







# Open Information Extraction as Additional Source for Kazakh Ontology Generation

Nina Khairova<sup>1</sup> , Svitlana Petrasova<sup>1</sup> , Orken Mamyrbayev<sup>2</sup> ,  
and Kuralay Mukhsina<sup>3</sup> 

<sup>1</sup> National Technical University “Kharkiv Polytechnic Institute”,  
Kyrpychova Street, Kharkiv 61002, Ukraine

nina\_khajrova@yahoo.com, svetapetrasova@gmail.com

<sup>2</sup> Institute of Information and Computational Technologies, 125, Pushkin Street,  
050010 Almaty, Republic of Kazakhstan  
morkenj@mail.ru

<sup>3</sup> Al-Farabi Kazakh National University, 71 Al-Farabi Avenue, Almaty, Republic of Kazakhstan  
kuka\_ai@mail.ru

**Abstract.** Nowadays, structured information that obtains from unstructured texts and Web context can be applied as an additional source of knowledge to create ontologies. In order to extract information from a text and represent it in the RDF-triplets format, we suggest using the Open Information Extraction model. Then we consider the adaptation of the model to fact extraction from unstructured texts in the Kazakh language. In our approach, we identify lexical units that name the participants of the action (the Subject and Object) and semantic relations between them based on words characteristics in a sentence. The model provides semantic functions of the action participants via logical-linguistic equations that express the relations of the grammatical and semantic characteristics of the words in a Kazakh sentence. Using the tag names and some syntactic characteristics of words in the Kazakh sentences as the values of the predicate variables in corresponding equations allows us to extract Subjects, Objects and Predicates of facts from texts of Web content. The experimental research dataset includes texts extracted from Kazakh bilingual news websites. The experiment shows that we can achieve the precision of facts extraction over 71% for Kazakh corpus.

**Keywords:** Open Information Extraction · RDF-triplets · Unstructured text · Logical-linguistic equations · Kazakh bilingual news websites

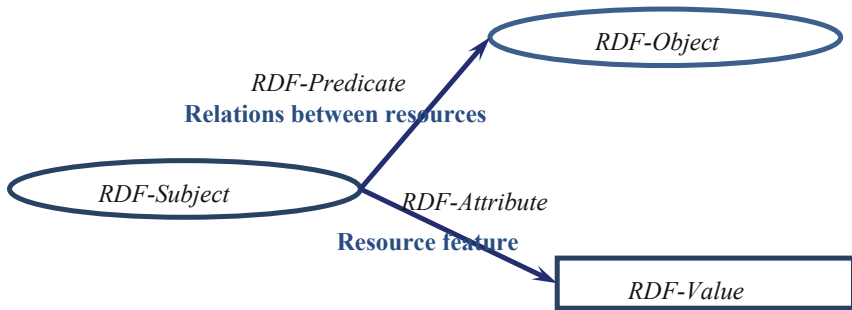
## 1 Introduction

Nowadays, the problem of information and fact extraction remains unsolved. Existing models and algorithms for fact extraction depend on the degree of a document structuring. In this way, we can divide text documents into: (1) well-structured texts, which often content tabular data; (2) semi-structured text documents described a specific domain, and (3) unstructured text document of any domain [1].

Generally, there are robust algorithms [2, 3] for fact extraction from well-structured text documents. At the same time, despite the constant growth of interest in researches

of information extraction from Web content, there is no general and well-grounded approach for structured information extraction from unstructured texts [4, 5]. This growing interest is primarily caused by the huge volumes of unstructured text information available in corporate and Internet networks (according to some sources, there are more than 85% of such texts). Additionally, increasing interest in researches of fact identification and extraction from unstructured texts is largely due to the expansion of areas of their use.

For instants, fact extraction from unstructured texts can be a serious additional source for ontologies generation based on Web content knowledge. Recent approaches of Open Information Extraction (Open IE) extract a fact as a triplet of Subject-Predicate-Object, where the Object and Subject are usually represented by nouns or noun phrases, while the Predicate is mostly expressed by a verb. This view of fact corresponds to an RDF graph (Fig. 1).



