UDK 410

M. Petic, PhD., **R. Dumbrăveanu**, PhD.,
S. Chilat

DEVELOPING A TOOL FOR TEXTS WITH HETEROGENEOUS STRUCTURE PROCESSING

Abstract. *This article presents our approach in the elaboration of the system for processing Romanian unstructured text data. The project aims to elaborate the SoFTcrates tool, a software system for processing unstructured text data in order to create structured data output as computer linguistics resources. We described some mathematical aspects in text representation and presented some stages in unstructured text data processing. Also, the interface of the application is illustrated. In the future we will try to implement mechanisms of diversification of the founded words by means of derivation and WordNet semantic net. More over we will optimize the interface to have the possibility to find not only by a single word, but also by several words that the user consider more relevant to the text.*

Keywords: *lemma, web service, online Web application, unstructured text data, extraction algorithm*

М. В. Петик, канд. комп. наук,
Р. Н. Думбрэвяну, канд. матем. наук,
С. В. Килат

РАЗРАБОТКА ИНСТРУМЕНТА ДЛЯ ОБРАБОТКИ ТЕКСТОВ С НЕОДНОРОДНОЙ СТРУКТУРОЙ

Аннотация*. Представлен подход в разработке системы для обработки неструктурированных текстовых данных румынского языка. Целью данного проекта является разработка инструмента SoFTcrates, системы программного обеспечения для обработки неструктурированных текстовых данных для создания структурированных выходных данных в качестве компьютерных лингвистических ресурсов. Мы описали некоторые математические аспекты в текстовом представлении и представили некоторые этапы обработки неструктурированных текстовых данных. Кроме того, интерфейс приложения проиллюстрирован. В будущем мы постараемся реализовать механизмы диверсификации основанных слов с помощью словообразовательных механизмов и семантической сети WordNet. Более того, мы будем оптимизировать интерфейс, чтобы иметь возможность искать не только с помощью одного слова за раз, но и с помощью нескольких слов, которые пользователь считает более подходящими к тексту.*

Ключевые слова: *лемма, веб-сервис, онлайн веб-приложение, неструктурированные текстовые данные, алгоритм извлечения*

М. В. Петік, канд. комп. наук,
Р. М. Думбревяну, канд. матем. наук,
С. В. Килат

РОЗРОБКА ІНСТРУМЕНТУ ДЛЯ ОБРОБКИ ТЕКСТІВ З НЕОДНОРІДНОЮ СТРУКТУРОЮ

Анотація*. Представлен підхід в розробці системи для обробки неструктурованих текстових даних румунської мови. Метою даного проекту є розробка інструменту SoFTcrates, системи програмного забезпечення для обробки неструктурованих текстових даних для створення структурованих вихідних даних в якості комп'ютерних лінгвістичних ресурсів. Ми описали деякі математичні аспекти в текстовому поданні і представили деякі етапи обробки неструктурованих текстових даних. Крім того, інтерфейс програми проілюстровано. В майбутньому ми постараємося реалізувати механізми диверсифікації заснованих слів за допомогою словотворчих механізмів і семантичної мережі WordNet. Більш того, ми будемо оптимізувати інтерфейс, щоб мати можливість шукати не тільки завдяки одному слову за раз, а й завдяки кількох слів, які користувач вважає більш відповідними до тексту.*

Ключові слова: *лемма, веб-сервіс, онлайн веб-додаток, неструктуровані текстові дані, алгоритм вилучення*

1. INTRODUCTION

The development of information technologies, computer networks, and communication led to huge volumes of available information. This data is very difficult to process, as it is mostly unstructured, but it contains markers for social, cultural and security processes.

The project aims to elaborate the SoFTcrates tool, a software system for processing un-

structured text data in order to create structured data output as computer linguistics resources.

The described software system will be based on mechanisms of natural language processing and will create a great potential of interpretation of human-created text data in a systematic way with different purposes, like contextual analysis of the activity of software users from a statistical point of view or context and meaning detection of natural language.

The aim of this article is to present our approach in the elaboration of the system for processing unstructured text data in order to create structured data output as computer linguistics resources by means of lexicon of markers.

This paper has the following structure: section 2 presents the methodology used in our research; in section 3 we describe the mathematical aspects of the research; section 4 illustrates the application interfaces and the working process.

