An Efficient Statistical Approach for Automatic Organic Chemistry Summarization

Florian Boudin[1], Juan-Manuel Torres-Moreno[1,2], and Patricia Velázquez-Morales[1]

[1] Laboratoire Informatique d'Avignon
339 chemin des Meinajariès, BP1228
84911 Avignon Cedex 9, France
{florian.boudin,juan-manuel.torres}@univ-avignon.fr
[2] École Polytechnique de Montréal
Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada

Abstract. In this paper, we propose an efficient strategy for summarizing scientific documents in Organic Chemistry that concentrates on numerical treatments. We present its implementation named YACHS (Yet Another Chemistry Summarizer) that combines a specific document preprocessing with a sentence scoring method relying on the statistical properties of documents. We show that YACHS achieves the best results among several other summarizers on a corpus made of Organic Chemistry articles.

1 Introduction

Over 1.7 million new Chemistry articles were published in 2007[1], thereby most of scientists today are on *information overload*. Information extraction technology arose in response to the need for efficient processing of documents in specialized domains. Scientists, especially chemists, want to be able to promptly access information concealed in a document in addition to the author's abstract that is often too concise or not satisfying. Automatically producing summaries from Organic Chemistry documents is a challenging but critical task for chemical information retrieval. *Text Summarization* is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user and task [1]. There are many uses of text summarization in everyday activities, we are familiar with summaries such as headlines, reviews or digests. Introduced by [2] in the late 1950's, text summarization was characterized by the use of a surface level approach (i.e. exploiting term frequencies). The first entity-level approaches based on syntactic analysis appeared in the early 1960's [3] while the use of location features and cue phrases was not developed until later [4]. The investigations reported by [5] at the Chemical Abstracts Service

[1] See Chemical Abstracts Publication Record, *http://www.cas.org/*

A. Ranta, B. Nordström (Eds.): GoTAL 2008, LNAI 5221, pp. 89–99, 2008.

(CAS) provide further insight into the effectiveness of automatic summarization in particular domain areas. Corpus-based approaches were introduced by [6] with a trainable summarization system using a collection of text/summaries pairs as training set. A Bayes classifier algorithm takes each sentence and, based on features such as cue phrases, sentence length or location, computes a probability that it should be included in the summary. Thereafter, [7] have extended this model using decision tree rules instead of bayesian classifiers. Rhetorical status was proposed by [8] to summarize scientific articles (Computational Linguistic conference articles) that can highlight the new contribution of the source article. The main drawback of this approach is that it depends on manually constructed resources (metadiscourse features are manually annotated). [9] proposes to combine semantic-based and frequency-distribution approaches for extractive text summarization in biomedical documents. However, this approach requires a difficult concept identification process. Benefits of automatic abstracting are now clearly identified: it is inexpensive compared to human effort and, unlike humans, it is consistent and avoid subjectivity and variability observed in human abstracts. Typically, summarization systems are two-phased, consisting of a content selection step followed by a generation step. Firstly, text fragments (most often sentences) are assigned a score that reflects how important they are. The highest-ranking material can then be arranged and displayed as an "extract". This paper presents YACHS (Yet Another Chemistry Summarizer), a summarization system that generates extracts from scientific articles in a specialized domain, Organic Chemistry. The motivation behind this work is to allow non-experts users to access information contained in high-end scientific documents by dynamically generating extracts. Specifically, through statistical entity level approaches, we seek to produce highly informative extracts that can stand in place of the original author's abstracts as surrogates.

2 Method

2.1 Pre-processing

The first question we are concerned with is whether classical Natural Language Processing (NLP) tools are reasonably consistent across the Organic Chemistry domain (no significant performance loss). The answer is clearly no. Tools such as parsers, taggers or chunkers achieve very poor on these documents without requiring a strenuous, costly and often manual adaptation phase. Issues encountered by classical tools are due to domain specificity: very wide vocabulary, long sentences containing *noise* (citations, chemical formulas, tables, pictures references, etc.), high quantity of *hapax legomena*[2], etc.

