Creating a Thesaurus "Crime-Related Web Content" Based on a Multilingual Corpus

Galiya Ybytayeva[1], Orken Mamyrbayev[2], Nina Khairova[3,4], Nina Rizun[5], Sanzharsultan Berdali[1], Kuralai Mukhsina[2]

[1] *Satbayev University, 22a Satbayev str., Almaty, 050013, Republic of Kazakhstan*
[2] *Institute of Information and Computational Technologies, 125, Pushkin str., Almaty, 050010, Republic of Kazakhstan*
[3] *National Technical University"Kharkiv Polytechnic Institute", 2, Kyrpychova str., Kharkiv, 61002, Ukraine*
[4] *Umeå University, 901 87 Umeå, Sweden*
[5] *Gdansk University of Technology, 11/12 Gabriela Narutowicza Street, Gdańsk, Poland*

Abstract

An overview of the most common ontological resources and methods of their construction and application is given. For purposes of scientific research we analyzed the characteristics of ontologies in the public domain and corpus containing criminal context. Additionally, we have recently developed a Flask-based web application that generates ontologies using the Anytree library.

The authors also developed a multilingual basic ontology called "Illegal Web content" based on a corpus of texts in criminal context in English, Ukrainian, Kazakh and Russian languages. The development of this ontology was motivated by the need for effective analysis and prevention of criminal activities based on textual information disseminated on the internet.

The newly developed web application allows users to create ontologies by importing text files in different languages, and then automatically generates an ontology based on the text. The application is user-friendly, and allows users to customize the ontology by adding or removing nodes, changing the labels of nodes and edges, and setting the weight of edges.

Overall, the development of the "Illegal Web content" ontology and the web application represents a significant contribution to the field of ontology development and text processing for criminal investigation and prevention. The main characteristics of the Web application, including its ease of use and customizability, make it a valuable tool for researchers and practitioners alike.

Keywords

Multilingual basic ontology, criminal topics, multilingual corpus, Kazakh-Russian parallel corpus, Web application.

1. Introduction

The issue of creating software tools and mechanisms to assist in the detection and prevention of criminal activities through online textual information remains a major challenge today. Although numerous applications have been developed to tackle this problem, such as hate speech detection, crime modeling, crime prediction, identification of crime-related topics, and many more, the challenge of text-based crime prevention and investigation in Computer-Mediated Communication (CMC) and social networking communications still requires more efficient and effective methods of criminal content analysis. This can be achieved through the application of subject matter knowledge, which will enable the development of more advanced techniques for detecting and preventing criminal activities based on

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine
EMAIL: ybytayeva.galiya@gmail.com (G. Ybytayeva); morkenj@mail.ru (O. Mamyrbayev); nina.khairova@gmail.com (N. Khairova); nina.rizun@pg.edu.pl (N. Rizun); sanchess.berdali@gmail.com (S. Berdali); kuka_ai@mail.ru (K. Mukhsina)
ORCID: 0000-0002- 4243-0928 (G. Ybytayeva); 0000-0001-8318-3794 (O. Mamyrbayev); 0000-0002-9826-0286 (N. Khairova); 0000-0002-4343-9713 (N. Rizun); 0000-0001-5856-4982 (S. Berdali); 0000-0002-8627-1949 (K. Mukhsina)

online textual information. As technology continues to advance, it is important to continue to develop innovative solutions to tackle this problem and ensure that the online space remains safe for all users..

Criminal investigations involve the collection and analysis of large volumes of data from various sources. The data include crime scene information, witness statements, forensic evidence, and criminal records. The complexity and size of these data make it difficult for investigators to identify and analyze the relevant information efficiently. Ontologies have been proposed as a solution to this problem.

Software tools that enable knowledge extraction provide essential support for linguistic research. In order to make the automatic processing of texts more qualitative and reliable, it is necessary to use knowledge both about language and about the surrounding world. Knowledge about the world can be represented using ontologies.

Thesauri as formalized information resources have been known for quite a long time. In the last 15 years such type of information resources as ontologies has been actively discussed.

The word ontology has two meanings:

- Ontology 1. – A philosophical discipline that studies the most general characteristics of being and entities;
- Ontology 2. – An artifact, a structure that describes the meanings of the elements of some system.

In this article we will use the word ontology in its second meaning as some computer resource, which is some description of a view of the world as applied to a particular area of interest.

On a formal level, an ontology is a system consisting of a set of concepts and a set of statements about these concepts, based on classes, objects, relations, functions and theories can be constructed.

