Experimenting with Maximal Frequent Sequences for Multi-Document Summarization*

Yulia Ledeneva, René Arnulfo García-Hernández,
Anabel Vazquez-Ferreyra, Nayely Osorio de Jesús

Autonomous University of the State of Mexico
Santiago Tianguistenco, México
ynledeneva@uaemex.mx
renearnulfo@hotmail.com

Abstract. In this work we address the task of multi-document text summarization which consists in reducing the contents of a collection of documents to a single short text, so that the user can judge about the contents of the whole collection upon reading only this short text. The required short text is composed of entire sentences selected from the documents of the collection which is called extractive summarization task. This task consists of weighting some sentences from the whole collection. We improve a language-independent method of weighting the sentences of the document collection looking for new terms which permits to improve results for multi-document task. Our results show that this method is among the best systems in the existing language-independent state-of-the-art methods.

1 Introduction

The text summarization tasks can be classified into single-document and multi-document summarization. In multi-document summarization, the summary of a whole collection of documents is built, in contrast to single-document summarization where the summary of only one document is to be built. So, a summary of a collection of documents is a single short text that contains the most important information from this collection of documents.

There are extractive and abstractive summaries. An abstractive summary is a short text that describes in different words the contexts of the source collection of documents [1]. Abstractive summarization process consists of re-phrasing the original collection of documents in fewer sentences. Usually, an abstractive summarization method uses linguistic methods to re-phrase and re-write documents. While this may seem better alternative to obtain a summary; really, in state-of-the-art, the abstractive methods offer worse quality than extractive methods.

An extractive summary is a selection of sentences or even paragraphs from the original collection of documents. It means, an extractive summarization method only decides, for each sentence, whether or not it will be included in the summary, and

* Work done under partial support of Mexican Government (CONACyT, SNI, PROMEP, SEP). The authors thank UAEM for its assistance.

Received 30/04/10
Final version 19/05/10
Accepted 14/05/10

then, it is usually presents to the user in the same order as in the source document collection. The resulting summary is read it too awkward; however, simplicity of the extractive summarization methods offers an attractive, robust and language-independent option, in contrast to more complicated abstractive methods.

A typical extractive summarization method consists in several steps [2], at each of them different options can be chosen. We will assume that the units of selection are sentences. Thus final goal of the extractive summarization process is sentence selection.

One of the ways to select the appropriate sentences is to assign some numerical measure of usefulness of a sentence for the summary and then select the best ones; the process of assigning these usefulness weights is called sentence weighting.

One of the ways to estimate the usefulness of a sentence is to sum up usefulness weights of individual terms of which the sentence consists; the process of estimating the individual terms is called term weighting. For this, one should decide what the terms are: for example, they can be words; deciding what objects will count as terms is the task of term selection. Different extractive summarization methods can be characterized by how they perform these steps.

For term selection step, several works employ domain-dependent terms like key-phrases, lexical chains, proper names and anaphors; and other ones employ word-based terms like words or n-grams [3-11]. Word-based terms have the advantage of depend less or nothing on the domain, since they are extracted from the source text to be summarized. Recently, the so-called Maximal Frequent Sequences (MFSs) have shown to be good terms for single-document extractive summarization with advantages over n-grams since each MFS only depends on the structure of the text, preserving more the sequential nature of the text [2, 12-14]. In each MFS not only is recovering the frequency of the term but also the size (number of words) of the term which provide more information about the importance of such term. Also, a maximal frequent sequence is a meaningful term since all its sequences are frequent too.

It is very useful the automatic construction of summaries for different collection of documents. For example, summaries of collection of news articles, sport reviews, research papers, description of products, politic debates, information from web pages, all today's news or all search results for a query. Such summaries present brief information about different points of view, opinions, sources of knowledge, controversial or agreement discussions on any event, fact, topic, history, book, web page, etc. It will help to the user not only quickly find the necessary information, but also understand what is the overall situation on a specific topic.