**Fig. 1.** The RDF diagram of a fact triplet, which corresponds to the concept of a fact in Open IE approach

Yet, the current approaches to structured information extraction from unstructured texts are either based on a limited number of predefined facts (IE) or use existing NLP tools for each specific language (Open IE). Solving both IE and Open IE tasks, a large labeled corpora of a particular language are required.

In our approach, we propose a logical-linguistic model for fact expression in a sentence of the natural language. This model implements the general approach of Open IE, namely, extraction of the unlimited domain-independent number of facts from texts.

This study focuses on the adaptation of our fact extraction model for the texts of the Kazakh language, which is the language with limited linguistic resources and, obviously, demands additional sources for Kazakh ontologies generation.

In order to estimate the effectiveness of the model, we utilize relatively small experimental corpus included texts extracted from Kazakh bilingual news websites.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related works, corresponding with IE and Open IE challenges. Section 3 describes the usage of our model within the general approach of Open IE and its implementation for the Kazakh language. Section 4 introduces the working corpus and describes its usage in our experiment. In the last Sect. 5, the scientific and practical contributions of the research, its limitations and future work are discussed.

## 2 Related Work

The problem of information extraction from unstructured texts can be divided into two basic approaches: Information Extraction (IE) and Open IE. Both of these technologies allow considering large volumes of texts that contain relatively small amount of factual information.

Herewith IE can be thought as a special kind of Information Retrieval (IR), when the query is formulated in advance. However, IE creates a data structure, describing facts, from a set of processed documents, whereas the result of IR is a set of links to documents that match the query.

First IE systems were mostly domain-oriented and based on the knowledge, generated in advance. The example of such an approach is one of the first IE systems. Working with texts on Latin American terrorism, it exploited pre-developed morphological and semantic patterns [6, 7]. Modern IE systems also use a predefined set of rules to extract information from texts [8]. Mainly, IE systems extract and present information as tuples of two objects with a predefined type of relations [9]. Thus, IE approaches are aimed to create predefined knowledge structures as a result and they do not allow working with Web content of unlimited knowledge texts where the target relations cannot be predetermined [8].

IE technologies usually exploit statistical methods as well as supervised and unsupervised machine learning methods [10]. Recognition of specific domain objects (faces, company names, etc.), parsing and semantic tagging are utilized as well [11, 12].

The new knowledge extraction paradigm that appeared in 2007, Open IE [7], allows identifying an unlimited number of relations and, therefore, does not depend on an application domain. Open IE includes a wide range of tasks: (1) identification and tracking of entities, (2) identification of their relations and characteristics, (3) detection and characterization of events.

The most of Open IE applications use NLP tools such as POS-tagging and Dependency parsing [13, 14], employing lexical restrictions [15] or semantic annotations [16] to minimize the large number of possible specific relations [17].

The reasons for ineffectiveness of statistical methods in solving Open IE problems are as follows. Due to the fact that statistical methods consider the document as an unordered “bag-of-words” in IR and text classification or clustering tasks [18], some knowledge, related to grammar and semantics, is lost. The second reason is the obvious need to extract facts not from the whole text but from sentences. This approach is associated with the fact presentation as a triplet: Subject–Relation–Object. In this paradigm, knowledge of a certain domain is a collection of information about the objects or subjects of this domain, their essential properties and relations presented in separate sentences. The third reason for the low effectiveness of using statistical methods is the synonymy and ambiguity of language units, which leads to the frequent occurrence of hidden facts in the text.

Today, the problem of fact extraction is studied for all languages; it has a high level of implementation not only for English texts but also for many others. For example, an experiment was conducted in [19] for assessing the adequacy of measuring the factual density of 50 randomly selected Spanish documents in the CommonCrawl corpus. In a

recent study [20], densities of simple and complex facts were considered as characteristics of measuring the quality of Russian Wikipedia articles. In [20], the first Open IE system was introduced to extract fact triplets from Chinese texts.

