INTERNATIONAL CONFERENCE

COMPUTATIONAL LINGUISTICS AND INTELLIGENT SYSTEMS

COLINS 2017

PROCEEDINGS

KHARKIV, UKRAINE

21 APRIL 2017

COLINS 2017

ISSN 2523-4013

The 1st International Conference

COMPUTATIONAL LINGUISTICS AND INTELLIGENT SYSTEMS

Proceedings of the Conference

21 April 2017

Kharkiv, Ukraine

The 1st International Conference
COMPUTATIONAL LINGUISTICS AND
INTELLIGENT SYSTEMS

PROCEEDINGS

Kharkiv, Ukraine
21 April 2017

ISSN 2523-4013

Preface

The volume contains the papers presented at COLINS 2017: the 1st International conference "Computational linguistics and intelligent systems".

The main purpose of the CoLInS conference is a discussion of the recent researches results in all areas of Natural Language Processing and Intelligent Systems Development.

The conference is soliciting literature review, survey and research papers comments including, whilst not limited to, the following areas of interest:

- mathematical models of language;
- artificial intelligence;
- statistical language analysis;
- data mining and data analysis;
- social network analysis;
- speech recognition;
- machine translation, translation memory systems and computer-aided translation tools;
- information retrieval;
- information extraction;
- text summarization;
- computer lexicography;
- question answering systems;
- opinion mining;
- intelligent text processing systems;
- computer-aided language learning;
- corpus linguistics;

The language of COLINS Conference is English.

The conference took the form of oral presentation by invited keynote speakers plus presentations of peer-reviewed individual papers. There was also an exhibition area for poster and demo sessions. A Student section of the conference for students and PhD students run in parallel to the main conference.

This year Organization Committee received 13 submissions, out of which 12 were accepted for presentation as a regular papers. The papers are submitted to the following tracks: corpus linguistics (1 paper), computational lexicography (2 papers), automatic ontology building (2 papers), morphological analisys (2 papers), content analysis (2 papers), intelligent computer systems building (2 papers) and problem of classification (1 paper). The papers directly deal with such languages: Ukrainian, Russian, Spanish, French, English, Polish and Danish.

COLINS 2017 Organization Committee:

Olga Kanishcheva (National Technical University "KhPI", Ukraine)
Olga Cherednichenko (National Technical University "KhPI", Ukraine)
Natalia Borysova (National Technical University "KhPI", Ukraine)
Victoria Vysotska (Lviv Polytechnic National University, Ukraine)

Organization Committee:

Olga Kanishcheva (National Technical University "KhPI", Ukraine)
Olga Cherednichenko (National Technical University "KhPI", Ukraine)
Natalia Borysova (National Technical University "KhPI", Ukraine)
Victoria Vysotska (Lviv Polytechnic National University, Ukraine)

Programme Committee:

Wolfgang Kersten (Institut für Logistik und Unternehmensführung, Germany)
Natalia Grabar (CNRS UMR 8163 STL, France)
Galia Angelova (Bulgarian Academy of Sciences, Bulgaria)
Thierry Hamon (LIMSI-CNRS & Université Paris 13, France)
Svetla Boytcheva (Sofia University, Bulgarian Academy of Sciences, Bulgaria)
Natalia Sharonova (National Technical University "KhPI", Ukraine)
Victoria Vysotska (Lviv Polytechnic National University, Ukraine)
Fadila Bentayeb (ERIC Laboratory, University of Lyon 2, France)
Olga Cherednichenko (National Technical University "KhPI", Ukraine)
Iryna Yevseyeva (Newcastle University, England)
Vitor Basto-Fernandes (University Institute of Lisbon, Portugal)
Oleg Garasym (Volvo IT, Poland)
Nina Khairova (National Technical University "KhPI", Ukraine)
Vasyl Lytvyn (Lviv Polytechnic National University, Ukraine)
Mikhail Godlevsky (National Technical University "KhPI", Ukraine)
Zoran Cekerevac ("Union – Nikola Tesla" University, Serbia)
Volodymyr Pasichnyk (Lviv Polytechnic National University, Ukraine)
Oleg Bisikalo (Vinnytsia National Technical University, Ukraine)
Oleksandr Gozhyi (Petro Mohyla Black Sea National University, Ukraine)
Volodymyr Lytvynenko (Kherson National Technical University, Ukraine)
Victoria Bobicev (Technical University of Moldova, Moldova)
Lidia Pivovarova (University of Helsinki, Finland)
Olena Levchenko (Lviv Polytechnic National University, Ukraine)
Scheller-Boltz Dennis (Vienna University of Economics and Business, Austria)
Oksana Bihun (Mathematics University of Colorado, Colorado Springs USA)
Evelin Krmac (University of Ljubljana, Slovenia)
Antonina Savka (Openet, Ireland)
Yuriy Myronovych (Sky UK, United Kingdom)
Waldemar Wojcik (Lublin University of Technology, Lublin, Poland)
Dmitry V. Lande (Institut for Information Recording of NAS of Ukraine, Ukraine)
Sergii Babichev (Jan Evangelista Purkinje University in Usti nad Labem, Chech Republic)
Michael Pokojovy (University of Memphis (TN, USA))
Silakari Sanjay (Rajiv Gandhi Technical University, India)
Andrzej Smolarz (Lublin University of Technology, Poland)
Yaroslav Novytskyy (NetDevLabs IT Consulting Inc., Canada)
Viktor Mashkov (Jan Evangelista University in Ústí nad Labem, Czech Republic)
Opeyemi Olakitan (Cornell University, United Kingdom)
Yurii Biurher (Zoral Labs, Ukraine)
Klaus ten Hagen (University of Applied Science Zittau/Goerlitz, Germany)
Danuta Zakrzewska (Lodz University of Technology, Poland)
Manik Sharma (DAV University, India)
Nastasiya Osidach (Grammarly, Ukraine)

Keynote Speakers:

Mariana Romanyshyn (Grammarly, Ukraine)
Taras Hnot (SoftServe, Ukraine)
Dmitry Lande (Institute for Information Recording of NAS of Ukraine)
Thierry Hamon (LIMSI-CNRS & Université Paris 13, France)
Vladimir Shalimov (Fortifier, Ukraine)
Wolfgang Kersten (Institut für Logistik und Unternehmensführung, Germany)
Vladyslav Kolbasin (Grid Dynamics, Ukraine)
Yaroslav Protsenko (Fortifier, Ukraine)

Conference Program

9.30	***Registration***	
10.00	***Conference Opening***	
10.20	Grammatical Error Correction: why commas matter *Mariana Romanyshyn* (Grammarly)	
10.45	Qualitative content analysis: expertise and case study *Taras Hnot* (SoftServe)	
11.10	Creation of subject domain models on the basis of monitoring of network information resources *Dmitry Lande* (Institute for Information Recording of NAS of Ukraine)	
11.30	Biomedical text mining *Thierry Hamon* (LIMSI-CNRS & Université Paris 13)	
12.00	***Poster section and coffee-break***	
13.30	Big Data – Revolution in Data Storage and Processing *Vladimir Shalimov* (Fortifier)	
13.55	The Digital Transformation of the Industry – the Logistics Example *Wolfgang Kersten* (Institut für Logistik und Unternehmensführung)	
14.20	AI trends, or brief highlights of NIPS 2016 *Vladyslav Kolbasin* (Grid Dynamics)	
14.45	Intuition on modern deep learning approaches in computer vision *Yaroslav Protsenko* (Fortifier)	
15.10	***Coffee-break***	
15.30	***Stream 1 (Paper Presentations)***	***Stream 2 (Student Section)***
15.30	Unsupervised acquisition of morphological resources for Ukrainian *Natalia Grabar, Thierry Hamon*	Intelligent data processing in creating targeted advertising *Stanislav Kirkin, Karina Melnyk*
15.45	Intelligent System Structure for Web Resources Processing and Analysis *Vasyl Lytvyn, Victoria Vysotska, Lyubomyr Chyrun, Andrzej Smolarz, Oleh Naum*	Use of Linguistic Criteria for Estimating of Wikipedia Articles Quality *Anastasia Kolesnik, Nina Khairova*
15.55	Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security) *Olena Orobinska, Jean-Hugues Chauchat, Natalya Sharonova*	Gamification: Today and Tomorrow *Katherine Yukhno, Eugenia Chubar*
16.10	NLP Resources for a Rare Language Morphological Analyzer: Danish Case *Mykhailo Kotov*	Analysis of Existing German Corpora *Inna Olifenko, Natalia Borysova*
16.20	Content Analysis of some Social Media of the Occupied Territories of Ukraine *Volodymyr Lytvynenko, Iryna Lurie, Svitlana Radetska, Mariia Voronenko, Natalia Kornilovska, Daria Partenjucha*	Search Optimization and Localization of the Website of Department of Applied Linguistics *Vsevolod Pidpruzhnikov, Margarita Ilchenko*
16.35	An Index of Authors' Popularity for Internet Encyclopedia *Dmitry Lande, Valentyna Andrushchenko, Iryna Balagura*	Improving Communication in Enterprise Solutions: Challenges and opportunities *Vitaliy Gorbachov, Olga Cherednichenko*
16.45	Methods of comparing interval objects in intelligent computer systems *Gennady Shepelev, Nina Khairova*	Development and Computerization of an English Term System in the Fields of Drilling and Geology *Herman Hordienko, Margarita Ilchenko*
17.00	Evaluation of Adequacy of the Emergency Situations Classification Formalized Model *Vera Titova, Ielizaveta Gnatchuk*	Statistical methods usage of descriptive statistics in corpus linguistic *Valeriy Didusov, Zoia Kochueva*
17.10	Semantic State Superpositions and Their Treatment in Virtual Lexicographic Laboratory for Spanish Language Dictionary *Yevgen Kuprianov*	
17.25	Discursive units in scientific texts *Yulia Verbinenko*	

17.35	Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation *Thierry Hamon, Natalia Grabar*
17.50	The Method of Automated Building Basic Ontology *Vasyl Lytvyn, Victoria Vysotska, Waldemar Wojcik, Dmytro Dosyn*
18.00	***Conference Closing***

Table of Contents

PAPER PRESENTATIONS

Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation

Natalia Grabar[1] and Thierry Hamon[2]

[1] CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France

natalia.grabar@univ-lille3.fr

[2] LIMSI-CNRS, Orsay, Université Paris 13, Sorbonne Paris Cité, France

hamon@limsi.fr

Abstract. The question on creation of linguistic resources (such as corpora, lexica or terminologies) occupies an important place in the research areas related to linguistics, Natural Language Processing, Computer Sciences, psycholinguistics, etc. In this paper, we propose the description of a multilingual corpus in which Ukrainian is the target language, while source languages are Polish, French and English. The corpus contains literary texts and a small subset built with texts provided by medical area. On the whole, the corpus is composed of 62 literary texts and 129 medical texts. The corpus counts over 1 million words in the tar-get Ukrainian language, and at least as much in the source languages taken all together. This is a directional corpus aligned at the level of sentences. After the description of this corpus, we introduce some possible exploitations and first results. We then conclude and indicate some directions for future work. The corpus presented in this work is available for the research purposes: http://natalia.grabar.free.fr/resources.php

1 Introduction

The question on creation of linguistic resources (such as corpora, lexica or terminologies) occupies an important place in the research areas related to linguistics, Natural Language Processing, Computer Sciences, psycholinguistics, etc. Indeed, the availability of such resources provides the possibility to design, develop and evaluate methods and tools specific to several contexts and applications (information retrieval, acquisition of lexica, machine translation, question/answering, categorization of documents...). As a matter of fact, different applications may require the availability of different kinds of resources.

The purpose of this work is to introduce and describe multilingual parallel and aligned corpus, in which the target language is Ukrainian, while the current source languages are Polish, French and English.

In what follows, we describe first the existing resources and NLP tools developed for the Ukrainian language (section 2), and then present our method for collection

(section 3) and building (section 4) of the parallel and aligned corpus. We then present some possible exploitations of this aligned corpus and the currently obtained results (section 5). We conclude with directions for future research (section 6).

2 Existing resources and methods for Ukrainian

Ukrainian language is part of the Slavic family of languages. Currently, little resources are freely available for Ukrainian, especially when looking at NLP tools and resources. We propose here a short review of some existing resources and tools: corpora, morphological resources, dictionaries and terminologies, and NLP tools.

2.1 Corpora

We have found several corpora dedicated to the description of the modern Ukrainian language: national corpus of the Ukrainian language [33] which is available online[1], literary corpus with the work by Ivan Franko [29] built for the research and educational purposes, and corpus with dialectal texts [38].

Besides, several parallel corpora involving Ukrainian have been proposed, such as Polish-Ukrainian [16] and Bulgarian-Ukrainian [23] corpora. Let's also notice a platform for the development and repository of comparable corpora in several languages including Ukrainian [4].

Although it started recently, there is an ongoing research on building of the electronic corpora [14, 13, 34], and on the related research questions such as representativeness of corpora [35], general methodological basis for the creation of corpora [26], creation of signed corpora [39], morphological annotation of corpora [32], methods for frequency studies [30].

2.2 Morphological resources

Two sets of morphological resources dedicated to Ukrainian can be mentioned: Multex-East Ukrainian lexicon for the general language[2] with morphological features [8, 17], and a corpus-based lexicon with pairs of morphologically related words from general and medical area languages [9]. Let's also notice the Mondilex infrastructure with digital resources in Slavic lexicography [7], that gathers resources for several Slavic languages.

2.3 Dictionaries and terminologies

Several dictionaries exist and describe the general language and specialized areas in Ukrainian. Yet, such dictionaries are mostly available in traditional paper format.

[1] *http://www.mova.info/corpus.aspx?l1=209*
[2] *https://www.clarin.si/repository/xmlui/handle/11356/1041*

Nevertheless, we can notice the frequency dictionary of texts written by Ivan Franko [28], and an electronic dictionary of fire security [40]. Besides, some of the existing dictio-naries can be queried online[3].

Notice that the current research in Ukraine increasingly addresses the use of electronic corpora for the building of dictionaries and terminologies [37, 27, 31], and the transformation of traditional dictionaries in electronic format [36]. As for the terminology-related research, a short review has been proposed [10].

2.4 NLP tools

Among the existing NLP tools, we can mainly mention two Part-of-Speech (POS) taggers: the UGtag POS tagger [18] which does not perform the syntactic and morphological disambiguation, and a TNT model for Ukrainian [1].

3 Collection of texts

For the purpose of our objectives, we use two kinds of texts, one covering general and one covering specialized languages:

– *Literature.* The literary corpus in Ukrainian is collected from the *Укр.Літ*[4] and *UkrLib*[5] websites which purpose is to promote literature in Ukrainian, with both original and translated works. According to the policy of these websites, these works are publicly available and can be used as far as they are cited. For some translated works, we could collect publicly available originals from websites like *Project Gutenberg*[6]. Three source languages are thus covered: Polish, French and English. This set of data contains the literary work written in Polish, French or English, and then translated in Ukrainian. These data provide a good basis for the creation of parallel corpora;
– *MedlinePlus.* Medical documents are obtained from the MedlinePlus of the National Library of Medicine[7]. These documents contain patient-oriented brochures on several medical topics, such as body systems, disorders and conditions, diagnosis and therapy, demographic groups, health and wellness. These brochures have been created in English and then translated in several languages, among which Ukrainian. These works, produced by the MedlinePlus team, are not copyrighted under U.S. law and can be freely used. Here again, Ukrainian is the target language.

[3] *http://lcorp.ulif.org.ua/dictua/*
[4] *http://ukrlit.org*
[5] *http://www.ukrlib.com.ua/*
[6] *http://www.gutenberg.org/*
[7] *www.nlm.nih.gov/medlineplus/healthtopics.html*

Table 1. Size of the collected parallel texts per language, in terms of number of texts and word occurrences

Corpus	Occ_{wds}	Nb_{txt}
Literature/UK	3,111,656	110
Literature/FR	1,310,732	29
Literature/EN	2,203,350	51
Literature/PL	260,536	30
MedlinePlus/UK	43,184	129
MedlinePlus/EN	46,544	129

In Table 1, we indicate the size of the collected corpora for each language: Ukrainian *UK*, French *FR*, Polish *PL*, and English *EN*. This dataset contains parallel texts, while in each pair of languages Ukrainian is the target language. The other three languages (French, Polish and English) are the source languages. Among the English-language authors we can find Charlotte Bronte, Lewis Carroll, Izak Azimov, Raymond Chandler, Agatha Christie, James Joyce, Jack London, George Orwell and JRR Tolkien. Among the French-language authors we can find Honoré de Balzac, Albert Camus, Alexandre Dumas, Charles Perrault, Guy de Maupassant, Antoine de Saint-Exupéry and Jules Verne. The Polish-language texts have all been written by Stanislaw Lem.

These source languages have been chosen for their representativity and relation with the Ukrainian language:

- Polish is also a Slavic language, and is close to Ukrainian. Polish is now quite well researched within the NLP field. We assume that the methods and tools developed for the Polish language can be adapted to Ukrainian provided that there are suitable corpora and resources;
- English and French languages are well researched from the NLP point of view. We assume, it is possible to take advantage of this research using the transfer methodologies [24, 21], provided that there are suitable parallel and aligned corpora, and resources.

As indicated in Table 1, the Ukrainian part of the corpus is the most extensive because in covers the works in the three source languages. We can also observe that specialized subset of texts contains greater number of documents but smaller number of word occurrences. This subset is much smaller than the literary work subset.

4 Building of corpus

The documents indicated in Table 1 are all converted in the text format and the UTF-8 encoding. The original documents can be in different formats (text, word, pdf, html...). We use Linux tools for converting them into text, such as **pdftotext**, **antiword** or homemade **perl** program **html2txt**. For managing the encoding, we use the Linux tool **recode**. Once these two aspects are homogeneous, these text files are segmented in sentences in each language, for which we use strong punctuation and upper-cased characters. Specific **perl** scripts have been created for each of the processed languages.

Ideally, such segmentation should provide corpus aligned at the sentence level.

Yet, it is necessary to verify the correctness of the segmentation in sentences and the parallelism between the source and target versions of a given document. Indeed, during the translation process, the organization of the sentences and their segmentation can be modified by the translator in order to better convey the meaning. Besides, some sentences can also be omitted. For instance, in Charlotte Bronte's Jane Eyre, the source sentence in Example (1) is segmented in two sentences during its translation in Ukrainian (by Петро Соколовський), as indicated in Example (2).

(1) *I was glad of it: I never liked long walks, especially on chilly afternoons: dreadful to me was the coming home in the raw twilight, with nipped fingers and toes, and a heart saddened by the chidings of Bessie, the nurse, and humbled by the consciousness of my physical inferiority to Eliza, John, and Georgiana Reed.*

(2) *Щодо мене, то я була рада: я страх не любила довгих пообідніх прогулянок, а надто взимку.*
Жахливо було вертатися додому в холодному присмерку, коли заходять зашпори в руки і в ноги, а серце ниє від сердитого бурчання Бесі, нашої няні, та від принизливого усвідомлення фізичної переваги наді мною Елізи, Джона та Джорджіани Рід.

Hence, the manual control and correction during the alignment at the sentence level is necessary. This is a very long and thorough process, which guarantees the quality of the aligned corpora. Notice that the human annotator must understand the source and target languages involved in order to be able to control the correct alignment of sentences.

Table 2. Currently aligned corpora, size indicated in word occurrences in each language

Corpus	*Source*	*Target*
Literature/FR	507,063	419,479
Literature/EN	502,393	424,730
Literature/PL	260,536	264,200
Medline/EN	46,544	43,184

In Table 2, we indicate the size of the currently aligned texts, each of which has undergone manual verification. On the whole, the aligned corpus provides 1,151,593 word occurrences in the target Ukrainian language. As we can see, all medical texts and all literary texts in the Polish/Ukrainian pair has been aligned and verified, while only part of French and English source texts is operational up to now. The current version of this parallel and aligned corpus is intended to grow with new texts: other texts are being checked for the correct alignment. In Table 2, we can also observe that the Ukrainian texts translated from English and French are usually shorter in number of words then the original texts, while the translation from Polish contains slightly higher number of words.

5 Exploitation of aligned corpus

In Figures 1 and 2, we present two excerpts from the English/Ukrainian sentence-aligned corpora: literary corpus from Charlotte Bronte's Jane Eyre and medical corpus, respectively.

These aligned corpora can be used for instance for the acquisition of bilingual lexica for the general and medical languages, for the acquisition of paraphrases [3, 5, 15], for the stylistic analysis of the source and target languages, for the contrastive studies, and for the machine translation. For instance, we have started to use the *Medline* aligned corpus for the acquisition of bilingual medical terminology in Ukrainian thanks to the use of the multilingual transfer [11]. Hence, in Figure 3, we underline the terms extracted in the English text and then transferred on the Ukrainian text thanks to their further alignment at the word level with the **GIZA++** algorithm [22].

English	*Ukrainian*
"What does Bessie say I have done?" I asked.	– Що вам Бесі наговорила на мене? – спитала я.
"Jane, I don't like cavillers or questioners;	– Джейн, я не люблю, коли чіпляються до слів і допитуються.
besides, there is something truly forbidding in a child taking up her elders in that manner.	Дитина не сміє так розмовляти зі старшими!
Be seated somewhere; and until you can speak pleasantly, remain silent. "	Іди сядь собі десь і, поки не навчишся бути чемною, мовчи.
A breakfast-room adjoined the drawing-room, I slipped in there.	З вітальні був хід у невеличку їдаленьку; отож я й шмигнула туди.
It contained a bookcase:	Там стояла шафа з книжками.
I soon possessed myself of a volume, taking care that it should be one stored with pictures.	Я вибрала собі одну з них, спершу подивившись, чи вона з малюнками.

Fig. 1. Example of the sentence-aligned literary corpus (English/Ukrainian), from Charlotte Bronte's Jane Eyre

English	*Ukrainian*
Cancer cells grow and divide more quickly than healthy cells.	Ракові клітини ростуть і діляться швидше, ніж здорові клітини.
Cancer treatments are made to work on these fast growing cells.	При лікуванні раку здійснюється вплив на ці клітини, що швидко ростуть.
– Tiredness	– Втома
– Nausea or vomiting	– Нудота або блювота
– Pain	– Біль
– Hair loss called alopecia	– Втрата волосся, що називається алопецією

Fig. 2. Example of the sentence-aligned MedlinePlus corpus (English/Ukrainian), file *CANCERTREATMENTSIDEEFFECTS.TXT*

English	*Ukrainian*
Cancer cells grow and divide more quickly than healthy cells.	Ракові клітини ростуть і діляться швидше, ніж здорові клітини.
Cancer treatments are made to work on these fast growing cells.	При лікуванні раку здійснюється вплив на ці клітини, що швидко ростуть.
– Tiredness	– Втома
– Nausea or vomiting	– Нудота або блювота
– Pain	– Біль
– Hair loss called alopecia	– Втрата волосся, що називається алопецією

Fig. 3. Example of the transferred terminological units using sentence-aligned MedlinePlus corpus (English/Ukrainian), file *CANCERTREATMENTSIDEEFFECTS.TXT*

Besides, parallel and aligned corpora can provide other interesting insights on language and grammar, typically issued from contrastive linguistics studies and Natural Language Processing. Let's mention some existing works:

- study of grammatical verbal constructions in English and Norwegian [12];
- crosslingual disambiguation [2], which shows that, depending on its context of occurrence, the English noun *plant* can be translated as French *plante* ("living thing in soil") or *usine* ("factory"). Such disambiguation of the source text improves the overall results of word sense disambiguation by up to 25%;
- improving the quality of lexicon bootstrapping in one language using translations in other languages [25], which shows that the results with German and English data are improved by 25%;
- semantic study of morphological units [6], in which the semantics of agentive suffixes in French *-iste* and Italian *-ista* rely on translation data obtained from an Italian-French bilingual dictionary and corpora;
- study of translations for out-of-dictionary words and expressions, such as translation of evaluative prefixes [19] or argumentative and discourse-organizing sequences [20]. Hence, in the study on translation of evaluative prefixes [19], the authors found out that several situations are possible: (1) translation with a derivative containing an evaluative prefix {*sous-estimer*, *underestimate*}; (2) translation with a derivative containing a non-evaluative prefix {*sous-utilisé*, *unused*}; (3) translation with a non-prefixed word (which can be a simplex word, a suffixed word or a compound) {*sous-alimenté*, *starving*}, {*sous-équipé*, *ill-equipped*}, {*surpoids*, *obesity*}; (4) translation with a periphrasis {*ultra-concurrence*, *competition taken to extremes*}, {*hyper-fédéraliste*, *extremely federalist*}; (5) zero translation, when the prefixed word is not translated in the target text.

Hence, the availability of parallel and aligned corpora provides several research possibilities for creating and enriching resources for Ukrainian.

6 Conclusion and Future Work

In this work, we propose parallel and aligned corpus involving Ukrainian language. The corpus is aligned at sentence level. This is a directional corpus because the source and target languages, as well as the translation direction are identified: Ukrainian is the target language, while Polish, French and English are source languages. The corpus contains texts from the general language (literary texts) and medical area. On the whole, the aligned corpus contains over 1 million words in the target Ukrainian language, and at least as much in the source languages.

In the future, we plan to extend the currently available aligned corpus with new sentence-aligned texts. The current three source languages (Polish, French and English) will be given advantage. This will allow to efficiently design and exploit transfer methodologies [24, 21] and statistical approaches such as those used in word alignment and machine translation [22]. Besides, several experiments, such as those cited in Section 5, can be performed and open the way to creation and enrichment of terminologies, lexica and contrastive studies involving Ukrainian language.

Another direction for future work consists of creation of parallel corpora, in which Ukrainian is the source language.

In order to make the alignment process and verification easier, we will test and exploit automatic sentence alignment tools. Currently, only one human annotator (NLP researcher) is involved in the building of corpus. If several human annotators are involved in the manual alignment, we will be able to compute the inter-annotator agreement, which will be indicative of the sophistication and difficulty of this task.

This sentence-aligned corpus is freely available for the research purposes: http://natalia.grabar.free.fr/resources.php

References

1. Babych, B.: Representation and interpretation of ambiguous deep syntactic structures. Ukrainian Linguistics 21, 89--100 (1997), in Ukrainian
2. Banea, C., Mihalcea, R.: Word sense disambiguation with multilingual features. In: Interna-tional Conference on Computational Semantics (ICCS 2011). pp. 25--34 (2011)
3. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: ACL. pp. 597--604 (2005)
4. Benko, V.: Aranea: Yet another family of (comparable) web corpora. In: Text, Speech and Dialogue. pp. 247--256 (2014)
5. Callison-Burch, C., Cohn, T., Lapata, M.: Parametric: An automatic evaluation metric for paraphrasing. In: COLING. pp. 97--104 (2008)
6. Cartoni, B., Namer, F.: Linguistique contrastive et morphologie : les noms en -iste dans une approche onomasiologique. In: CMLF. pp. 1245--1259 (2012)
7. Dimitrova, L., Koseska-Toszewa, V., Garabik, R., Erjavec, T., Iomdin, L., Shyrokov, V.: MONDILEX - Towards the Research Infrastructure for Digital Resources in Slavic Lexi-cography, pp. 147--162 (2010)
8. Erjavec, T.: MULTEXT-East: Morphosyntactic resources for central and eastern european languages. Language Resources and Evaluation 46(1), 131--142 (2012)
9. Grabar, N., Hamon, T.: Acquisition non supervisée de ressources morphologiques en ukrainien. In: Atelier Traitement Automatique des Langues Slaves (TASLA). pp. 1--10 (2015)

10. Grabar, N., Shyshkina, N., Zorko, H., Hamon, T.: Terminological research in ukraine. In: Terminologie et Intelligence Artificielle (TIA) (2015)
11. Hamon, T., Grabar, N.: Acquisition of medical terminology for Ukrainian from parallel cor-pora and Wikipedia. In: Terminologie et Intelligence Artificielle (TIA) (2015)
12. Hantson, A.: English gerund clauses and norwegian det + infinitive / at clause constructions. In: Granger, S., Lerot, J., Petch-Tyson, S. (eds.) Corpus-based Approaches to Contrastive Linguistics and Translation Studies, pp. 75--90. Rodopi, New-York, Amsterdam (2003)
13. Kelih, E.: Preliminary analysis of a slavic parallel corpus. In: Corpus based Grammar research. pp. 173--183 (2009)
14. Kelih, E., Buk, S., Grzybek, P., Rovenchak, A.: Project description: designing and construct-ing a typologically balanced ukrainian text database. In: Методи аналізу тексту. pp. 125--132 (2009)
15. Kok, S., Brockett, C.: Hitting the right paraphrases in good time. In: NAACL. pp. 145--153 (2010)
16. Kotsyba, N.: Polukr (a polish-ukrainian parallel corpus) as a testbed for a parallel corpora toolbox. Philological Studie LXIII, 181--196 (2012)
17. Kotsyba, N.: Overview of the ukrainian language resources within the multilingual european MULTEXT-East project. Інформаційні системи та мережі 770, 122--129 (2013)
18. Kotsyba, N., Mykulyak, A., Shevchenko, I.V.: UGTag: morphological analyzer and tagger for the Ukrainian language. In: Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009) (2009)
19. Lefer, M., Grabar, N.: Evaluative prefixes in translation: From automatic alignment to seman-tic categorization. Linguistic Issues in Language Technology journal 11(6), 169--187 (2014)
20. Lefer, M.A., Grabar, N.: N-grams in multilingual corpora: extracting and analyzing lexical bundles in contrastive studies. In: EUROPHRAS 2015 (2015)
21. Lopez, A., Nossal, M., Hwa, R., Resnik, P.: Word-level alignment for multilingual resource acquisition. In: LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Data. Las Palmas, Spain (2002)
22. Och, F., Ney, H.: Improved statistical alignment models. In: ACL. pp. 440--447 (2000)
23. Siruk, O., Derzhanski, I.: Linguistic corpora as international cultural heritage: The corpus of Bulgarian and Ukrainian parallel texts. Digital Presentation and Preservation of Cultural and Scientific Heritage 3, 91--98 (2013)
24. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: HLT (2001)
25. Ziering, P., van der Plas, L., Schütze, H.: Multilingual lexicon bootstrapping. Improving a lexicon induction system using a parallel corpus. In: International Joint Conference on Natural Language Processing. pp. 844--848 (2013)
26. Бобкова, Історичні та концептуальні передумови корпусної лінгвістики. Філологічні науки 2, 13--17 (2014)
27. Бугаков, О.: Создание семантического словаря предложных конструкций на основе украинского национального лингвистического корпуса. Tech. rep., Украинский языково-информационный фонд НАН Украины, Киев, Украина (2006)
28. Бук, Ровенчак, Частотний словник роману Івана Франка "Перехресні стежки", pp. 138--369 (2007)
29. Бук, Лінгводидактичний потенціал корпусу текстів Івана Франка у викладанні української мови як іноземної. In: Theory and Practice of Teaching Ukrainian as a Foreign Language. pp. 70--74 (2010)
30. Бук, Сучасні методи дослідження мови письменника у слов'янознавстві. Проблеми слов'янознавства 61, 86--95 (2012)
31. Глибовец, А., Решетнев, І.: Метод итеративного построения терминологии в коллекциях научных текстов на украинском языке. Кибернетика и системний анализ 50(6), 53--62 (2014)

32. Дарчук, Н.: Морфологічне анотування Корпусу української мови. In: Комп'ютерна лінгвістика: сучасне та майбутнє. pp. 16--18 (2012)
33. Дарчук, Дослідницький корпус української мови: основні засади і перспективи. ВІСНИК Київського національного університету імені Тараса Шевченка 21, 45--49 (2010)
34. Демська, О.: Текстовий корпус: ідея іншої форми. ВПЦ НаУКМА, Київ, Україна (2011)
35. Демська-Кульчицька, О.: Репрезентативність як ознака текстового корпусу. Українська мова 3, 100--107 (2005)
36. Левченко, О., Кульчицький, І.: Технологія перетворення п'ятимовного словника порівнянь в електронну форму. In: Інформаційні системи та мережі. pp. 129--138 (2013)
37. Монахова, Т.: Застосування прийомів корпусної лінгвістики в лексикографії. Наукові праці 98(85), 55--60 (2009)
38. Сірук, Підготовка діалектних текстів для корпусного опрацювання. In: Комп'ютерна лінгвістика: сучасне та майбутнє. pp. 43--45 (2012)
39. Тищенко, Засади створення корпусу української жестової мови глухих. Лексикографічний бюлетень 13, 47--52 (2006)
40. Шуневич, Українсько-англійський комп'ютерний словник пожежно-технічних термінів: лексичні матеріали, програмне забезпечення. In: Комп'ютерна лінгвістика: сучасне та майбутнє. pp. 46--48 (2012)

Unsupervised acquisition of morphological resources for Ukrainian

Thierry Hamon[1] and Natalia Grabar[2]

[1] LIMSI-CNRS, Orsay, Université Paris 13, Sorbonne Paris Cité, France

hamon@limsi.fr

[2] CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France

natalia.grabar@univ-lille3.fr

Abstract. Availability of morphological resources is an important and recurrent need because they allow the development of NLP tools and applications for a given language. Indeed, such resources provide basic information which is necessary for such tools for performing more sophisticated treatments (information retrieval, morpho-syntactic tagging, etc). We propose to acquire morphological resources for Ukrainian language. The method proposed exploits corpora in order to extract words that are related morphologically between them. The method has two versions: without and with processing of prefixes. The association strength between these words indicates their probability to have a morphological and se-mantic relation between them. We use three corpora (literary, medical and general-language) and evaluate the results obtained. According to the corpora, precision varies between 67% and 86%. The results from different corpora are also com-pared, which shows that there is little redundancy between the corpora. The currently available resource contains 3,315 fully validated pairs of words.

1 Introduction

Morphological resources provide basic and crucial knowledge upon which many Natural Language Processing (NLP) applications are built. During the *Part-of-Speech tagging and the lemmatization*, morphological lexicon helps to analyze words and to recognize inflected forms of a given word and then to deduce its lemma. Thus, in morphologically rich languages, the recognition of affixes and of inflections allows to disambiguate and to deduce the Part-of-Speech category. In *information retrieval and extraction*, the needs and the goals go beyond the inflectional morphology. Indeed, such applications often re-quire the identification of relations between derivations and even compounds. Generally, this information is helpful to collect higher number of relevant answers or documents, and then to increase the recall of automatic systems. The *processing of unknown words* is relevant for many NLP applications because the existing dictionaries or lexical resources are known to be uncompleted. In that

respect, if morphological information on words is available, it can be useful to induce the grammatical or syntactical categories, as well as the semantics. In *speech recognition*, the resources providing groups of words with the same morphological root, are useful for the disambiguation of a spoken sequence and for selection of the most relevant candidate.

Nowadays, such resources are available and widely used in several languages, e.g. CELEX [4] for German, English and Dutch, Démonette [14], lexique.org[8] and Lefff [27] for French, Morph-it [34] for Italian. Those resources provide inflected information associated with words, such as singular and plural forms of noun ({*president*, *presidents*}), adjectives ({*présidentiel*, *présidentielle*}) or verbs {*preside*, *presided*}. It is less common to find resources that also provide relations between derivational forms ({*president*, *presidential*}) or between compounds and their basis ({*president*, *presidoscopy*}). Moreover, the general-language morphological resources often show partial coverage in specialized domains because such documents involve specific lexicon.

Various methods have been proposed to acquire morphological resources. Among them, various kinds of information can be used in isolation or combined: associations between words in corpora [33, 35]; distributional properties of words in corpora [5]; distribution of letters in words in order to identify frontiers of morphemes [7, 32, 28]; analogy between the word formation in order to deduce or generate new constructed words [24, 12, 15]; frequency of the suffix couple, which insures the reliability of the semantic link between two words [10]; exploitation of dictionaries in order to identify words which are semantically and morphologically related within a given entry [19, 15]; semantically related pairs of terms [12]; samples used by supervised methods in order to induce morphological rules [2, 30, 24].

Several tools are available for morphological analysis in several languages: Flemm and Derif for French [23], Morphisto for German[9], tools for Nguni [3, 25], Indian [1], or Macedonian [17]. Building of morphological resources is an active research topic, including specialized and low-resourced languages. Besides, several methods have been proposed for the automatic acquisition of morphological resources and consequently various types of data can be processed for the acquisition of such resources.

In our work, we propose to tackle the creation of morphological resources for Ukrainian. We propose to take advantage of freely available text corpora which are not an-notated syntactically neither semantically. Our method relies on previous works [33, 35] and computes corpus-based associations between words. However, several adaptations are performed to take into account particularities of the Ukrainian language: text encoding, text segmentation and morphological specificities.

In the following, we first propose a description of Ukrainian language and mention some existing works on this language (Section 2). We then present the material in Section 3 and the method in Section 4. Results are described and discussed in Section 5. We conclude and indicate some future works (Section 6).

[8] *www.lexique.org*
[9] *https://code.google.com/p/morphisto*

2 Specificities of the Ukrainian Language

Ukrainian language is part of the Slavic family of languages. It uses Cyrillic alphabet composed of 33 letters and of apostrophe. One particularity of Ukrainian is that the apostrophe plays phonetic role and is not a word separator.

Similarly to other Slavic languages, Ukrainian has a rich inflectional morphology, with seven cases and three genres for common and proper names, numerals, adjectives pronouns and some verbal forms. The derivational morphology plays a key role in the building of the lexical and grammatical structures (e.g. aspect, tense). To illustrate the word creation, we present set of words coined on *walk* (in (1)).

(1) *хід* (*walk*), *вхід* (*entrance*), *вихід* (*exit*), *захід* (*East, sunset, event*), *прихід* (*arrival*), *перехід* (*crossing, cross walk*), *відхід* (*start, departure*), *підхід* (*approach*), *дохід* (*approach still closer of the goal, incomes*), *прохід* (*passing through an obstacle such as woods or a hedge*), *обхід* (*passing by, walk around*)

Thanks to the rich morphology, even if sentences obey to the canonical order subject-verb-object (SVO), the word order is free without introducing stylistic effects. Ambiguities exist at lemma and inflection levels, and it is common to find inflected forms which correspond to different lemmas. These particularities may lead to difficulties with the classical NLP methods, and above all with the POS-tagging. However, they can also facilitate the sentence parsing since inflected information provides useful clues for this task [6]. Concerning the NLP work, we can mention some existing works: since 2010, the Ukrainian language is integrated in the Multex-East POS tagset[10] [9]; UGtag POS tagger has been developed [18]. It exploits dictionary and rules, but does not perform syntactic and morphological disambiguation of words; recognition of named entities [16]; sentiment analysis [26]. As for the corpora, there were endeavor to build the Ukrainian National corpus[11], Polish-Ukrainian [31] and Bulgarian-Ukrainian [29] parallel corpora. However the access to these corpora is restricted and, to our knowledge, there is no free existing corpora or lexicon. Our objective is to contribute to the development of NLP methods and resources dedicated to the Ukrainian language. Such methods and resources are necessary for boosting the development of NLP applications.

3 Linguistic Resources

We use three types of resources: (1) corpora (Section 3.1) that allow to acquire morphological resources; (2) set of stopwords (Section 3.2) used in order to remove grammatical and invariable words, and (3) set of prefixes (Section 3.3).