One of the most famous definitions of ontology formulated by T. Gruber is [1]: Ontology is the formal specification of a coherent conceptualization. For all the differences to the definition of ontology, many authors agree on a set of basic components of ontology.

The main components of an ontology are classes or concepts, attributes, relations, axioms, instances. A very broad treatment of classes (concepts) of ontology is often used. With a broad interpretation it is stated that classes (concepts of ontology) can be abstract and concrete, elementary and composite, really existing and imaginary. In other words, a class (concept) can be any entity about which any information can be given.

Ontologies enable the use of knowledge about the world, which is necessary to perform many stages of textual analysis. Ontologies have been used in various domains, including healthcare, finance, and e-commerce. In the context of criminal investigations, an ontology can be used to represent crime-related knowledge and information.

This paper presents the development of a crime ontology based on the Flask web framework and the Anytree library. The ontology was designed to capture and represent crime-related knowledge and information. The Flask web framework was used to create a web-based interface for the ontology, while the Anytree library was used to manage the ontology's hierarchical structure. The ontology was evaluated using a case study of a criminal investigation, which demonstrated the effectiveness of the ontology in facilitating the investigation process.

The presented paper consists of three parts. The first part analyzes the characteristics of ontologies, identifies the dynamics of the development of these resources and evaluates the possibilities of their application to the tasks of automatic text processing. In the second part, the authors review the methods of term extraction and the relationships between them used by modern researchers in the construction of lexical resources in order to determine the main trends and possible directions of development in this area, primarily for the construction of ontologies. The third part describes its own experience in constructing a multilingual basic ontology. In the conclusion the results of the work are summarized.

2. Related Works

The number of projects aimed at fulfilling the demands for ontology development and maintenance is on the rise.

MicroComos ontology (later called OntoSem) is one of the best-known ontological resources. This ontology is developed within an approach called "ontological semantics" [2]. The ontology is intended for use in automatic text processing applications and the construction of a semantic, language-

independent representation of the content of text sentences. For incoming text, preprocessing, morphological analysis, syntactic analysis, semantic analysis are done, the results of which are presented as Text-Meaning Representation (TMR).

As the authors have made significant efforts to limit the size of the ontology, the size of MicroCosmos ontology (OntoSem) is about 6 thousand concepts, each of which is described by an average of 16 properties. The lexicon of the system is several tens of thousands of words and expressions.

One of the currently well-known projects in the field of lexical semantics description is the FrameNet linguistic resource, which was created under the guidance of the famous linguist Charles Fillmore [3] within the concept of frame semantics. The goal of the project is to create an online lexical resource based on frame semantics and to provide it with a base in the form of a text corpus. The project aims to describe the semantic and syntactic combinability of words - valences - for each word in each known sense.

In 2009 the resource contained 960 hierarchically organized frames with over 11,000 lexical units associated with them.

The largest thesaurus in Russian by volume is the RuThes thesaurus. This project is being developed by the Laboratory of Information Research. It is based on the principles of WordNet, but the model for describing entities is different. The unit of the thesaurus is a concept, equipped with a set of terms, the values of which correspond to a given concept. As terms can be words and phrases, the number of which may be large enough, such as 20 or more. Words and word combinations relating to one concept are called ontological synonyms.

Currently, the RuThes thesaurus contains 55 thousand concepts, 158 thousand words and expressions, 210 thousand relations between these concepts.

Another active project to create a thesaurus of the Russian language is Yet Another RussNet (YARN). The developers use a model completely corresponding to WordNet, which is based on synsets - groups of synonyms and quasi-synonyms, united by a common lexical meaning. Synsets are linked by hierarchical relationships and by antonymy relationships, which are established between terms with opposite meanings. Additionally, there are relations between words as well as inter-lingual relations between YARN and WordNet synsets.

Information from the Wiktionary was used as initial content. A distinctive feature of this project is the organization of the addition of the thesaurus based on a crowdsourcing approach: anyone, after registering on the YARN site, can participate in adding and editing data.

Currently YARN contains 143 508 words, 69 799 synsets, 104 906 unedited pairs of synonyms, and 29 764 unedited genus-species relations.

There are also crime-related ontologies for interpreting terms and relations in this context, and these are further used to represent knowledge in some existing systems. Examples include the Multi-Modal Situation Assessment and Analytics Platform (MOSAIC) project [4]; CAPER [5], which simultaneously uses a European LEAs interaction ontology and a multilingual crime ontology; and the ePOOLICE project [6], COPKIT project [7], ASGARD project [8], TENSOR project [9].

