N E W S
OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 5, Number 315 (2017), 149 – 155

UDC 007.3

O.Zh. Mamyrbaev, K.Zh. Muhsina

Institute of Information and Computing Technologies
of the Ministry of Education and Science of the Republic of Kazakhstan
E-mail: morkenj@mail.ru, kuka_ai@mail.ru

ANALYSIS OF EXISTING SYSTEMS FOR DETERMINATION OF TONNITY OF TEXT

Abstract. The article considers the existing systems for the analysis of the test. The development of computer technology and the increasing role of information on the present day has given the methods of text analysis a special privileged role. Methods of text analysis are used to search, systematize, evaluate, select information, diagnose, analyze and predict the events or behavior of the subject, which is why these methods have been widely used in security systems.

Keywords: text, text analysis, text analysis program.

ӘОЖ:007.3

О.Ж. Мамырбаев, Қ.Ж. Мухсина

Ақпараттық және есептеуіш технологиялар институты, Алматы қ., Қазақстан

МӘТІН ҮНДЕСІТІЛІГІН АНЫҚТАУҒА АРНАЛҒАН ҚОЛДАНЫСТАҒЫ ЖҮЙЕЛЕРДІ ТАЛДАУ

Аннотация. Мақалада мәтінді талдауға арналған қолданыстағы жүйелер қарастырылған.Компьютерлік технологиялардың дамуы мен ақпараттың рөлінің артуы жағдайында қазіргі таңда мәтінді талдау әдістеріне ерекше көңіл бөлінді.Мәтінде талдау әдістері іздеу, жүйеге келтіру, бағалау, ақпаратты іріктеу,оқиғаларды немесесубъектінің әрекетінанықтауда, талдау және болжау кезінде қолданылатындықтан, бұл әдістерқауіпсіздік жүйесіндекең қолданыс.

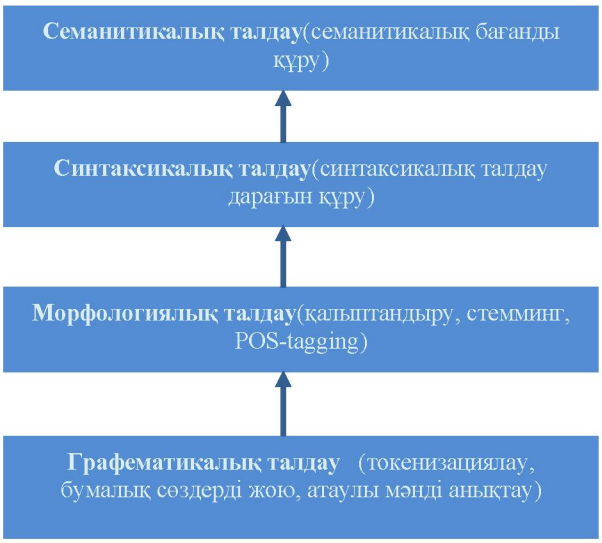
Тірек сөздер: мәтін, мәтін талдауы, мәтін талдауға арналған бағдарламалар.

Кіріспе. Компьютерлік технологиялардың дамуы мен ақпараттың рөлінің артуы жағдайында қазіргі таңда мәтінді талдау әдістеріне ерекше көңіл бөлінді. Мәтінде талдау әдістері іздеу, жүйеге келтіру, бағалау, ақпаратты іріктеу. Қазіргі кезде қоғам ақпараттың үлкен көлемін генерациялайды. Сол ақпараттың маңызды бөлігі – кітап, мақала, хат, хабарлама, түрлі құжаттар мен басқалары сияқты мәтіндік түрде жүреді. Осы мәтіндердің көпшілігі оңай басқарылмайтын құрылымға ие. Сонымен қатар мәтінді кластерлеу, мәтіндерді түйіндеу, екі мәтін релеванттылығын бағалау, сұраныс бойынша іздеу және т.б. тапсырмалар да бар. Сондықтан, осы тапсырмаларды автоматты түрде шешу үшін алгоритмдер мен құралдар қажет. 1960 жылдардан бастап ақпаратпен жұмыс істейтін автоматтандырылған іздеу жүйелері пайда бола бастады. Осы кезден бастап ақпараттық іздеудің принциптері мен әдістерін іске асыруға бағытталған белсенді жұмыстар жүріп жатыр.

