6. Сапарбаев З. Кыргыз мифологиясы. Мектеп. 1980.
7. Сапарбаев З. Эпикалык чыгармалар, легендалар. Мектеп 1975.
8. Сагынбай Орозбаковдун вариантында айтылган.

References:

1. Bakchiev T. "Manaschilar" Bishkek. 2012.
2. Manas encyclopedia. Volume 1. Chief editor of the Kyrgyz encyclopedia. Bishkek. 1995.
3. Muratova S. "The Queen of Alai" BAITIK". 2016
4. Mirchi Eliade Shamanism: Archaic technical ecstasy. Per.s engl.-K.: "Sofia, 2000.
5. Roerich N. K. "Oh forever" - M.: Politizdat, 1991.
6. Saparbaev Z. Kyrgyz mythology. A school. 1980.
7. Saparbaev Z. Epic works, legends. School 1975.
8. Said in Sagynbay Orozbakov's version.

УДК 82(575.2)
DOI 10.33514/BK-1694-7711-2022-1(2)-27-31

Муканбетова Айкерим, Шаршембаев Бакыт
Кыргыз-Түрк Манас Университети, Компьютердик инженерия кафедрасы, магистрант,
Кыргыз-Түрк Манас Университети, Компьютердик инженерия кафедрасы, доцент
Муканбетова Айкерим, Шаршембаев Бакыт
Кыргызско-Турецкий университет Манас, кафедра компьютерной инженерии, магистрантка,
Кыргызско-Турецкий университет Манас, кафедра компьютерной инженерии, доцент
Mukanbetova Aikerim, Bakit Sharshembaev
Kyrgyz-Turkish University Manas, Department of Computer Engineering, graduate student,
Kyrgyz-Turkish University Manas, Department of Computer Engineering, Associate Professor

КЫРГЫЗ ТИЛИ ҮЧҮН МОРФОЛОГИЯЛЫК АНАЛИЗАТОРДУН МОДЕЛДЕРИ ЖАНА АЛГОРИТМДЕРИ
МОДЕЛИ И АЛГОРИТМЫ МОРФОЛОГИЧЕСКОГО АНАЛИЗА КЫРГЫЗСКОГО ЯЗЫКА
MODELS AND ALGORITHMS OF MORPHOLOGICAL ANALYSIS OF THE KYRGYZ LANGUAGE

Аннотация: Илимдин, техниканын жана маданияттын өнүгүшүнүн азыркы этабында коомдук өндүрүштун процессинде алдыңкы орунду ээлеген жана адамдын ишинин бардык чөйрөлөрүнө кирип жаткан информацияны иштеп чыгуу процесстерине артыкчылык берилип жатат. Адамзатка ынгайлуулукту түзүү үчүн маалыматты табигый тилде иштетүү каражаттарына көбүрөөк маани берилүүдө. Бүгүнкү күндө табигый тилди иштетүү (Natural Language Processing NLP) көптөгөн тармактарда колдонулат, анын ичинде үн жардамчылары, автоматтык текст котормолору жана текстти фильтрлөө. Негизги үч багыт болуп төмөнкүлөр саналат: кеп таануу (Speech Recognition), табигый тилди түшүнүү (Natural Language Understanding) жана жаратылыш тилинин мууну (Natural Language Generation). Көптөгөн ушул сыяктуу системаларда табигый тилди иштетүүнүн негизин морфологиялык анализ түзөт. Ал эми илимпоздор буга чейин кеңири таралган тилдер үчүн табигый тилди

иштетүүчү көптөгөн системаларды иштеп чыгышса да, кыргыз тили үчүн көлөмдүү базасы бар системалар жок жана бул иштин актуалдуулугун көрсөтүп турат. Бул изилдөөдө кыргыз тилинин жетиштүү базасы бар жаңы анализатору сунушталган. Колдонуучулук жана жеткиликтүүлүк жагынан окшош системаларда бир катар тесттер жүргүзүлдү жана алардын натыйжалары бул анализаторду жакшыртуу үчүн колдонулган.

