case no textual unit is chosen from the set provided by the classification module, a second set is built with all the units in the document, ranked by order of occurrence.

As for informativity, the DUC-submitted version of the system determined the inclusion in the summary of a not sentence constituent relying exclusively on syntactical requirements or discursive particles. In this improved version, each lexical item in the chosen TU has been assigned an informativity status, so that words belonging to a strong LC have been considered most informative, and frequent, nonempty words have been assigned a secondary relevance status. Decisions as to the inclusion of sentence constituents in the summary are now taken considering syntactic, rhetoric and lexical information.

4.3. Assessment of the goodness of improvements

Every one of the presented modifications has been evaluated with the methodology presented in Section 2.2., in order to assess whether they introduced significant improvements in the performance of the development system, and include them in the stable version of the system.

What we observed was that emphasizing the informativity of summaries always yielded improvements in performance. This seems to be a side-effect of the fact that comparisons are established by unigram overlap of the sum-

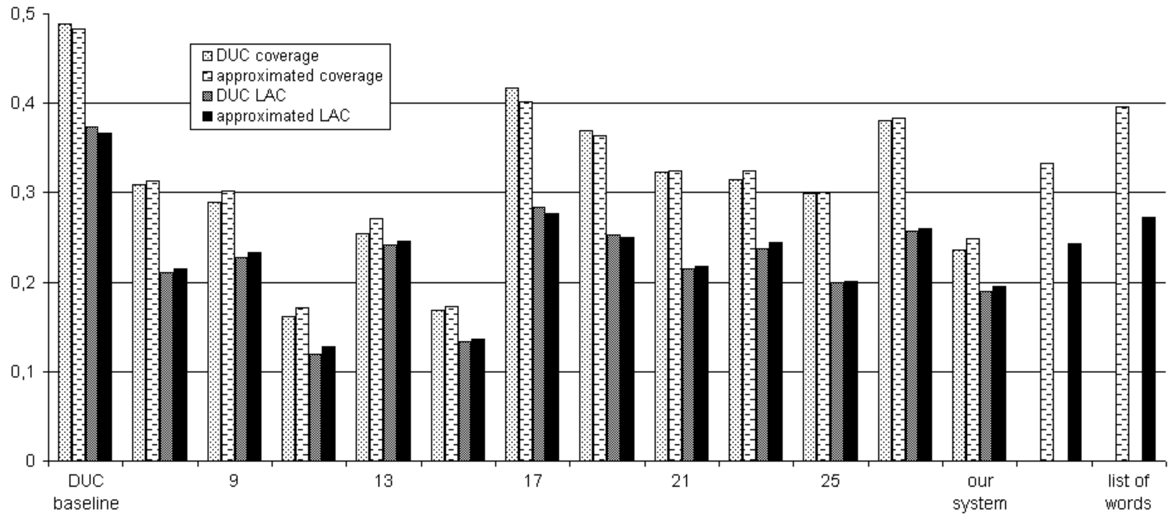

Figure 2: Results obtained by DUC 2003 participating systems, measures are provided for coverage and LAC, both real and calculated by approximation (see Section 2.2.).

maries to be compared. This also explains why the list-of-words baseline outperforms the scores obtained by the system.

It must be said that informativity is not the only target of this system, but also grammaticality and readability of the produced summaries. However, this aspect of texts is not taken into account in unigram-based evaluations.

5. Conclusions and Future Work

We have presented a method to provide high-quality evaluation of automatically produced summaries at a very low cost. This method follows the line of current evaluation efforts in the area of automatic summarization (DUC, 2004), with two main differences from standard methods (Lin, 2004): it is based in word-form similarities between *extractive* summaries, instead of *abstracts*, which may present lexical variations. Secondly, it establishes comparisons between a very high number of summaries, which allows to obtain safer conclusions, since the chance to find very similar summaries increases with the number of available summaries. The evaluation of automatic summaries provided by DUC is crucial to obtain this high number of summaries.

An example application of this methodology has been presented, leading to significant improvements on a system that participated in DUC 2003. Using this methodology, we could evaluate the goodness of every change in the system, and take decisions accordingly.

Future work will be aimed at improving this method by trying to capture grammaticality and readability of automatic summaries. N-gram based measures will be applied, but also features like structural and lexical complexity.

6. Acknowledgements

This research has been partially supported by grant PB98-1226 and ALIADO project (TIC2002-04447), both from the Spanish Research Dept, and also by the European Comission project CHIL (IST-2004506969). Our research group, TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.

7. References

Barzilay, Regina, 1997. *Lexical Chains for Summarization*. Master's thesis, Ben-Gurion University of the Negev.

Conroy, John M., Judith D. Schlesinger, Dianne P. O'Leary, and Mary Ellen Okurowski, 2001. Using HMM and Logistic Regression to generate extract summaries for DUC. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*. New Orleans, Louisiana.

DUC, 2004. DUC–document understanding conference. http://duc.nist.gov/.

Fuentes, Maria, Marc Massot, Horacio Rodríguez, and Laura Alonso, 2003. Headline extraction combining statistic and symbolic techniques. In *DUC03*. Edmonton, Alberta, Canada: Association for Computational Linguistics.

Fuentes, Maria and Horacio Rodríguez, 2002. Using cohesive properties of text for automatic summarization. In *JOTRI'02*.

Lin, Chin-Yew, 2004. http://www.isi.edu/˜ cyl/ROUGE/.

Lin, Chin-Yew and Eduard Hovy, 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Marti Hearst and Mari Ostendorf (eds.), *HLT-NAACL 2003: Main Proceedings*. Edmonton, Alberta, Canada: Association for Computational Linguistics.

MINIPAR, 1998. www.cs.ualberta.ca/˜ lindek/minipar.htm.

Morris, Jane and Graeme Hirst, 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics*, 17(1):21–48.

SIPINA, 2000. http://eric.univ-lyon2.fr/˜ ricco/sipina.html.

SUMMAC, 1998. SUMMAC, the final report. http://www.itl.nist.gov/iaui/894.02/ related_projects/tipster_summac/.