cogent engineering

COMPUTER SCIENCE | RESEARCH ARTICLE

Logical-linguistic model for multilingual Open Information Extraction

Nina Khairova[1], Orken Mamyrbayev[2]*, Kuralay Mukhsina[3] and Anastasiia Kolesnyk[1]

Received: 23 October 2019
Accepted: 05 January 2020
First Published: 14 January 2020

*Corresponding author: Orken Mamyrbayev, Computer Science, Institute of Information and Computational Technologies, Kazakhstan
E-mail: morkenj@mail.ru

Reviewing editor:
Saurabh Pratap, Mechanical Engineering, Indian Institute of Information Technology Design and Manufacturing, India

Additional information is available at the end of the article

Abstract: Open Information Extraction (OIE) is a modern strategy to extract the triplet of facts from Web-document collections. However, most part of the current OIE approaches is based on NLP techniques such as POS tagging and dependency parsing, which tools are accessible not to all languages. In this paper, we suggest the logical-linguistic model, which basic mathematical means are logical-algebraic equations of finite predicates algebra. These equations allow expressing a semantic role of the participant of a triplet of the fact (Subject-Predicate-Object) due to the relations of grammatical characteristics of words in the sentence. We propose the model that extracts the unlimited domain-independent number of facts from sentences of different languages. The use of our model allows extracting the facts from unstructured texts without requiring a pre-specified vocabulary, by identifying relations in phrases and associated arguments in arbitrary sentences of English, Kazakh, and Russian languages. We evaluate our approach on corpora of three languages based on English and Kazakh bilingual news websites. We achieve the precision of facts extraction over 87% for English corpus, over 82% for Russian corpus and 71% for Kazakh corpus.

Subjects: Computer Science; Algorithms & Complexity; Computer Engineering; Computer Science; General

Keywords: Open Information Extraction; fact extraction from unstructured texts; Kazakh bilingual news websites; criminal subject; logical-linguistic model; finite predicates algebra

ABOUT THE AUTHOR

Nina Khairova

Nina Khairova has been working as a Professor in the Department of Intelligent computer systems of National Technical University “Kharkiv Polytechnic Institute”. She makes research in the international team of scientists from Kazakh and Ukraine that concerns the problems of Open Information Extraction from a large number of unstructured texts in different languages. The main research activity of the international scientific group is the analysis of the criminal contained information, that unstructured texts of social networks, blogs and the others web-sources can contain.

PUBLIC INTEREST STATEMENT

Nowadays, computer networks contain a huge amount of texts, which in turn can provide an abundance of information. Information found in the text can be transformed into the structured format: the target facts, objects, and the relations useful for further automatic processing. However, in order to extract this information and represent it in the formal structure of the facts, special methods and approaches are required. This problem of the information extraction from a text can be considered both as the artificial intelligence task and the natural language processing task. This paper suggests the logical-linguistic model, which basic mathematical means are logical-algebraic equations of finite predicates algebra. The model allows extracting the unlimited number of domain-independent facts from Web-content of different languages.

1. Introduction

In recent years, there has been a growing interest in the research included in the overall artificial intelligence task, which focuses on the ways of Information Extraction (IE), Open Information Extraction and Fact Extraction from unstructured and semi-structured texts. Such technologies for information extraction from texts in natural language allow automatically looking through a large number of the texts containing a small amount of required information. Information found in the text will be transformed into the structured format: the target facts, objects and the relations useful for further automatic processing are identified (statistical processing, visualization, the search of patterns in data, etc.). The results of such studies can be utilized for enhancing machine reading by creating knowledge bases in Resource Description Framework (RDF) or ontology forms.

Unfortunately, the most part of the IE approaches is able to handle only a limited number of facts types. Current Open IE approaches based on such NLP techniques as POS tagging and dependency parsing, depend on the availability of the special methods and tools for each particular language.

We propose the Open IE model that allows extracting the unlimited domain-independent number of facts from sentences of different languages. The core of the model is not dependent on the language; however, it requires the implementation of the constructed logical-linguistic equations for each particular language. These equations must represent how morphological and semantic words features in a particular language express relations between the participants and attributes of an action.

The remainder of the paper is organized as follows: Section 2 gives an overview of the related works, corresponding challenges, motivations, and derived researches questions associated with Information Extraction and Open Information Extraction. Section 3 describes our approach to Open IE. We present the basic mathematical means of the model in Subsection 3.1 and the implementation of our model for the English, Russian, and Kazakh languages in Subsections 3.2, 3.3 and 3.4 accordingly. Section 4 introduces the corpora and describes them corresponding to the usage in our experiments. In the last Section 5, there are discussions about scientific and practical contributions of the research, its limitations, and future work.

2. Related work

Sometimes IE is considered as a specific kind of information retrieval. At the same time, differences between IE and IR lie in the fact that inquiries are known in advance and, as a result of IE, there is created structure of data describing the relevant facts from a set of documents, while IR gets a set of references to documents.

Usually, a base of such systems is a set of text extraction rules that identifies the information to be extracted. In a structured text, the rules specify a fixed order of relevant information and the labels or HTML tags that identify strings for further extraction. But the IE system needs more steps in addition to extraction rules when it comes to free texts (Fader, Soderland, & Etzioni, 2011).

Typically, such IE systems include several tasks, namely: (1) Name Entity Recognition (NER); (Duc-Thuan & Bagheri, 2016) the process of determining whether two noun phrases refer to the same real-world entity or concept (Zhou, Qian, & Fan, 2010); cross-document co-reference resolution (Bondarenko & Shabanov-Kushnarenko, 2007); semantics role recognition (Khairova, Petrasova, & Gautam, 2016); entity relation recognition (finding the relation between entities and the relation that is possibly written with a semantic role) (Starostin, Bocharov, & Alexeeva, 2016). For the task of fact extraction from free texts, statistical methods, and learning methods are even more often used (Shinzato & Sekine, 2013), (Liu et al., 2017), (Wang, Zhang, & Chen, 2018). It is prevalent to use additional integration of syntactic analysis, semantic tagging, and recognizers for domain objects (person, company names, etc.) in IE systems.

However, the IE approach is based on the produce of a set of target knowledge structures as output. The most IE systems extract and represent information in a tuple of two entities and a given type of relationship between them. Usually, there are several predefined types of these relations in a specific preselected domain (Duc-Thuan & Bagheri, 2016). This approach does not scale corpora where the number of target relations is very large or where the target relations cannot be specified in advance (Fader et al., 2011).

At the same time, the Open IE system identifies an unlimited number of relations, which are domain-independent. The task of Open Information Extraction subsumes a broad range of tasks, including entity detection and tracking, Relation Detection and Characterization (RDC), and event detection and characterization (Zhou et al., 2010).