The basic idea is to represent the document within the vector space model introduced by [10] and apply specific numeric treatments to select the most salient sentences. An n-dimensional term-space Γ, where n is the number of different terms found in the document, is constructed. One convenient way to represent

[2] Terms which only appears once in a document.

the document in Γ is a matrix $M = [a_{x,y}]_{\ x=1...m;\ y=1...n}$ where m is the number of sentences and n the number of different terms. In this interpretation, every row of M is a vector $\vec{s}_x$ representing the sentence x in which each component is the term frequency within the sentence.

In order to reduce the size of the matrix M and accordingly cut down the computational complexity, sentences are filtered and normalized (see Table 1). In written language, some words carry more *meaning* than others. Thereby, a stop-words elimination phase is performed **(1)** to delete non representative words (words such as '*the*', '*of*', '*in*'... are removed). One standard pre-processing would normalize character case, remove punctuation and special characters **(2)**. However, important information about chemical compounds may be lost during the filtering process (e.g. '1,2-*dienes*' is transformed into '*dienes*'). Besides if word normalization (in our case stemming[3]) is applied afterwards **(3)**, erroneous information is brought in the sentence (e.g. '1,2-*dienes*' is transformed into '*dien*'). We propose to perform a chemical compounds detection to protect these terms during the normalization process **(2′)**. Finally stemming is performed only on non-chemical terms **(3′)**. Chemical compounds are detected within sentences

Table 1. Example of sentence pre-processing

Original	Cycloalkynes are known to isomerize to the 1,2-dienes under basic conditions.
(1)	Cycloalkynes known isomerize 1,2-dienes under basic conditions.
(2)	cycloalkynes known isomerize dienes under basic conditions
(3)	cycloalkyn know isomer dien under basic condit
(2′)	**cycloalkynes** know isomerize **1,2-dienes** under basic conditions
(3′)	**cycloalkynes** *know isomer* **1,2-dienes** *under basic condit*

using a combination of two classifiers. The first one is a Bayes classifier trained on 3-grams of letters whereas the second one uses pattern matching with a small number of manually written rules (7 rules). Each sentence is tokenized in words and each word is classified by the two classifiers, precision is prioritized by using the AND combination (a word has to be classified as chemical compound by the two classifiers). This hybrid approach (statistical and symbolic) for chemical term recognition achieves very good results on a test corpus composed by Organic Chemistry articles [12].

2.2 Sentence Ranking

Once sentences are pre-processed, a combination of features (also called metrics) is used to assign a score to each sentence. That score reflects how important the sentences are in relation to the whole document. The main advantage of this approach is that *zero knowledge* is required and that makes the system fully

[3] The Porter Stemmer algorithm [11] is used to normalize words by removing commoner morphological and inflexional endings from words.

adjustable to any language and/or domain. This section formally describes the metrics calculated by YACHS.

Authors normally conceive titles as circumscribing the topic of the document. Sentences sharing words, containing words related to or similar with the title are likely to be relevant. Following this assumption, two metrics computing similarity measures between a sentence and the title have been implemented. The first measure is the well known cosine angle [10] between a sentence and the title vectorial representations in Γ. The main weakness of *cosine* and more generally of all similarity measures using words for tokens is that they are relying too much on term normalization. Their performance dramatically decrease with wrongly or non normalized words. We propose a second similarity measure based on the Jaro-Winkler distance [13] that can bridge morphologically similar words in order to smooth normalization and misspelling errors. The original Jaro-Winkler measure, denoted JW, uses the number of matching characters and transpositions to compute a similarity score between two terms, giving more favourable ratings to terms that match from the beginning (see examples in Table 2). We have extended this measure to calculate the similarity between a sentence s_x and the title t (see Table 3):

$$\mathrm{JW}_e(s_x, t) = \frac{1}{|t|} \cdot \sum_{w_i \in t} \max_{w_j \in S'} \mathrm{JW}(w_i, w_j) \tag{1}$$

where S' is the term set of s_x in which the terms w_j that already have maximized $\mathrm{JW}(w_i, w_j)$ are removed.