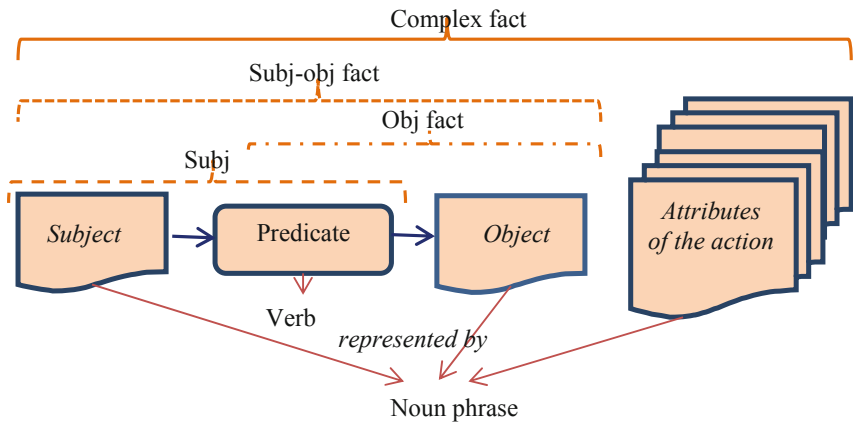
Despite the available research results, however, there are no multilingual standard Open IE methods and approaches [19], in particular, for languages with limited linguistic resources such as the Kazakh language.

### 3 Using the Model Within General Approach of Open Information Extraction

#### 3.1 Mathematical Means of the Model

The Open IE approach extracts triplets of Subject - Predicate - Object 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.

We distinguish four semantic types of facts extracted from the text. Each of them is expressed by different structures of natural languages [21] (Fig. 2).



**Fig. 2.** The scheme of the formalization of four semantic types of facts in unstructured text.

The first semantic type of facts, called *subj-fact*, is expressed by the smallest grammatical sentence. It includes only a noun phrase that defines the Subject as the initiator of the action, and the Predicate that defines the action. The second type of facts, called *obj-fact*, is also identified in the smallest grammatical sentence. It includes the Predicate and a noun phrase defining the participant of the action, i.e. the Object of the action. The third semantic type of facts, called *subj-obj*, is expressed by a sentence including two noun groups that name both the Subject and Object of the action, and the Predicate. The last type of facts is a *complex fact* extracted from a sentence that includes more

than two noun groups. It names the Subject, Object and action Predicate as well as some attributes of the action (time, place, mode of action, etc.).

To set semantic relations in a fact, we suggest applying semantic functions expressed as the ratio of morphological and semantic categories of sentence participants by means of the algebra of finite predicates. Its formulas consist of symbols, predicate variables, signs of disjunction, conjunction, negation, logical constants 0 and 1. The predicate is basic in this algebra for recognizing the subject  $a$  by the predicate variable  $x_i$ :  $x_i^a = 1$  if  $x_i = a$ , and  $x_i^a = 0$  otherwise, where  $i = \{1, 2, \dots, n\}$ ,  $n$  is the number of variables [22].

Identifying an object, the predicate is introduced on a given finite universe  $U$  of elements,  $a \in U$ . In set theory, the universe means the concept of universal set, which contains all the entities. The universe  $U$  of the complex Kazakh language system, considered in our model, includes predefined sets of words, morphemes, collocation characteristics, and many similar discrete and finite linguistic objects. In particular, the universe  $U$  includes a finite, discrete, deterministic set of grammatical and lexical characteristics of words of the Kazakh sentence, influencing their semantic roles  $M = \{m_1, \dots, m_n\}$ , where  $n$  is the number of these characteristics. The relations between these characteristics can be represented as a Cartesian product.

Let us introduce the predicate system  $S$  on the set  $M$  so that any predicate  $P(x_i) \in S$  equals 1 on the set of sentence words with grammatical and semantic information corresponding to a certain semantic role and equals 0 otherwise.

The  $n$ -dimensional predicate  $P(x_1, \dots, x_n)$  defines the semantic role of a participant of the action through subject variables that name the grammatical and semantic characteristics of the sentence:

$$P(x_1, \dots, x_n) \rightarrow P(x_1) \wedge \dots \wedge P(x_n) \quad (1)$$

The predicate  $P(x_1, \dots, x_n) = 1$  if the analyzed word has certain morphological and semantic characteristics of a given language, performing some semantic function. The relations of grammatical characteristics, described by the equation, are independent of a particular word.

In practice, a subset of the coherent morphological, syntactic, and semantic characteristics of the action participants does not coincide with the Cartesian product of all the characteristics.