[10] *http://nl.ijs.si/ME/V4/*
[11] *http://ulif.org.ua/*

3.1 Corpora

The corpora are issued from three sources and represent three different genres:

-- *literary corpus*, composed of texts from Taras Shevchenko's Kobzar (89,289 words);
-- *medical corpus*, composed of medical articles and brochures issued from Medline-Plus [22] (46,230 words). We use those that are translated in Ukrainian;
-- *general corpus*, composed of articles from the Ukrainian Wikipedia pages (1,201,585 articles totaling 246,368,411 words are available in the used version).

Obviously, Wikipedia is the biggest corpus while MedlinePlus is the smallest.

3.2 Stopwords

We use a set of 385 stopword forms issued from an existing resource dedicated to the localization of graphical interfaces[12], such as in (2). However, we observe that this list is incomplete and needs to be augmented through the corpora analysis.

(2) *зі* (*with*), *ми* (*we*), *на* (*on*), *та* (*and*), *ти* (*you*), *ще* (*still*), *що* (*that/what*), *їй* (*to her*), *їм* (*to them*)

3.3 Set of Prefixes

The set of 73 prefixes is issued from the existing dictionary [36]. An example is given in (3) together with approximate translations. The prefixes are used for associating words with common bases that may occur after these prefixes, such as in Examples (1)).

(3) *без* (*without*), *від* (*from*), *екстра* (*extra*), *з* (*perfective meaning*), *за* (*behind*), *до* (*up to*), *об* (*around*), *най* (*the most*), *пере* (*re*), *понад* (*over*)

4 Approach for Building Morphological Resources

The corpora are first pre-processed (Section 4.1). We then apply our method to extract morphologically and semantically related pairs of words (Section 4.2), and to process the prefixes (Section 4.3). Besides, the reliable morphological rules from the first set of validated resources are used to prevalidate some of the remaining word pairs (Section 4.4). Finally, the evaluation of the acquired word pairs is performed (Section 4.5).

[12] *https://github.com/fluxbb/langs/blob/master/Ukrainian/stopwords.txt*

4.1 Corpus Pre-processing

The corpora are first converted in UTF-8, then a word and sentence tokenization is done. This step takes into account the role of the apostrophe character: inside words, it is not considered as word separator, while at frontiers of words, it shows its traditional quote meaning. Empty words are removed to avoid noise generation during the next step.

4.2 Extraction of Morphologically Related Pairs of Words

The method aims the identification of words which are related morphologically and semantically. We use for this the notion of thematic continuity. Indeed, thematic links exist at the lexical level within a piece of text: words and lexemes from a given semantic field tend to be used together (e.g., *hospital*, *physician*, *operate*). Consequently, words from the same morphological family may also co-occur (e.g., *operate*, *operation*). This provides the possibility to automatically find morphologically-related words in corpora.

Similarly to previous works [35], the notion of thematic continuity is approximated with a graphical window of *W* words. The morphological proximity between two words is then identified through the *n* first characters of each word. To summarize this first method (henceforth, standard method) [11], we collect words which share the same initial string with a length superior to n characters and which co-occur in the same window of W words. This last criteria will be measured with a statistical association measure which measures the frequency of the co-occurrence in comparison to a random association. We use the likelihood ratio [21] i.e. the ratio $\lambda = \frac{L(H_1)}{L(H_2)}$ between the probability to observe the number of co-occurrences of the word w_1 and the word w_2 according to the hypothesis H_1 where words are independent and the probability to observe the number of co-occurrences according to the hypothesis H_2 where words are dependent each other (we compute $-2\log\lambda$). This ratio is computed as follows:

-- Probability to observe the H_1 hypothesis (independence):
$L(H_1) = b(c_{12}, c_1, p)\ b(c_2 - c_1, N - c_1, p)$;

-- Probability to observe the H_2 hypothesis (dependence):
$L(H_2) = b(c_{12}, c_1, p_1)\ b(c_2 - c_1, N - c_1, p_2)$;

-- Binomial law (probability of a sequence of k success among n draw):
$b(k, n, p) = C_k^n p^k (1-p)^{n-k}$;

-- Elementary probabilities:

- $p = \frac{c_{12}}{N};\ p_1 = \frac{c_{12}}{c_1};\ p_2 = \frac{c_2 - c_{12}}{N - c_1}$;
- c_1 is the number of occurrences of the word w_1,
- c_2 is the number of the windows where the word w_2 occurs,
- c_{12} is the number of windows in which w_1 and w_2 occur,

- N is the number of words of the corpus.

This association measure is asymmetric and depends on frequency of each word. For instance, there is a higher probability to observe a noun such as *canal* in the neighbor of its adjective *canalized* then the opposite. Given the two possible directions, the higher association score is kept. We process and evaluate the proposed pairs independently on their scores: even pairs with low scores may convey correct morphological relations.

The proposed method is applied on the three corpora. We use a window of 10 words on the left and on the right of the pivot word ($W = 21$). The minimal length of the initial string is fixed to 3 characters ($c = 3$) because it allows to keep pairs which may share common roots or bases.

4.3 Processing of Prefixes

Prefixes are very frequent in Ukrainian and play an active role in word formation. Pre-fixes may introduce two problems with the standard method:

-- they prevent from associating words with the same basis, which are preceded by such prefixes, such as in Examples (1);
-- they associate words that do not have the same bases (and that have no morphological or semantic relations) but share only the prefix, such as in Examples (4);

(4) {заплануйте; запізнюйтесь} ({*to plan*, *being late*}), {відповідає; відстань} ({*answer*, *don't bother*}), {переставляйте; перевірте} ({*move*, *verify*})

In the modified version of the method, we propose to inhibit the prefixes and to focus on the bases of words, event if they occur after these prefixes. In this modified version of the method, the prefixes undergo the following processing:

-- the known prefixes (Section 3.3) are temporarily removed from the words within a given word pair, starting from the longest prefix: for instance, *за* (*behind*) is applied before *з* (*perfective meaning*);
-- the standard method (Section 4.2) is applied to the remaining strings, with the minimal initial string set to 3 characters ($c = 3$): *закритий* (*closed*), *відкритому* (Dative of *open*), *відомим* (*unknown*) temporarily become *критий*, *критому* and *омим*;
-- if a given pair contains the common string with at least 3 characters, the prefixes are restored and this word pair is kept as candidate. Thus, the pair {*закритий*, *відкритому*} is now a candidate, while the pair {*відкритому*, *відомим*} is not proposed. Notice the latter word pair is proposed by the standard method, but not by the modified method. On the contrary, the modified method taking into account the prefixes proposes the former word pair but not the latter one, as expected.

4.4 Rule-based Prevalidation

The proposed method induces very large number of word pairs. Their manual validation is a very tedious task. We propose to exploit the first set of the validated word pairs for prevalidating the remaining word pairs. We use for this the morphological rules issued from these validated pairs. For instance, the word pair {*брат*, *брата*} (*brother)* provides the morphological rule /a, for the Genitive inflection of nouns. If this rule is always reliable in the validated dataset, then it can be used safely for prevalidating other word pairs that show the same rule, such as {*імператор*, *імператора*} (*imperator*), {*благовєщенськ*, *благовєщенська*} (*Blahoveshesk*), {*трилисник*, *трилисника*} (*clover*). We use rules that have at least three correct occurrences in the validated dataset.

4.5 Evaluation

The results are evaluated manually by a native speaker. For the evaluation of the results, the task is to check whether the words from a given pair share the same morphological basis and are morphologically related.

5 Results

The corpora are pre-processed and the proposed method permits to extract a large set of word pairs which are expected to be morphologically related. In Table 1, we indicate

Table 1. Number of word pairs and precision

Corpora	Standard		Modified/Prefixes	
	# pairs	Precision	# pairs	Precision
Kobzar	2,546	76,4%	2,550	75,6%
MedlinePlus	1,961	68,8%	1,757	76,7%
Wikipedia (evaluated sample)	29,968	65,4%	--	--

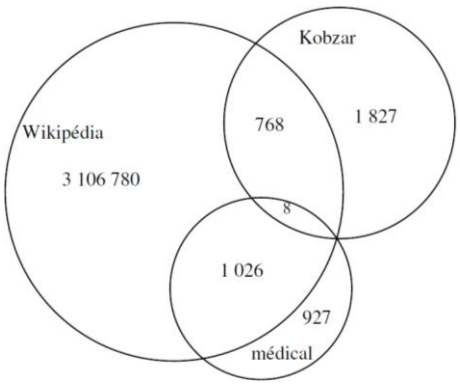

Fig. 1. Intersection between the resources acquired on the three corpora

the numbers of the evaluated word pairs. Due to the large number of extractions from Wikipedia (3,108,591) only a sample of 29,968 word pairs is evaluated by now. Precision obtained varies between 65% on Wikipedia, and up to 75% on literary and 77% on medical corpora. Our results are comparable with those obtained in a previous work on the medical French [35]. The second set of results is obtained with the version of the modified method dealing with prefixes. We can see that this modification of the method is suitable for the quality of the results: precision is improved on medical corpus and shows a slight decrease on literary corpus. Several new and correct word pairs are extracted. Same processing will be applied to Wikipedia corpus. The validated set (31,476 pairs) provides 23,326 correct words pairs and will be made available for the research community.

Our results also show that the method can be used in different languages when corpora are available, in order to bootstrap acquisition of morphological resources. Hence, validated word pairs permit to induce 1,176 morphological rules with at least 3 correct occurrences. These rules have been applied to the remaining set with 2,991,890 word pairs and permitted to prevalidate 723,553 word pairs. The first analysis of these pairs indicates that they are correct. Such approach allows to increase the size of the available resource.

In Figure 1, we present the coverage between the resources acquired from the three corpora. We can observe that the intersection is low and that most of the pairs are issued from Wikipedia. 52.6% of word pairs acquired on the medical corpus are also acquired on Wikipedia, while this ratio represents 29.6% of pairs from the literary corpus. Only 8 word pairs are acquired on the three corpora. They correspond to the inflected forms of words (Example (5)). This suggests that several corpora should be processed to reach a good coverage of morphological resources.

(5) {*руки*, *руку*} (*arm*), {*серця*, *серцем*} (*heart*), {*кров*, *крові*} (*blood*), {*ліжка*, *ліжку*} (*bed*), {*новими*, *нові*} (*new*), {*одна*, *одну*} (*alone*), {*стало*, *стали*} (*become*), {*кров*, *кров'ю*} (*blood*)

Some of the acquired pairs contain ambiguous words which can correspond to different lemmas according to the domain. The following examples illustrate this point:

-- {*поділися*, *поділось*}: this pair contains forms of the verb *to disappear*. However, the word *поділися*, with a different stress accent, corresponds to the verb *to share*;
-- the main meaning of the pair {*дітись*, *діти*} is *to put somewhere*. However, the word *діти* is also an inflected form of *child*;
-- the main meaning of the pair {*гори*, *горить*} is *to burn* while *гори* is also an inflected form of *mountain*.

We consider that such pairs are correct because at least one of their meanings is correct. Among the correct pairs, an important part contains inflections, even if lemmas occur seldom. We also observe derivations (Examples (6)) and compoundings (Examples (7)).

(6) {*алергійна, алергія*} ({*allergic, allergy*}), {*братерська, брате*} ({*brotherly, brother*}), {*вакцинацію, вакцина*} ({*vaccination, vaccine*}), {*дитину, дитячий*} ({*child, childish*})

(7) {*ангіопластика, ангіограми*} ({*angioplasty, angiogram*}), {*бронхіоли, бронхіт*} ({*bronchiole, bronchitis*}), {*газованих, газоутворення*} ({*gaseous, production of gas*})

By comparison with French or English, we identify two specific morphological phenomena in Ukrainian: diminutive forms such as those presented in (8), patronymic forms such as those presented in (9).

(8) {*ангеляточко, ангел*} *(angel)*, {*біленькі, білих*} *(white)*, {*Богданочку, Богдане*} *(Bohdan)*, {*воленьки, волі*} *(freedom)*, {*годину, годиночку*} *(hour)*

(9) {*Іван, Іванович*} ({*Ivan, son of Ivan*}), {*Микола, Миколайович*} ({*Mykola, son of Mykola*})

Main errors occur when words with the same initial string have no morphological or semantic relations (Examples (10)), and in case of allomorphies which occur within the initial string (the first c = 3 characters) such as in (11).

(10) {*криза, криму*} ({*crisis, Crimea*}), {*проблем, прокурорської*} ({*problem, prosecutor (adj)*})

(11) {*хід, хода*} *(walk)*, {*воля, вільний*} ({*freedom, free*})

6 Conclusion and Future Work

We presented an approach for the acquisition of morphological resources for Ukrainian. An unsupervised method is proposed, which does not require annotations or dedicated resources and only relies on the use of corpora. Two versions of method are designed and tested: standard method and modified method with the processing of prefixes. Statistical association metrics between words are used to assess the probability of semantic and morphological relations between words. Our approach has been applied to three Ukrainian corpora: literary, medical and general language. For now, a set of 23,326 word pairs have been judged correct. 723,553 more word pairs are prevalidated with reliable morphological rules. This set will be progressively enriched with more validated data and made freely available for the research purposes. The method allows to acquire word pairs with precision which varies between 65% and 77% according to corpora.

One limitation is related to allomorphy which occurs within the initial string. Specific methods or rules will be tested for the processing of such situations. Another problematic point is that manual evaluation is a time-consuming task. To reduce the validation time, we plan to use other association and statistical measures [13, 20], and metrics from the graph theory [8]. Morphological rules induced with the method can

also be used for enriching this resource. We plan to use this resource for POS-tagging and in-formation retrieval.

References

1. Abeera, V., Aparna, S., Rekha, R., Kumar, M., Dhanalakshmi, V., Soman, K., Rajendran, S.: Morphological analyzer for Malayalam using machine learning. Data Engineering and Management, LNCS 6411, 252--254 (2012)
2. van den Bosch, A., Daelemans, W., Weijters, T.: Morphological analysis as classification: an inductive-learning approach. In: International Conference on Computational Linguistics (COLING) (1996)
3. Bosch, S., Pretorius, L., Fleisch, A.: Experimental bootstrapping of morphological analysers for Nguni languages. Nordic Journal of African Studies 17(2), 66--88 (2008)
4. Burnage, G.: CELEX - A Guide for Users. Centre for Lexical Information, University of Nijmegen (1990)
5. Claveau, V., Kijak, E.: Generating and using probabilistic morphological resources for the biomedical domain. In: Proceedings of the Ninth International Conference on Language Re-sources and Evaluation (LREC'14). pp. 3348--3354 (2014)
6. Collins, M., Hajic, J., Ramshaw, L., Tillmann, C.: A statistical parser for czech. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. pp. 505--512. Association for Computational Linguistics, College Park, Maryland, USA (June 1999), http://www.aclweb.org/anthology/P99-1065
7. Déjean, H.: Morphemes as necessary concept for structures discovery from untagged corpora. In: Workshop on Paradigms and Grounding in Natural Language Learning. pp. 295--299. Adelaide (1998)
8. Diestel, R.: Graph Theory. Springer-Verlag Heidelberg, New-York (2005)
9. Erjavec, T.: MULTEXT-East: Morphosyntactic resources for central and eastern european languages. Language Resources and Evaluation 46(1), 131--142 (2012)
10. Gaussier, E.: Unsupervised learning of derivational morphology from inflectional lexicons. In: Kehler, A., Stolcke, A. (eds.) ACL workshop on Unsupervised Methods in Natural Lan-guage Learning. College Park, Md. (Jun 1999)
11. Grabar, N., Hamon, T.: Acquisition non supervisée de ressources morphologiques en ukrainien. In: Atelier Traitement Automatique des Langues Slaves (TASLA). pp. 1--10 (2015)
12. Grabar, N., Zweigenbaum, P.: Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In: Traitement Automatique de Langues Naturelles (TALN). pp. 175--184 (1999)
13. Hamon, T., Engström, C., Manser, M., Badji, Z., Grabar, N., Silvestrov, S.: Combining com-positionality and pagerank for the identification of semantic relations between biomedical words. In: BIONLP NAACL. pp. 109--117 (2012)
14. Hathout, N., Namer, F.: La base lexicale Démonette: entre sémantique constructionnelle et morphologie dérivationnelle. In: TALN. pp. 208--219 (2014)
15. Hathout, N.: Analogies morpho-syntaxiques. In: Traitement Automatique des Langues Naturelles (TALN). Tours (2001)
16. Katrenko, S., Adriaans, P.: Named entity recognition for Ukrainian: A resource-light approach. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. pp. 88--93. Association for Computational Linguistics, Prague, Czech Republic (June 2007), http://www.aclweb.org/anthology/W/W07/W07-1712
17. Kostov, J.: Le verbe macédonien : pour un traitement informatique de nature linguistique et applications didactiques (réalisation d'un conjugueur). Thèse de doctorat, INaLCO, Paris, France (2013)

18. Kotsyba, N., Mykulyak, A., Shevchenko, I.V.: UGTag: morphological analyzer and tagger for the Ukrainian language. In: Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009) (2009)
19. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 191--202 (1993)
20. Loukachevitch, N., Nokel, M.: An experimental study of term extraction for real information-retrieval thesauri. In: TIA. pp. 1--8 (2013)
21. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA (1999)
22. Miller, N., Lacroix, E., Backus, J.: MEDLINEplus: building and maintaining the national library of medicine's consumer health web service. Bull Med Libr Assoc 88(1), 11--7 (2000)
23. Namer, F.: Morphologie, Lexique et TAL : l'analyseur DériF. TIC et Sciences cognitives. Hermes Sciences Publishing, London (2009)
24. Pirrelli, V., Yvon, F.: The hidden dimension: a paradigmatic view of data-driven NLP. JETAI 11, 391--408 (1999)
25. Pretorius, L., Bosch, S.: Exploiting cross-linguistic similarities in Zulu and Xhosa computa-tional morphology. In: AFLAT. pp. 96--103 (2009)
26. Romanyshyn, M.: Rule-based sentiment analysis of ukrainian reviews. International Journal of Artificial Intelligence & Applications (IJAIA) (2013)
27. Sagot, B., Clément, L., Villemonte de la Clergerie, E., Boullier, P.: The Lefff 2 syntactic lexicon for french: architecture, acquisition, use. In: LREC (2006)
28. Schone, P., Jurafsky, D.: Knowledge-free induction of inflectional morphologies. In: Work-shop NA de ACL (2001)
29. Siruk, O., Derzhanski, I.: Linguistic corpora as international cultural heritage: The corpus of Bulgarian and Ukrainian parallel texts. Digital Presentation and Preservation of Cultural and Scientific Heritage 3, 91--98 (2013)
30. Theron, P., Cloete, I.: Automatic acquisition of two-level morphological rules. In: ANLP. pp. 103--110 (1997)
31. Turska, M., Kotsyba, N.: Polish-ukrainian parallel corpus and its possible applications. In: GmbH, P.L. (ed.) Practical Applications in Language and Computers. Lódź (April 2007)
32. Urrea, A.M.: Automatic discovery of affixes by means of a corpus : a catalog of Spanish affixes. Journal of quantitative linguistics 7(2), 97--114 (2000)
33. Xu, J., Croft, B.W.: Corpus-based stemming using co-occurrence of word variants. ACM Transactions on Information Systems 16(1), 61--81 (1998)
34. Zanchetta, E., Baroni, M.: Morph-it! a free corpus-based morphological resource for the ital-ian language. Corpus Linguistics 2005 1(1) (2005)
35. Zweigenbaum, P., Hadouche, F., Grabar, N.: Apprentissage de relations morphologiques en corpus. In: Traitement Automatique des Langues Naturelles (TALN) (2003)
36. Клименко, Карпіловська, Карпіловський, Недозим, Словник Афіксальних Морфем Української Мови. Інститут Мовознавства ім. О.О. Потебні Національної Академії Наук України, Київ, Україна (1998)

NLP Resources for a Rare Language Morphological Analyzer: Danish Case

Mykhailo Kotov (orcid.org/0000-0001-8327-5197)

Envion Software / V.N. Karazin Kharkiv National University, Kharkiv, Ukraine

mykhailo.kotov@gmail.com

Abstract. The paper discusses the characteristics and practical aspects of application of the natural language processing resources available for developing a rare language morphological analysis solution. The case under consideration reveals the pipeline design needed to prepare the grammatical resources for Danish. Being rare not only in terms of distribution, but also in the amount of natural language resources available, the Danish language represents a significant problem in terms of application of third-party tools to help solve various NLP-related issues. The paper focuses on part-of-speech tagging and lemmatization, typical but indispensable tasks at the pre-processing stage within the framework of developing a morphological analyzer as a custom NLP solution.

Keywords: morphological analyzer, lemmatization, part-of-speech tagging, Hunspell, OpenNLP, Snowball stemmer, SyntaxNet, word-list.

1 Introduction

Developing a morphological analyzer is a multi-staged complex process, whose starting point is preparation of word-list(s) with indicated grammatical meanings of entries. Such word-list(s) preparation includes the essential pre-processing measures – tokenizing, cleaning, permuting, part-of-speech (POS) tagging, lemmatizing or stemming. At the same time, this initial stage, aiming at obtaining a "lemma-derivative-grammatical meaning" format out of an unsortedword-list rich in inflected and suppletive forms, appears to be resource-consuming and demands cost-efficient solutions and optimization.

The task of proper word-list processing becomes no less difficult in case of its implementation for a rare language. In our classification, rare is a language with not only few speakers and/or limited distribution [8], but also, what is more important, a limited number of natural language processing (NLP) and linguistic resources available.

In this paper, the focus is on Danish – a language rare in terms of availability of NLP solutions. Being a typical representative of the group of analytic Germanic

languages, Danish is characterized by a decent number of inflected forms for all parts of speech, among which the postpositive usage of the definite article with nouns (e.g. *en hund – hund**en***) is of special interest. Certain word-forms are marked by suppletion (e.g. *god – bedre – bedst*). In Section 2, POS tagging resources available for Danish are analyzed and compared. Section 3 deals with the description of properties of the lemmatizing/stemming solutions. The criteria used for the analysis and comparison includethe declared quality (per token accuracy), operating system (OS) compatibility, and licensing policy. Finally, Section 4 outlines a possible approach on how to handle the word-list processing issue. In the outline, nonetheless, I do not concentrate in detail on the cleaning and permuting, as to a large extent the above can be accomplished without using any additional external solutions.

2 Part-of-Speech Tagging Solutions for the Danish Language

Being critical for almost every natural language processing system, POS tagging is still receiving a great deal of attention. Traditionally, based on the type of information in use, several approaches to implementation of POS tagging solutions are distinguished: rule-based, stochastic/probabilistic, and the combination of the two (hybrid). The first type, as evident from its name, is based on the linguistic models which "range from a few hundreds [sic] to several thousand rules, and they usually require years of labour" [11]. The prototypical representative of the rule-based approach among POS-taggers is Brill's tagger [4]. The stochastic and hybrid approaches are becoming more popular nowadays, in particular in spheres concerning but not limited to neural networks application [1, 9].

Considering the Danish language, it is important to point to the fact that the part-of-speech resources available are to a large extent based on stochastic approaches and represented by domain leaders. These are Apache OpenNLP POS-tagger and SyntaxNet by Google. Among the rule-based taggers, it is necessary to mention Brill's adapted CST's POS tagger [5].

Apache OpenNLP project provides models for part-of-speech tagging. The OpenNLP POS tagger uses a probability model in order to predict the precise part-of-speech tag out of the tag-set. In order to limit the possible tags, one can make use of a tag dictionary aiming as well at increasing the tagging precision and runtime performance. For testing, it is advised, to try out the part-of-speech tagger via the command line tool. But the API is also available for embedding into an application. Licensed under the Apache License, OpenNLP POS tagger shows decent results when language models match the input text, and the latter is correctly decoded [3]. The tagger is OS-independent. Judging by tests [6], if pre-trained on the training part of the Danish Dependency Treebank, with part-of-speech tags converted to the Google universal tag-set, the POS-tagger can achieve the accuracy of 96.8% on the test portion of the Danish Dependency Treebank.

Another novel representative of the stochastic approach family is Google's SyntaxNet, an open-source implementation of the method discussed in [1]. SyntaxNet has been integrated in the TensorFlow framework and accompanied by ParseyMcParseface parser, "tuned for a balance of speed, simplicity, and accuracy". For the latter, there is a set of syntactic models available. The models are pretrained

on Universal Dependencies datasets. The Danish language model boasts the accuracy of 95% over all tokens, including punctuation. Despite the high quality of the performance and favorable licensing policy (available under Apache License), SyntaxNet's limitations are connected with OS compatibility. At present, as integrated into the TensorFlow framework, the solution is functioning under UNIX systems.

CST's POS-tagger, available under GNU General Public License, is represented by the adapted version of Brill's tagger. As stated, the distribution comprises Brill's original; distribution and the archive with CST's software adaptations (reformatting to C++ standard, better handling of capitals in headings, making the source code independent of language etc.) [12]. The tagger was trained on DSL's publicly accessible PAROLE Corpus. In terms of architecture, as stated in [3], CST's tagger is characterized by the restricted access to a web version only, which can hardly be suited for a large amount of text.

The overall comparison of the discussed solutions is represented in Table 1.

Table 1. Comparison of POS-taggers for the Danish language

Solution	**Accuracy**	**OS Compatibility**	**Licensing Policy**
OpenNLP	96.8%	any	Apache License
SyntaxNet	95%	Unix	Apache License
CST's	N/A	any	GPL License

3 Stemming and Lemmatization Solutions for the Danish Language

A different but equally important procedure within the pipeline for a word-list processing is returning the base word-form. Such "dictionary" word-forms can be obtained from two similar but at the same time different in nature processes – stemming and lemmatization. As given in [10], "Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma."

The choice between the solutions is predetermined by the available resources and specification, and often can be characterized by the fact that stemming solutions are easier to get and make use of; they tend to increase recall, but hurt precision. However, lemmatizers, being more precise, but thus more "knowledgeable" and complex in design, are under stricter limitations in terms of licensing policy, let alone the case with rare languages. The available and noteworthy solutions for the Danish language are Hunspell, CST's lemmatizer, and NLTK compatible Snowball stemmer.

A well-known solution for primarily spelling checking, Hunspell has proven efficient as a stemmer with varying degrees of accuracy for different languages. Available as GPL/LGPL/MPL tri-license, Hunspell, at the same time, can be used with any type of OS and employs Unicode character encoding. Hunspell delivery includes the dictionary files: one with base forms and links to declension/conjugation

patterns, and the other describing the mentioned patterns to generate derived and inflected forms for lemmas.

The Snowball stemmer, or, to be more precise, stemming algorithm, together with the stop word list and Snowball compiler, represent, in case of Danish, a hardly viable alternative for Hunspell. Although the major advantage is compatibility with the Natural Language Tool Kit (NLTK) and BSD licensing policy, which eases largely the application of the stemmer.

Finally, CST's lemmatizer solution [7] consists of a rules set and a dictionary. The rules set is derived from the Large Computational Dictionary, STO, and the verification of the output boasts 94%-98% accuracy, depending on whether the entries have been supplied with the proper grammatical meanings or not. CST's lemmatizer is freely available for non-commercial applications, but special permission is required for commercial usage.

The overall comparison of the discussed stemming and lemmatizing solutions is represented in Table 2.

Table 2. Comparison of Stemmers and Lemmatizers for the Danish language

Solution	Accuracy	OS Compatibility	Licensing Policy
Hunspell	N/A	any	GPL/LGPL/MPL tri-license
Snowball	N/A	any	BSD
CST's	94-98%	any	Non-commercial use – free, special permission for commercial use

4 Word-List Processing Pipeline

As discussed earlier, the final aim of the application of multiple third-party resources is to construct a word-list of a specific form for further usage as a linguistic resource for a custom NLP solution. In this case, a base form (lemma) of a certain lexeme stands next to its derived or inflected word forms, and each and every word form, including lemma, receives a proper grammatical description. Such an outcome makes further manual processing of the linguistic information (verification with subsequent further more elaborated classification) much easier.

The suggested pipeline consists of several stages, which immediately follow one another (Fig 1.). Resulting from the preliminary Stage 0 (tokenization, cleaning, and permuting), as the input for further processing we have a list of tokens, one per line. During Stage 1, the word-list is verified for consistency, minor mistakes and omissions, resulting from Stage 0 and influencing further activities, are fixed.

The next two stages (Stage 2 and 3) are the key for subsequent classification. By applying third-party resources available we add extra features – grammatical meaning, in the form of part-of-speech tags, and base word forms for each entry. Such extra information significantly enhances our capability for correct automatic juxtaposing within the triplet lemma – derived/inflected word-form – part-of-speech tag.

It is important to mention the fact that Stages 2 and 3 can be iterated using different available resources to reach the optimum for precision and recall. Each stage can be followed by manual revision to tweak the list for the next procedure.

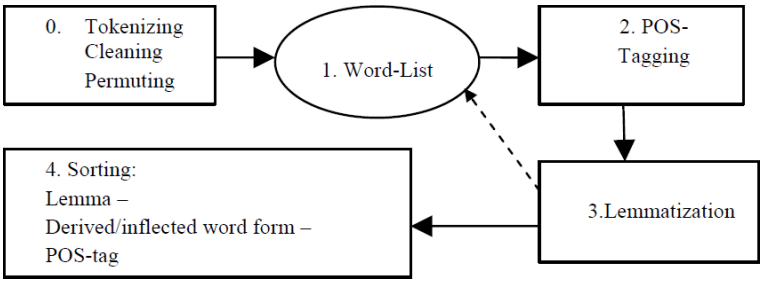

Fig. 1. Pipeline for Word-List Processing

Finally, Stage 4 presupposes sorting procedures. Taking into consideration the extra features obtained resulting from Stages 2 and 3, we build up the list in the desired format. Typically, we first group by lemma, and then by the relevant part-of-speech tag. Lemmatizers, let alone the notorious homonymy issue, may yield nothing at all, leaving an empty row, thus manual revision and verification here are indispensible. In case of stemmers, the output, depending on language specific features and, surely, a stemmer's accuracy, may turn out to be of significantly inferior quality, thus demanding more efforts for revision and improvements introduction.

5 Conclusion

The presented pipeline for processing the word-lists, one of the initial stages in developing a rare language (Danish) morphological analyzer, opens the way for significant reduction of manual labor. This happens due to automation of part-of-speech tagging and lemma ascribing processes.

The list of the available resources and their comparative analysis aimed to help solve the discussed above issue. The assistance consisted in both providing references to existing libraries for part-of-speech tagging, lemmatization or stemming of the Danish language, and indicating possible stumbling blocks for the resource application, e.g. operating system compatibility, licensing policy, accuracy of the solution.

The discussed pipeline is the first step in designing the framework for automation of manual labor and optimization of the available resources allocation.

References

1. Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (p. 2442–2452).
2. *Apache OpenNLP Developer Documentation*. (2017). Retrieved from https://opennlp.apache.org/documentation/1.5.2-incubating/manual/opennlp.html.
3. Asmussen, J. (2015). *Survey of POS taggers. Approaches to making words tell who they are* (Technical Report DK-CLARIN WP 2.1). Retrieved from http://korpus.dsl.dk/clarin/corpus -doc/ pos-survey.pdf

4. Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguist.*,*21*, 543–565.
5. Hansen, D.H. (2000). *Træningogbrugaf Brill-taggerenpådansketekster* (Ontoquery Technical report). Retrieved from https://cst.dk/online/pos_tagger/Brill_tagger.pdf
6. Johannsen, A. (2014). A trainable Part-of-Speech Tagger and Dependency Parser for Danish. Available at: https://github.com/andersjo/danish_dependency_parser/blob/master/README.md
7. Jongejan, B., and Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. (pp. 145–153).
8. Lewis, M.P., Simons, G.F., and Fennig, C.D. (eds.). (2013). *Ethnologue: Languages of the World*. Dallas, Texas: SIL International.
9. Ling, W., Dyer, C., Black, A.W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L, and Luis, T. (2015). Finding function in form: Compositional character models for open vocabulary word representation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1520– 1530).
10. Manning, C.D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press
11. Marquez i Villodre, L. (1999). *Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees* (Doctoral dissertation). Retrieved from https://upcommons.upc.edu/ bitstream/handle/2117/93974/TLMV1de2.pdf
12. taggerXML [Computer software and its adaptations]. Retrieved from http://cst.dk/download/uk/index.html#tagger

Semantic State Superpositions and Their Treatment in Virtual Lexicographic Laboratory for Spanish Language Dictionary

Yevgen Kuprianov

Ukrainian Lingua-Information Fund, NAS of Ukraine, Holosiivskiy av. 3, 03039 Kyiv, Ukraine

cuprijanow.eugen@yandex.ua

Abstract.The paper is devoted to ambiguities of Spanish language units: their formal modelling and treatment in the virtual lexicographic laboratory VLL DLE 23. The final goal is to find optimum solution for lexicographic treatment and research of ambiguities in the laboratory. As a theoretical base for developing ambiguity model, the theory of semantic states was selected. The ambiguity, i.e. the acquisition of different meanings by the unit at the same time in a given context, is represented in the model as a superposition of respective semantic states.Based on literature materials, the formal model of superpositions describing ambiguity formation mechanism in Spanish units was built. The model was further used to make out the interface intended for treating semantic state superpositions in VLL DLE 23.

Keywords: ambiguity, semantic state, superposition, virtual lexicographic laboratory, computer lexicography.

1 Introduction

One of the main problems of computer linguistics and lexicography isdevelopingmethods for language substance modelling. As an objectof modelling can be any unitof phonetic, morphological, lexical and other levels.A special aspectwe'd like to stay on here concerns ambiguitieswhich the lexical units (plain words, collocations) display in speech or text.Studyingambiguities hasbecome an object of research intheoretical [1, 2, 3, 4] and applied (computer) linguistics [5, 6, 7]. The main problems to be touched on by the researchers are: 1) the nature of ambiguity, including its position among other phenomena like homonymy and polysemy; 2) ambiguity classification; 3) ambiguity behavior in different discourses; 4) developing natural-language processing systems to deal with texts which contain ambiguities etc. The problem we'd like to deal with in our paper is ambiguities in the context of creating specialsystem (software) to maintain monolingual dictionaries in

digital environment. This systemis intended for lexicographers to process dictionary material and for scholars to conduct their different investigations based on the dictionary. Conducting lexicographic works and linguistic researches requireselaboration of a special software complex called virtual lexicographic laboratory[13] (herein after VLL). At present Ukrainian lingua-information fund develops a VLL for Spanish monolingual dictionary "Diccionario de la lengua española, 23ª edición" (herein after VLL DLE 23) and additional module to it for treatingand researching ambiguities of Spanish lexical units.

Prior to developing VLL DLE 23 it is important to get a strictly formalized microstructure of the dictionary in question using the theory of semantic states developed by Ukrainian Academic VolodymyrV. A. Shyrokov and his colleaguesand successfully proved on the materials of the Ukrainian language. The main idea laid in this theory is that any monolingual dictionary contains a set of possible semantic states which units can havein a language.When used in a context, collocation, sentence or text, the language unit is supposed to acquire one of semantic states from the set. The formalization of semantic state comes to building the model consisting of grammar (relation to a part of speech, grammar category) and lexical component (semantics) of state as well as additional parameters like homonymy index, context number etc. The whole set of semantic states for any lexical unit, registered in the dictionary, can be represented in the form of a chain. This chain is possible to be reduced to one element owing to the given context in which the unit functions. In this case we can assert that the unit has a "pure" semantic state. However there can be other contexts where it acquires two or more semantic states at the same time. In traditional linguistics such a phenomenon has different names like amphiboly [8], polysemy games [9], language game [10], lexical ambiguity [11, 12, 13] etc. In Shyrokov's interpretation this is a superposition of semantic states in other words, i.e. a chain of two or more elements. The semantic state and its superposition model are described in respective subsections. It should be noted the formal model of semantic state serves us as a basis for developing the database and user interface for VLL DLE 23.

Based on the above the goal of our research is to elaborate solutions concerning the user interface for lexicographic treatment of the superpositions in VLL DLE 23. To achieve the goal we are to: 1) study all possible superposition cases revealed in factual materials; 2) build up a formal model of superpositions, defining its parameters to be accessible through the interface; 3) outline the diagram showing interface with its main components.

1.1 Semantic State and its Formal Model

It's Russian mathematic A. M. Kolmogorov who was the first to introduce the notion of language unit state when attempting to give a formal definition to the case in the Russian language. But he hadn't published his linguistic works; theresults of his

[13] VLL (short for virtual lexicographic laboratory) is a digital environment where a dictionary exists as a language-information object designed to facilitate comprehensive informational description of lexical-grammar structures of a language or a set of languages [15].

research were published later by his disciple V. A. Uspenskiyin his paper [14]. Afterwards they were actually forgotten about.The second life to Kolmogorov and Uspenskiy's conception was given by V. A. Shyrokov in his works [16, 17] whereit got profound developmentas the theory of semantic states. According to this theory, semantic state represents a sum of grammar and lexical semantics and generalizes the notions of grammar and lexical meaning. As a basic statement we'll consider the existence of the correlation between a language unit and its state:

$$s: X \rightarrow (X), \tag{1}$$

Where X is a unit belonging to a certain class of language units, s is the correlation between X and a formal object $s(X)$, which is a content of the unit X. So, it is this object that will be named as semantic state. Let us assume decomposition of sematic state $s(X)$ into grammar and lexical components:

$$s(X) = g(X)l(X), \tag{2}$$

where $g(X)$ is grammar component of the state and $l(X)$ represents lexical meaning of the unit X. Decomposition (2) shows the dichotomy of language sign which is interpreted in traditional linguistics as a relation between form and content of language unit.

Let us analyzethe peculiarities of representing semantic states of Spanish language unit in DLE 23.The principles of their representation have been elaborated by the authors taking into account grammatical and lexical features of headwords: part-of-speech variation, dependence between lexical and grammaticalsemantics, special cases when lexical meaning has a limited use due to some grammar characteristics of a word etc.

For distinguishing grammatical and lexical componentsof semantic statesthe DLE 23 authors adopted the following designation system: 1) two vertical parallel lines ("‖") to separate lexical meanings (definitions) corresponding to one grammatical meaning (part of speech, grammatical category); 2) black circle ("●") to separate blocks of lexical meanings corresponding to different grammatical meanings; 3) white circle ("○") to separate lexical meanings corresponding to some grammatical categories of a headword. The adjectives, adverbs and pronouns are marked as "adj.", "adv." and "pron.", respectively. The nouns are identified with gender and number marks ("m.", "f.", "m. y f.", "m. o f.", "pl."). Fig. 1 shows the example of a DLE 23 entry*cómico*.

cómico, ca. (Del lat. *comĭcus*, y este del gr. κωμικός *kōmikós*).
adj. 1. Que divierte y hace reír. *Situación cómica*. ‖ 2. Perteneciente o relativo a la comedia. ‖ 3. Dicho de un actor: Que representa papeles **cómicos**. U. t. c. s. ‖ 4. Dicho de un autor antiguo: Que escribía comedias. U. t. c. s. ● m. y f. 5. **comediante** (‖ actor). ○ f. 6. *Pan.* **historieta** (‖ serie de dibujos). U. m. en pl. ‖ 7. *Pan.* **dibujos animados**. ■

Fig. 1. Entry of the headword cómico in DLE 23

The headword in consideration has three blocks of lexical meanings; the first is related to adjective (adj.), the second, to the noun of common gender (m. y f.) and the third, to the noun of feminine gender. The first block consists of four lexical meanings (1-4), the second, of one meaning (5) and the third, of two meanings (6-7). So, the sum of semantic states $S(X)$ can be formalized in the following way:

$$S(X) = \sum_{i,j} g_i(X) l_j(X), \quad (3)$$

Thus, grammatical states and respective lexical states of the language unit X = *cómico* are as follows:

1) g_1(*cómico*) = "adj.":l_1(*cómico*) = "Que divierte y hace reír. *Situación cómica*", l_2(*cómico*) = "Perteneciente o relativo a la comedia", l_3(*cómico*) = "Dicho de un actor: Que representa papeles cómicos. U. t. c. s.", l_4(*cómico*) ="Dicho de un autor antiguo: Que escribía comedias. U. t. c. s."[14];

2) g_2(*cómico*) = "m. y f.": l_5(*cómico*) ="**comediante** (|| actor)"[15];

3) g_3(*cómico*) = "f.": l_6(*cómico*) = "*Pan.* **historieta** (|| serie de dibujos). U. m. en pl.", l_7(*cómico*) = "*Pan.* **dibujos animados**"[16].

