# Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation

Elena Lloret and Manuel Palomar
Department of Software and Computing Systems, University of Alicante, Spain
E-mail:{elloret, mpalomar}@dlsi.ua.es

**Keywords:** natural language processing, automatic summarization, relevance detection, quality-based evaluation

**Received:** April 12, 2009

*This paper is about the Automatic Summarization task within two different points of view, focusing on two main goals. On the one hand, a study of the suitability for* "The Code Quantity Principle" *in the Text Summarization task is described. This linguistic principle is implemented to select those sentences from a text, which carry the most important information. Moreover, this method has been run over the DUC 2002 data, obtaining encouraging results in the automatic evaluation with the ROUGE tool. On the other hand, the second topic discussed in this paper deals with the evaluation of summaries, suggesting new challenges for this task. The main methods to perform the evaluation of summaries automatically have been described, as well as the current problems existing with regard to this difficult task. With the aim of solving some of these problems, a novel type of evaluation is outlined to be developed in the future, taking into account a number of quality criteria in order to evaluate the summary in a qualitative way.*

*Povzetek: Razvita je metoda za zbirni opis besedila, ki temelji na iskanju najpomembnejših stavkov.*

## 1 Introduction

The high amount of electronic information available on the Internet increases the difficulty of dealing with it in recent years. Automatic Summarization (AS) task helps users condense all this information and present it in a brief way, in order to make it easier to process the vast amount of documents related to the same topic that exist these days. Moreover, AS can be very useful for neighbouring Natural Language Processing (NLP) tasks, such as Information Retrieval, Question Answering or Text Comprehension, because these tasks can take advantadge of the summaries to save time and resources [1].

A summary can be defined as a reductive transformation of source text through content condensation by selection and/or generalisation of what is important in the source [2]. According to [3], this process involves three stages: *topic identification*, *interpretation* and *summary generation*. To identify the topic in a document what systems usually do is to assign a score to each unit of input (word, sentence, passage) by means of statistical or machine learning methods. The stage of interpretation is what distinguishes extract-type summarization systems from abstract-type systems. During interpretation, the topics identified as important are fused, represented in new terms, and expressed using a new formulation, using concepts or words not found in the original text. Finally, when the summary content has been created through abstracting and/or information extraction, it requires techniques of Natural Language Generation to build the summary sentences. When an extractive approach is taken, there is no generation stage involved.

Another essential part of the Text Summarization (TS) task is how to perform the evaluation of a summary. Methods for evaluating TS can be classified into two categories [4]. The first, intrinsic evaluations, test the summary on itself. The second, extrinsic evaluations, test how the summary is good enough to accomplish some other task, for example, an Information Retrieval task. However, to determine whether an automatic, or even a human-made summary, is appropriate or not, is a subjective task which depends greatly on a lot of factors, for instance, what the summary is intended for, or to whom the summary is addessed [2].

In this paper, we focus on single-document[1] Text Summarization from an extractive point of view, and we set out two goals for this research. On the one hand, the first goal is to present a method to detect relevant sentences within a document, and therefore, select them to make up the final summary. On the other hand, the second aim of this piece of work is to discuss the current problems the automatic evaluation of summaries in a quantitative way have, so that we can outline a novel approach to measure the quality of a summary to be developed in further research.

The paper is structured as follows: Section 2 gives an overview of the Text Summarization task, describing the main criteria that have been used to determine the relevance of a sentence within a document. In Section 2.1, a new mechanism for detecting important sentences in a text, based on *"The Code Quantity Principle"* [5], is explained.

[1] Single-document differs from multi-document summarization in the number of input documents a system has, just one document or more than one, respectively.

Then, in Section 3 we analise the experiments performed and the results obtained for the approach we have proposed (Section 3.1). We also discuss current problems for evaluating summaries (Section 3.2), proposing a new qualitative model for the evaluation, by means of several quality criteria (Section 3.3). Finally, Section 4 draws the main conclusions and explains the work in progess.