The creation and maintenance of ontologies have become a popular area of focus in various projects. One such project is the development of the CAPER ontology, which was designed to address drug crimes by following INTERPOL classification patterns. This multi-lingual ontology consists of four main concepts - "Crimes", "Techniques", "Essential Conditions", and "Countries" - represented as classes within its structure. The current version of the ontology contains 346 nodes, and efforts are ongoing to collect translations, synonyms, and slang terminologies for Italian, Spanish, English, and Hebrew.

In addition, several other projects have been created to support law enforcement agencies in analysing and preventing criminal activities. For instance, the COPKIT project has developed data-driven policing technologies to aid in investigative and strategic analysis work. Its toolkit enables the production and exploitation of knowledge, supporting the Early Warning/Early Action paradigm at both operational and strategic levels.

The ASGARD project has also developed a best-of-class tool set for the extraction, fusion, exchange, and analysis of Big Data, including cyber-offense data for forensic investigation. Another project, TENSOR, has created a unified semantic infrastructure for the fusion of terrorism-related content and threat detection on the web. This framework includes an ontology and an adaptable semantic reasoning

mechanism, providing Law Enforcement Agencies with planning and prevention functionalities for the early detection of terrorist activities, radicalisation, and recruitment. However, these tools usually have two main drawbacks. First, they are not multilingual, and existing annotated crime corpora are mostly available for English. Second, some of them have limited visualization capabilities, namely the graphical representation of recognized relationships between named entities.

Analysis has shown that there are currently no publicly available ontologies of Russian and Kazakh languages related to criminal topics.

The criminal domain has its own vocabulary and narrative and assimilates a writing style. Therefore, crime experts advise the inclusion of crime news and official websites in the corpus, arguing that they follow the same narrative form and similar requirements as crime reports. This paper presents a parallel Kazakh-Russian corpus that is used to search for information in crime-related documents.

Studies aimed at creating and describing corpuses of texts containing criminal contexts are few. The best known is the Old Bailey Corpus [10], which presents socio-linguistically, pragmatically, and textually annotated late-modern English texts based on the proceedings of the Central Criminal Court, which were published from 1674 to 1913. The 2163 volumes of the corpus contain the records of nearly 200,000 lawsuits, totaling some 134 million words. Unfortunately, it is evident that the Old Bailey corpus, though it includes a large vocabulary related to criminal activity, is more related to the period of privately based policing in London than to the vocabulary of modern professional policing.

A second interesting corpus containing illegal content is the US Supreme Court Opinions corpus, which contains about 130 million words included in 32,000 court decisions from the 1790s to the present. The corpus was released in March 2017 [11]. Texts are taken from FindLaw.com and Justia and subsequent comparison with Cornell University information.

The British Law Report Corpus is a corpus of 8.5 million words of legal texts from 1,228 court decisions handed down by British courts between 2008 and 2010. The corpus was compiled and classified by Dr. Maria Jose Marin, a professor of legal English at the University's LACELL Research Group. Murcia, Spain. The text contains morphological and syntactic markup, using the Penn Treebank tag set and runs on the web platform sketchEngine [12].

It is much rarer to find similar corpora for all other languages excluding English. For example, there are no publicly available corpora containing illegal or crime-related textual information for Kazakh and Russian [13].

The authors of a promising study [14] provided a corpus of extremist texts in Kazakh. They considered the automatic calculation of the weight function TF-IDF, which determines the list of key words of this corpus with preservation of their inflective forms. However, this corpus is too small for practical use. In our study we are based on the use of the Kazakh-Russian parallel corpus, which includes texts of criminal content of the news sites of the Republic of Kazakhstan [15].

Thus, the review shows that although there are currently enough studies aimed at the search and analysis of illegal content, mainly the existing developments involve the processing of texts in English, French, Chinese and some other European languages.

The complete absence of available multilingual subject area ontologies shows the current direction of work for ontology creators. It should also be noted that the openness of both such lexical resources themselves and models for their development may allow improving the quality of ontologies, as well as making a significant contribution to solving the problems of automatic text processing in the relevant fields.

3. Multilingual corpus

In this study, in order to extract specific lexical resources from the texts for the multilingual ontology being created, two corpuses dedicated to criminal topics were used.

The first multilingual corpus includes texts in Russian, Ukrainian and English. The information for its filling was obtained from Internet news websites, using the Python library parser BeautifulSoup from June 2018 to October 2020. Each text in the corpus is related to a crime topic and placed in one of the three directories Ukr_texts, Eng_texts, Ru_texts.

The texts in Ukrainian were automatically downloaded from the official website of Ukrainska Pravda, as well as from the website of Glavcom. The Ukrainian subcorpus contains 3,147 texts.