Taking into account the relation between grammatical and lexical semantics, the formula (3) will get another element $I(i; j; x)$ displaying this relation:

$$S(X) = \sum_{i,j} g_i(X)(i; j; X) l_j(X), \quad (4)$$

where i is grammatical meaning index of the headword X having semantic state $S_i(X)$; j is lexical meaning index corresponding to index i;$I(i; j; X)$ is the function providing relation between grammatical and lexical components of semantic state.

1.2 Language Unit Ambiguity and Semantic State Superpositions

The formula to display the whole set of a unit semantic states, is as follows:

$$S(X) = \alpha_1 s_1(X) + \alpha_2 s_2(X) + \cdots + \alpha_n s_n(X), \quad (5)$$

where$X$ is a language unit; $s_1(X)$, $s_2(X)$,…, $s_n(X)$ being partial semantic states the structure of which comprises grammatical and lexical components of the unit X;α_1, $\alpha_2$and$\alpha_n$being weighting factors the values of which can get different values depending on the context where the unit X is used. In other words, the recipient (lexicographer, reader or computer program) during context processing by his "intelligence-communication apparatus"assigns respective values to these factors based on his subjective ideas aboutsemantic functioning of the unit X in the given context.The calculation of weighing factors is a subject of another research. But the important condition to be fulfilled is that the sum of their values $\alpha_1 + \ldots + \alpha_n$ should be equal to 1. The semantic state that has got maximum coefficient will be looked

[14] l_1(*cómico*) = "Entertaining and causing laughter. *Comic situation*", l_2(*cómico*) = "Relating to a comedy", l_3(*cómico*) = "An actor playing comic parts. A. u. as a n. [also used as a noun]",l_4(*cómico*) = "An ancient author who wrote comedies. A. u. as an."

[15] l_5(*cómico*) = "Comedian (|| actor)".

[16] l_6(*cómico*) = "*Pan.* Comics (|| comic strips).M. f. u.in pl." [more frequently used in plural].

upon as the most relevant. In this way the equation (5) applied to a certain context is supposed to be reduced to one element. Let us give some text fragments where Spanish word *banco* acquires different semantic states:

1. "Como no tenía nada que hacer, después de desayunar un jugo de naranja en una cafetería me dediqué a leer el periódico sentado en un *banco*, …"[17];
2. "Los *bancos* del mundo deciden bloquear cualquier transacción financiera proveniente de Haití"[18].

According to DLE 23, the word analyzed has a set of 10 semantic states which in the given contexts undergoes reduction to one element: in (1), to s_1(*banco*) = "Asiento, con respaldo osinél,enquepuedensentarsedosomáspersonas"[19] (a bench); in (2), to s_5(*banco*) = "Empresa dedicada a realizar operaciones financieras con el dinero procedente de sus accionistas y de los depósitos de sus clientes"[20] (a bank). The process of reduction is known in linguistics as word-sense disambiguation. The semantic states given above are considered to be "pure" since they don't contain grammatical and lexical componentsof other semantic states of the word *banco*. Consequently the recipient can identify them easily in these contexts.

However a language unit doesn't alwayshave "pure" semantic states as it was shown in the examples above. The cases of unit functioning in different sematic states at same time and in the same text (context) are attributedto the superposition of semantic states. In this situation we have an ambiguous word and the context analysis was unsuccessfulin identifyingthe sense it is used in. For example in the sentence "Su desgracia fue quebrarle la mano"[21] the word *mano* can be interpreted as "as a part of human body" (a hand) or "a pointer on a clock" (a hand), or "an act or right of playing first" (lead). Besides that, the pronoun *su* can denote "his", "her", "its" or "your".

Thus, the whole set of semantic states are reduced not to one but to the sum of two or more elements. Let us take a closer look at the phenomenon of semantic state superposition on the examples from Spanish literature and build up the formal model of semantic state superposition.

2 Semantic State Superpositions Occurrence in Speech

Ambiguity in texts can arise either naturally due to internal peculiarities of semantic nature of the word (1), or be made artificially by language speakers to express irony, achieve a comic effect etc. (2), simulate misunderstanding (3), veil the meaning of a word (4) or combine direct and figurative meanings in the same word (5):

1. "En tanto que don Quijote *pasaba* el libro, *pasaba* Sancho la maleta, sin dejar rincón en toda ella, ni en el cojín que no buscase, escudriñase e inquiriese, ni costura que no deshiciese, ni vedija de lana que no escarmenase, porque no se

[17] After having orange juice for breakfast at acafeteria, with nothing elseto do I set on a *bench* and devoted myself to reading a newspaper.

[18] The world *banks* decided to block any financial transactions from Haiti.

[19] A seat, with or without back, that can sit two or more persons on.

[20] An establishment engaged in financial operations with money incoming from its shareholders or deposited by its clients.

[21] He / She / You had amisfortune to break his / her / its hand / pointer / lead.

quedase nada por diligencia ni mal recado; tal golosina había despertado en él los hallados escudos, que *pasaban* de ciento"(I, 23, p. 284)[22];

2. "Salió de la carcel con tanta honra, que le acompañaron doscientos *cardenales*; salvo que a ningunallamaban eminencia"(Fco. de Quevedo, La vida del buscón llamado Pablos)[23];
3. "P: ¿Y si finalmente te quedarás para vestir *santos*? / R: Yo pués *los* vestiría con un Lacroix"(Tamara Falcó, Entrevista concedida a lecturas)[24];
4. "*Cruzados* hacen *cruzados* / *escudos* pintan *escudos*, / y tahúresmuy desnudos / con dados ganan condados"(Luis de Góngora, Dineros son calidad)[25];
5. "El carnaval logra *enmascararlo* todo. Salvo la belleza femenina" (Jnj "melibeo", en Flickr)[26].

Let us analyze the context (1). By using the verb *pasar* author meant that (a) Sancho *studied thoroughly* the content of the bag, (b) Don Quijote*looked through* the bookand (c) Sancho found *more than* hundred coins. So the superposition includes three semantic states (lower indexes of the states correspond to definition number in DLE 23): s_{22}(*pasar*) = "Leer o estudiar sin reflexión" (to look through smth.), s_{21}(*pasar*) = "Recorrer, leyendo o estudiando…" (to study thoroughly) and s_8(*pasar*) = "Exceder, aventajar, superar" (exceed, to be more than).

The context (3) contains a short fragment of the interview with a woman working as a fashion designer. In journalist's question (P) the word *santos* acquires its sematic state as a component of the Spanish collocation *vestir santos*and means"to be left on the shelf (of a woman)". He actually wants to find out what she will do when she passes the age in which she may have an opportunity to marry. But she (R) understood this phrase in its literal meaning: "which clothing styleshe would select for saints". She mighthave interpreted word *santos* in direct meaning and that's why her answer was "I would dress them [the saints] in Lacroix style".

As for the context (4), the word *cruzado* has been applied in two semantic states at the same time: s_3(*cruzado*) = "Dicho de un caballero: Que trae la cruz de una orden militar. U. t. c. s." (crusader) and s_7(*cruzado*) = "Monedade Castilla, de plata o de vellón, mandada acuñar por Enrique II, yque tenía una cruz en el anverso, en el caso de la de plata" (coin, money). The same situation happens to the word *escudo*: s_2(*escudo*) = "Superficie o espacio generalmente en forma de escudo, en que serepresentan los blasones de un Estado, población, familia, corporación, etc."

---

[22] While Don Quixote examined the book, Sancho examined the valise, not leaving a corner in the whole of it or in the pad that he did not search, peer into, and explore, or seam that he did not rip, or tuft of wool that he did not pick to pieces, lest anything should escape for want of care and pains; so keen was the covetousness excited in him by the discovery of the crowns, which amounted to near a hundred [http://pd.sparknotes.com/lit/donquixote/section27.html].

[23] He was going out of prison with a great honor in the company of two hundred cardinals, though none of them was addressed as Eminence.

[24] Q: And what if you finally remain *to dress the saints* [left on the shelf, never get married] /A: Well, I would dress themin Lacroix style.

[25] *Money* makes *money* [*knights* make *knights*] / *gold pieces* paint *escutcheons* [*escutcheons* paint *escutcheons*] / and gamblers nude / with dice they win counties.

[26] The carnival can *mask* [*disguise*] everything, except for feminine beauty.

(escutcheon) and s_9(*escudo*) = "Unidad monetaria antigua de distintos países y épocas" (ancient monetary unit).

The phrase (5) evidences the superposition consisting of semantic states, one representing direct figurative meanings, respectively: s_1 = "Cubrir el rostro con máscara. U. t. c. prnl." (direct: to cover the face in a mask) and s_2 = "Encubrir o disimular algo. U. t. c. prnl." (figurative: conceal smth. from view). Consequently, the formal model of semantic state superposition in the contexts (1), (3), (4) and (5) is as follows:

$$S(X) = \sum_p s_p(X) , \quad (6)$$

where p is the index of a semantic state composing the superposition. The ambiguity shown in the context (2) is based on the homonymity of the word *cardenal*: $s_1^{[1]}$ = "Cada uno de los prelados que componen el colegio consultivo del papa y forman el cónclave para su elección" (a cardinal) and $s_1^{[2]}$ = "Mancha amoratada, negruzca o amarillenta de la piel a consecuencia de un golpe u otra causa" (a bruise). So the superposition of semantic states for the word analyzed will have the following model:

$$S(X) = \sum_{p,[k]} \alpha_p{}^{[k]} S_p{}^{[k]}(X), \quad (7)$$

where k is the index of a homonym used in meanings 1, 2… N. The formula above is also applicable to homography cases when different parts of speech coincide with each other by their grammatical form. For example, the popular shampoo in Argentina had slogan "para la caspa".When it was promoted by TV the viewers couldn't catch whether the word *para* was referred to the verb *parar* (3rd person singular in the indicative moodof present tense) or to the preposition *para* (for). The slogan in question could be interpreted ambiguously: the shampoo stops dandruff or the shampoo is intended for dandruff.

For the Spanish language is also natural to have grammatical ambiguity, in particular in nouns. With the same grammatical form the can function as an adjective in a sentence. This can also lead to ambiguous understanding of a word. For example: "Soy *cómico*, sí lo confieso, pero de corazón no malo y aún sincero cuando me lo propongo" (Chabaud Jaime, Divino pastor Góngora)[27].

The above sentence shows grammatical ambiguity of the word *cómico*. It can be either a noun or an adjective. It's worth noting that the main sign of the noun in the sentence is a definite (*el, la*) or indefinite (*un, una*) article. But it isn't observed in the context. That's why the phrase "Soy *cómico*" has two interpretations: "I'm a comedian" or "I'm funny".The superposition of semantic states of lexical and grammatical homonyms is formed in the way it is shown by the equation (7).

3 Lexicographic Treatment ofSuperpositions in VLL DLE 23

Within the scope of works on creating VLL DLE 23, we develop the interface intended for treating "pure" semantic states and their superpositions. If compare

[27] I'm *comic* [*a comedian*], I confess it from the bottom of my heart and I'm even sincere when I set my mind to it.

ordinary electronic dictionaries, including online dictionaries, VLL proposes software interface to perform:

1. *Access administration function*: user authorization and identification; addition and deletion of new users; managing access modes (only reading, reading and edition of dictionary material);
2. *Lexicographic works*: editing dictionary entries; creating dictionary on the basis of DLE 23; entry representation in any mode;
3. *Research works*: researching language levels covered by DLE 23 (grammar including word formation; vocabulary including semantics and pragmatics); researching the interaction of the language levels: grammar and semantics, word formation and semantics, semantics and pragmatics etc.

In the context of our research topic we propose a special module "Superpositions" to be provided for VLL DLE 23 (fig. 2) and had the following interface elements:

1. The table enlisting all semantic states of a headword superpositions revealed in authentic texts (belles-lettres, journalism, advertising etc.), with the following columns:
   (a) "Canonic form" where the initial form of the word having ambiguity is indicated;
   (b) "Superposition" which contains the index of word superposition, for example: "1+5", "3+2" etc.; the figures indicating the numbers of semantic states represented in DLE. In case of homonyms their numbers to be given in square brackents;
2. Text field "Context" to enter a text fragment where a word shows its ambiguity and, respectively, forms semantic state superposition;
3. The table containing the values of parameters constituting the equation (7):
   (a) "State": semantic state index (corresponding to definition number in DLE) in the superposition;
   (b) "Weight":the factor indicating the relevance of semantic state in superposition;
   (c) "Homonym": if semantic state belongs to a homonym the index of the homonym (lexical or grammatical) should be indicated;
   (d) "Part of speech" containing the relation of the word to the part of speech or grammatical category. The designation of parts of speech will be taken from DLE 23;
   (e) "Definition" to insert a definition text from DLE 23. This parameter corresponds to lexical component of semantic state;
4. Text field "Comment" to enter some notes containing the users' interpretations of word ambiguity (for example: language game or polysemy game etc.).

The window of the program module "Superpositions" provided for VLL DLE 23 shows the example of treating ambiguous word *cardenal* and representing its semantic state superposition is shown on figure 2.All the tables and text fields are filled by a specialist (lexicographer and / or linguist) after thorough examination of the contexts extracted by him from different literature sources. On finding a context the user determines word ambiguity by looking through all the definitions of the word (semantic states) enlisted in VLL DLE 23. By the results of the context analysis the user defines the superposition of the word and fills it in the upper table in a form of the index combination. Further he fills details related to semantic states composing

the superposition. The weight factors are determined based on the user's own considerations or by using separate software.

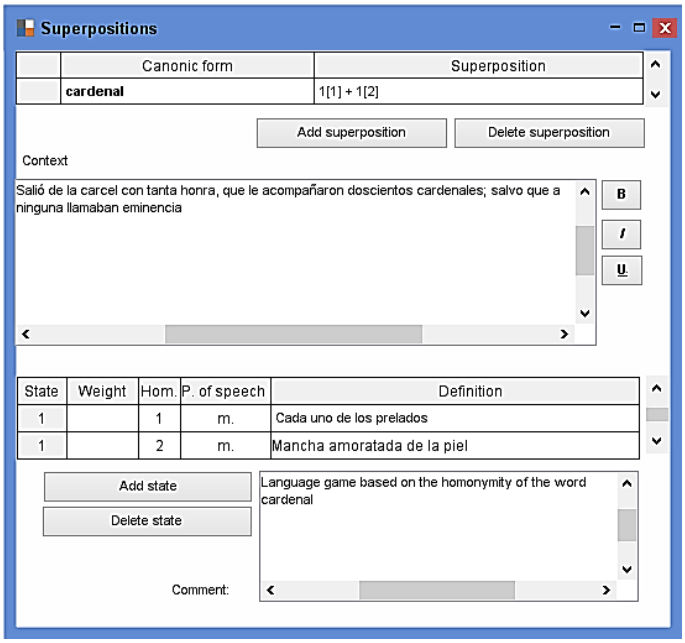

Fig. 2. Diagram of VLL DLE 23 interface for treating semantic state superpositions

4 Conclusions

1. Ambiguity is a property of a language unit to function in several semantic states at the same time in a context. Ambiguity can arise naturally (i.e. caused by the nature of a language) or can be caused deliberately by language speakers to create a particular effect.
2. The examples given above testify that any ambiguous Spanish unit can form semantic state superposition either on one or several language levels.
3. Using the theory of semantic states we built a formal model of superposition which represents the ambiguity of Spanish language unit. The parameters of the model (7) are going to be used for semantic state indexing, searching and displaying in respective form in VLL DLE 23.
4. In our opinion, the program module to be included in VLL DLE 23will facilitate conducting an ample range of linguistic researches, among them are logical-linguisticstudy of texts, analysis of speech acts including language games, semantic analysis etc. The module will be also planned to be used in linguo-didactic applications.

References

1. Cruise, A.: Meaning in language: An introduction to semantics and pragmatics. Oxford University Press (2000)
2. Lyons, J.: Introduction to theoretical linguistics. Cambridge University Press (2001)
3. Weinreich, J.: Languages in contact: findings and Weinreich. London; The Hague : Mouton and co (1964)
4. Apressian Yu.: Selected works, vol. 1:Lexical semantics, Moscow (1995)
5. Mihalcea, R.:Knowledge-Based Methods for WSD. In: Word Sense Disambiguation: Algorithms and Applications. 33, pp. 107--131 (2006)
6. Schvaneveldt, Roger W., Meyer David R.: Lexical ambiguity, semantic contextand visual word recognition. Journal of Experimental Psychology: Human Perception and Performance. 2, pp. 243--256 (1976)
7. Onifer, W., and SwinneyD.: Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. In: Memory and Cognition. 9, pp. 225-236 (1981)
8. Peña, H.: La ambigüedad, http://www.humanidades.uach.cl/documentos_linguisticos/ docannexe.php? id=437(1982)
9. Miranda, A. Aquellas sonadas soñadas invenciones. Algunos juegos fónicos como recurso de oralización en el *Quijote*, http://ru.ffyl.unam.mx/bitstream/handle/10391/3067/ 10_Hor_Cult_Quijote_2010_Miranda_101-110.pdf?sequence=1(2010)
10. Budor, K.: Aproximación lingüístistica a los juegos de palabras,hrcak.srce.hr/file/179799
11. Hogaboam, Th.: Lexical ambiguity and sentence comprehension. Journal of Verbal Learning and Verbal Behavior. 14, pp. 265--274 (1975)
12. Jastrzembski, J. E.: Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. In: Cognitive Psychology. 13, pp. 278--305 (1981)
13. Onifer W.: Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. In:Memory & Cognition. 9, pp. 225--236 (1981)
14. Uspenskii V. A. Concerning the definition of the case by A. N. Kolmogorov. К определению падежа по А. Н. Колмогорову. In: Bul. of the Association for machine translation problems. 5, pp. 11--18 (1957)
15. Ostapova I. V.: Etymological dictionary: lexicographic structure and representation in digital environment. Лексикографическая структура этимологического словаря и его представление в цифровой среде. In: Proceedings of the Annual International Conference "Dialogue". 8(15), pp. 359--364 (2009)
16. Shyrokov V. A.: Elements of lexicography. Елементи лексикографії. Kyiv (2005)
17. Shyrokov V. A.: Linguistic and technological fundamentals of explanatory lexicography. Лінгвістичні та технологічні основи тлумачної лексикографії. Kyiv (2010)

An Index of Authors' Popularity for Internet Encyclopedia

Dmitry Lande, Valentyna Andrushchenko, Iryna Balagura

Institute for information Recording of NAS of Ukraine, Kyiv

dwlande@gmail.com

Abstract. The new index of the author's popularity estimation is represented in the paper. The index is calculated on the basis of Wikipedia encyclopedia analysis (Wiki-Index–WI). Unlike the conventional existed citation indices, the suggested mark allows to evaluate not only the popularity of the author, as it can be done by means of calculating the general citation number or by the Hirsch index, which is often used to measure the author's research rate. The index gives an opportunity to estimate the author's popularity, his/her influence within the sought-after area "knowledge area" in the Internet – in the Wikipedia. There are proposed algorithms and the technique of the Wiki-Index calculation through the network encyclopedia sounding, the exemplified calculations of the index for the prominent researchers, and also the methods of the information networks formation – models of the subject domains by the automatic monitoring and networks information reference resources analysis.

Keywords: Wikipedia, Author's popularity estimation, Wiki- Index, Information networks, Subject domains

Introduction

Today scientometric mostly uses several indices, according to which the scientists' rate and their impact on science and society are calculated. Thus, the simplest index is the number author's publications. It is clear that this index does not depict the qualitative parameters that are better reflected in another index – the number of citations. In 2005 the physician Jorge E. Hirsch from the California University established the most popular index – Hirsch Index [1]. The principle of its calculation is quite simple, while it combines the advantages of the first and second approaches. The index calculation is based on the distribution of citations of the researcher work. According to Hirsch scientist has index *h*, if *h* of his *Np* papers cited at least *h* times each, while both articles remaining (*Np* – *h*) quoted no more than h times each. This index gained the support and is used in such scientometric systems as Scopus, Web of Science, and Google Scholar Citations.

At the same time this indicator, which is focused on the scientific importance, significance of the author, not quite fully reflects the overall importance of the results

that he/she received. For such an assessment it is appropriate to use non-fiction and open access systems. As one of the approaches to solve this problem, the authors proposed methodology for calculating the new index - the Wiki-Index of authors' popularity [2].

This index can appear unimportant tool in combination and with other indices can provide a complete picture of influential scientific achievements of the author, not only in the research community, but the overall impact on the formation of perspective and fully understanding of research information by the users.

Today Wikipedia (https://www.wikipedia.org/, https://en.wikipedia.org/ – English version, https://zh.wikipedia.org/ – Chinese version etc.) is the most visited site in the Internet, and one of the most popular encyclopedic resources covering all the disciplines, it provides answers to the most search engines queries. At this time only the English version of Wikipedia contains more than 5 million articles (German - more than 2 million, Chinese, Russian – more than 1 million).

It is known that Wikipedia does not publish original research results, but at the same time all the information and references are verified according to the Wikipedia citing policy [3,4]. All Wikipedia articles are open to be edited, so it can improve information, but no new information which need to be approved will not be published freely.

A sufficient amount of works and publications are dedicated to the research of subject areas as well as to the Wikipedia service that prove the relevance of the conducted studies [5]. The methods of building networks of co-authors, the definition of significant nodes of the network structure, research citations and appropriate buildings are among them [6]. Also authors have studied the array of publications relating to the approaches to the assessment of citations and other aspects of the update, existence, filling, editing of the encyclopedic resource Wikipedia [7-9].

Based on the results of the processed data, we can assume the uniqueness of the proposed indices and value of the information that will be obtained by the computations to evaluate the level of certain data in the system of science popularization and accessibility of provided research information on specific issues.

The use of indices is appropriate in different directions of evaluation and analysis of scientific activity, can also act as an additional tool for decision making, forming educational programs etc.

The rule of Wiki-Index computation

The authors suggested the following rules for calculating Wiki-Index of author's popularity. It is supposed that the references on the author are found in *N* Wikipedia articles.

Sorted by decreasing number of parameters that determine how many times author's name happens in bibliographic references of these articles we will denote as:

Wiki-Index of author's popularity (*WI*) corresponds to the maximum number of articles (*WH*) of Wikipedia, in which the number of references no more than the *WH* value, which is multiplied by a certain integral function, which is not decreasing (e.g., the square root is considered below) the *N*, that is:

Wiki-index of author popularity is ideologically close to the Hirsch index;

however, it doesn't take into account the number of articles that refer to the author's article and citations to the work of the author and the number of articles from Wikipedia, which contain these data links. Another difference from the Hirsch index is the multiplication by a function of *N*, reflecting the consideration it provides greater popularity and the more spread of index values for different authors.

It should be noted that the author popularity level must be attached to his subject domain on one hand in order to avoid false counting for homonyms, and on the other – to ensure completeness on subject area.

Example:

Let assume that the Wikipedia article with the highest number of references to author George Smith (in a given subject area) contains 100 references. The second – 20 documents, a third - 10, fourth – 5, fifth – 5, 4 more – only one link. So we have a number of values:

=100, =20, =10, =5, =5, =1, =1, =1, =1
1 article contains the number of references least=100;
2 articles contain the number of references least=20;
3 articles contain the number of references least=10;
4 articles contain the number of references least=4.
5 articles contain the number of references least=5.
There are no 6 articles that contain the number of references least 6.
In this case:

As follows,

Algorithm

In the process of the Wiki-Index calculating there should be provided the procedure of Wikipedia resources scanning, corresponding to the subject area in which the author works. Accordingly, as "adverse product" of the Wiki-Index computations, a model of the subject domain is being built, the model – is the network – nodes are concepts that represent articles from Wikipedia, and edges – are the hyperlinks between articles.

The process of the subject domain model of the author forming is possible in two ways:

- The use of Wikipedia dump database (not really relevant, but the link is available) by which the full range of all possible concepts and relationships. The advantage of this approach - completeness of information, disadvantage - possible loss of accuracy due consideration of homonyms, going beyond the subject area, considerable calculation time;

- The use of the principle of network services sounding (small sample volume of important contents of large information networks for technological reasons cannot be subjected to a complete scan). The advantage of this approach – getting accurate information strictly within a several subject domain, solvation of the homonyms problem and a short calculation time.The main drawback – the possible slight completeness, which may be assessed by additional experiments.

Authors chose the second approach for the Wiki-Index computation while

building its corresponding domain model chose the second approach, which was implemented as a software as a service.

Formation of subject domain by sounding Wikipedia

To implement calculation of Wiki-Index authors considered the following algorithm to form subject domains according to Wikipedia, avoiding the effect of topic drift:

On the main national Wikipedia page in the search line the initial word is given, e.g. (for English version - «Albert Einstein», for Chinese one –阿尔伯特·爱因斯坦, etc.).

- The search window opens. It contains information about concept, according to the task on the Step 1. The initial word/word combination is a graph vertex, which will be formed as the result of scanning.
- All terms-concepts corresponding the hyperlinks on the chosen page, are added to the formed graph. All the words/words combinations are the nodes of the graph. The edges to them are formed from the initial node.
- The next transition is made by the first not involved hyperlink from the examining pages.
- In text on the page to which the transition has been made the search of shortened researcher's name (e.g., Einstein, 爱因斯坦) or tag (e.g., physics, relativity, 物理学, 相对性) is to be carried out.
- In case, if there is a shortened researcher's name or tag is found, the transition to the Step 4 is made and accordingly from the node – word/word combination of the current search the new nodes are built.
- If there is no word/word combination in the text – the given graph branch is considered to be built.
- The next transition presumes pass to the page, which had been scanned – the word is not added as a graph node, and the feedback to the created node is formed.
- All the operations under steps 4-9 repeat until the not involved hyperlinks, chosen from the page, are left. In another case the graph is considered to be built.

According to the suggested algorithm the data collection process in Wikipedia from the first node-notion is stopped when according to the algorithm transition to the new node is impossible (there are no more basic nodes for transition), so the "loop" is impossible.

Calculation of the Wiki-Index of author's popularity

To compute the Wiki-Index it is necessary to make some changes to the suggested above algorithms, that is on the page, transition to which had been made by the hyperlink (5th Step of the algorithm), the search of author mentions in Publications, References, Further Reading sections (or in sections «**参考文献**»,

«**外部链接**», «**参考資料**», «资料来源», «**参考**资料» for the Chinese Wikipedia) is provided.

Herewith, the number of these mentions, which correlates values, is counted. If , the article is not important, the concept is defined as the endnote and the transition to the Step 4 is provided. Of course, this rule narrows the scanning of Wikipedia pages list and results the completeness loss, though, as the real computations prove, has little effect on the overall results. Pages dedicated to the scientific concepts and those, which don't contain relevant publications, can be ignored – just skipped. Therewith, the time of Wikipedia target segment is significantly reduced.

As a result of the full network sounding, the sequenceis formed, which is used to calculate Wiki-Index, according to the rules above.

Experimental section

The represented algorithms were implemented as a software system, through which the subject domains models and Wiki-Index are formed. Here are some examples of calculating Wiki-Indices for three authors: Albert Einstein, Enrico Fermi, Benoit Mandelbrot.

In Fig. 1 shows the Gephi (http://gephi.org) visualization of domain model fragment that were obtained by sounding Wikipedia according to the above algorithm. The parameters of obtained networks (subject domain models); nodes-concepts of Wikipedia are following.

For a network that meets the model of authors' subject domain:

Albert Einstein:

- nodes– 718,
- edges – 22111,
- the largest nodes (Table 1):

Table 1. Description of the largest nodes for the Albert Einstein subject domain

Concept	The node degree
Quantum_nonlocality	188
Alain_Aspect	181
Hermann_Weyl	177
Paul_Dirac	174
Electromagnetic_radiation	174
Isaac_Newton	169
Galileo_Galilei	169
Wolfgang_Pauli	169
General_relativity	167
Antimatter	167

(128 articles with the references, *WH* = 12)

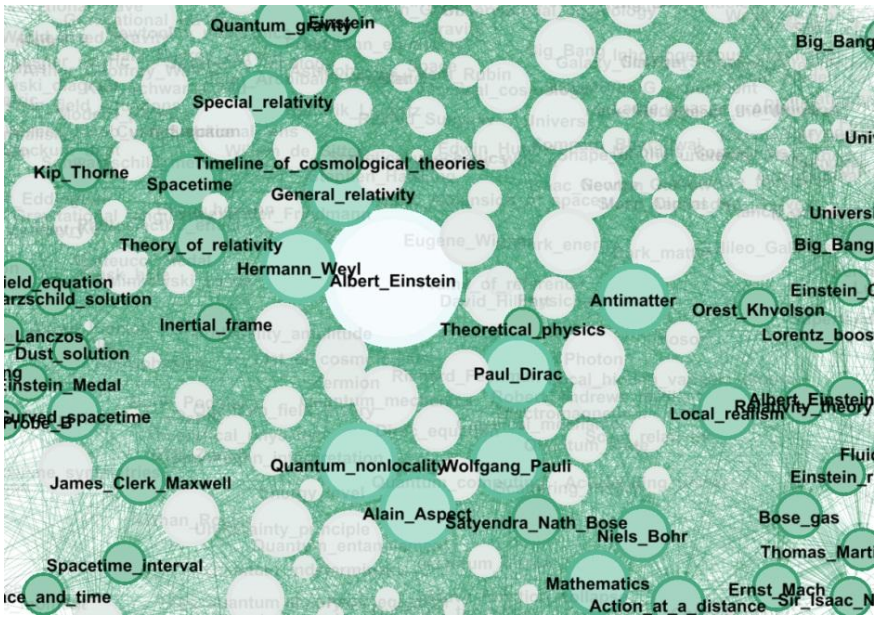

Fig. 1. Fragments of subject domains

Enrico Fermi:
- nodes – 605,
- edges – 22079,
- the largest nodes (Table 2):

Table 2. Description of the largest nodes for the Enrico Fermi subject domain

Concept	The node degree
Enrico_Fermi	440
Nobelium	206
Transuranic_element	206
Particle_physics	204
Mendelevium	204
Einsteinium	204
Berkelium	203
Radioactive_decay	195
Radioactive	190
Particle_accelerators	188

(92 articles with the references, WH=7)

Benoit Mandelbrot:
- nodes – 34,
- edges – 259,
- the largest nodes (Table 3):

Table 3. Description of the largest nodes for the Benoit Mandelbrot subject domain

Concept	The node degree
Benoit_Mandelbrot	22
Pattern	20
Chaos_theory	18
Patterns_in_nature	18
Hausdorff_dimension	17
Patterns	16
Fractal	15
Fractal_dimension	15
Fractal_geometry	15
Fractals	15

(11 articles with the references, WH = 6)

There is an example on the Figure 2 of the subject domain fragment, which responds “A. Einstein” for the Chinese Wikipedia.

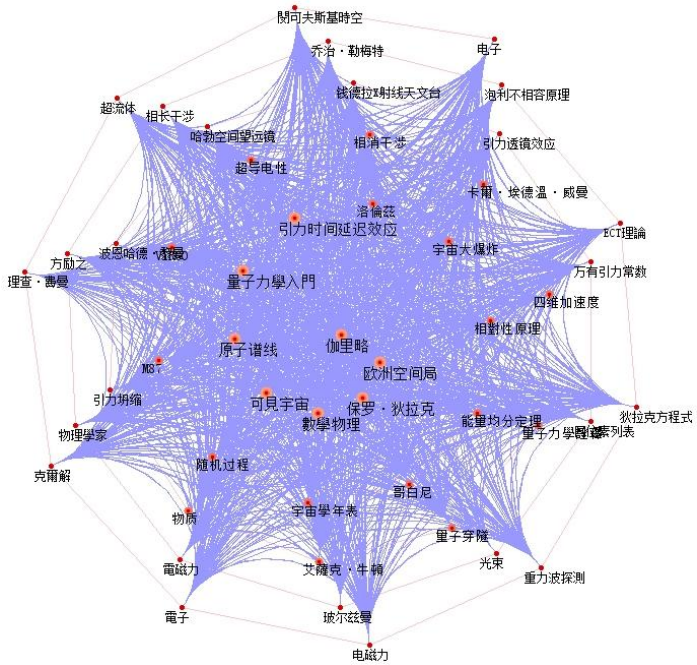

Fig 2. The fragment of subject domain for the Chinese Wikipedia

There were provided comparisons of the results – Wiki-Index, calculated on the research and the Hirsch-index, represented by the world’s leading scientometric resources Scopus, Web Of Science and Google Scholar Citations. Results are depicted in Table 4.

There are also were calculated the appropriate Wikipedia indices for the Chinese language Wikipedia that appeared an average 10-20% less.

Table 4. Comparison of Wiki-Indices values with the Hirsch-index (Scopus, Web Of Science and Google Scholar Citations)

N	Scientist	Wiki-Index	h-index Scopus	h-index Web Of Science	h-index Google Scholar Citations
1.	Albert Einstein	141	36	6	110
2.	Enrico Fermi	67	26	1	49*
3.	Benoit Mandelbrot	20	31	36	90

*Profile missing, the value was calculated for: "E. Fermi" according to the Google Scholar (Google Scholar Calculator) service.

By comparison, we can see and estimate the role of information on research and publications on open-access resources in comparison with data that consider purely scientific information with a certain restrictions set.

Conclusions

As a result of calculations and proposed approaches tests to the formation of popular author index due to the presence of references to his/her work and references in the largest encyclopedic resource – Wikipedia, following conclusions can be made:

1 The principle of Wiki-Index forming differs primarily from those, which currently is used in scientometrics with consideration of citation from not only scientific papers but popular service Wikipedia (separately for each language version). This way the index of author's popularity within this service can be obtained. This is an important issue, considering the fact that Wikipedia is currently the largest and most popular encyclopedic resource.

2 There is suggested the technique of the Wiki-Index quick calculation, which allows to realize computation as a separate service, and also automatically form the subject domain.

3 Due to the use and promotion of proposed indices there can be a significant expansion of open access resources (available to be edited by Internet users).

4 Provided work may be continued by analyzing other resources and the formation of indicators to estimate and analyze the influence in a particular environment.All the obtained results can be compared with those, which can be reached by analyzing other resources and the open access research on-line systems.

It is also necessary to note a fundamental difference between the proposed approach of automatic subject domains models formation and those that already exist, based on direct participation of experts in selecting specific nodes and links. In cases, as it depicted in the work, the researcher uses only a small share of knowledge represented by the name of the scientist, his writing abbreviated names of several key terms, concepts to construct an appropriate network. After that, the program uses the

knowledge that is implanted by different languages Wikipedia articles' authors, tags defined by internal hyperlinks. This way expert area is widely extended.

References

1. Hirsch, Jorge E., An index to quantify an individual's scientific research output // E-preprint ArXiv. arxiv.org/abs/physics/0508025
2. Lande D.V., Andrushchenko V.B., Balagura I.V. Wiki-index of authors' popularity // E-preprint ArXiv. arxiv.org/abs/1702.04614
3. M. Pei, K. Nakayama, T. Hara and S. Nishio, Constructing a Global Ontology by Concept Mapping Using Wikipedia Thesaurus,22nd International Conference on Advanced Information Networking and Applications - Workshops (workshops 2008), Okinawa, 2008, pp. 1205-1210.
4. https://en.wikipedia.org/wiki/Wikipedia:Verifiability
5. Zareen Saba Syed, Tim Finin, Anupam Joshi. Wikipedia as an Ontology for Describing Documents, Proc. 2nd Int. Conf. on Weblogs and Social Media, AAAI Press, March 2008., pp. 136-144.
6. Fei Wu and Daniel S. Weld. Automatically refining the wikipediainfobox ontology. In Proceedings of the 17th international conference on World Wide Web (WWW '08). ACM, New York, NY, USA, 2008, pp. 635-644.
7. Norlidah Alias, Dorothy DeWitt, SaedahSiraj, Sharifah Nor Atifah Syed Kamaruddin, MohdKhairulAzmanMdDaud, A Content Analysis of Wikis in Selected Journals from 2007 to 2012, Procedia - Social and Behavioral Sciences, Volume 103, 2013, pp. 28-36
8. Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, James R. Curran, Learning multilingual named entity recognition from Wikipedia, Artificial Intelligence, Volume 194, 2013, pp. 151-175
9. F. Abbas, M. K. Malik, M. U. Rashid and R. Zafar, WikiQA — A question answering system on Wikipedia using freebase, DBpedia and Infobox,2016Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, 2016, pp. 185-193.

Intelligent System Structure for Web Resources Processing and Analysis

Vasyl Lytvyn[1], Victoria Vysotska[2], Lyubomyr Chyrun[3], Andrzej Smolarz[4], Oleh Naum[5]

[1,2] Information Systems and Network Department, Lviv Polytechnic National University, Bandery str., 12, Lviv, Ukraine, 79013

vasyl.v.lytvyn@lpnu.ua[1], victoria.a.vysotska@lpnu.ua[2]

[3]Computer-Aided Design Department, Lviv Polytechnic National University, Bandery str., 12, Lviv, Ukraine, 79013

chyrunlv@mail.ru[3]

[4]Institute of Electronics and Information Technology, Lublin University of Technology, Nadbystrzycka str., 38A, Lublin, Poland, 20618

smolan64@gmail.com[4]

[5]Information Systems and Technologies Department, Drohobych Ivan Franko State Pedagogical University, I. Franko str., 24, Drohobych, Ukraine, 82100.

oleh.naum@gmail.com[5]

Abstract. The paper describes the general detailed and formal description of intelligent system of information resources processing (ISIRP) based ontology. The content life cycle phase implementation of ISIRP structure is improved. The general principles of ISIRP designing structures enable automated information resource processing to increase regular user text content realization, reducing the production cycle, saving time and increasing the e-commerce capabilities.

Keywords: content analysis, information resources, rating evaluation, content management system, ontology, knowledge base, machine learning, intelligent agent, stemming, parser.

1 Introduction and review of current research on the topic

Information resource in ISIRP – data set with a property set (Tab. 1), which is the action object of IT content transformation. The result of one IT usage may be

information resource of another [1]. Content in IT is formalized information and knowledge, placed in IS environment and, unlike data without detailed properties specification, formalization methods and regulation. Transformation of heterogeneous data in an organic centralized information resource is one of the most important problems of ISIRP construction and operation. The procedure of information resources formation and usage in ISIRP (Fig. 1) are determined by primary source data select method, data fixation, filtering, conversion to a specified format to create content and placement in the database [2].