2. THE METHODOLOGY

Therefore, our main purpose is searching for texts of news or articles by using a set of terms (query) extracted from an unstructured document. The large amount of unstructured data leads to the importance of its processing. One of the ideas is to extract the keywords that are important for the text understanding [1]. Having a list of keywords it will help us to find out which is the main topic of the text [2].

The task was to lemmatize the words from the text and then to make a lexicon of frequency. To solve these problems, we used the PoS Tagger for Romanian web service[1] developed by the Alexandru Ioan Cuza University from Iasi [3]. The first part of the program took the text document at the input and at the exit we have obtained other text document, where text was divided into sentences and the last one in words, by means of introducing of XML language specifications [4].

The application represents a language processing model based on both rule-based information extraction and statistical methods. The rules are used as constraints against ambiguity difficulties, creation of rules being facilitated by the use of Graphical Grammar Studio, an open-source software product which facilitates the identification and matching of tokens and sequences that can be annotated as well as the simple tokens. PoS Tagger, as well as most of the other NLP utilities developed at UAIC, is published as a web service using the WSDL specification on top of the SOAP protocol [3].

Each word was wrapped in tag `<w> ... </w>`, which consists of some attributes indicating the word in his initial form and its parts of speech [5]. For the future vocabulary, we choose adjectives, nouns and verbs, because only words of these parts of speech have individual meaning. The second part of our application has the following purpose:

- Taking a word at the initial form and it's part of speech;
- Word rating and sort descending for verbs, nouns and adjectives.

Then, if the word was often used, it is important and it must be introduced in the vocabulary. Computing many texts, we can come closer to vocabulary of the topic. In practice, difficulties arise concerning how many order words we must take for the vocabulary if percentage of useful words come down with taking a lot of words.

The result consists in a lexicon of text frequency and the possibility to find the relevant texts in a searching engine of the most frequent words of the processed text.

However not all the most frequent words express the main meaning, it is correct to give the possibility to the user to select those words that are important. For example, we often find the verb to be that is of general meaning and is not a keyword for a specific topic.

3. MATHEMATICAL ASPECTS

The motivation of SoFTcrates program[2] is to facilitate the process of understanding natural language texts (unstructured documents) written in Romanian [6].

To accomplish this task, an unstructured document is converted into a form that computer can "understand". This step is called text representation [7]. One way to achieve text representation is to assign to each unstructured doc-

[1] http://nlptools.infoiasi.ro/WebPosRo/ - Simionescu Radu, UAIC Romanian Part of Speech Tagger, 2011

[2] softcrates.usarb.md -

ument D, a vector v of finite real numbers of length k, as follows:

$$v = (w_1, w_2, w_3, \ldots, w_k), \quad (1)$$

where $w_i \in \Re$, or $0 \leq w_i \leq 1$ (i.e. w_i is normalized) denotes the weight of term $t_i \in T$, $1 \leq i \leq k$. The set $T = \{t_1, t_2 \ldots t_k\}$ of terms is predefined and is called *index set* or *vocabulary*.

This approach is called Vector Space Model (VSM). The idea of the VSM (as can be seen from above) is to represent each document in a collection as a point in a space (a vector in a vector space) [8]. Respectively the closeness of points in this space correspond to semantically similar documents and conversely. In this setting are suitable many mathematical tools from vector spaces, thus we can calculate the similarity of two documents as follows:

$$\cos\Theta = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}, \quad (2)$$

where the ·(dot) is the scalar product of vectors and $\|.\|$ is the norm. The triple (*DS*, *QS*, *sim*), where *DS* is the document space, *QS* is the query space and *sim* is the similarity mapping is called retrieval system.

Although the main goals regarding the SoFTcrates are to compute the weights for the document vectors and calculate similarities between these vectors and a given query, the authors would like also to investigate topological aspects of retrieval systems. Their motivation is due to already existing results in this area, e.g. in [9] was proved that under certain conditions the *DS* may be viewed as Haudorff space; in [10] was proved that algebraic information retrieval spaces can be regarded as a Hilbert space.

4. THE APPLICATION INTERFACE

Taking into account the methodology from the section 2 and the needed functionalities our application has three interfaces, corresponding to three stages in processing input texts:

- getting input text and the number of resulting words (in Romanian *Numarul de rezultate necesare*) (Fig. 1);
- displaying the most frequent resulting words with its frequency (Fig. 2);
- the searching results in a searching engine (e.g. Google).