One possible scenario is when a researcher tries to understand a specific science area; in this case, the summary will return short definitions about this area. Another scenario, for example, when we want to know what books are in the library of a particular author. Given the name of the author, will be very interesting not only to see the names of the books of this author but also a short description of each book.

In this paper, we are experimenting with MFSs for multi-document task in order to improve the quality of extractive summaries.

The paper is organized as follows. Section 2 summarizes the state-of-the-art of multi-document summarization methods. In Section 3, some notions used for term selection in our method are introduced. Section 4 presents our experimental setting. Sections 5 and 6 describe the obtained experimental results for different term selection and term weighting schemes, respectively, which are compared in Section 7 with those of existing methods. Section 8 concludes the paper.

2 Related Work

One of the most competitive methods of the-state-of-the-art methods are presented by Mihalcea [15] in the form of a clear graph-based formalism. In this method, the words that have closer relationships with a greater number of "important" words become more important themselves, the importance being determined in a recursive way similar to the PageRank algorithm used by Google to weight web pages. The latter idea can be applied directly to sentence weighting without term weighting: a sentence is important if it is related to many important sentences, where relatedness can be understood as, say, overlap of the lexical contents of the sentences [15]. The two methods presented in [15, 16] are those that currently give the best results and with which we compare our suggested method.

Another relevant works are developed by Wei *et al.* [17] where the relevance of a term is derived from an ontology constructed with formal concept analysis. Song *et al.* [3] weight a word calculating the number of lexical connections, such as semantic associations expressed in a thesaurus that the word has with its neighboring words; along with this, more frequent words are weighted higher. Passages are retrieved using a language model [18] with the objective to predict the probability of word sequences actually occur and low probability on word sequences that never occur. The n-gram model is used as a basis for the proposed language model.

A special procedure is designed by Nenkova et al. [19, 20] for comparative analysis of the content of several texts. In these works special terms are annotated using the pyramid scheme. The presence of each term in all documents of the collection accumulates the importance of this term. The more documents have the term, the more important is this term, and consequently will be included in the summary.

One of the most implemented sentence selection methods are supervised learning methods which consider sentence selection as a classification task. These methods train a classifier using a collection of documents supplied with existing summaries. As features of a sentence such methods can consider text units (in which case we can speak of term selection) or other, non-lexical characteristics. Different lexical and non-lexical features have been used in [21-23]. Most of these features are "heuristically motivated", since they tend to emulate the manual creation of extracts.

In a work of Kupiec [21], the following features were proposed: sentence position, sentence length, the presence of key phrases and overlap with the title of the document. More recent works [22, 23] extend these features incorporating information about the occurrence of proper names and the presence of anaphors. The "heuristically motivated" features allow extract very precise summaries. However, they have a very big disadvantage of being highly linked to a specific domain. This condition implies that the change for one domain to another, it may be necessary to redefine or even eliminate some features. For instance, key phrases, which are particular for each domain, require being modified, while the overlap with the title, which has no sense in all topics, may be eliminated.

In order to increase the domain (and language) independence of machine learning summaries, Villatoro [24] eliminates all kind of "heuristically motivated" attributes and substitute them by word-based features. In particular, he uses word sequences (n-grams) as terms. Although the first attempt to use n-grams is exceeded the results of other methods, it has some disadvantages. One is that they are always sequences of a fixed size, which was previously defined by the user. The big part of the problem in such techniques lies in defining the size of the sequence to be extracted, which usually depends on the analysis of the text.

A very old and very simple sentence weighting heuristic in single-document summarization does not involve any terms at all: it assigns highest weight to the first sentences of the text. Texts of some genres—such as news reports or scientific papers—are specifically designed for this heuristic: *e.g.*, any scientific paper contains a ready summary at the beginning. This gives a baseline in single-document summarization [25] that proves to be very hard to beat on such texts. Similar to this heuristic in multi-document summarization is employed the first sentence of each document to conform the summary, each sentence is added until to reach the desired length [26]. However, comparing term-based methods with such position-based baseline is not fair in the sense that this baseline only works on text of specific genres and uses information (the position of the sentence) not available to term-based methods. It is worth noting that in Document Understanding Conference (DUC) competitions [25] only five systems performed above this baseline, which does not demerit the other systems because this baseline is genre-specific.