Table 1. The main properties of information resources in ISIRP

Name	Property
Heterogeneity	Components of different origin, content and format of presentation.
Consistency	Absence of conflicting or converse content values.
Format accessibility	Accessibility for all users on the basis of standardized methods, tools and interfaces.
Openness	The ability to interact, exchange and share values with external resources.
Dynamism	Quick update under the terms of a system or environment.
Scalability	Ability to change the logical / physical volume of content (values / concepts and their designations).
Manageability	Changes identification /content usage and its implications on the IS processes.

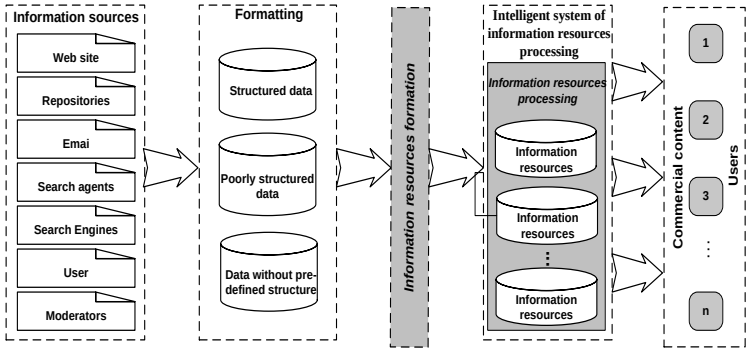

Fig. 1. The procedure for information resources formation and usage in ISIRP

Suppose there is some pre-defined set of content n_X primary sources $Sourse(x_i)$ with fixed or variable composition, where x_i is *i*-content from a source at $i=\overline{1,n_X}$, generates a certain set of values containing information / knowledge / facts from the domain ISIRP (Tab. 2).

Table 2. Objectives and tasks of research

Objective	Scientific innovation	Conclusion compliance
perform ISIRP analysis and evaluation based on compatibility option detalization of these systems to improve their classification;	improved ISIRP classification based on theoretical studies and compatibility option detalization of these systems by analyzing features of the system creation and usage.	studied and improved ISIRP classification based on the analysis and evaluation of such systems, allowing us to determine, detail and justify choice of their compatibility options.
develop a text content forming method through lifecycle improvement for management requirement determining of content flow;	for the first time developed the method of text content formation by its life cycle stages improvement through information resources detailed study.	for the first time developed content formation method by its lifecycle stages improvement for requirement determining of content flow control.
improve the text content control method on the basis of its system formation and analysis for determining text content control parameters;	further developed methods for text content managing on the basis of its system formation and analysis, ensuring control parameter regulation and requirements of content formation.	improved text content control method on the basis of its system formation and analysis for determining text content control parameters.
develop a method of content tracking based on statistical analysis of ISIRP to change the control parameters and the content forming requirements ;	for the first time developed a method of text content tracking based on statistical analysis of ISIRP allowing us to determine the content managing parameters.	developed a method of text content tracking based on statistical analysis of ISIRP to change the control parameters and content forming requirements which makes it possible to increase turnover volumes of content at 9%.
improve the structure of ISIRP by analyzing the information resources processing to develop typical system design recommendations;	improved the structure of ISIRP based on information resources processing specification and through the sharing of the formation processes, management and content tracking, ensuring the implementation its life cycle stages and development of typical system design recommendations;	improved the structure of ISIRP based on information resources processing different than the current with subsystems for formation, management and text content tracking; elaborated recommendations for ISIRP structure design
carry out results evaluation through the implementation of information resources processing technological software in ISIRP to reduce the time and cost of the textual content formation, management, and tracking .	elaborated recommendations for ISIRP structure design different than the current for stage specification and information resources processing subsystems existence that enable content lifecycle support; developed and implemented software for textual content formation, management, and tracking to increase the constant user text content volume at 9%.	developed and implemented software for textual content formation, management, and tracking to increase the potential user active attraction and the target audience expansion by 11%.

2 Information flows in intelligent system of information resources processing

As a conversion result of ISIRP certain technological means to the source $Sourse(x_i)$ is generating a range set $X = \{x_1, x_2, ..., x_{n_x}\}$ through Web-source information parser, perceived and presented figurate. In the process of selection and fixation of generated values according to the technological features of the system, every source of information generated range set is converted into input content set $C = \alpha(u_f, x_i, t_p)$ of defined format c_r, where $r = \overline{1, n_C}$. The main objective of the ISIRP development project is creating information resource information architecture by creating up to date text content that is formed on the reverse reaction of the users according to the type of content distribution (Fig. 2). Each content set is presented in a structured, semi structured data or data without description of the structure and stored in a text content database. Content structuring involves the formation for each set describing its composition, methods of combining elements and their regulation, i.e. condition set $U = \{u_1, u_2, ..., u_{n_U}\}$, where u_f is content formation condition at $f = \overline{1, n_U}$. Source data set is a combination of set values in a given format and condition set $\langle X, U \rangle$ when forming a content input set without structure description $U = \varnothing$.

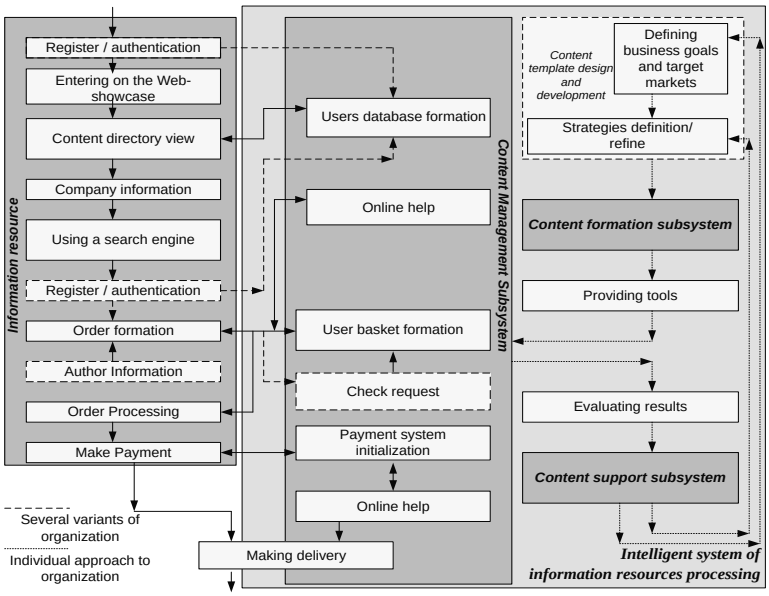

Fig. 2. Data flow diagram in ISIRP

The resulting content before saving is verified / validation for its formal / substantial accuracy / relevancy confirmation upon the system demand. In case of inconsistency to required criteria piece of content is removed from further use. Filtered content is formatted and stored, then the relevant information and knowledge

$\langle C,H\rangle$ become available to users through information resource ISIRP, i.e. $Sourse(x_i) \rightarrow x_i \in X \rightarrow X \rightarrow \langle X,U\rangle \rightarrow \alpha(u_f,x_i,t_p) \rightarrow c_r \rightarrow C \rightarrow DataBase(C) \rightarrow \beta(q_d,c_r,h_k,t_p) \rightarrow \langle C,H\rangle$, where $i=\overline{1,n_X}$, where n_X – content source number, $Sourse(x_i)$ - i-content source, $x_i \in X$ - i-source content, $Sourse(x_i)$; $X=\{x_1,x_2,...,x_{n_x}\}$ - data set as a result of the source selection, $Sourse(x_i)$; $\langle X_i,U_i\rangle$ – data set with condition set, $\alpha(u_f,x_i,t_p)$ – forming content operator, c_r – formed content, C – generated content set, $DataBase(C)$ – content preserving statement in the database, $\beta(q_d,c_r,h_k,t_p)$ – content control statement, $\langle C,H\rangle$ – ISIRP information resources composed of text content sets and content steering conditions [1].

The text content formation process submitted on this coupling scheme: $Sourse(x_i) \rightarrow x_i \in X \rightarrow X \rightarrow \langle X,U\rangle \rightarrow \alpha_1(Downloading(\langle X,U\rangle),T) \rightarrow \alpha_2(Verification(\langle X,U\rangle),T) \rightarrow \alpha_3(Conversion(\langle X,U\rangle),T) \rightarrow \alpha_4(\langle X,U\rangle,T) \rightarrow \alpha_5(Qualification(\langle X,U\rangle),T) \rightarrow \alpha_6(\langle X,U\rangle,T) \rightarrow \alpha_7(\langle X,U\rangle,T) \rightarrow c_r \in C$, where $X=\{x_1,x_2\ldots,x_{n_x}\}$ – income data set $x_i \in X$ from different information resources or moderators under $i=\overline{1,n_X}$; α_1 – content collecting statement from different sources, α_2 – content duplication identification statement, α_3 – content formation statement, α_4 – content key words and concepts identification statement [3], α_5 – content automatic categorization statement [4], α_6 – content digest forming statement, α_7 – content selective distribution statement, $T=\{t_1,t_2\ldots,t_{n_T}\}$ – transaction time $t_p \in T$ of text content formation under $p=\overline{1,n_T}$, $C=\{c_1,c_2\ldots,c_{n_C}\}$ – text content set $c_r \in C$ under $r=\overline{1,n_C}$, $Verification(\langle X,U\rangle)$ – content verification statement, $Qualification(\langle X,U\rangle)$ – content qualification statement, $Conversion(\langle X,U\rangle)$ – content transformation statement, $Downloading(\langle X,U\rangle)$ – content uploading statement. Improved ISIRP structure is through the addition of text content technological formation, management and support software (Fig. 3) [1].

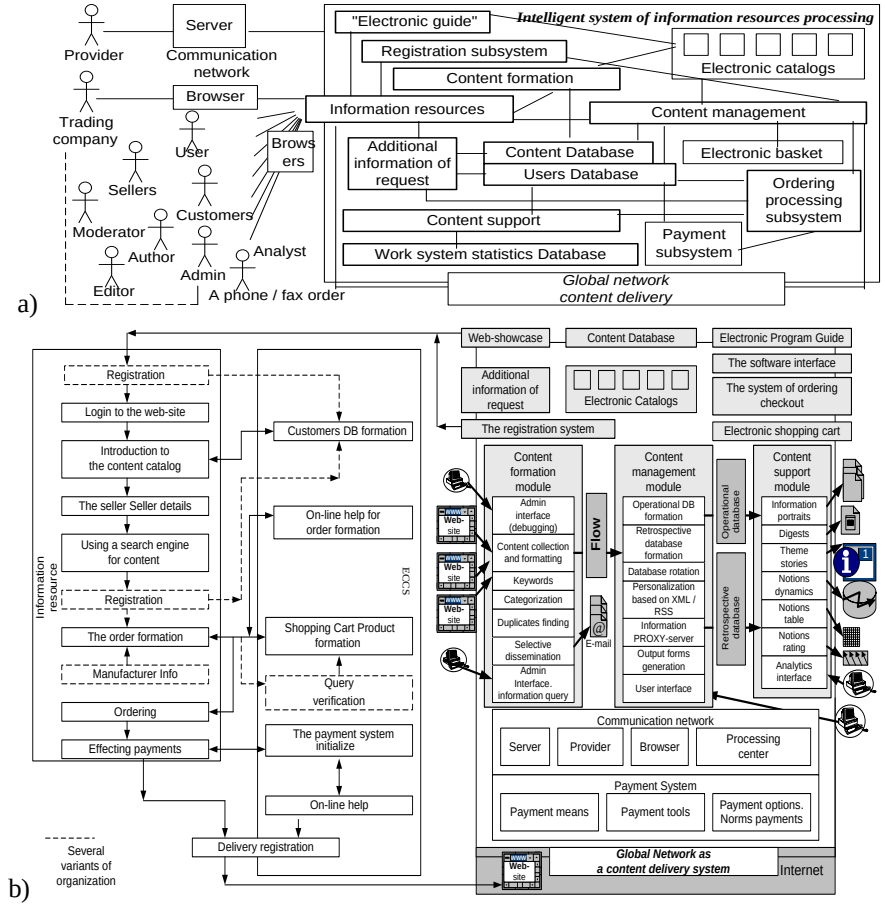

Fig. 3. Improved ISIRP structure

Text content formation - a set of measures to ensure data processing control from various sources (Fig. 4) to create a content with an additional set of values such as relevance, authenticity, uniqueness, completeness, accuracy and so on.

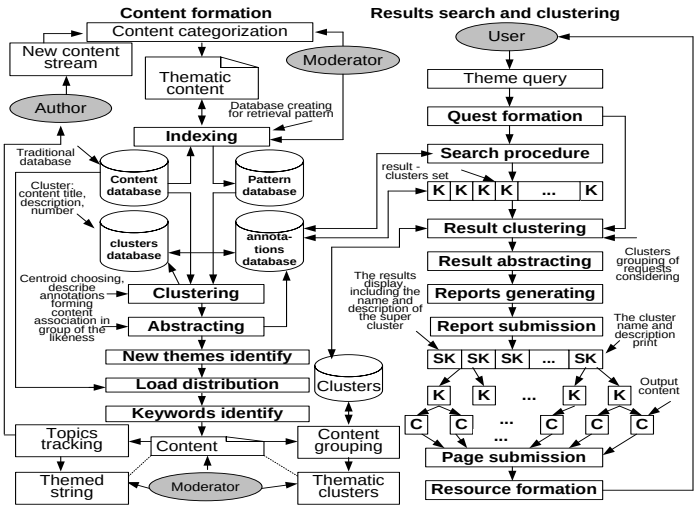

Fig. 4. Usage scheme of annotation database for text content search

Text content managing is a set of measures to provide text content defining parameter value support as topicality, completeness, relevance, authenticity, reliability to determined requirements by a criteria set (Fig. 5). Text content tracking is a set of measures to ensure the ISIRP functioning under certain requirements and any subsequent changes to these requirements (Fig. 6). ISIRP information resources processing allows to get current and objective data about the system operation and for competition level evaluation on the segment of the content financial market; estimate competitor level and measure their competitiveness across the financial landscape for content distribution. The main classes of information resource users/characters (clients, managers and administrators) define the information resource design and decision-making process. ISIRP necessarily includes Web-mart (information resource) with a text content catalog (searchable) and the necessary interface elements to enter registration data, the formation of orders, making payments over the Internet, delivery handling (e-mail / on-line), obtaining data about the company and on-line help. Registration/user authorization happens while making order or entering the system. The interaction is carried out over a secure channel SSL for protection purposes. The whole process is recorded in the content management subsystem for ISIRP functioning statistic formation and offers as a list of popular content topics for content forming subsystem.

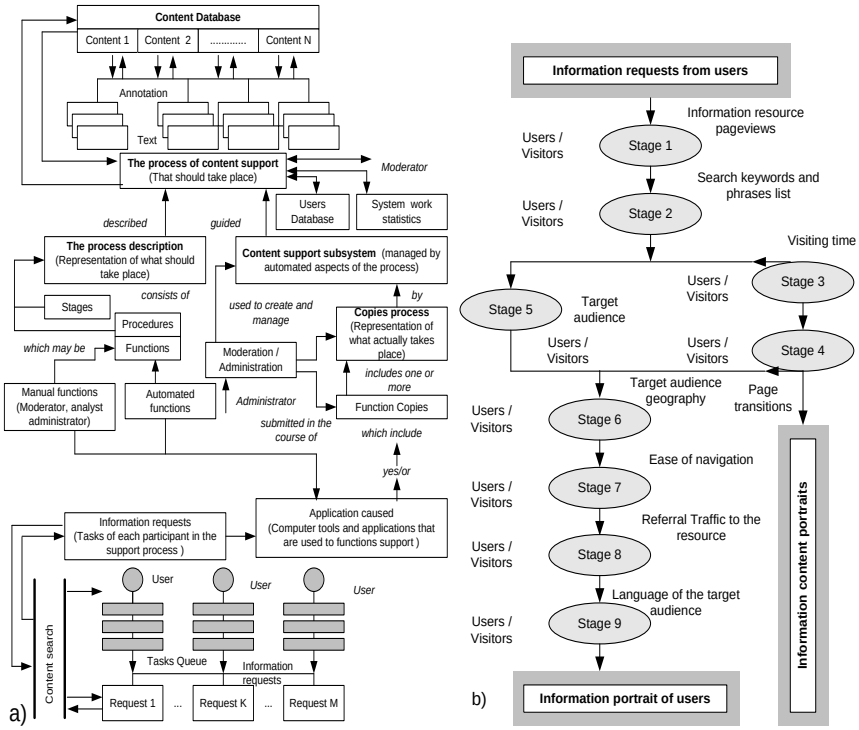

Fig. 5. The dependence scheme of a) component and b) stages of the text content tracking

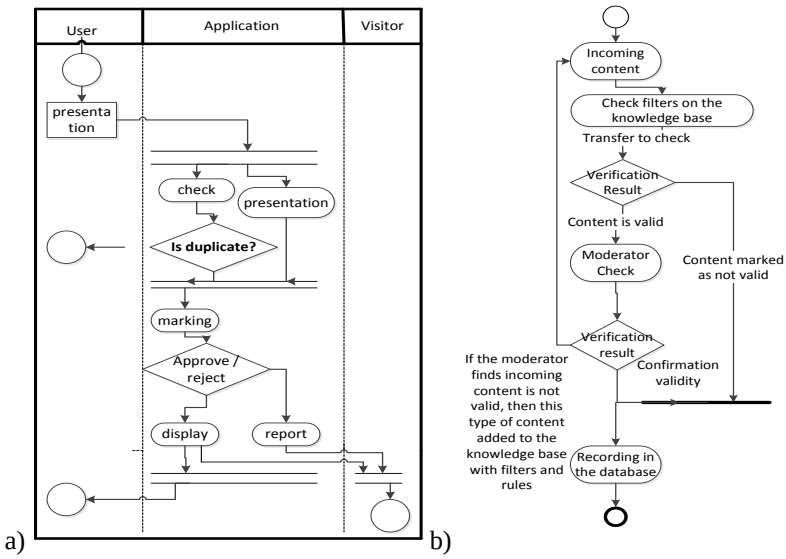

Fig. 6. The detailed scheme of a) tracking and b) a set of measures to ensure data processing control from various sources

3 Method of Web Resources Processing and Analysis

The process of content formation executes a transformation between a set of input data from different sources and a set of formatted and saved content elements: $S(x_i) \rightarrow x_i \rightarrow X \rightarrow \alpha(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow D(C)$, where $S(x_i)$ – is a source of data, $D(C)$ – content database. The formation of content $\alpha : X \rightarrow C$ is presented as a superposition of functions

$$\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_0, \text{ or } \alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \quad (1)$$

where α_0 – is an operator of content creation; α_1 – operator of gathering content from different sources through Web-source information parser; α_2 – operator of content deduplication; α_3 – operator of content formatting; α_4 – operator of keywords and concepts elucidation; α_5 – operator of automatic categorization; α_6 – operator of compilation of content digests; α_7 – operator of discretionary content sharing. The process of content formation is presented as

$$\alpha = \langle X, T, U, C, \alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7 \rangle. \quad (2)$$

1. The operator for content creation is a mapping of input data from various sources into content, which is actual and commercially worthy $\alpha_0 : (X, U_C, T) \rightarrow C_0$.

2. The operator of content gathering through Web-source information parser is a mapping between input data obtained from authors or moderators into content which is actual and authentic $\alpha_1 : (X, U_G, T) \rightarrow C_0$.

3. The operator of content deduplication is a mapping of initial content into content having no duplicate elements $\alpha_2 : (C_0, T, U_B) \rightarrow C_1$.

4. The operator of content formatting is a changing of content's format $\alpha_3 : (C_1, U_{FR}, T) \rightarrow C_2$.

5. The operator of keywords elucidation define an addition to content in the form of the set of keywords, which generally describe it $\alpha_4 : (C_2, U_K, T) \rightarrow C_3$.

6. The operator of content categorization – is a transformation of content via analysis and validation into a new state where content is assigned to some thematic category $\alpha_5 : (C_3, U_{CT}, T) \rightarrow C_4$.

7. The operator of compilation of digests based on content is a transformation of content to new state having a short content digest $\alpha_6 : (C_4, U_D, T) \rightarrow C_5$.

8. The operator of discretionary sharing of content adds to content a target audience definition and sharing to this audience $\alpha_7 : (C_5, U_{Ds}, T) \rightarrow C_6$.

The process of content formation is described by operator $c_{r+1}(t_{p+1}) = \alpha(c_r, t_p, X, u_f)$, where $u_f = \{u_{1f}, u_{2f}, \ldots, u_{n_U f}\}$ – is a set of conditions for content c_r formation:

$$c_r = \left\{ \bigcup_i^{n_X} x_i \middle| \begin{array}{l} \forall x_i \in X_{u_f}, x_i \notin X_{\overline{u_f}}, \exists u_f \in U_{x_i}, u_f \notin U_{\overline{x_i}}, \\ X = X_{u_f} \cup X_{\overline{u_f}}, U = U_{x_i} \cup U_{\overline{x_i}}, f = \overline{1, n_U} \end{array} \right\}.$$

The process is going through data transformation stages into a set of relevant, formatted, categorized and validated content elements: $x_i \in X \rightarrow \alpha_0(X, U_C, T)$ або $\alpha_1(X, U_G, T) \rightarrow \alpha_2(C_0, T, U_B) \rightarrow \alpha_3(C_1, U_{FR}, T) \rightarrow \alpha_4(C_2, U_K, T) \rightarrow \alpha_5(C_3, U_{Ct}, T) \rightarrow \alpha_6(C_4, U_D, T) \rightarrow \alpha_7(C_5, U_{Ds}, T) \rightarrow c_r \in C$.

Keywords discovery in content. Textual content (articles, commentaries, books, etc.) contains a lot of information in natural language, some of which is abstract. The text is presented as a sequence of character units, the basic properties of which are informational, structural and communicative connectivity / integrity that reflects the informational and structural nature of the text. The method of text processing is the linguistic analysis of content. This process splits the text on lexical tokens using finite automata (Fig. 7).

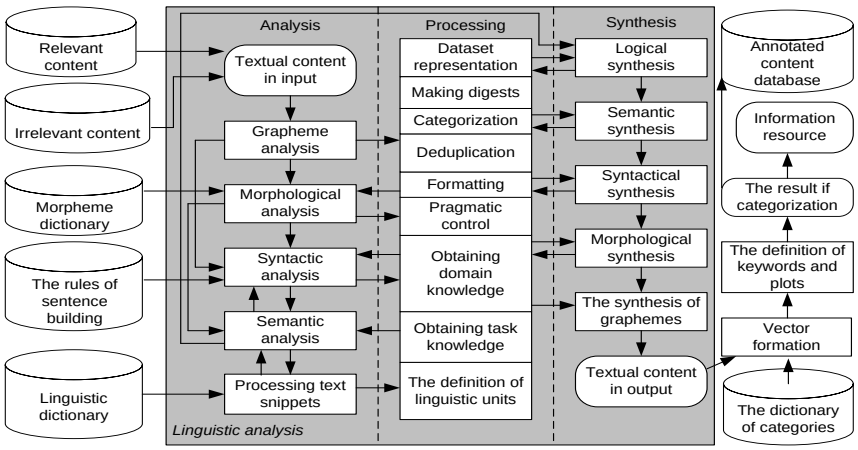

Fig. 7. The structure of linguistic analysis of textual content

The operator of keyword discovery in content is a mapping of content into a new state, which is different from the previous state by the presence of keywords describing this content. During the process of analysis are explored: a multi-layered structure of content; a linear sequence of characters; a linear sequence of morphological structures; a linear sequence of sentences. The discovery of keywords in text snippet is accomplished using such process. On the compositional level the sentences, paragraphs, sections, chapters and pages which are not related to the text snippet are isolated. Therefore, they are not considered. Next, using database of terms

/ morpheme units and text analysis rules, the search of the term is performed. Using the rules of generative grammar, the term is corrected according to the rules of its use in the context. After parsing, the text is drawn into a data structure, such as a tree, which corresponds to the syntactic structure of the input sequence, and is best suited for further processing. After analyzing the text snippet and term, a new term is synthesized as a keyword using a database of terms and their morphemes. Our approach to keyword identification is based on Zipf law and comes to the choice of words with an average frequency of occurrence (most used words, found in stop dictionaries, and rare words are ignored). Keyword detection module is implemented on the website Victana (available at address http://victana.lviv.ua/index.php/kliuchovi-slova).

The process of content categorization. The analysis of the lexical, grammatical and pragmatical structure of a text is used to automatically categorize content and build its digest. The operator of categorization of content is a mapping of the content to a new state via its validation. The new state is different from the previous one by availability of its assignment to some set of contents themes $\alpha_5 : (C_3, U_{CT}, T) \to C_4$. The analysis of content's meaning is performed by the process of pulling grammatical data from the word using grapheme analysis and the correction of morphological analysis results by analyzing the grammatical context of linguistic units. The process of categorization $C_4 = \alpha_5(\alpha_4(C_2, U_K), U_{CT})$ via procedure of automatic content indexing C_3 is divided sequentially into following blocks: morphological analysis, syntactical analysis, semantical and syntactical analysis of linguistic structures, the variation of textual content's written representation.

Creating a digest of content. The digest is a summary of a publication in ISIRP. In order to form it, the weighted content analysis is used and the frequency of words usage from a previously created dictionary of terms is considered. The operator of a digest formation is a mapping of content into a new state. This state is different from a previous one by the availability of additional part (digest) which adds to the content value. The process of a digest formation creates the set of the brief annotations and main points of the content for a specified time. This is convenient for a quick familiarization with the content's basics for a specified subject, and also when doing research on some topic.

Content distribution process. Our implementation of distribution process of content shares a workload between authors and moderators while contributing to increase of the reading audience and the volume of content. In the beginning, the system obtains digests of sources via RSS. Next the digests are distributed among authors according to author's rating. Digests are sent first to authors with higher rating. The rating reflects the performance of each author and is influenced by such criteria as uniqueness of content, the number of views (either direct or referenced), user ratings. The rating assessment system utilizes many criteria, so final result is rather objective and it stimulates authors to improve their performance. After this the system goes into the standby mode until some new content will be added. The workload for moderator is reduced, especially in such tasks as sorting, assessment, rating, evaluation and analysis of content. Therefore, the content is created faster, and has higher quality because of objective process of evaluation of the content.

As a result of analysis of ISIRP functioning S and content support C the set

$Y = \{Y_P, Y_T, Y_C, Y_R\}$ is created according to conditions $V = \{V_P, V_T, V_C, V_R\}$, where $Y_P = Y_{Pc} \vee Y_{Pq}$ - the subset of informational snapshots of content Y_{Pc} and users Y_{Pq}, Y_T - the subset of thematic plots, Y_C - the subset of content dependencies tables, Y_R - the subset of content ratings, $V_P = V_{Pc} \vee V_{Pq}$ - the subset of conditions for information snapshots, V_T - the set of conditions for thematic plots discovery, V_C - the set of conditions for creating content dependencies tables, V_R - the set of parameters used in content rating assessment. The set of informational snapshots of content Y_{Pc} is represented as $Y_{Pc} = BuInfPort(V_{Pc}, C, H, Q, T)$, and the set of user's snapshots Y_{Pq} as $Y_{Pq} = BuInfPort(V_{Pq}, Q, H, Z, T)$, where $V_P = V_{Pc} \vee V_{Pq}$ - the set of conditions for creating snapshots, *BuInfPort* - the operator of snapshot creation $Y_P = Y_{Pc} \vee Y_{Pq}$.

The set of thematic plots Y_T is presented as $Y_T = IdThemTop(C, H, Q, V_T, T)$, where V_T - is the set of conditions for plot discovery. *IdThemTop* - the operator of thematic plot discovery Y_T. The set of content dependencies tables Y_C is shown as $Y_C = ConCorrTablConc(C, V_c, T)$, where V_C - the set of conditions for creating content dependencies tables, *ConCorrTablConc* - the operator for creation of content dependencies tables. The set of ratings Y_{Rc} is represented as $Y_{Rc} = CalRankConc(C, Q, H, Y_C, V_{Rc}, Spam, Tonality, T)$, where $V_R = V_{Rc} \vee V_{Rm}$ - the set of parameters used in rating assessment, $Tonality(Q^+, Q^0, Q^-, T, H)$ - criteria of content tonality, $Spam(Q, T)$ - the operator for comments filtering, *CalRankConc* - the operator of rating definition for content and moderators $Y_R = Y_{Rc} \vee Y_{Rm}$. The set of output statistical data Y is presented by:

$$Y = \{Y_P, Y_T, Y_C, Y_R\} = Support(V, C, Q, H, Z, T, \Delta T)$$
$$Y = \{Y_P, Y_T, Y_C, Y_R\} = Support(V_P, V_T, V_C, V_R, C, Q, H, Z, T, \Delta T)$$

where $Y_P = Y_{Pc} \vee Y_{Pq}$ is the set of informational snapshots of content and its users, Y_T - the set of thematic plots, Y_C - the set of content dependencies tables, $Y_R = Y_{Rc} \vee Y_{Rm}$ - the set of ratings for content and moderators, *Support* - the operator for content support.

4 Basic idea and structure of web-source information parser

The use of standard software libraries avoids unjustified time, financial and human resources overrun for their re-development. That is why a wide range of active similar to developed one projects were analyzed, most of which are based on open source code concept and software distribution on terms of free licensing. Leading development team provide their projects by means of API (Application Programming Interface), thanks to which the functionality of these projects can be effectively used

by simple cataloged and well documented procedures and functions with corresponding arguments. Software development community worked out application software package use policies under different license provisions as well as participation in existing projects promotion so that each developer may receive, establish for personal use and develop as possible in one's direction these projects or included software environment libraries. Internet portals such as SourceForge.net contain all the necessary toolkit range o for deployment, documentation and project maintaining of arbitrary degree of complexity, development stage, access level and popularity among users. Developers make heavy use of development version-control special-purpose servers which provide collective (though thousands of participants) software development. The most popular among these is Git server. It can be installed separately as an individual or corporate server, and as well can be used by the global public Git-server GitHub. As studies have shown, among programming languages the vast majority of developments in the field of human language text document processing and almost all developments the field of ontology construction and education are written on Java. In addition, Java retains dominance among project languages, placed on the website SourceForge (Fig. 8) [5].

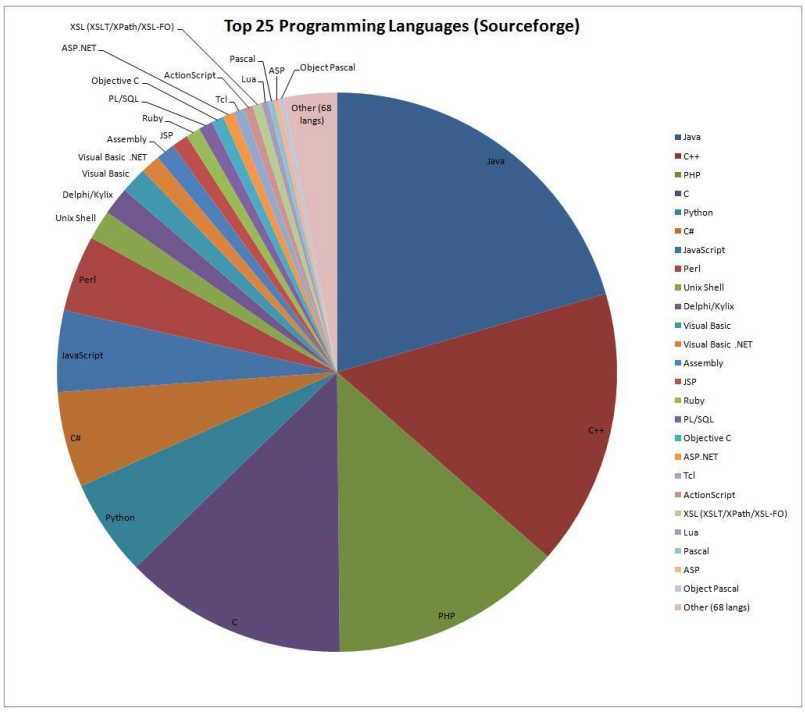

Fig. 8. The shares of 25 most popular programming language among developers that provide access to their portal projects through SourceForge.net.

The winning argument in favor of Java as a project programming language was the availability and accessibility of Java API in a projects at Stanford University (USA) Protege-OWL, as it was Stanford Research Center of Biomedical Informatics,

who has become the practical studies flagship in the field of development tools, knowledge base and ontology editing and teaching with knowledge representation language OWL [6-19]. Projects developed in Java as well:

- Gate [http://gate.ac.uk/] -set of text document processing tools to identify new knowledge.
- owlapi.sourceforge.net - another Java-project, which is a library of Java-classes with broad functionality of OWL- document processing.
- Pellet [http://clarkparsia.com/pellet/] - a software tool - the Java inference machine to implement arguments (knowledge creation) from knowledge base in OWL 2.0 language.

Since there was no unified system for online content search that would bear the value and novelty, the goal was to create one. In other words to implement a system which would use user key words and return only relevant content as result for further processing. The value of such a system lies in scholar labor saving during search of necessary material or document that clearly increases his productivity.

In the service of the aim the following technical challenges were placed [7]:

- Creation of unified information search system, which includes relevant content;
- The system must be implemented as a program (independent module), which should be written in a modern programming language;
- The program must include components for information web source communication (scientific web resources, libraries, archives, data storage, etc.);
- The program must include a component that gives web page content;
- The program must include a component which provides the article content reading and processing and related information from a remote Web resource;
- The program should be built on such a design pattern that allows to expand its capabilities without significant changes of the existing code base;
- The program must include technical documentation that gives an insight into operating principle and simplify further development;
- The program must meet the basic criteria of the Basics of Occupational Safety and general accepted development standards;
- Program should be cross-platform and contain minimum dependencies on third-party libraries of chosen programming language.

During the system creation process the syntactic analysis mechanism of Java integrated web page content was evaluated. There are two such mechanisms or rather two interfaces: DOM (Document Object Model) and SAX (Simple API for XML). When employing DOM document, wherein analysis is necessary, it is systematized in tree-type (the tree hierarchy).The elements of such hierarchy are conveniently accessed and processed. Also DOM element search possesses fast response, but the interface mechanism requires more memory than SAX.

SAX search for relevant information takes place with help of iterative method in other words enumerative technique of all elements in document from the first to the last. That is, per one iteration loop only one element is given for processing and it is possible to refer to it only when iteration reaches the last element of the document and start a new cycle. SAX mechanism has a much lower response speed than DOM, but uses less memory. So, for the system implementation the DOM mechanism was selected as it is much faster to refer to necessary elements in document if they are

classified in the DOM model. There are a lot of implementations and add-ons for DOM. Main of them are: Jericho HTML Parser; Java XML Parser; JTidy; HTML Parser. Within this set of analyzers Jericho HTML Parser was chosen, as its main advantages are:

- Not valid HTML document does not cause errors during the analysis;
- HTML document is being analyzed even if it contains server's tags;
- Option analysis is held with help of StreamedSource class, which allows memory to process large files efficiently. It provides additional functions that are not available in other on-stream analyzers;
- The row and column number of each position in the original document are easily available.

Jericho HTML Parser is a Java open source library, distributed on two licenses: Eclipse Public License (EPL) and the GNU Lesser General Public License (LGPL). So you can use it for commercial purposes in accordance with the license agreement conditions. Javadocs contain comprehensive information on the entire API.

Classes, methods and fields used to create a system:

- Source is class, the object of which contains HTML document.
- fullSequentialParse () is class method Source, which analyzes the output tags.
- Segment - class, the object of which parses the Source object content into segments.
- getAllElements (StartTagType) is class method Segment, which creates a DOM source document hierarchy according to specified conditions.
- getTextExtractor () is method, which returns the clean text and removes all tags in a given document.
- HTMLElementName is a class that contains static methods for choosing right tags.
- getAttributeValue (String attributeName) is method, which returns the decoded attribute value with the specified name.
- getParentElement () is method, which returns the parent towards given in the DOM hierarchy.
- getName () is method, which returns the name of the element.
- getChildElements () is returns a list of the direct descendants of this element in the document hierarchy.

The system consists of three main modules (interfaces) and auxiliary interfaces that provide iteration procedure (Fig. 9a). These modules are: iConnectionProvider, iWebInfoParser, iWebInfoSource. Auxiliary interfaces that provide iteration: Iterable <Publication>, Iterable <String>. IConnectionProvider module is an interface, which is implemented through StraightConnection or ProxyConnection classes depending on chosen connection option (direct inclusion or inclusion with server authentication).

StraightConnection class includes two major public method that implement the iConnectionProvider interface: connect (URL), isConnectible (URL). The method connect (URL) is provides a connection to the web information source. The method isConnectible (URL) is checks for a connection. ProxyConnection class includes fields that contain the data required to make server authentication authorization (Fig. 9b): proxyHost (server authentication name in the form of domain or IP address); proxyPort (server authentication port); proxyUserName (authentication server user login); proxyUserPassword (authentication server user password). All of

these fields are closed due to encapsulation concept. They are accessed using the methods set and get. Also ProxyConnection class includes two major public method that implement the iConnectionProvider interface: connect (URL), isConnectible (URL). The method connect (URL) is provides a connection to the information web source. Method isConnectible (URL) is checks a connection. Also ProxyConnection also includes the following key methods: connectAuthProxy (URL), connectNoAuthProxy (URL), testAuthProxy (), testNoAuthProxy (). IWebInfoSource component consists of classes (Fig. 10): AbstractWebInfoSource, ScienceDirectWIS, CiteSeerXWIS, WileyOnlineLibraryWIS. AbstractWebInfoSource class is an abstract class and implements the interface iWebInfoSource. Classes ScienceDirectWIS, CiteSeerXWIS and WileyOnlineLibraryWIS implement receiving of "raw» not analyzed data from web sources http://www.sciencedirect.com, http://citeseerx.ist.psu.edu and http://onlinelibrary.wiley.com accordingly.