The texts in Russian were obtained from the news website "Redpost", a Kharkiv socio-political regional publication, namely from the section "Crime and incidents". This part of the corpus contains 5506 texts. The English-language texts were obtained from the Corpus Christi, Texas newspaper Caller Times, Crime section. This sub-corpus currently contains 300 texts. The development of this corpus will continue in the next phases of this study.

The second multilingual corpus, which is used as a base for ontology generation, is a parallel Kazakh-Russian corpus, which has been developing for more than three years [15]. In this regard, it should be noted that the creation of high-quality parallel multilingual corpus of texts is one of the most relevant and progressive directions of modern linguistics.

Expansion and addition of this parallel corpus was carried out by parsing four news sites of the information Internet space of Kazakhstan zakon.kz, caravan.kz, lenta.kz, nur.kz during the period from March to November 2021. These sites contain a huge number of news articles related to criminal information, including about such crimes as robbery, murder, traffic accidents and others. At the moment the volume of parallel Kazakh-Russian corpus contains 3000 texts in Russian and 3000 texts in Kazakh, including two thousand texts with aligned Kazakh-Russian sentences [16].

After conducting research based on previous studies [17-19], we identified and extracted Crime-Related Events (CRE) from a collection of news articles relating to police and criminal activities. In contrast to previous research that focused on specific types of crimes (such as drug-related crimes or traffic incidents), we considered a broad range of events that involved unlawful actions, including Traffic Accidents, Hate Crimes, and Police Activities. Our focus was on three types of events - TRANSFER, CRIME, and POLICE - and their seven subtypes, which are listed in Table I.

Typically, all CRE involve two participants and several attributes of the action or event. The Agent is the participant who initiates the event, while the Object is the person, organization, or vehicle to which the event action is directed. To determine the participants of CRE, we used the Coplink project [20], which distinguishes between three types of entities that can be involved in a criminal action: people, organizations, and vehicles. However, different types and subtypes of CRE may involve different entity types as Agents and Objects.

All the types and subtypes of events we considered have traditional TIME-ARG and PLACE-ARG attributes. In some cases, we also looked for the Instrument or device used to carry out the event, such as a weapon. In rare cases, we used an optional slot called WHY-ARG to describe the reason for the event.

A CRIME CRE occurs when a person or organization engages in a criminal or unlawful act. There are two subtypes of CRIME events: INJURE and OFFENSE. An INJURE subtype occurs when a person is physically harmed or affected by a criminal action. In this case, the Object is the harmed person(s), while the Agent is the initiator of the attacking action, whether it is a person or an organization.

An OFFENSE subtype occurs when the object of the criminal action is not a person directly. In this case, the Agent is the initiator of the offense, while the Object can be an inanimate object or may not be mentioned.

A TRANSFER CRE includes two subtypes: MOVEMENT and TRAFFIC ACCIDENT. A MOVEMENT subtype occurs when an inanimate object or a person is moved from one location to another. However, if the movement is carried out to steal or commit a crime, it is considered a CRIME CRE. The other subtype, TRAFFIC ACCIDENT, occurs when a vehicle is involved in an accident, and the Agent is the person or vehicle that caused the accident.

The final type of event we considered as a CRE is a POLICE Event, which occurs when an action is carried out by police or other officials. This type includes three subtypes: ARREST, TRIAL, and PD. An ARREST occurs when a person's movement is restricted by a state actor such as a policeman or a judge. The Agent is the person or organization that initiated the detention, while the Object is the detained person.

A TRIAL occurs when a court or government organization accuses a person or organization of a crime, and a PD subtype occurs when a police officer carries out official duties. In this case, the Agent is a policeman or a police department as an organization.

4. A method of thesaurus creation

The basic sources of "Illegal Internet Content" ontology generation are the developed parallel Kazakh-Russian corpus, which includes texts containing criminal news [15] and Web application of multilingual basic ontology "Illegal Web Content".

The Web Application is a multilingual lexical database. The scope of the Web application is integrated law enforcement information and forensic systems, as well as integrated information and analytical systems of other government agencies used for information search and analysis of illegal information in the open part of the Internet content, namely, in group online communities of discussion (Computer-Mediated Communication (CMC)) of various blogs, social networks, web media in Kazakh, Russian, English and Ukrainian languages.

The Web application is also freely and publicly available at http://94.247.129.230:5000/. The structure of the Web application makes it a useful tool for the development of computational linguistics and artificial intelligence approaches in modeling the meaning of text documents, as well as developing methods of semantic analysis of texts, automatic generation of ontologies based on unstructured texts and information search.