Мәтінді компьютерлік талдау есептерін шешудің қазіргі заманғы тәсілдері компьютерлік лингвистикалық талдаудың тұтас бірсатылы процедура еместігін көздейді. Мәтінді компьютерлік талдау процедурасы бірнеше өңдеу деңгейлерін қамтиды. Кейбір деңгейдегі өңдеуіштің мәтін

талдау нәтижелері келесі деңгейдегі өңдеуішке беріледі. Осылайша, өңдеуіштер әрбір буыны белгілі бір мәтіндік ақпаратты өңдеу сатысына жауап беретін тізбек түзеді. Лингвистикалық процессордың өңдеуіштер құрамы ол тағайындалған есептермен анықталады. Мәтінді талдауда қажет болатын түсініктер, қатынастар және шектеулермен, яғни қарапайым және түсінікті онтологияны ажырату қиын болатын күрделі грамматикалы әлсіз құрылымды мәліметтермен жұмыс жасаған кезде жоғарыда аталған мәселелерді шешу жолдары болады.

1. Қолданыстағы мәтіндік ақпаратты талдау жүйелеріне шолу жасау, жүйелерді салыстыру критерийлерін анықтап, оларға салыстырмалы талдау жасау.
   a. Графематикалық жүйелер.
   b. Мәтіннің морфологиялық белгілері мен канонизациясын анықтау жүйелері.
   c. Синтаксистік өзара байланысжәне синтаксистік парсингті талдау жүйелері.
   d. Мәтіннің үндестігін талдау жүйелері.
   e. Барлық тіл деңгейінде мәтінді талдауға мүмкіндік беретін интегралды пакеттер.
2. Осы ұсыныстармен жұмыс істейтінақпарат мәтін мен әдістер ұсыну тәсілдеріне шолу жасау
   a. raw textқұрылмаған мәтін пішіні
   b. Vector Space Model (VSM) - векторлық деректер моделі
   c. Vector Semantic Spaces - ықтималды векторлық деректер моделі
   d. Синтаксистік талдау дарағы
3. Мәтіндік ақпараттың классификациялау және жіктеу әдістерінің қазіргі күйіне шолу жасау.
   a. K-means, K-medoids, X-means
   b. Аңғал Байесов сыныптаушысы (Naive Bayes)
   c. K-NN жіктелімі
   d. Иерархиялық жіктелімдер(аггломеративті және дивизимді)
   e. Нейрондық желілер: көп қабатты персептрондар, рекурсивті нейрондық желілер, терең оқыту нейрондық желілері (Deep Learning)
4. Мәтінді алдын ала өңдеудің әдістерін салыстырмалы талдау және іске асыру.
   a. Мәтінді канонизациялау алгоритімі: стемминг, лемматизация
   b. Морфологиялық белгілерін анықтау алгоритімдері:
   c. Синтаксистік парсерлер
5. Деректер моделін, жүйе компоненттері мен олардың өзара байланысын жобалау.
   a. Бағдарламалық құралдарды таңдау
   b. ДҚ жобалау
   c. Жүйе архитектурасын жобалау



1 сурет - Мәтіндік ақпаратты талдау үдерісі

Зерттеу әдістері.Осылайша, жұмыс қойылымдық шолу, теориялық және тәжірибелік құрамдас бөліктерін қамтитын бірнеше бөлімдерден тұрады.

Мәтінді талдау күрделілігі тілденгейініңарту шамасына қарай жоғарылайды. Жоғары деңгейдегі талдауды алдыңғы деңгейлерде жүргізілген талдаусыз өткізу мүмкін емес. Мысалы, синтаксистік талдау жүргізу (синтаксистік талдау дарағын құру) морфологиялық талдаусыз мүмкін емес. Өз кезегінде, синтаксистік талдау дарағын құрмай тұрып, семанитикалық байланысты анықтау мүмкін емес.