Аннотация: В наши дни, когда развитие науки, техники и культуры происходит быстрыми темпами, все же первые позиции занимают процессы обработки информации, которые изо дня в день проникают во все сферы деятельности человека. Для создания удобства человечеству, все большее значение отдается средствам обработки информации на естественном языке. Сегодня обработка естественного языка (Natural Language Processing, NLP) применяется во многих сферах, в том числе в голосовых помощниках, автоматических переводах текста и фильтрации текста. Основными тремя направлениями являются: распознавание речи (Speech Recognition), понимание естественного языка (Natural Language Understanding) и генерация естественного языка (Natural Language Generation). Во многих таких системах основой обработки естественного языка является морфологический анализ. И хотя, для распространенных языков, учеными были уже разработаны множество систем обработки естественного языка, для кыргызского языка нет систем с достаточным объемом базы и это показывает актуальность задачи. В этом исследовании был предложен новый анализатор кыргызского языка с достаточной базой. Над схожими системами был проведен ряд тестов, как удобство в использовании и доступности и их результаты были использованы для улучшения анализатора.

Abstract: In our days, when the development of science, technology and culture is taking place at a rapid pace, nevertheless, the first positions are occupied by the processes of information processing, which day by day penetrate into all spheres of human activity. To create convenience for mankind, more and more importance is given to the means of processing information in natural language. Today, Natural Language Processing (NLP) [1] is used in many areas, including voice assistants, automatic text translations, and text filtering. The main three areas are: speech recognition (Speech Recognition), natural language understanding (Natural Language Understanding) and generation of natural language (Natural Language Generation). In many such systems, the basis of natural language processing is morphological analysis. And although scientists have already developed many natural language processing systems for common languages, there are no systems with a sufficient base for the Kyrgyz language, and this shows the relevance of the task. In this study, a new analyzer of the Kyrgyz language with a sufficient base was proposed. A number of tests were carried out on similar systems, both in terms of usability and accessibility, and their results were used to improve this analyzer.

Негизги сөздөр: табигый тилди иштетүү, морфологиялык анализ, корпус, Кыргыз тили.

Ключевые слова: обработка естественного языка, морфологический анализ, корпус, Кыргызский язык.

Keywords: natural language processing, morphological analysis, corpus, Kyrgyz language.

Морфологиялык анализатор – сөздүктөгү (тактап айтканда, лексикадагы) жеке сөздөрдү жана сөз формаларын салыштырып, сөздөрдүн грамматикалык белгилерин аныктоочу алгоритмдердин жыйындысы.

Синтаксистик талдоочу сүйлөмдү талдоодо сөздөрдүн морфологиялык анализинин жыйынтыгын активдүү колдонот, бирок тааныгычсыз иштей алат. Ошондой эле баштапкы текстти грамматикалык маалымат менен белгилөө талдоо эрежелерин түзүүнү бир топ жеңилдетет. [2]

Орус тилинин морфологиясын формалдаштыруунун ар кандай аракети ар кандай грамматикалык формалардын дал келиши сыяктуу көйгөйгө сөзсүз түрдө дуушар болот. Мисалы, “китептер” деген сөзгө карап, “эски китептер үстөлдүн үстүндө” деген номинативдик сөздүн көптүк түрүн, “ошол эски китептерди ыргытып сал” деген айыптоочту сүйлөп жатабызбы же жокпу, так айтууга болбойт. "китептин башталышы" генитивдик учурдун сингуляры. Таануучунун милдеттерине сөздүн грамматикалык формасын тааануунун бардык мүмкүн болгон варианттарын табуу, ошондой эле бул контекстте ачык жана так тыюу салынган тааануу варианттарын алып салуу кирет. Бул омонимияны алып салуу көйгөйүн бир нече өзүнчө тапшырмаларга бөлүүгө болот, алар өздөрүнүн адистештирилген алгоритмдери менен чечилет. Таануучунун көп убакыт бериле турган акыркы милдети белгисиз сөздөрдүн мүмкүн болуучу грамматикалык белгилерин аныктоо болуп саналат. Башкача айтканда, талданган сөз лексикадан табылбаса, анда тааныучу анын грамматикалык ролунун мүмкүн болгон варианттары жөнүндө болжолдоолорду жасоого тийиш.

