The Bulgarian Summaries Corpus

Viktoriya Petrova
Bulgarian Academy of Sciences
v.k.petrova@abv.bg

Abstract

This article aims to present the Bulgarian Summaries Corpus, its advantages, its purpose and why it is necessary. It explains the selection of texts and process of summarization and the tool used, in addition of a quick overview of the current situation in Bulgaria. The paper also presents a general outline of the market needs, the use of this kind of tools and a short list of examples of a variety of corpora around the world both in language and field.

1. Introduction

Web content[1] has become a science with a list of new jobs[2] because of its growing importance. This has increased the volume of information available online and with it the need of its quick processing in a rapid and effective way. The necessity of extracting the most valuable parts of documents has also grown slightly, although papers and studies in this direction have been written for more than twenty years. Since more information is becoming available, more tools are needed to handle it (Mani, I. and Maybury, M.T. 1999). Some summarization-related technologies attracted substantial investment companies.

Universally, a "summary" is to be understood as a text that is produced from another bigger text, and that conveys the most important information from the original text. It should be no longer than half of the original.

Nowadays summarization is applied in multiple areas: from scientific articles to web pages content, to the creation of large and especially designed corpora. They all adopt different methods and techniques, such as deleting textual units that are considered unimportant for the main message (it often happens by using a discursive structure of text) or the structure trees that compute different segments of the text or sentence compression (consists in removing lexical units that are not important enough in the sentence to change or distort its main meaning).

2. Corpora around the world

A large variety of summarization corpora has been developed. Each of them stresses on a particular point of what the texts can be used for: length of the document, interpretation of the text (especially in

[1] Even though it is generally divided into textual, visual and aural, the focus in this paper is only on the first one and will be understood as such in the entire paper.
[2] Web content writer, Web content Manager, Web content Editor etc.

Keywords: Bulgarian Summaries Corpus, texts selection, summarization, tools

the area of politics), whether they are multi or monolingual, or are related to a particular area. Various examples are:

- The Japanese Text Summarization Corpus, especially developed to be able to judge the credibility of the information collected on the web. Another purpose is also the preparation of gold standard data to evaluate smaller sub-processes within the extraction and summary generation process, and the investigation of the summaries made by human summarizers (Nakano M., Shibuki H and al., 2010).
- The composed entirely on French Human Reference Corpus for Multi-Document Summarization and Sentence Compression whose purpose is the development of automatic methods for multi-document summarization including text, audio and video.
- The multi-document multilingual summarization corpus made for Arabic, English, Greek, Chinese, Romanian and others, whose aim was to evaluate a series of language-independent algorithms and the problem of summarizing news topics.
- TweetMotif, a tool created specifically for the search and topic summarization of Twitter messages.
- The Polish Summaries Corpus, created for the support of the tools for automated single-document summarization of texts in the Polish language.

3. Origins and purpose of the corpus

The Bulgarian Summaries Corpus was created under the guidance of the Institute for Bulgarian Language of the Bulgarian Academy of Science. It is the first corpus of its kind in the country and it is a part of the Bulgarian National Corpus[3]. The aim is not to be left behind the rest of the world, and to help in the application of the different linguistic areas and other research purposes. It is also expected to become a resource of the Bulgarian language on the internet, especially since it represents a peripheral language[4].

4. Text selection

When choosing texts, the type of the corpus that has to be taken into consideration. In some cases randomly selected documents are acceptable, but in others they are not. Due to its particular nature, a variety of articles was selected for the Bulgarian Summary Corpus. They cover different journalistic domains and a large variety of styles. The main subjects vary between political analysis and newspaper articles, followed by health issues, diseases and their possible cures. The documents were subjected to an additional filter, where interviews and files with more than one text inside were deleted. In this way, every text was put on a different file.

The texts are divided in two main groups:

- Texts containing 1000 to 1999 words;
- Texts containing 2000 to 2999 words.

After the process of summarization is completed, two more files are created – one made up by sentences and another containing only the main closes from the file with the sentences. In this way the total number of files is 3. When there are both computer and human summarizations of texts, it is possible to compare the results from the machine and the people – it is a process similar in some extent

[3] It was developed between 2001 and 2009, with over 240 000 text samples. Access to the corpus: http://dcl.bas.bg/bulnc/en/

[4] According to the The Global Language System of de Swaan and its hierarchy of four levels (the peripheral, central, supercentral and hypercentral languages) the linguistic dimension of the world goes hand in hand with the political and economic aspects. The present global situation of languages is the product of prior conquest and domination and of ongoing relations of power and exchange. “Peripheral languages” are 98% of the world's languages and spoken by less than 10% of the world’s population.

to the machine translation evaluation. The only difference here is that there are no translations, but purposely omitted parts of a text.