Table 2. Examples of Jaro-Winkler distance (JW) between words

Word 1	**Word 2**	JW
nucleophile	nucleophilic	0.94515
nucleophile	electrophile	0.47643
diphenyl	1,1-Diphenylmethanone	0.35516
1,1-Diphenylmethanone	nucleophile	0.11038

Experiments have shown that sentence position within the document is a very important feature [1]. Indeed, the information is not homogeneously spread across the document but scattered tidily by the author respecting universally accepted writing rules. Document beginnings and endings usually contain sentences that are highly relevant because their original goals are to present and sum up the topic. Sentence position is therefore used as metric, denoted P (Equation 2), by computing a normalized parabolic function depending on the total number of sentence m in the document.

$$P_x = \frac{(x - \lceil \frac{m}{2} \rceil)^2}{\lfloor \frac{m}{2} \rfloor^2} \tag{2}$$

Table 3. Example of similarity measures between the title and a sentence ($\mathbf{T}_{preproc.}$ and $\mathbf{S}_{preproc.}$ are the pre-processed title and the pre-processed sentence)

Title	Generation of Cycloalkynes by Hydro-Iodonio-Elimination of Vinyl Iodonium Salts
Sentence	Cycloalkylidenecarbene can provide a ring-expanded cycloalkyne via 1,2-rearrangement.
$\mathbf{T}_{preproc.}$	*generat* **cycloalkynes** *hydro-iodonio-elimination* **vinyl iodonium salt**
$\mathbf{S}_{preproc.}$	**cycloalkylidenecarbene** *provid ring expand* **cycloalkyne** *via rearrang*
cosine	0 (no co-occurrencies)
JW_e	0.43348

where $\lceil x \rceil$ is the ceiling function that returns the smallest integer not less than x and $\lfloor x \rfloor$ is the floor function that returns the highest integer less than or equal to x.

We have implemented four other metrics relying on numerical treatments, they are computed on the matrix M (previously introduced in section 2.1). The first one is the sum of word frequencies, denoted F (Equation 3), that uses the frequencies of words in sentences. Sentences that are containing a high number of *informative* words (words remaining after pre-processing) are considered relevant.

$$F_x = \sum_{y=1}^{n} a_{x,y} \tag{3}$$

The second metric, denoted C (Equation 4), relies on the number of chemical compounds detected in the sentence giving a penalty to sentences that do not contain any chemical compounds.

$$C_x = \begin{cases} 1 & \text{if } x \text{ contains at least one chemical compound} \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

The third metric, denoted I (Equation 5), represents the interaction relationship between sentences. The underlying idea is that sentences containing words that are used in other sentences are statistically more representative for the document [14].

$$I_x = \sum_{\substack{y=1 \\ a_{x,y} \neq 0}}^{n} \sum_{\substack{z=1 \\ z \neq x}}^{m} a_{z,y} \tag{5}$$

The last metric, denoted H (Equation 6), is the sum of the Hamming distances computed on the sentence pair words [14]. The idea is to give more weight to pairs of words that appears independently in sentences. Synonyms and topic-related words generally are, according to the Hamming distance, high weighted. In order to compute this metric, a second matrix denoted M_h is constructed from M. M_h is a $n \times n$ triangular matrix constructed from word co-occurrences between sentence pairs:

$$
\begin{aligned}
M_h &= [h_{i,j}]_{\ i=1\ldots n;\ j=1\ldots n} \\
h_{i,j} &= \sum_{x=0}^{m} \begin{cases} 1 & \text{if } a_{x,i} \neq a_{x,j} \\ 0 & \text{Otherwise} \end{cases} \\
H_x &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \begin{cases} h_{i,j} & \text{if } a_{x,i} \neq 0 \text{ and } a_{x,j} \neq 0 \\ 0 & \text{Otherwise} \end{cases}
\end{aligned} \tag{6}
$$

Sentences are scored by using a equiprobable linear combination[4] of the normalized metrics (i.e. ranged in $[0, 1]$) described above. A ranked sentences list is produced by the system allowing to construct the extract by arranging the high scored sentences until the desired size is reached.