Заманауи мәтіндік ақпаратты талдау құралдарын екі үлкен санатқа бөлуге болады:

1.Мамандандырылған құралдар - нақты тіл деңгейінде талдау жасауға арналған аспаптар (морфологиялық анализатор, синтаксистік парсеттер және т.б.);

2. Интегралды пакеттер –әртүрлі тіл деңгейлерінде мәтіндік талдау жүргізуге мүмкіндік беретін бағдарламалық құралдар;

Зерттеу нәтижелері.

Семантикалық талдау – семанитикалық түйін мен семанитикалық қарым-қатынастан тұратын семанитикалық сөйлем құрылымды құруға бағытталған талдау. Талдау жүргізу мақсаты – бастапқы сөйлем сөздерінен құралған осы түйіндерді құру болып табылады.Семантикалық түйін құрылымына қатысты гипотез жасау негізінсинтаксистік талдау нәтижесінде қол жеткізген ақапарат құрайды [1]. Талдау нәтижелері бірқатар кезеңнен тұратын (семанитикалық түйіндер мен синтаксикалық фрагмент нұсқаларын инициализациялау, көптеген сөздік интерпретация түйіндерін құрылу, уақыт топтарынжасау, жақша ішіндегі түйіндержасау және т.б.) семанитикалық баған түрінде ұсынылады. Семантикалық талдау әртүрлі әдістемелер арқылы жүргізілуі мүмкін, мәселен, PROTAN және басқа көптеген әдістеме түрлерімен. Мысалы, семантикалық талдау T-LAB ToolsforTextAnalysis әдістемесімен жүргізілген, аталмыш әдістеме тақырыптық талдау, салыстырмалы талдау, шектестік талдау сынды үш түрлі талдау жүргізуге мүмкіндік беретін, сонымен қатар сөздердің мағыналық паттерндері мен мәтіннің негізгі идеяларын айқындайтын компьютерлік әдістеме болып табылады. Бұл әдістемеде мәтінмен жұмыс істеу процессі өзіне мәтінді саралауды, түйін сөздер іріктемесін, сондай-ақ талдаудың үш типтерін жүзеге асыруға арналған процедураларды қамтиды.

Кесте 1 - Орыс тіліндегі мәтіндерді синтаксистік талдауға арналған құралдар

Атауы	Әдістері	Лицензиясы	Платформасы	Console	API	Модуль-діліг і	Құны (ком. тұлға.)
Еркін таратылатын							
AOT	Сөздікті	LGPL	GNU/Linux, Microsoft Windows	+	+	AOT модулі	
Link Grammar Parser	Байланыс грамматикасы	BSD	GNU/Linux, Microsoft Windows	+	+		
AGFL	Соңғы торлар аффиксінің граматикасы	GPL	GNU/Linux, Microsoft Windows	+			
MaltParser	Машиналықоқу	өзіндік	Java	+			
NLTK	Машиналық оқу	Apache Lisence	Python		+		
Pattern	Ереже, тұрақты айту	BSD	Python				
			Проприетарлық				
ABBYY Compreno	Ереже	Коммерциялық	Microsoft Windows			модуль ABBYY SDK	н/д
DictaScope	Ереже	Коммерциялық	FreeBSD, Microsoft Windows		+	Кітапхана	н/д

Синтаксикалық талдау – сөздер рөлдері мен олардың өзара байланысын нәтижесінде осындай байланыстар көрсететін дарақтар жинағына қол жеткізе отырып, анықтайды. Міндеттерді орындау кіріс деректерінің көп мағыналығы және сол сияқтыталдау ережелерінің бірмәнді еместігімен байланысты талдау барысында туындайтын балама нұсқалардың көптігімен күрделене түседі [2]. 1 кестеде орыс тіліндегі мәтіндердісинтаксистік талдауғаарналған құралдар келтірілген.

