Automatic Multilingual Ontology Generation Based on Texts Focused on Criminal Topic

Nina Khairova[a], Anastasiia Kolesnyk[a], Orken Mamyrbayev[b], Galiya Ybytayeva[c] and Yuliia Lytvynenko[a]

[a] *National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., Kharkiv, 61002, Ukraine*
[b] *Institute of Information and Computational Technologies, 125, Pushkin str., Almaty, 050010, Republic of Kazakhstan*
[c] *NJSC "Kazakh National Research Technical University named after K. I. Satpayev", 22a Satpaev str., Almaty, 050013, Republic of Kazakhstan*

Abstract
Nowadays, the explosive growth of textual information on computer networks has made the automatic ontology generation from the text a very up-and-coming research area. The main reason for this is that usage of ontologies can produce efficient and beneficial in such different applications as information extraction, question answering systems, information retrieval and many others. However, the manual creation of ontologies is a time-consuming and costly process. Accordingly, over the past few years, many approaches tried to automate ontologies generation based on textual data have appeared.This paper suggests the approach to automated multilingual ontology generation that covers the domain focused on the criminal topic.The approach is based on the three basic components: multilingual synonym dictionary, themultilingual and parallel text corpora focused on criminal topics and the logical-linguistic model of facts extraction from texts.This paper shows these basic components created for four languages: English, Ukrainian, Kazakh and Russian.In addition, it also discusses the ontology construction process that includes all of these three mentioned essential components.

Keywords 1
Automatic ontology generation, criminal topics, multilingual English-Ukrainian-Russian corpus, Kazakh-Russian parallel corpus, multilingual synonym dictionary, logical-linguistic model, facts extraction

1. Introduction

Nowadays, law enforcement and government agencies tend more and more to focus on preventing crime and terrorism before it takes place than on dealing with crime after it had been committed [1]. In order to prevent crime it necessary to analyse a huge amount of information, including text information, exploit advanced data mining and text mining technologies and additionally NLP tools and approaches.

Many researchers realize the seriousness of the problem of the possibility to use Internet tools for illegal and extremism actions and try to establish methods of automatic detection of texts, which include various types of illegal and criminal information, in online Computer-Mediated Communication (CMC) [2]. Such communications can be blogs, Facebook, Twitter, Instagram, YouTube and etc. However, there is no information in open resources about real working systems for automatic information retrieval and identification of illegal Internet content so far.

1COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: nina.khairova@gmail.com (N. Khairova); kolesniknastya20@gmail.com (A. Kolesnyk); morkenj@mail.ru (O. Mamyrbayev) ; gali17@mail.ru (G. Ybytayeva); julialytvynenko12@gmail.com (Yu. Lytvynenko)
ORCID: 0000-0002-9826-0286 (N. Khairova); 0000-0001-5817-0844 (A. Kolesnyk); 0000-0001-8318-3794 (O. Mamyrbayev); 0000-0002-4243-0928 (G. Ybytayeva); 0000-0001-6087-8155 (Yu. Lytvynenko)

The main challenges remain such problems as a practical impossibility to find and trace contents of every slide, which potentially can point to intend and prepare some crime and blurring of text markers or digital footprints of a crime presented on the Internet.

Most studies that are aimed at using statistical, stylometric, and even lexical approaches traditionally lead to the low efficiency of information retrieval of this kind of information. The main reason for this lies with the short texts size in online CMCs groups.

In the models of information retrieval and analysis of potential criminal and illegal content, it is necessary to consider the connections and relations between words or concepts that are available only via an ontology that similar to WordNet. Obviously, the reliability and efficiency of these models will increase with the use of additional lexical bases and ontologies [3]. However, unfortunately, the linguistic base WordNet does not relate to this topic.

In our study, we suggest the approach to automated multilingual ontology generation. Our approach is based on the three basic components: (1) multilingual synonym dictionary for English, Russian, Kazakh and Ukrainian languages; (2) the text corpora focused on criminal topics in the four above-mentioned languages; (3) logical-linguistic model of facts extraction from texts [4]. The created ontology has to cover the domain that involves illegal and criminal information contented texts on the Web.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related works, corresponding to the automatic identification of texts that include illegal and criminal information and the review of the corpora focused on criminal topics challenges. Section 3 introduces three steps that underlie our approach to automatic ontology generation such as: used corpora, the multilingual synonym dictionary and the logical-linguistic model of facts extraction from texts. Section 4 describes a suggested method to multilingual ontology generation based on the corpora that include criminal and illegal text content. In the last Section 5, the scientific and practical contributions of the research, its limitations and future work are discussed.

2. Related work

2.1. The problem of automatic identification of texts that include illegal and criminal information

Today, Web-content has become an important source of information both for law-enforcement authorities and for special government security forces [5]. The existing scientific study on the problem of automatic identification of texts on the Internet, which include various types of illegal and criminal information, can be divided into two main areas: (1) the so-called psycholinguistic approach, which is often based on the task of Sentiment Analysis and (2) an approach based on keywords or, on rare occasions, ontologies.