3 Experimental Settings

Considerable interest has been expressed and effort expended in attempting to evaluate automatically the quality of the summaries. There exists two different types of evaluation: extrinsic and intrinsic [15]. Extrinsic evaluations measure the quality of a summary based on how it affects certain tasks. In intrinsic evaluations, summary's quality is evaluated by an analysis of its content. Most existing automated evaluation methods work by comparing the produced summaries to one or more reference summaries (ideally, produced by humans). In order to evaluate our system, we have collected a testing set from *http://pubs.acs.org*. The testing set is composed by 100 pairs of articles/abstracts coming from different journals (Organic Letters, Accounts of Chemical Research and Journal of Organic Chemistry) of different years (respectively 2000-2002, 2005-2007 and 2007-2008), different authors and topics. Each document has been cleaned up manually from the PDF (or HTML) version (figures, bibliographic references, special characters, etc. have been removed). By ways of comparison the corpus used in the Document Understanding Conference (DUC)[5] 2005 competition was also composed of 100 sets. Table 4 shows some statistics about the testing set.

3.1 Performance Measures

To evaluate the quality of our generated summaries, we choose to use the ROUGE[6] [16] evaluation toolkit, that has been found to be highly correlated with human judgments [17]. ROUGE-N is a N-gram recall measure calculated between a candidate summary and one or more reference summaries. In our experiments ROUGE-1, ROUGE-2 and ROUGE-SU4 will be computed. Each generated extract will be

[4] Other combinations might be considered, but a large training corpus is required to tune the parameters.

[5] Document Understanding Conferences are competitions on text summarization conducted since 2000 by the National Institute of Standards and Technology (NIST), *http://www-nlpir.nist.gov*

[6] ROUGE is available at *http://haydn.isi.edu/ROUGE/*

Table 4. Testing corpus description

Journal	Year	Number	Sentences	Words
Organic Letters	2000-2008	63	5.313	104.588
Accounts of Chemical Research	2005-2006	10	979	18.337
The Journal of Organic Chemistry	2007-2008	27	2.631	66.242
Total	-	100	8.923	189.167

evaluated by comparison with the author's abstract. The size of the produced extracts is set at 5% of the original document (in sentence number) with a minimum of three sentences. This value corresponds to the average compression rate observed on the evaluation corpus (average compression rate is $5,39\%$).

4 Results

The first experiment is focused on the study of metrics. Figure 1 shows the ROUGE results of each metric alone and their combination. As we can see from these results, the combination, denoted by Combi., always outperforms the best metric alone. The most discriminant metrics are the similarity measures with the title (JW_e and *cosine*) and the interaction relationship between sentences (I). The title similarity measures allow to focus the summary on the document main topic, delineated by the author. The similarity measure JW_e that we propose is globally the most discriminant metric, its ability to bridge morphologically similar words is well adapted for Organic Chemistry documents. The interaction metric uses the networks built by words within the document to compute a relevance score, sentences that are constructed with terms appearing in many other sentences are selected. These sentences are judged as being the most representative to the document because they are containing most of the information.

A second evaluation compares YACHS to a generic statistical summarizer and a baseline on the corpus of manually segmented documents (see Figure 2). We use

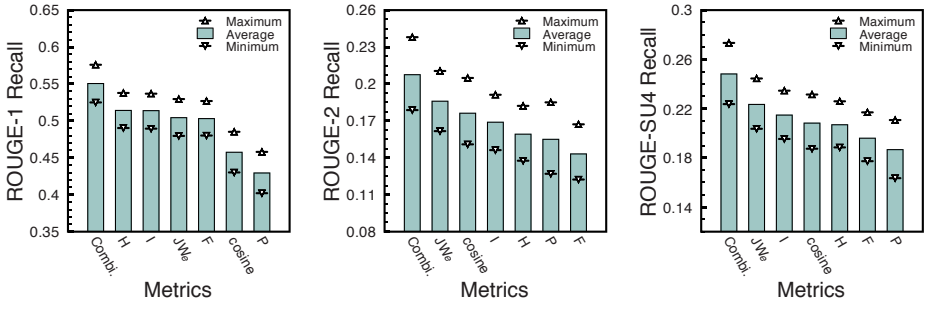

Fig. 1. ROUGE-1, ROUGE-2 and ROUGE-SU4 recall scores for each metric independently and for their combination (denoted Combi.)

the Cortex summarizer [14] which is based on the same approach that YACHS, namely a combination of relevance metrics, but without the chemical compounds detection process and the powerful JW_e metric. The baseline is generated by arranging n sentences selected randomly from the document, n being 5% of the document sentence number with a minimum of three sentences. In order to smooth the baseline results, the average of 100 baseline evaluations is used in our experiments. YACHS achieves the best results among the ROUGE evaluations. It confirms that the specialized pre-processing and sentence scoring are well adapted to process domain specialized (Organic Chemistry) documents.