# 2 Determining sentence's relevance in text summarization

Although there has been increased attention to different criteria such as well-formedness, cohesion or coherence when dealing with summarization [6], [7], most work in this NLP task is still concerned with detecting relevant elements of text and presenting them together to produce a final summary. As it has been previously mentioned, the first step in the process of summarization consists of identifying the topic of a document. To achieve this, the most common things systems do is to split the text into input units, usually sentences, and give them a relevance score to decide on which ones are the most important. Criteria such as *sentence position* within texts and *cue phrase indicators* [8], *word and phrase frequency* [9], [10], *query and title overlap* [11], *cohesive or lexical connectedness* [12], [13] or *discourse structure* [14] are examples of how to account for the relevance of a sentence. Furthermore, the use of a graph to obtain a representation of the text has proven effective, especially in multi-document summarization [15], [16], [17].

In contrast to all this work, this paper suggests a novel approach for determining the relevance of a sentence based on *"The Code Quantity Principle"* [5]. This principle tries to explain the relationship between syntax and information within a text. The first goal of this paper is to study whether this principle can be suitable or not as a criterion to select relevant sentences to produce a summary. This idea will be explained in detail in the next Section.

## 2.1 The code quantity principle within the text summarization task

*"The Code Quantity Principle"* [5] is a linguistic theory which states that: (1) a larger chunk of information will be given a larger chunk of code; (2) less predictable information will be given more coding material; and (3) more important information will be given more coding material. In other words, the most important information within a text will contain more lexical elements, and therefore it will be expressed by a high number of units (for instance, syllables, words or phrases). In [18], this principle have been proven to be fulfilled in written texts. Moreover, *"The Code Quantity, Attention and Memory Principle"* [19] states that the more salient and different coding information used within a text, the more reader's attention will be caught. As a result, readers will retain, keep and retrieve this kind of information more efficiently. There exists, then, a proportional relation between the relevance of information and the amount of quantity through it is coded. On the basis of this, a coding element can range from characters to phrases. A noun-phrase is the syntactic structure which allows more flexibility in the number of elements it can contain (pronouns, adjectives, or even relative clauses), and is able to carry more or less information (words) according to the user's needs. Furthermore, the longer a noun-phrase is, the more information it carries for its nucleus. For example, if a text contained two distinct noun-phrases referring to the same entity (*"the Academy of Motion Pictures Arts and Sciences"* and *"the Academy"*), the second one would lead to ambiguities. Therefore, if a summary selected this noun-phrase without having previously given more specific information about the concept, the real meaning could be misunderstood.

Starting from these principles, the approach we suggest here is to study how *"The Code Quantity Principle"* can be applied in the summarization task, to decide on which sentences of a document may contain more relevant information through its coding, and select these sentences to make up a summary. In this particular case, the lexical units considered as encoding elements are words inside a noun-phrase, without taking into account stopwords. The hypothesis is that sentences containing longer noun-phrases will be given a higher score so, at the end, the highest ranked sentences will be chosen to appear in the final summary. To identify noun-phrases within a sentence the *BaseNP Chunker*[2], which was developed at the University of Pennsylvania, was used. One important thing to take into consideration is that the use of a chunker (as well as any other NLP tool) can introduce some error rate. This tool achieves recall and precision rates of roughly 93% for base noun-phrase chunks, and 88% for more complex chunks [20]. For the experiments performed, the score for a sentence was increased by one unit, each time a word belonged to a sentence's noun-phrase. The way we compute the score of a sentence according to the length of the noun-phrase is shown in Formula 1.

$$Sc_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |w| \, . \qquad (1)$$

where:

#NPi = number of noun-phrases contained in sentence i,
|w |= 1, when a word belongs to a noun-phrase.