5. Summarization process

Normally, the summarization of a text may be language-dependent (when an algorithm is specifically designed for a certain language) or independent (mostly based on algorithms for which it is not important. Over the years many scientific papers describe different processes and techniques for the summarization of information and the different purposes of its use. Some examples are:

- The PageRank algorithm used by Google Search to rank websites in their search engine results, with its graph-based ranking algorithms;
- Classifier4J[5] with its micro service "Summarizer"[6]. Its purpose is to "extract sentences from a text document, determine which are most important, and return them in a readable and structured way."
- A linear-time algorithm for lexical chain computation. It makes lexical chains as an intermediate representation for automatic text summarization. By lexical chains is understood the cohesion among an arbitrary number of words that can be computed in a source document by grouping sets of words that are semantically related and have a sense flow.

Although it is undeniable that the current technological advancements are remarkable, much is still to be done in this area. For this reason is not recommended to give full credibility to algorithms. The same applies if they are language-dependent or independent. No matter the improvement, they are still not able to match human judgment, especially about the nuances that each word contains within itself. This is one of the reasons why automatic text summarization is a very difficult task.

This said, here is also another point of view to consider: when a human summarizes a piece of text, he or she usually reads it with the purpose of developing his or her understanding. Then, when he or she writes the summary, there is a tendency to highlight points that are related to the person's own background. This implies rating as "important" information that other individuals might consider superfluous. In order to avoid this risk, when engaged in the process of summarizing, it is strongly recommended to give the task to more than one person.[7]

For the Bulgarian Summarization Corpus was chosen human summarization.

For the process of summarization was necessary the following:

- Texts containing 1000 to 1999 words had to be reduced respectively reduced by 40%, 20% and 10% of their initial volume when using entire sentences and by 32%, 16% and 8% when reducing by leaving only simple closes.
- Texts containing 2000 to 2999 words to be reduced by up to 24%, 12% and 6% of their initial volume when using entire sentences and by 20%, 10% and 5% when using simple closes.

[5] A Java library designed to do text classification. http://classifier4j.sourceforge.net/

[6] The tool may be found on the following website: https://algorithmia.com/algorithms/nlp/Summarizer

[7] An example is the Polish Summaries Corpus, where manual summarization was conducted by 11 annotators. Texts were randomly assigned.

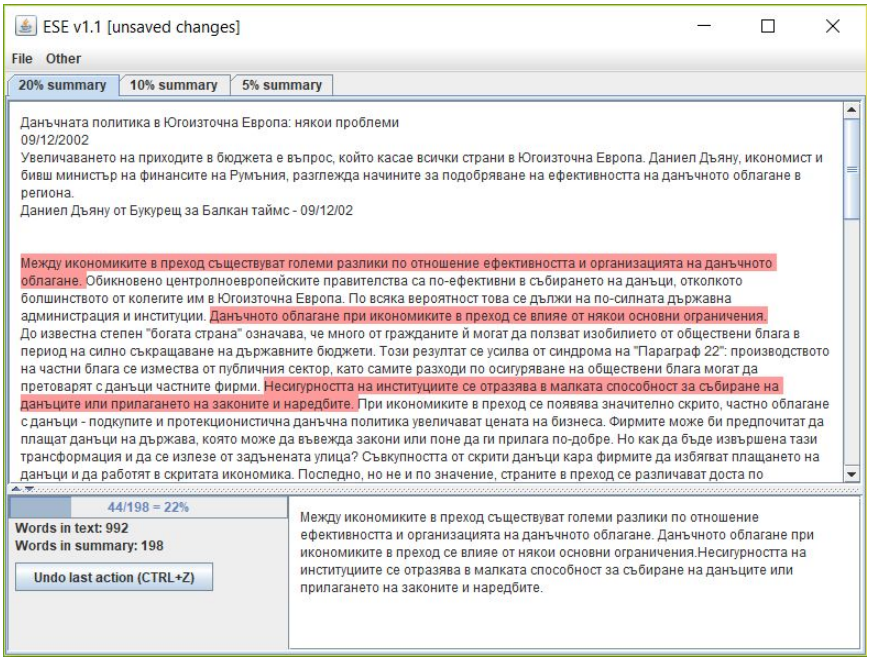

Figure 1: The interface of the ExtraSumAnnotator

For both groups, the reduction of simple clauses was done using the file containing the sentences selected in the first stage. All superfluous information in the texts such as author's name, dates or résumés as well as titles and abstracts had to be ignored in the process of summarizing the information in order to eliminate any possible interference with the results.