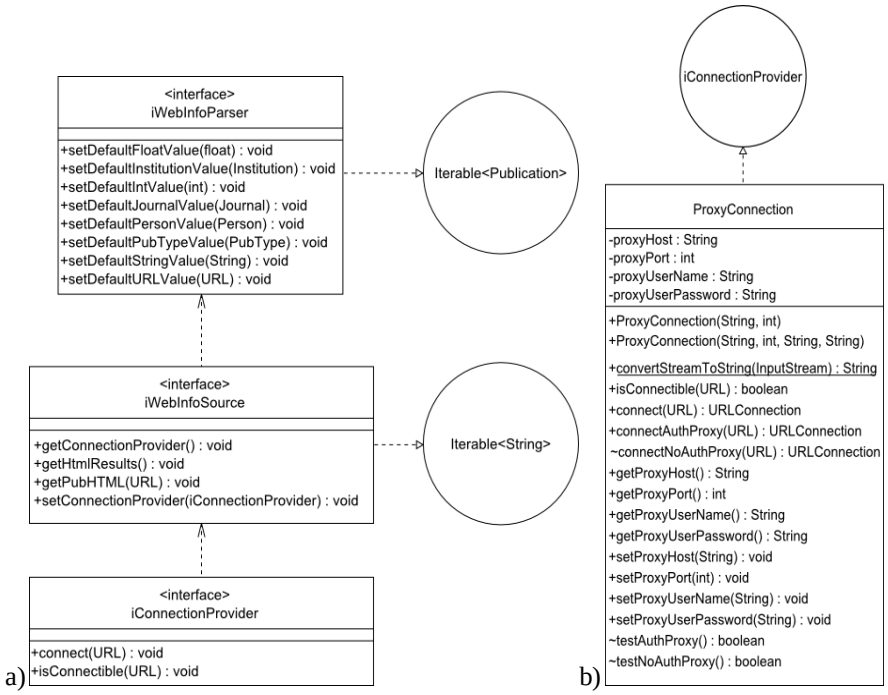

Fig. 9. a) General diagram of the parser structure and b) ProxyConnection UML-class diagram

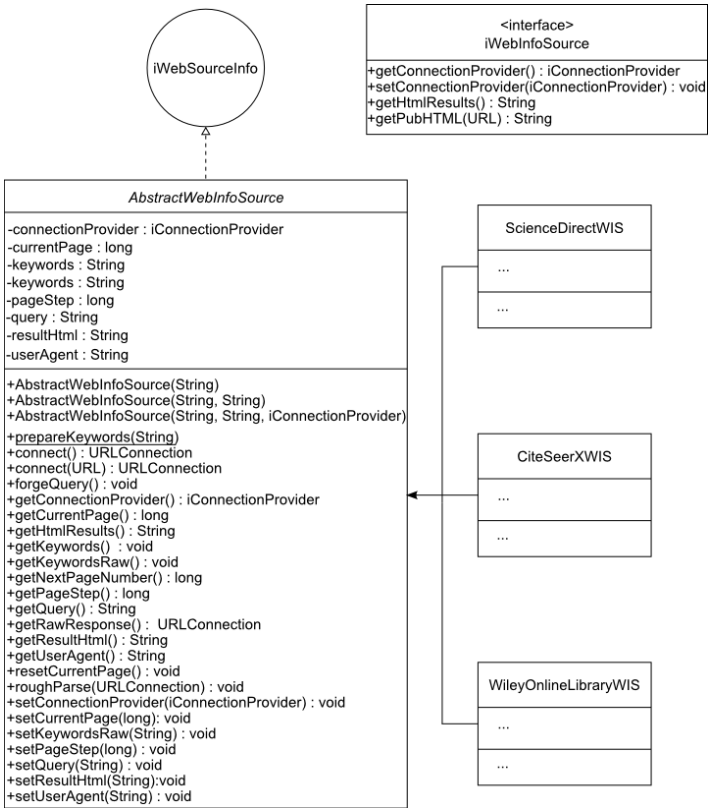

Fig. 10. Schematic diagram of iWebInfoSource component

IWebInfoParser module consists of the following classes: AbstractParser, ScienceDirectParser, CiteSeerXParser, WileyOnlineLibraryParser. AbstractParser class is an abstract class and implements the interface iWebInfoParser.
Classes ScienceDirectParser, CiteSeerXParser WileyOnlineLibraryParser analyzes raw» not analyzed data from component iWebInfoSource.

5 Conclusions

The tremendous growth rates of internet and volume of content stored on its servers requires the creation of tools for automatic content analysis and processing. Intelligent systems for content processing aim to simplify and automate such tasks as content analysis and classification, finding keywords and building digests, distributing content according to specified criteria. This paper presents the results of a study of patterns, characteristics and dependencies in automatic text processing of content. Data and information flows in processes of content transformation are elucidated and formalized. The methods of linguistic analysis are proposed for automation of all operations of content processing. A content management system built using

developed methods is constantly monitoring content from various sources, gathers and integrates content and distributes it to customers. The implementation of proposed methods and procedures allows effectively create and distribute content for targeted social audience and individual customers.

Content forming is implemented as a content-monitoring system collecting the content from various data sources. This enables the creation of a database of content information according to user needs. The result of collecting and primary processing of initial content is new content reduced to a single format, classified according to certain rubricator, and having attributed some descriptors and keywords.

The subsystem of content support provides the formation of information snapshots; the detection of thematic plots; the building of content dependencies tables; the calculation of content rating, the identification of new events in content flows, their tracking and clustering. The analysis of maintenance process of content allows us to determine the causes of the target audience formation using the set of parameters. By adjusting the themed set of content, its uniqueness, its formation efficiency and providing the adequate management according to the individual needs of regular users, the boundaries of targeted social audience and the number of unique visitors from search engines can simulated.

The article describes following results:

- Has been analyzed the development problem of ontology-based information resources processing intellectual systems.
- Has been proposed a new evaluating method of a text document relevance according to the information needs of the information system client, which is based on building information need model in the form of intelligent agent optimal strategy, assessment of expected usefulness and its change due to refinement of the plan by adding information from the investigational document.

The automated ontology synthesis information technology was implemented as software CROCUS, which can be used for proposed method of relevance assessment in human language text information search and knowledge extraction systems.

References

1. Vysotska V., Chyrun L., Lytvyn V., Dosyn D. (2016). Methods based on ontologies for information resources processing : Monograph. LAP Lambert Academic Publishing. Saarbrucken, Germany.
2. Lytvyn V., Pukach P., Bobyk I., Vysotska V. (2016). The method of formation of the status of personality understanding based on the content analysis, *Eastern-European Journal of Enterprise Technologies*, no5/2(83), 4–12.
3. Bisikalo O.V., Vysotska V.A. (2016). Identifying keywords on the basis of content monitoring method in ukrainian texts, *Journal «Radio Electronics, Computer Science, Control»*, No 1, Zaporizhzhya National Technical University, 74-83, Access mode: http://ric.zntu.edu.ua/article/view/66664/0.
4. Lytvyn V., Vysotska V., Veres O., I Rishnyak., and Rishnyak H. (2017). Classification Methods of Text Documents Using Ontology Based Approach, *Advances in Intelligent Systems and Computing* 512, Springer International Publishing AG: 229-240.
5. Ourania Hatzi, Dimitris Vrakas, Nick Bassiliades, (2010). Dimosthenis Anagnostopoulos, and Ioannis Vlahavas. The PORSCE II Framework: Using AI Planning for Automated

Semantic Web Service Composition the Knowledge Engineering Review, Cambridge University Press, Vol. 02:3, 1–24 p. (In English)
6. Lytvyn V. (2013). Design of intelligent decision support systems using ontological approach, *An international quarterly journal on economics in technology, new technologies and modelling processes*, Krakiv-Lviv, Vol. II, No 1, 31 – 38 (In English).
7. Lytvyn V., Dosyn D., Smolarz A. (2013). An ontology based intelligent diagnostic systems of steel corrosion protection, *Elektronika*, Lodzj. – No. 8. – 2-13. – Pp. 22-24 (In English).
8. Lytvyn V. (2011), The similarity metric of scientific papers summaries on the basis of adaptive ontologies , *Proceedings of VIIth International Conference on Perspective Technologies and Methods in MEMS Design*, Polyana, Ukraine, pp. 162. (In English)
9. Link Grammar – Carnegie Mellon University, available at: http://bobo.link.cs.cmu.edu/link.
10. Qiu Ji, Peter Haase, and Guilin Qi (2008). Combination of Similarity Measures in Ontology Matching using the OWA Operator, In Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Base Systems.
11. Gruber T. A. (1993). Translation approach to portable ontologies. Knowledge Acquisition, № 5 (2):199–220.
12. Guarino N. (1995). Formal Ontology, Conceptual Analysis and Knowledge Representation. International Journal of Human-Computer Studies, 43(5-6):625–640.
13. Sowa J. (1992). Conceptual Graphs as a universal knowledge representation. In: Semantic Networks in Artificial Intelligence, Spec. Issue of An International Journal Computers & Mathematics with Applications. (Ed. F. Lehmann), № 2–5:75–95.
14. Montes-y-Gómez M. (2000). Comparison of Conceptual Graphs. Lecture Notes in Artificial Intelligence, Vol. 1793. – Springer-Verlag, Access mode: http://ccc.inaoep.mx/~mmontesg/publicaciones/ 2000/ComparisonCG.
15. Muller H.M., Kenny E.E., Sternberg P.W. (2004). "An Ontology-Based Information Retrieval and Extraction System for Biological Literature". PLoS Biol. 2(11):e309. doi:10.1371/journal.pbio.0020309.
16. Knappe R., Bulskov H., Andreasen T. (2004). Perspectives on Ontology-based Querying // International Journal of Intelligent Systems, Access mode: http://akira.ruc.dk/~knappe/publications/ijis2004.pdf.
17. Jacso, Peter. (2010). "The impact of Eugene Garfield through the prizm of Web of Science,". Annals of Library and Information Studies, Vol. 57, p. 222.
18. Christoph Meinel Serge Linckels (2007). Semantic interpretation of natural language user input to improve search in multimedia knowledge base, Information Technologies, 49(1):40–48.
19. Giorgos Stoilos, Giorgos Stamou, and Stefanos Kollias (2005) A String Metric For Ontology Alignment, Proc. of the 4rd Int. Semantic Web Conf. (ISWC), vol 3729 of LNCS, p. 624–637, Berlin. Springer.

A Method of Construction of Automated Basic Ontology

Vasyl Lytvyn[1], Victoria Vysotska[2], Waldemar Wojcik[3], Dmytro Dosyn[4]

[1,2]Information Systems and Network Department, Lviv Polytechnic National University, Bandery str., 12, Lviv, Ukraine, 79013

vasyl.v.lytvyn@lpnu.ua[1], victoria.a.vysotska@lpnu.ua[2]

[3]Institute of Electronics and Information Technology, Lublin University of Technology, Nadbystrzycka str., 38A, Lublin, Poland, 20618

waldemar.wojcik@pollub.pl[3]

[4]Systems Analysis Laboratory, Karpenko Physico-Mechanical Institute of the NAS of Ukraine, Naukova str., 5, Lviv, Ukraine, 79060

dmytro.dosyn@gmail.com[4]

Abstract. The paper describes an approach to development of a computer system that automatically constructs an ontology base. Basic modules of the system and its operation are described, as well as the choice of software tools for implementation. Application of the proposed system allows to fill the domain ontology in an automatic mode. Therefore, this paper introduces an approach to development of an automated basic ontology composition. An architecture of synthesis of the ontology system is created using CROCUS (Cognition Relations or Concepts Using Semantics) software model. The main system modules and their functions are described. A decision of SDK for system realization is justified. Application of the proposed system can fill an ontology of subject area automatically.

Keywords: computer system, ontology, knowledge base, database, text document, machine learning, intelligent agent, utility, semantic, logic of predicate.

1 Introduction and review of current research on the topic

A study of results and developments in the area of intellectual information systems and Internet services [1-8] led us to suggest the following soft/hardware decisions:

- Realization of the ontology synthesis system as a subsystem of the Internet portal system;
- Using OWL as a knowledge presentation language;

- Using HTN and OWL-S as structures of the automated knowledge base planning language;
- Java API for Protégé OWL as the API and the library processing classes, in particular for the machine learning (reinforcement learning) of the OWL-ontology and knowledge bases;
- Link Grammar Parser as an instrument of the grammatically-semantic analysis of English text documents [2];
- Apache-PHP-MySQL as a software tool to build a web-portal based user interface;
- Wget as a web service for automated access to search engines with a query, formed from the keywords;
- SWRL as a logical new knowledge output language with deductive and inductive methods;
- WordNet as the basic glossary of English.

An ontology in OWL language contains a high-level meaning automation of the subject area [3]. A high-level ontology provides [4]:

- A logical output of new knowledge with the addition of new messages with the context [5];
- The verification of the validity of obtained statements [6];
- The evaluation of the probability of the message sources [7];
- The ensuring of the knowledge base logical integrity [8].

The machine learning is provided with the means of Java API Protégé-OWL. These means contain libraries of classes, which realize methods to work with OWL-structures like reading and addition. Therefore, the tools of machine learning work in addition to the OWL-ontology. Java API Protégé-OWL takes templates of grammatically semantic structures to recognize statements (the first order predicates) into the research and/or educated texts including new elements as the result of such recognition [9]. Link Grammar Parcer [2] divides a grammatically correct definition of sentence into interconnected pairs of words. LGB contains a table that has all the conformities between grammar constructions of English and syntax-semantically links between words (intellections). LGP API allows to link this table to OWL-ontology, so the table can adapt in the process of learning the given object area dynamically. Java API Protege-OWL based means of machinery education contain a generalized description of the semantic link, which serves as the template for generating new types of semantic link during studying. In addition, it forms appropriate vectors and indications of these links to form and identify semantic links in a text. Herewith, properly classes of links and their properties adding to the OBP. Exemplars of those classes are for the description of existing and new classes of ontology by their use as first rank predicates.

ZMN ontologies make sense only as the part some intellectual system. Optimal solutions in our opinion are that, where such an intellectual system in the information search system for which an adaptive ontology is an instrument for information research, analysis and classification on the one hand, which uses search instrumentalities to provide data for its filling, new predicates and rules synthesis, learning new means and semantic links on the other [10-11]. An intellectual system of information searching based on the adaptive ontology, material science knowledge base, a database of scientific publications became such a decision [12-13].

A developed architecture of the ontology synthesis system was realized with the usage of selected and decrypted means and program-technically solutions as a CROCUS (Cognition Relations or Concepts Using Semantics) [14-17].

2 Main modules of CROCUS

The overall concept of CROCUS at the fig. 1 is introduced. A subsystem of the ontology learning uses training texts of annotations of scientific publications from article DB. The system forms a plural of key words to fill the DB in. It chooses the main metadata about the publications in the defined subject area in Internet (ScienceDirect, CiteSeer, Wiley Online Library, and Springer) including their annotations, which become the core of analysis and ontology learning.

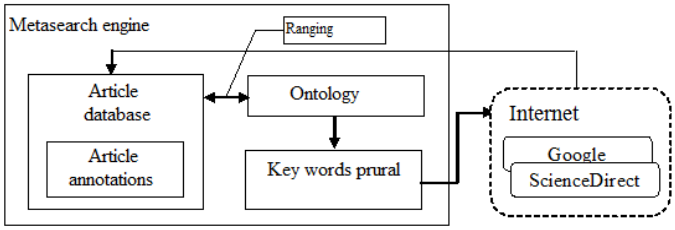

Fig. 1. CROCUS concept schema

The essence of the knowledge extraction method from the natural text document is into building of the intellectual agent activity strategy – an informational model of the recognition subject of its specification based on dedicated from recognized text document data. A plan is considered to be a specific optimal strategy realization of some task, which has an intellectual agent within the subject area.

The plan is built with the same with informational model formal knowledge representation language – a database of an intellectual agent. Considering that, such a knowledge base is already an overall plan of intellectual agent functioning, build basing on the natural text recognition is a sub-plan. It means that it is a specification of an overall plan and it bases on it. A value of information, received as a result of recognition the context of a text document is determined as increasing of the updated intellectual agent functioning plan expected utility. Scientific publications range for the relevance to the users informational demands, for the conformity to ontology, which displays these demands. An analysis of each annotation as natural text is made, builds its image in the terms of ontology as predicates and rules in this purpose. These predicates and rules are added in the knowledge base of the system and the expected utility of an intellectual agent is calculated again. A system puts those publications nearer to the beginning of the list, which data including leads to the greater reliability change with such a type of ranging.

A system can adapt to the users requirements by saving his preference system in the DB. Each user can perform an education of his ontology. The system saves the data about this process, leads the session statistics, and provides the possibility to correct the education errors and does backtracking to previous versions of ontology.

CROCUS system modules are shown at the picture 2. A client has a possibility to control the priority of document ranging, to correct their order in the list of the most important (relevant to the client informational requirements) document and classify them with the help of graphic interface. The most important documents are used for ontology learning and building of the efficient sets of key schemas and new, received from Internet, articles (their metadata including annotation) insert into the DB of publications with the link with preferences of the user and other prerequisites of the document receiving.

An annotation processing after its previous processing, conversion into the massive of predicates as a result of grammatically-syntax analysis of Link Grammar Parseris executed. Formed annotation models are supplemented with semantically near ontology predicates – the context of this annotation. Supplemented annotation models are compared between themselves to calculate the semantic length between their semantic weight midpoints and so the nearest by content documents are chosen with their further ranging and classification. Two main functions of CROCUS:

1. An interactive automatic building of the problem area ontology;
2. Searching, saving and classification (ranging) of scientific publications as in interactive semiautomatic as in automatic mode.

Each of these functions is realized with its base set of functionality modules but a part of them was a double appointment. CROCUS is realized as an object oriented paradigm by using Java as a hierarchy of code classes, which copies call each other with determined at that moment parameters or they interact through throw events and/or handlers. Most of them have a Swing graphical interface and AWT libraries. All the connected libraries have an open source status. Project has a full functionality and has all the necessary means for its development (evolution). A functional assignment of the main CROCUS system modules is shown at the fig.2.

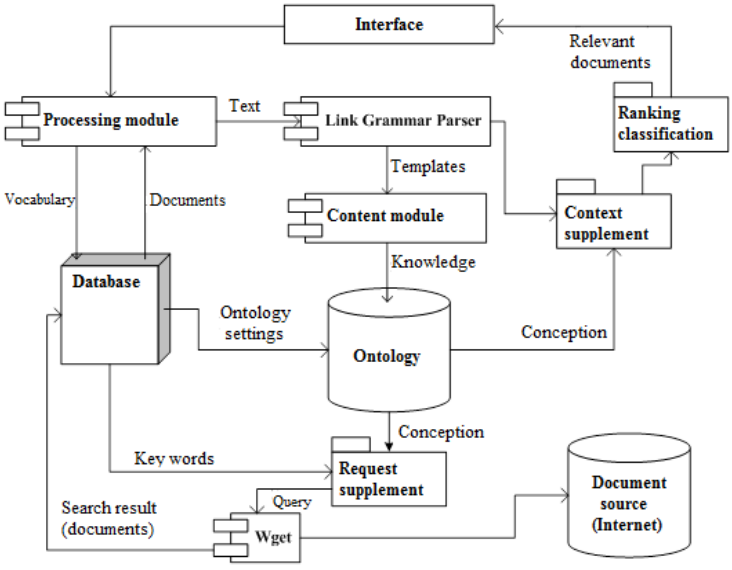

Fig. 2. CROCUS modules

Detailed analysis of similar systems is made in [14-17].

3 Functionality of CROCUS modules

Functionality of the main CROCUS modules is described in table 1.

Table 1. Functional purpose of the main CROCUS modules

№	Modules	Functionality
1	Attribute.java	A subprogram to detect attributive links in sentences, or rather verbal links. Each adjectival link is a form of verbal, actually: "A tomato is green = a tomato**'s color is** green". Herewith an ontology has to contain a binary predicate "**color is**" (<domain> = <tomato>, <range of acceptable values> = <a green color>)
2	BinaryLink.java	Determination of the horizontal binary link weight
3	CGrammarObject.java	Breaks sentences into words-tokens which are accompanied by a letter, which means a type of word (n-noun, v-verb etc.), so it is possible to get words and its language type using methods **getName**() or **getType**(). Type is determined by LinkGrammarParser
4	CGrammarObjectArr.java	A copy constructor of this class forms the massive of grammatical objects from the text line 'sentence'
5	CGrammarObjectType.java	Class of elements type massive CGrammarObjectArr
6	CLinkType.java	Class of methods of breaking link type arrays between pairs of words in sentence at the main type of link **getType**() and its subtypes **getSpecification**()
7	ControlGUI.java	The main program control window
8	CSOLink.java	Breakes the results of Link Parcer work – determines the conformity of specified by Parser types of link to the words, marked by brackets into the previous string
9	CSOLinksArr.java	Puts exemplars of CSOLink type into a dynamic array
10	Descriptor.java	Forms a prural of patterns into one array
11	DParsedData.java	Forms text lines with the results of LGP work
12	FunctionGUI.java	A graphical interface template
13	GSL.java	Forms a pattern of semantic link type to recognize semantic links by their patterns
14	Is_a.java	Procedure of adding subclasses to existing classes
15	MainProc.java	The main procedure in 'non-graphical variant' – it is not supported since 02.05.2011 12:38. 1. Creates constant parameters to initialize an external for this package Link Grammar Parses; 2. Opens 3 threads: to enter Parser, to output its data and errors; 3. The online URL-address of an ontology is indicated; 4. The name of text file, which will extend an

№	Modules	Functionality
		ontology, is read from the string of program initialization (1st parameter in string); 5. An ontology is read from the specified address; 6. Its contents is put into standard output channel; 7. Contents is put into the resulting.owl file using the DumpOWLModel procedure; 8. Searching Is-a links; 9. Searching Consist-of links; 10. Executing an Attribute procedure.
16	MetaObject.java	Creates a structure to describe the semantic object – a subject of action or an action as a type of link between subject and object
17	OntologyMapClass.java	Adds classes to the indicated level of an ontology **(addClassesToLevel)**
18	OWLEvaluation.java	The most applied procedure of this class (its objects) • Calcuting the weight of concepts and link of an ontology • Output an antology into <file.owl> using DumpOWLModel(String file_name) • ShowAll – making the procedure of visualization into the standart output channel.
19	Parser.java	Initialization of Link Grammar Parser, its customization, inputting the text file with sentences, saving the result of Parser execution into it.
20	Part_of.java	Recognizes semantic links 'Part-of' type into English sentences (after their processing into Link Grammar Parser).
21	Pattern.java	A class Pattern constrictor, in which creates a semantic link pattern **for its recognition into sentences using static method.** 3 main elements are collected "subject -> meta link -> object", in particular "subject -> metalink" and "metalink -> object". They became during education of this semantic link so they play the main role (because they are the most common). Statistics saves as exemplars of the semantic link class in an ontology: default_<semantic_link_name>_ <metalink:{D, S, O,...}>_<Subject|Object>_ <subject/object_name>
22	SemLinkDescriptArr.java	Chooses a sequence of SemLinkDescriptor type descriptors from the investigated sentence and creates a dynamic array of them.
23	SemLinkDescriptor.java	Triplet semantic link descriptor: metaSubject (string) -> link-type (string) -> metaObject (string) A semantic link signs vector has to consist of such triplets. Each triplet of each semantic link type has its value coefficient. A size of such a vector is 10.

The basic system control element in CROCUS is ControlGUI module (user graphical control interface). This module has a graphical interface by which user can

execute procedures, which are provided by the system functionality. Module has the main menu and its main functions are in the toolbar. Control output is carried out at the appropriate text panel. There is an output field and an input field at the underside of the main window to specify the semantic link type at the process of ontology learning using learning sentences (fig. 3).

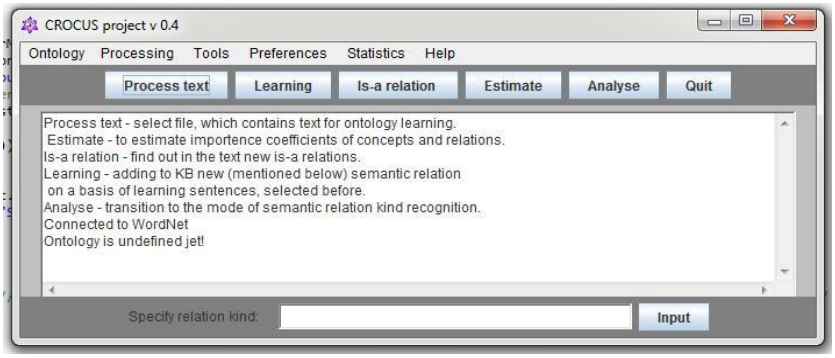

Fig. 3. The main window of CROCUS user interface

Despite the importance of the effective dialog between system and user, a great attention during developing was to the graphical interface design, its intuitiveness and pithiness with a functional completeness of project tasks realization. In addition, there is a possibility to zoom interface to expand its functionality. Despite the intense competition between the similar projects, a great importance was to create a recognizable logotype, which will be replied to the content of the word CROCUS. An experience of such foreign projects confirms an efficiently of such an approach (Protégé to work out with OWL-Ontology, project GATE, etc.). That is why all the windows in dialogue with user are decorated with CROCUS logotype – an illustration of 6-petal saffron (crocus) flower. Professional designers developed an image and the design of main windows interface.

System has an internationalization of whole text dialogues. User can choose a comfortable interface language from four available. There is no problem in language addition. A dialog language file MessagesBundle_xx_XX.properties has to be translated, where XX – the code of a language (RU – Russian, UA – Ukrainian etc.). To choose the dialogue language you have to choose a subparagraph ‘language’ into the paragraph of the main menu ‘Preferences’.

4 Justification of the SDK choose

A using of SDK common libraries gives a chance to avoid unjustified overrun of time, finances and human resources for their redevelopment. Therefore, there is a wide list of investigated currently working analogue projects in this work. Most of them use the concept of open source code and free licensing. The leading developers groups provide their projects with API (Application Programming Interface) means, so the functionality of these objects may be efficiently used by cataloged and well-

documented procedures and functions with appropriate settings.

Co-authorship of SDK developments worked out principles of software application usage with different license agreements. In addition, they can take part in support and development of existence projects, so each developer has a possibility to get and install these project or libraries and use them as he or she wants. Such internet portals as SourceForge.net contain all the necessary instrumentals for documentation and support projects of every level of difficulty, readiness, access level and popularity between users. Developers actively use special foundation servers, which provide collective (let it be 1000 developers) software developing. The most popular foundation server is Git. It can be installed separately as individual or corporative server and it is possible to use a global GitHub server.

Researches show, that most it develops in text documents natural processing, almost all develops in ontology learning are performed on Java. Moreover, Java is dominating between projects languages at SourceForge resource.

A decisive argument to Java usage is an accessibility of Protégé-OWL Java API of Stanford University (USA), because Stanford Center for Biomedical Informatics Research became a flagman of practice developments in OWL SDK-s.

Projects made by Java:

- Gate [http://gate.ac.uk/] – a couple of text documents processing means to find a new knowledge;
- owlapi.sourceforge.net – another one Java project, which is an OWL documents processing Java classes library with broad functionality;
- Pellet [http://clarkparsia.com/pellet/] – a logical output machine to realize thinking (new knowledge output) from OWL 2.0 knowledge base.

5 Conclusions

Therefore, this work shows an approach to develop an automated basic ontology building. An architecture of ontology synthesis system as CROCUS (Cognition Relations or Concepts Using Semantics) software model is created. The main system modules and their appointment are described. A decision of SDK for system realization is substantiated. A usage of such a system can fill an ontology of subject area automatically.

References

1. Ourania Hatzi, Dimitris Vrakas, Nick Bassiliades, (2010). Dimosthenis Anagnostopoulos, and Ioannis Vlahavas. The PORSCE II Framework: Using AI Planning for Automated Semantic Web Service Composition the Knowledge Engineering Review, Cambridge University Press, Vol. 02:3, 1–24 p. (In English)
2. Link Grammar – Carnegie Mellon University, available at: http://bobo.link.cs.cmu.edu/link.
3. Qiu Ji, Peter Haase, and Guilin Qi (2008). Combination of Similarity Measures in Ontology Matching using the OWA Operator, In Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Base Systems.

4. Lytvyn V. (2013). Design of intelligent decision support systems using ontological approach, *An international quarterly journal on economics in technology, new technologies and modelling processes*, Krakiv-Lviv, Vol. II, No 1, 31 – 38 (In English).
5. Gruber T. A. (1993). Translation approach to portable ontologies. Knowledge Acquisition, № 5 (2):199–220.
6. Guarino N. (1995). Formal Ontology, Conceptual Analysis and Knowledge Representation. International Journal of Human-Computer Studies, 43(5-6):625–640.
7. Sowa J. (1992). Conceptual Graphs as a universal knowledge representation. In: Semantic Networks in Artificial Intelligence, Spec. Issue of An International Journal Computers & Mathematics with Applications. (Ed. F. Lehmann), № 2–5:75–95.
8. Montes-y-Gómez M. (2000). Comparison of Conceptual Graphs [Електронний ресурс]. Lecture Notes in Artificial Intelligence, Vol. 1793. – Springer-Verlag. Режим доступу до журналу: http://ccc.inaoep.mx/~mmontesg/publicaciones/ 2000/ComparisonCG.
9. Muller H.M., Kenny E.E., Sternberg P.W. (2004). "An Ontology-Based Information Retrieval and Extraction System for Biological Literature". PLoS Biol. 2(11):e309. doi:10.1371/journal.pbio.0020309.
10. Knappe R., Bulskov H., Andreasen T. (2004) Perspectives on Ontology-based Querying // International Journal of Intelligent Systems. – http://akira.ruc.dk/~knappe/publications/ ijis2004.pdf.
11. Jacso, Peter. (2010). "The impact of Eugene Garfield through the prizm of Web of Science,". Annals of Library and Information Studies, Vol. 57, p. 222.
12. Christoph Meinel Serge Linckels (2007). Semantic interpretation of natural language user input to improve search in multimedia knowledge base, Information Technologies, 49(1):40–48.
13. Giorgos Stoilos, Giorgos Stamou, and Stefanos Kollias (2005) A String Metric For Ontology Alignment, Proc. of the 4rd Int. Semantic Web Conf. (ISWC), vol 3729 of LNCS, p. 624–637, Berlin. Springer.
14. Lytvyn V., DosynD., Smolarz A. (2013). An ontology based intelligent diagnostic systems of steel corrosion protection, *Elektronika*, Lodzj. – No. 8. – 2-13. – Pp. 22-24 (In English).
15. Lytvyn V. (2011), The similarity metric of scientific papers summaries on the basis of adaptive ontologies , *Proceedings of VIIth International Conference on Perspective Technologies and Methods in MEMS Design*, Polyana, Ukraine, pp. 162. (In English)
16. Lytvyn V., Pukach P., Bobyk I., Vysotska V. (2016). The method of formation of the status of personality understanding based on the content analysis, *Eastern-European Journal of Enterprise Technologies*, no5/2(83), 4–12.
17. Lytvyn V., Vysotska V., Veres O., Rishnyak I., and Rishnyak H. (2017). Classification Methods of Text Documents Using Ontology Based Approach, *Advances in Intelligent Systems and Computing* 512, Springer International Publishing AG: 229-240.

Content Analysis of some Social Media of the Occupied Territories of Ukraine

Volodymyr Lytvynenko[1], Iryna Lurie[2], Svitlana Radetska[3], Mariia Voronenko[4], Natalia Kornilovska[5], Daria Partenjucha[6]

Informatic and Computer Science Department, Kherson National Technical Univesity, Bereslav Shosse, 24, Kherson, Ukraine, 73008

immun56@gmail.com[1], iil@rambler.ru[2], rad_svete@rambler.ru[3], mary_voronenko@i.ua[4], knv06061971@rambler.ru[5], partenjukha@rambler.ru[6]

Abstract. The paper analyzes the activities of the Internet publications in the temporarily occupied territories. The analysis of the tools for the detection and monitoring of free and paid services, data analysis of social media to monitor references and comments on social networks are presented in the paper. This paper includes an example of a practical text analysis and extraction of information online. It is shown that the use of computing environment KNIME opens new pathways in a variety of social media and determine the user behavior. This technique improves the process of analytical studies.

Keywords: Social Media, KNIME, Content Analysis, Network Analytics, Text Mining, Occupation

1 Introduction

Analysis of the texts is one of the most common types of the scientific and the scientific and practical analyses. Content analysis relates to the researches affecting the text, images, audio and video and allows to write content in a social context. Understanding the social meaning of the document involves the process of identifying of the responses the document may receive in public life, the degree of originality of the document, its difference from other documents of a particular kind. Today there is a popular database based on the archives of the media. Despite of the valuable role played by the Internet in the media, the reasons for its development and the role of market information has been studied insufficiently.

To determine the general orientation of the electronic media source vectors it is needed to consider their similarities and differences in order to identify social factors that affect this orientation, to investigate them from a scientific point of view.

The essence of this research is due to an attempt to analyze the specific functions of the electronic media in the temporarily occupied territories of Ukraine. Taking into

consideration the fact that Ukraine is currently in terms of the information war, it is necessary to obtain full and objective view of the events, which take place there.

The aim of this paper is to consider the role of the information technology use for the analysis of the electronic media in the temporarily occupied territories of Ukraine.

Monitoring and analysis of the electronic media is the systematic tracking of the news reports from the Internet resources, which enables timely to identify and predict the trends in the emergence of competitive situations using quantitative and qualitative analyses. Quantitative analysis is used for the primary assessment of the reports on the objective activities of analysis in the mass media. It gives the possibility to estimate the overall amount of time to cover the subjects and themes selected for analysis in the electronic media.

Qualitative analysis is used to examine the main characteristics of the information field, forms of its presentation, its emotional orientation or absence of infringement of the legislation and to evaluate the information concerning the object of analysis as positive, negative or neutral according to its content.

When it is impossible to conduct qualitative and quantitative analyses we are able to use media effects, such as plots designed to manipulate public opinion. The main criterion for the use of this assessment is the lack of relevance, accuracy, transparency, facts, balance, diversity, timeliness, clarity.

2 State of modern informational on-line publications

There is no clear definition of the Internet media. It is believed that the Internet media (Internet edition online newspaper) is the regularly updated information site that serves as the media and has its permanent audience. It differs from the traditional media only in the field of activity whereas the functions and purposes are identical [1, 2]. Most online publications, which work as the news agencies are not registered due to the lack of the legal regulations of the Internet publications. This fact leads to misunderstanding of the situation by journalists who are beginning to think that there is a complete lack of permissiveness and responsibility for propagating information use [3, 4].

Media often faces the shortcomings of the national legal framework. This makes the media vulnerable to the pressure from various institutions (public, private and criminal), manipulation and intimidation of the political elite, low-culture audience of regional media and, conversely, very high level of self censorship by the media representatives in relation to coverage and analysis of the political sphere [5, 6]. Massive social and political actions against the Ukrainian authorities to protect the Russian language, during anti-government, federalist, pro-Russian and separatist slogans which were shared at the end of February – beginning of March 2014 in the cities of the south-eastern Ukraine after the power change of government, exacerbated the conflict between the West and East Ukraine and led to the instability and separating of the society [7].

As a result the Crimean Peninsula was annexed by the Russian Federation, and the unrecognized states, the so called Donetsk and Luhansk Republics were proclaimed. Citizens involved in the activities of the republics were called separatists and terrorists. They were engaged in extremist propaganda and military action. Due to this the network media representing a radical resources aimed at inciting ethnic hatred, the implementation of pro-Russian views, opinions and ideas emerged.

List of the electronic media in the temporarily occupied territories of Ukraine, is given in Table 1.

Table 1. SWOT-analysis of the electronic media

Name	Options	Advantages	Disadvantages
"RIA News of the Crimea" http://crimea.ria.ru/	Providing of timely information and news about events in the Crimea, Russia, world	The speed of information. International distribution. Easy navigation of the system	Trying to present impartial news, but the problem is the contradiction between global and national interests.
"News of the Crimea" http://news.allcrimea.net/	Independent online publication that covers events that happen in the Crimea	Support news archive, functioning of keyword search, the opportunity to discuss the news on social networks.	Many embarrassing and news layouts, making it difficult to view.
News media "Russian Spring" http://rusvesna.su/	Informative and analytical portal that provides quick information about events in New Russia, in Ukraine and in the world	Represent the views of the experts, politicians, reports from the places of military events, information is presented in 5 languages, multimedia information.	Submission of false information. Too much advertising. At this website in section "Help" fundraising services for the sustenance of the resource is held. But, it is not clear who gets the money
Internet publication "DNR24" http://dnr24.su/	Information about events in eastern Ukraine, Donetsk and Lugansk regions.	The speed of information. Simple and easy navigation of the system.	The lack of censorship. The radical nature of news. Images to the users. Worsening political and armed conflict
News Agency "Antifa-shyst" http://antifashist.com/	Anti-Fascist Forum of Ukraine, which aims at presenting the news on political and military action.	Multimedia. A large number of articles.	Inconvenient online design of the resource. The subjectivity of the information provided. Shocking news content.
News Agency "The news front" http://news-front.info/	An independent agency network. Truthful information about the events in the New Russia, Ukraine and the world.	The ability to view the site in the 5 languages. Apps for mobile platforms IOS and Android.	Conducting surveys from the front line and exaggeration of the events taking place in the temporally occupied territories, distorted information.
"The truth of the NPT" http://dnr-pravda.ru/	Independent publications of Donetsk Republic.	Presentation of current political news.	The propaganda nature of the news.
"DAN" - Donetsk News Agency http://dan-news.info/	The agency covering events in the Donbass region and the Donetsk Republic.	Convenient ways of navigation and search of the information.	Misleading false news. Incitement to hatred and discrimination.

In today's world of mass media, there is a huge amount of available data. The problem is to convert them into useful and relevant form. Under these conditions, the actual content is the personification of the global system WWW, which is impossible without the development and implementation of the appropriate techniques and tools [8]. A large number of messages, comments which reflect the mood of the users in terms of escalation and information confrontation have been considered [9].

KNIME is easy to use as a graphical tool for all types of the process analyses - data access, data conversion, initial study, the powerful mining, visualization and report's generating. This open integrated platform offers more than 1,000 modules (nodes), developed by the community KNIME. The procedures implemented through the workflows and workflow consists of nodes. The Nodes (Units) are responsible for the implementation of various procedures in the working process, can be found in the "repository nodes." Data View are the components which serve to visualize data (graph, charts). Each node has a configuration setting window. These tools can give an initial overview. They are not suitable for a deeper understanding of the behavior, needs, problems, desires or tendencies of the individuals as these tools and services do not actually provide any data for presenting of the summarized data [10]. Combining of text mining and network analysis have been conducted for the Information Agency "anti-fascist" - a news resource that aims to present reports about political, economic and military action in the temporarily occupied territories of Ukraine. In general, the community contains about 7,000 comments, 10,930 articles on politics and about 85,000 users. Community of members is very active with more than 100 responses to the topic. The majority of people leave their comments and are registered under their nickname, but some comments are anonymous. In particular, the Twitter account of the online edition "anti-fascist", which has 1,389 readers, followers and 28,839 tweets has been analyzed. With the help of the Follower wonk the data about users, followers of the pages were obtained. Predictive text analysis, network analysis methods and inverse transformation of the raw data into the usual information were applied basing on the information that is the methods of clustering and modeling were used. Twitter-appearance of the account is presented in Fig. 1.

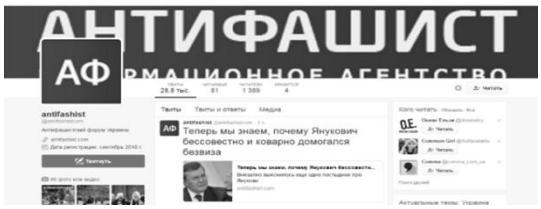

Fig. 1. Twitter account of the anti-fascist news agency

To understand the mood of the user it is necessary to determine the level of his/her relations up the nature of his/her comments, accordingly they can be positive or negative. The level of relationship can also be used to classify users in future. In order to classify the mood of the lexicon containing words (tips) in addition to other information, their polarity is used to indicate the words as positive or negative. There polarity: positive, neutral or negative has been used. For each user-follower the

frequency of positive and negative words use has been determined and calculated in accordance with the participation in the anti-fascist page retweets. The difference between these frequencies determining the attitude of the user is the following:

$$\lambda (u) = f_{pos} (u) - f_{neg} (u), \quad (8)$$

where f_{pos} (u) – the frequency of the positive words use; f_{neg} (u) - the frequency of the negative words use. Positive λ – defines positive users and negative λ – negative users.

In the first stage of the process, several traditional anti-fascist processing reading units for data mining are used. Then each post is converted into the data type of the document for further text analysis operations. Finally, the node "dictionary Tagger» tag links polarity of each word in the document column (Fig. 2).