The psycholinguistic approach allows analyzing the person mental state on the basis of texts produced by this person [6]. The majority of this type studies focus on the detection of activity behavioral markers on Internet content. Such markers can be expressed via some linguistic features that can determine attitudes, motives, intentions and even the possibility of potential criminal or radical violence. In the study [7] the markers are defined as linguistic markers for radical violence. A set of such markers can signal preparation for a criminal act in such forms as an attack on a politician, a terrorist attack, etc. Additionally, these kinds of markers can inform about illegal activities that have already carried out. That can be such forms criminals as financial fraud, copyright infringement, distribution of child pornography, hacking etc. [8, 9].

For instance, the hypothesis of the existence of a correlation between the use of the phrase and the psychological state of the author is proved in the study [10]. In the work [5] authors investigated and proved the possibility of tracing the behavioral markers of radical violence in written texts of social networks or blogs, based on the so-called “warning behaviors” occurred in texts. In the paper [11], in order to evaluate the behavior pattern of a "lone wolf terrorist" [7] authors exploited 198 variables associated with linguistic markers of their social activity, which express the ability to commit an act of violence, and also with other not linguistic markers of the actor's behavior.

However, nowadays, the studies based on the linguistic markers of behaviors remain mainly experimental and are included in applications that work with the Deep Web.

Often researches, which connect with the psycholinguistics approach, applied Sentiment Analysis methods to compare the levels of anger, hatred and racism that can be traced in the texts of various forums [12]. At the same time, the use of Sentiment Analysis approaches to identify radical and criminal-colored texts on Internet content is still not reliable and accurate [13]. In many cases, such studies are still experimental.

Along with Sentiment Analysis, some studies use Machine Learning classification approaches (naive Bayesian algorithm, logistic regression, linear SVM, random forest, gradient boosting) to automatic identify texts, which include various types of illegal and criminal information. Additionally, many papers consider the possibility of the complementary use of differentiating lexical and semantic features to improve the quality of the classification [14]. For example, in the paper [15], in order to classify Twitter messages contained terrorist support, the authors used such stylometric features as functional words, frequency words, features of punctuation, bigrams and so forth.

The second approach to the task of searching and extracting illegal and extremist information is based on the use of keywords for text analysis. For instance, in the papers [16, 17] authors provided dictionaries that contain keywords and phrases typical for various types of extremist activity. In the paper [18], the keywords of the hatred and violence topic were utilized to create so-called Mapping Websites linked through actors - users. Most often such users refer to themselves as an alias. This method was proposed to identify the authors of illegal and extremist textual information. The authors emphasized that the effectiveness of the method would increase if it could be possible to use an ontological representation of the relationship between the concepts of the subject area instead of simple keywords.

Detailed examination of the relevant research shows that despite the existence of the various approaches to search and identify illegal and criminal included content on the Web, it is too early to talk about a universal model for identifying messages or documents with criminal content and its effective utilization in applications. Rather, the existing diversity of approaches demonstrates the activities of theoretical research carried out in this scientific area.

2.2. Text corpora focused on criminal topics

There are few studies aimed at creating and describing text corpora containing some criminal context. The Old Bailey Corpus [19] is a sociolinguistically, pragmatically and textually annotated corpus based on the proceedings of the Old Bailey (Central Criminal Court). The proceedings of the Old Bailey were published from 1674 to 1913 and constitute a large corpus of Late Modern English texts. 2163 volumes contain materials of almost 200,000 lawsuits, a total of approximately 134 million words. Since the proceedings were taken down in shorthand by scribes in the courtroom, the verbatim passages are near the spoken words of the period.

Obviously, The Old Bailey Corpus includes a big lexicon that is associated with criminal activities of the period the transformation of policing in London from a system that relied on private individuals to a modern professional police system.

The Corpus of US Supreme Court Opinions contains approximately 130 million words in 32,000 Supreme Court decisions from the 1790s to the present. This corpus was released in March 2017 [20]. Texts were taken from FindLaw.com and Justia. The corpus developers also compared the texts with information from Cornell University to make sure there was no shortage of texts.

British Law Report Corpus is English court corpus. It consists of 8.5 million words of legal texts from 1,228 court decisions handed down by British courts and tribunals between 2008 and 2010. It was compiled and classified by Dr. Maria Jose Marin, a teacher of legal English from the LACELL research group at the University of Murcia, Spain. It is marked by parts of speech as well, using the Penn Treebank Tagset and is operated at the sketchengine Web platform [21].

Much less often we can see similar corpora for non-English languages. For instance, there are not in the public domain any corpora that comprise illegal and crime-connected text information for Ukrainian and Russian languages [22].