In a previous work [27], we analyzed several options for simple language-independent statistical term selection and corresponding term weighting, based on units larger than one word. In particular, we showed that so-called MFSs, as well as single words that are part of bigrams repeated more than once in the text, are good terms to describe a collection of documents. In this paper, we experiment with some minimum-frequency thresholds of MFSs in order to improve the quality of the multi-document summaries.

3 Proposed Method

The research on multi-document summarization is less developed than single-document summarization because summarizing a collection of thematically related documents is more difficult, than summarizing a single text. In order to avoid repetitions, one has to identify and locate thematic overlaps. In other words, has to locate the most important terms for representing a collection of documents. One also has to decide what sentences are more important, and to arrange events from various sources along a single timeline.

One of our hypotheses is that MFSs with higher threshold should generate summaries with better quality than MFSs with lower threshold. It can be explained on reason that there would exist in the language a multiword expression that can express the same content in the more compact way which can be detected more precisely using higher threshold (see Experiment 1).

Another hypothesis is, in contrast to MFSs, FSs is important if there would exist in the language a single word or at least an abbreviation to express it. Such single words or abbreviations should be considered as bearing the more important meaning with lower threshold because we need to extract more single words or abbreviations to know if they can be used for composing a summary (see Experiment 2)..

The third hypotheses we explore in this work, is that MFSs represent in a better way the summarized content of collection of documents than FSs because their (MFSs) probability to bear important meaning is higher. It can happen because there are too many non-maximal FSs in comparison to MFSs (see Experiment 3).

Our proposed method followed the sequence of steps as follows:

1. Term selection: the main idea is to utilize the MFSs and its derived FS as the main terms.
    1.1 Frequency threshold
2. Term weighting: each MFS can be weight based in its frequency or in its length;
3. Sentence weighting: we test the option calculating the sum of the weights of the terms contained in the sentence.
4. Sentence selection: we test the option when the sentences with greater weight are selected until the desired size of the summary (100 words) is reached.

3.1 Term Selection

An n-gram is a sequence of *n* words. We say that an n-gram occurs in a text if these words appear in the text in the same order immediately one after another. We call an n-gram frequent, if it occurs more than β times in the text, where β is a predefined threshold. Frequent n-grams—we will also call them frequent sequences (FSs)—often bear important semantic meaning: they can be multiword expressions (named entities: *The United States of America*, idioms: *kick the basket*) or otherwise refer to some idea important for the text (*the President's speech*, *to protest against the war*).

An n-gram can be a part of another, longer n-gram. All n-grams contained in an FS are also FSs, for example, the if MFS is *The United States of America* then *The United States or States of America* is a FS too. In this case, *The United States* tends to be synonymous to the longer expression, and the author of one document would choose one or another way to refer to the same entity. FSs that are not parts of any other FS are called Maximal Frequent Sequences (MFSs) [28, 29]. For example, in the following collection of documents

D1: ... *Mona Lisa <u>is the most beautiful</u> picture of Leonardo da Vinci ...*
D2: ... *Eiffel tower <u>is the most beautiful</u> tower ...*
D3: ... *St. Petersburg <u>is the most beautiful</u> city of Russia ...*
D4: ... *<u>The most beautiful</u> church is not located in Europe ...*

the only MFS with $\beta = 3$ is *is the most beautiful*, while the only MFS $\beta = 4$ is *the most beautiful* (it is not an MFS with $\beta = 3$ since it is not maximal with this β). As this example shows, the sets of MFSs with different thresholds do not have to, say, contain one another.

The notions of FSs and MFSs are closely related to that of repeating bigrams. This set is conceptually simpler, but for computational implementation MFSs could be more compact.