Then we can define the predicate  $P(x_1, \dots, x_n)$  as:

$$P(x_1, \dots, x_n) = \gamma_k(x_1, \dots, x_n) \times P_1(x_1) \times \dots \times P_n(x_n), \quad (2)$$

where  $k \in [1, h]$ ,  $h$  is the number of participants and attributes of the action. The predicate  $\gamma_k(x_1, \dots, x_n) = 1$  if the conjunction of the grammatical characteristics of the sentence words shows a certain semantic role of the participant (the Subject, Object) or the attribute of the action, and  $\gamma_k(x_1, \dots, x_n) = 0$  otherwise. Thus, if the relations between the grammatical characteristics of the Kazakh sentence words do not express any fact element, they are removed from the formula (2) by the predicate  $\gamma_k(x_1, \dots, x_n)$ .

The semantic functions of the participants and attributes of the action are expressed by the relations of the grammatical and semantic characteristics of the words in the sentence of particular natural language. However, due to the existing difference in morphology and syntax, there are features of the model implementation for each specific language.

We analyzed the implementation of our logical-linguistic OIE model for English [23] and Russian [20] languages. In this study, we consider its adaptation for the texts of the Kazakh language.

### 3.2 Implementation of the Model for the Kazakh Language

Adapting the developed OIE model for fact extraction from Kazakh texts, we introduce the irreducible set  $M$  of ten predicate variables that define the grammatical and semantic features of sentence words. They affect the semantic role of action participants. Most of these features are expressed by affixes in the language structure.

The Kazakh language model is represented by a large number of predicate variables due to the agglutinateness of the language. This means that each word-forming morpheme has its own specific morphological or semantic meaning (for example, person, case, number). Using a large number of predicate variables in the model is also based on the need to distinguish not only action participants but also different types of actions (Action or Predicate) in the Kazakh language.

Table 1 shows the predicate variables and their values ranges defined in the model.

Using the set of predicative variables  $\{x, f, z, a, n, c, y, d, m, b\}$  introduced for the Kazakh language, we can transform Eq. (2) to the following form:

$$P = \gamma_k \times P_x(x) \times P_y(y) \times P_z(z) \times P_f(f) \times P_m(m) \times P_n(n) \times P_a(a) \times P_b(b) \times P_c(c) \times P_d(d) \quad (3)$$

Then, we define the predicate of the action initiator or the Subject of the fact as  $\gamma_{1K}$ :

$$\gamma_{1K} = (x^1 \vee x^2 \vee x^3)z^{Nom}(c^{tar} \vee c^{ter} \vee c^{dar} \vee c^{der} \vee c^{lar} \vee c^{ler} \vee c^0) \quad (4)$$

The semantic role of the Object of the fact in the Kazakh phrase, i.e. the person or object of the action is defined as  $\gamma_{2K}$ :

$$\gamma_{2K} = (x^0 \vee x^2 \vee x^3)(z^{Gen} \vee z^{Acc})(y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^0) \wedge (c^{tar} \vee c^{ter} \vee c^{dar} \vee c^{der} \vee c^{lar} \vee c^{ler} \vee c^0)a^{NSim} \quad (5)$$

Forming the logical-linguistic equation of the Predicate of the action in the Kazakh phrase is based on the definition of the fact. According to the definition, a fact is a real, concrete single event that happened or will happen. Thus, we consider only the indicative mood of verbs and do not take into consideration the imperative, optative, conditional moods that exist in the Kazakh language.

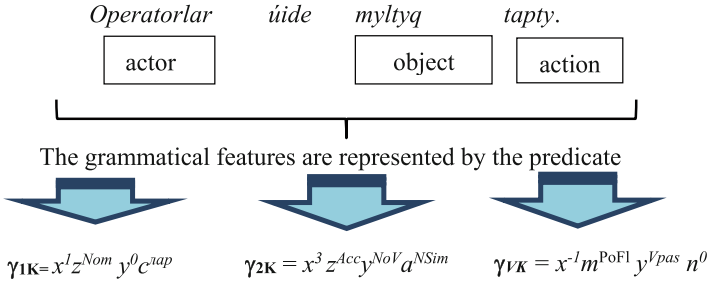
The predicate  $\gamma_{VK}$  defines a combination of semantic and grammatical features of the central part of a fact triplet, namely an action or a fact Predicate:

$$\gamma_{VK} = (x^{-1} \vee x^{-2} \vee x^{-3})((f^{tur} \vee f^{otur} \vee f^{jaty} \vee f^{jur})m^{PrFzVad} \vee (y^{Oad} \vee y^{FuCo})m^{PrFl} \vee y^{FuCo}(m^{PrFl} \vee m^{PrFlfedi}) \vee y^y(f^{edi} \vee f^{eken}) \vee (y^{Vad}m^{PrFl}(p^{mic} \vee p^0)) \vee m^{PoFl}((y^{Vart} \vee y^{Vpa} \vee y^{Vpas}) \vee f^{edi}(n^{joq} \vee n^{emes} \vee n^{me} \vee n^0) \vee (y^{Part} \vee y^{Vad} \vee f^{otur} \vee f^{tur} \vee f^{jaty} \vee f^{jur} \vee f^{ParP} \vee f^{UnFu}))) \quad (6)$$

**Table 1.** The predicate variables and their values ranges defined in the Open IE model for the Kazakh language

Variables	Features	Values
<i>x</i>	The location of the analyzed word in a phrase	Shows a word position in a sentence, “minus” means the start of the count from the end of the sentence; 0 shows any other position of the word except the first three and the last three words in the sentence
<i>f</i>	The feature of an auxiliary verb in the phrase	<i>aux</i> shows the existence of any of 35 auxiliary verbs of the Kazakh language in the analyzed phrase
<i>z</i>	The grammatical case of the Kazakh noun	<i>Nom</i> – nominative, <i>Gen</i> – genitive, <i>Dat</i> – dative, <i>Acc</i> – accusative, <i>Ela</i> – local, <i>Ins</i> – instrumental, <i>Abl</i> – ablative
<i>a</i>	The types of the Kazakh nouns declensions	<i>NSim</i> is a simple declension of nouns, <i>NPos</i> is a possessive declension of nouns
<i>n</i>	The feature of the negative sentence	<i>me</i> and <i>emes</i> are signs of a negative sentence, represented by two different lists of words or particles
<i>c</i>	The feature of plural suffixes	<i>tar, ter, dar, der, lar, ler</i> show the presence of a plural suffix with the same name in the analyzed word
<i>y</i>	The derivational suffixes for verbs, nouns, participles, adverbials	<i>UnFu, FuCo</i> are features of a suffix of uncertain future tense and future conjecture tense in the analyzed word; <i>Psuf</i> and <i>Usuf</i> are features of one of 189 productive or one of 65 unproductive suffixes from specific lists in the analyzed verb; <i>NoN, NoV, Ncom, Nder</i> are features of the noun generation ( <i>NoN</i> – from a noun, <i>NoV</i> – from a verb, <i>Nder</i> is a feature of some expression); <i>Part, ParP</i> are features of the participle generation by means of two different lists of suffixes; <i>VaP, Oad, Vad</i> are features of the verbal participle generation by means of three different lists of suffixes; <i>Vpas</i> is a feature of one of 20 verb suffixes in the analyzed word; <i>y</i> is a sign of the existence of suffix of the infinitive verb form; 0 is a sign of a verb stem
<i>d</i>	The subjunctive action of the analyzed verb	<i>shi</i> shows a suffix of the subjunctive in the analyzed verb and 0 shows lack of such suffixes
<i>m</i>	A personal predicative or possessive flexion of the analyzed verb and verbal forms	<i>PrFl/PoF</i> show a personal predicative/possessive flexion of analyzed participles, verbal adverbs, main and auxiliary verbs
<i>b</i>	The supplementary semantics of the analyzed action	<i>mic</i> denotes the guessed action, <i>se</i> denotes the conditional mood and 0 denotes the lack of some supplementary semantics of the analyzed verb

Figure 3 shows an example of the model implementation for the Kazakh sentences. In the Kazakh phrase “Operatorlar úide myltyq tapy”, according to formula (6), the verb “tapy” represents an action (past perfect tense). According to Eq. (5), the noun “Operatorlar” is identified as the subject of the action or the subject of the fact. The predicate  $\gamma_{2K}$  (5) identifies the noun “mylty”, as an object of the fact.



**Fig. 3.** An example of the fact identification in the Kazakh phrase. The predicate  $\gamma_{1K}$  defines the grammatical features of the Subject action, the predicate  $\gamma_{2K}$  defines the Object and  $\gamma_{VK}$  is the Predicate of the fact.