The authors of promising study [23] provided a corpus of extremist content texts in the Kazakh language. They considered automatic computation of the weight function tf-idf that determined a list of keywords of this corpus with maintaining their inflectional forms. However, unfortunately, the list is rather small for its practical use.

The paper [24] regarded a Kazakh-Russian parallel corpus that was focused on criminal topics and included texts with criminal-content from news websites of the Republic of Kazakhstan.

3. Our approach to basic components of automated multilingual ontology generation

3.1. The multilingual synonyms dictionary with criminal-connected lexis

The manually established and filled by substantive lexis dictionary is the basis for our approach to automatic ontology generation.

The lexis for our XML dictionary of synonyms we have obtained by hand from texts on crime-related topics in English, Ukrainian, Kazakh and Russian languages. Three main thematic categories were selected for the terms, namely road traffic accidents, homicide and disappearance or abduction. This choice of categories was conditioned by the fact that the information resources from which the corpus texts were taken contained the most data on these three criminal areas. This made it possible to make our dictionary narrowly focused. All terms have also been separated into their parts of speech, that is, only nouns, verbs and adjectives have been included in the dictionary. Figure 1 shows its structure scheme.

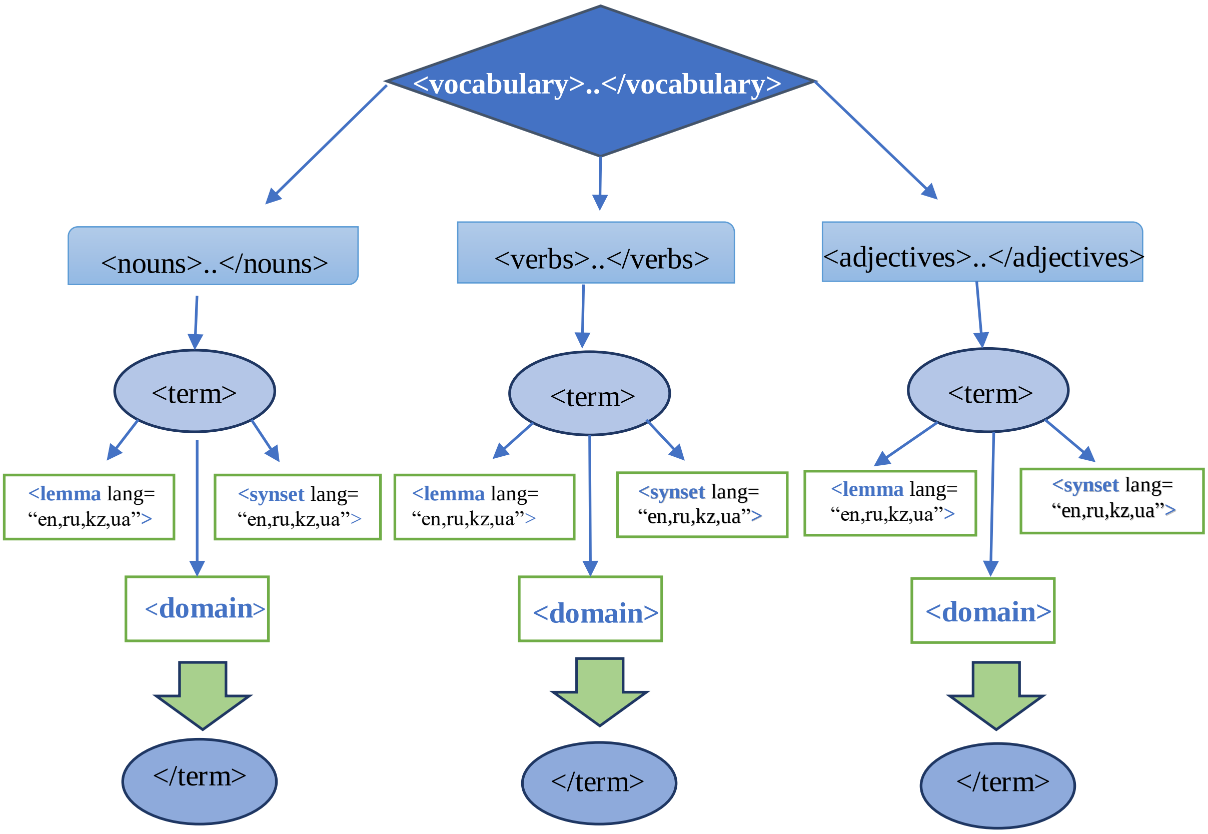
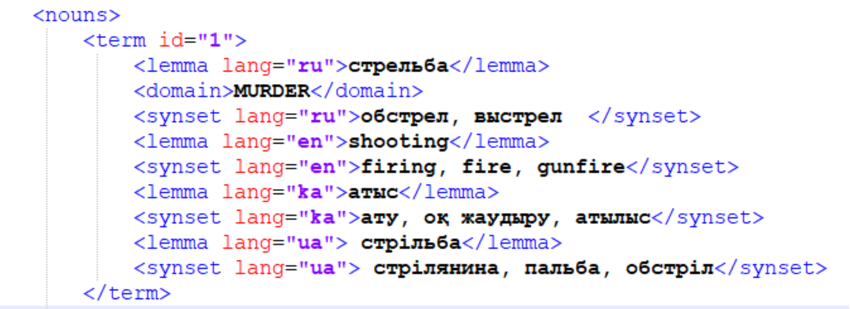