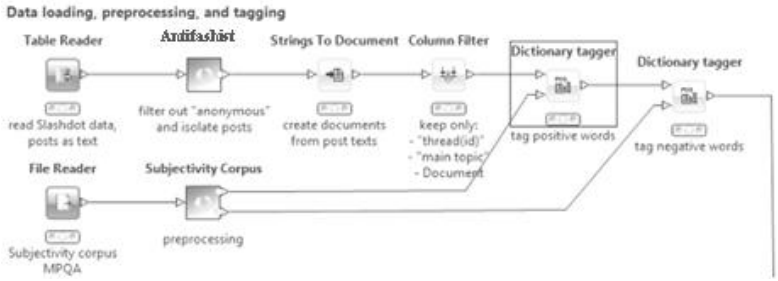

Fig. 2. Preliminary processing of messages in the workflow

Now all the words in tweets and retweets are marked as positive or negative. It is possible to begin the estimation of the relationship of each user. There are two types of important nodes in a category Text Processing KNIME:

- Node “BoW- creator” accumulating the words (Bow) for a set of documents consists of two columns: one containing the document and another containing the terms occurring in certain tweets.
- Node «TF» calculates a relative term frequency (TF) of each part of the documents and adds a column containing the term frequencies, calculated by the division of the absolute frequency of the term usage, which is found in the post, the number of terms of the post.

The frequency of negative and positive terms is aggregated above the user’s ID to obtain the general frequency of negative and positive words use. Then the level of each user relationship is calculated as the difference between the frequencies of terms (Fig. 3).

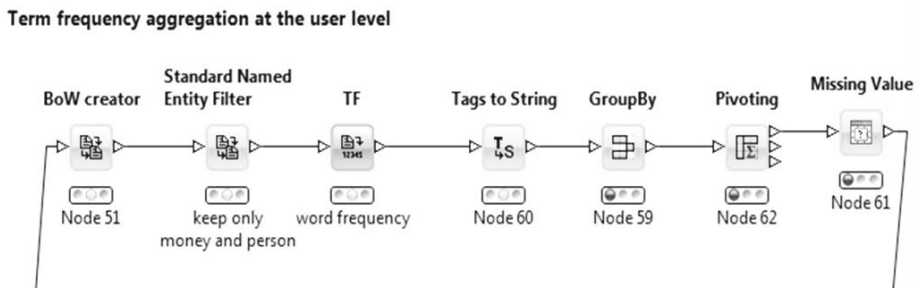

Fig. 3. Part of the process of calculating the frequency of terms at the user level

Members were divided into two categories namely "positive" and "negative." Suppose that a custom of the level relations are Gaussian distribution around the mean with variance $\mu\lambda$ and $\sigma\lambda$ and the most users around $\mu\lambda$ are neutral. Therefore, it can be foreseen that the users with the level of relationships λ within $\mu\lambda \pm \sigma\lambda$ are neutral, while the users with λ in the left turn of Gauss distribution ($\lambda \leq \mu\lambda - \sigma\lambda$) are negative and the users with λ in the right turn of Gauss distribution ($\lambda \leq \mu\lambda + \sigma\lambda$) are positive. Based on the calculated values for $\mu\lambda$ and $\sigma\lambda$, the results of binning process is 807 negative users and 582 positive users.

In Fig. 4 a graph distribution of all known users and followers is shown. The X-axis represents the frequency of the positive words use, Y represents the frequency of the negative words use by the users. Negative users are painted red, positive users are painted green.

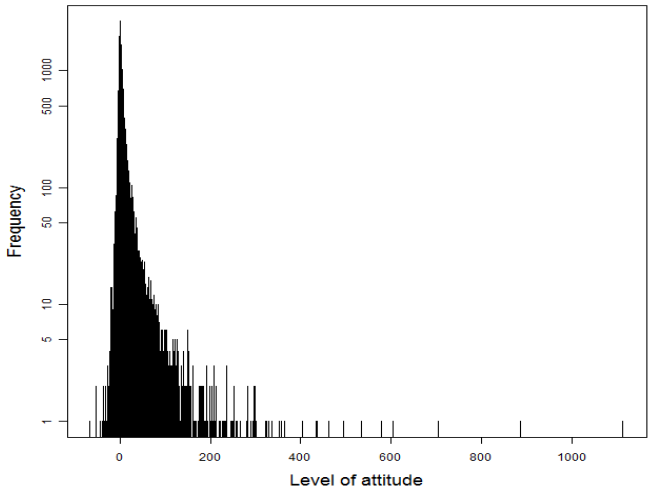

Fig. 4. Distribution of the users attitude

The user who uses a great amount of different words (positive and negative) is «Turmalay», it can be seen in the upper right corner in Fig. 5. However, it is not a user with the highest level of relations, as the level of its social authority is 60.

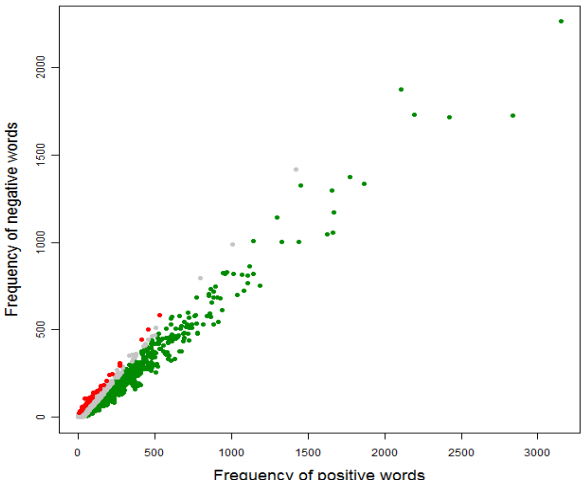

Fig. 5. Diagram of frequency of the negative words use in comparison to positive ones use for all users

The average frequency of words (positive and negative) used by positive users is 418, which is almost twice as bigger than the negative users – 217. Thus, the negative users often do not write.

The final workflow for processing and retweets records is shown in Fig. 6.

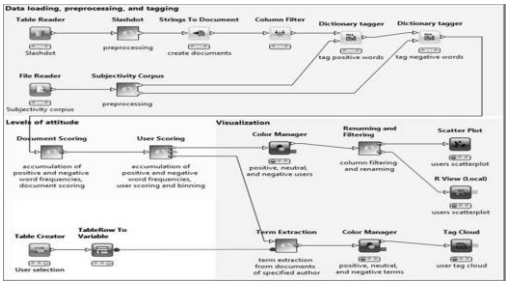

Fig. 6. KNIME workflow word processing

The main purpose of the network analysis is to identify the leaders and followers based on the status and structural state of the users in the current network. After filtering of all articles and comments of the users, 26 non-interconnected components were created. 25 contained only three smaller tops and one - 24.055 tops from 98.150 countries. The network, created on the basis of anti-fascist data page is extremely complex (Fig. 7). The importance of networking visualization becomes clear, the focus on specific areas with identified leaders and followers is considered to be of great importance. Leaders are the users who create their own tweet or comment, which becomes the topics for discussion, for example tweets concerning vital political issues. These users may be of interest to those individuals who are involved in the formulation of the public opinion, as they attract a lot of attention to their

publications. Followers are the users who creates retweets of the leaders' entries but they do not get the comments by themselves.

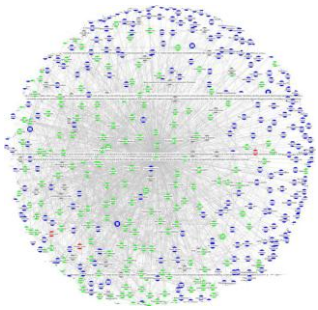

Fig. 7. Full network of the Internet publication (интернет-издание)"Antifascist"

In order to identify the leaders and the followers let us use centrality index from web analytics. This figure is based on assigning each vertex of two different values, combining the weight of the authority and the weight of the node. Peak meets the high weight of the hub when it refers to large numbers and high level of credibility. Therefore, the high weight is assigned by the hub users who often react to the articles published by others.

Fig. 8 shows the diagram of leaders scattering against the weight for all users belonging to the major components. X-axis represents the follower account in the calculation of the mass of the hub, while the Y axis represents the account of the leader per weight of authority.

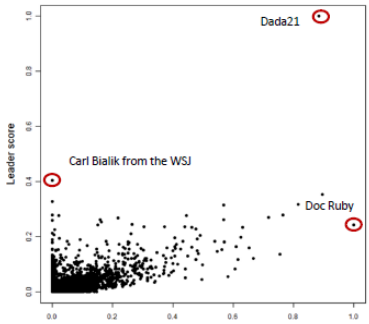

Fig. 8. Plot band of the leaders against the followers among the total number of users

User «Vadim Stolicin» attractes immediate attention and the highest level of the authority equals to 1, and high concentration estimates 0.9. So he gets a lot of comments from other users to his / her message (high score leader) and at the same time he/she often comments the articles and comments of other users. This person really is one of the most active at discussions of political issues in the anti-fascist page. Another user «neokomm» also can be of great interest for the investigation. The user has the greatest weight of authority 4, but a very small concentration of estimates 0, meaning that it has a very large number of followers, but never responds to their

records.

On the opposite side of the scattering diagram, you can see user's «nyepx» data, who has the highest concentration of 1 ratings and moderate the authority of 0.2, which means that he leaves a lot of comments on the reports of others, but rarely writes his/her own posts, and even if he / she does write posts, they rarely get comments. He/she is one of the best followers.

Fig. 9 shows the workflow KNIME, which filters out all anonymous posts and users, creating a network of people based on the anti-fascist data set, it draws the largest component, calculates the weight of the authority for each user using R Networks of obtaining 7 plugins, and visualizes the weight and the credibility into the dispersion chart.

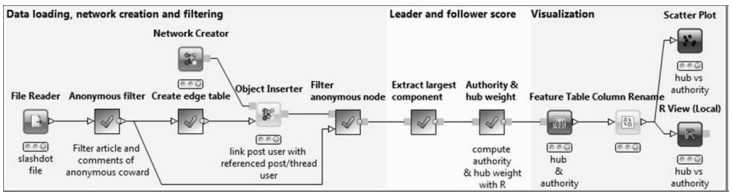

Fig. 9. Hybrid KNIME and R workflow to extract information from the network

Information about the actual author of the text, sentiment expressed in it, as well as graphs and numbers of readers and defendants can not reveal the position of the person in relation to all others in the community, and can not detect the interaction between this user and others.

The network analysis is suitable for detection of anomalies, such as "I will vote for you, and you will vote for me," which is a classic problem to extract text used in the analysis of sentiment. The workflow which combines the text mining results with the results of the network analysis can be seen on Fig. 10

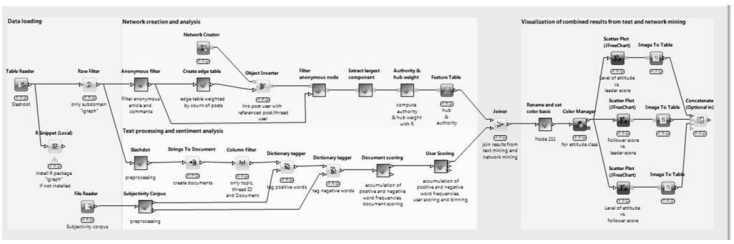

Fig. 10. Workflow KNIME combining the analysis of connections and text mining

Based on the analysis of the attitudes and the extracted from the texts information it is possible to place each user to schedule scattering with level of orientation on the X axis and the pusher or the account leader on the axis Y.

Not all the positive attitudes of thinkers on the right of the diagonal will matter for our marketing campaigns. In fact, despite the positive thoughts in the information they mostly react to someone else's original thought (positive or negative). On the other hand, there are some friendly-oriented users, whose attitude is above the

diagonal, they are definitely the leaders. These users can be considered as those whose positive contributions are determined by the community attitude.

3 Conclusions

The main trends and online publications in the temporarily occupied territories were analyzed. There is a wide range of tools for identifying, monitoring and analysis, there are both paid and free services for analyzing social media data. All these tools allow to track the comments and references in the social networks.

The combination of intelligent text analysis and extraction of information from various offers new possibility to penetrate into social media and determine the user's behavior that would be impossible using each approach separately.

This approach can also be improved by including additional sources of relevant data relating to specific priority areas, such as company and product names, political parties, well-known users. Additional data will further strengthen the capabilities of the method of identification, segmentation and development of interesting groups. This technique introduces additional features into the data intended for the user to improve the analytical research process.

References

1. Manoylo, A.V., Petrenko, A.I.,Frolov,D.B.: Gosudarstvennaya informatsionnaya politika v usloviyakhinformatsionno-psikhologicheskoy voyny: monografiya. M., Goryachaya liniya, Telekom, 2003, 541 s (In Russian)
2. Safina, A.R.: Osobennosti zhanrov Internet SMI.In: Izvestiya Samarskogo nauchnogo tsentra Rossiyskoy akademii nauk,2013, №2-1, tom15, S. 226-229 (In Russian)
3. Shevchenko, T.: Pravoviy status Ínternet -ZMÍ v Ukraĭní : problemi, perspektivi vregulyuvannya (In Ukrainian) http://www.yur-gazeta.com/ru/oarticle/1120/
4. Domarev, V.V.: Zashchita informatsii i bezopasnost' komp'yuternykh sistem.K., DiaSoft, 1999480 s (In Ukrainian)
5. Kaparini, M.: Mass-media, sektor bezopasnosti i vlast'. Rol' novostnykh sredstv massovoy informatsii v kontrole i podotchetnosti sektora bezopasnosti: Nauchnoye posobiye. K., 2005, 280 s. (In Russian)
6. Luman, N.: Real'nost' massmedia. M., Praksin, 2005, 256 s (In Russian)
7. Paniotto, V.: Ukraina. Yevromaydan. Vestnik obshchestvennogo mneniya. № 3–4, 2013, S. 17–24(In Russian)
8. Berezko, O.L., Peleshchishin, A.M.: WWW yak sotsíal'na merezha.In:Proc. oftheSecondIntern. Conf. onComputerScienceandEngineering (CSE'2007),Lviv, 2007, P. 29-30 (In Ukrainian)
9. Azarov, A.A., Brodovskaya, Ye.V., Dmitriyeva, O.V., Dombrovskaya, A.YU., Fil'chenkov,A.A.:Strategii formirovaniya ustanovok protestnogo povedeniya v seti Internet: Opyt primeneniya kibermetricheskogo analiza (na primere Yevromaydana,2013 (In Russian)
10. Mason,W.A., Conrey,F.R., Smith,E.R.: Situating social influence processes:Dynamic, multi directional flow so fin fluence within social networks. Personality and Social Psychology Review, Vol. 11, P. 279 – 300(2007) (In English)
11. Killian Thiel et al, "Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining" KNIME 2012. (In English)

https://www.knime.org/files/knime_social_media_white_paper.pdf
12. M. Hofmann , A. Chisholm Text mining and visualization: case studies using open-source tools/ CRC Press, 2016. — 336 p. — (data mining and knowledge discovery). — ISBN: 9781482237580, 148223758X (In English)
13. T. Wilson, J. Wiebe, and P. Ho_mann. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing, pages 347{354. Association for Computational Linguistics, 2005. (In English)

Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security)

Olena Orobinska[1], Jean-Hugues Chauchat[2], Natalya Sharonova[3]

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

olena.orobinska@univ-lyon2.fr[1],
jean-hugues.chauchat@univ-lyon2.fr[2], sharonova@kpi.kharkov.ua[3]

Abstract. We propose a hybrid, semi-automatic approach that uses the intersection of semantic classes of nouns and verbs built on the domain lexicon and builds kernel ontology from a list of initial concepts and then completes this kernel ontology by new entities detected in a large corpus of texts of international standards of Radiological Safety. The results confirm the important role of initial linguistic modeling and show that the external lexical resources available online can contribute effectively to the resolution of the problem of lexical disambiguation.

Keywords: ontology learning, text processing, semantic analysis, terms extraction

1 Introduction

In the community of ontology researchers and computer scientists, is recognized that the construction of an ontology involves some steps. It's common to start with the installation of a kernel ontology which includes either the simple enumeration of the denominations of the concepts or, in addition, a hierarchy of these concepts. The kernel ontology is used to extract new candidates.

We propose to include the extraction of semantic relations to the first step. In other words, we propose to give the same importance to the concepts and the relations that joins them. Thus, we start with the anticipated conceptualization of the modeling domain and, at the same time, the anticipated linguistic modeling of the input corpus and introduce the notion of predicative framework.

2 Semantic and Linguistic Modeling of Kernel Ontology

The specific terminology of a certain domain is univocal. Thus the denominations of the concerned phenomena, physical quantities, units of measurement are strictly

defined and listed in the specialized glossaries. It does not vary much in technical and scientific texts. On the other hand, a concept always presents a class of objects possessing similar properties. There are two ways of defining a concept: either by its intention, i.e. the explicit definition restricting its properties, or by its extension, i.e. by the enumeration of objects that posses its characteristic properties. We have chosen the representation of concepts by their extensions.

Hence, the following definition of the kernel ontology.

Definition. The kernel ontology is the combination of the list of semantic classes of names; each class corresponds to the extension of a concept, and the predicative framework modeling their semantic relations.

For the construction of an ontology we propose the following operating mode:

1. In consultation with the experts, de ne a limited list of general terms and categories of semantic relations between these terms.
2. Taking each term as a reference for a concept, grouping around them its synonyms to constitute the semantic classes representing the concepts through their semantic extensions.
3. Form the predicative framework in the form of the set of lexical-semantic classes of verbs.
4. Apply the predicative framework for the extraction of new candidates-terms to populate the ontology.

Note that the order of items 2 and 3 is exchangeable.

2.1 Initial List of Concepts

The selection of the initial concepts was carried out in 4 steps.

1. At the beginning, we extracted from the two corpuses, French and Russian, the 100 “best” candidates-terms according to the TF-IDF index.
2. To select the general concepts of the domain, we used the RISK framework, which summarizes the situations related to risks of any kind.
3. The final validation by the expert allowed to retain a list of ten words, these becoming the initial denominations of the concepts. This list includes the following terms (in French and Russian): *damage*, *exposure*, *control*, *personnel*, *population*, *protection*, *radiation*, *risk*, *safety* and *source*.

We perform a first grammatical analysis to recover in the corpus the pairs of the type (*w*, *v*), where *w* is the name and *v* is an “characteristic” verb in the same sentence. In the complete list of all noun-verb pairs, we keep those that contain predefined terms or their synonyms suggested by the dictionary. A module in Java has been written for this step.

The evaluation of the synonyms of the initial terms was carried out according to the FCA method: two names are considered to be true synonyms if they are associated with the same characteristic verbs. In order to select the characteristic verbs which form the formal context of each concept, we proposed to measure the degree of association between each general term and each of the verbs associated with it in the corpus with the coefficient **K** (1) that is the product of the Mutual Information (MI) and the Jaccard Coefficient.

$$K = MI(c_i, v_j) \cdot JaccardCoefficient(c_i, v_j) \tag{9}$$

where

$$MI(c_i, v_j) = |W| \cdot \log TF(c_i, v_j) / (TF(c_i) \cdot TF(v_j))$$

and

$$JaccardCoefficient(c_i, v_j) = TF(c_i, v_j) / (TF(c_i, \bar{v}_j) + TF(\bar{c}_i, v_j))$$

2.2 Predicative Framework

Relationships contribute to the construction of an ontology in the same way as concepts.

Definition. By predicative framework we mean the set of lexical indices which explain the relations between the concepts and make it possible to detect them in the corpus.

We focus on verbs as they are the main predicative agents: each semantic relation category corresponds to a certain predicate and each predicate can be realized using several verbs which in this case form a semantic class.

The diversity of the grammatical and lexical means of a language to express the relations between the objects of the real world complicates their emphasis in the texts. One of the most explicit ways of doing this is using verbs. In this method, we use a superficial analysis of sentences to extract the subject-verb-object (SVO) triplets, subject and object being represented by terms designating the concepts. As a rule, the subject is expressed by a nominal group to the left of the verb, while the object is a nominal group to the right of the verb. In the case of a passive construction, these places are reversed. The use of lemmas makes it possible to reduce the sensitivity of the method to this inversion.

In the first step, as for the names in the previous method, we retrieve in the corpus the potential synonyms of the verbs, selected using the CRISCO Dictionary of Synonyms. But this operation is not sufficient to constitute the semantic classes because most verbs are polysemic and because the dictionary does not explicitly distinguish the different types of semantic similarity, notably the hierarchy (or subsumption) and equivalence, which are realized by different predicates and have different properties in logical theory.

The justification for choosing a good criterion to evaluate the semantic similarity of two words is non-trivial [1]. In order to quantify and measure the degree of synonymy between verbs, we tested the Cosinus measure (2).

$$simCos = \frac{V_i^C \cap V_j^C}{\sqrt{|V_i| \times |V_j|}} \tag{10}$$

Here $V_i^C \cap V_j^C$ is the number of co-occurrences of verb v_i and v_j with the same concept; and $|V_i|$ and $|V_j|$ are the co-occurrences of these verbs with the other names

of the corpus.

2.3 Terminology Pattern Method

The working hypothesis of this method is that the domain lexicon can be detected in the specialized corpus using linguistic analysis. By having a list of generic terms and by empirically discovering the frequent syntactic structures in which these terms appear, we can extend the kernel ontology by new terms, forming the taxonomy, [2]. For example, the term dose is part of the lexicon of the Radiation Security domain. Varied terms, such as *effective dose*, *effective collective dose*, etc. are formed around it.

According to [3], terms are formed by hierarchical syntactic structures. And to enrich the kernel ontology, it's possible to use terminological patterns, which we de ne as the morpho-syntactic structure with one of the generic terms at the head of each. Our goal is to establish these patterns. Terminological pat-terns are formed in two ways: from the analysis of the frequencies of syntactic structures in the corpus; then from the syntactic analysis of the terms of the domain glossary. The fragments of sentences that correspond to the patterns are extracted automatically from the corpus and then validated by the expert. By construction, all extracted fragments contain generic terms that form the kernel ontology: one of the generic terms is the radical of each new term. After validation, terms derived from the same root form a partial taxonomy. They are added in ontology as corresponding concepts.

Initially the patterns are N-grams of grammatical tags that have replaced the words in the corpus. We use N-grams varying from 2 to 6 and extract from the corpus all the fragments of sentences corresponding to these N-grams. The selection of potentially relevant patterns was made from the initial list of generic terms.

3 Conclusion

During our work we have proposed and implemented a coherent algorithm for the construction of ontology in the domain of Radiation Security. These include the formation of semantic classes representing concepts and their relationships, the learning of morpho-syntactic patterns and the installation of partial taxonomies of terms.

All methods are integrated, starting from a limited list of general terms, previously defined with the domain expert. The implementation of this approach required the installation of two corpuses specialized in the domain of Radiation Security, in French and Russian, with 1,500,000 and 600,000 lexical units respectively. A broad synthesis on the state of the art preceded the experimental stage. It covers the various aspects of ontology learning: the theoretical foundations of knowledge representation, natural language modeling, the extraction of terms and relations, the conceptualization phase and the panorama of available tools.

The results have been published in 13 national and international journals and proceedings, between 2010 and 2016, including IMS-2012, TIA-2013, TOTH-2014, Bionica Intellecta, Herald of the NTU “KhPI”.

References

1. Nokel, M., Loukachevitch, N.: An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri In: Proceedings of 10th International Conference on Terminology and Arti cial Intelligence TIA 2013 pp.69-76. Paris (2013)
2. Orobinska, O., Chauchat J.-H., Sharonova N.: Enrichissement d'une ontologie de domaine par extension des relations taxonomiques a partir de corpus specialis In: Proceedings of the 10th International Conference on Terminology and Arti cial Intelligence TIA 2013, pp.129{137. Paris (2013)
3. Cabre, M.T., Cormier, M.C., Humbley, J.: La terminologie: theorie, methode et applications. Presses de l'Universit d'Ottawa, (1998)

Methods of comparing interval objects in intelligent computer systems

Gennady Shepelev[1] and Nina Khairova[2]

[1]Institute for Systems Studies of Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

gis@isa.ru

[2]National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine

khairova@kpi.kharkov.ua

Abstract. Problems of expert knowledge representation by means of generalized interval estimates approach and using methods of comparing interval alternatives in the framework of intelligent computer systems are considered. The problems are common in economy, engineering and in other domains. Necessity of multi criteria approach to comparing problem that is taking into account both preference criteria and risk ones is shown. It is proposed to use a multi-steps approach to decision-making concerning choice of preferable interval alternatives. It is based on consistent using of different comparing methods: new collective risk estimating techniques, "mean-risk" approach (for interval-probability situations) and Savage method (for full uncertainty situations).

Keywords: interval alternatives, risk estimating techniques, collective risk assessment, "mean-risk approach", generalized interval estimates

1 Introduction

Intelligent systems are a staple of artificial intelligence research. Characteristic features of intelligent computer systems are, among others, availability of the subsystem of knowledge representation and subsystem of problem solving. Problem solving should also include decision making as the process of the best suitable alternative choice out of multiple alternatives set that due to complexity and uncertainty decision-making requires human involvement in such processes. An important role in practice plays problems of comparing and choice of alternatives with numerical quality indicators, which due to uncertainty have interval representations (so-called interval alternatives – IA). To include such objects in the knowledge and model bases of intelligent systems they should be describe by special methods of representation and analysis.

As to representation problem it should be borne in mind that in many cases quite difficult to expressexpert knowledge of an uncertain parameter byusing the only interval estimate. Indeed, a too-wideinterval reduces the value of expert knowledge while atoo-narrow interval often causes errors of forecasting. To overcome this difficulty an approach of generalized interval estimates (GIE) was proposed [1, 2]. The approach provides expert tools for expressing expert knowledge about problem parameters by specifying a set of intervals. The set characterizes the inaccuracy of lengthand location of the interval estimate for a parameter of problem. To represent knowledge about parameters of a problem an expert firstly forms apolyinterval estimate (PIE). For this purpose theexpert specifies several characteristic intervals of the set of interval estimates and then constructs a PIE from these intervals. The simplestexample of PIE construction is the specification ofsome interval as the initial estimate of a parameter and extending it (not necessarily symmetrically) on bothsides of the initial boundaries. It is natural to assume that allintervals between the initial and the extended intervals are included in the set of interval estimates and characterize the expert knowledge of the parameter. Theseintervals are possible realizations (scenarios) of theanalyzed parameter.In this case, the PIE of a parameter D is visuallyrepresented by the curvilinear trapezoid $X = D$, $Y = h$in the plane, which is constructed from the expertestimates for the trapezoid bases. On the axis of ordinates $h \in [0, 1]$ the ordered left boundaries of theintervals from the set are marked. The greatest base ofthe trapezoid (the base interval) corresponds to $h = 0$, and the least base (the miniinterval) corresponds to$h = 1$. Thus, the PIE is determined by the positionsand lengths of the least (upper) interval $[D_{lu}, D_{ru}]$) andthe base (lower) interval $[D_{ld}, D_{rd}]$) in the set of intervals, as well as by the shape of the lateral sides of thetrapezoid. In applications these sides can be assumedto be rectilinear. The h axis can be interpreted not onlyas the axis of marks of the intervals forming the PIE.Its interpretation depends on the problem under consideration. So this axis has an obvious physicalmeaning in problems with dependent variables. Forexample, in the problem of estimating the dependenceof the amount of extracted oil on their price each point estimate for the price on the h axis of thePIE corresponds to an interval estimate of reserves onthe D axis.The expert judgement about the chance that a certain value of the analyzed parameter is realized may be expressed by specifying a density of the joint probability distribution function $f(D, h) = f_1(h)f_2(D|h)$ on the PIE. We refer tothus obtained construction as a generalized interval estimate for the quantity D. In the general case thedensities $f_2(D|h)$ defined on different (with respect to$h$) strips of the PIE may belong to different families ofdistributions.There are some ways of applying the GIEapproach to decision making problems. We will focus here only on one possibility. Specifically, we can obtain anaveraged density $f(D)$ (density of marginal distribution) on the baseandthus go to the model of well-known monointerval case. One can see that $f(D)$on the baseinterval is a probability mixture of distributions onGIE intervalscenarios with mixing function determined by the PIE distribution on the h axis. The most common in practice boundariesof PIE arerectilinear. Note that averaged probability distributions received by this way generalize known distributions. Forexample, if to set uniform distributions on both PIEaxes, averaging yields a generalized uniform distribution whose properties are much richer that those of thestandard uniform distribution.

The GIE approach is at the junction of several scientific domains, such as the

theory of decision making, knowledge engineering, probability theory, andsupport systems for expert decisions. This approachexpands the possibilities of more completely revealingexpert knowledge of initial data, provides a more adequate allowance for uncertainty, and improves thequality of decisions.

Thus speaking about the problem of decision-making in relation to the choice of a preferable object from a set of IA we can confine ourselves to the task of comparing mono interval estimates. Quite often the problems of comparing IA belong to a class of problems of unique (non-repeating) choice. Therefore preference chances of tested alternatives in comparing with others are as a rule unknown.In these circumstances expert/DM may either abandon the use of hypotheses about mentioned chances or involve different approaches for the formalization of the knowledge.Such well-known methods of decision-making under the game with nature, or methods under pure uncertainty, as methods by Wald, Hurwicz, Savage may be used in the first case [3].Such methods of comparingascomparison on value of mathematical expectations of quality indicators of IA, method of stochastic dominance, "mean– risk" method [3] and the method of collective risk estimation [4]become available for the usein the second case.The purpose of the further part of this paperis to compare the various approaches and methods of comparing IA and to identify their place in the decision-making process for the choice of the preferred interval objects.

2 Comparing interval alternatives under interval-probabilistic uncertainty

It is natural to assume that each interval estimate includes all possible, up to the available knowledge, point implementations of studied parameter. But in the future, when the uncertainty is removed, this quantity will receive certain the only numeric value. Let assume also for definiteness that such situations take place when the greater values of the quality indicators are preferable than small values. Note that problem of IA comparing due to its nature cannot be exhaustively solved by purely mathematical methods. Indeed in the general case when compared alternatives have non-zero intersection in principle one cannot with certainty conclude, which alternative in their set will be preferable. Any alternative may be "better" in the future, at the time of "removal" of uncertainty, when the interval estimates are replaced by point (exact) values of quality indicator. So at the time of the comparison can be judged only on the chances that one alternative will be preferable to others. Therefore always there is an irremovable risk that in fact namely another alternative but not tested one would be better.Thus formal methods of comparison cannot guarantee choice of truly the best object in the process of comparing. It means that such problems are problems of the decision-making theory as the decision-making processes have to include preferences of decision makers (DM) and take account of their risk tolerance. Therefore comparison, which is adequate to essence of the problem, should be based on at least two criteria, measure of preference of alternative in relation to others in their group and risk measure. Human involvement in the decision-making process can determine not only the choice of the tools of describing the uncertainty but also, due to previous experience and knowledge of human being, the choice of the comparison methods that lead to the result indicators, which are

familiar to the expert or DMs.

Each of the available methods of comparison allows calculate its measures of alternative preference in relation to the other alternatives in their group and (but not always) their risk measures. For some configurations, i.e. relative locations, of compared alternatives and types of uncertainty describing predictions of different methods are equivalent but for others not. In this regard it should be stated that at present there is no approach transcending all others in quality of recommendations obtained on its basis. Each method has its advantages and disadvantages.Joint using of different methods on various successive stages of the decision-making process is probably the best way to combine the power of formal methods and knowledge of experts and DMs. Therefore decision-making for comparing of IA is both science and art.

Let assume that all alternatives are comparable on preference (system of alternatives is full). If disjunction containing comparable interval alternatives is not rigorous then choice of preferred objects depends on the chances of preferences of the disjunction members. Similar relations of disjunction members are based on the degree of assurance in the truth of the hypotheses about preference an alternative from their set, which is tested by DM. Such relations may be called by relations including the risk. Assume that from all possibilities of uncertainty description the tools of distribution functions, similar to tools that used in the probability theory, is selected here for the quantification of preference chances for compared interval alternatives or subsets of values contained in them. This apparatus is the most familiar, in our experience, expert practitioners. It is important because expert analysis of practical problems is most productive when it is conducted in the usual for domain experts' language with using terminology understandable to them.

Let us first briefly describe methods for interval-probability comparing. In the method of collective risk estimating [4], when direct calculations of preference chances of alternatives in comparison with others in their group is produced, compared objects are viewed as interconnected community. Because tool of distribution functions was selected for quantification of preference chances and associated risks, the problem of comparing can be analyzed in the framework of probability logic approach [5]. In accordance with this approach in addition to the truth or falsity of logical statements intermediate logical values are possible. They are interpreted as chances of truth. The use of this approach to IA comparing allows calculating both the chances of alternative preferences and associated risks. Risk of choice of an alternative in their group as the preferred depends on the relative position of alternatives (configuration of alternatives) and on the number of compared objects. An interaction of compared objects leads to a "collective effect", which consists in the fact that the properties of objects of interacting components of the system are significantly different from those of relatively independent objects. Therefore risk of making the wrong decision during choosing a preferred object increases with the growing number of compared alternatives. The matter looks at a rigorous language as follows.Suppose that there are KIAI_i, $i = 1, 2,\ldots, K$ with the same interval quality indicators and dimensionless quantity $C(I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K))$ is the chances in the truth of a testable hypothesis of preference that the alternative I_i is more preferable than all at once alternatives $(I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K)$ from initially given their set (I_i is "better" of others "as a whole"), $\succ$ is symbol of preference. The term "all at once" means here that

$I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K) \equiv (I_i \succ I_1) \wedge (I_i \succ I_2) \wedge (I_i \succ I_3) \wedge \ldots \wedge (I_i \succ I_{i+1}) \wedge \ldots \wedge (I_i \succ I_K)$,

where $\equiv$ and $\wedge$ are symbols of equivalence and conjunction respectively.Risk that I_i would not preferred in reality will be measured by means of dimensionless quantity $R_s(I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K))$ complementing previous chances to unity so that

$$Rs(I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K)) = 1 - C(I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K)).$$

As can be seen $R_s(I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K))$ is degree of assurance in the truth of a hypothesis, which is opposite to the testable hypothesis of preference. Equivalently $R_s(I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K)) = C(\neg(I_i \succ (I_1, I_2,.., I_{i-1}, I_{i+1},\ldots, I_K)))$, where $\neg$ is symbol of negation. One may show that the following relations hold for chances

$C(I_1 \succ (I_2, I_3,\ldots, I_K)) + C(I_2 \succ (I_1, I_3,\ldots, I_K)) + C(I_3 \succ (I_1, I_2, I_4,\ldots, I_K)) + ... + C(I_K \succ (I_1, I_2,\ldots, I_{K-1})) = 1$

and for risks

$R_s(I_1 \succ (I_2, I_3,\ldots, I_K)) + R_s(I_2 \succ (I_1, I_3,\ldots, I_K)) + R_s(I_3 \succ (I_1, I_2, I_4,\ldots, I_K)) + ... + R_s(I_K \succ (I_1, I_2,\ldots, I_{K-1})) = K - 1.$

The natural desire of DM is reduce the risk when deciding. A possibility to do so is to reduce the number of compared alternatives as the method of collective risk estimating suggests. Therefore before deciding on preferred alternative choice it's useful to conduct a preliminary analysis of their initial set to reduce the calculated risk. Some recipes for this are given some later. By reducing the number of intervals in their initial set one may increase the calculated preference chances of analyzed alternative and decrease risks.Can other methods of comparing help to reduce the dimension of the problem?

As to mathematical expectation of quality indicator as random variable it should note that this criterion is adequate for problems of repetitive choice. At the same time the problems under uncertainty deal mainly with situations of unique choice. This requires, in general, rejection of average estimates, or, if they are used, the mandatory accounting as no less important criterion estimations of risk calculated on the basis of certain indicators. Method "mean – risk" does so. Compared alternatives are considered here as isolated, not "interactive" objects. Value of preference criterion is calculatedfor each alternative separately and then a risk indicator is computed again separately for each object. The problem of comparing is solved then as a two-criterion task. Mathematical expectation value is here the criterion of preference. Such indicators as variance, left and right semivariances, left and right mean semideviations and the others act as risk criteria [6]. Let note that the calculated values of the comparison criteria for these methods do not depend on the number of analyzed alternatives. Method "mean – risk" can be used to reduce number of compared IA and to diminish risk of false decision about preferences [4].

In the methods of stochastic dominance pairwise comparison of alternatives carried out only on the basis of the behavior of the distribution functions defined on

intervals-carriers, without taking into account the numerical characteristics of the firsts. If to define interval objects $I_i = [L_i, R_i]$ by the left L_i and right R_i boundaries, $L_i<R_i$, then there are, up to permutations, the four different configurations of compared alternatives pairs: coinciding intervals; intervals without intersection; configurations of right shift and embedded intervals. As to comparing problem the second configuration has no interest. One can show thatalready for a pair of the compared intervals this comparison method does not allow to reduce the dimensionality of the problem.They say that the IAI_1, where integral distribution F_1 of the random variable X_1 is given, dominates (by probability) IA I_2, where distribution F_2 of the random variable X_2 is given, if for a set of possible point implementations $I_1 \cup I_2$ for any point implementation x chances $F_1(X_1<x)$ are not more than chances $F_2(X_2<x)$, and at least for one point implementation they are smaller. In other words the graph of the distribution function F_1 for alternative I_1 lies always below the graph of the distribution function F_2, possibly coinciding with the first in some parts. In the case of right shift configurations, when $L_2<L_1<R_2<R_1$, for uniform distributions alternative I_1 dominates alternative I_2 by probability. Indeed if by definition $\Delta I_i = R_i - L_i$, $i = 1, 2$ then distribution functions F_i are intersected at a point $I_{int} = (L_2\Delta I_1 - L_1\Delta I_2)/(L_2\Delta I_1 - L_1\Delta I_2)$ besides the case $\Delta I_1 = \Delta I_2$ when they are parallel. One can verify that the inequality $L_1<I_{int}<R_2$ is not met for the right shift configurations, and therefore in area $[L_2, R_1]$ $F_1 \le F_2$. Therefore the first alternative dominates the second by probability. This conclusion is valid for any scope of the uncertainty zone $[L_1, R_2]$ for point implementations of IA that is a significant disadvantage of the method. May DM therefore always select as the preferred first alternative due to the dominance of the second by probability? It seems that not because the adoption of this requirement means the neglect of the risk of making a wrong decision on the preference. DM can but should not make such a choice. Thus using of the dominance by probability principle to eliminate certain alternatives from their set for decreasing their number to reduce collective risk is problematic.

Thus in the framework of methods for interval-probabilistic uncertainty may recommend using of the method of collective risk estimating to evaluate integral risks for each alternative in the group and find a subset of the “best” alternatives as the alternatives with the highest chances at pairwise comparisons in the set of compared alternatives. Then this narrowed set of alternatives may be evaluated according to the criteria of preference and risk, which are based on methodof “mean – risk” approach.

3 Comparing interval alternatives under pure uncertainty

Expert does not attempt to specify the distributions of chanceson interval-carriers under pure uncertainty. Instead values of the quality indicators are forecasted usually for small number of possible different states of nature.These values are in a certain interval of values. By this, however, is limited the similarity of comparing methods under pure uncertainty and other methods based on interval representation parameters of the problem when, relatively speaking, the number of the states of nature, which are taken into account, is infinitely large. For the first case the uncertainty is given by means of indication of the nature states, which are essential for the expert, but values of the quality indicators for the statesare calculated as deterministic.This leads to a

certain coarsening of the real problem and to refusal of the possibility for quantitative estimating risk that does not allow for an adequate analysis of the problem.At the same time these methods may be useful in some cases as a means of express-analysis of the problem. The results of such analysis have usually simple interpretation that is essential for DM- practitioners.