The web application can be commercialized using the developed models, methods and algorithms in such applications of automatic processing of Kazakh, Russian, English and Ukrainian text documents as automatic rewriting, paraphrasing, automatic synonymizer, machine translation systems (Kazakh - Russian - English - Ukrainian), systems of automatic opinion analysis (Sentiment analysis), systems of automatic text clustering and other NLP applications.

The Web Application of the Multilingual Basic Ontology "Illegal Web Content" is an online tool that works on any device with a Web browser, including cell phones, tablets, and desktops.

Nouns, verbs, and adjectives are grouped into sets of cognitive synonyms (synsets), each expressing a different concept. The resulting network of meaningfully related words and concepts can be navigated using a browser.

The basic sources of the Web application of the multilingual basic ontology "Illegal Web Content" are the developed parallel Kazakh-Russian corpus, which includes texts containing criminal news.

The basic vocabulary of the thesaurus was obtained manually from crime texts in English, Ukrainian, Kazakh and Russian. Seven basic thematic categories were selected: "Movement", "Traffic Accident", "Injure", "Offense", "Arrest", "Trial" and "Police Department", which allowed making the thesaurus narrowly thematic. This choice of categories is conditioned by the subject matter of the information resources used to fill the thesaurus. The investigated criminal news items are mostly related to the three criminal areas: "Police" (Police), "Transfer" (Traffic), "Crime" (Crime) and their subtypes mentioned above.

The thesaurus includes nouns, verbs, and adjectives of the four parts of speech. Figure 1 shows the structural diagram of the developed thesaurus, whose XML document includes three main elements: <nouns>, <verbs> and <adjectives>, which in turn includes child elements <term>. Each <term> element represents a word of this part of speech and its synonyms in English, Ukrainian, Kazakh and Russian, in the corresponding child elements <lemma> and <synset>. The <domain> element of the dictionary denotes one of the seven thematic categories related to crimes and illegal actions. Each element specified in the <term> tag represents a word of the given part of speech with its synonym series (synsets), definitions, examples, hyponyms and hypernyms in the four languages, represented by the XML child elements.

The most numerous relations between the synsets of nouns is the generic relation, while the specific synset is called a hyponym, and the generic hypernym. This is a transitive hierarchical relationship, similar to the ISA relationship in artificial intelligence research.

Synset *X* is called a hyponym of synset *Y* if native English speakers consider normal sentences like "An *X* is a (kind of) *Y*".

When hierarchical systems are built on the basis of generic relations, it is usually assumed that the properties of superior concepts are inherited onto inferior ones - the so-called inheritance property. Thus, the nouns in the thesaurus are organized as a hierarchical system with inheritance. The developers have made a systematic effort to find for each lemma its generic concept, its hypernym and hyponym.

Figure 1: Structural scheme of the "Crime-related Web content" thesaurus

Figure 2 shows a fragment of the thesaurus, which currently includes about 600 main words (about 330 nouns, about 107 adjectives and 168 verbs) and more than 2500 synonyms of main words. The thesaurus will be supplemented in the next stages of the study, with an extended use of bigrams.

This paper presents the development of a crime ontology based on the Flask web framework and the Anytree library. The Flask web framework is a lightweight web application framework written in Python. It is commonly used for web development, including the creation of RESTful APIs. The Anytree library is a Python library for working with tree data structures. It provides an easy-to-use API for creating, navigating, and modifying tree structures.

The development of the crime ontology involved several steps, including ontology design, ontology implementation, and ontology evaluation. The ontology design involved the identification of relevant concepts and their relationships. The ontology was designed to represent crime-related knowledge and information, including crime types, crime locations, suspects, victims, witnesses, and evidence.

The ontology was implemented using the Flask web framework and the Anytree library. The Flask web framework was used to create a web-based interface for the ontology. The interface allowed users to browse and search the ontology and to add, modify, and delete ontology elements. The Anytree library was used to manage the ontology's hierarchical structure.

The ontology was evaluated using a case study of a criminal investigation. The case study involved a burglary investigation. The ontology was used to represent the crime-related knowledge and information, including the crime location, the suspects, the evidence, and the witnesses. The ontology was also used to identify the relationships between the different elements of the investigation.

