УДК 004.823
МРНТИ 20.19.27

ҚАЗАҚ ТІЛІНДЕГІ КЛИНИКАЛЫҚ МӘТІННІҢ СЕМАНТИКАЛЫҚ РӨЛІН БЕЛГІЛЕУ

А.Б. ДЖАКСЫЛЫКОВА, А.А. ЗИЯДЕН, Ж. РАХЫМБЕКҰЛЫ, А. ҚАЛИЕВА, П. КОМАДА

Институт информационных и вычислительных технологий КН МОН РК
Казахский Национальный университет им. аль-Фараби
КГКП «Региональная инфекционная больница г.Талдыкорган»
Технический университет «Люблинская политехника», Польша

Аңдатпа: *Семантикалық рөлдік таңбалау (SLR) әртүрлі мәтіндерден олардың мағыналары мен қатынастарын үстірт бейнелейді, семантикалық қабат табиғи тілді түсіну үшін маңызды. Медицина саласында негізінен аннотацияланған клиникалық корпустардың, әсіресе қазақ тілінде болмауына байланысты, семантикалық рөлдердің таңбалануы бойынша бірнеше зерттеулер жүргізілді. Бұл жұмыстың мақсаты – өнімділікті жоғарылату және шығындарды үнемдеу, сонымен қатар болжамды медицинаның сапасын жақсарту үшін тәжірибелік клиницист дәрігерлердің құрған корпусын қолдана отырып, клиникалық корпустарға арналған семантикалық рөлдерді белгілеу үшін негіз жасау. Материалдар мен әдістер: клиникалық тәжірибелер, атап айтқанда, асқазан-ішек жолдары, жүрек-қан тамырлары аурулары және басқалары негізінде жасырын жасалынған мәліметтер базасы доменнің деректер жиынтығы ретінде пайдаланылды. Жазбалар қолмен талданып, белгіленді. Семантикалық белгілеу шеңбері және семантикалық рөлдердің қолданылуы мен олардың нақты клиникалық жағдайларға қатынасы туралы мәліметтер келтірілген.*

Түйінді сөздер: *семантикалық рөлді белгілеу, таяз семантикалық талдау, табиғи тілді клиникалық өңдеу, доменге бейімделу*

SEMANTIC ROLE MARKING FOR CLINICAL TEXT IN KAZAKH

Abstract*: semantic role labeling (SLR) extracts a superficial representation of meanings and their relationships from various texts, the semantic layer is important for understanding natural language. Few studies in the labeling of semantic roles have been conducted in the field of medicine, mainly due to the lack of annotated clinical buildings, especially in Russian. The aim of this work is to develop a framework for marking semantic roles for clinical notes using the corps created by practicing clinicians to increase productivity and save costs, as well as improve the quality of predictive medicine. Materials and methods: an anonymous database, collected on the basis of clinical practice, in particular, diseases of the gastrointestinal tract, cardiovascular system and others, was used as a data set of the target domain. Records were manually analyzed and tagged. The framework of semantic markup is presented and the analysis of the applicability of semantic roles and their relationships with respect to real clinical cases is given.*

Keywords: *semantic role marking, shallow semantic analysis, clinical processing of natural language, domain adaptation*