Морфологиялык анализатордун маанилүү өзгөчөлүгү анын белгилүү бир тилге байланбаганында. Сөздүктүн түпнуска тексттеринде түрдүү эрежелер белгиленген жана морфологиялык анализдин жүрүшүн тигил же бул тилдин өзгөчөлүктөрүнө ийкемдүү өзгөртүүгө мүмкүндүк берет. Маселен, орус тили үчүн көптөгөн эрежелер жеке сөздөрдүн морфологиялык анализинин жыйынтыгын фильтрациялап, зат атоочтор менен сын атоочтордун грамматикалык саны, учуру боюнча келишүүсүн эске алат. Сөздүк кураштыруу учурунда атайын программа (сөздүк түзүүчү) эрежелердин тексттик көрүнүшүн окуйт жана аны сөздүк базасында сакталган ыңгайлуураак экилик сүрөттөлүшкө айландырат. Колдонмо программалары жана грамматикалык сөздүктүн компоненттери, зарыл болсо, эрежелерди эс тутумга жүктөйт жана аларды тааныу опцияларын жок кылуу же басуу үчүн колдонушат.

Макалада тексттин сандык мүнөздөмөлөрүн аныктоо үчүн колдонула турган морфологиялык текст анализаторлору каралып чыккан. Морфологиялык анализаторлорду колдонуунун өзгөчөлүктөрү баяндалып, артыкчылыктары жана кемчиликтери белгиленген.

Тексттин сандык мүнөздөмөлөрү төмөнкү милдеттерди чечүү үчүн баштапкы чекиттер болуп саналат: тексттин авторлугун аныктоо; тексттин жанрын жана стилин аныктоо, тексттердеги адистиктин тилин тандоо.

Тексттердин сандык мүнөздөмөлөрүн эсептөөнүн тактыгы бул маселелерди чечүүдөгү катага таасирин тийгизет. Тексттин сандык мүнөздөмөлөрүн эсептөө процесси эмгекти жана убакытты талап кылгандыктан автоматташтырылууга тийиш , бирок баштапкы сандык мүнөздөмөлөрдү аныктоо үчүн модулдар катары колдонулган морфологизаторлор өздөрүнүн катачылык пайызына ээ.

Mystem жана PHPMorphy морфологиялык анализаторлорунун үстүнөн салыштырма анализи жүргүзүлдү .[2][3]

Mystem – Яндекстен коммерциялык эмес колдонуу үчүн орус тилинин акысыз морфологиялык анализатору. Морфологиялык анализатор Си тилинде жазылган өз алдынча тиркеме катары иштейт.Программа морфологизация үчүн маалымат алынган текст файлдары менен же сөздөрдүн стандарттык киргизүү/чыгарылышы менен иштейт. Морфологиялык анализатор түпнуска сөздөрдүн бардык мүмкүн болгон формаларын көрсөтөт.

PHPMorphy PHP платформасында ишке ашырылган акысыз морфологиялык анализ китепканасы. PHPMorphy төмөнкү милдеттерди аткарууга мүмкүндүк берет:

- лемматизация (сөздүн нормалдуу формасын алуу);
- Сөздүн бардык түрлөрүн алуу;
- Сөзгө грамматикалык маалымат алуу (сөздүн мүчөсү, учур ж.б.);
- Берилген грамматикалык мүнөздөмөлөргө ылайык сөздүн формасын өзгөртүү;
- Берилген үлгү боюнча сөздүн формасын өзгөртүү.

Анализатор камтыган тилдер: орус, англис, немис , украин, эстон (ispell негизинде). Myspell сөздүгүн колдонуу менен башка тилдерди да кошууга болот.

Бул максатка жетүү үчүн төмөнкүдөй эксперимент жүргүзүлдү: ошол эле текст PHPmorphy жана Mystem морфологдорунун киргизүүсүнө берилди. Эксперименттин натыйжалары 1-таблицада көрсөтүлгөн.

Таблица 1. Тексттин сандык мүнөздөмөсү

Сөз түркүмдөрү	phpMorphy		MyStem	
	Ачык-айкын чечмелөө	Түшүнүксүз чечмелөө	Ачык-айкын чечмелөө	Түшүнүксүз чечмелөө
Этиш	8	4	9	5
Зат атооч	13	3	10	11
Сын атооч	8	2	6	3
Тактооч	0	1	0	1
Байламталар	0	3	0	3
Бөлүкчө	0	1	0	5
Атоочтук	2	2	2	2
Тууранды сөздөр	0	3	0	3
Сан атооч	1	0	1	0

Эксперименттин жыйынтыгын салыштырып талдоодо, сөз мүчөсүнүн автоматташтырылган аныктоосу менен MyStem морфологизаторду колдонууда түшүнүксүз чечмелөө көбүрөөк пайда болоорун көрсөттү жана төмөнкүдөй тыянак чыгарууга болот: сандык параметрлерди автоматташтырылган аныктоо үчүн сөздүн бөлүктөрүн аныктоочу модул катары PHPMorphy морфологизатору дагы жакшыраак жыйынтыктарды берет.