For term selection, we compared MFSs with more traditional word-based features such as single words and n-grams. Namely, we considered the following variants of term selection**:**

– *M*: the set of all MFSs with some threshold *ß*. In the example from Section 3, *M*={*is the most beautiful*}. Also, we denote by M_2 the set of all MFSs with *ß* = 2.
– *W*: single words (unigrams) from elements of M. In our example, *W*={*is*, *the*, *most*, *beautiful*}.

Optionally, stop-words were eliminated at the pre-processing stage; in this case our bigrams (or MFSs) could span more words in the original text.

3.2 Term Weighting, Sentence Weighting and Sentence Selection

For term weighting, the frequency of the term was used; for sentence weighting, the sum of the weights of the terms contained in the sentence was used; for sentence selection, the sentences with greater weight were selected until the desired size of the summary (100 words) is reached.

Optionally, stop-words were eliminated at the pre-processing stage. For term weighting, different formulae were considered containing the following values:

– *f*: frequency of the term in MFSs, *i.e.*, the number of times the term occurs in the text within some MFS. In our example, *f*(*is*) = 3 since it occurs 3 times in the text within the MFS *is the most beautiful*. Under certain realistic conditions (MFSs do not intersect in the text, words do not repeat within one MFS) *f* is the number of times the term occurs in the text as part of a repeating bigram. In our example, *f*(*is*) = 3 since it occurs 3 times in a repeating bigram *is the* (and one time in a non-repeating context *church is not*).
– l: the maximum length of an MFS containing the term. In our example, l(is) = 4 since it is contained in a 4-word MFS is the most beautiful.
– 1: the same weight for all terms.

For sentence weighting, the sum of the weights of the terms contained in the sentence was used. For sentence selection, the following options were considered:

– best: sentences with greater weight were selected until the desired size of the summary (100 words) is reached. This is the most standard method.

4 Experimental Setting

We realized several experiments in order to verify our hypotheses formulated in the previous section. The specific settings for each step varied between the experiments and are explained below for each experiment.

Test data set. We used the DUC collection provided [25]. In particular, we used the data set of 60 document collections which consist of 567 news articles of different length and with different topics. Each collection of documents in the DUC collection is supplied with a set of human-generated summaries provided by two different experts. While each expert was asked to generate summaries of different length, we used only the 100-word variants.

Evaluation procedure. We used the ROUGE evaluation toolkit [30] which was found to highly correlate with human judgments [31]. It compares the summaries generated by the program with the human-generated (gold standard) summaries. For comparison, it uses n-gram statistics. Our evaluation was done using n-gram (1, 1) setting of ROUGE, which was found to have the highest correlation with human judgments, namely, at a confidence level of 95%.

As a kind of statistical significance check, we randomly divided our test data into two halves and ran this (and most of the other) experiments separately on each subset. These experiments confirmed the qualitative observations reported in this paper.

Previous results. We tried term selection options, such as M and W, with the term weighting option 1, l, the options related to f, and their combination (Table 1). For sentence selection, we tried the *best* combination. Term selection W gave a slightly better result than M. The best results are highlighted in boldface. (See more details in [27]. We borrow this table from [27] to compare with Tables 2-4).

Table 1. Results for different term selection and term weighting options for multi-document summarization.

Term Selection	Term Weighting	Sentence Selection	Results		
			Recall	Precision	F-measure
M	F	best	0.31372	0.31986	0.31660
	f^2		0.31162	0.31870	0.31499
	1		0.30620	0.31347	0.30965
	L		**0.31411**	**0.32199**	**0.31786**
	l^2		0.29184	0.30275	0.29706
	$f \times l$		0.31329	0.32103	0.31696
	$f \times\times l$		0.28328	0.29592	0.28933
W	F	best	**0.31919**	**0.32494**	**0.32192**
	1		0.26413	0.27828	0.27072
	f^2		0.30056	0.30764	0.30391

Experiment 1. For this experiment, we use the configuration of the algorithm of previous results (see Table 1) of this section. Then we tested the algorithm with $\beta = 2$ (see Table 2).