Figure 1: The structure scheme of the multilingual synonyms dictionary with criminal-connected lexis

This XML document includes three types of basic elements: <nouns>, <verbs> and <adjectives>, which, in turn, consist of child elements <term>. Each element <term> presents a word in a given part of speech with its synonyms in English, Ukrainian, Kazakh and Russian languages.Each element

<term> presents a word in a given part of speech with its synonyms in four languages via accordingly child elements <lemma> and <synset> with attribute "lang".

To date, our completely manually created multilingual dictionary of synonyms for crime-related terms is comprised of more than 500 words (about 301 nouns, 100 adjectives and 130 verbs). Figure 2 shows a fragment of our XML dictionary of synonyms.



```
<nouns>
    <term id="1">
        <lemma lang="ru">стрельба</lemma>
        <domain>MURDER</domain>
        <synset lang="ru">обстрел, выстрел  </synset>
        <lemma lang="en">shooting</lemma>
        <synset lang="en">firing, fire, gunfire</synset>
        <lemma lang="ka">атыс</lemma>
        <synset lang="ka">ату, оқ жаудыру, атылыс</synset>
        <lemma lang="ua"> стрільба</lemma>
        <synset lang="ua"> стрілянина, пальба, обстріл</synset>
    </term>
```

Figure 2: The fragment ofour synonyms XML-dictionary

In order to make the dictionary easy to use and complete, a special application has been developed that allows you to quickly add and search for new terms in the dictionary without having to open the XML file itself.

In the application, it can be easily changed languages, added a new word, its translations and synonyms, choose a subject area and a part of speech. In this way, the developed application, which is shown in picture 3, has an interface that allows fully managing the contents of the dictionary and changing its size.

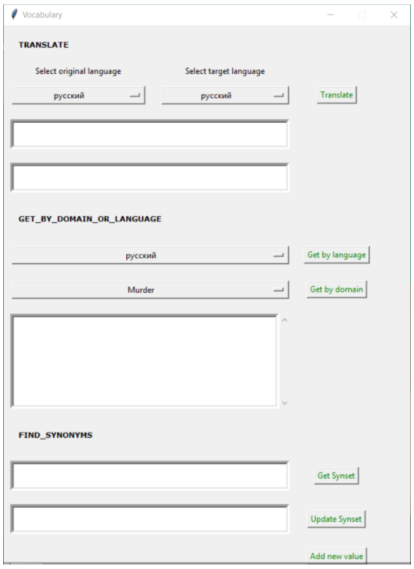


Figure 3: The program interface of the synonyms multilingual dictionary

3.2. The established text corpora focused on criminal topics

Obviously, the automatic highly specialized ontology construction from texts should be based on domain-oriented textual corpora. In order to extract some particular lexical resources for our multilingual ontology, we provide two corpora focused on criminal topics.

The first multilingual corpus comprises texts in Russian, Ukrainian and English languages. The information for its filling was taken from the Internet news sites and was collected automatically by a scraping procedure that is based on the Python BeautifulSoup library from June 2018 to October 2020 and then was placed into three directories - Ukr_texts, Eng_texts, Ru_texts. Every text in the corpus associates with the criminal topic.

The Ukrainian-language texts were automatically downloaded from the official website "Ukrainska Pravda", as well as from https://glavcom.ua and unian.ua. The Ukrainian subcorpus contains 3,147 texts. Texts in Russian were scraped from the news website "Redpost", which is the Kharkiv socio-political regional edition, namely, from the section "Crime and Accidents". This part of the corpus contains 5,506 texts. The English texts were taken manually as well as scraped from "Caller Times" - the newspaper of record for Corpus Christi, Texas, namely from it`s website, in the Crime section, and this part contains 300 texts. It is under active development.