The crime ontology developed in this study was effective in representing crime-related knowledge and information. The ontology was able to capture the hierarchical relationships between different elements of the investigation, including the crime location, the suspects, the evidence, and the witnesses.

```
<nouns>
    <term id="1">
        <lemma lang="ru">стрельба</lemma>
        <domain>OFFENSE</domain>
        <synset lang="ru">обстрел, выстрел</synset>
        <definition lang="ru">учебные занятия по ведению огня из различных видов оружия;  ведение огня, применение огнестрельного
оружия для выполнения поставленной задачи (Пулевая с., стендовая с., с. из пистолета, с. из лука)</definition>
        <example lang="ru">Два человека получили ранения при стрельбе в Таразе</example>
        <hypernims lang="ru">['приведение в действие', 'движение']</hypernims>
        <hyponims lang="ru">['контрвыстрел', 'разряд', 'отстрел', 'выстрел', 'выстрел из пистолета', 'выстрел в голову', 'выстрел из
снаряда', 'перестрелка']</hyponims>
        <lemma lang="en">shooting</lemma>
        <synset lang="en">firing, fire, gunfire</synset>
        <definition lang="en">the act of firing a projectile</definition>
        <example lang="en">his shooting was slow but accurate</example>
        <hypernims lang="en">['actuation', 'propulsion']</hypernims>
        <hyponims lang="en">['countershot', 'discharge', 'firing', 'firing off', 'gunfire', 'gunshot', 'headshot', 'potshot',
'shellfire', 'shoot']</hyponims>
        <lemma lang="ka">атыс</lemma>
        <synset lang="ka">ату, оқ жаудыру, атылыс</synset>
        <definition lang="ka">оқ атылғанда шығатын дыбыс, тарсыл; көздесу, нысанаға алып, оқ тигізу</definition>
        <example lang="ka">Алматының Ақбұлақ мөлтекауданында атыс болып, бес адам қаза тапты</example>
        <hypernims lang="ka">['іске қосу','қозғаушы күш']</hypernims>
        <hyponims lang="ka">['қарсы атыс', 'ату', 'басына ату', 'снарядтан ату', 'атыспақ', 'атысу', 'атысып қалу']</hyponims>
        <lemma lang="ua">стрілянина</lemma>
        <synset lang="ua">стрільба, пальба</synset>
    </term>
```

Figure 2: Fragment of multilingual thesaurus with criminal vocabulary.

The ontology was also able to identify the relationships between different elements of the investigation, such as the relationship between the suspects and the evidence.

The web-based interface developed using the Flask web framework allowed users to browse and search the ontology. The interface was user-friendly and intuitive, allowing users to easily navigate the ontology and find the relevant information.

The case study demonstrated the effectiveness of the ontology in facilitating the investigation process. The ontology allowed investigators to quickly identify the relevant information and to identify the relationships between different elements of the investigation. This, in turn, enabled investigators to make more informed decisions and to solve the case more efficiently.

The development of a crime ontology based on the Flask web framework and the Anytree library provides a useful tool for criminal investigations. The ontology can be used to capture and represent

Capabilities of the web application with ontology visualization:

1) Domain search;

2) Multilingual search by keyword;

3) Tree depth;

Thus, the developed web application, shown in Figure 2, has an interface that allows full visualization of the contents of the thesaurus.

A more detailed process of the web application is described in the steps below:

1) Dictionary pre-processing.

The application opens a file with the words in the environment, then converts the file into an operational table with which the system components interact throughout the work. The table is a pandas dataframe with all the fields involved. The fields are accessed by both defined fields and undefined fields - that is, fields that have not been used before. Despite the limited number of fields used on each keyword, adding additional fields as attributes is supported by the application. The application is designed fault-tolerant - missing fields will be empty, so its lemma is sufficient to add a new word.

Data processing is an essential part of modern scientific research, and efficient algorithms are needed to transform raw data into a format suitable for analysis. This paper presents an algorithm for processing data from a raw dictionary file and transforming it into a JSON format file that can be used for further analysis.

Algorithm (The algorithm of dictionary preprocessing is shown in Figure 3.)

Initialize – Load all needed libraries and find the raw dictionary file.

The first step is to load all the necessary libraries, including the library for working with XML files. Then, the raw dictionary file is located and opened.

Open file and transform input into an appropriate format for further processing.

The next step is to read the data from the raw dictionary file and transform it into a format that can be processed by the algorithm. This might involve parsing the data, converting it to a specific data type, or extracting specific information.