Table 2. Results for different term selection and term weighting options with $\beta = 2$.

Term Selection	Term Weighting	Sentence Selection	Results		
			Recall	Precision	F-measure
M	*F*	best	0.29038	0.29749	0.29373
	f^2		0.29038	0.29749	0.29373
	1		0.29038	0.29749	0.29373
	L		**0.29881**	**0.30714**	**0.30279**
	l^2		0.28837	0.29912	0.29351
	$f \times l$		**0.29881**	**0.30714**	**0.30279**
	$f \times\times l$		0.28455	0.29606	0.29006
W	*F*	best	**0.32238**	**0.32902**	**0.32557**
	1		0.26413	0.27828	0.27072
	f^2		0.29559	0.30361	0.29796

Experiment 2. For this experiment, we use the configuration of the algorithm of previous results (see Table 1) of this section. Then we tested the algorithm with $\beta = 3$ (see Table 3).

Table 3. Results for different term selection and term weighting options with $\beta = 3$.

Term Selection	Term Weighting	Sentence Selection	Results		
			Recall	Precision	F-measure
M	*F*	best	0.30601	0.31302	0.30933
	f^2		0.30601	0.31302	0.30933
	1		0.30601	0.31302	0.30933
	L		**0.31574**	**0.32286**	**0.31911**
	l^2		0.29904	0.30852	0.30356
	$f \times l$		**0.31574**	**0.32286**	**0.31911**
	$f \times\times l$		0.28309	0.29270	0.28768
W	*F*	best	**0.32279**	**0.32826**	**0.32538**
	1		0.26413	0.27828	0.27072
	f^2		0.30934	0.31602	0.31253

Experiment 3. For this experiment, we use the configuration of the algorithm of previous results (see Table 1) of this section. Then we tested the algorithm with $\beta = 4$ (see Table 4).Comparison of the results for the proposed method is shown in Tables 5 and 6. The results of the state-of-the-art methods [25] for multi summarization are shown in Table 7.

Table 4. Results for different term selection and term weighting options with $\beta = 4$.

Term Selectio n	Term Weight-ing	Sentence Selection	Results		
			Recall	Precision	F-measure
M	*F*	best	0.30964	0.31648	0.31288
	f^2		0.30964	0.31648	0.31288
	1		0.30964	0.31648	0.31288
	L		**0.32326**	**0.32980**	**0.32636**
	l^2		0.30825	0.31812	0.31296
	$f \times l$		**0.32326**	**0.32980**	**0.32636**
	$f \times\times l$		0.29651	0.30533	0.30079
W	*F*	best	**0.31855**	**0.32329**	**0.32076**
	1		0.26413	0.27828	0.27072
	f^2		0.30934	0.31602	0.31253

5 Conclusions

We observed that MFSs with higher threshold generate summaries with better quality than MFSs with lower threshold. It can be explained on reason that there exist in the language multiword expressions that can express the same content in the more compact way which can be detected more precisely using higher (see Table 5).

Then, we observed that, in contrast to MFSs, FSs is important if are extracted with lower threshold. It can be explained because there exist in the language a lot of single word or at least an abbreviation to express an important meaning.

Such single words or abbreviations should be considered as bearing the more important meaning with lower threshold because we need to extract more single words or abbreviations for knowing if they can be used for composing a summary (see Experiment 2).

The third hypotheses we explore in this work, is that MFSs represent in a better way the summarized content of collection of documents than FSs because their (MFSs) probability to bear important meaning is higher. It can happen because there are too many non-maximal FSs in comparison to MFSs (see Experiment 3).

Table 5. Comparison of results using different thresholds (terms are MFS).

Method	Recall	Precision	F-measure
M where $\beta = 2, 3, 4$	0.31411	0.32199	0.31786
M where $\beta = 2$	0.29881	0.30714	0.30279
M where $\beta = 3$	0.31574	0.32286	0.31911
M where $\beta = 4$	**0.32326**	**0.32980**	**0.32636**

Table 6. Comparison of results using different thresholds (terms derived from MFS).