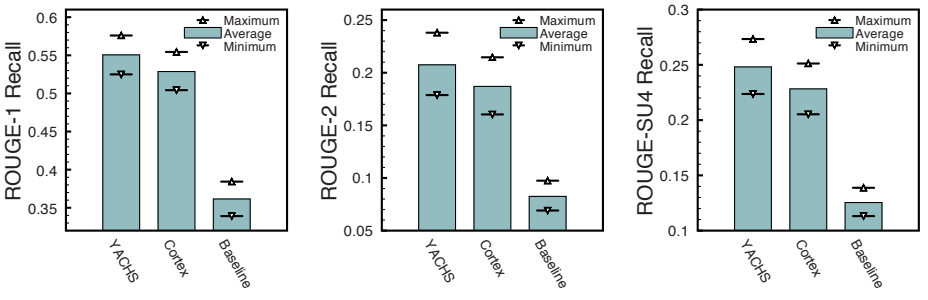

Fig. 2. ROUGE-1, ROUGE-2 and ROUGE-SU4 recall scores of YACHS, Cortex and the random baseline

The last evaluation models a real world summarization task: a plain text is given as input (without manual sentence segmentation), each summarizer has to produce an extract of size equals to 5% of the original document (in sentence number). We compare YACHS to six extractive summarizers and one baseline, results are shown in Figure 3. YACHS, Cortex and the baseline use the same automatic sentence segmentation process which consists in a standard sentence boundaries detection system enriched with lists of abbreviations. The other systems using their own sentence splitters. The baseline is generated by arranging n sentences selected randomly from the document, n being 5% of the document sentence number with a minimum of three sentences. Again, the average of 100 baseline evaluations is used in our experiments. MEAD[7] is a centroid based summarizer [18] that extract sentences according to three features: sentence centrality within the cluster, sentence position within the document and weighted similarity with the title. Open Text Summarizer[8] (OTS) [19] is an Open Source project that, similarly to MEAD, use statistical word-frequency methods to score sentences that are beforehand parsed. It also incorporates an English language lexicon with synonyms and cue terms. Pertinence Summarizer[9] performs linguistic processing of a document to generates an extract,

[7] Available at *http://www.summarization.com/mead/*
[8] Available at *http://libots.sourceforge.net*
[9] Available at *http://www.pertinence.net/ps/*

the sentence scoring method considering general and specialized (Chemistry) linguistic markers. Besides, two frequency-based summarizers are evaluated: Copernic[10] summarizer and the AutoSummarize feature of Microsoft Word. Exact details of their algorithms are unfortunately not documented.

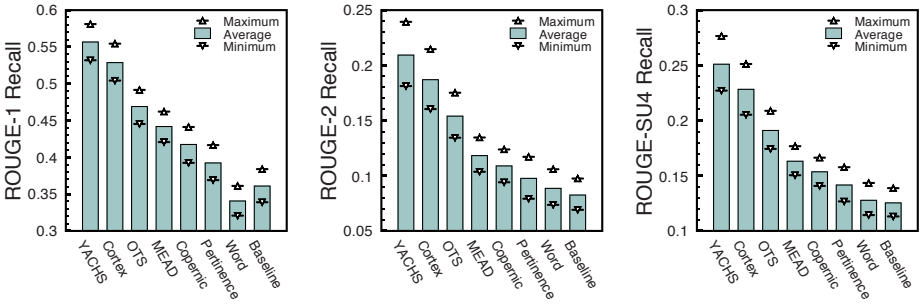

Fig. 3. Comparison of the ROUGE-1, ROUGE-2 and ROUGE-SU4 recall scores for the seven summarizers and the random baseline

YACHS and Cortex clearly stand out from the crowd. These two methods perform significantly better than the other systems (and the baseline) confirming that these statistical techniques work well for Organic Chemistry documents. YACHS achieves the best results among all summarizers proving that specialized pre-processing and adapted sentence scoring are features allowing to generate better specialized extracts.