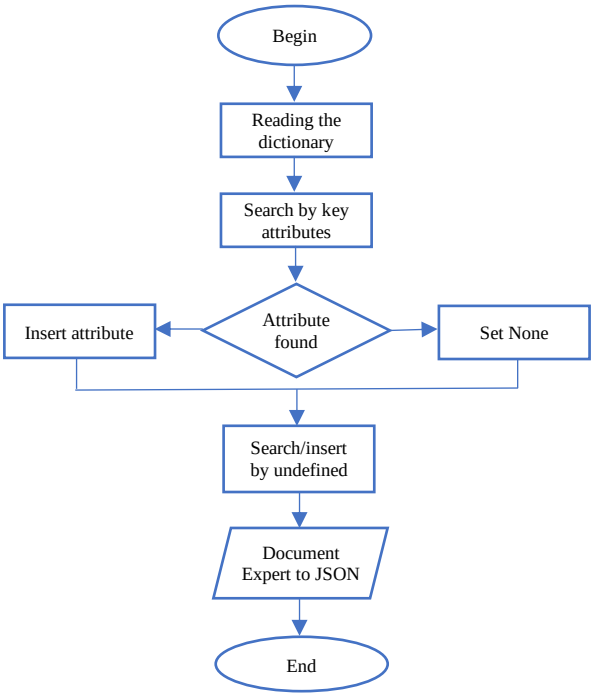


Figure 3: Dictionary preprocessing algorithm

Search by key attributes according to the architecture of the tree.

The algorithm searches for specific attributes in the data using the architecture of the tree. This might involve searching for specific keys or values in the data, or using a more complex search algorithm to find the desired attributes.

If attribute is found then insert needed attribute, if not, set in None.

Once the desired attribute is located, the algorithm inserts the relevant information into the data structure. If the attribute is not found, the algorithm sets the value to None.

Search and insert target attributes.

The algorithm searches for the target attributes in the data and inserts the relevant information. This might involve searching for specific keys or values in the data, or using a more complex search algorithm to find the desired attributes.

Export document to JSON format file which will be processed by other functions in further steps.

Finally, the algorithm exports the processed data to a JSON format file that can be used for further processing.

The algorithm presented in this paper provides a framework for processing data from a raw dictionary file and transforming it into a JSON format file that can be used for further processing. The algorithm is flexible and can be adapted to different data formats and structures, making it a valuable tool for scientific research.

2) Launching the web application.

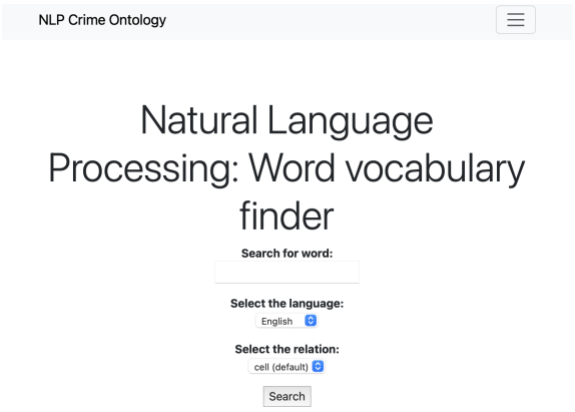
The web application has forms with fields used in the interface. They are described in a separate file with forms, where the client interface is described, part of the design is described in separate files.

The following windows are implemented in the web application:

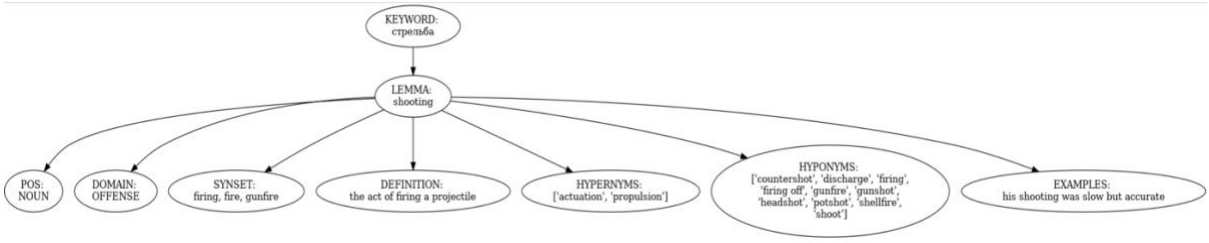
1) Main request page (see figure 4a);
2) Search result page (see figure 4b).

In the scenario where the term entered was not found in the live table, a push notification is sent to the client in the browser (with its support on the client side) and is redirected to the main query page.

The software is built on a client-server architecture. On the client side, a web browser is used where the user opens the web site by accessing the domain name. On the server side, an application built on the Flask framework is used, while Gunicorn is used as the web server itself.



a)



b)

Figure 4: Web-application and its parts: a) main page of the Web application; b) search result page

The "Illegal Web Content" multilingual basic ontology web application is developed in Python and uses a file in XML format as its data source.