## 4 Source Data and Experimental Results

The experimental research dataset includes a pilot parallel corpus of Russian-Kazakh texts extracted from Kazakh bilingual news websites *inform.kz*, *azattyq.org*, *patrul.kz*, *zakon.kz*, *caravan.kz*, *lenta.kz*, *nur.kz* by the parser, based on the Python BeautifulSoup library. Information collection time: from June 2018 to June 2019. The choice of sites is grounded on: (1) the reliability of the sites; (2) the ability to select specialized criminal texts; (3) the ability to switch between Kazakh and Russian languages. The volume of the corpus is about 500 thousand words. Since this study considers the implementation of the logical-linguistic model for the Kazakh texts, in the experiment we used only the Kazakh part of the corpus, which includes about 225 thousand words.

In order to get the values of the subject variables of formulas (4)–(6), tokenization and POS-tagging of texts were carried out. Tokenization was conducted by the `tokenize` module of the NLTK Python library. For POS-tagging of Kazakh texts, we developed a tagger based on the `RegexpTagger` class of the NLTK Python package. Figure 4 shows a fragment of a regular expression that allows identifying some forms of nouns in Kazakh sentences.

```
patterns=[(r'.*бeн$', 'NN'), (r'.* пeнeн$', 'NN'), (r'.* бaсшылыќ$', 'NN'), (r'.*
iпкoнy$', 'NN'), (r'.* тapмeн$', 'NN'), (r'.* гepлepмeн$', 'NN'), (r'.* здap$', 'NN')]
```

**Fig. 4.** A fragment of a regular expression that allows identifying some noun forms in Kazakh sentences.

Using the tag names and some syntactic characteristics of words in the Kazakh sentences as the values of the predicate variables in corresponding equations allows us to extract Subjects, Objects and Predicates of facts from texts of Web content.

Since the training corpus is created using the model, only the precision indicator was used to evaluate the effectiveness of the model. To assess the results of the model, approximately a thousand facts were randomly selected from the list of facts that were automatically extracted from the corpus. The expert evaluated the extracted fact as true if the fact triplet was identified correctly and false otherwise. A fact is considered to be



correct if all three elements of the fact were identified correctly: the initiator of the action is the Subject, the object or targeted person of the action is the Object, and the Predicate names the action and unites its participants. If at least one of the three elements of the fact was detected incorrectly, the expert assessed this fact as false.

In addition, to identify how well two annotators made the same annotation decision for a certain fact, the inter-annotator agreement was measured according to Cohen's Kappa [24].

Table 2 shows the obtained precision and agreement of the developed model for the Kazakh text corpus.

**Table 2.** Evaluation of the experimental results.

Language	Size, words	Precision	Agreement
Kazakh	225 000	71.0%	0.72

## 5 Conclusions and Future Works

The main result of this research is the adaptation of the developed logical-linguistic model for fact triples extraction from unstructured Kazakh texts. This model, created within Open IE approach, allows extracting the unlimited domain-independent number of facts from sentences of the Kazakh Web content.

Representing the structured information extracted in our model as the Subject-Predicate-Object fact allows exploiting it to form automatic RDF triplets, i.e. automatic ontology generation. In this case, in the ontology RDF graph, the word, whose semantics is described by Eq. (6), will form the RDF-Predicate, the noun corresponding to Eq. (4) will form the RDF-Subject, and the noun described by Eq. (5) will represent the RDF-Object of the triplet.

Extracted from Kazakhstan news websites, the constructed text corpus and conducted experiment show that the precision of the model is more than 71%, with the agreement coefficient of about 72%. The precision is thought to be slightly increased by improving the results of POS-tagging of texts.

In future studies, we intend to formulate and experimentally verify the logical-linguistic equations that identify the attributes of the fact in the Kazakh sentence, such as time, place of the action, etc. This problem is a more complex challenge than the fact core formation (the Subject-Predicate-Object triplet). The reason is the lack of strict determinacy of grammatical features expressing semantics of the fact attributes. We can assume that the solution to this problem will require the integrated use of regular expressions and logical-linguistic equations.

In addition, to increase the experimental reliability, our further work will extend the domain of the texts studied and compare the results of the model implementation for Russian and Kazakh texts of the constructed parallel corpus.