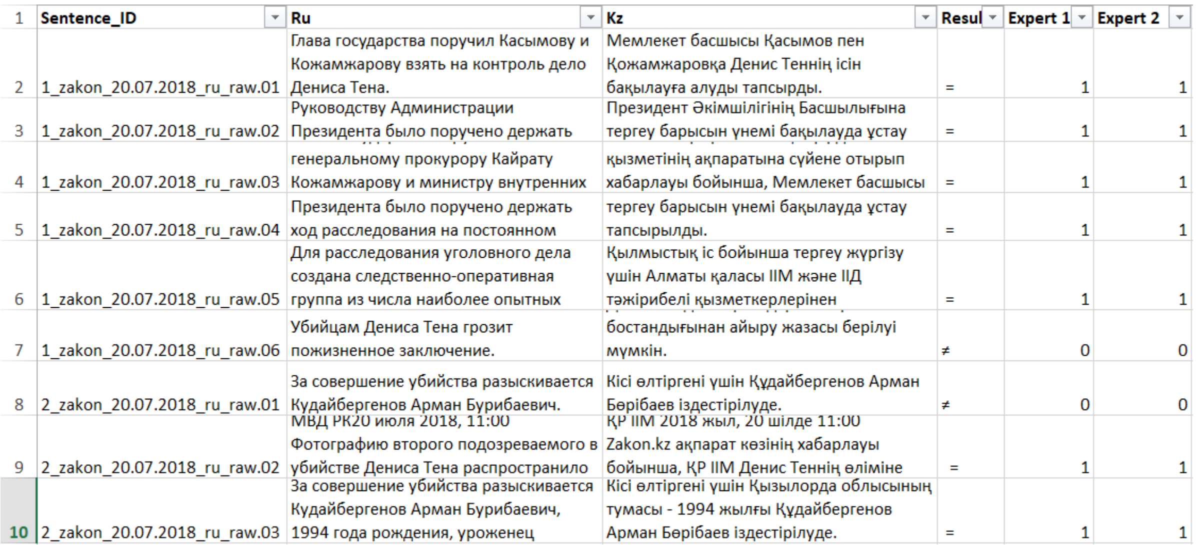
The second multilingual corpus that we use in our study is a parallel Russian-Kazakh corpus that has been developing for more than three years [24]. In this regard, it should be noted that the creation of high-quality parallel multilingual text corpora is one of the most relevant and progressive areas of modern linguistics.

Now our parallel Kazakh-Russian corpus consists of texts from four Kazakh news websites for the period 2018 - 2020. The websites from Kazakhstan's information Internet space, which we scrape with our special parsing software are zakon.kz, caravan.kz, lenta.kz, nur.kz. They contain a huge number of articles with criminal information, for example various crimes such as robbery, murder, traffic accidents and others. At the moment, the volume of the parallel Kazakh-Russian corpus is 3,000 texts in Russian and 3,000 in Kazakh.

After collecting the data, we applied our own automatic text alignment application [24]. It is based on a dictionary algorithm we developed to search for translated equivalents of words in two languages.

In the next step, in order to assess the credibility of the alignment of sentences in the parallel corpus, we checked the correctness of the automatic alignment process by the experts-philologists of two languages: Russian and Kazakh.

Each expert was given sentences in Russian and Kazakh with the result of the program evaluation, namely whether it considers them parallel or not, after which the experts had to mark their agreement or disagreement with this conclusion of the developed application with 0 - disagreement, 1 - agreement. Figure 4 shows an excerpt from the peer review results of the automatic alignment process of our parallel Kazakh-Russian corpus.



1	Sentence_ID	Ru	Kz	Resul	Expert 1	Expert 2
2	1_zakon_20.07.2018_ru_raw.01	Глава государства поручил Касымову и Кожамжарову взять на контроль дело Дениса Тена.	Мемлекет басшысы Қасымов пен Қожамжаровқа Денис Теннің ісін бақылауға алуды тапсырды.	=	1	1
3	1_zakon_20.07.2018_ru_raw.02	Руководству Администрации Президента было поручено держать	Президент Әкімшілігінің Басшылығына тергеу барысын үнемі бақылауда ұстау	=	1	1
4	1_zakon_20.07.2018_ru_raw.03	генеральному прокурору Кайрату Кожамжарову и министру внутренних	қызметінің ақпаратына сүйене отырып хабарлауы бойынша, Мемлекет басшысы	=	1	1
5	1_zakon_20.07.2018_ru_raw.04	Президента было поручено держать ход расследования на постоянном	тергеу барысын үнемі бақылауда ұстау тапсырылды.	=	1	1
6	1_zakon_20.07.2018_ru_raw.05	Для расследования уголовного дела создана следственно-оперативная группа из числа наиболее опытных	Қылмыстық іс бойынша тергеу жүргізу үшін Алматы қаласы ІІМ және ІІД тәжірибелі қызметкерлерінен	=	1	1
7	1_zakon_20.07.2018_ru_raw.06	Убийцам Дениса Тена грозит пожизненное заключение.	бостандығынан айыру жазасы берілуі мүмкін.	≠	0	0
8	2_zakon_20.07.2018_ru_raw.01	За совершение убийства разыскивается Кудайбергенов Арман Бурибаевич.	Кісі өлтіргені үшін Құдайбергенов Арман Бөрібаев іздестірілуде.	≠	0	0
9	2_zakon_20.07.2018_ru_raw.02	МВД РК20 июля 2018, 11:00 Фотографию второго подозреваемого в убийстве Дениса Тена распространило	ҚР ІІМ 2018 жыл, 20 шілде 11:00 Zakon.kz ақпарат көзінің хабарлауы бойынша, ҚР ІІМ Денис Теннің өліміне	=	1	1
10	2_zakon_20.07.2018_ru_raw.03	За совершение убийства разыскивается Кудайбергенов Арман Бурибаевич, 1994 года рождения, уроженец	Кісі өлтіргені үшін Қызылорда облысының тумасы - 1994 жылғы Құдайбергенов Арман Бөрібаев іздестірілуде.	=	1	1