Let us restrict ourselves to the three states of nature: unfavorable states, which correspond to the left borders L_i of IA quality indicators; favorable, which correspond to the right boundary $R_i$of quality indicators; and neutral ones, which correspond to some internal points$In_i<R_i$of IA quality indicators.If an expert has decided to use Wald method for IA comparing, then$I_i \succ I_j$if$L_i>L_j$.The advantage of the Wald method consists in rapidity and visibility of the results. The disadvantages are rooted in the incomplete using of information about IA, information about the problem concerning only unfavorable state of nature, as well as in the actual refusal of the risk estimating during decision-making.But the risk, which isdepend on the configuration of compared IA, may be quite large. So IA $I_1$is chosen as the preferred on the Wald method in the configuration of the right shift.However entering the point implementations in the interval $[L_1, R_2]$ (area of I_1 and I_2intersection) does not guarantee that I_1 will be the best. Here the greater the length of the mentioned interval the higher the risk of error in the choice of the preferred object.In the case of a pair of embedded intervals Wald method selects I_2: $I_2 \succ I_1$. If the left boundary of I_2 is positive and DM avoids the risk one can agreewith this conclusion. But if the length ofinterval $[R_2, R_1]$ is rather big and additionally to take into account that part of the point implementations in $[L_2, R_2]$ favored to choice of I_1 then, even with a small propensity of DM to risk, he may prefer I_1.Thus the use of the Wald method for comparing IA requires additional analysis and taking into account the specifics of mutual location of interval estimates.The absence in the framework of the method of a risk indicator, which is required by content of the problem, is a disadvantagethat reduces the possibility of applying the method.

By resorting to Hurwicz method, expert uses two estimates that delimit values of IA quality indicators and correspond to unfavorable and favorable states of nature.However Hurwicz approach is not limited by considerationof quality indicators only for these boundary states of nature.In fact the method takes into account all possible states corresponding to the values of quality indicators within the interval estimation.By this Hurwicz approach differs from all other methods of pure uncertainty. According to Hurwicz interval indicator $I = [L, R]$ is replaced by a point indicator $T(\lambda)$, which is equivalent to the initial interval estimate on expert opinionwhen IAare compared.The value of T is determined by expert choice of the parameter λ, which reflects the expert knowledge and referred to as Hurwicz "pessimism – optimism"factor. Then for $0 <\lambda< 1$

$$T(\lambda) = (1 – \lambda)L + \lambda R.$$

In a situation where the larger value of the quality indicator corresponds to a more preferred state $\lambda = 1$ corresponds to unrestrained optimism of DM and $\lambda = 0$ to pessimism. These limit values of λ should be reversed to the opposite situation. It is believed that under comparing of IA preference should give to the alternative with the best (highest or lowest) value of $T(\lambda)$. In the Hurwicz method also no place for

estimating risk of making a wrong decision about preference of IA. Here there is also a disadvantage associated with the complexity of justifying the value of λ in concrete problems of IAcomparing.Let us take attention in this connection on the fact that using recommended sometimes values of the "pessimism – optimism" factor, for example, λ = 1/3, in some cases insufficiently productive.So $I_1 \succ I_2$for all permissible identical for the compared IA values of $\lambda \in [0, 1]$in the configurations of the right shift ($L_2<L_1<R_2<R_1$).Indeed since

$$T_1 - T_2 = \lambda(R_1 - R_2) + (1 - \lambda)(L_2 - L_1),$$

then $T_1>T_2$ for these configurations. Therefore some experts have to show their knowledge use various different values of λfor differentcompared IA. This opens the way for arbitrariness in the choice of the preferred IA. These remarks apply also to other configurations in the case of application of Hurwicz method for comparing IA.

Let expert decided to bring all available information about IA that meets all three states of nature and to use Savage method.In accordance with this method we have for pair of IA in configuration of right shift: for the unfavorable state of nature $MAX_{UF}(I_1, I_2) = L_1$; forthe neutral state of nature $MAX_N(I_1, I_2) = MAX(In_1, In_2)$;for the favorable state of nature $MAX_F(I_1, I_2) = R_1$. We havethen the following regret matrix (Table 1) for configuration of right shift (RS):

Table 1. Regret matrix(for RS)

	UF	N	F
I_1	0	$\mathrm{MAX}(In_1, In_2) - In_1$	0
I_2	$L_1 - L_2$	$\mathrm{MAX}(In_1, In_2) - In_2$	$R_1 - R_2$

One can see now that $I_1 \succ I_2$ in accordance with the Savage criterion for $In_1 \geq In_2$(natural case ofuniform changing the value of the quality indicator with changing states of nature). We receive for$In_1<In_2$ (such condition can be set by an expert):

$$I_1 \succ I_2, \text{ if } In_2 - In_1 < MAX(L_1 - L_2, R_1 - R_2)$$

$$I_2 \succ I_1, \text{ if } In_2 - In_1 > MAX(L_1 - L_2, R_1 - R_2).$$

Similarly, in the configuration of embedded intervals we have: forthe unfavorable state of nature $MAX_{UF}(I_1, I_2) = L_2$; for the neutral state of nature $MAX_N(I_1, I_2) = MAX(In_1, In_2)$;for the favorable state of nature $MAX_F(I_1, I_2) = R_1$. We havethen the following regret matrix(Table 2) for configuration of embedded intervals (EI):

Table 2. Regret matrix(for EI)

	UF	N	F
I_1	$L_2 - L_1$	$\mathrm{MAX}(In_1, In_2) - In_1$	0
I_2	0	$\mathrm{MAX}(In_1, In_2) - In_2$	$R_1 - R_2$

Hence, if$In_1 \geq In_2$then

$$I_1 \succ I_2, \text{ if} L_2 - L_1 < MAX(In_1 - In_2, R_1 - R_2),$$

$$I_2 \succ I_1, \text{ if} L_2 - L_1 > MAX(In_1 - In_2, R_1 - R_2).$$

Similarly, for $In_1<In_2$ one can receive:

$$I_1 \succ I_2, \text{ if} R_1 - R_2 < MAX(L_2 - L_1, In_2 - In_1),$$

$$I_2 \succ I_1, \text{ if} R_1 - R_2 > MAX(L_2 - L_1, In_2 - In_1).$$

For coincide intervals Wald method leads to the conclusion that compared objects are equivalenton preference. The results for Savage method depend on the position of the point estimates corresponding to the neutral state of nature. Namely,

$$I_1 \succ I_2, \text{ if } In_1 \geq In_2; I_2 \succ I_1, \text{ if } In_1 < In_2.$$

Thus one can see that Savage method is the most suitable for comparing IA under pure uncertainty.The method permits to use the knowledge of expertsbetter than by other methods of this class. It takes into account the values of quality indicators for the many states of nature, as well as DM preferences in predicting values of quality indicators at interior points of the interval estimates.

4 Conclusion

Formal methods of comparison of interval alternatives as components of intellectual computer systems for information-analytical support of the decision-making process cannot guarantee choice of truly the best object as the result of comparing procedure. The results using of such methods can serve for DM only as a guideline, kind of a hint in the decision-making. At present there is no approach transcending all others in quality of recommendations obtained on its basis. Each of the available methods has its advantages and disadvantages. Each of method allows calculating its measures of alternative preference in relation to the other alternatives in their set and some as well as their risk measures.Presence of the collective effect in groups of compared alternatives is manifested primarily in reducing value of preference chances for each alternative with respect to its chances under pair-wise comparison. This leads to a quantitative increasing risk value of selection as preferred alternative such one, which may not actually be per se later, at the time of removal of uncertainty.However seeing that the perception of risk is individual and can vary from one DM to another the risk value resulting from the use of formal methods is nothing more than a calculated risk, which can serve only as an estimate for the DMs. To reduce the calculated risk shouldtry to reduce the number of comparable alternatives, to take into account a possibility of joint using of different comparing methods as well as to combine the power of formal methods and knowledge of experts and DMs.In this regard the following procedure can be recommended. Firstly, to use the method of collective risk estimating to evaluate integral risks for each alternative in the group and find the “best” alternatives as the alternatives with the highest chances at pairwise comparisons in the set of compared alternatives. Then this narrowing set of alternatives should be evaluated according to the criteria of preference and risk, which are based on methods of “mean – risk” approach. For situation of pure uncertainty Savage method may be recommended as tool of comparing because the method permits more fully to use the knowledge of expertsthan in the framework of other methods of this class.

References

1. Sternin, M., Shepelev, G. Generalized interval expert estimates in decision making. Doklady Mathematics, vol. 81 issue 3, pp. 1–2 (2010)
2. Khairova, N., Shepelev, G., Sternin, M. Comparing interval alternatives in decision-making. In: Proceedings of International Conference MCS-2012, pp.158. Simferopol, Ukraine. (2012)
3. Kochenderfer, M. J. Decision Making Under Uncertainty. MIT Press. (2015).
4. Shepelev, G. Decision-making in groups of interval alternatives. International journal "Information theories and applications", vol. 23, issue 4, pp. 303 – 320 (2016)
5. Nilsson, N. J. Probabilistic logic. Artificial Intelligence, vol. 28, issue 1, pp. 71-87 (1986)
6. Ogryczak, W., Ruszczyński, A. From stochastic dominance to mean-risk models: semideviations as risk measures. European journal of operational research, vol. 116, pp. 33 – 50 (1999)

Evaluation ofa Formalized Model for Classification of Emergency Situations

Vera Titova and IelizavetaGnatchuk

Khmelnitsky National University, Khmelnitsky, Ukraine

sobaka2032@rambler.ru

Abstract. Formalization of conditions that characterize the problem of classification of emergency situations is considered in this paper.This formalization is the basis for the Formalized Model of the emergency situations classificationproblem. Intelligent methods are used to solve this problem. These methods are also the basis for the development of the Neural Network Model for emergency situation classification. In this paper wedevelop the structure of the model and determine the number of network layers, the types of neurons and its membership functions. Using the Neural Network Model as decision support for the dispatchers of emergency services makes it possible to improve the quality of emergency situations classification.

Keywords: formalization of conditions, fuzzy neural networks, emergency situations, problem of classification of emergency situations, Formalized Model, Neural Network Model.

1 Introduction

The dispatchers of emergency services have often to deal with emergency situations in their everyday professional activity and make decisions to eliminate such situations.

Nowadays, there are many geoinformation systems for emergency services that help dispatchers in their work [1, 2].However, most of these systems just visualize the information about the situation; displaythe locations of forces and resources at dispatcher's disposal; simulate the situation. The decision to eliminate the situation is entirely madeby the dispatcher, who makes it based on visual informationavailable.

Events that characterize any emergency situation can changeduring the time neededfor the dispatcher to make a decision. Moreover, various situations differ by the type of event and the set of possible decisions.

The more possible decisions and events characterize the situation, the more resources and time will be neededfor the dispatcher for making the final decision to eliminate this situation.

Based on these facts, we can conclude that the dispatcher of emergency services

is the decision maker. And in his or her professional activity he or she needs a decision support, which targets at situation elimination rather than at information visualization. Therefore, providing of such decision support for the dispatcher of emergency services is an urgent scientific problem.

2 Problem formulation

A primary problem that the dispatcher should solve to make a decision is the problemof emergency situation classification.

Emergency situation is violation of normal living conditions and human activities on the similar or territory caused by accidents, natural disasters, epidemic, epizootic or epiphytotic processes, large fire, using the weapons of mass destruction that resulted or can result in human casualtiesand material losses. Any emergency situation is characterized by the lethal casualties, asignificant deterioration of living conditions, the significant deterioration of people's health, economic losses [3].

The decision to eliminate any emergency situation includes the forces and resources necessary to minimize the human casualties, material losses and restore the normal living conditions.

To make a decision the dispatcher has to analyze the available information and assign the situation to one of the known classes. The quality of the decision depends on its correct classification.

Emergency situations that can occur on the territory of Ukraine are classified by the nature of occurrence and by the level of possible consequences [3].

By the nature of occurrencean emergency situation can be [3]:

- technogenic;
- natural;
- social and socio-political;
- military.

By the level of possible consequences an emergency situation can be [3]:

- national;
- regional;
- local;
- building or property.

Any situation is characterized by two parameters: the place of occurrence and conditions. The conditions of the situation are theevents that are inherent to this situation. The place of the situation is the geographical location on the country map.

Events directly influencethe situation class as they determine its nature. Geographical location does not influence the situation classdirectly, but there are several parameters that characterize the situation place and can have influence on the situation class. The first parameter is the situation area.It determines which area is covered by the situation at the time of receivingthe information about it.

The second parameter is the danger of the situation place. It means the presence of buildings that can worsen the situation. For example, a nuclear power plant, a military arsenal, etc.

The thirdparameter is the number of peopleat the situation place.The more people are at the situation place, the more can be injured or die.

Another parameter that influences the nature of the situation and its possible consequences is the time elapsed from beginning of the situation to receiving the information about it.The later information of the situation isreceived, the more dangerous consequences the situation will have and the more time and resources will be needed to fix it.

The problem of classifying emergency situations is characterized by the following attributes:

- a human is the source of information about it, so that the input data can be inaccurate, incorrect, contradictory or subjective;
- there is athe large number of hidden relationships between the input and output data;
- the input data can be changed during solvingthe problem;
- the solution of the problem cannot be reduced to mathematical calculations.

Thus, emergency situations classification is a poorly formalized problem [4], which complicates its solving. The formalization of this problemenables to simplify its solution.

3 Formalized Model of emergency situations classification

To develop the formalized model, we introduce the following indexes:

- I_1 – the number of peopleat the situation place. The more people are at the situation place, the higher is the index value;
- I_2 – the danger of the situation place. The larger isthe number of dangerous buildings at the situation place, the higher is the index value;
- I_3 – the situation area. The larger is the area covered by the situation, the higher is the index value;
- E – the set of events that characterize the situation, $E=[e_1, e_2, e_3 .. e_i]$, i – the number of all possible events that can be the cause of an emergency situation. We have identifiedthe following subsets in this set: $E_1=[e^1_1, e^1_2, .. e^1_a]$ – the subset of events that characterize the technogenic emergency situation; $E_2=[e^2_1, e^2_2, .. e^2_b]$ – the subset of events that characterize the natural emergency situation; $E_3=[e^3_1, e^3_2, .. e^3_c]$ – the subset of events that characterize the social and socio-political emergency situation; $E_4=[e^4_1, e^4_2, .. e^4_d]$ – the subset of events that characterize the military emergency situation; $E_5=[e^5_1, e^5_2, .. e^5_f]$ – the subset of events that characterize the emergency situation at the national level; $E_6=[e^6_1, e^6_2, .. e^6_g]$ – the subset of events that characterize the emergency situation at the regional level; $E_7=[e^7_1, e^7_2, .. e^7_h]$ – the subset of events that characterize the emergency situation at the local level; $E_8=[e^8_1, e^8_2, .. e^8_j]$ – the subset of events that characterize the emergency situation at the building or property;
- T – time elapsed from the beginning of the situation to receivingthe information about it;
- S – the set of emergency situations classes; s_1 – atechnogenic emergency situation; s_2 – a natural emergency situation; s_3 – asocial or socio-political emergency situation; s_4 – amilitary emergency situation; s_5 – a emergency situation at the national level; s_6 – a emergency situation at the regional level; s_7 – a emergency situation at the local level; s_8 – a emergency situation at the building or property.

Any situation belongs to one of the classes s_1 .. s_4 and to one of the classes s_5 .. s_8. For example, the situation that belongs to s_1 and s_7 is atechnogenic emergency situation at the local level.

Taking into account the dependencies between the input parameters and the situation classes, we obtained the following equations:

$$s_1=f_1(e^1_1, e^1_2, .. e^1_a, T) \quad (11)$$

$$s_2=f_2(e^2_1, e^2_2, .. e^2_b, T) \quad (12)$$

$$s_3=f_3(e^3_1, e^3_2, .. e^3_c, T) \quad (13)$$

$$s_4=f_4(e^4_1, e^4_2, .. e^4_d, T) \quad (14)$$

$$s_5=f_5(e^5_1, e^5_2, .. e^5_f, T, I_1, I_2, I_3) \quad (15)$$

$$s_6=f_6(e^6_1, e^6_2, .. e^6_g, T, I_1, I_2, I_3) \quad (16)$$

$$s_7=f_7(e^7_1, e^7_2, .. e^7_h, T, I_1, I_2, I_3) \quad (17)$$

$$s_8=f_8(e^8_1, e^8_2, .. e^8_j, T, I_1, I_2, I_3) \quad (18),$$

where f1..f8 are conversions that have be performed on input data to classify a situation.

The analysis of the subject area displays that five events are enough to classify a situation. Therefore, we introduce a set $E`= [e`_1, e'_2, e'_3, e'_4, e'_5]$, where $e`_1..e'_5$ are the events characterizing the situation that is classified.

Also, the following conclusions are madefrom the analysis of the subject area:

- the situation belongs to the class s_1 if it is characterized by events from the corresponding set E_1. This means the situation belongs to the class s_1 if $E` \subset E_1$, $E` \not\subset E_2$, $E` \not\subset E_3$, $E` \not\subset E_4$;
- the situation belongs to the class s_2 if it is characterized by events from the corresponding set E_2. This means the situation belongs to the class s_2 if $E` \subset E_2$, $E` \not\subset E_1$, $E` \not\subset E_3$, $E` \not\subset E_4$;
- the situation belongs to the class s_3 if it is characterized by events from the corresponding set E_3. This means the situation belongs to the class s_3 if $E` \subset E_3$, $E` \not\subset E_1$, $E` \not\subset E_2$, $E` \not\subset E_4$;
- the situation belongs to the class s_4 if it is characterized by events from the corresponding set E_4. This means the situation belongs to the class s_4 if $E` \subset E_4$, $E` \not\subset E_1$, $E` \not\subset E_2$, $E` \not\subset E_3$;
- the situation belongs to the class s_5 if it is characterized by events from the corresponding set E_5, if the situation place is characterized by the large number of people, the high danger,the large areaand also if a lot of timehas passed since the beginning of the situation. This means the situation belongs to the class s_5 if $E` \subset E_5$, $E` \not\subset E_6$, $E` \not\subset E_7$, $E` \not\subset E_8$, $I_1, I_2, I_3, T \to \max$;
- the situation belongs to the class s_6 if it is characterized by events from the corresponding set E_6, if the situation place is characterized by thesmall number of people, the low danger, the large area and also if a lot of timehas passed since the beginning of the situation. This means the situation belongs to the class s_6 if $E` \subset E_6$, $E` \not\subset E_5$, $E` \not\subset E_7$, $E` \not\subset E_8$, $T, I_3 \to \max, I_1, I_2 \to \min$;

– the situation belongs to the class s_7 if it is characterized by events from the corresponding set E_7, if the situation placeis characterized by thesmall number of people, the low danger, the small area and also if a lot of timehas passed since the beginning of the situation. This means the situation belongs to the class s_7 if $E` \subset E_7, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_8, T \to \max, I_1, I_2, I_3 \to \min$;

– the situation belongs to the class s_8 if it is characterized by events from the corresponding set E_8, if the situation placeis characterized by thesmall number of people, the low danger, the small area and also if alittle of timehas passed since the beginning of the situation. This means the situation belongs to the class s_8 if $E` \subset E_8, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_7, I_1, I_2, I_3, T \to \min$.

As a result we have a system of rules with denoting s – the situation to be classified:

$s=f_1(e^1_1, e^1_2, .. e^1_a, T) + f_5(e^5_1, e^5_2, .. e^5_f, T, I_1, I_2, I_3)$, if $E` \subset E_1, E` \not\subset E_2, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_5, E` \not\subset E_6, E` \not\subset E_7, E` \not\subset E_8, I_1, I_2, I_3, T \to \max$;

$s=f_1(e^1_1, e^1_2, .. e^1_a, T) + f_6(e^6_1, e^6_2, .. e^6_g, T, I_1, I_2, I_3)$, if $E` \subset E_1, E` \not\subset E_2, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_6, E` \not\subset E_5, E` \not\subset E_7, E` \not\subset E_8, T, I_3 \to \max, I_1, I_2 \to \min$;

$s=f_1(e^1_1, e^1_2, .. e^1_a, T) + f_7(e^7_1, e^7_2, .. e^7_f, T, I_1, I_2, I_3)$, if $E` \subset E_1, E` \not\subset E_2, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_7, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_8, T \to \max, I_1, I_2, I_3 \to \min$;

$s=f_1(e^1_1, e^1_2, .. e^1_a, T) + f_8(e^8_1, e^8_2, .. e^8_j, T, I_1, I_2, I_3)$, if $E` \subset E_1, E` \not\subset E_2, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_8, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_7, I_1, I_2, I_3, T \to \min$;

$s=f_2(e^2_1, e^2_2, .. e^2_b, T) + f_5(e^5_1, e^5_2, .. e^5_f, T, I_1, I_2, I_3)$, if $E` \subset E_2, E` \not\subset E_1, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_5, E` \not\subset E_6, E` \not\subset E_7, E` \not\subset E_8, I_1, I_2, I_3, T \to \max$;

$s=f_2(e^2_1, e^2_2, .. e^2_b, T) + f_6(e^6_1, e^6_2, .. e^6_g, T, I_1, I_2, I_3)$, if $E` \subset E_2, E` \not\subset E_1, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_6, E` \not\subset E_5, E` \not\subset E_7, E` \not\subset E_8, T, I_3 \to \max, I_1, I_2 \to \min$;

$s=f_2(e^2_1, e^2_2, .. e^2_b, T) + f_7(e^7_1, e^7_2, .. e^7_f, T, I_1, I_2, I_3)$, if $E` \subset E_2, E` \not\subset E_1, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_7, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_8, T \to \max, I_1, I_2, I_3 \to \min$;

$s=f_2(e^2_1, e^2_2, .. e^2_b, T) + f_8(e^8_1, e^8_2, .. e^8_j, T, I_1, I_2, I_3)$, if $E_2 \neq \{\}, E` \subset E_2, E` \not\subset E_1, E` \not\subset E_3, E` \not\subset E_4, E` \subset E_8, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_7, I_1, I_2, I_3, T \to \min$;

$s=f_3(e^3_1, e^3_2, .. e^3_c, T) + f_5(e^5_1, e^5_2, .. e^5_f, T, I_1, I_2, I_3)$, if $E` \subset E_3, E` \not\subset E_1, E` \not\subset E_2, E` \not\subset E_4, E` \subset E_5, E` \not\subset E_6, E` \not\subset E_7, E` \not\subset E_8, I_1, I_2, I_3, T \to \max$;

$s=f_3(e^3_1, e^3_2, .. e^3_c, T) + f_6(e^6_1, e^6_2, .. e^6_g, T, I_1, I_2, I_3)$, if $E` \subset E_3, E` \not\subset E_1, E` \not\subset E_2, E` \not\subset E_4, E` \subset E_6, E` \not\subset E_5, E` \not\subset E_7, E` \not\subset E_8, T, I_3 \to \max, I_1, I_2 \to \min$;

$s=f_3(e^3_1, e^3_2, .. e^3_c, T) + f_7(e^7_1, e^7_2, .. e^7_f, T, I_1, I_2, I_3)$, if $E` \subset E_3, E` \not\subset E_1, E` \not\subset E_2, E` \not\subset E_4, E` \subset E_7, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_8, T \to \max, I_1, I_2, I_3 \to \min$;

$s=f_3(e^3_1, e^3_2, .. e^3_c, T) + f_8(e^8_1, e^8_2, .. e^8_j, T, I_1, I_2, I_3)$, if $E` \subset E` \subset E_3, E` \not\subset E_1, E` \not\subset E_2, E` \not\subset E_4, E_8, E` \not\subset E_5, E` \not\subset E_6, E` \not\subset E_7, I_1, I_2, I_3, T \to \min$;

$s=f_4(e^4_1, e^4_2, .. e^4_d, T) + f_5(e^5_1, e^5_2, .. e^5_f, T, I_1, I_2, I_3)$, if $E` \subset E_4$, $E` \not\subset E_1$, $E` \not\subset E_2$, $E` \not\subset E_3$, $E` \subset E_5$, $E` \not\subset E_6$, $E` \not\subset E_7$, $E` \not\subset E_8$, $I_1, I_2, I_3, T \to \max$;

$s=f_4(e^4_1, e^4_2, .. e^4_d, T) + f_6(e^6_1, e^6_2, .. e^6_g, T, I_1, I_2, I_3)$, if $E` \subset E_4$, $E` \not\subset E_1$, $E` \not\subset E_2$, $E` \not\subset E_3$, $E` \subset E_6$, $E` \not\subset E_5$, $E` \not\subset E_7$, $E` \not\subset E_8$, $T, I_3 \to \max, I_1, I_2 \to \min$;

$s=f_4(e^4_1, e^4_2, .. e^4_d, T) + f_7(e^7_1, e^7_2, .. e^7_f, T, I_1, I_2, I_3)$, if $E` \subset E_4$, $E` \not\subset E_1$, $E` \not\subset E_2$, $E` \not\subset E_3$, $E` \subset E_7$, $E` \not\subset E_5$, $E` \not\subset E_6$, $E` \not\subset E_8$, $T \to \max, I_1, I_2, I_3 \to \min$;

$s=f_4(e^4_1, e^4_2, .. e^4_d, T) + f_8(e^8_1, e^8_2, .. e^8_j, T, I_1, I_2, I_3)$, if $E` \subset E_4$, $E` \not\subset E_1$, $E` \not\subset E_2$, $E` \not\subset E_3$, $E` \subset E_8$, $E` \not\subset E_5$, $E` \not\subset E_6$, $E` \not\subset E_7$, $I_1, I_2, I_3, T \to \min$;

This system is our proposed Formalized Model of emergency situation classification. Analyzing the model, we conclude:

- the problem of emergency situations classification is characterized by a large number of possible decisions and a large number of closely related characteristics. Thus, it is difficult to solve by the exhaustive search of all decisions available;
- this problem cannot be solved by algebraic methods because its solution cannot be reduced to numerical calculations;
- with some input data, the situation does not correspond to any rule completely. It can belongs to several classes with different membership degrees, which makes the classification fuzzy.

Given the conclusions above, we propose to solve this problem by methods of artificial intelligence, for example, hybrid neural network [5,6].This approach has the following advantages [6]:

- its training is carried out using neural network learning algorithms, which have advantages in processing unreliable data;
- all conclusions are made on the basis of fuzzy logic, in the function and expression of which it is easy to transform the system of rules of the Formalized Model of emergency situation classification;
- hybrid neural networks, as expert systems, enable to add expert information to the learning process and explain the results of solving problems, which enables to trace or alter decision making process.

4 Neural Network Model of emergency situations classification

Given the relations and interactions of the Information Model, itis determined that input data vector consists of 9 elements. So, the Neural Network Model has 9 inputs: I1 – the number of people at the situation place; I2 – the danger of the situation place; I3 – the situation area; T – time elapsed from the beginning of the situation to the receiving the information about it; e1 .. e5 – events that characterize the situation.

I1 and I2 take values in the range [0..3], where 0 – no people/ no danger, 1 – a few people/ low danger, 2 – moderate number of people/ medium danger, 3 – many people/ high danger. I3 takes values in the range [0..10], where 0 – minimal area that can be covered by the situation within the building or property, 10 – maximal area that can be covered by the situation within several regions. T takes values in the range

[0..3], where 0 – the situation began less than 2 hours ago, 1 – the situation began 2-12 hours ago, 2 – the situation began 12-24 hours ago, 3 – the situation began over 24 hours ago. e1..e5 take values in the range [1..100], 100 – the maximum number of possible events. An event number is determined by the facts of the event. For example, 22 – fire, 1 – earthquake, 25 – building collapse, etc.

The number of model outputs is two. The first output determines the emergency situation class by nature. It takes values in the range [0..1], where the values in the range [0..0.35] correspond to naturalemergency situations, the values in the range [0.2..0.55] representtechnogenicemergency situations, the values in the range [0.4..0.75] describemilitary emergency situations, the values in the range [0.6..1] are attached to social or socio-political emergency situations.

The second output determines the emergency situation class by the level of possible consequences. It takes values in the range [0..1], where the values in the range [0..0.35] represent to emergency situations at the building or property, the values in the range [0.2..0.55] describe emergency situations at the local level, the values in the range [0.4..0.75] are attached to emergency situations at the regional level, the values in the range [0.6..1] cover emergency situations at the national level.

The Neural Network Model of emergency situation classification is displayed in Fig. 1. It consists of three layers of neurons. The neurons of the first layer determine the matching degree of input variables to fuzzy membership functions as illustrated in [7, Fig. 4-7]. The neurons of the second layer determine the degree of truth of each of the system rules described above. The neurons of the third layer determine the emergency situation class.

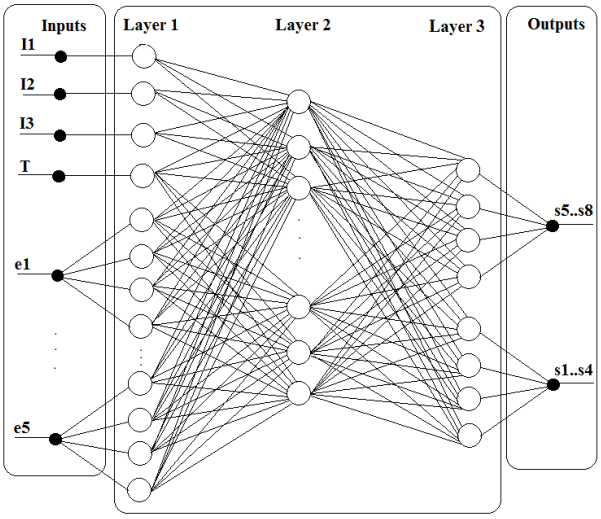

Fig. 1. Neural Network Model of emergency situations classification

5 Evaluating the accuracy and adequacy of the Formalized Model

The Neural Network Model of emergency situations classification isimplemented using the software package Fuzzy Logic Toolbox of Matlab. The implementation of the model is given in [7].

The model is tested on the following set of input values, (Table 1, columns I_1-I_3, T, e_1-e_5), where the number of test sets was 20. The columns R_1, R_2 (Table 1) display the actual results of the model operation, R_1 determines the class by nature, R_2 determines the class by the level of possible consequences. The columns $R`_1$, $R`_2$ (Table 1) display the expected results corresponding to the incoming sets, $R`_1$ determines the class by nature, $R`_2$ determines the class by the level of possible consequences.

Table 1. Results of Neural Network Model application

No.	I_1	I_2	I_3	T	e_1	e_2	e_3	e_4	e_5	R_1	R_2	$R_1`$	$R_2`$
1	0	0	0	3	25	-	-	-	-	0.15	0.55	0.15	0.45
2	1	3	0	1	25	-	-	-	-	0.25	0.55	0.25	0.45
3	2	3	1	0	25	-	-	-	-	0.25	0.15	0.25	0.15
4	3	1	6	0	1	-	-	-	-	0.25	0.55	0.25	0.55
5	3	1	6	1	1	-	-	-	-	0.25	0.55	0.25	0.55
6	3	1	6	2	1	-	-	-	-	0.85	0.55	0.85	0.55
7	3	1	6	3	1	-	-	-	-	0.85	0.75	0.85	0.75
8	3	1	6	0	1	34	35	39	40	0.35	0.55	0.35	0.55
9	3	1	6	1	1	34	35	39	40	0.35	0.55	0.35	0.55
10	3	1	6	2	1	34	35	39	40	0.85	0.85	0.85	0.75
11	3	1	6	3	1	34	35	39	40	0.85	0.85	0.85	0.75
12	1	3	1	0	36	-	-	-	-	0.35	0.25	0.35	0.25
13	2	3	1	1	36	-	-	-	-	0.35	0.25	0.35	0.35
14	1	3	1	2	36	-	-	-	-	0.35	0.55	0.35	0.55
15	2	3	1	3	36	-	-	-	-	0.85	0.85	0.85	0.75
16	3	2	3	0	35	39	20	-	-	0.35	0.35	0.35	0.35
17	3	2	3	1	35	39	20	-	-	0.35	0.35	0.35	0.35
18	3	2	3	2	35	39	20	-	-	0.85	0.45	0.85	0.45
19	3	2	3	3	35	39	20	-	-	0.85	0.45	0.75	0.45
20	3	0	3	0	35	39	20	-	-	0.35	0.35	0.35	0.35
21	3	0	3	1	35	39	20	-	-	0.35	0.35	0.35	0.35
22	3	0	3	2	35	39	20	-	-	0.35	0.45	0.35	0.45
23	3	0	3	3	35	39	20	-	-	0.85	0.45	0.85	0.45
24	2	2	3	0	75	-	-	-	-	0.85	0.35	0.85	0.35
25	2	2	3	1	75	-	-	-	-	0.85	0.35	0.85	0.35
26	2	2	3	2	75	-	-	-	-	0.85	0.35	0.75	0.45
27	2	2	3	3	75	-	-	-	-	0.55	0.45	0.55	0.45
28	2	0	0	0	25	-	-	-	-	0.15	0.15	0.15	0.15
29	2	1	0	1	25	-	-	-	-	0.25	0.15	0.25	0.15
30	2	2	0	2	25	-	-	-	-	0.25	0.25	0.25	0.25

Input values and expected results are determined as follows:

- **situation №1:** fire in an uninhabited area, the area of firebeing small andthe starting time of the fire being unknown. So, I_1=0, I_2=0, I_3=0, T=3, e_1=25. The

analysis of situation classification statistics enables us to conclude that situation isa natural emergency situation at the local level. $R`_{1} \in [0..0.35]$, $R`_{2} \in [0.2..0.55]$.

- **situation №2:** fire near to a nuclear power plant, area of fire being small andthe starting time of the fire being less than 12 hours. So, $I_1=1$, $I_2=3$, $I_3=0$, $T=1$, $e_1=25$. The analysis of situation classification statistics leads to conclusion that situation can be viewed asatechnogenic emergency situationat the regional level, $R`_{1} \in [0.2..0.55]$, $R`_{2} \in [0.4..0.75]$.
- **situation №3:** an earthquake in a densely inhabited area, the area covering the region, the starting time of the earthquakebeing less than 2 hours. So, $I_1=3$, $I_2=1$, $I_3=6$, $T=0$, $e_1=1$. An earthquake can generate mudslides ($e_i=5$), building collapsing ($e_i=34$), power network accidents ($e_i=35$), gas pipeline accidents ($e_i=40$), water pipeline accidents ($e_i=39$), etc. This can lead to panic, robbery, looting as time increases. Thus, the analysis of situation classification statistics enables us to conclude that situation be classifies as atechnogenic or thesocial emergency situation of the regional level, $R`_{1} \in [0,2..0.55]$ or $R`_{1} \in [0.6..1]$, $R`_{2} \in [0.4..0.75]$.

Actual results falling into a wrong interval are marked with a dash (Table 1). From the analysis of actual results, we conclude that the accuracy of the mathematical model is 94% for the class by nature and 90% for the class by the level of possible consequences. This error appears acceptable [8].

The adequacy of the model iscalculated by the absolute error.It is 0.1 for the class by nature and 0.3 for the class by the level of possible consequences, which is also acceptable [8].

The accuracy and adequacy evaluation performed enables us to conclude that the Formalized Model developed isapplicable to solvingthe problem.

6 Conclusions

The problem of emergency situation classification solved by a dispatcher of emergency services after receiving the information about the beginning of the situation is considered.

We come to conclusion that this problem is poorly formalized, and, therefore it is impractical to apply mathematical methods to solve it.

The Formalized Model of the emergency situation classification is developed on the basis of the analysis of relations and interactions between the parameters characterizing the situation and its possible classes.

The class of any emergency situation depends on the events, the area of the situation location, the danger of the situation location and the time that elapsed from the situation beginning. These parameters were the basis for the Formalized Model represented by a system of rules.

After the analysis of the Formalized Model we propose to use hybrid neural network for solving of the problem of emergency situation classification.The novelty of this approach and its advantages over other methods consist in the fact that such networks not only use standard algorithms for training neural networks, but they can acquire new knowledge and are logically transparent to the user.

Hybrid neural network is implemented in the Matlab environment. The results of its applicationenable us to conclude that developed model is appropriate for the

problem consideration.

The model above can serve as a basis for the development of a decision support system for the dispatchers of emergency services.

References

1. Michael N. DeMers. Fundamentals of geographic information systems. 4th ed. – Hoboken, NJ : Wiley, 2009 – pp. 1-443.
2. V. Seredovych. Geoinformatsionnyie sistemyi (naznachenie, funktsii, klassifikatsiya)/ V. Seredovych, V. Klyushnychenko, N. Tymofeeva. – Novosibirsk: SGGA, 2008. – pp. 1-192. [in Russian]
3. Ya.Bedriy.Bezpekazhyttyediyal'nosti. – L'viv: Vydavnycha firma «Afisha», 1998. – pp. 1-298. [in Ukrainian]
4. V. Lokazyuk. Intelektual'ne diahnostuvannya mikroprotsesornykh prystroyiv ta system/ V. Lokazyuk, O. Pomorova, A. Dominov. – Kyyiv: "Takispravy", 2001. – pp. 1-286. [in Ukrainian]
5. J. Hawkins. On Intelligence /J. Hawkins,S. Blakeslee. – New York, NY: Owl Books, 2005. – pp. 1-262.
6. Kruglov V. Nechetkaya logika i iskusstvennyie neyronnyie seti/ V. Kruglov, M. Dli, R. Golunov. – M.: FIZMATLIT, 2001. – pp. 1-201. [in Russian]
7. V. Titova. Realizatsiya nechitkoyi neyronnoyi merezhi dlya rozpiznavannya nadzvychayny khsytuatsiy u paketi Matlab. / Proceedings of international scientific conference "Theoretical and applied aspects of program systems development – 2014", Kyyiv, 2014. – pp. 231-238. [in Ukrainian]
8. N. Nazarov. Metrologiya. Osnovnyie ponyatiya i matematicheskie modeli. – M.: Vyisshayashkola, 2002. — pp. 1-352. [in Russian]

Discursive units in scientific texts

Verbinenko Yulia

Ukrainian Lingua-Information Fund of NAS of Ukraine, Kyiv, Ukraine

yulia_verbinenko@yahoo.co.uk

Abstract. Discursive units are text elements that ensure its coherence, direct attention to the context, make text clear etc. Undeveloped theory of semantic description and its lexicographical representation complicates the description of the discursive units. There are also difficulties in dictionary definitions formulating, as discursive units are often very integrated into the context. Because of this, it is difficult to define system boundaries and build up the correct classification. The main criterion for merging of heterogeneous units into one class of discourse units is their joint function of regulation and organization of the communication process. It is impossible to classify discursive units only by grammatical (morphological and syntactic) features. In terms of morphology, these units are also difficult to combine into one class. In our opinion, it is functional feature that is the most relevant for determining discursive units in the text. Therefore, semantic-pragmatic characteristics are most relevant for the determination of the discursive units in the text.