Fig. 1. The interface of the Softcrates for input text

As it is shown in Fig. 2, the most frequent words are displayed as list links. By clicking on each link, one could initiate a search with this word using Google.

Fig. 2. The interface of the Softcrates for input text

Besides adding a second search engine by which one could search news articles in Romanian language, we intend to make the search interface more flexible by adding to each word a checkbox. This feature will provide the user with the possibility to choose what words should be used for the search, i.e. build custom queries.

5. CONCLUSIONS AND FUTURE WORK

The combination of methods from the field of information technology with the ones from linguistics would resolve the problem of computational linguistics, namely structured information retrieval of unstructured texts.

SoFTcrates will represent a substantial contribution in the field of natural language pro-

cessing software as a synthesis, intermediation and enhancement of existing instruments in the field.

In the future, we will try to implement mechanisms of diversification of the founded words by means of derivation and WordNet semantic net. Moreover we will optimize the interface to have the possibility to find not only by means of a single word, but also by several words that the user considers more relevant to the text.

6. **ACKNOWLEDGEMENTS.**

This article is carried out as part of the project ”Developing of a text processing system with heterogeneous structure” supported by Supreme Council for Science and Technological Development from the Republic of Moldova.

References

1. Pankratova N.D., and Dozirtsiv V.O., (2016), Application of Methods for text Analysis of the Emotional tone to Identify Social Disasters, *In System analysis and Information Technology: 18-th International Conference SAIT 2016,* Kiev, Ukraine, May 30 – June 2, 2016, *Proceedings*. 38 p.

2. Bolea C., (2015), Vocabulary, Synonyms and Sentiments of Hazard-related Posts on Social Networks. An Analysis for Romanian Messages, *Proc. IEEE Conf. SPED 2015*, Bucharest, Oct. 2015.

3. Simionescu R., (2011), Hybrid POS Tagger, *Proceedings of “Language Resources and Tools with Industrial Applications” Workshop (Eurolan 2011 Summer School),* Cluj-Napoca, Romania, pp. 21 – 28.

4. Petic M., and Raciula L., (2014), Computer Based Identification of Lines with Romanian Chromatic Words from Poems, *In: Electrotechnic and Computer Systems Journal,* No. 13 (89), *Section Systems of Artificial Intelligence,* Odessa, Ukraine, pp. 114 – 119.

5. Petic M., and Gîfu D., (2014), Transliteration and Alignment of Parallel Texts from Cyrillic to Latin, *In: Proceedings of LREC-2014,* Reykjavik, Iceland, 26 – 31 May 2014, pp. 1819 –1823.

6. Petic M., Chilat S., Dumbraveanu R., and Osoian E., (2015), Methods of Processing Unstructured Texts, *Book of Abstract CAIM*, Suceava, Romania, 74 p.

7. Wang D., and Zhang H., (2013), Inverse-Category-Frequency Based Supervised Term Weighting Schemes for Text Categorization, *Journal of Information Science and Engineering*, Vol. 29, pp. 209 – 225.

8. Turney P.D., and Pantel P., (2010), From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, Vol. 37, pp. 141 – 188.

9. Egghe L., and Rousseau R., (1998), Topological Aspect of Information Retrieval, *Journal of the American Society for Information and Technology,* Vol. 49, No. 13 pp. 1144 – 1160.

10. C. J. van Rijsbergen, (2004), The Geometry of Information Retrieval, *Cambridge University Press, Cambridge.*

Received 31.05.2016

Petic
Mircea, Associate Profesor,
PhD. in Computer Science,
Departament of Mathematics and Computer Science,
Alecu Russo Balti State University Republic of Moldova
38 Puskin street, Balti, MD-3100.
Tel.: +373 231 52488.
E-mail: petic.mircea@gmail.com

Dumbrăveanu
Radu, Researcher,
PhD. in Mathematics, Departament of Mathematics and Computer Science, Alecu Russo Balti State University Republic of Moldova.
38 Puskin street, Balti, MD-3100.
Tel.: +373 231 52488.
E-mail: vundicind@gmail.com

Chilat
Sergiu, Senior Researcher,
PhD. Student in Computer Science,
Departament of Mathematics and Computer Science,Alecu Russo Balti State University Republic of Moldova.
38 Puskin street, Balti, MD-3100.
Tel.: +373 231 52488.
E-mail: chilatsergiu@gmail.com