Кесте - 2. Орыс тіліндегі мәтіндерді талдауға арналған құралдар

Название	Методы	Лицензия	Платформа	Console	API	Модульность	Стоимость (ком.лиц.)
Свободно распространяемое							
AOT	словарный	LGPL	GNU/Linux, Microsoft Windows		+		
MAnalyzer	словарный	MIT	GNU/Linux	-	-	Библиотека	-
Myaso	алгоритм Витерби	MIT	Ruby	-	+	Библиотека	-
mystem	словарный	Некоммерческая	GNU/Linux, Microsoft Windows	+	+		
phpmorphy	словарный	LGPL	PHP	-	+	Библиотека	-
Pullenti SDK	н/д	Условно бесплатная	NET		+	модуль SDK	100 000 руб
pymorphy	словарный	MIT	Python	-	+	Библиотека	-
RussianMorphology	словарный	Apache License	Java	-	+	Библиотека	-
RussianPOSTagger	словарный	GPL	Java	+	+	модуль GATE	-
Snowball	алгоритм Портера	BSD	GNU/Linux, Microsoft Windows	+	+		
Stemka	словарный	Собственная	GNU/Linux, Microsoft Windows	+	+		
SVMTool	метод опорных векторов	LGPL	Perl		+	Библиотека	
TreeTagger	деревья принятия решений	Некоммерческая	GNU/Linux, Microsoft Windows	+	+		
FreeLing	словарный	Условно платная	GNU/Linux	-	+	Библиотека	
Проприетарное							
RCO	словарный	Коммерческая	Microsoft Windows		+	Пакет для СУБД OracleRCO	от 35 000 $р_{уб}$

Мәтінді өңдеудің қиынырақ және ресурс сыйымды сатысы синтаксистік талдау болып табылады. Есептерге байланысты лингвистикалық процессорларда үстіртін (shallow) немесе терең (deep) синтаксистік талдау пайдаланылады [3]. Үстіртін талдау дегеніміз қарапайым, рекурсив енгізілмеген синтаксистік топтардың бөлінуі, бұл тәсіл шетелдік әдебиет көздерінде «chunking» терминімен белгілі. Бұл синтаксистік талдаудың ең қарапайым есебі. Қоспа синтаксистік дарақ

құруды көздейтін үстіртін талдау үшін грамматикалық сынды синтаксистік әдістер де қолданылады. Олардың ішінде стохастистік контекстік-еркін грамматика құруға негізделген әдістер кеңінен танымал [4]. Терең синтаксистік талдау мынадай экстралингвистикалық білімдерді тарта отырып, ережелер жүйесін пайдалануға негізделеді: семантикалық сөздіктер, тезаурустар, топтастыру [5]. Орыс тілін терең синтаксистік талдау жүйелеріне ЭТАП-3 [4], Abby Compreno [6] жатады. Ал ағылшын және басқа да еуропа тілдері үшін терең синтаксистік талдау әдістері Xerox XLE жүйесінде [7], RASP [8] и ENJU [9]. іске асырылады. Қазіргі синтаксистік талдау әдістері өте жоғары дәлдігі мен синтаксистік байланыстар орнату толықтығын көрсетеді: тілге байланысты орта есеппен 75% - дан 90% - ға дейін [3].

Мәтінді сөздерге және сөйлемдерге бөлу бірінші кезектегі міндет болып табылады. Ол әдетте эвристикаларды қолдану арқылы тұрақты өрнектер мен түпкілікті автоматтар көмегімен шешіледі [10].

Морфологиялық талдау – аталмыш сөзтұлға жасалынған қалыпты форма мен осы сөзтұлғамен тіркелінген параметрлер жинағын анықтаумен қамтамасыз етеді[11]. 2 - кестеде орыс тіліндегі мәтіндерді талдауға арналған құралдар көрсетілген.

Морфологиялық талдау басқа да мәтін талдау түрлері үшін негіз болып табылатындықтан, көптеген әдістемелерде іске асырылады.

Морфологиялық талдау мынадай сатыларды қамтиды:

1. Мәтінде көптеген мүмкін ықтимал морфологиялық сөз қолдану түсіндірмесін анықтау (леммалар мен сөзтұлғаның морфологиялық сипаттамалары);
2. Омонимге рұқсат беру – мүмкін морфологиялық түсіндірмелерден мәтінде сөз қолдану бағытына сәйкес келетін түсіндірмені бөліп көрсету.