In Figure 1, an example of how we compute the score of a pair of sentences is showed. Firstly, two sentences that belong to the original document can be seen. Then, chunks of these sentences are identified and stopwords are removed from them. Lastly, scores are calculated according to Formula 1. These sentences have been extracted from the DUC 2002 test data[3]. Once we have the score for

[2]This resource is free available in ftp://ftp.cis.upenn.edu/pub/chunker/
[3]Document Understanding Conference: http://duc.nist.gov/

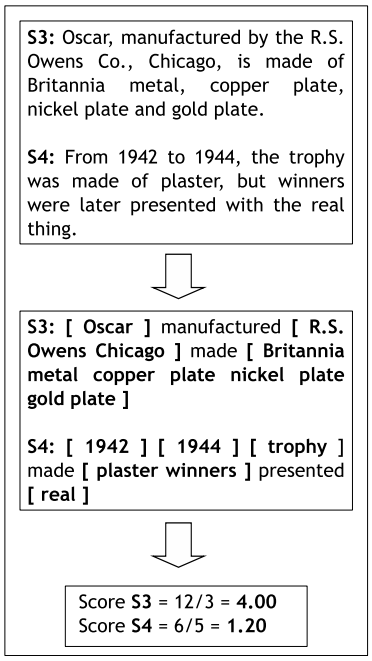


Figure 1: Example of sentence's scoring for document AP880217-0100

The motion picture industry's most coveted award, Oscar, was created 60 years ago and 1,816 of the statuettes have been produced so far. Oscar, manufactured by the R.S. Owens Co., Chicago, is made of Britannia metal, copper plate, nickel plate and gold plate.
According to the Academy of Motion Pictures Arts and Sciences, the only engraving mistake was in 1938 when the best actor trophy given to Spencer Tracy for "Boy's Town" read: "Best Actor: Dick Tracy." The Academy holds all the rights on the statue and "reserves the right to buy back an Oscar before someone takes it to a pawn shop," said Academy spokesman Bob Werden.

Figure 2: Automatic summary for document AP880217-0100

each sentence of the entire document, the sentences with the highest scores will be selected to form part of the final summary, presenting them in the same order as they were in the original text, to keep the order of the text. Figure 2 shows an example of 100-word summary using the proposed scoring method aforementioned. One particular remark of the approach suggested is how pronouns are dealt with. The use of pronouns is very common in written texts, and they substitute somebody/something that has been previously mentioned. Although they can sometimes carry important information, depending on what they are referring to, we decided not to consider them, and consequently they were treated as stopwords. The reason for taking such decision was mainly because they refer to entities previously mentioned in a document, so we strengthened the importance of those mentioned entities instead of noun-phrases containing pronouns.

# 3 Evaluating automatic summarization

Evaluating summaries, either manually or automatically, is a hard task. The main difficulty in evaluation comes from the impossibility of building a fair gold-standard against which the results of the systems can be compared [13]. Furthermore, it is also very hard to determine what a correct summary is, because there is always the possibility of a system to generate a good summary that is quite different from any human summary used as an approximation to the correct output [4]. In Section 1, we mentioned the two approaches that can be adopted to evaluate an automatic summary: instrinsic or extrinsic evaluation. Instrinsic evaluation assesses mainly coherence and summary's information content, whereas extrinsic methods focus on determining the effect of summarization on some other task, for instance Question Answering.

Next, in Section 3.1, we show how we evaluated the novel source of knowledge and the results obtained. Afterwards, in Sections 3.2 and 3.3, we present the problems of the evaluation and the automatic methods developed so far, and we propose a novel idea for evaluating automatic summaries based on quality criteria, respectively.

## 3.1 The code quantity principle evaluation environment

For the approach we have suggested taking into consideration *"The Code Quantity Principle"*, we have chosen an intrinsic evaluation because we are interested in measuring the performance of the automatic summary by itself. To do this, we used the state-of-the-art measure to evaluate summarization systems automatically, ROUGE [21]. This metric measures content overlap between two summaries (normally between a gold-standard and an automatic summary), which means that the distance between two summaries can be established as a function of their vocabulary (unigrams) and how this vocabulary is used (n-grams).

