МРНТИ 28.23.11

*О.Ж. Мамырбаев[1], А.С.Шаяхметова[1], Ж.А. Курманбек[2]*

[1]Институт информационных и вычислительных технологий,
г. Алматы, Казахстан
[2]Казахский национальный аграрный университет, г. Алматы, Казахстан

# К ОСНОВНЫМ ВИДАМ РАЗМЕТКИ ТЕКСТОВЫХ КОРПУСОВ

**Аннотация.** В исследовании рассматриваются разработка корпусов современного веб-контента казахского, русского и английского языков, а также модель грамматического выражения идентичности смысла факта побуждения к действию в английском языке и технологии автоматической экстракции синонимичных пар коллокаций из текстов корпусов. Актуальным для решения различных производственно-хозяйственных задач является развитие технологии поиска, извлечения и анализа криминально-значимый информации из массивов неструктурированных и слабоструктурированных данных.
**Ключевые слова:** автоматическая обработка естественного языка, корпус текстов, криминально-значимая информация.

• • •

**Түйіндеме**. Зерттеуде қазақ, орыс және ағылшын тілдерінің заманауи веб-контентінің корпустарын әзірлеу, сондай-ақ ағылшын тіліндегі іс-әрекетке итермелеу фактісі мәнінің грамматикалық сәйкестілігінің моделі және корпус мәтіндерінен синонимдік коллокация жұптарын автоматты түрде экстракциялау технологиясы қарастырылады. Әр түрлі өндірістік және экономикалық мәселелерді шешудің өзекті мәселесі құрылымданбаған және нашар құрылымдалған мәліметтер массивінен қылмыстық маңызды ақпаратты іздеу, алу және талдау технологиясын дамыту болып табылады.
**Түйінді сөздер**: табиғи тілді автоматты түрде өңдеу, мәтіндер корпусы, қылмыстық маңызды ақпарат.

• • •

**Abstract.** The research examines the development of modern web content corpora in Kazakh, Russian and English, as well as the model of grammatical expression of the meaning identity of the fact that prompts to action in English and the technology of automatic extraction of synonymous collocation pairs from the texts of the corpora. The development of technology for searching, extracting and analyzing forensically meningfull information from unstructured and weakly structured data is relevant for solving various industrial and economic problems.
**Keywords:** automatic natural language processing, corpus of texts, criminally significant information.

**Введение.** В последние десятилетия, в связи с распространением сетевых компьютерных технологий, мобильной связи и Интернета, информационные ресурсы современного общества подвергаются растущему числу угроз, чреватых экономическим ущербом и ставящих под угрозу безопасность национальной информационной инфраструктуры. Подобным атакам подвергаются как государственные, так и коммерческие системы, в то время как рост криминальной активности в глобальных сетях (в таких формах, как финансовые мошенничества, нарушения авторского права, распространение детской порнографии, хакерство и т.д.) создаёт угрозы безопасности личности и общества в целом [1]. Таким образом, открытость глобальной сети обусловливает её большую уязвимость от преступных посягательств. Чем больше расширяется Интернет, тем больше сетевых преступлений регистрируется. Кроме того, благодаря компьютерным сетям насильственный экстремизм может глобально распространятся, сохраняя низкую стоимость и высокую скорость. В тоже время, открытость и глобальность Интернета, представляющего собой всемирную телекоммуникационную сеть, создают огромные потенциальные возможности для криминалистов и работников правоохранительных органов. Существующие в настоящее время технологии обработки текстов позволяют специалистам по анализу разведывательных данных и полиции осуществлять превентивную обработку текстовых данных компьютерной сети, собирая, соединяя и анализируя ‘слабые сигналы’ или ‘цифровые следы’ огромных текстовых массивах, которые присутствуют в Интернете. В некоторых случаях такой анализ может помочь обнаружить потенциал противоправного действия прежде, чем оно будет осуществлено. В то же время, одной из главных проблем такой превентивной обработки текстов, наряду с громадным объемом информации Интернет, подлежащей анализу [2], является проблема слабой «окрашенности» криминальных текстов для использования традиционно принятых подходов классификации, кластеризации и выделения шаблонов NaturalLanguageProcessing (NLP). Разработка моделей и методов поиска и анализа криминально значимой информации в неструктурированных и слабоструктурированных текстовых массивах является актуальной научно-практической проблемой.