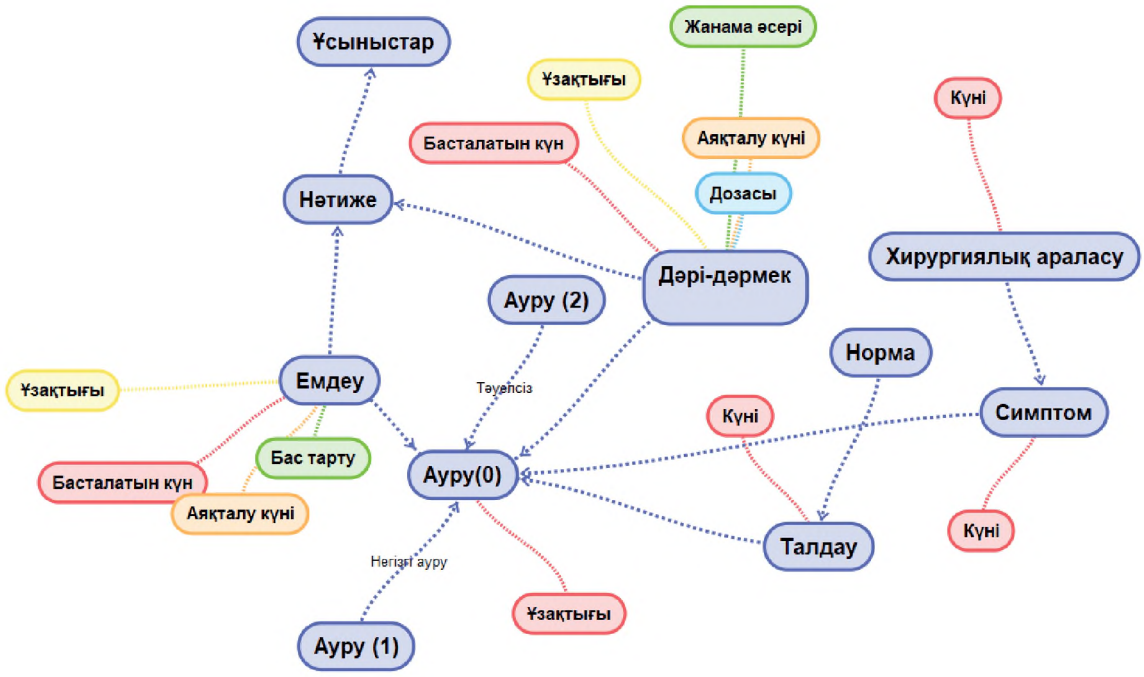
СЕМАНТИЧЕСКАЯ РОЛЕВАЯ МАРКИРОВКА ДЛЯ КЛИНИЧЕСКОГО ТЕКСТА НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация: *Маркировка семантических ролей (SemanticRoleLabeling, SLR) извлекает поверхностное представление смыслов и их отношений из различных текстов, семантический слой важен для понимания естественного языка. Исследований в маркировке семантических ролей в области медицины было проведено мало, в основном, из-за отсутствия аннотированных клинических корпусов, особенно на русском языке. Целью настоящей работы является разработка каркаса маркировки семантических ролей для клинических заметок с использованием корпуса, созданного практикующими врачами клиницистами, для повышения производительности и экономии затрат, а также повышения качества предиктивной медицины. Материалы и методы: в качестве набора данных целевого домена использовалась обезличенная база данных, собранная на основе клинической практики, в частности, по болезням желудочно-кишечного тракта, сердечно сосудистой системы и других. Записи были вручную проанализированы и помечены. Представлен каркас семантической разметки и приведен разбор применимости семантических ролей и их отношений относительно реальных клинических случаев.*

Ключевые слова: *семантическая ролевая маркировка, неглубокий семантический анализ, клиническая обработка естественного языка, адаптация домена*