In order to assess the performance of our novel approach based on *"The Code Quantity Principle"* and show that it is suitable for Text Summarization, we evaluated the summaries generated from the DUC 2002 data, consisting of 567 newswire documents. As a preprocessing step, we converted the HTML files into plain text, and we kept only the body of the news. In the DUC 2002 workshop[4], there was a task whose aim was to generate 100-word length sum-

[4]http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

maries. A set of human-made summaries written by experts was also provided. We evaluated our summaries against the reference ones, and we compared our results with the ones obtained by the systems in the real competition. Moreover, the organisation developed a simple baseline which consisted of taking the first 100 words of a document. In [22], the participating systems in DUC 2002 were evaluated automatically with the ROUGE tool, and we set up the same settings[5] for it, so that we could make a proper comparison among all the systems.

In Table 1 we can see the results of the top 3 performing DUC 2002 systems (S28, S21, S19), the baseline proposed in that workshop, and the approach we have suggested in this paper (CQPSum), only for the ROUGE-1, ROUGE-2, ROUGE-SU4 and ROUGE-L recall values. As it is shown in Table 1, the system 28 performed the best at DUC 2002, according to the ROUGE evaluation. From the 13 participating systems, there were only two systems (S28 and S21) that obtained better results than the baseline. The CQPSum approach performed slightly worse than the best system, but it performed, however, better than the rest of the participating systems in DUC 2002, including the baseline, except for the ROUGE-2 value. In S28 [23] two different algorithms, a Logistic Regression Model and a Hidden Markov Model were merged together to develop a single-document summarization system. The features this system used were: position of the sentence in the document, number of tokens in the sentence (stopwords discarded), and number of terms which were more likely to occur in the document (called "pseudo-terms"). They used a machine learning approach to train the data and afterwards, generate the final summary. In contrast, our proposal do not use any machine learning approach, and it is based on a linguistic principle using just one feature (the number of coding words that takes part in a noun-phrase) to discriminate the relevance among sentences. We have shown that this simple idea on its own performs well in the state-of-the-art of single-document summarization task. If more sources of knowledge were combined together, it could be expected that our approach would obtain better results.

## 3.2 Current difficulties in evaluating summaries automatically

The most common way to evaluate the informativeness of automatic summaries is to compare them with human-made model summaries. However, as content selection is not a deterministic problem, different people would chose different sentences, and even, the same person may chose different sentences at different times, showing evidence of low agreement among humans as to which sentences are good summary sentences [24]. Besides the human variability, the semantic equivalence is another problem, because two distinct sentences can express the same meaning but not using the same words. This phenomenon is known as paraphrase. In [25], we can find an approach to automatically evaluating summaries using paraphrases (ParaEval). Moreover, most summarization systems perform an extractive approach, selecting and copying important sentences from the source documents. Although humans can also cut and paste relevant information of a text, most of the times they rephrase sentences when necessary, or they join different related information into one sentence [26].

For years, the summarization community research has been actively seeking an automatic evaluation methodology. Several methods have been proposed, and thanks to the conferences carried out annually until 2007 within the DUC context[6], some of these methodologies, for instance, ROUGE [21] or the Pyramid Method [27] have been well adopted by the researchers to evaluate summaries automatically. Although ROUGE is a recall-oriented metric, the latest version (ROUGE-1.5.5) can compute precision and F-measure, too. It is based on content overlap and the idea behind it is to assess the number of common n-grams between two texts, with respect to different kinds of n-grams, like unigrams, bigrams or the longest common subsequence. In order to address some of the shortcomings of the comparison of fixed words n-grams, an evaluation framework in which very small units of content were used, called Basic Elements (BE) was developed [28].

The idea underlying the Pryamid method is to identify information with the same meaning across different human-authored summaries, which are tagged as *Summary Content Units* (SCU) in order to derive a gold-standard for the evaluation. Each SCU will have a weight depending on the number of summarizers who expressed the same information, and these weights will follow a specific distribution, allowing important content to be differentiated from less important one. The main disadvantages of this method are (1) the need to have several human-made summaries, and (2) the labourious task to annotate all the SCU. An attempt to automate the annotation of the SCUs in the pyramids can be found in [29].