Several years ago, Open Information Extraction became a novel extraction paradigm that tackles an unlimited number of relations, eschews domain-specific training data and scales linearly (Etzioni, Banko, Soderland, & Weld, 2008), (Schmitz, Bart, Soderland, & Etzioni, 2012). In contrast to traditional IE systems, Open IE systems extract facts, which are usually represented in the form of surface subject-relation-object triples. Open IE was introduced by Banko et al. in 2007 (Etzioni et al., 2008). Since then, many different Open IE systems have been proposed. The most part of them was based on NLP techniques such as POS tagging and dependency parsing (Gamallo, Garcia, & Fernandez-Lanza, 2012), (Akbik & Loser, 2012). These systems tried to avoid overly specific relations by using lexical constraints (Fader et al., 2011) and delete all sub constituents connected by certain typed dependencies (Angeli, Premkumar, & Manning., 2015) or use minimized extractions with semantic annotations (Gashteovski, Gemulla, & Del Corro, 2017).

However, unfortunately, to date, there are not such NLP techniques for all languages. Open IE is a challenge for low resourced languages. The multilingual methods need to be developed for many of such languages (Gamallo & Garcia, 2015).

In our study, we suggest the logical-linguistic model that allows extracting facts from the texts of different Web-resources in different languages.

Recently, the interest in the researches, which focuses on the ways of the identification and extractions of the facts in unstructured texts, has been growing constantly. This is due to the recent proposal to utilize a statistical measure called factual density to assess the quality of content and indicate the informativeness of a document from the Internet (Khairova, Lewoniewski, Węcel, Mamyrbayev, & Mukhsina, 2018), (Lex et al., 2012).

The problem of the fact extraction is very popular absolutely for all languages and it has a high level of realization not only for English. For example, Horn at el. conducted experiments to estimate the adequacy of the application of factual density to the informativeness of 50 randomly selected documents in the Spanish language from CommonCrawl corpus (Nivre, 2016). In the recent study (Khairova, Lewoniewski, & Wecel, 2017), densities of simple and complex facts as features to measure the quality of articles in Russian Wikipedia were considered. The study (Yuen-Hsien Tseng et al., 2014) presents the first Chinese Open IE system that is able to extract entity-relation triples from Chinese free texts.

Traditionally, we consider a fact as a triplet: Subject—Predicate—Object, where the Predicate expresses some semantic action, the Subject expresses a doer of the action and the Object expresses some participant of the action, for which the action is aimed. Usually, the Predicate is represented by a verb; nouns or noun phrases represent the Object and the Subject. Additionally, a fact can comprise a few attributes like time, location, mode of action and belonging (or possessing) and others, which may be represented by noun phrases too. In this context, when we use "Subject" we mean a doer of the action and we use "Object" that involves an entity or a person who the action is aimed at.

3. Our approach for multilingual open information extraction

3.1. Basic mathematical means of the model

Basic mathematical means of our model are logical-algebraic equations of the finite predicates algebra. We input *U* as a universe of elements that contains various elements of the language system: sentences, phrases, words, grammatical and semantic features, collocations features, etc. Based on the fact that the sets of the elements regarded U are finite and determined, we can say that the universe is finite and determined (Bondarenko & Shabanov-Kushnarenko, 2007).

The set M = {m_1, …, m_n} is a subset of grammatical and semantic features of words in sentences of a particular language, where *n* is a number of system characters.

Variable x_i^a is called a predicate variable and it describes whether there is a particular grammatical feature *a* of the word *i*. According to the algebra of predicates, x_i^a equals 1 if the word *i* possesses the particular grammatical feature *a* and it equals 0 otherwise:

$$x_i^a = \begin{cases} 1, & if\ \ x_i = a \\ 0, & if\ \ x_i \neq \end{cases} \quad (1 \le i \le n), \tag{1}$$

For instance, the equation $x_i^{gen} = 1$ means that the grammatical case of the word *i* is *genitive* and the equation $x_i^{gen} = 1 \vee x_i^{nom} = 1$ means that the word *i* has either genitive or nominative case.

The next step, we input the system of predicates *S*. In our model, the predicate $P_i(x_i) \in S$ equals 1 if the grammatical and semantic features belong to the word that can be a part of a triplet. The predicate $P_i(x) = 0,$ otherwise. Multi-place predicate $P(x_{i,\ldots,}x_n)$ defines the semantic role of a noun via predicate variables that describe grammatical features of the words in a sentence:

$$P(x_{1,\ldots,}x_n) \rightarrow P(x_1) \wedge \ldots \wedge P(x_n) \tag{2}$$

The predicate $P(x_{1,\ldots,}x_n)$ if the features of the nouns in the sentence have certain values. It means that a word, which conjunction of grammatical features is described by the predicate (2), represents the participant (the Subject or the Object) or attributes of the action. It is obvious that relations between morphological and syntactic features of the noun do not depend on the particular token.

In practice, the subset of agreed morphological and syntactic features of the action participants does not coincide with a Cartesian product over the set of all features. Let us define the predicate over a Cartesian product $S \times \mathrm{S}$:

$$P(x_{1,\ldots,}x_n) = \gamma_k(x_{1,\ldots,}x_n) \times P_1(x_1) \times \ldots. \times P_n(x_n), \tag{3}$$

here $K \in [1, h]$ where *h* is the number of considered participants and attributes of facts in the model. The predicate $\gamma_k(x_{1,\ldots,}x_n) = 1$ if the certain morphological and syntactic characteristics of the sentence words express the certain semantical meaning of the participant or attribute of the action, and $\gamma_k(x_{1,\ldots,}x_n) = 0$ if the conjunction of grammatical categories does not represent any semantic role. In this case, if the relationships between morphological and syntactic characteristics of the sentence words do not represent any fact elements, they are excluded from the formula (3) by the predicate $\gamma_k(x_{1,\ldots,}x_n)$.

We provide our model for facts identification and extraction from the English, Russian, and Kazakh text corpora. The semantic cohesion between participants of the action is explicitly expressed by grammatical relations of the words in the sentences in all these languages.

However, with regard to the fact that there is a differentiation among syntax and morphology of the English, Russian, and Kazakh languages, we obtained some distinctions between the implementations of the model for the various languages. The main reason for this differentiation is that semantic cohesion is represented: (1) by the order of words and existence of prepositions in English; (2) by a range of grammatical cases in Russian; (3) and by the order of words as well as a range of grammatical cases in Kazakh.