Method	Recall	Precision	F-measure
W where $\beta = 2, 3, 4$	0.31919	0.32494	0.32192
W where $\beta = 2$	**0.32238**	**0.32902**	**0.32557**
W where $\beta = 3$	0.32279	0.32826	0.32538
W where $\beta = 4$	0.31855	0.32329	0.32076

We compared the following results (see Table 7):

– **State of the art:** The best top 5 systems from 17 systems in DUC 2002 for multi-document summarization task are listed in Table 7 [25].

– **Baseline:** DUC collection has a configuration denotes as *Baseline,* which selects the first sentence in the first, second, third, and so on document in chronological sequence until you have the target summary size [25]. This baseline gives good results on the kind of texts (news reports) that we experimented with. Thus we can compare with this configuration of baseline. Also we believe this configuration to be a more realistic baseline for the types of texts.
– **Recent work:** As shown in Table 1 (see [27]), using the configuration the W term selection scheme and the f term weighting scheme. We call it as the 4th best method.
– **Our proposal:** We compare these methods with the best results obtained in this paper: *M* term selection scheme with $\beta = 4$ and the *l* term weighting scheme, as shown in Table 5. W term selection scheme with $\beta = 2$ and the *f* term weighting scheme, as shown in Table 6.

We tested new method for the automatic generation of text summaries for a multi-document summarization based on the discovery of MFSs, specifically we tested different combinations of term selection, term weighting, sentence weighting and sentence selection schemes with different thresholds. Comparing to other methods, we did not receive the best results but considering that the proposed method is language- and domain-independent, we think that the results are very encouraging. Also we improve the obtained results using MFSs (see 4th best method and proposed in Table 7).

Table 7. Comparison of results with other methods.

Method	F-measure
1st best method	0.3578
2nd best method	0.3447
Best proposed	**0.3264**
3rd best method	0.3264
4th best method	0.3219
5th best method	0.3056
6th best method	0.3047
Baseline	0.2932