Figure 4: Excerpt from the peer review results of the automatic alignment process of our parallel Kazakh-Russian corpus

The measure of the agreement of the experts' opinions was calculated using the Kappa Cohen coefficient, which showed almost complete agreement among the experts (agreement ≈0.98) and with the results of the application. From this, we can conclude that the multilingual Kazakh-Russian corpus that we have developed can be called a parallel corpus.

3.3. The logical-linguistic model of facts extraction from texts

The next stage of the ontology building is its automatic filling and extension. This stage is based on our logical-linguistic model of information extraction from unstructured texts [4].The model allows representing a fact from a text by the RDF-triplet format without defining specific relation types in advance. Since this kind of facts is usually expressed by various unregulated constructions of the natural language, we identify lexical units that name the participants of the action (the Subject and Object) and semantic relations between them in the sentence. For this purpose, we define semantic functions of the action participants via logical-linguistic equations that describe the relations of the grammatical and semantic characteristics of the words in a sentence. In a general way, such a logical-linguistic equation can be represented via the multi-place predicate P ($x_1,\ldots, x_n$):

$$P(x_1,\ldots,x_n) = \gamma_k(x_1,\ldots,x_n) \times P_1(x_1,\ldots,x_n) \times \ldots \times P_n(x_1,\ldots,x_n), \qquad (1)$$

where k∈[1,h], h is the number of participants and attributes of the action. The predicate $\gamma_k(x_1,\ldots,x_n) = 1$, if the conjunction of the grammatical characteristics of the sentence words shows a certain semantic role of the participant (Subject or Object) and the attribute of the action, and $\gamma_k(x_1,\ldots,x_n) = 0$, otherwise. Therefore, if the relations between the grammatical characteristics of the words in the particular sentence in the specific language do not express any fact element, they are removed from the formula (1) by the predicate$\gamma_k(x_1,\ldots,x_n)$.

Using POS-tagging and some syntactic characteristics of words in the sentence as the values of predicate variables in corresponding equations allows us to extract Subjects, Objects and Predicates of facts and from the texts corpora. By now, we have adapted our model to English [4], Ukraine, Russian [25] and Kazakh [26] languages.

4. The approach to (semi-) automatic filling and extension of the ontology

Our suggested approach to semi-automatic filling and extension of the ontology is based on several well-known hypotheses. The main statistical semantics hypothesis states that statistical patterns of human word usage can be used to figure out what people mean [27]. In other words, human intelligence can understand words according to their surroundings. This general hypothesis underlies the more specific distributional hypothesis in linguistics. According to [28], the hypothesis states that words that occur in similar contexts tend to have similar meanings.

However, unlike the traditional VSM approach, which handles the window of words, in our study, we consider syntactic relations in a sentence. This approach can be based on the hypothesis that the meaning of an entity is limited by possible combinations of this entity with other concepts or entities. Therefore, in order to define the belonging of words to the common semantic area, it is necessary to consider not only the words in the surrounding context but exactly the grammatically related words of the context.

In practice, based on our above-mentioned logical-linguistic model, we extract a fact from a sentence. In the most common case, the fact is the triplet of the Subject, Object and Predicate.We consider these concepts as semantic categories [4].The subject names the actor of the action that is described in a sentence. The object names an item or person, on which the action is directed. And the predicate, in turn, names the action of the sentence.

According to previous studies [24, 25, 26], we created two multilingual corpora that are focused on criminal topics and the synonyms dictionary with basic criminal-connected lexis in Ukrainian, Kazakh, English and Russian languages. With the help of the developed application for automatic extraction of triplet facts, we were able to develop an algorithm that allows us to fill automatically our dictionary with terms in several languages.

Figure 5 shows the general scheme of our approach to filling and extension of the dictionary and creating the multilingual ontology.