3.2. The use of model for the English language

Based on the definitions of predicate variables (1), we can input the finite set of grammatical and syntactic features of words in English sentences. The set includes seven variables {*x, z, m, f, n, p*} [5]

The predicate P_z *(z)* identifies the syntactic feature of having a certain preposition after a verb in phrases:

$$P_z(z) = z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee z^{in} \vee z^{between} \vee z^{for} \vee z^{from} \vee z^{over} \vee z^{out}, \tag{4}$$

where predicate variable z^{prep} shows the certain preposition existence after a verb in an English phrase, *prep* = {*to, by, with, about, of, on, at, in, between, for, from, over*} and z^{out} shows the lack of any preposition after the verb.

The predicate P_x *(x)* identifies the order of words in the phrase, in particular, a place of the analyzed noun in the sentence:

$$P_x(x) = x^f \vee x^i \vee x^{kos}. \tag{5}$$

In this equation: x^{f} = 1 if the analyzed noun locates before the main verb in the phrase, x^{l} = 1 if the analyzed noun locates after the main verb and x^{kos} = 1 if it locates after the indirect object.

The subject variable *y* defines the possessive case of a noun via the existence of the apostrophe:

$$P_y(y) = y^{ap} \vee y^{aps} \vee y^{out} = 1, \tag{6}$$

where y^{ap} and y^{aps} show the usage of the apostrophe or apostrophe with s (‘s) at the end of an analyzed word to identify its possessive case; y^{out} shows the lack of any apostrophe at the word end.

The predicate P_m *(m)* identifies whether there is any form of the verb “to be” in the phrase:

$$P_m(m) = m^{is} \vee m^{are} \vee m^{havb} \vee m^{was} \vee m^{were} \vee m^{out}, \tag{7}$$

In this equation, the superscript of the variable *m* identifies the form of the verb “to be” or the lack of it in the phrase. For example, m^{havb} = 1 if there is an auxiliary verb “has been” in the phrase.

The predicates P_f (*f*) and P_n(*n*) identify likewise whether there is any form of modality or negation accordingly in the phrase:

$$P_f(f) = f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee f^{would} \vee f^{out}, \tag{8}$$

$$P_n(n) = n^{not} \vee n^{out}. \tag{9}$$

Additionally, we input the predicate P_p(*p*) that identifies forms of the main verb at the English phrase:

$$P_p(p) = p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II}, \tag{10}$$

here p^{I} = 1 if there is the base form of the verb in the phrase (used as the infinitive form, with or without “to”); p^{ed} = 1 if there is the past form of the verb (used for the past simple tense) in the analyzed phrase; p^{III}, p^{II}, and p^{ing} show the past participle form, the past form of an irregular verb and—*ing* form of a verb in the phrase accordingly.

Table 1 shows the predicate variables and their values ranges defined in the model for the English language. So then, based on the existing Equations (4)–(10) for English, the formula (3) can be converted to the following form:

$$P(x,y,z,m,p,f,n) = \gamma_k P(x,y,z,m,p,f,n) \times P_x(x) \times P_y(y) \times P_z(z) \times P_m(m) \times P_p(p) \times P_f(f) \times P_n(n). \tag{11}$$

The predicate γ_{1E} can define the semantic relation that distinguishes the Subject of the fact or the actor of the action in an English phrase:

Table 1. The predicate variables and their values range defined in the Open IE model for the English, Russian, and Kazakh languages

vari-ables	features	values
	English language	
z	a certain preposition after a verb in phrases	{*to, by, with, about, of, on, at, in, between, for, from, over, out*}
x	the order of words in the phrase	*f, l, kos*
y	the usage of an apostrophe at the end of the analyzed word	*ap, aps, out*
m	any form of the verb “to be” in the phrase	{*is, are, have been, has been, had been, was, were, am, out*}
f	any form of modality in the phrase	{*can, may, must, should, could, need, might, would, out*}
n	the negation in the phrase	*not*—negative phrase, *out*—affirmative sentence
p	a form of the main verb in the English phrase	I, II, III, *ing* end *ed*—four the base forms of the verbs and irregular verbs
	Russian language	
z	six grammatical cases	{*nom*—nominative, gen—genitive, *dat*—dative, *acc* —accusative, *ins*—instrumental, *loc*—prepositional}
x	animacy	*anim*—animate noun, *inan*—inanimate noun
y	noun semantic characteristics	{*device, hum, tool, pc:hum, space, time:moment, time:period, s:loc*}
	Kazakh language	
x	the location of the analyzed word in the phrase	{-1, −2, −3, 0, 1, 2, 3}
f	the feature of an auxiliary verb in the phrase	*aux* shows the existence of any of 35 auxiliary verbs of the Kazakh language in the analyzed phrase, 0
z	the grammatical case of the Kazakh noun	*Nom*—nominative, *Gen*—genitive, *Dat*—dative, *Acc* —accusative, *Ela*—local, *Ins*—instrumental, *Abl*—ablative
a	the types of the Kazakh nouns declensions	*NSim, NPos*
n	the feature of the negative sentence	*me, eme, joq*, 0
c	the feature of plural suffixes	*tar, ter, dar, der, lar, ler*
y	the derivational suffixes for verbs, nouns, participles, adverbials	*UnFu, FuCo, Psuf, Usuf, NoN, NoV, Ncom, Nder, Part, ParP, VaP, Oad, Vad, Vpas, y*, 0
d	the subjunctive action of the analyzed verb	shi, 0
m	a personal predicative or possessive flexion of the analyzed verb and verbal forms	*PrFl, PoF*, 0
b	the supplementary semantics of the analyzed action	*mic, se*, 0

$$\gamma_{1\mathrm{E}}(\mathrm{z},\ \mathrm{y},\ \mathrm{x},\ \mathrm{m},\ \mathrm{p},\ \mathrm{f},\ \mathrm{n}) = \mathrm{y}^{\mathrm{out}}\left(\left(\mathrm{f}^{\mathrm{can}} \vee \mathrm{f}^{\mathrm{may}} \vee \mathrm{f}^{\mathrm{must}} \vee \mathrm{f}^{\mathrm{should}} \vee \mathrm{f}^{\mathrm{could}} \vee \mathrm{f}^{\mathrm{need}} \vee \mathrm{f}^{\mathrm{might}} \vee \mathrm{f}^{\mathrm{would}} \vee \mathrm{f}^{\mathrm{out}}\right) \left(\mathrm{n}^{\mathrm{not}} \vee \mathrm{n}^{\mathrm{out}}\right) \left(\mathrm{p}^{\mathrm{I}} \vee \mathrm{p}^{\mathrm{ed}} \vee \mathrm{p}^{\mathrm{III}}\right) \mathrm{x}^{\mathrm{f}}\mathrm{m}^{\mathrm{out}} \vee (\mathrm{x}^{\mathrm{l}}(\mathrm{m}^{\mathrm{is}} \vee \mathrm{m}^{\mathrm{are}} \vee \mathrm{m}^{\mathrm{havb}}\mathrm{m}^{\mathrm{hasb}} \vee \vee \mathrm{m}^{\mathrm{hadb}} \vee \mathrm{m}^{\mathrm{was}} \vee \mathrm{m}^{\mathrm{were}} \vee \mathrm{m}^{\mathrm{be}} \vee \mathrm{m}^{\mathrm{out}})\ \mathrm{z}^{\mathrm{by}}\right). \tag{12}$$