5 Conclusion

In this paper we have described an efficient approach for automatically generating extracts from documents in Organic Chemistry. Through experiments performed on a corpus composed of scientific articles, we have showed that our approach (implemented in the YACHS[11] system) achieves promising results. This work represent a good starting point but do show a critical point: a lot of information is lost during document pre-processing. Indeed, pictures, tables or captions, that are removed during PDF (or HTML) to text conversion, are containing salient information that can be used to enhance extracts. Among the others, there are several points that would be worthy of further investigation:

- Use multi-media information (pictures, texts, tables, etc.) to generate extracts.
- Fuse text summarization and Question Answering (QA) to model real-world complex QA, in which a question cannot be answered by simply stating a name, date, quantity, etc.

[10] Available at *http://www.copernic.com/en/products/summarizer/index.html*

[11] An demonstration version of YACHS is available at `http://daniel.iut.univ-metz.fr/yachs`

Acknowledgement

We are grateful to Pr. Alain Krief and Julie Henry for our useful talks. This work was partially supported by the *Laboratoire de chimie organique de synthèse*, FUNDP (*Facultés Universitaires Notre-Dame de la Paix*), Namur, Belgium and by the *Agence Nationale de la Recherche*, France, project RPM2.

References

1. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)
2. Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2(2), 159 (1958)
3. Climenson, W.D., Hardwick, N.H., Jacobson, S.N.: Automatic Syntax Analysis in Machine Indexing and Abstracting. American Documentation 12(3), 178–183 (1961)
4. Edmundson, H.P.: New Methods in Automatic Extracting. Journal of the ACM (JACM) 16(2), 264–285 (1969)
5. Pollock, J.J., Zamora, A.: Automatic Abstracting Research at Chemical Abstracts Service. Journal of Chemical Information and Computer Sciences 15(4), 226–232 (1975)
6. Kupiec, J., Pedersen, J., Chen, F.: A Trainable Document Summarizer. In: 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 68–73. ACM Press, New York (1995)
7. Mani, I., Bloedorn, E.: Machine Learning of Generic and User-focused Summarization. In: 15th National Conference on Artificial intelligence (AAAI), pp. 820–826. AAAI Press, Menlo Park (1998)
8. Teufel, S., Moens, M.: Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. Computational Linguistics 28(4), 409–445 (2002)
9. Reeve, L.H., Han, H., Brooks, A.D.: The use of Domain-Specific Concepts in Biomedical Text Summarization. Information Processing and Management 43(6), 1765–1776 (2007)
10. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM 18(11), 613–620 (1975)
11. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14, 130–137 (1980)
12. Boudin, F., Torres-Moreno, J.M.: Mixing Statistical and Symbolic Approaches for Chemical Names Recognition. In: Gelbukh, A. (ed.) CICLing 2008. LNCS, vol. 4919, pp. 334–349. Springer, Heidelberg (2008)
13. Winkler, W.E.: The State of Record Linkage and Current Research Problems. Statistics of Income Division 4, 73–79 (1999)
14. Torres-Moreno, J.M., Velazquez-Morales, P., Meunier, J.G.: Condensés de textes par des méthodes numériques. In: Journées internationales d'Analyse statistique des Données Textuelles (JADT), vol. 2, pp. 723–734 (2002)
15. Spärck Jones, K., Galliers, J.R.: Evaluating Natural Language Processing Systems: An Analysis and Review. Springer, Heidelberg (1996)
16. Lin, C.Y.: Rouge: A Package for Automatic Evaluation of Summaries. In: Workshop on Text Summarization Branches Out, pp. 25–26 (2004)

17. Dang, H.T.: Overview of DUC 2005. In: Document Understanding Conference (DUC) (2005)
18. Radev, D.R., Blair-Goldensohn, S., Zhang, Z.: Experiments in Single and Multi-Document Summarization Using MEAD. In: Document Understanding Conference (DUC) (2001)
19. Yatsko, V.A., Vishnyakov, T.N.: A Method for Evaluating Modern Systems of Automatic Text Summarization. Automatic Documentation and Mathematical Linguistics 41(3), 93–103 (2007)