Морфологиялық талдау үшін сөзтұлғамен бірге оның леммасы, морфологиялық қасиеттері сақталынатын тілдің сөз тұлғасы арнайы сөздікке орналастырылатын әдіс кеңінен қолданылады. Талдау осы сөздіктегі аталмыш сөзтұлғаны іздеуге түйістіріледі. Бұл тәсіл АОТ[12] және Freeling[13] мәтіндерді клмпьютерлік талдау жүйелерінде жүзеге асырылған.

Кейбір жағдайларда тек сөздіксіз тәсіл - стемминг (stemming) пайдаланылады, онда белгілерін болжау аффикстер кестесі көмегімен, ал нормалау – аффикстерді мүмкін сөз негізіне дейін қию арқылы жүзеге асырылады [14]. Морфологиялық талдау нәтижесінде пайда болатын омонимдерге рұқсат юеру үшін жасырын Марков моделін [15, 16] және сөздердің морфологиялық сипаттамаларын анықтауға арналған сөз пайдаланулар мәнмәтінін пайдаланатын ережелер жүйесін (мәселен, машиналық оқыту көмегімен қалыптасатын) қолданады [17].

Графематикалық талдау. Морфологиялық мәтін талдауын бастау үшін бастапқы құрылымсыз мәтінді сөйлем мен сөзге бөлу қажет. Бұл бір қарағанда, жеңіл міндеттің өзіндік ерекшеліктері бар және әрі қарай мәтін талдау кезінде маңызды рөлге ие.

Графематикалық талдау мыналарды қамтиды:

- ✓ бастапқы мәтінді элементтерге бөлу (сөздерге, айырғыштарға);
- ✓ мәтінге жатпайтын элементтерді жою (белгілерге, метаақпаратқа);
- ✓ стандарттан тыс элементтерді бөліп рәсімдеу:
- ✓ құрылымдық элементер: тақырыптар, абзацтар, ескертулер;
- ✓ сандар, күндер, әріптік-сандық кешендер ;
- ✓ аттары, әкесінің аты, тегі;
- ✓ пішіндеу элементтері: курсивті, астын сызу, қалың шрифт;
- ✓ электрондық мекенжайларды бөлу;
- ✓ файл аттарын бөлу;
- ✓ орнықты айналымды бөлу, бір бірінен бөлек жазылмайтын сөздер ;

Ағылшын тіліндегі дереккөздерінде tokenization (токенизация) анықтамасын кездестіруге болады, ол өзінің мазмұны бойынша графематикалық талдауға ұқсас. Токенизация – мәтін ағынын токендерге бөлу процессі: сөздер, сөз тіркестері және сөйлемдер[18].Осылайша, графематикалық талдау мәтінді әрі қарай өңдеу үшін қажетті ақпарат қалыптастыратын қандайда бір кодтамада нышандар тізбегі түрінде ұсынылған құрылымы жоқ мәтіннің бастапқы талдауы болып табылады.

Іс жүзінде графематикалық талдауға мамандандырылған құралдар жоқ. Көбіне, графематика интегралды NLTK, Stanford CoreNLP, Apache NLP, AOT, MBSP мәтінді талдау пакетіне және де т.б. мәтіндерге енгізілген.Сондай-ақ токендерге бөлу міндеті мәтіннің белгілеу бағдарламасына

енгізілген, мысалы, part- of-speech taggers. Көп жағдайларда бөлу амалдарын тривиалді тәсілмен айырғыш сөздік және тұрақты өрнек сөздігін пайдалана отырып,шешуге болады. Сонымен қатар, бұл міндетті тұрақты өрнектер көмегімен де шешеді.