We can also explicitly distinguish the Object or the participant, which an action is directed at, via the particular disjunction of conjunctions of the subject variables, that identify morphological and syntax features of the English sentence words (4)–(10):

$$\gamma_{2E}(z,\ y,\ x,\ m,\ p,\ f,\ n) = y^{out}\left(n^{not} \vee n^{out}\right)\left(f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee \vee f^{would} \vee f^{out}\right)\left(z^{out}x^{l}m^{out}\left(p^{I} \vee p^{ed} \vee p^{III}\right) \vee x^{f}\left(z^{out} \vee z^{by}\right)\left(m^{is} \vee \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{be} \vee m^{out}\right) \left(p^{ed} \vee p^{III}\right). \tag{13}$$

Apart from the main participants of the action we also distinguish attributes of the fact. They can be the attributes of time, location, mode of action, affiliation with the Subject or the Object, etc.

According to our previous articles, the attributes of the action in English simple sentence can be represented by nouns that were defined by the logical-linguistic Equations [5–6]. For instance, we can distinguish the attribute of time via the predicate γ_{3E} that shows the disjunction of conjunctions of grammatical features of the noun denoting the time of the fact:

$$\gamma_{3\mathrm{E}}(z,\ y,\ x,\ m,\ p,\ f,\ n) = \left(y^{out}z^{on}x^{kos} \vee y^{out}z^{in}x^{kos} \vee y^{out}z^{at}x^{kos}\right)\left(n^{not} \vee n^{out}\right)\left(f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee f^{would} \vee f^{out}\right)\left(z^{out}x^{l}m^{out} \left(p^{I} \vee p^{II} \vee \vee p^{ed} \vee p^{III}\right)\left(m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee \vee m^{be} \vee m^{out}\right) \tag{14}$$

The predicate γ_{4E} identifies spatial relationships. That means that a noun with grammatical features that are relevant to the conjunctions of the predicate denotes the location or direction of the action.

$$\gamma_{4\mathrm{E}}(z,\ y,\ x,\ m,\ p,\ f,\ n) = \left(\left(z^{to} \vee z^{from} \vee z^{between}\right)\left(p^{I} \vee p^{II} \vee p^{ed} \vee p^{III}\right)\left(m^{is} \vee m^{are} \vee m^{havb} \vee \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{be} \vee m^{out}\right)\left(x^{kos} \vee x^{l} \vee x^{f}\right) \vee z^{by} \left(p^{I} \vee p^{II} \vee p^{ed}\right) m^{out}\left(x^{kos} \vee x^{f}\right)\right) y^{out}\left(n^{not} \vee n^{out}\right)\left(f^{can} \vee f^{may} \vee f^{must} \vee \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee f^{would} \vee f^{out}\right)\left(n^{not} \vee n^{out}\right) \tag{15}$$

Figure 1 shows an example of the model implementation for the phrase: "... *the steamer moved slowly from the dock*". In the phrase, the verb "move" identifies the action types as "the movement". Then, according to Equation (12), we can identify the noun "steamer" as the doer of the action or the Subject of the fact. The predicate $\gamma_{4\mathrm{E}}$ (15) provides the direction attribute "from the dock" of the action of the movement in the phrase.

3.3. The use of the model for the Russian language

In the case of the adaptation of our model for Russian text we input the set of grammatical and semantic features of words in Russian sentences $M = \{z,\ y,\ x\}$, where z is the finite subset of morphological features that describes the grammatical cases of Russian nouns, y is the finite subset of semantic features of nouns and x is the finite subset of the characteristic of animacy (Khairova et al., 2017).

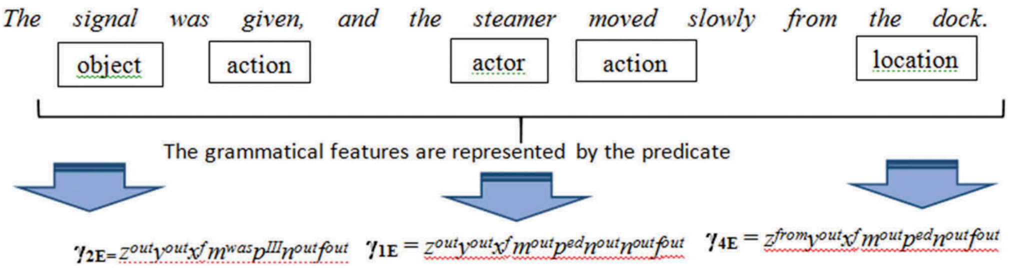
The signal was given, and the steamer moved slowly from the dock.
object
action
actor
action
location
The grammatical features are represented by the predicate
$\gamma_{2E} = z^{out}y^{out}x^{f}m^{was}p^{III}n^{out}f^{out}$
$\gamma_{1E} = z^{out}y^{out}x^{f}m^{out}p^{ed}n^{out}n^{out}f^{out}$
$\gamma_{4E} = z^{from}y^{out}x^{f}m^{out}p^{ed}n^{out}f^{out}$

Figure 1. Example of the fact identification from English sentences. The predicate γ_{1E} defines grammatical features of the Subject, the predicate γ_{2E} defines grammatical features of the Object and γ_{4E} the predicate defines grammatical features of the location attribute of the fact.

The predicate P_z identifies such six cases in Russian as nominative, genitive, dative, accusative, instrumental, and prepositional.

$$P_z(z) = z^{nom} \vee z^{gen} \vee z^{dat} \vee z^{acc} \vee z^{ins} \vee z^{loc} \tag{16}$$

The predicate $P_x(x)$ identifies animacy (antonym is inanimacy), which represents a grammatical and semantic feature, expressing how sentient and alive the referent of a noun is.

$$P_x(x) = x^{anim} \vee x^{inan} \tag{17}$$

The predicate $P_y(y)$ can identify such specific semantic characteristics of the noun as *device, tool* (a noun refers to a concept that belongs to the group of devices or tools), *space, time:moment, time: period* and others[1]:

$$P_y(y) = y^{\text{device}} \vee y^{hum} \vee y^{tool} \vee y^{\text{pc:hue}} \vee y^{\text{space}} \vee y^{\text{time.moment}} \vee y^{\text{time.period}} \vee y^{\text{s:loc}} \tag{18}$$

here index *hum* means belonging to the semantic class “person”, index *pc:hum* means belonging to the semantic class “part of the body” and index *s:loc* means belonging to the semantic class “destination”. We employ meta-marking of taxonomic relations from Russian National Corpus in selecting the semantic indexes of predicate variables of the model. The predicate variables and their values ranges defined in the model for the Russian language are summarized in Table 1.