**Постановка задачи формируется следующим образом**: разрабатывается информационно-лингвистическая технология автоматического определения, выделения, поиска и анализа криминально значимой составляющей в неструктурированных и слабоструктуриро-

ванных тестовых массивах различных языков. Для этого необходимо решить следующие задачи:

- разработать и аннотировать корпуса текста казахского и русского языков социальных сетей и веб-медиа;

- разработать технологию поиска семантически близких коротких фрагментов текста.

**Разработка и аннотирование корпусов текстов казахского и русского языков социальных сетей и веб-медиа.** Разрабатываемый корпус казахских и русских текстов Веб-медиа представляет собой файловую структуру, показанную на рисунке 1.

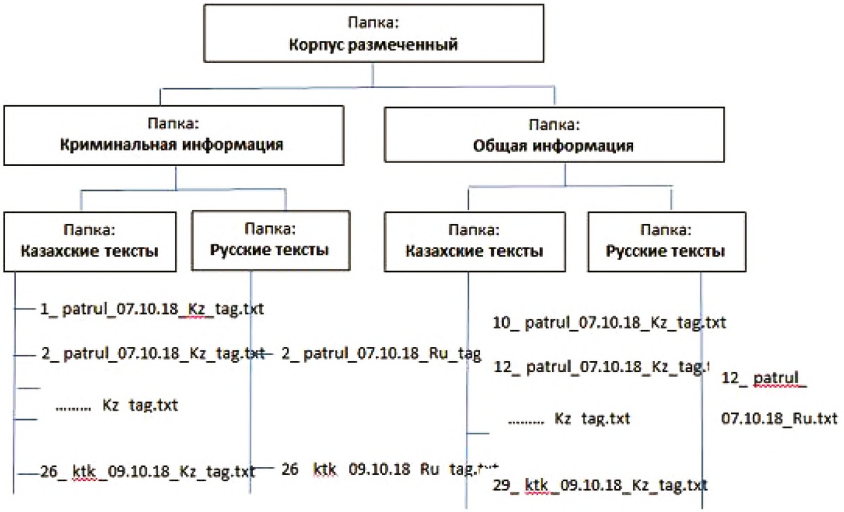


Рисунок 1 - Файловая структура разрабатываемого корпуса

Имя текстового файла должно соответствовать шаблону: *порядковыйНомер_названиеАгенства_дата_язык_tag|row.txt*

Например, размеченный текстовой файл порядкового номера 49 на казахском языке, полученный на сайте информационного агентства partrul седьмого сентября 2018 г. должен иметь имя: 49_ partrul_07.09.18_Kz_tag.txt. Критериями оценки текстового корпуса, помимо его репрезентативности и размера, является используемая система разметки и правильность кодирования метаданных корпуса. Разметка представляет собой добавление в текстовый корпус некоторой дополнительной лингвистической информации. Эта информация

может быть морфологическая (POS-tagging), синтаксическая, семантическая и т.д.

Структурная разметка разрабатываемого корпуса включает следующие теги:

a) заголовок текста выделяется тегом: *<head type=main>заголовок</head>*

b) если в тексте есть подзаголовки они выделяются тегами: *<head type=h1>заголовок</head>*

c) дата публикации выделяется: *<date></date>*

d) сайт информационного Веб-агенства, откуда взят текст, выделяется тегом: *<site></site>*

e) если есть автор, то он отмечается тегом: *<author> автор < /author>*

При разработке tagset учитывались критерии выбора меток: краткость (conciseness) – короткие метки более удобны, чем более подробные и, соответственно, длинные; понятность (perspicuity) – легко интерпретируемые метки; анализируемость (analysability) – метки должны легко декомпозироваться на логические части, как легко читаемые при машинной обработке, так и понимаемые человеком.