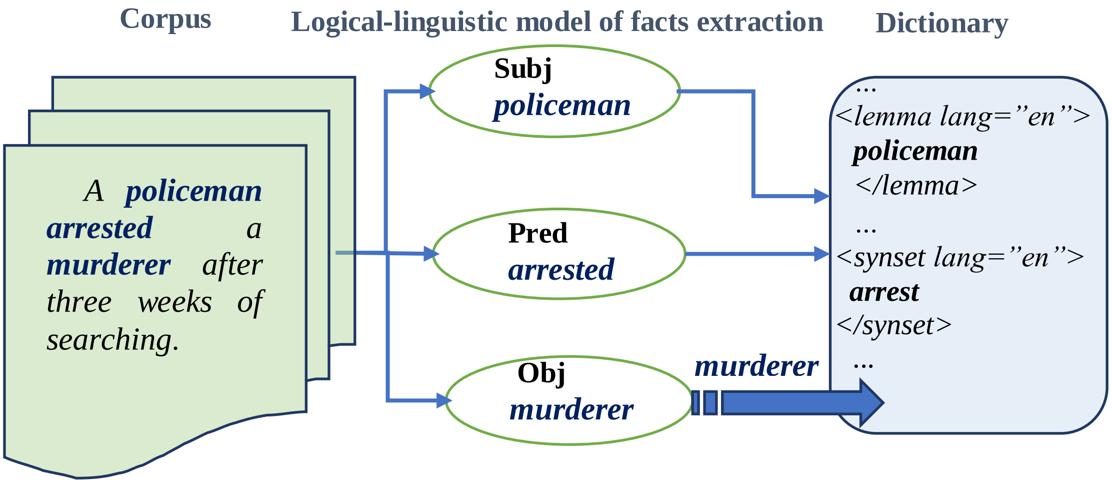


Figure 5: The general scheme of our approach to filling and extension of the dictionary and creating the multilingual ontology

With the help of the developed application for automatic extraction of triplet of facts, we were able to develop an algorithm that allows us to automatically fill our dictionary with terms in several languages. In the first step of the algorithm, we analyse each individual sentence of a text and find RDF-fact that is represented by the predicate, the subject and the object. In the next step, we check if three found elements are in our dictionary for every particular language. If two components of the triplet are found as elements of <*lemma*> tag or <*synset*> tag and one component is not found in the dictionary, the last one is automatically placed in it. Table 1 shows the examples of automatically extracted facts from the sentences, which are included a subject, an object and a predicate, the result of searching for these lemmas in the dictionary and the adding of the missing lemmas into the dictionary.

Table 1
The examples of automatically extracted facts from the sentences, the result of searching for lemmas of the subject, object and predicate in the dictionary and the addition of the one missing lemma into the dictionary

Sentence	Automatically extracted triplet of the fact	Lemmas in dictionary	Lemma is placed into dictionary
The governor stormed into the hospital and demanded to know how many children died.	*Subj*: governor		governor
	Obj: hospital	hospital	
	Pred: demanded	demand	
Police are searching for a person of interest in the murder of an 18-year-old woman in November.	*Subj*: police	police	
	Obj: person of interest		person of interest
	Pred: searching	search	
Police encountered a distraught woman crying that her baby had died.	*Subj*: police	police	
	Obj: distraught woman		distraught woman
	Pred: encountered	encountered	

In the last step, a native speaker has to check the result of the automatic filling of the dictionary so that it fully corresponds to its thematic focus, namely, the criminal related information. In this way, the XML-vocabulary of the terms is expanded in parallel with the expansion of the corpus and becomes the basis of ontology.

5. Conclusions and future works

In this work, we propose the approach to automated multilingual ontology generation. Our approach is based on the exploitation of the synonym dictionary, two multilingual text corpora and the logical-linguistic model of facts extraction from a sentence. Since all these components are established for English, Ukrainian, Kazakh and Russian languages and cover criminal-related information, the created ontology has to address the domain that involves illegal and criminal information contented texts in these four languages.

Hence, in our research, we focused on a narrow thematic area for our ontology, namely criminal terms and information. We consider 2 designed corpora: a parallel Russian-Kazakh corpus and a multilingual corpus in 3 languages (English, Ukrainian and Russian). Each corpus contains criminally related vocabulary and reflects the current state of affairs in the field.

The present work resulted in the creation of an automatically filled four-language dictionary of terms and synonyms on criminal topic. Whereas the main relation in our XML-dictionary is synonymy relation, we can consider the dictionary as a basis for the ontology.

In future work, in order to fill <synset> elements of the ontology, we plan to apply automatic extracted semantic similar words from texts of our corpora. This step will be based on the VSM (Vector Space Model) and the word2vec algorithm.

Further, it will be possible to choose the fourth element of the <domain> tag, which will cover terms that are thematically overlapping and not be clearly assigned to road traffic accidents, homicide and disappearance or abduction classes.

Additionally, in the future, the proposed approach can be reconfigured to suit different corpus and research topics.

6. Acknowledgements

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic Kazakhstan (Grant No. AP09259309).

7. References

[1] J. Mena, Machine Learning Forensics for Law Enforcement, Security, and Intelligence. Auerbach Publications, 2011,349 p.

[2] J. Ware, Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos. ICCT Policy Brief, 2020. doi: 10.97812345/2020.4.2.

[3] B. S. Iskandar, Terrorism detection based on sentiment analysis using machine learning, Journal of Engineering and Applied Sciences, 2017 № 12 691–698.