According to our model, we can define the semantic roles of the Doer and the Object of a fact in Russian sentences via the following predicates γ_{1R} and γ_{2R}, respectively.

$$\gamma_{1R}(x, y, z) = x^{anim} z^{nom} (y^{device} \vee y^{tool} \vee y^{pchue}) \tag{19}$$

$$\gamma_{2R}(x, y, z) = z^{acc} (x^{inam} \vee x^{anim}) \tag{20}$$

We can distinguish the attributes of location, time, destination, beneficiary, and tool of action via logical-linguistic equations in a very similar way. For instance, the predicate γ_{3R} can denote semantic and grammatical features of the action tool or the action reason.

$$\gamma_{3R}(x, y, z) = z^{\text{ins}} z^{inam} (y^{tool} \vee y^{pc:hum} \vee y^{device}) \tag{21}$$

3.4. The use of model for the Kazakh language

Unlike the Russian and English languages, Kazakh is the agglutinative language. It means that a word is composed of morphemes number, each of which has a specific meaning. This is the opposite of inflectional language where every morpheme has several inseparable meanings at once (for example, a case, gender, number, etc.) and analytical language where there are almost no inflexions. When we adjust our model for the Kazakh language, we input the set *M* of ten grammatical features of words in Kazakh sentences, most of which are one or the other types of suffixes bearing particular semantic meaning. They are such features as a position of the analyzed word in a phrase, the existence of an auxiliary verb at the phrase, the presence or absence of plural suffixes, a grammatical case of the analyzed noun, the semantic meaning represented by certain suffixes and some others.

The fact that the set *M* of grammatical features of Kazakh language is much more than a comparable set of grammatical features of the Russian or English languages is connected with two main reasons. First, the cause is the complexity of Kazakh, in which there are a lot of morphological and syntactic characteristics and every feature is usually expressed by a particular affix. The second reason for employing such a large number of grammatical features is that in case of Kazakh we consider and analyze not only participants of the action but also different action types.

The predicate $\mathrm{P_x}$ identifies the location of the analyzed word in a phrase

$$P_x(x) = x^1 \vee x^2 \vee x^3 \vee x^{-1} \vee x^{-2} \vee x^{-3} \vee x^0, \tag{22}$$

where 1, 2, 3, −1, −2, −3 show 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.

The predicate $\mathrm{P_f}$ identifies whether or not there is an auxiliary verb in the phrase:

$$P_f(f) = f^{aux} \vee f^0, \tag{23}$$

where *aux* shows the indication of the existence of any verb from the list of 35 auxiliary verbs of Kazakh language in the analyzed phrase.

The predicate $\mathrm{P_z}$ identifies such seven cases in Kazakh as nominative, genitive, dative, accusative, local, instrumental and ablative:

$$P_z(z) = z^{nom} \vee z^{Gen} \vee z^{Dat} \vee z^{Acc} \vee z^{Ela} \vee z^{Ins} \vee z^{Abl} \tag{24}$$

The predicate $\mathrm{P_a}$ identifies two possible types of Kazakh nouns declensions such as simple and possessive:

$$P_a(a) = a^{\mathrm{Nsim}} \vee z^{\mathrm{NPos}}, \tag{25}$$

where *NSim* is a sign of simple declension of nouns, *NPos* is a sign of possessive declension of nouns. The predicate P_n identifies the feature of the negative sentence:

$$P_n(n) = n^{me} \vee n^{\mathrm{emes}} \vee n^{\mathrm{joq}} \vee n^0, \tag{26}$$

where *me* is a sign of negative sentence, which is represented by the existence of the particle from the list *[ma, me, ba, be, pa, pe]* in the sentence; *emes* and *joq* are a sign of negative sentence, which is represented by the existence of “*emes*” and “joq” words, accordingly, in the sentence; 0 shows the lack of any negation in the sentence.

The predicate $\mathrm{P_c}$ identifies the presence or absence of plural suffixes:

$$P_c(c) = c^{\mathrm{tar}} \vee c^{\mathrm{ter}} \vee c^{\mathrm{dar}} \vee c^{\mathrm{der}} \vee c^{\mathrm{lar}} \vee c^{\mathrm{ler}} \vee c^0, \tag{27}$$

where *tar, ter, dar, der, lar, ler* show the presence of plural suffix with the same name in the analyzed word.

The predicate $\mathrm{P_b}$ identifies the presence of some supplementary semantics or meaning of an analyzed verb:

$$P_b(b) = b^{\mathrm{se}} \vee b^{\mathrm{mic}} \vee b^0, \tag{28}$$

where *mic* denotes the guessed action, *se* denotes the conditional mood and 0 denotes the lack of some supplementary semantics of the analyzed verb.

The next few features are related to semantic meaning represented by certain suffixes. The predicate $\mathrm{P_y}$ identifies derivational suffixes, which establish verbs, nouns, participles, adverbials:

$$\begin{aligned}P_y(y) = {} & y^{\mathrm{ParP}} \vee y^{\mathrm{Vpas}} \vee y^{VaP} \vee y^{UnFu} \vee y^{FuCo} \vee y^{VAd} \vee y^{OAd} \vee y^{\mathrm{Psuf}} \vee y^{Usuf} \vee \\ & \vee y^{Part} \vee y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^{y} \vee y^{0},\end{aligned} \tag{29}$$

where:

- *UnFu, FuCo* are features of inclusion of a suffix of uncertain future tense and future conjecture tense, accordingly, in analyzed word;
- *Psuf* and *Usuf* are features of including of one of 189 productive or one of 65 unproductive suffixes, accordingly, from specific lists in an analyzed verb;
- *NoN, NoV* are features of the noun generation (*NoN*—from a noun, *NoV*—from a verb);
- *Ncom* is a feature of including a complex suffix of noun formation in analyzed word;
- *Nder* is a feature of the existence of some expression (diminutive and derogatory shades);
- *Part, ParP* are features of the participle generation by means of two different lists of suffixes;
- *VaP, Oad, Vad* are features of the verbal participle generation by means of three different lists of suffixes;
- *Vpas* is a feature of including one of 20 verb suffixes in analyzed word;
- *y* is a sign of the existence of suffix of the infinitive verb form;
- *0* is a sign of a verb stem (the form of the second person singular, future imperative time).

The predicate $\mathrm{P_d}$ identifies whether or not there is a subjunctive action of the analyzed verb:

$$P_d(d) = b^{\mathrm{shi}} \vee d^{0}, \tag{30}$$

where *shi* shows the inclusion of a suffix of the subjunctive into the analyzed verb and 0 shows lack of such suffixes.