More methods that perform the evaluation of automatic summaries can be found in [30] and [31]. In the former, Relative Utility (RU) is proposed as a metric to evaluate summaries, where multiple judges rank each sentence in the input with a score, giving them a value which ranged from 0 to 10, with respect to its suitability for inclusion in a summary. Highly ranked sentences would be very suitable for a summary, whereas low ranked ones should not be incuded. Like the commonly used information retrieval metric of precision and recall, it compares sentence selection between automatic and reference summaries. The latter have developed an evaluation framework, called QARLA, which provides three types of measures for the evaluation under the assumption that the best similarity metric should

[5]ROUGE version (1.5.5) run with the same parameters as in [22]: ROUGE-1.5.5.pl -n 2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -l 100 -d

[6]The summarization workshop will no longer be referred as DUC. From 2008, the new workshop is called Text Analysis Conference (TAC) and includes other NLP tasks apart from summarization (http://www.nist.gov/tac/).

Table 1: Results for the CQPSum approach

| SYSTEM | ROUGE-1 | ROUGE-2 | ROUGE-SU4 | ROUGE-L |
|---|---|---|---|---|
| S28 | 0.42776 | 0.21769 | 0.17315 | 0.38645 |
| **CQPSum** | **0.42241** | **0.17177** | **0.19320** | **0.38133** |
| S21 | 0.41488 | 0.21038 | 0.16546 | 0.37543 |
| DUC baseline | 0.41132 | 0.21075 | 0.16604 | 0.37535 |
| S19 | 0.40823 | 0.20878 | 0.16377 | 0.37351 |

be the one that best discriminates between manual and automatically generated summaries. These measures are: (1) a measure to evaluate the quality of any set of similarity metrics, (2) a measure to evaluate the quality of a summary using an optimal set of similarity metrics, and (3) a measure to evaluate whether the set of baseline summaries is reliable or may produce biased results.

Despite the fact that many approaches have been developed, some important aspects of summaries, such as legibility, grammaticality, responsiveness or well-formedness are still evaluated manually by experts. For instance, DUC assessors had a list of linguistic qualitity questions[7], and they manually assigned a mark to automatic summaries depending on what extent they accomplished each of these criteria.

## 3.3 Evaluating summaries qualitatively

The main drawback of the evaluation systems existing so far is that we need at least one reference summary, and for some methods more than one, to be able to compare automatic summaries with models. This is a hard and expensive task. Much effort has to be done in order to have corpus of texts and their corresponding summaries. Furthermore, for some methods presented in the previous Section, not only do we need to have human-made summaries available for comparison, but also manual annotation has to be performed in some of them (e.g. SCU in the Pyramid Method). In any case, what the evaluation methods need as an input, is a set of summaries to serve as gold-standards and a set of automatic summaries. Moreover, they all perform a quantitative evaluation with regard to different similarity metrics. To overcome these problems, we think that the quantitative evaluation might not be the only way to evaluate summaries, and a qualitative automatic evaluation would be also important. Therefore, the second aim of this paper is to suggest a novel proposal for evaluating automatically the quality of a summary in a qualitative manner rather than in a quantitative one. Our evaluation approach is a preliminary approach which has to be studied more deeply, and developed in the future. Its main underlying idea is to define several quality criteria and check how a generated summary tackles each of these, in such a way that a reference model would not be necessary anymore, taking only into consideration the automatic summary and the original source. Once performed, it could be used together with any other automatic methodology to measure summary's informativeness.