**Технология поиска семантически близких коротких фрагментов текста.** Слова, описывающие преступные деяния, имеют свою специфику и часто именно они являются индикативным признаком, по которому осуществляется отбор документов, предназначенных для последующей аналитической обработки. Понятны словосочетания: ножевое ранение, признаки насилия, огнестрельное ранение, взрывчатое вещество, наркотическое вещество, угон автомобиля, завладение имуществом, умышленный поджог, кража денег и т.п. Однако, иногда интересно выявить менее привычные, но более эффективные сочетания слов для поиска криминально значимой информации, например, «винт солянка». Каждое из приведенных словосочетаний у профессиональных работников правоохранительной системы вызывает ассоциации с определенным видом преступления, а, следовательно, их наличие в тексте требует, по крайней мере, глубокого изучения этого текста. В рамках семантико-синтаксического подхода коллокации (устойчивые словосочетания) рассматриваются как синтаксически связанные, лексически определённые элементы грамматических структур, которые характеризуются семантической, синтаксической и дистрибутивной регулярностью.

Как показывают проведенные исследования, в криминально значимых текстах особый интерес представляют именные коллокации. По-

этому на первом этапе обработки массива разнородных текстов необходимо выделить именные словосочетания, используемые в качестве объектов или характеристик данных объектов, которые определяются через взаимное информационное влияние слов в предложении. При этом анализируются только предложения, подчиняющиеся закону проективности, то есть предложения «делового стиля». Содержательный смысл условия проективности предложения состоит в том, что синтаксически связанные слова близки друг к другу и по положению в предложении. Например, именная группа может быть образована только из смежных слов. Проективность не допускает разрыва именной группы [3]. Интерес представляет не только выделение коллокаций в тексте, но и поиск синонимичных коллокаций, обозначающих близкие понятия. В последние несколько лет растёт число исследований, связанных с семантическим подобием различных по уровню текстовых элементов (слов, словосочетаний, коллокаций, коротких текстовых фрагментов различной длины). Это связано, прежде всего, с расширением границ использования семантически близких фрагментов текста в различных NLP-приложениях. Вторая причина роста интереса к идентификации семантически сходных элементов в текстах заключается в ежедневной публикации в социальных сетях миллиардов небольших текстовых сообщений, каждое из которых состоит из 30-40 слов, в то время как традиционные популярные алгоритмы, такие, как, например, Tf-Idf не работают на текстах такого размера [4]. Для текстов такой длины часто необходимы алгоритмы, отличающиеся от статистических.

Далее, на рисунке 2 показана структурная схема используемой технологии, включающая несколько шагов. На первом этапе для того, чтобы правильно разметить обрабатываемые тексты применяется POS-tagging и UDпарсер. Основной причиной использования парсера UD является то, что его древовидные структуры централизованно организованы вокруг понятий субъекта, объекта, клаузального дополнения, определителя существительного, модификатора существительного и т. д. Поэтому синтаксические отношения, соединяющие слова предложения друг с другом, которые определяются UD парсером, могут выражать семантический контент, необходимый для получения семантических характеристик коллокантов. Мы используем шесть типов синтаксической разметки (compound, nmod, nmod:possobj, obj (dobj), amod и nsubj ) из имеющихся грамматических отношений UD, для выделения направленных отношений между двумя существительными, глаголом и существительным и существительным и прилагательным.