The predicate $\mathrm{P_m}$ identifies whether there is a personal flexion of the analyzed word:

$$P_m(m) = m^{\mathrm{Pr\,F1}} \vee m^{\mathrm{PoF1}} \vee m^{0}, \tag{31}$$

where *PrFl* shows the occurrence of a personal predicative flexion of analyzed participles, verbal adverbs, main and auxiliary verbs, and *PoFl* shows the occurrence of a personal possessive flexion of analyzed participles, verbal adverbs, main and auxiliary verbs.

All predicate variables and their values ranges defined in the model for the Kazakh language are summarized in Table 1.

For the Kazakh language, according to the Equations (22)–(31), we can convert the predicate of agreed morphological and syntactic features of the words affect a fact represented in formula (3), into the equation:

$$P() = \gamma_k \times P_x(x) \times P_y(y) \times P_z(z) \times P_f(f) \times P_m(m) \times P_n(n) \times P_a(a) \times P_b(b) \times P_c(c) \times P_d(d). \tag{32}$$

We can determine the predicate of the action actor in a Kazakh sentence as following:

$$\gamma_{1k} = (x^1 \vee x^2 \vee x^3) z^{Nom} (c^{\mathrm{tar}} \vee c^{\mathrm{ter}} \vee c^{\mathrm{dar}} \vee c^{\mathrm{der}} \vee c^{\mathrm{ter}} \vee c^{\mathrm{ler}} \vee c^{0}) \tag{33}$$

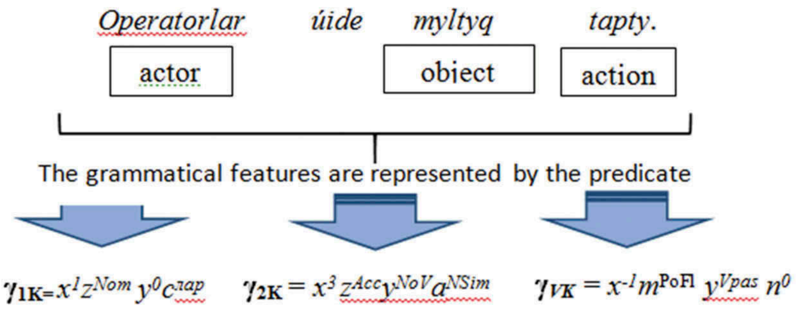
Operatorlar
úide
myltyq
tapty.
actor
obiect
action
The grammatical features are represented by the predicate
$\gamma_{1K}=x^{1}z^{Nom}y^{0}c^{лар}$
$\gamma_{2K}=x^{3}z^{Acc}y^{NoV}a^{NSim}$
$\gamma_{VK}=x^{-1}m^{PoFl}y^{Vpas}n^{0}$

Figure 2. Example of the fact identification from the Kazakh phrase. The predicate γ_1 defines grammatical features of the Doer, the predicate γ_2 defines grammatical features of the Object and γ_{VK} is the predicate of the Action (the Predicate) of the fact.

Then, we can determine the semantic roles of the Object of a fact in Kazakh phrase by means of the following γ_{2K} predicate:

$$\begin{aligned}\gamma_{2K} = {} & (x^0 \vee x^2 \vee x^3)(z^{Gen} \vee z^{Acc})(y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^0)\wedge \\ & \wedge (c^{tar} \vee c^{ter} \vee c^{dar} \vee c^{der} \vee c^{lar} \vee c^{ler} \vee c^0)a^{NSim},\end{aligned} \tag{34}$$

In order to determine the logical-linguistic equation of the formal action in a Kazakh phrase, we base on the hypothesis that a fact is a very real event that really happened or will happen. On this basis, we identify only the indicative mood of the verbs and we do not take into consideration the imperative, optative, conditional moods that exist in the Kazakh language. The predicate γ_{VK} denotes semantic and grammatical features of the key part of the fact triplet, namely the Action or Predicate of the fact:

$$\begin{aligned}\gamma_{VK} = {} & (x^{-1} \vee x^{-2} \vee x^{-3})((f^{tur} \vee f^{otur} \vee f^{jatyr} \vee f^{j\,ur})m^{PrF}z^{Vad} \vee (y^{Oad} \vee y^{FuCo})m^{PrFl}\vee \\ & y^{FuCo}(m^{PrFl} \vee \left(m^{PrFl}f^{edi}\right)) \vee y(f^{edi} \vee f^{eken}) \vee (y^{Vad}m^{PrFl}(p^{mic} \vee p^0))\vee \\ & \vee m^{PoFl}((y^{Vart} \vee y^{Vpa} \vee y^{Vpas}) \vee f^{edi}(n^{joq} \vee n^{emes} \vee n^{me} \vee n^0)\wedge \\ & \wedge (y^{Part} \vee y^{Vad} \vee f^{otur} \vee f^{tur} \vee f^{jatyr} \vee f^{jur} \vee f^{ParP} \vee f^{UnFu})))\end{aligned} \tag{35}$$

Figure 2 shows an example of the model implementation for the Kazakh language. In the Kazakh phrase “Operatorlar úide myltyq tapty”, the verb “tapty” represent action (remote past tense) of finding. Then, according to Equation (33), we can identify the noun “Operatorlar” as the actor of the action or the Subject of the fact. The predicate γ_{2K} (34) identifies the noun “mylty” and provides the Object of the action in the phrase.

Therefore, in order to extract a semantic role of the participant of the action in every phrase and construct a triplet of the fact, we utilize the common approach for three languages. In the first stage, we define every possible grammatical and semantic words characteristics in the sentence. For this, we use available tools analyzing every language. For example, Stanford Dependencies (SD) parser analyses English texts, Russian texts are POS labelled by pymorphy2 Python packet and Kazakh texts are processed via regular expressions. In the next step, according to the obtained logical-linguistic equations (accordingly (12–15) for English, (19–21) for Russian, (33–35) for Kazakh), we define correlations between words characteristics that influence on the express of the semantic role of the participant of the action in the phrase.

4. Implementation aspects and experimental results

Our dataset comprises three corpora of Russian, English, and Kazakh texts. All texts are obtained by means of the developed parser that is based on BeautifulSoup library of Python language from June 2018 to June 2019.

We consider various texts from bilingual websites *inform.kz, azattyq.org, patrul.kz, zakon.kz caravan.kz, lenta.kz, nur.kz in* order to collect text material for Russian and Kazakh corpora. The main reason why we chose these websites for our study is the fact that they are well-known and reliable news websites of Kazakhstan. There are a lot of articles that correspond to the topic of the study—criminal information. And furthermore, these websites can switch text information between two languages: Russian and Kazakh.