КІРІСПЕ

Табиғи тілді өңдеу (NLP) технологиялары электронды медициналық есепке алу жүйелерінде клиникалық есептерде сақталған ақпаратты өңдеу үшін маңызды, дегенмен, әртүрлі компьютерлік қосымшаларды, мысалы, био бақылау және клиникалық шешімдерді қолдау сияқты компьютерлік медициналық қосымшаларды қолдана отырып, семантикалық маңызды ақпаратты алу үшін әртүрлі NLP жүйелері жасалған. Дәстүрлі жүйелер семантикалық рөлдік таңбалауды (SRL) қолданады [1], (семантикалық деп те аталады) синтаксистік талдау) [2], бұл предикаттар мен олардың дәлелдерін әртүрлі мәтіндерден алады. Қазіргі SRL жүйесі ашық домендерде және әртүрлі биомедициналық субдомендерде ақпаратты алу үшін жасалынған және қолданылған [3–12]. Алайда клиникалық салада, SRL зерттеулері өте аз [13,14], мүмкін, бұл аннотацияланған үлкен корпустардың болмауына, әсіресе қазақ тілі үшін осындай корпустардың болмауы мәселенің туындауына себепші болып отыр. Мұндай SRL клиникалық корпусын жасау көп уақытты қажет етеді және қымбатқа түседі. Бұл зерттеуде біз SRL-ді клиникалық белгілеуге сілтеме ретінде предиктивті медицинаның мәселесі ретінде қарастырамыз. Мақсатымыз – SRL әдістемесін қазақ тілі үшін клиникалық аймағына бейімдеу [15,16]. SRL-нің міндеті – сөйлемдегі семантикалық қатынастарды сөйлемдерді ұсынуға арналған предикативті актантикалық құрылым (PAS) ретінде белгілеу [17]. PAS анықтамасы семантика теориясындағы сөйлемдерді білдіру үшін негізгі логикадан туындады. Ағылшын тіліне арналған биомедициналық мәтіндерде семантикалық қатынастарды алу туралы көптеген жұмыстар бар [4-12,18-22]. Әдетте, медицина сияқты жабық доменде бастапқы семантикалық типтердің шектеулі саны және әртүрлі семантикалық аргументтердің мағыналық қатынастарды қалай байланыстыруға болатындығын анықтайтын шектеулер бар. Екі жоба бар, атап айтсақ, лингвистикалық құрылым жобасы (Linguistic String Project, LSP) [20] және медициналық тілдерді шығару және кодтау жүйесі (Medical Language Extraction and Encoding System, MedLEE) [21], олар тілдік грамматиканың қосалқы жүйесін қолданады – бұл NLP екі ерте жүйе медициналық салада семантикалық қатынастарды алуға арналған. SemRep – бірыңғай медициналық тіл жүйесінде анықталған семантикалық болжамдарды шығаратын тағы бір биомедициналық семантикалық қарым-қатынас жүйесі. Биомедициналық әдебиеттің семантикалық желісі. PennBioIE корпусы, анықтаған болатын және семантикалық тұрғыдан шектеулі доменде де синтаксистік вариациялар көп



1-сурет. Клиникалық мәтіндерді белгілеуге арналған мағыналық рөлдердің, предикаттар мен атрибуттардың байланыс диаграммасы

кездесетінін және әртүрлі болатындығын анықтады. Қазіргі уақытта, NLP -ге негізделген көптеген клиникалық жүйелер, көбінесе ережеге негізделген әдістерді қолдана отырып заңдылықтарды қолмен алу арқылы семантикалық қатынастарды таниды. Ол семантикалық фреймдерде алдыңғы қайта іздеуден [25] және семантикалық рөлдер мен синтаксистік іске асырудың арасындағы байланысты шабыттандырады [23, 24]. [25] Қазіргі кездегі SRL тәсілдері негізінен ашық жерлерде дамытылғанымен (осылайша семантикалық рөлдер немесе дәлелдемелер түрлері жеткіліксіз немесе медициналық практика үшін жарамсыз болуы мүмкін), оларды балама түрде қамтамасыз етіп, медициналық салада кеңейтуге болады [13,14] немесе клиникалық-семантикалық қатынасты алу үшін қосымша тәсілдер.

КЛИНИКАЛЫҚ МӘТІНДЕРДІҢ СЕМАНТИКАЛЫҚ БЕЛГІЛЕНУІ

SRL-де предикат, әдетте атрибутты немесе қатынасты көрсететін сөзге сілтеме жасайды, ал дәлелдер предикат үшін әртүрлі семантикалық рөлдер ретінде әрекет ететін синтаксистік компоненттерге сілтеме жасайды. Негізгі дәлелдер – бұл предикаттың негізгі аргументтері, ал қосымша дәлел уақыт пен орын сияқты предикаттың жалпы қасиеттерін білдіреді.

1-суретте келесі элементтер үшін семантикалық рөлдердің диаграммасы көрсетілген:

Symptom –клиникалық немесе басқа зерттеулерде анықталған симптомдар, белгілер мен ауытқулар;

Illness – халықаралық аурулар жіктеуіші бойынша нозологиялық өлшем (МКБ -10);

Drug – пациент қабылдаған дәрілер;

Drug_Duration –қабылдау ұзақтығы;

Drug_Start_date – препаратты қабылдауды бастаған күн;

Drug_End_date – препаратты қабылдауды тоқтатқан күн;

Side_effects – жанама әсерлері;

Dousage – препараттың дозасы;

P_Cause – аурудың себебі;