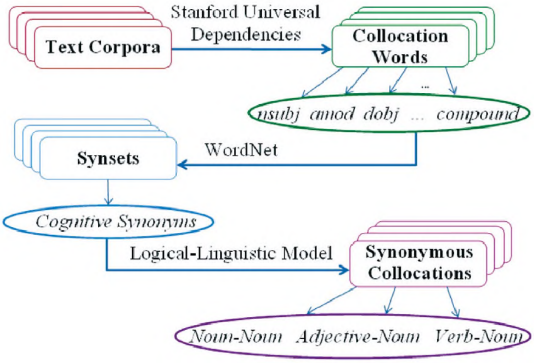


Рисунок 2 – Структурная схема технологии поиска семантически близких коллокаций

На следующем этапе для формализации семантически сходных фрагментов текста посредством конъюнкции грамматических и семантических характеристик коллокантов используется разработанная логико-лингвистическая модель [5-7]. Семантико-грамматические характеристики определяют роль слов в субстантивных, атрибутивных и вербальных коллокациях. В модели множество грамматических и семантических характеристик слов коллокаций определяется двумя предметными переменными ai и ci. Во всех трех типах коллокаций возможные грамматические и семантические характеристики для главного слова коллокации определяются через предикат P(x), а возможные грамматические и семантические характеристики зависимого слова коллокации определяется предикатом P(y). Двухместный предикат P(x,y) описывает бинарное отношение, определенное на декартовом произведении P(x)*P(y) и определяет корреляцию семантической и грамматической информации первого хи второго у слов коллокации:

$$P(x,y) = (x^{NSubAg} \vee x^{NSubOfAg} \vee x^{VTr})(y^{NObjAtt} \vee \vee y^{NObjPac} \vee y^{AAtt} \vee y^{A\Pr}) \quad (1)$$

Используя данное уравнение, определяем предикат семантической эквивалентности между двумя двух словными коллокациями как:

$$P(x_1,y_1) \times P(x_2,y_2) = \gamma_i(x_1,y_1,x_2,y_2) \wedge \wedge P(x_1,y_1) \wedge P(x_2,y_2) \quad (2)$$

где: x - обозначает семантическое сходство; ^ – декартовое произве-

дение, а предикат γi исключает коллокации, между которыми семантическая эквивалентность не может быть идентифицирована.
На следующим этапе для того, чтобы получить синонимы слов, входящих в заданные типы коллокаций, используется WordNet. Для каждого типа коллокации (субстантивного, атрибутивного и вербального) осуществляется поиск WordNetсенсете.

**Заключение.** Таким образом, в работе рассмотрены специфические особенности извлечения криминально-значимой информации из текстов. Так же рассмотрена технология поиска семантически близких коротких фрагментов текста, имплементация, которая позволяет повысить полноту выдачи системы информационного поиска криминально-значимой информации, приводится структура и tagset создаваемых корпусов казахского и русского языков.

**Список литературы**

1 *Бондарева Л. В., Борисенко Т. И., Валентей Т. В.* Современный терроризм: сущность, причины, модели и механизмы противодействия. – М.: Импульс, 2013. – 252 с.

2 *Meloy J. R., Hoffmann J., Guldimann A., James D.* The role of warning behaviors in threat assessment: An exploration and suggested typology. Behavioral Sciences & the Law. – 2012. - № 30(3). – P. 256–279.

3 *Гладкий А. В.* Грамматики деревьев: опыт формализации преобразований синтаксических структур естественного языка. Информ. вопросы семиотики, лингвистики и автоматического перевода. – 1971. – Вып. 1. – С. 16–41.

4 *De Boom C., Canneyt S. V., Bohez S., Demeester T., Dhoedt B.* Learning Semantic Similarity for Very Short Texts. Pattern Recognition Letters. – 2016. - Vol. 80. – P. 150–156.

5 *Khairova N., Petrasova S., Lewoniewski W., Mamyrbayev O., Mukhsina K.* Automatic Extraction of Synonymous Collocation Pairs from a Text Corpus. FedCSIS. Proceedings of the Federated Conf. on Computer Science and Information Systems. – 2018. - Vol. 15. - P. 485–488.

6 Dependen See: A Dependency Parse Visualisation/Visualization Tool // http://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/: 15.04.18

7 NLTK 3.3 documentation: Source code for nltk.stem.wordnet // http://www.nltk.org/_modules/nltk/stem/wordnet.html :27.06.2018

***Мамырбаев О.Ж.*** - PhD, e-mail: morkenj@mail.ru
***Шаяхметова А.С.*** - PhD, e-mail: asemshayakhmetova@mail.ru
***Курманбек Ж.А.*** - магистрант, e-mail: Zhansaya_kurmanbek@mail.ru