In addition, we chose three sources that represent news websites such as edition.cnn.com, news.sky.com and foxnews.com. These websites are popular and up-to-date informational resources and they have reliable articles that correspond to the topic of our study.

As a result of this program, we have received a general set of 6000 texts in three languages: English, Russian, and Kazakh. The corpus size is more than 700 000 words, about 275 000 of them in Russian and about 225 000 words in Kazakh and approximately 200 000 words in English.

Processing English texts, in order to correctly identify links between words we utilize the syntactic dependency relations. We exploit Stanford Dependencies (SD) parser because its treebanks can analyze verb groups, subordinate clauses, and multi-word expressions for many languages most sufficiently. Syntactic relations in SD model are centrally organized around notions of a subject, object, clausal complement, noun determiner, noun modifier, etc. (Nivre, 2016). These relations, which connect words of a sentence to each other, often express some semantic content. In Dependency grammar, a verb is considered to be the central component of a phrase and all other words are either directly or indirectly connected to it. This corresponds to an idea that the verb is a core component of a triplet of the fact and all participants of the action, that are represented by nouns, depend on the Predicate (action), which calls the fact and is represented by a verb.

For our analysis, we used 7 out of 40 grammatical relations between words in English sentences, which UD v1[2] contains. They are *subj, nsubjpass, csubj,obj, iobj, dobj, ccomp. Nsubj* label denotes the syntactic subject dependence on a root verb of a sentence, *csubj* label denotes the clausal syntactic subject of a clause, and *nsubjpass* label denotes the syntactic subject of a passive clause. *Obj* denotes the entity acted upon or which undergoes a change of state or motion. The labels *iobj, dobj* and *ccomp* are used for more specific notation of action object dependencies on the verb.

For example, Figure 3 shows the graphical representation of SD for the part of the sentence "He referred to the deaths as "the cost of doing business on that particular engagement"., which is obtained using a special visualization tool for dependency parser—DependenSee.[3]

Grammatical and semantic characteristics realized through the syntactic relations tags corresponding to the value of the variables in the predicates of the Equations (12)–(15). Table 2 shows an example of facts extracted from the English corpus of texts.

For experimental verification of the model on the Russian and Kazakh corpora, we used POS tagging allowing defining explicitly the value of the relevant substantive variables entered into the models. For Russian corpus labelling, we chose the pymorphy2 Python packet,[4] which is specially developed for morphological analysis of Russian and Ukrainian texts. The libraries of the packet use the OpenCorpora dictionary and make hypothetical conclusions for non-recognized words.

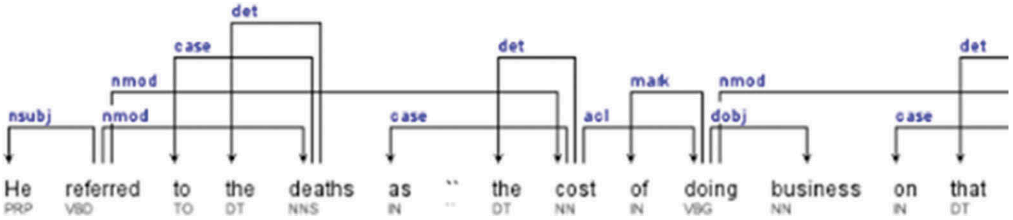

Figure 3. Graphical representation of Universal Dependencies for the sentence "He referred to the deaths as "the cost of doing business on that particular engagement." Source: DependenSee.

Table 2. Fragment of the table with facts extracted from the English corpus of texts

Num. of sent.	Predicate verb	Actor nsubj	Object			
			Advcl	Dobj	Ccomp (object)	Xcomp
1	Consisted	War	Fighting			
2	Lasted, took	War, majority		Place		
3	Focused	Insurgents	Featured, ambushing			
4	Featured	Fighting		Warfare		
5	Killed	Iraqis				Many
6	Saw	Anbar		Fighting		
7	Relinquished	Army		Command		
8	Struggled	Sides				Secure
9	Secure			Valley		
10	Escalated	Violence	Struggled			
11	Became, turned	Qaeda		Capital		Group
12	Issued	Corps	Declaring	Report		
13	Began	Tribes				Form
14	Turn			Tide		
15	Maintained	Forces		Role		
16	Flew	Bush	Celebrating, congratulate			
17	Known	Fallujah	Secular			
18	Held	System		Influence		
19	Favored	Conditions		Insurgency		
20	Lost	That		Power		
21	Claim				Defeated	

In turn, the complexity, structural and typological characteristic of Kazakh marking are connected with the fact that it belongs to agglutinating languages and languages that are difficult for formalization or that do not have enough linguistic resources for the moment. For this reason, we make the POS-tagging of Kazakh texts via the regular expression tagger based on RegexpTagger class of nltk Python package.

For example, we can identify some types of nouns in Kazakh texts via the following list of regular expressions

```
patterns=[(r’.*бен$','NN'), (‘r.* пенен$','NN'), (‘r.* басшылық$','NN'),
(r’.* іпқону$','NN'), (r’.* тармен$','NN'), (r’.* герлермен$','NN'),
(r’.* здар$','NN')]
```

Additionally, to increase recall and precision of our POS-tagging of Kazakh texts we combine regular expressions with the system including several rules. For instance, “If a word followed by words from the special list—the word is marked as Verb”.

In our experiments, we use precision to assess the validity of our approach. The main reason why we could not evaluate recall of the results of the experiment is that we did not have a training corpus with correctly identified triplets of fact. In order to obtain the number of correctly found facts or triplets of “*Subject- Predicate- Object*” in the corpora of three different languages, we use an expert opinion. We interviewed two experts, which native language corresponded to the language of the corpus, for every language.

Table 3. Precision and agreement of the developed logical-linguistic model for the corpora of three different languages

The language of corpus	The size of corpus	Precision of the model	Alignment
English	200 000	87,2 %	0,91
Russian	275 000	82,4 %	0,78
Kazakh	225 000	71,0 %	0,72

About 1000 automatically determined facts were randomly extracted from each list of three corpora of different languages and presented for consideration. The purpose of the evaluation was to obtain opinions on whether the triplets of “Subject- Predicate-Object” found in the texts were correct. The experts needed to assess a fact as 1 if it was correct and 0, otherwise. A fact is correctly identified if all its components (three components) are correctly identified: the initiator of an action—Doer, the participant to whom action is directed—Object and Predicate which calls an action and unites its participants.

Table 3 shows the precision of the developed logical-linguistic model for the English, Kazakh, and Russian languages.