Кесте 3 - Мәтін талдау әдістері

Атауы	Әдісі	Тілдер	Лицензиясы	Платформасы
Tokenizer	Ереже	орыс, ағылшын, неміс	GPL	C
Greeb	тұрақты өрнектер	орыс, ағылшын,	MIT	Ruby
Twitter NLP and Part-of- Speech Tagger	Машиналық оқыту	ағылшын	GPL	Java
Lemmatizer	Сөздік	орыс, ағылшын,	GPL	GNU/Linux

Нәтижелерін талқылау

Жоғарыда аталған әдістерден басқа мәтін талдаудың мынадайтәсілдері бар: құрылымдық талдау,семиотикалық (семиологиялық) талдау, жүйелік талдау, символикалық (мифологиялық) талдау, әлеуметтік индикаторлар мен түйін сөздер наррациясын (желі) талдау, әлеуметтік-рөлдік талдау,риторикалық талдау, перформативтікталдау, жанрлықталдау, сөйлеу қызметі талдауы, психоаналитикалық талдау, сыни талдау,тарихи талдау, мәдени талдау, интертекстуалды талдау, феноменологиялық талдау типтері; коммуникативтік стратегиялар мен еркін қауымдастық талдауы,прагма-, психо-, социо-, этно-, когнитивті-лингвистикалықталдау және т.б.түрлері бар.

Қазіргі таңда лингвистикалық мәтін талдау әдістері ойлап табылып, мәтін қосымшаларына автоматты морфологиялық, синтаксистік және семантикалық талдау жүргізуге мүміндік беретін бағдарламалық құралдар әзірленген: AOT [19, 20], Solarix [21], NLTK [22, 23], FreeLing [24] және басқалар. Бұл жүйелердің есептеу тиімділігі мен лингвистикалық талдау сапасының деңгейі оларды үлкен мәтіндер топтамасын өңдеу үшін қолдануға мүмкіндік береді.

Қорытынды. Қазіргі таңда автоматты морфологиялық, синтаксистік және семантикалық талдаулар жүргізуге мүмкіндік беретін бірқатар мәтіндік ақпаратты лингвистикалық талдау әдістері ойлап табылған. Сонымен қатар, ЕЯ мәтіндерін автоматты талдауға арналған бағдарламалық құралдар (олардың көпшілігі еркін бағдарламалық қамтамасыз ету лизензиясы бойынша таралады) әзірленген. Осыған қарамастан, өнеркәсіптік ақпараттық-аналитикалық жүйелерде лингвистикалық мәтіндер ақпараты (лексикалық, морфологиялық, синтаксистік және семантикалық) кешенжі түрде қолданылмайды.

Белгілі қосымшаларда мәтіндік ақпараттың статистикалық сипаттамалары бар лексика векторы түріндегі қарапайым ұғымын қолданылады немесе мәтіннің өз моделін (мәселен, синтаксистік құрылымдар мен тек семантикалық мағыналарын есепке алатын) пайдалана отырып, жеке есептер (мысалы, сұрақ-жауапты немесе фразалық іздену) шешіледі.

Кез келгенталдауды жүргізу барысындағы ең басты параметрлер бірі – қол жеткізген мәліметтердің нақтылығы (яғни талданатын мәтін толықтығымен және оның көрнекілігімен қамтамасыз ететін) және талдау бірлігінің толықтай зерттеуші біліктілігі мен негізі болып табылатын теориялық моделіне байланысты болатын интеркодтау дәйектілігі. Мәтіндік талдауды пайдалануды шектеу зерттеушінің субъективті әрекетінеықпал тигізіп, өз таңдауын жасауға әсерін тигізеді.

ӘДЕБИЕТ

[1] Berelson, B. Content Analyses in Communication Research / B. Berelson. –Glencoe, 1952. – 220с.

[2] Большакова Е.И., Клышинский Э.С., ЛандэД.В.,Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. Пособие //. – М.: МИЭМ, 2011. – 272 с.

[3] Смирнов И. В., Шелманов А. О. Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов // Искусственный интеллект и принятие решений. - 2013. - Т. 1. - С. 41–54.

[4] Л.Л. Иомдин, В. В. Петроченков, В. Г. Сизов, Л. Л. Цинман . Синтаксический анализатор системы ЭТАП: современноесостояние. / Papers from the Annual International Conference "Dialogue" (2012). – 2012.