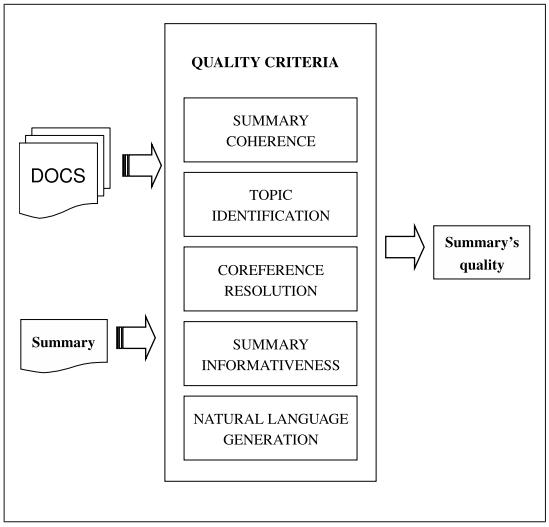


Figure 3: Quality criteria for evaluating summaries in a qualitative way

Attempts to measure the quality of a summary have been previosuly described. In [32] indicativeness (by means of document topics) and sentence acceptability were evaluated by comparing automatic summaries with model ones. More recent approaches have suggested automatic methods to determine the coherence of a summary [33], or even an analisys of several factors regarding readability, which can be used for predicting the quality of texts [34].

As can be seen in Figure 3, the quality criteria aforementioned for the proposed methodoloy will include, among others, coherence within the summary, how anaphoric expressions have been dealt with, whether the topic has been identified correctly or not, or how language generation has been used. The final goal is to set up an independent summarization evaluation environment suitable for generic summaries, which tests a summary's quality, and decides on whether the summary is correct or not, with respect to its original document. Having available a methodology like the one proposed here, would allow automatic summaries to be evaluated automatically in an objective way on their own, without comparing them to any gold-standard in terms of more linguistic and readability aspects.

[7] http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt

# 4 Conclusions and future work

In this paper we presented two main contributions. First of all, we studied *"The Code Quantity Principle"*, which is a linguistic theory about how humans codify the information in a text, depending on what they want a reader to pay more attention to. We presented an approach in which this principle was developed, and we ran it within a newswire domain document set, taking profit of the data provided by DUC 2002 workshop. The evaluation of this method was performed with the ROUGE tool, which made possible the comparison between automatic summaries and reference ones. The results obtained showed that our approach can be suitable for selecting important sentences of a document, and therefore can be a good idea to take this feature into account when building a summarization system. Secondly, owing to all the difficulties the summarization evaluation have, a novel manner of performing the evaluation of an automatic summary was also outlined. What we suggested was to define some quality indicators in order to assess an automatic summary in a qualitative way, rather than in a quantitative one, and therefore, determine if the generated summary can be suitable or not, with regard to its original source.

In future work, we plan to combine, on the one hand, the approach developed to select sentences according to their relevance with other sources of knowledge, such as the word-frequency, and extend this approach to multi-document summarization. Moreover, we are interested in exploring discourse structures in summarization and also, how other human languages technologies can affect the summarization process. Another research line to bear in mind for the future is to provide approaches to be developed with a Natural Language Generation module, in order to try to generate a real summary (that is an abstract, how humans would do summarization) and not only an extract. On the other hand, our second goal for the immediate future is to develop the idea outlined in this paper about evaluating automatic summaries qualitatively, with regard to specific quality criteria, starting from defining such criteria and studying how they can contribute to the evaluation of a summary.

## Acknowledgement

This research has been supported by the FPI grant (BES-2007-16268) from the Spanish Ministry of Science and Innovation, under the project TEXT-MESS (TIN2006-15265-C06-01).

# References

[1] Hassel, M.: Resource Lean and Portable Automatic Text Summarization. PhD thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden (2007)

[2] Spärck Jones, K.: Automatic summarising: The state of the art. Information Processing & Management **43**(6) (2007) 1449–1481

[3] Hovy, E.: Text Summarization. In: The Oxford Handbook of Computational Linguistics. Oxford University Press (2005) 583–598

[4] Mani, I.: Summarization evaluation: An overview. In: Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL). Workshop on Automatic Summarization. (2001)

[5] Givón, T.: Isomorphism in the Grammatical Code. Simone, R. ed. Iconicity in Language (1994)

[6] Alonso i Alemany, L., Fuentes Fort, M.: Integrating cohesion and coherence for automatic summarization. In: EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics. (2003) 1–8

[7] Hasler, L.: Centering theory for evaluation of coherence in computer-aided summaries. In (ELRA), E.L.R.A., ed.: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco (2008)