Result – емдеудің динамикалық дамуы;

Result_Duration – соңғы нәтиженің ұзақтығы;

Recomm – дәрігердің ұсынысы;

Surgical_intervention – хирургиялық араласу;

Surgery_Date – ота жасалған күн;

Treatment – емдеу, атап айтқанда шаралар жиынтығы;

Treatment_Start_date – емдеудің басталу күні;

Treatment_End_date – емдеудің аяқталған күні;

Denial – науқастың дәрігер тағайындаған қажетті шаралардан бас тартуы;

Treatment_Duration – емдеудің ұзақтығы;

Dispencery – науқасты (созылмалы) аурулары бойынша есепке алу;

Duration_Illness – аурудың ұзақтығы;

Illness_date – аурудың басталу күні;

Symptom_date – белгілердің басталу күні;

Analyses – талдау түрлері;

Date_Medical_Operation – медициналық тексеру күні;

Norma – норма

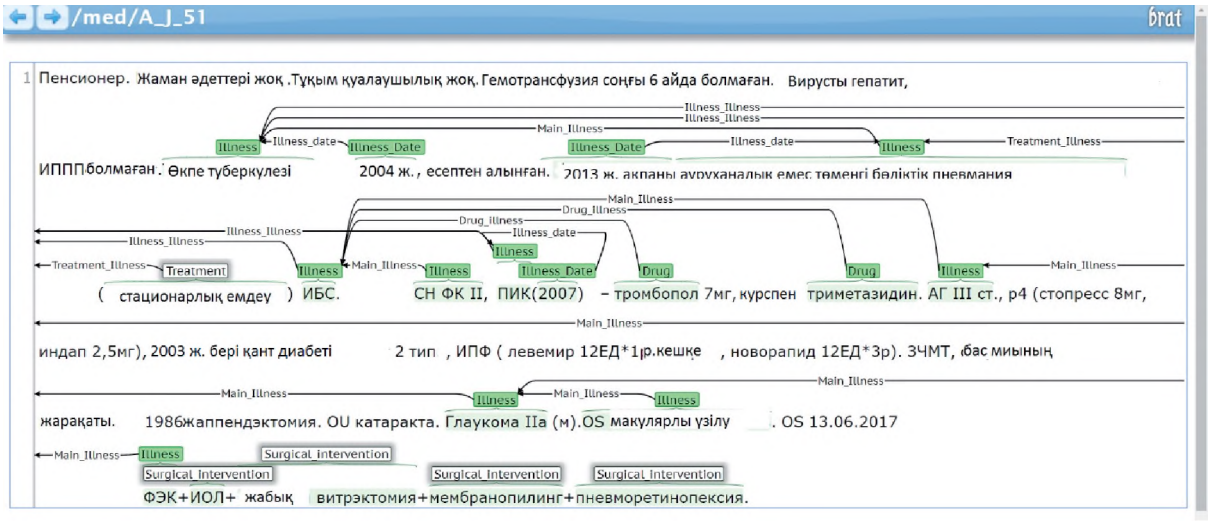
Symptom_Duration – симптомдардың ұзақтығы

Төмендегі 2-суретте дискуляциялық энцефалопатияның нақты клиникалық жағдайларына түсініктеме беру мысалы келтірілген.

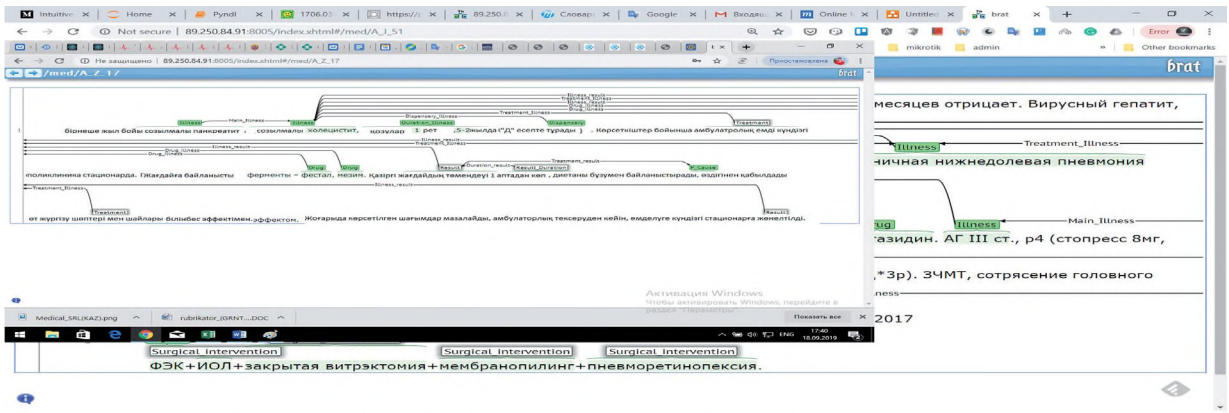
Бұл мысалда келесі аңғартпалар пайдаланылады: Illness және Duration_Illness, сонымен қатар Main_Illness қатынасы.