Бул этапта сөздүн грамматикалык касиеттерин тааануу үчүн лексикадагы маалымат колдонулат.

Лексика - сөздүк маалыматтар базасынын бир бөлүгү, анда сөздүк жазуулары жана парадигмалар сакталат. Башкача айтканда, ар бир сөз үчүн грамматикалык форма катары кайсы сөздүккө киреринаныктоого болот.

Орус тили үчүн грамматикалык формалардын тизмесинде кайталоосу бар 3 миллионго жакын форма бар. Албетте, мындай чоң тизмеден издөө түйшүктүү болуп калбашы үчүн өзгөчө көңүл бурууну талап кылат. Процедуралык API аркылуу лексика базасы менен иштеген колдонмо коду үчүн лексикон издөө алгоритмдеринин бардык деталдары толугу менен жашырылган. Ыкчам издөөнү камсыз кылган структуралар сөздүк

түзүүдө кыймылдаткычтын өзү тарабынан түзүлөт жана эч кандай конфигурацияны талап кылбайт.

Эгерде реляциялык маалымат базасы лексикондун репозиторийлери катары колдонулса, анда базанын стандарттык куралдары катары - таблицаны индекстөө колдонулат.

Дал ушул сыяктуу эле, бул макалада сунушталган системанын базасы, индекстелген таблица катары даярдалган. Ар бир сөз бир гана жолу кездешет, башкача айтканда кайталанбастык бар. Буга чейин морфологиялык анализ үчүн

Германия сайтына [4] жүктөлгөн кыргыз тилинин корпусу сунуш кылынган. Бул корпус 1 млндон ашык сөздөн турат. Ар бир сөзгө тегдер коюлуп чыккан, мисалы кайсы сөз түркүмү, кайсы жак, жеке же көптүк ж.б.у.с. Бирок бул корпуста жыштык эске алынган эмес, 1 млндон ашык болсо да, мисалы “деп” деген сөз 500 жолу кездешиши мүмкүн. Ошондуктан, эң аз 30 миңден турган индекстелген сөздүк даярдоо пландалган жана бул сан азыркы ушул тапта даяр. (план боюнча 70 миңге жакын сөздү индекстөө, учурда улануууда). Бул система кыргыз тилинин грамматикасындагы эрежелери [5] эске алынып түзүлгөн Ал мындан ары өнүктүрүүнү талап кылат.

Колдонулган адабияттар:

1. Natural Language Processing with Pythonhttps://www.nltk.org/book_1ed/
2. Морфологиялык анализатор жана Part-Of-Speech Tagger http://www.solarix.ru/for_developers/docs/morphology_analyzer.shtml
3. Логинов, Е.В. Маалымат издөө системаларынын морфологиялык анализаторлорун изилдөө / Э.В. Логинов, А.А. Рыбанов. — Текст : электронный // NovaInfo, 2015. — № 32. — URL: https://novainfo.ru/article/3394 (дата обращения: 23.05.2022).
4. CQPweb https://corpora.clarin-d.uni-saarland.de/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=kyrgyz_20190418&why=6
5. Кыргыз тилинин грамматикасы https://kyrgyzchasy.blogspot.com/p/1_20.html

References:

1. Natural Language Processing with Pythonhttps://www.nltk.org/book_1ed/
2. Morphological analyzer and Part-Of-Speech Tagger http://www.solarix.ru/for_developers/docs/morphology_analyzer.shtml
3. Loginov, E.V. Study of morphological analyzers of information retrieval systems / E.V. Loginov, A.A. Rybanov. - Text: electronic // NovaInfo, 2015. - No. 32. - URL: https://novainfo.ru/article/3394 (date of access: 05/23/2022).
4. CQPweb https://corpora.clarin-d.uni-saarland.de/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=kyrgyz_20190418&why=6
5. Grammar of the Kyrgyz language https://kyrgyzchasy.blogspot.com/p/1_20.html