[8] Edmundson, H.P.: New methods in automatic extracting. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1969) 23–42

[9] Luhn, H.P.: The automatic creation of literature abstracts. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1958) 15–22

[10] Lloret, E., Ferrández, O., Muñoz, R., Palomar, M.: A Text Summarization Approach Under the Influence of Textual Entailment. In: Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008) 12-16 June, Barcelona, Spain. (2008) 22–31

[11] Radev, D.R., Blair-Goldensohn, S., Zhang, Z.: Experiments in single and multi-document summarization using mead. In: First Document Understanding Conference, New Orleans, LA. (2001) 1–7

[12] Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1999) 111–122

[13] Fuentes Fort, M.: A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language. PhD thesis (2008) Adviser-Horacio Rodríguez.

[14] Marcu, D.: Discourse trees are good indicators of importance in text. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1999) 123–136

[15] Mihalcea, R.: Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. (2004) 20

[16] Radev, D.R., Erkan, G., Fader, A., Jordan, P., Shen, S., Sweeney, J.P.: Lexnet: A graphical environment for graph-based nlp. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia. (July 2006) 45–48

[17] Wan, X., Yang, J., Xiao, J.: Towards a unified approach based on affinity graph to various multi-document summarizations. In: Proceedings of the 11th European Conference, ECDL 2007, Budapest, Hungary. (2007) 297–308

[18] Ji, S.: A textual perspective on Givon's quantity principle. Journal of Pragmatics **39**(2) (2007) 292–304

[19] Givón, T.: A functional-typological introduction, II. Amsterdam : John Benjamins (1990)

[20] Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Proceedings of the Third ACL Workshop on Very Large Corpora, Cambridge MA, USA. (1995)

[21] Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003). (2003) 71–78

[22] Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K.: Two uses of anaphora resolution in summarization. Information Processing & Management **43**(6) (2007) 1663–1680

[23] Schlesinger, J.D., Okurowski, M.E., Conroy, J.M., O'Leary, D.P., Taylor, A., Hobbs, J., Wilson, H.: Understanding machine performance in the context of human performance for multi-document summarization. In: Proceedings of the DUC 2002 Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia. (2002)

[24] Nenkova, A.: Summarization evaluation for text and speech: issues and approaches. In: INTERSPEECH-2006, paper 2079-Wed1WeS.1. (2006)

[25] Zhou, L., Lin, C.Y., Munteanu, D.S., Hovy, E.: Paraeval: Using paraphrases to evaluate summaries automatically. In: Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006). New York, NY. (2006) 447–454

[26] Endres-Niggemeyer, B.: Summarizing Information. Berlin: Springer (1998)

[27] Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing **4**(2) (2007) 4

[28] Hovy, E., Lin, C.Y., Zhou, L., Fukumoto, J.: Automated summarization evaluation with basic elements. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy. (2006)

[29] Fuentes, M., González, E., Ferrés, D., Rodríguez, H.: Qasum-talp at duc 2005 automatically evaluated with a pyramid based metric. In: the Document Understanding Workshop (presented at the *HLT/EMNLP* Annual Meeting), Vancouver, B.C., Canada. (2005)

[30] Radev, D.R., Tam, D.: Summarization evaluation using relative utility. In: CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management. (2003) 508–511

[31] Amigó, E., Gonzalo, J., Peñas, A., Verdejo, F.: QARLA: a framework for the evaluation of text summarization systems. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. (2005) 280–289

[32] Saggion, H., Lapalme, G.: Selective analysis for automatic abstracting: Evaluating indicativeness and acceptability. In: Proceedings of Content-Based Multimedia Information Access (RIAO). (2000) 747–764

[33] Barzilay, R., Lapata, M.: Modeling local coherence: An entity-based approach. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, Association for Computational Linguistics (June 2005) 141–148

[34] Pitler, E., Nenkova, A.: Revisiting readability: A unified framework for predicting text quality. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, Association for Computational Linguistics (October 2008) 186–195