References

1. Lin, C.Y. and Hovy, E. Automated Text Summarization in SUMMARIST. In Proc. of ACL Workshop on Intelligent, Scalable Text Summarization, Madrid, Spain, 1997.
2. Ledeneva Y., Gelbukh A., García-Hernández R. Terms Derived from Frequent Sequences for Extractive Text Summarization. LNCS 4919, pp. 593-604, Israel, Springer-Verlag, ISSN 0302-9743, 2008.
3. Song, Y., et al. A Term Weighting Method based on Lexical Chain for Automatic Summarization. CICLing 2004, LNCS, vol. 3878, Springer-Verlag 2004.
4. Cristea D., et al. Summarization through Discourse Structure. CICLing 2005, LNCS, vol. 3878, Springer-Verlag 2005.
5. Liu, D., He, Y., Ji, D., Hua, J. Multi-Document Summarization Based on BE-Vector Clustering. CICLing 2006, LNCS, vol. 3878, Springer-Verlag 2006.
6. Xu, W., Li, W., et al. Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis. CICLing 2006, LNCS, vol. 3878, Springer-Verlag 2006.
7. HaCohen-Kerner, Y., Zuriel, G., Asaf, M. Automatic Extraction and Learning of Keyphrases from Scientific Articles. LNCS, vol. 3878, pp. 645–657, Springer-Verlag 2005.
8. Gelbukh A., Sidorov G., Sang Yong Han, Hernandez-Rubio E. Automatic Enrichment of Very Large Dictionary of Word Combinations on the Basis of Dependency Formalism. In: Raúl Monroy, et al. (Eds.). ISSN 0302-9743, vol. 2972, pp. 430–437, 2004.
9. Sidorov, G., Gelbukh A., Procesamiento Automático del Español con enfoque en recursos léxicos grandes. IPN, ISBN 970-36-0264-9, 2006.
10. Bolshakov, I.A. Getting One's First Million…Collocations. CICLing-2004, Seoul, Korea, LNCS 2945, p. 229-242, Springer-Verlag 2004.
11. Filippova K., Mieskes M., Nastase V., et al. Cascaded Filtering for Topic-Driven Multi-Document Summarization. Proc. of Document Understanding Conference 2007. http://duc.nist.gov/pubs.html#2007.
12. Ledeneva Y. Effect of Preprocessing on Extractive Summarization with Maximal Frequent Sequences. LNAI 5317, pp. 123-132, Springer-Verlag, ISSN 0302-9743, 2008.
13. García-Hernández R., Ledeneva Y., Gelbukh A., Rendon E., Cruz R. Text Summarization by Sentence Extraction Using Unsupervised methods. LNAI 5317, pp. 133-143, Mexico, Springer-Verlag, ISSN 0302-9743, 2008.
14. Ledeneva Y., Gelbukh A., García-Hernández R. Keeping Maximal Frequent Sequences Facilitates Extractive Summarization. In: G. Sidorov et al (Eds). Advances in Computer Science and Engineering, Research in Computing Science, vol. 34, pp. 163-174, ISSN 1870-4069, 2008.
15. Mihalcea, R. Random Walks on Text Structures. CICLing 2006, LNCS, vol. 3878, pp. 249–262, Springer-Verlag 2006.
16. Mihalcea, R., Tarau, P. TextRank: Bringing Order into Texts, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, 2004.
17. Xu Wei, Li Wenjie, et al. Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis. CICLing 2006, LNCS, vol. 3878, Springer-Verlag, pp. 480–489, 2006.
18. Ying J.-C., Yen S.-J., Lee Y.-S. Language Model Passage Retrieval for Question-Oriented Multi Document Summarization. Proc. of Document Understanding Conference 2007. http://duc.nist.gov/pubs.html#2007.
19. Nenkova A. Understanding the process of multi-document summarization: content selection, rewriting and evaluation. Ph.D. Thesis, Columbia University, 2006.
20. Nenkova A., Passonneau R. Evaluating content selection in summarization: The pyramid method. In: Proc. of NLT/NAACL–2004, 2004.

21. Kupiec, J., Pedersen, J.O., Chen, F. A Trainable Document Summarizer. In Proceedings of the 18th ACM-SIGIR Conference on Research and Development in Information Retrieval. Seattle, pp. 68–73, 1995.
22. Chuang T.W., Yang J. Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms. Proc. of the ACL-04 Workshop. Barcelona, España, 2004.
23. Neto L., Freitas A., et al. Automatic Text Summarization using a Machine Learning Approach. Proc. of the ACL Workshop, España, 2004.
24. Villatoro-Tello, E., Villaseñor-Pineda, L., Montes-y-Gómez, M. Using Word Sequences for Text Summarization. TSD, LNAI Springer 2006.
25. DUC. Document understanding conference 2002; www-nlpir.nist.gov/projects/duc.
26. Lin Chin-Yew and Eduard Hovy, From Single to Multi-document Summatization: A Prototype System and its Evaluation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic (ACL), 2002. pp.457-464.
27. Ledeneva, Y., García-Hernández, R.A., Gelbukh A. Multi-document Summarization using Maximal Frequent Sequences. Research in Computer Science, pp.15-24, vol. 47, ISSN 1870-4069, 2010.
28. García-Hernández, R.A., Martínez-Trinidad J. F., Carrasco-Ochoa J. A. A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text, CIARP’2004, LNCS vol. 3287 Springer-Verlag 2004. pp. 478–486.
29. García-Hernández, R.A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A. A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection, CICLing’2006, LNCS vol. 3878 Springer-Verlag 2006. pp. 514–523.
30. Lin C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization of ACL, Spain, 2004.
31. Lin C.Y., Hovy E. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In Proceedings of HLT-NAACL, Canada, 2003.