## References

1. Sint, R., Schaffert, S., Stroka, S., Ferstl, R.: Combining unstructured, fully structured and semi-structured information in semantic wikis. In: Proceedings of the 4th Semantic Wiki WorkShop (SemWiki) at the 6th European Semantic Web Conference, ESWC (2009)
2. Crestan, E., Pantel, P.: Web-scale knowledge extraction from semi-structured tables. In: WWW 2010 Proceedings of the 19th International Conference on World Wide Web, pp. 1081–1082 (2010)
3. Wong, Y.W., Widdows, D., Lokovic, T., Nigam, K.: Scalable attribute-value extraction from semi-structured text. In: 2009 IEEE International Conference on Data Mining Workshops, pp. 302–307 (2009)
4. Phillips, W., Riloff, E.: Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP) (2002)
5. Jones, R., Ghani, R., Mitchell, T., Riloff, E.: Active learning with multiple view feature sets. In: ECML 2003 Workshop on Adaptive Text Extraction and Mining (2003)
6. ARPA. Proceedings of the 3rd Message Understanding Conference (1991)
7. Etzioni, O., Banko, M., Soderland, S., Weld, D.: Open information extraction from the web. *Commun. ACM* **51**(12), 68–74 (2008)
8. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, pp. 1535–1545 (2011)
9. Duc-Thuan, V., Ebrahim, B.: Open information extraction. In: Encyclopedia with Semantic Computing and Robotic intelligence, vol. 1, no. 1 (2016)
10. Shinzato, K., Sekine, S.: Unsupervised extraction of attributes and their values from product description. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, pp. 1339–1347 (2013)
11. Liu, L., Ren, X., Zhu, Q., et al.: Heterogeneous supervision for relation extraction: a representation learning approach. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 46–56 (2017)
12. Wang, X., Zhang, Y., Chen, Y.: A survey of truth discovery in information extraction (2018)
13. Gamallo, P., Garcia, M., Fernandez-Lanza, S.: Dependency-based open information extraction. In: Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, pp. 10–18 (2012)
14. Akbik, A., Loser, A.: KrakeN: N-ary facts in open information extraction. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pp. 52–56 (2012)
15. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545 (2011)
16. Angeli, G., Premkumar, M.J., Manning, C.D.: Leveraging linguistic structure for open domain information extraction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pp. 344–354 (2015)
17. Gashteovsk, K., Gemulla, R., Del Corro, L.: MinIE: minimizing facts in open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2630–2640 (2017)
18. Mooney, R.J., Bunescu, R.: Mining knowledge from text using information extraction. *ACM SIGKDD Explor. Newslett.* **7**(1), 3–10 (2005). Natural language processing and text mining
19. Gamallo, P., Garcia, M.: Multilingual open information extraction. In: Portuguese Conference on Artificial Intelligence, pp. 711–722 (2015)

20. Khairova, N., Lewoniewski, W., Węcel, K.: Estimating the quality of articles in russian wikipedia using the logical-linguistic model of fact extraction. In: Abramowicz, W. (ed.) BIS 2017. LNBIP, vol. 288, pp. 28–40. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-59336-4\\_3](https://doi.org/10.1007/978-3-319-59336-4_3)
21. Khairova, N., Lewoniewski, W., Węcel, K., Orken, M., Kuralai, M.: Comparative analysis of the informativeness and encyclopedic style of the popular web information sources. In: Abramowicz, W., Paschke, A. (eds.) BIS 2018. LNBIP, vol. 320, pp. 333–344. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93931-5\\_24](https://doi.org/10.1007/978-3-319-93931-5_24)
22. Khudhair, A.T.: The intelligence theory mathematical apparatus formal BASE. *Adv. Inf. Syst.* **1**(1), 38–43 (2017)
23. Khairova, N.F., Petrasova, S., Gautam, A.P.S.: The logical-linguistic model of fact extraction from English texts. In: Dregvaite, G., Damasevicius, R. (eds.) ICIST 2016. CCIS, vol. 639, pp. 625–635. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46254-7\\_51](https://doi.org/10.1007/978-3-319-46254-7_51)
24. Regneri, M., Wang, R.: Using discourse information for paraphrase extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 916–927 (2012)