Keywords: discursive units, discursive markers, scientific text, formal characteristics of discursive units

Discursive units (DU) are text elements that ensure its coherence, direct attention to the context, make text clear, focus reader's attention on different of context elements. Discursive units provide clarity, structure to the language, regulate emotional coloring, and make text more clear. Their marker functions in the context are varied from statement organizing, shifting from one topic to another to expressing of text macrostructure and autoreflection (individual point of view) etc.

Discursive units "regulate the flow of discourse" [6] and have composition-structural, regulatory and modal-assessment functions. There are no texts that do not contain discursive units, scientific texts are no exception. They accompany author's main communicative intentions. According to M. Kozhina [5], discursive units are "specially created" for the scientific style.

Introductory words, modal words and phrases (безсумніву, власнекажучи), conjunctions (якщо, але), particles (ж, просто, якраз), phrases (в зв'язку з цим, в данному випадку) and even sentences (підсумуємо найбільш важливе, список можна продовжити) could function as discursive units. Clearly, that from the lexical and morphological (and even syntactical) point of view DU are very dissimilar, and

that is why it is almost impossible to classify them by lexical-grammatical parameters, i.e. refer them to a specific part of the language. [2]

There are also difficulties in dictionary definitions formulating, as discursive units are very oftenintegrated into the context. Undeveloped theory of semantic description and its lexicographical representation complicates the description of the discursive units. Because of this, it is difficult to define system boundaries and build up the correct classification.

The main criterion for merging of heterogeneous units into one class of discursive units is their joint function of regulation and organization of the communication process. N. Bogdanova [9] thinks that DUs are units of the functional-pragmatic level, despite the fact that they have different semantics and structure. Note that DUs have no denotative meaning. The fuzziness, semantic complexity of the units of this class usually does not allow to use the traditional lexicographic method of definition decomposition onto semantically deterministic components.

Different DUs could have the same discursive function. For example, you could start your speech with such discursive units as first of all, therefore, and often the choice of a particular unit in such a situation is difficult to motivate formally. A large number of DUs are interchangeable, that's why we think that they are contextually synonymous, but that synonymy, however, cannot be always semantically classified at the level of lexical meanings. Due to such uncertainty of meaning discursive units are difficult to describe linguistically [3].

Nowadays there is no even minimal list of discursive markers that might help to determine discursive units in the text, as well as the complexity of their system signs.

It is impossible to classify discursive units only by grammatical (morphological and syntactic) features. For example, syntactic characteristics are not enough to determine DUs in the text, although DUs have some syntactic features. In terms of morphology, these units are also difficult to combine into one class, because morphologically similar words might be discursive units or not. In our opinion, functional feature are the most relevant fordetermining discursive units in the text.

Therefore, semantic-pragmatic characteristics are most relevant for the determination of the discursive units in the text. For example, a unit that has formal noun characteristic, having DU function in the context, it may lose some of its characteristics and get characteristics of another part of speech. Similar phenomenon is described by V. Ivanov in his book "Linguistics of the third millennium": "It seems especially difficult to select noun (and especially noun group as a part of a sentence, separate from the verbal group) in polysynthetic languages where noun is often appears only truncated morph that is incorporated into the verbal form. Native American, who taught me the irokez language Onondaga, had refused to translate word "tree" into English, saying that morphs with a similar meaning could only be a part of verbal form [4].

E. V. Khachaturyan [7] made an attempt to determine the main formal characteristics of DUs:

- isolated discursive units cannot form an answer to a question;
- they are not used with negation (unless negation is a part of a discursive unit);
- they are usually omitted in indirect speech;
- they cannot be repeated in echo-question;
- unlike parts of sentences, position of a discursive unit (that has no syntactic

function in sentence) is not fixed, but is determined by a semantic criteria;

- usually, discursive unit or the entire construction with it in a speech is distinguished by lexical means (like pauses).

Forexample (examples are from «Ukrainian-Russian-English Dictionary for physicists» by S. M. Yudina):

1. *Более или менее / Більш або менш*

This situation is *more or less* appropriate for a liquid solution.

Цяситуація*більш-менш*придатна для рідкогорозчину.

Эта ситуация более или менее подходит для жидкого раствора.

2. *Брать на себя смелость / Брати на себе сміливість*

We dare suggest that there is no real scientific reason for such situation: instead, it occurs due to excessive conservativeness and inertia of thought.

Беремо на себе сміливість припустити, що для такої ситуації немає реальної наукової причини: навпаки, це відбувається через надмірну консервативність і інерцію думки.

Берем на себя смелость предположить, что для такой ситуации нет реальной научной причины: напротив, это происходит из-за чрезмерной консервативности и инерции мысли.

3. *Якобы / Нібито*

In fact, Minkowski preferred to ignore recent results that *allegedly* refuted the theory of relativity.

Насправді, Мінковський надавав перевагу ігноруванню нещодавніх результатів, які *нібито* спростовують теорію відносності.

На самом деле, Минковский предпочитал игнорировать недавние результаты, которые *якобы* опровергают теорию относительности.

4. *Эквивалентно/или, что эквивалентно/Еквівалентно/або, що еквівалентно*

Using Equations (A4) [or, *equivalently*, Equations (B2)], we obtained the following expressions.

Використовуючи рівняння (A4) [або, *що еквівалентно*, рівняння (B2)], ми отримали наступні вирази.

Используя уравнения (A4) [или, *что эквивалентно*, уравнения (B2)], мы получили следующие выражения.

5. *На самом деле (действительно) / Насправді (дійсно)*

Actually, neither silicon nor germanium crystals have been satisfactory for this application.

На самом деле, ни кристаллы кремния, ни германия не были удовлетворительными для этих целей.

Насправді, ані кристали кремнію, ані германію не були задовільними для цих цілей.

The last example represents special semantic condition – superposition [8], where discursive or non-discursive meaning depends from the position in the sentence. (Actually, neither silicon nor germanium crystals have been satisfactory for this application. – A blackbody does not really exist in nature.)

A. N. Kolmogorov firstly used the concept of semantic condition. V. A. Shyrokov developed the theory of semantic condition, and according to it any word (any language unit) in a context or in a language stream is in some semantic condition. For the units of a lexical level it is a combination of characteristics of

grammatical and lexical semantics, since grammatical and lexical meanings are the two main language aspects [9].

The repertoire of discursive units, the frequency of their usage and formal grammatical structure are associated not only with the structure of a particular language, but also with individual linguistic view of the world. Since language and culture are inseparable, using of foreign languages in isolation from the culture is impossible; the difference between cultures usually has no clear recordings in dictionaries, so researchers point out that cross-language cultural barrier creates additional problems to the lingual communication. It is very important for the modern scientific communication. In addition, it is of great interest to study functioning of discursive units in languages for special purposes.

Stated determines the relevance of this research area and its perspectives in theoretical and applied linguistics.

References

1. Bogdanova N. O. On the draft of a discourse units dictionary Russian language (on corpus material) // Computer Linguistics and Intelligent Technologies: Proceedings of the international conference "Dialogue", Bekasovo, May 30 – June 3, 2012. – Moscow: 2012. – P 71-80.
2. Viktorova, E. Yu. Does gender influence on the usage of discursive units? (Based on the of written scientific discourse) // Izvestia of the Saratov University. – 2011. – No. 3. – P. 8-14
3. Discursive words of the Russian language: the experience of contextual-semantic description / Ed. K. KiselevaandD. Payar. –M: Metatext, 1998. – 447 p.
4. Ivanov V. V. Linguistics of the Third Millennium: Questions for the Future. – M.: Languages of the Slavs. Cultures, 2004. – 208 p.
5. Kozhina M. Scientific style // Stylistic encyclopedic dictionary of the Russian language. – M., 2003. – P. 242-247
6. Sirotinina, O. B. On the syntactic status of some components of discourse// Oamenisiidei: Studiidefilologie. – Cluj-Napoca, 2005. – P. 342-348
7. Khachaturyan, E. V. Semantics and syntactics of discourse words of verbal origin in modern Italian: dis. ... cand. Philol. Sciences: 10.02.05 / E. Yu. Khachaturyan. – M., 2000. – 171 s.
8. Shyrokov V.A. Computer lexicography. Monograph. / V.A. Shyrokov; NAS of Ukraine. Ukrainian Lingua-Information Fund. – K .: Naukova Dumka, 2011. – 351 p
9. ShyrokovV.A. Semantic conditions of language units and the irusagein cognitive lexicography // Movoznavstvo. – 2005. – № 3-4. – P. 47-62

STUDENT SECTION

Statistical Methods Usage of Descriptive Statistics in Corpus Linguistic

Valeriy Didusov and Zoia Kochueva

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

valeradidusiov@gmail.com, kochueva@kochuev.com

The purpose of this study: the development of software to determine the criteria of Pearson's chi-squared on the basis of both scientific and literary texts. In the first part of the study it is conducted the analysis of contemporary corpus linguistics. Statistical studies and their views of corpus linguistics are also described. Held his own description of the body of fiction and non-fiction texts. In the second part of the paper it is focused on statistical methods in corpus linguistics, namely the descriptive statistics. The description of the operation of the developed software.

Research in various linguistic areas has the subject a text or collection of texts and imply at first selection of the material, and then analysis and processing of large amounts of text to identify some language patterns. Traditional methods of linguistic analysis of the text can perform all of these tasks, but their low efficiency makes the usage of methods of computer analysis of the text more frequent. This reduces the work of linguist and considerably increase the amount of processing data and avoid inaccuracies and errors in calculations. A computer text analysis enables the establishment of speech patterns based not on theoretical but empirical data [1].

A statistical hypothesis is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability – the significance level. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by identifying two conceptual types of errors (type 1 & type 2) [2].

The major purpose of hypothesis testing is to choose between two competing hypotheses about the value of a population parameter. For example, one hypothesis might claim that the wages of men and women are equal, while the alternative might claim that men make more than women.

The hypothesis actually to be tested is usually given the symbol H_0, and is commonly referred to as the null hypothesis. As is explained more below, the null hypothesis is assumed to be true unless there is strong evidence to the contrary –

similar to how a person is assumed to be innocent until proven guilty. The other hypothesis, which is assumed to be true when the null hypothesis is false, is referred to as the alternative hypothesis, and is often symbolized by H_A or H_1. Both the null and alternative hypothesis should be stated before any statistical test of significance is conducted. In other words, you technically are not supposed to do the data analysis first and then decide on the hypotheses afterwards [3].

Algorithm:

1. State the null hypothesis and the alternate hypothesis.
2. Select the appropriate test statistic and level of significance.
3. State the decision rules.

The decision rules state the conditions under which the null hypothesis will be accepted or rejected. The critical value for the test-statistic is determined by the level of significance. The critical value is the value that divides the non-reject region from the reject region.

4. Compute the appropriate test statistic and make the decision.

Compare the computed test statistic with critical value. If the computed value is within the rejection region(s), we reject the null hypothesis; otherwise, we do not reject the null hypothesis.

5. Interpret the decision.

Based on the decision in Step 4, we state a conclusion in the context of the original problem [4].

In order to carry out some statistical research and apply statistical methods in linguistics you must have a representative corpus that meets the task. In this study it was used the corpus "Scientific and fiction" in order to study deeper statistical hypotheses testing methods in linguistics. Used corpus is a sample that contains only some of the necessary material in contrast to the general population. Since the usage of whole general population is not possible it was decided to have a representative sample of all elements of the population and objects that appear frequently in the general population, often manifested in it. The purpose of this corpus is to create the most informative sampling on this topic for further statistical analysis.

As the test material it was used corpus containing a collection of scientific and literary texts. Testing statistical hypotheses were carried out in accordance with established Pearson's chi-squared criteria. Knowing such values as number of sentences, nouns, conjunctions help to test statistical hypotheses. The analysis allowed us to determine the null hypothesis. As a result of the program, the user can see the results of treatment. The main indicator is the value of chi-squared criteria.

References

1. Lehmann E. Testing Statistical Hypotheses / E. Lehmann, L. Romano, P. Joseph. – New York: Springer, 2005.
2. Gosall K. Doctor's Guide to Critical Appraisal / K. Gosall, N. Kaur, S. Gurpal. – Knutsford: PasTest, 2012.
3. Schervish M. Theory of Statistics / M. Schervish. – New York: Springer, 1996.
4. Wikipedia: Statistical hypothesis testing: [Electronic source]. – Access mode: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Improving Communication in Enterprise Solutions: Challenges and opportunities

Vitaliy Gorbachov and Olga Cherednichenko

National Technical University "Kharkiv Polytechnic Institute",
Kirpichova str., 2, Kharkiv, Ukraine

raikerian@me.com, olha.cherednichenko@gmail.com

This study constitutes of current state in enterprise solutions and massive possibilities of integrating chatbots in real world products.

The current level of development of artificial intelligence allows you to develop not only programs that solve the same type of applied problems. This class of software serves to support decision making [1]. For example, to implement a virtual interlocutor, who could answer certain questions. This is called a chatbot.

The relevance of the development of chatbot is justified by the widespread popularity of such systems. Thanks to the chatbots, you can reduce the number of calls to support users, as the chatbot can answer most of the simple questions put to it.

Chatbots models and domains are being looked into. The main discussion is a possibility to integrate chatbots into a commercial network and choose appropriate model and domain.

An enterprise network is an enterprise's communications backbone that helps connect computers and related devices across departments and workgroup networks, facilitating insight and data accessibility. An enterprise network reduces communication protocols, facilitating system and device interoperability, as well as improved internal and external enterprise data management [2].

A chatbot is a service, powered by rules and sometimes artificial intelligence, that you interact with via a chat interface. The service could be any number of things, ranging from functional to fun, and it could live in any major chat product.

Chatbots have two different models: retrieval-based and generative models.

Retrieval-based models use a repository of predefined responses and some kind of heuristic to pick an appropriate response based on the input and context. The heuristic could be as simple as a rule-based expression match, or as complex as an ensemble of Machine Learning classifiers. These systems don't generate any new text, they just pick a response from a fixed set.

Generative models don't rely on pre-defined responses. They generate new responses from scratch. Generative models are typically based on Machine Translation techniques, but instead of translating from one language to another, we "translate" from an input to an output (response).

Chatbots can be based on either open or closed domain.

In an open domain setting the user can take the conversation anywhere. There isn't necessarily have a well-defined goal or intention. Conversations on social media sites like Twitter and Reddit are typically open domain – they can go into all kinds of directions. The infinite number of topics and the fact that a certain amount of world

knowledge is required to create reasonable responses makes this a hard problem.

In a closed domain setting the space of possible inputs and outputs is somewhat limited because the system is trying to achieve a very specific goal. Technical Customer Support or Shopping Assistants are examples of closed domain problems. These systems don't need to be able to talk about politics, they just need to fulfill their specific task as efficiently as possible. Sure, users can still take the conversation anywhere they want, but the system isn't required to handle all these cases – and the users don't expect it to.

Chatbots can improve communication in most of today's enterprise solutions. As an example, we look into integrating chat bots into commercial network. This is closed domain solution, but the model is still needs to be determined. Chatbots potentially a huge business opportunity for anyone willing to jump headfirst and build something people want.

Generative model is the hardest one, but is more rewarding. It can learn in time and its optimizing itself for each use-case. But there is a common problem with generative systems is that they tend to produce generic responses like "That's great!" or "I don't know" that work for a lot of input cases.

While retrieval-based model is easiest one, it can only answer using pre-determine checks. It cannot be optimized on-run.

To produce sensible responses systems may need to incorporate both linguistic context and physical context [3]. In long dialogs people keep track of what has been said and what information has been exchanged.

The main purpose is to develop chatbot on closed domain, using generative model for commercial network.

Considering all the models we have researched, an open domain based retrieval model is impossible, because you can never collect enough responses to cover all cases. The generative system of the open area is almost artificial intelligence (AI), since it must handle all possible scenarios. We are very far from such systems (but there is a lot of research that is going on in this area).

Chatbot code will be hosted using AWS Lambda technology with a specific API gateway.

References

1. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate.arXiv preprint arXiv:1409.0473, 2014.
2. S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficultyof learning long-term dependencies, 2001.
3. K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Arxiv preprint arXiv:1406.1078, 2014.

Development and computerization of an English term system in the fields of drilling and drilling rigs

Herman Hordienko and Margarita Ilchenko

National Aerospace University "Kharkiv Aviation Institute",
Chkalova str., 17, Kharkiv, Ukraine

ggerman1995@gmail.com

The goal of the research is to modify and expand English term system, which will be used in the field of drilling, to develop a system of connecting a MultiTran system to SDL Trados to use it for translation of technical documents. Besides that we aim to make a paper reference book for people who work in the sphere of geology.

The object of the research is terms which form a term system in the sphere of drilling rigs and drilling proper. Subject of the analysis is represented by logic and semantic relation among components of term system.

The purpose and task define the choice of main research methods that include cognitive linguistics methods, method of component analysis of terms meaning, analysis of terms definitions within the terminological linguistic dictionaries (an analytical method), as well as descriptive and comparative method and empirical sources systematization in the drilling field.

Term is a word or a phrase which can be compared with distinct concept of a certain field of science, technique, arts, social and political life. It relates with other similar language units, forming a special system – a terminology. Thus, terminology is a set of terms – words or phrases – that express some specific concept in certain field of science, technique or art, as well as a set of all terms in any language.

The first stage of the research is to expand a term system of highly specialized drilling sphere based on the methods of cognitive analysis and addition of geology terms to this system. It is necessary to expand term system and add several new frames which would describe geological processes, phenomena and materials. We have added several frames (mineralogy, geochemistry, geodynamics, geohydrology) and expanded our dictionary as twice as the previous version up to 1000 dictionary entries.

The second stage of the research involves advantages of SDL Trados in translation of technical documentation. The main aim at the second stage is to connect the expanded term system based on SDL MultiTerm to the SDL Trados 2014 for further translations in the sphere of drilling.

The third step ensures making a paper reference book in geology for needs of drilling rig workers. An effective English-Ukrainian book will be useful for workers and they can refer to it for theoretical help.

Reference

1. Балог В.О. Функціонально-стилістична характеристика термінологічних

словосполучень (на матеріалі Словника української мови) // Українська термінологія і сучасність: Зб. наук. праць / Відп. ред. Л.О. Симоненко. – Вип. IV. – К.: КНЕУ, 2001. – С. 301-304.
2. Сікорська З.С. Структурно-граматична характеристика української словотвірної термінології // Українська термінологія і сучасність: . наук. праць / Відп. ред. Л.О. Симоненко. – Вип. IV. – К.: КНЕУ, 2001. – С. 267-271.

Intelligent Data Processing in Creating Targeted Advertising

Stanislav Kirkin and Karina Melnyk

National Technical University "Kharkiv Polytechnic Institute"
Kirpichova str., 2, Kharkiv, Ukraine

skirkin@ukr.net, melnikkv@kpi.kharkov.ua

For successful conduct and development of any business, many conditions must be met. One of the most important conditions affecting business is the proper conduct of advertising campaigns. Advertising campaign – is a deliberate system of planned promotional activities, united by one idea and concept in order to achieve the specific marketing goals within an advertiser's coherent marketing strategy [1].

The greatest effect from advertising is achieved when it is shown to the interested audience, i.e. to those people who can purchase the advertised product or service. To do this, the advertiser compiles a list of requirements that potential customers must meet. This method of compiling and maintaining an advertising campaign is effective and economically advantageous and is called targeted advertising. Thus, targeted advertising – are advertisements about the provided services or goods that are demonstrated only to the target audience.

There are several ways to distribute targeted advertising: the first is the advertising in social networks, the second is the mass mailing distribution of advertising messages. The first method is effective because every sixth user of the Internet has an account in a social network. But each click on the advertisement by the usual Internet user is chargeable for the advertiser. As a result advertising in social networks involves large financial investments, so in this work it is proposed to consider the second way of advertising messages in more details.

Targeted advertising assume the delivery of advertising messages only to those customers who are potentially interesting in that, so there is a need to make a decision whether or not an advertisement will be sent to a particular client. This process is the separation of client database into two segments: the first – is the target group of customers, and the second – is that part of the customers, who are unlikely will be interested in the advertised object. This formulation of the problem relates to the problem of binary classification.

Formally the problem of classification of customers can be presented in the following way. Let K – is a set of clients of the organization under consideration, where $K = \{k_1,\ldots,k_m\}$ and m – is a number of clients in the database of the considered company. We introduce Y – that is a set of groups or classes of clients. The existing entire client database should be split into such classes, i.e. $Y = \{y_1, y_2\}$. Then it is necessary to find an algorithm or mapping of one set to another, when each element of the first set put into correspondence with a particular element of the second set: $a: K \to Y$.

Consider the solution of the problem of binary classification in the context of the client database of "Atlant Shina" company, which sells automobile tires and related

products, and is also the truck tire market leader. Advertised objects have various aspects, as well as the clients have their own preferences, so at first it is necessary to make a list of input variables, based on which the decision about object demonstration to the target customers will be taken. It is proposed to use the following inputs: the characteristics of the advertised product, for example, tire dimensions, load capacity for which tires are designed; client preferences; client shopping list; number of client cars.

For intelligent processing of such information, various data processing algorithms can be applied. But due to the fact that the data have different nature, and the company "Atlant Shina" has a lot of experience in doing business and it stores sales statistics for many years, it is proposed to use the artificial neural networks, which work well with various kinds of problems [2, 3].

The first step in using neural networks for solving the classification problem is the development of a network architecture. In this work, the architecture of an artificial neural network has been developed and investigated using MATLAB software package. Further stages of the neural network approach for finding solution of the classification problem are:

- preparation of data for network training;
- training the network;
- testing the network;
- network modeling (use the network for solving the problem).

The preparation of input data is the normalization of data for the leveling of emissions or abnormal data. To solve the problem of classification of clients three-layered feed-forward neural network of perceptron type was developed:

- input layer includes five neurons such as the number of input parameters (tire dimensions; load capacity for which tires are designed; client preferences; client shopping list; number of client cars);
- hidden layer consists of seven neurons;
- output layer contains two neurons – as the number of customer groups (one group is to be sent the advertising message, and the second one is not to be sent the advertising message);
- as the activation function the sigmoid function is proposed to be used;
- as the learning algorithm of the neural network with considered topology it is suitable to use the backpropagation algorithm.

Thus the current work proposes the solution of classification problem of customers for the company that sells tires and related products, in order to be used in targeted advertising campaigns. Artificial neural networks were considered as the mathematical apparatus for solving the considered problem.

References

1. Bagiev G.L. Marketing: uchebnik / G. L. Bagiev, B. M. Tarasevich, H. Ann. – 2-nd ed., updated and revised. – M. : Ekonomika, 2001. – 354 p.
2. Neyronnyie seti dlya obrabotki informatsii / S. Osovskiy. – 2-nd ed., updated and revised. – M. : Finansyi i statistika, 2016. – 448 p.
3. Cherkassky V. Learning from data: concepts, theory, and methods / V. Cherkassky, F.M. Mulier. – 2-nd ed., updated and revised – M. : Wiley-Interscience, 2007. – 538 p.

Use of Linguistic Criteria for Estimating of Wikipedia Articles Quality

Anastasiia Kolesnik and Nina Khairova

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

kolesniknastya20@gmail.com, nina_khajrova@yahoo.com

As far as the question of texts and articles quality is urgent today, in process of a research, the concept of quality for Wikipedia articles was analysed. There were marked out linguistic criteria of quality for technical documentation and scientific articles.

Nowadays everyone knows about such informational resource as Wikipedia. Since that day when Wikipedia was just an offshoot of Nupedia (project to produce a free encyclopedia), it has become the most well-known and popular internet encyclopedia with 282 active language editions such as German, French, Russian and Polish and of course the biggest one is English edition, that has more than 5 million articles. It is multilingual, web-based, free content encyclopedia project. It takes the 5th place according to the list of the most popular websites [6].

Wikipedia is written collaboratively by largely anonymous volunteers who write without pay. Anyone, with Internet access, can write and make changes to Wikipedia articles, except in limited cases where editing is restricted to prevent disruption or vandalism. Users can contribute anonymously, under a pseudonym, or, if they choose to, with their real identity. Some users visit Wikipedia to share their knowledge, others to get (acquire) [6].

Every day, hundreds of thousands of visitors from the various parts of the world collectively make tens of thousands of edits and create thousands of new articles to augment the knowledge held by the Wikipedia encyclopedia. All users, old or young, with different backgrounds and people of all cultures can make changes in articles or add their own one.

Wikipedia's greatest strengths, weaknesses, and differences all arise because it is open to anyone, it has a large contributor base, and its articles are written according to editorial guidelines and policies. According to the *Nature* (the first to use peer review that compares Wikipedia and Britannica's coverage of science), Wikipedia's strongest suit is the speed at which it can updated, a factor not considered by *Nature*'s reviewers. Of course it has large amount of uncovered flaws, different kinds of factual errors, omissions or misleading statements.

Quality issues, however, concern the creators of Wikipedia. That's why, in 2006 during the Opening plenary at Wikimania, Jimmy Wales suggested to concentrate on quality of the articles instead of their number [2]. They created assessment system *WP: ASSESS*. It uses a letter scheme which estimates how complete the article is, assigning to the definite article its grade. According to this system, Wikipedia has 9 grades: FA (Featured Article) [4], A, GA (Good Article), B, C, Start, Stub, FL (Featured List), List. Each of these grades has special criteria. Featured articles are

considered to be the best articles in Wikipedia [2]. This kind of article must be well-written, comprehensive, well-researched, neutral and stable.

The article with A grade is well-written, clear, appropriately structured, well referenced and it contains complete description of the topic. A good article (GA) [5] also must be well written, its spelling and grammar are correct, it complies with the manual of style guidelines and it mustn't contain copyright violations or plagiarism. The prose of the Start article is not fully un-encyclopedic but it should satisfy fundamental content policies.

All experts admit that there are some difficulties in determining the quality of the Wikipedia articles [3]. Such not easy task is connected with large number of articles (3.7 million articles). It is obvious that it is not easy task to search and evaluate all of them, especially when their amount keeps growing every day. Wikipedia isn't static resource. Anyone can make changes and it can well affect article quality.

The following linguistic resources, which are well-recommended at estimating of technical documentation quality, are proposed to be used for quality evaluation of Wikipedia articles [1]:

- Writing of digits from 1 through 9 in words.
- Use of numerals for 10 and greater.
- Use of numerals for all measurements, even if the number is less than 10.
- Use of one-word verbs instead of verb phrase
- Use of only international writing of terms.
- Use of only one gap after the punctuation mark.
- There is no coma in MMMM YYYY date format.
- Use punctuation mark without extra gap.
- Use of *(from* і *through*) instead of (*between* і *and).*
- Slash cannot be a substitute of *"or".*
- Use of MMMM DD, YYYY date format.
- No abbreviation of months (only full names).
- Use of italic formatting instead of upper-case.

Given linguistic criteria for estimating Wikipedia articles quality can be easily formalized, that will allow to raise efficiency of automatic estimating of articles quality.

References

1. Microsoft Manual of Style 4th edition / Published by Microsoft Press. – 2012. – 439 p.
2. Giles G. Internet encyclopedias go head to head. Nature, 438 (2005), 900-901. Wikipedia: Manual of Style: [Electronic source]. – Access mode: http://en.wikipedia.org/wiki
3. Wikipedia: WikiProject Articles for creation/Assessment: [Electronic source]. – Access mode: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Articles_for creation/ Assessment
4. Wikipedia: featured articles: [Electronic source]. – Access mode: http://en.wikipedia.org/ wiki/Wikipedia:featu redarticles.
5. Wikipedia: good_articles: [Electronic source]. – Access mode: http://en.wikipedia.org/ wiki/ Wikipedia:good_articles.
6. Stvilia B., Twidale M.B., Gasser L., Smith L.C. Information quality discussions in Wikipedia // In Proc. ICKM, 2005. –P. 101-113.

Analysis of Existing German Corpora

Inna Olifenko and Natalia Borysova

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

innaolifenko@gmail.com, borysova.n.v@gmail.com

Nowadays, almost all modern languages have linguistic corpora. The most popular German linguistic corpora, available on the Internet and can be used during the various linguistic studies, are Das Deutsche Referenzkorpus, German political speeches Corpus and Visualization, the NEGRA corpus, the TIGER Treebank, corpus of the Berlin-Brandenburg Academy of Sciences, Limas corpus.

Das Deutsche Referenz korpus (*DeReKo* or *COSMAS II*) [1] is the corpus of contemporary written German of the Institute for German Language in Mannheim. It has 5,4 billion words and constitutes the biggest collection of machine-readable written German. It consists of fiction, science and popular-science texts, a large number of newspaper texts and other written texts. They are constantly extended. DeReKo has 3 subcorpuses: Mannheimer Korpus 1 (mk1), Mannheimer Korpus 2 (mk2) [2], Bonner Zeitungskorpus (bzk) [3]. Mk1 has 293 texts from the period 1950-1967, and about 2,2 million running text tokens. Mk2 has 52 texts from 1949, 1952, 1960-1974, and about 0,3 million running text tokens. Bzk has 10840 texts from 1949, 1954, 1959, 1964, 1969 and 1974, and about 3,1 million running-text tokens.

German political speeches Corpus and Visualization [4] contains the Presidency subcorpus and the Chancellery subcorpus. The Presidency subcorpus has a total of 1442 texts (2392074 tokens), from the period 01.07.1984-17.02.2012. It contains speeches of the presidents: Richard von Weizsäcker (1984-1994), Roman Herzog (1994-1999), Johannes Rau (1999-2004), Horst Köhler (2004-2010) and Christian Wulff (2010-2012). The speeches were got from the online archive of the German Presidency (bundespraesident.de). The Cancellary subcorpus has a total of 1831 texts (3891588 tokens), from the period 11.12.1998-06.12.2011. It contains not only speeches by the chancellors Gerhard Schröder and Angela Merkel), but also a number of other state ministers and a few unrelated speeches of other politicians. The speeches were got from the online archive of the German Chancellary (bundesregierung.de).

The NEGRA corpus [5] version 2 consists of 355096 tokens (20602 sentences) of German newspaper texts. The texts are taken from the Frankfurter Rundschau. The corpus is part-of-speech tagged and completely annotated with syntactic structures. The corpus is stored in an SQL database. Alternatively, the annotations can be represented inline-oriented export format or in PennTreebank format. The different types of information are coded in the corpus: part-of-speech tags (Stuttgart-Tübingen-Tagset (STTS)); morphological analysis (only for the first 60000 tokens, the expanded STTS); the grammatical function in the directly dominating phrase; the category of nonterminal nodes (phrases).

The TIGER Treebank (Version 2.1) [6] consists of app. 900000 tokens

(50000 sentences) of German newspaper texts. The texts are taken from the Frankfurter Rundschau. The corpus is part-of-speech tagged and completely annotated with syntactic structures. The annotations can be represented in NEGRA export format or in TIGERXML format. The different types of information are coded in the corpus: part-of-speech tags and morphological analysis (based on STTS, but modified);the grammatical function in the directly dominating phrase; the category of nonterminal nodes (phrases).The TiGer Dependency Bank (TiGer DB) covers 8000-10000 sentences of the TIGER Corpus and is created as a dependency-based gold standard for German parsers. Its annotation is close to the annotation of PARC 700 dependency bank.

Das Wortauskunfts system zur deutschen Sprache in Geschichte und Gegenwart (DWDS-Corpus) [7] is a dictionary digital system, based on very large electronic texts and corpora. It is based on six-volume German dictionary (WDG) and links it with its own texts and dictionaries. This corpus provides the user with information about the correct spelling, pronunciation of sound files and meaning of words with the help of available tags.

Limas corpus [8] contains 500 sources, the total of which has more than 1 million word forms. The collection contains the full texts and passages of different genres. You can not only read sources but also perform keyword search. Today there are three search strategies: simple search, contextual search and phrase search. Data use conditions can be such as text collections can be used for scientific and commercial purposes, provided that citation is done.

In addition to the above corpora, German is processed by such corpora as: The European Parliament hearing corpus, EU documents corpus, InterCorp, Multilingual corpus of the Oslo University, Korpora-Links auf dem Linguistik-Portal LINSE, Lehren un Lernenmit Korpora im DaF-Unterricht and Bibliotheca Augustana, IULA's UPF Textual, plurilingual, specialized Corpus and others.

All considered German corpora certainly have their advantages, but they also have some disadvantages that need be removed.

References

1. Das Deutsche Referenz korpus: [Electronic source]. – Access mode: http://www.ids-mannheim.de/cosmas2/
2. Mannheimer Korpus 1 und Mannheimer Korpus 2: [Electronic source]. – Access mode: http://www.ids-mannheim.de/kl/projekte/korpora/archiv/mk.html
3. Bonner Zeitungskorpus: [Electronic source]. – Access mode: http://www.ids-mannheim.de/kl/projekte/korpora/archiv/bzk.html
4. German political speeches Corpus and Visualization: [Electronic source]. – Access mode: http://perso.ens-lyon.fr/adrien.barbaresi/corpora/index.html
5. The NEGRA corpus: [Electronic source]. – Access mode: http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
6. The TIGER corpus, Treebank and dependency bank: [Electronic source]. – Access mode: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
7. Корпус Берлинской Бранденбургской академии наук: [Electronic source]. – Access mode: http://www.dwds.de/pages/pages_textba/dwds_textba.htm
8. Корпусы института Немецкого языка LIMAS-Korpus: [Electronic source]. – Access mode: http://www.korpora.org/Limas/

Search optimization and localization of the website of Department of Applied Linguistics

Vsevolod Pidpruzhnikov and Margarita Ilchenko

National Aerospace University "Kharkiv Aviation Institute",
Chkalova str., 17, Kharkiv, Ukraine

podpruzhnikov21@gmail.com

Localization is important to international search engine optimization (SEO) because it helps connect your product to a location using the words, terms, and behaviors of an audience in a particular region. Rather than simply use a generic term search in the hopes that it is used universally, one might prefer to use the terminology and language of the target audience.

Generally speaking, localization refers to aspects of the content of the page/website, and other marketing efforts that are tailored to a specific geographic place or region. This can mean anything from using terminology that is familiar to users within that region to establishing map locations for physical presences you have in the area.

Linguistic optimization is a set of special measures that are generally connected to the change of site content and links so that they correspond to potential users' inquiries. An ideal search advance is when the website ranks in the top three pages of the search results.

Practically any verbal phrase that makes some sense is typed into a search line as an inquiry will receive millions of links to the sources where this phrase is mentioned. It is natural that most users pay greatest attention to higher positions of the offered search results. According to statistics, no more than 85% of users follow even to the second link, and no more than 10% go further than the second link. Thus, any owner of a web resource (business company, private and public organizations, social networks, clubs etc.) wishes his/her website to be "shown through" to the Internet users and collect as many visitors as possible. In a nutshell it means they try to promote their websites to the first pages of search results (and ideally number one!). The solution of this task is provided through SEO i.e. ensuring a website can be found in search engines for words and phrases relevant to what the site is offering.

Our research consists in two stages. Firstly, we aim to improve the website of Applied Linguistics Department at National Aerospace University through reorganization of its semantic kernel using optimization techniques and meta-tags.

Secondly, we are going to localize the Ukrainian version of the website into the English version. This will add up to the department popularity and attract foreign students that might be interested in becoming an applied linguist. To illustrate, a Chinese speaking student may become our student and obtain profound knowledge of English and German.

Of great concern are the linguistic and technical aspects of localization. The first suppose perfect knowledge of translational transformations and good skills in content

management, the second focus on electronic tools such as SDL Passolo 2014. SDL Passolo is a specialized visual software localization tool intended to enable the translation of user interfaces and other software. It meets all the relevant demands of software localization.

Gamification: today and tomorrow

Yukhno Katherine and Chubar Eugenia

National Aerospace University "Kharkiv Aviation Institute",
Chkalova str., 17, Kharkiv, Ukraine

katherineyukhno@gmail.com, eugenia-pyzina@yandex.ru

Gamify your life is the motto of today. Some years ago, people tended to look for serious ways to solve serious problems. Today we came up with understanding that teenagers can manage projects; studying can be fun and easy; and gaming and gamification brings not only profit but also use and benefits.

10 years ago, games were thought to be inherently fun and not serious. People claimed that this field had no prospects in future and was a complete waste of time. Today, we see how different the situation is. Every day we meet our friends and colleges who play computer and mobile games and are not going to give up on them (no matter how old they are). A bright recent example is *PokemonGo* that proved to be a huge success last summer. Statistics says that 83 out of 100 respondents (aged from 19 to 30) played that game. Brief queries tell us that 40% of respondents use gamified apps for training and controlling their diet; 76% play games to relax or to concentrate (after work or studies); 90% support the idea that games can improve our life quality and help people. Today people are ready to take gamification as a perspective field.

Gamication is all about boosting motivation and engagement by giving small rewards and making boring and routine tasks more game-like. Games are used in huge projects that save life and make people happier and healthier, examples being NIKE +, Life is Strange, Fitocrasy, Saudi Girls Revolution, Foldit and many other games.

Psychologists appreciate gaming for motivating people, keeping them healthy, optimistic and goal-oriented. However, there are many people who are sarcastic about gaming. "Games are inherently fun and not serious", claims P.A. Newman. "The expectations from the gamification are much too high," insists Brayan Bruk. The same Brayan Bruk some years ago had forecasted that by the 2014th projects connected to games would neither bring money nor attract customers. Both proved to be wrong.

We have numbers to prove that the game market is not going to disappear soon. According to the Global Games Market Report provided by Newzoo agency the international games market reached $102.9 bln. in 2017 comparing to the $100 bln. in 2014. They reported the relentless growth of both Asian markets and mobile gaming. The market for (smart)phones and tablets rose from $17.6 bln. in 2013 to $35.4bn in 2017 – ultimately dominating one third of the global games market.

By assumptions, the worldwide games market will reach $113.3 bln. by 2018. This represents a 2014 to 2018 Compound Annual Growth Rate (CAGR) of +7.9%. The market for (smart)phones and tablets will rise from $30.0 billion this year to $44.2 billion in 2018, ultimately taking 39% of the global games market. By 2018, China and the US will grow to $32.8 billion and $24.1 billion games markets respectively, together claiming 50% of the world's games revenues.

The Computer Screen (PC/Mac) with $41.2 billion will account for 36% of the market by 2018 and will remain the most revenue generating screen, growing at a robust CAGR of +6.9% driven primarily by PC/MMO games. The Entertainment Screen (TV/Console, VR) with $26.8 billion will have 24% of the market by 2018, down from 27% in 2015. In 2018, the Personal Screen with $30.2 billion will account for 27% of the pie, leaving 13% for the Floating Screen (Tablets, Handhelds). Overall, China's games market will grow at a CAGR of +16.1% to reach $32.8 billion in 2018, while the US will grow at a CAGR of +3.1% to reach $24.1 billion in 2018. Together, they will account for 50.1%f of the global games market, up from 47.0% previous year.

Talking about Ukraine: after the 2015 video games grew at a CAGR of 12%. Recently, the number of Ukrainians taking part in gaming has grown considerably due to the development of free-to-play and "freemium" games which allow users to make small in-game purchases instead of purchasing a complete game in one go.

Sometimes weird and not serious methods bring results, and gamification is one of them. Many companies implement gamification platforms. Gamification strate-gies are used for building reputation and making people communicate outside their usual clusters and communities. Gamification helps people to interact for the sake of something bigger, solve real life problems. Gamification does motivate people to put in just a little extra effort. Even that is enough, to begin with…

Reference

1. Newzoo, Global games market investigation reports, 2016, Available at: https://newzoo.com/insights/articles/global-games-market

Author index