Қарапайым типтелген мәтін ауқымының категориясы реляциялық ақпаратты шығарудың қарапайым міндеттері үшін аталған нысандарды және бинарлық қатынастарды тану үшін аңғартпалар құруға жарамды.

Brat аннотациясы [сілтеме] сонымен қатар белгілі рөлдерге қатысқан басқа аннотациялардың кез келген санын біріктіре алатын n-дық қауымдастықтарының аннотациясын қолдайды. Бұл аннотация санатын, мысалы, келесі мысалда Illness сияқты оқиғаларға түсініктеме беру үшін пайдаланылады: басқа аннотациялардың ұқсас түрлері мен қасиеттерін аннотацияларда орнатуға болатын төл-



2-сурет. Белгілеу мысалы – Қант диабеті



3-сурет. Белгілеу мысалы – созылмалы холецистит және панкреатит

сипаттарды қолдану арқылы одан әрі анықтауға болады.

Медициналық деректерді белгілеудің негізгі буыны жарақат немесе ауру болып табылады (тэг Illness). Ауру жалғыз болуы мүмкін, сонымен қатар басқа аурулармен қатар жүруі мүмкін. Аңғартпаларда мұндай қатынастар Independent және Main_Illness арқылы көрсетіледі (1,3 суретті қарау).

1-кесте. Медициналық жазбаларды санат бойынша бөлу

Ауру аттары	64
асқазан-ішек жолдары	65
Инфекция	33
Онкология	7
Ішкі аурулар	21
Жүрек-қан- тамырлары	207
Барлығы	397

ҚОРЫТЫНДЫЛАР МЕН ТАЛҚЫЛАУ

Медициналық жазбалар мен ескертпелер дұрыс диагноз қою үшін негізгі ақпарат көздерінің бірі болып табылады және медициналық диагноз қою үшін сапалы болжам бола алады. Медициналық мәтіндердің белгіленуі медицинаның белгілі бір саласы үшін нақтыланған семантикалық рөлдерді білдіреді, бұл жұмыста біз сан түрлі ауруларды немесе әртүрлі нозологиялық бірліктерді қамтитын жалпыланған семантикалық түзуді ұсындық. Атап айтқанда, келесі медициналық категорияларға талдау жасалды:

Бастапқы домендерден алынған білімді клиникалық аймаққа бейімдеуге болатынына қарамастан, клиникалық мәтіннің бірегей сипаттамалары SRL жүйесін одан әрі жетілдіру үшін арнайы ресурстар мен шешімдерді қажет етеді. Ерекше белгілердің бірі – клиникалық лексика және олардың арасындағы семантикалық қарым-қатынас. «Аралас генездің дискулиративті энцефалопатиясы (тамырлы, дисметаболикалық) 1 тип», «дискуляторлық энцефалопатия» алтын стандартта ауру немесе жарақат ретінде көрсетілген. Клиникалық білім осы семантикалық қатынастарды дәл түсіндіру үшін қолданылуы керек. Клиникалық мәтіннің тағы бір ерекшелігі – үзінділердің жоғары жиілігі; яғни грамматикалық жағынан аяқталмаған сөйлемдер. Осылайша, қазақ тіліндегі клиникалық мәтіндерге семантикалық түзету нақты медициналық саланы ескере отырып, қосымша зерттеуді қажет етеді.