The tool ExtrSumAnnotator from the toolbox of the SummaryAnnotationTools[8] was used. It was originally created for the Polish Corpus mentioned in paragraph 2. Figure 1 shows its user friendly interface with the different stages of the summarization process. The selection of percentages is performed manually by the user, while the word count is automatic. All selected information is highlighted in red and shown in the little window under the main text. This information is then transferred to the next tab, where the process is repeated again until the end.

5.1 Selection of sentences

The selection of sentences was done as follows for every text:

In the first stage of the work the whole text was read;

Then started the selection of entire sentences respecting the selected percentages. The changes were saved as a new file;

This new file was then opened again and from it were selected only the simple clauses. This second file was then saved, again separately.

As mentioned above, the total number of files at the end was 3.

It followed a specific pattern during the entire process – mostly the first and last sentences were selected for each paragraph of the text. In specific cases the middle part was preferred, but this procedure is to be considered an exception.

It is noteworthy that for the majority of articles were chosen the first and last sentences of every paragraph. Also, in certain cases the summarization process was extremely difficult due to the impossibility of reduction of the volume to the percentage required. This happened mainly because of the risk of loss of information.

[8] They can be found on link of the Explore GitLab:
http://git.nlp.ipipan.waw.pl/summarization/SummaryAnnotationTools

The selection of simple clauses was a much easier process, since the message was almost always in the main sentence.

A practical example of the summarization process can be described as follows:

The main text contains 992 word (since it is in the first group, it has to be reduced respectively by 40%, 20% and 10% for entire sentences and by 32%, 16% and 8% for simple closes). In numbers this means that the reduction is:

For entire sentences:
992 words – 396 words (40%)
992 words – 198 words (20%)
992 words – 99 words (10%)

For simple clauses:
992 words – 317 words (32%)
992 words – 158 words (16%)
992 words – 79 words (8%)

Below parts of the text have been copy-pasted[9].

Step 1 is the reduction by 40%:

"Между икономиките в преход съществуват големи разлики по отношение ефективността и организацията на данъчното облагане. Обикновено централноевропейските правителства са по-ефективни в събирането на данъци, отколкото болшинството от колегите им в Югоизточна Европа. По всяка вероятност това се дължи на по-силната държавна администрация и институции. Данъчното облагане при икономиките в преход се влияе от някои основни ограничения.

До известна степен "богата страна" означава, че много от гражданите й могат да ползват изобилието от обществени блага в период на силно съкращаване на държавните бюджети. Този резултат се усилва от синдрома на "Параграф 22": производството на частни блага се измества от публичния сектор, като самите разходи по осигуряване на обществени блага могат да претоварят с данъци частните фирми. Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Но как да бъде извършена тази трансформация и да се излезе от задънената улица? Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно, но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. […]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Но добрите идеи и предприемаческият дух може да не са достатъчни, когато има нужда от банково финансиране, а банките изискват трудни за получаване гаранции. Новите наредби на Банката за международни разплащания за банковото осигуряване на заеми може да удари сериозно малките и средно големи фирми, освен ако банките не намерят творчески начини за финансиране. Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят

[9] Since the working language is English and the text are in Bulgarian, this is to be considered just as an example and the will focus only on the percentages and how the selection works visually. Because of this, there are only the first and the last paragraphs. The underlined sentences represents the red marks in the ExtraSumAnnotator.

специални финансови институции, които да задоволяват техните нужди. Положителен знак е, че ЕБВР е измежду спонсорите на банките, занимаващи се с МСП, създадени в региона."

Step 2 is a reduction of 20%:

"Данъчното облагане при икономиките в преход се влияе от някои основни ограничения. Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно, но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. […]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди."

Step 3 is a reduction by 10%:

"Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. […]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди."

As mentioned above, the selected sentences are saved in a separate file that is used for the simple clauses. Here are also visible the aforementioned difficulties and the risk of loss of information, since many of the sentences cannot be reduced to simple clauses.

Step 1 is a reduction of 32%

"Между икономиките в преход съществуват големи разлики по отношение ефективността и организацията на данъчното облагане. Обикновено централноевропейските правителства са по-ефективни в събирането на данъци, отколкото болшинството от колегите им в Югоизточна Европа. По всяка вероятност това се дължи на по-силната държавна администрация и институции. Данъчното облагане при икономиките в преход се влияе от някои основни ограничения.