[5] Federici S., Montemagni S., Pirrelli V. Shallow parsing and text chunking: a view on underspecification in syntax // Cognitive science research paperuniversity of Sussex CSRP. – 1996. – P. 35–44 75

[6] Syntactic and semantic parser based on ABBYYCompreno linguistic technologies / K.V.Anisimovich,K.Ju.Druzhkin,F.R.Minlosetal.//PapersfromtheAnnualInternationalConference"Dialogue"(2012).–Vol.2.– 2012.–P.91–103

[7] Speedandaccuracyinshallowanddeepstochasticparsing / RonaldM. Kaplan, StefanRiezler, Tracy H.Kingetal .// InproceedingsofHLT-NAACL'04. – 2004.

[8] Briscoe T., Carroll J. Robust accurate statistical annotation of general text. – 2002

[9] Miyao Y., Tsujii J. Feature forest models for probabilistic hpsg parsing // Comput. Linguist. – 2008. – Vol. 34, no. 1. – P. 35–80

[10] Урюпина О. Автоматическое разбиение текста на предложения для русского языка. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (4–8 июня 2008 г.). Вып. 7 (14). – М.: РГГУ, 2008.

[11] Большакова Е.И., Клышинский Э.С., ЛандэД.В.,Носков А.А., Пескова О.В., Ягунова Е.В. *Автоматическая обработка текстов на естественном языке и компьютерная лингвистика.*: учеб.пособие /. – М.: МИЭМ, 2011. – 272 с.

[12] Автоматическая Обработка Текста (АОТ). / [Электронный ресурс] URL: http://www.aot.ru (дата обращения 23. 01.2013)

[13] Freeling: An Open Source Suite Of Language Analyzers / [Электронный ресурс] URL: http://nlp.lsi.upc.edu/freeling // (дата обращения 05.03.2013).

[14] Porter M.F. An algorithm for suffix stripping, Program. №14(3), 1980. P.P. 130−137.

[15] Сокирко А. В., Толдова С. Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель маркова и синтаксический анализатор именных групп) // Сборник работ стипендиатов Yandex. – 2005. 64.

[16] Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. //

[17] Brill E. A simple rulebased part of speech tagger / Proceedings of the workshop on Speech and Natural Language // Association for Computational Linguistics. – 1992. – P. 112–116.

[18] Feinerer, I., Hornik, K. & Meyer, D. Text mining infrastructure in R. / Feinerer, I., Hornik, K. & Meyer, D. //Journal of statistical software, 25(5). - 2008. - American Statistical Association

[19] Автоматическая Обработка Текста (АОТ). / [Электронный ресурс] URL:http://www.aot.ru (дата обращения 23. 01.2013)

[20] А.Сокирко. Семантические словари в автоматической обработке текста (поматериалам системы ДИАЛИНГ) / Дисс канд.т.н. // [Электронный ресурс]URL: http://www.aot.ru/docs/sokirko/sokirko-candid-1.html (датаобращения 23.01.2013)

[21] Solarix: Компьютерная лингвистика. / [Электронный ресурс] URL:

[22] http://www.solarix.ru/ (дата обращения 05.03.2013)

[23] Natural Language Toolkit. / [Электронный ресурс] URL: http://nltk.org/

[24] (дата обращения 05.03.2013).

[25] Bird S. Natural Language Processing with Python. / O'Reilly Media Inc, 2009

[26] Freeling: An Open Source Suite Of Language Analyzers. / [Электронныйресурс] URL: http://nlp.lsi.upc.edu/freeling / (дата обращения 05.03.201).

УДК 007.3

О.Ж. Мамырбаев, Қ.Ж. Мухсина

Институт информационных и вычислительных технологий КН МОН РК

АНАЛИЗ СУЩЕСТВУЮЩИХ СИСТЕМ ДЛЯ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТА

Аннотация. В статье рассмотрены существующие системы для анализа теста. Развитие компьютерных технологий и увеличение роли информации на настоящий день отвело методам анализа текста особую привилегированную роль. Методы анализа текста применяются при поиске, систематизации, оценке, отборе информации, диагностике, анализе и прогнозировании событий или поведения субъекта, из-за чего эти методы получили широкое применение в системах безопасности.

Ключевые слова: текст, анализа текста, программы анализа текста.