[4] N. F. Khairova, S. Petrasova, A. P. Gautam, The logical-linguistic model of fact extraction from English texts. In Information and software technologies. Volume 639 of the series communications in computer and information science. Springer, Cham, 2016. doi:10.1007/978-3-319-46254-7_51.

[5] K. Cohen, F. Johansson, L. Kaati, J. Clausen Mork, Detecting Linguistic Markers for Radical Violence in Social Media. Terrorism and Political Violence V.26. № 1, 2014, pp.246–256. doi: 10.1080/09546553.2014.849948.

[6] W. J. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn,The development and psychometric properties of LIWC2015, University of Texas at Austin, UT Faculty Researcher Works, 2015.

[7] M.S. Hamm, R. FJ. Spaaij, S. Cottee,The age of lone wolf terrorism,New York, NY: Columbia University Press, 2017.

[8] K. Cohen, L. Kaati, A. Shrestha, T. Isbister, Linguistic markers of a radicalized mind-set among extreme adopters / cyber deviance detection, 2017.

[9] M. A. Kaufhold, C. Reuter, Cultural Violence and Peace in Social Media / Information Technology for Peace and Security, 2019, pp. 361-381.

[10] Y. Goldberg, O. Levy, Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, ArXiv e-prints, 2014.

[11] B. Schuurman, E. Bakker, P. Gill, N. Bouhana, Lone Actor Terrorist Attack Planning and Preparation: A Data-Driven Analysis, Journal of Forensic Sciences Vol. 63, No. 4.2018, pp. 1191-1120.doi.org/10.1111/1556-4029.13676.

[12] T. Fu, A. Abbasi, D. Zeng, H. Chen, Sentimental spidering: leveraging opinion information in focused crawlers, ACM Transactions on Information Systems (TOIS), V. 30 № 4 2012, pp. 1-30.

[13] R. Scrivens, G. Davies, R. Frank, Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors, Behavioral sciences of terrorism and political aggression, V. 10 № 1,2018, pp. 39-59.doi.org/10.1080/19434472.2016.1276612.

[14] M. Nouh, RC. J. Nurse, M. Goldsmith, Understanding the radical mind: Identifying signals to detect extremist content on twitter, 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), 2019, pp.98-103.

[15] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, N. Prucha, Detecting jihadist messages on twitter, Intelligence and Security Informatics Conference (EISIC), 2015, pp.161-164. doi: 10.1109/EISIC.2015.27.

[16] M. A. Finlayson, J. R. Halverson, S. R. Corman, The N2 Corpus: A Semantically Annotated Collection of Islamist Extremist Stories in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014, pp. 896-902.

[17] P. Wadhwa, MPS. Bhatia, Classification of radical messages in Twitter using security associations, Case studies in secure computing: Achievements and trends, 2014, pp. 273-294.doi: 10.5815/ijisa.2014.05.04.

[18] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, P. Svenson, Harvesting and analysis of weak signals for detecting lone wolf terrorists, V. 2. № 1, 2013, pp.11-26.

[19] Old Bailey Corpus, 2020. URL: https://www.oldbaileyonline.org/static/HowToReadTrial.jsp

[20] COSCO-US (Corpus of US Supreme Court Opinions), 2019. URL: https://www.english-corpora.org/scotus/

[21] BLaRC (British Law Report Corpus). URL: https://www.sketchengine.eu/blarc-british-law-reference-corpus/

[22] D. Devyatkin, I. Smirnov, M. Ananyeva, M. Kobozeva, A. Chepovskiy, F. Solovyev, Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts), 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), 2017.pp. 188-190.

[23] M.A. Bolatbek, Sh.Zh. Mussiraliyeva, U.A. Tukeyev, Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language, Journal of Mathematics, Mechanics and Computer Science, vol. 97, № 1, 2018, pp. 134–142.

[24] N. Khairova, A. Kolesnyk, O. Mamyrbayev, K. Mukhsina, The aligned Kazakh-Russian parallel corpus focused on the criminal theme in: CEUR Workshop Proceedings, 2019, pp. 116-125.

[25] N. Khairova, W. Lewoniewski, K. Węcel, O. Mamyrbayev, K. Mukhsina, Comparative Analysis of the Informativeness and Encyclopedic Style of the Popular Web Information Sources. In: Abramowicz W., Paschke A. (eds) Business Information Systems. BIS 2018. Lecture Notes in Business Information Processing, vol 320. Springer, Cham. doi.org/10.1007/978-3-319-93931-5_24.

[26] N. Khairova, O. Mamyrbayev, K. Mukhsina, A. Kolesnyk, Logical-Linguistic model for multilingual open information extraction, Cogent Engineering, 7:1, 1714829, 2020.

[27] P.D. Turnay, P. Pantel, From frequency to meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research 37, 2010, pp. 141-188.

[28] Harris Z., Distributional structure. Word, , 10 (23), 1954, pp.146-162.