До известна степен "богата страна" означава, че много от гражданите й могат да ползват изобилието от обществени блага в период на силно съкращаване на държавните бюджети. Този резултат се усилва от синдрома на "Параграф 22": производството на частни блага се измества от публичния сектор, като самите разходи по осигуряване на обществени блага могат да претоварят с данъци частните фирми. Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Но как да бъде извършена тази трансформация и да се излезе от задънената улица? Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно,

но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. […]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Но добрите идеи и предприемаческият дух може да не са достатъчни, когато има нужда от банково финансиране, а банките изискват трудни за получаване гаранции. Новите наредби на Банката за международни разплащания за банковото осигуряване на заеми може да удари сериозно малките и средно големи фирми, освен ако банките не намерят творчески начини за финансиране. Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди. Положителен знак е, че ЕБВР е измежду спонсорите на банките, занимаващи се с МСП, създадени в региона."

Step 2 is a reduction of 16%.

"Данъчното облагане при икономиките в преход се влияе от някои основни ограничения.

Несигурността на институциите се отразява в малката способност за събиране на данъците. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно, но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. […]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди."

Step 3 is a reduction of 8%

"Несигурността на институциите се отразява в малката способност за събиране на данъците. […]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди."

6. Conclusion

Summarization procedures and techniques will increase and improve in the following years due to the constant rise of information available on the internet. Considering the current situation and the growing market needs, it is to be expected that instruments such as corpora will be very useful and appreciated by professionals and common users, especially when they belong to peripheral languages like Bulgarian.

As for the Bulgarian Summaries Corpus, hopefully it will grow with more texts, which will cover additional fields.

Hopefully, this work will contribute to the future development of the Bulgarian Summaries Corpus and will increase the popularization of this sort of instruments.

References

Hristov, D. (2017). Automatic Text Summarization for the Bulgarian Language. Faculty of Mathematics and Informatics, Sofia University.

Mani, I. and Maybury, M.T. (1999). *Advances in Automatic Text Summarization*. (MITRE Corporation) Cambridge, MA: The MIT Press

Nakano, M., Shibuki, H., Miyazaki, R., Ishioroshi, M., Kaneko, K., Mori, T. (2010). *Construction of Text Summarization Corpus for the credibility of Information on the Web.* Graduate School of Environment and Information Sciences Yokohama National University
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.675.2386&rep=rep1&type=pdf

De Loupy, C., Guégan, M., Ayache, C., Seng, S., Torres Moreno, J-M. (2010). *A French Human Reference Corpus for Multi-Document Summarization and Sentence Compression*. Syllabs, Laboratoire Informatique d'Avignon (UAPV), Ecole Polytechnique de Montréal
http://www.lrec-conf.org/proceedings/lrec2010/pdf/919_Paper.pdf

Li, L., Forascu, C., El-Haj, M., Giannakopoulos, G. (2013). *Multi-document multilingual summarization corpus preparation, Part 1:Arabic, English, Greek, Chinese, Romanian.* BUPT, China, UAIC, Romania, Lancaster Univ., UK, NCSR Demokritos, Greece.
https://pdfs.semanticscholar.org/c4ba/4a4bc313a93850091b0b17ac27e2fbae569e.pdf

Ogrodniczuk, M., Kopéc, M.,(2014). *The Polish Summaries Corpus*. Institute of Computer Science, Polish Academy of Sciences
https://www.researchgate.net/publication/263087349_The_Polish_Summaries_Corpus

O'Connor, B., Krieger, M., Ahn, D. (2010). *TweetMotif: Exploratory Search and Topic Summarization for Twitter.* Carnegie Mellon University, Meebo, Inc. Microsoft, Inc
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.478.5512&rep=rep1&type=pdf

Mihalcea, R. (2004). *Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization.* Department of Computer Science, University of North Texas
http://www.aclweb.org/anthology/P04-3020

Silber, H. G., McCoy, K. F. (2002). *Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization.* University of Delaware
http://www.aclweb.org/anthology/W00-1438

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K., (2017). *Text Summarization Techniques: A Brief Survey.* Computer Science Department, University of Georgia, Institute of Bioinformatics, Department of Mathematics, Institute of Bioinformatics
http://cobweb.cs.uga.edu/~pouriyeh/Text_Summarization_Techniques_a_Brief_Survey.pdf

"Web content" – https://en.wikipedia.org/wiki/Web_content

"PageRank'' – https://en.wikipedia.org/wiki/PageRank

"Global language system" – https://en.wikipedia.org/wiki/Global_language_system