5. Conclusions

The complete absence of available ontologies of subject areas shows the current direction of work for the creators of ontologies. It should also be noted that the openness of both such lexical resources themselves and their development models can make it possible to improve the quality of ontologies, as well as make a significant contribution to solving the problems of automatic text processing in the relevant fields.

One approach to generating ontologies is through the use of machine learning methods, which can combine statistical and linguistic approaches. Another approach is through the use of specialized libraries such as the anytree library, which provides a powerful toolset for building, manipulating, and visualizing tree structures, such as those used in ontology construction.

Recently, a flask-based web application has been developed for ontology generation using the any tree library. This application provides an intuitive user interface for constructing and visualizing ontologies, as well as the ability to export ontologies in a variety of standard formats. The anytree library allows for easy manipulation of tree structures and provides a range of tools for constructing ontologies, including support for multilingual ontologies.

Thus, modern research in the field of ontologies is developing in several directions, studying both axiomatic ways of representing knowledge about the world and less formalized methods. The creation of ontologies based on strict formal principles is currently associated with problems of scalability of the description, with problems of understanding by users, and with the existence of other formal points of view on the same sphere of concepts. Therefore, when solving specific applied tasks, especially in broad subject areas, it is necessary to make an informed choice of the level of complexity of the formalism of knowledge representation about the subject area, and to consider available tools and resources, such as the any tree library and flask-based web applications, for ontology construction.

6. Acknowledgements

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

7. References

[1] A. Gomez-Perez, O. Corcho, M. Fernandez-Lopez, Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web, in: First Edition (Advanced Information and Knowledge Processing), Springer-Verlag, 2004.
[2] S. Nirenburg, V. Raskin, Ontological Semantics, MIT Press, 2004.
[3] C.J. Fillmore, R.L Miriam, J.R. Petruck, W. Abby. "Framenet in Action: The Case of Attaching". International Journal of Lexicography 16.3 (2003): 297-332.
[4] R. Adderley, P. Seidler, A. Badii, M. Tiemann, F. Neri, M. Raffaelli, Semantic Mining and Analysis of Heterogeneous Data for Novel Intelligence Insights, Fourth Int. Conf. Adv. Inf. Min. Manag, Volume 1, 2014, pp. 36–40.
[5] P. Casanovas, J. Arraiza, F. Melero, J. González-Conejero, G. Molcho, M. Cuadros. "Fighting Organized Crime Through OpenSource Intelligence: Regulatory Strategies of the CAPER Project". Front. Artif. Intell. Appl., 271 (2014): 189–198.
[6] B. Brewster, S. Andrews, S. Polovina, L. Hirsch, B. Akhgar, Environmental scanning and knowledge representation for the detection of organised crime threats, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 8577 LNAI, 2014, pp. 275–280.
[7] COPKIT project. URL: https://copkit.eu/
[8] ASGARD project. URL: https://www.asgard-project.eu
[9] TENSOR project. URL: https://tensor-project.eu
[10] Old Bailey Corpus, 2020. URL: https://www.oldbaileyonline.org/static/HowToReadTrial.jsp.
[11] COSCO-US (Corpus of US Supreme Court Opinions). URL: https://www.english-corpora.org/scotus/
[12] BLaRC (British Law Report Corpus). URL: https://www.sketchengine.eu/blarc-british-lawreference-corpus/
[13] D. Devyatkin, I. Smirnov, M. Ananyeva, M. Kobozeva, A. Chepovskiy, F. Solovyev, Exploring linguistic features for extremist texts detection, in: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), 2017, pp.188–190.
[14] 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 97.1 (2018): 134–142.
[15] 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.
[16] N. Khairova, S. Petrasova, O. Mamyrbayev, K. Mukhsina, Open Information Extraction as Additional Source for Kazakh Ontology Generation, in: Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, pp. 86–96 https://doi.org/10.1007/978-3-030-41964-6_8
[17] F. Rahma, A. Romadhony, Rule-Based Crime Information Extraction on Indonesian Digital News, in: Proc. of the 2021 International Conference on Data Science and Its Applications (ICoDSA), 2021, pp.10–16.
[18] A.M. Davani et al., Reporting the Unreported: Event Extraction for Analyzing the Local Representation of Hate Crimes, in: Proc. and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5753–5757.
[19] N.S. Mullah, W. M. N. W. Zainon, Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review, IEEE Access, vol. 9, 2021, pp. 88364–88376.
[20] H. Chen et al., COPLINK connect: information and knowledge management for law enforcement, Decision support systems, vol. 34, iss. 3, 2003, pp. 271–285.