Бұл жұмыс «AP05132760» жобасы бойынша гранттық қаржыландыру негізінде Қазақстан Республикасы Білім және ғылым министрлігі Ғылым комитеті ШЖҚ РМК «Ақпараттық және есептеу технологиялары институтында» жасалды.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР

1. Pradhan SS, Ward WH, Hacioglu K, et al. Shallow semantic parsing using support vector machines / In: Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.-Boston, Massachusetts, USA: Association for Computational Linguistics July 21-26, 2004.-P.233–240.
2. Allen J. Natural Language Understanding. /2nd ed. Menlo Park, CA: Benjamin /Cummings.,-1995.
3. Surdeanu M, Harabagiu S, Williams J, et al. Using predicate-argument structures for information extraction. // In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics July 7-12, 2003.-P.8–15.
4. Akane Y, Yusuke M, Tomoko O, et al. Automatic construction of predicate- argument structure patterns for biomedical information extraction. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics July 22-23, 2006.-P.284–292.
5. Yakushiji A, Miyao Y, Tateisi Y, et al. Biomedical information extraction with predicate-argument structure patterns. In: Proceedings of the First International Symposium on Semantic Mining in

Biomedicine. Hinxton, Cambridge, UK: European Bioinformatics Institute April 10-13, 2005.-P.60–69.
6. Nguyen NTH, Miwa M, Tsuruoka Y, et al. Open information extraction from biomedical literature using predicate-argument structure patterns. // In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine. Zurich, Switzerland December 12-13, 2013.
7. Wattarujeekrit T, Shah PK, Collier N. PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics 200.-P.155.
8. Kogan Y, Collier N, Pakhomov S, et al. Towards semantic role labeling & IE in the medical literature. AMIA Symposium October 22-26, 2005.-P.410–414.
9. Shah PK, Bork P. LSAT: learning about alternative transcripts in MEDLINE. Bioinformatics 2006.-P.857–865.
10. Bethard S, Lu Z, Martin JH, et al. Semantic role labeling for protein transport predicates. BMC Bioinformatics 2008.-P.277.
11. Barnickel T, Weston J, Collobert R, et al. Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. PloS One 2009.-P.e6393.
12. Paek H, Kogan Y, Thomas P, et al. Shallow semantic parsing of randomized controlled trial reports. AMIA Symposium November 11-15, 2006.-P.604–608.
13. Albright D, Lanfranchi A, Fredriksen A, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc. 2013.-P.922–930.
14. Wang Y, Pakhomov S, Melton GB. Predicate argument structure frames for modeling information in operative notes. Stud Health Technol Inform. 2013.-P.783–787.
15. Meyers A, Reeves R, Macleod C, et al. Annotating noun argument structure for NomBank. In: Proceedings of the Language Resources and Evaluation Conference. Lisbon, Portugal: European Language Resources Association May 26-28, 2004.-P.803–806.
16. Meyers A, Reeves R, Macleod C, et al. The NomBank Project: an interim peport. In: Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation. Boston, Massachusetts, USA: Association for Computational Linguistics May 6, 2004.-P.24–31
17. Palmer M, Gildea D, Kingsbury P. The proposition bank: an annotated corpus of semantic roles. Comput Linguist. 2005.-P.71–106.
18. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003.-P.462–477.
19. Kilicoglu H, Shin D, Fiszman M, et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012.-P.3158–3160.
20. Sager N. Natural Language Information Processing. UK; Addison-Wesley,1981.
21. Chen ES, Hripcsak G, Xu H, et al. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. Jam Med Inform Assoc. 2008.-P.87–98.
22. Simpson MS, Demner-Fushman D. Biomedical text mining: a survey of recent progress. In: Aggarwal CC, Zhai C, eds. Mining Text Data. USA; Springer 2012.-P.465–517.
23. Schuler KK. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. ProQuest Paper AAI3179808, 2005.
24. Ruppenhofer J, Ellsworth M, Petruck MRL, et al. FrameNet II: extended the ory and practice. http://framenet.icsi.berkeley.edu/ Accessed 10 March 2014.
25. Levin B. English Verb Classes and Alternations: a Preliminary Investigation. -Chicago, USA: University of Chicago Press 1993.