5. Conclusions and future works

The main result of this study is the logical-linguistic model for multilingual Open Information Extraction, which is based on the hypothesis that semantic roles of participants of the action can be represented by the logical conjunction of grammatical features of words in a phrase. The model allows extracting fact in the form of the triplet “Subject—Predicate—Object”, from Web-content of different languages.

In order to assess the model, we created the corpora of Russian, English, and Kazakh texts obtained from particular Kazakh and English websites that comprise specific information, namely criminal contained information.

The performed experiment showed that the precision of our Open IE model achieves a result over 87% for English corpus, over 82% for Russian corpus and 71% for Kazakh corpus.

Obviously, that greater precision is achieved with the implementation of the model for the English language, and the lowest with the using the model for extraction of the facts in the Kazakh corpus. We believe that the main reason for such a result consists in the existence of the means of well-developed POS-tagging, with high precision and parser for English and lack of them for Kazakh.

Funding
This work was supported by the research project provided by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (project No. AP05131073 – Methods, models of retrieval and analyses of criminal contained information in semi-structured and unstructured textual arrays); Ministry of Education and Science of the Republic of Kazakhstan [project No. AP05131073].

Author details
Nina Khairova[1]
E-mail: nina_khajrova@yahoo.com
ORCID ID: http://orcid.org/0000-0002-9826-0286
Orken Mamyrbayev[2]
E-mail: morkenj@mail.ru
ORCID ID: http://orcid.org/0000-0001-8318-3794
Kuralay Mukhsina[3]
E-mail: kuka_ai@mail.ru
ORCID ID: http://orcid.org/0000-0002-8627-1949
Anastasiia Kolesnyk[1]
E-mail: nina_kolesniknastya20@gmail.com
[1] Department of Intelligent Computer Systems, National Technical University ”Kharkiv Polytechnic Institute”, Kharkiv, 61002, Ukraine.
[2] Institute of Information and Computational Technologies, Almaty, 050010, Republic of Kazakhstan.
[3] Department of Information Systems, Al-Farabi Kazakh National University, Almaty, Republic of Kazakhstan.

Citation information
Cite this article as: Logical-linguistic model for multilingual Open Information Extraction, Nina Khairova, Orken Mamyrbayev, Kuralay Mukhsina & Anastasiia Kolesnyk, *Cogent Engineering* (2020), 7: 1714829.

Notes

1. http://www.ruscorpora.ru/old/en/corpora-sem.html.
2. http://universaldependencies.org/en/dep/.
3. http://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/.

4. https://pypi.org/project/pymorphy2/.

References

Akbik, A., & Loser, A. (2012). KrakeN: N-ary facts in open information extraction. AKBC-WEKEX'12: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (pp.52-66). Association for Computational Linguistics, Montréal, Canada.

Angeli, G., Premkumar, M. J., & Manning., D. (2015). Leveraging linguistic structure for open domain information extraction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 344–354. Beijing, China.

Bondarenko, M., & Shabanov-Kushnarenko, J. (2007). *The intelligence theory* (pp. 576). Kharkiv: "SMIT".

Duc-Thuan, V., & Bagheri, E. (2016). Open information extraction. *Encyclopedia with Semantic Computing and Robotic Intelligence*, *1*(1), 1630003.

Etzioni, O., Banko, M., Soderland, S., & Weld, D. (2008). Open information extraction from the web. *Communications of the ACM*, *51*(12), 68–74. doi:10.1145/1409360.1409378

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. *Proceedings of the conference on empirical methods in natural language processing*, 1535–1545. Edinburgh, Scotland, UK.

Gamallo, P., & Garcia, M. (2015). Multilingual Open Information Extraction. In Portuguese Conference on Artificial Intelligence, 711–722. Coimbra, Portugal. doi:10.1007/978-3-319-23485-4_72.

Gamallo, P., Garcia, M., & Fernandez-Lanza, S. (2012). Dependency-based open information extraction. *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, 10–18. Avignon, France.

Gashteovski, K., Gemulla, R., & Del Corro, L. (2017). MinIE: Minimizing Facts in Open Information Extraction. Proceedings of the Conference on *Empirical Methods in Natural Language Processing (EMNLP)*, 2630–2640. Copenhagen, Denmark. doi:10.18653/v1/d17-1278.

Horn, C., Zhila, F., Gelbukh, F., Kern, R., & Lex, E. (2013). Using factual density to measure informativeness of web documents. Proceedings of the 19th Nordic Conference of Computational Linguistics, (NODALIDA 2013); Linköping Electronic Conference Proceedings, 227–238.

Khairova, N., Lewoniewski, W., & Wecel, K. (2017). Estimating the quality of articles in Russian wikipedia using the logical-linguistic model of fact extraction. Poznan, Poland. Part of the Lecture Notes in Business Information Processing book series (LNBIP, volume 288), Springer, Cham 28–40.

Khairova, N., Lewoniewski, W., Węcel, K., Mamyrbayev, O., & Mukhsina, K. (2018). Comparative analysis of the informativeness and encyclopedic style of the popular web information sources. In W. Abramowicz & A. Paschke Eds., *Business information systems. BIS 2018. Lecture notes in business information processing* (Vol. 320, pp. 333–344). Berlin: Springer doi:10.1007/978-3-319-93931-5_24.

Khairova, N. F., Petrasova, S., & Gautam, A. P. (2016). 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. doi:10.1007/978-3-319-46254-7_51.

Lex, E., Voelske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., … Granitzer, M. (2012). Measuring the quality of web content using information. *Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality*, Lyon, France. 7–10. doi:10.1007/0-387-27727-7_15.

Liu, L., Ren, X., Zhu, Q., Zhi, S., Gui, H., Ji, H., & Han, J. (2017). Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 46–56. doi:10.18653/v1/d17-1005.

Nivre, J. (2016). Universal dependencies v1: A multilingual treebank collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 1659–1666.

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012). Open language learning for information extraction. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea. 523–534.

Shinzato, K., & Sekine, S. (2013). Unsupervised extraction of attributes and their values from product description. *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013*, Nagoya, Japan. 1339–1347.

Starostin, A. S., Bocharov, V. V., & Alexeeva, S. V. (2016). FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*, Moscow, Russia. 702–720.

Tseng, Y.-H., Lee, L.-H., Lin, S.-Y., Liao, B. S., Liu, M.-J., Chen, H.-H., … Fader, A. (2014). Chinese open relation extraction for knowledge acquisition. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden. volume 2: Short Papers*, 12–16. doi:10.3115/v1/e14-4003.

Wang, X., Zhang, Y., & Chen, Y. (2018). *A survey of truth discovery in information extraction*. Paper presented in ACM SIGKDD Explorations Newsletter.

Zhou, G., Qian, L., & Fan, J. (2010). Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *Information Sciences*, *180*, 1313–1325. doi:10.1016/j.ins.2009.12.006.