УДК: 81'322

Касиева А. А., ф. и. к., доц., aida.kasieva@manas.edu.kg
Байсалова А. А., 2-курс студенти, 1901.10014@manas.edu.kg
«Манас» КТУ

ТАБИГЫЙ ТИЛДИ ИШТЕТҮҮДӨ СӨЗ ТҮРКҮМДҮК ЭНТЕКТЕШТИРҮҮ

Бул макаланын негизги максаты – табигый тилди иштетүү процессинде сөз түркүмдүк энтектештирүү тууралуу маалымат берүү. Сөз түркүмдүк энтектештирүүнүн негизги үч түрү жана алардын ортосундагы негизги айырмачылыктар кылдаттык менен талкууланып, кыргыз жана англис тилдеринде мисалдар менен бекемделет. Киришүүдө сөз түркүмдүк энтектештирүү эмне экендиги, андан мурда орун алган процесстер жана эң негизгиси эмне үчүн сөз түркүмдүк энтектештирүү маанилүү болуп саналганы тууралуу маалымат берилет. Ал эми экинчи бөлүмдө сөз түркүмдүк энтектештирүүнүн негизги үч түрү: эрежеге таянган энтектештирүү, стохастикалык энтектештирүү жана трансформациялык энтектештирүү тууралуу сөз болот. Андан кийинки бөлүмдө сөз түркүмдүк энтектештирүү жүргүзүп жатканда пайда боло турган көйгөйлөр (алардын ичинен эң негизгиси – кош маанилүүлүк (ambiguity)) талкууланат.

Өзөктүү сөздөр: *кыргыз тилинин корпусу, ТТИ (Табигый тил иштетүү), сөз түркүмдүк энтектештирүү (POS Tagging), эрежеге таянган энтектештирүү, стохастикалык энтектештирүү, трансформациялык энтектештирүү, кош маанилүүлүк (ambiguity).*

Касиева А.А., к. ф. н., доцент, aida.kasieva@manas.edu.kg
Байсалова А.А., студент 2го курса, 1901.10014@manas.edu.kg
КТУ «Манас»

ЧАСТЕРЕЧНАЯ РАЗМЕТКА В ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА

Основная цель этой статьи предоставить информацию о частях речевых тегов в НЛП. В исследовании изучаются различия между тремя подходами POS Tagger, поддерживая их примерами на кыргызском и английском языках. Во вводной части мы объясняем, что такое POS-тегирование, а также процессы, предшествующие процедуре тегирования, и объясняем, почему тегирование POS-тегов считается важным этапом / процессом в NLP. Во второй части мы обсуждаем три основных типа тегов POS: теги на основе правил, стохастические теги и теги трансформации. В следующей части рассматриваются проблемы, которые могут возникнуть во время маркировки POS, что может привести к неправильной маркировке, то есть к устранению неоднозначности.

Ключевые слова*: корпус кыргызского языка, НЛП (обработка естественного языка), тегирование POS, тегирование на основе правил, стохастическое тегирование, трансформационное тегирование, двусмысленность (неясность).*

Aida Kasieva candidate of Philology Sciences, Associate Prof.,
aida.kasieva@manas.edu.kg
2nd grade student Baysalova A. 1901.10014@manas.edu.kg
Kyrgyz-Turkish 'Manas' University

PARTS OF SPEECH TAGGING IN NATURAL LANGUAGE PROCESSING

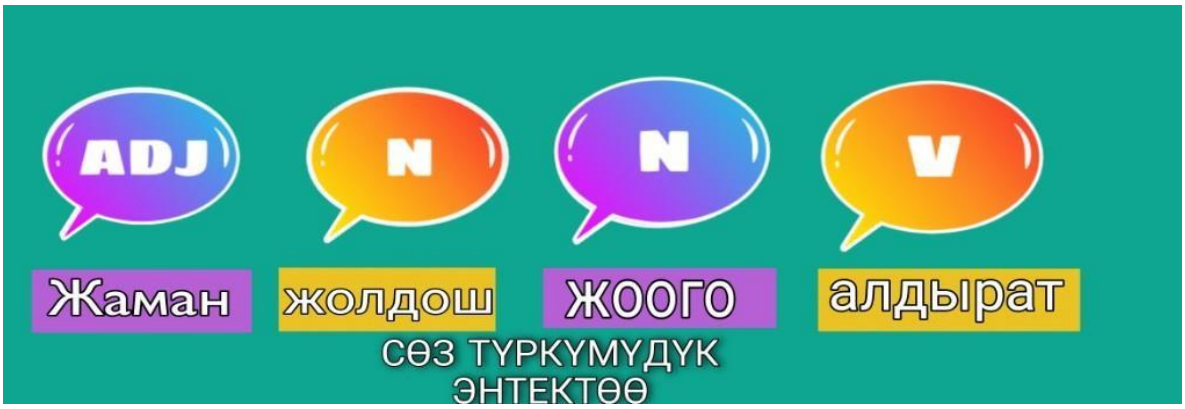
The main purpose of this paper is to provide information on Parts of Speech Tagging in NLP. The study examines the differences between three approaches of POS Tagger by supporting them with examples in the Kyrgyz and English languages. In the introductory part, we explain what POS Tagging

is along with the processes that precede tagging procedure through explaining why POS tagging is considered to be an important stage/process in NLP. In the second part we discuss three main types of POS tagging: rule-based tagging, stochastic tagging and transformation tagging. The next part deals with the issues that may occur during POS tagging, which can lead to mis-tagging i.e. disambiguation.

Keywords: *Kyrgyz corpus, NLP (Natural Language Processing), POS tagging, rule-based tagging, stochastic tagging, transformation tagging. ambiguity.*

Киришүү

Корпустук лингвистикада сөз түркүмдүк энтектештирүү (POS Tagging, грамматикалык энтектештирүү) – текстти автоматтык түрдө иштетүүнүн башкы этабы, анын милдети – тексттеги (же тилдик корпустагы) сөздөрдү маанисине жана контекстине жараша, грамматикалык мүнөздөмөлөрүн эске алуу менен, кайсы сөз түркүмүнө тиешелүү экендигин аныктап, ылайыктуу эн(tag) коюу [https://en.m.wikipedia.org/wiki/Part-of-speech_tagging].



Сөз түркүмдүк энтектештирүү табигый тилди иштетүүдөгү эң алгачкы жана жөнөкөй кадамдардын бири болуп саналат. Ал – семантикалык анализдөөдөн (semantic analysis) же синтаксистик парсингди (syntactic parsing) аткаруудан мурун, алдын ала иштелип чыгуучу абдан маанилүү процесс. Сөз түркүмдүк энтектештирүү өз алдынча кандайдыр бир конкреттүү ТТИ көйгөйлөрүн чечпесе да, корпустук лингвистика, текстти сөзгө айландыруу (text to speech conversion), айрым сөздөрдөгү кош маанилүүлүктү жоюу (word sense disambiguation), машина котормосу (machine translation) сыяктуу жогорку деңгээлдеги ТТИ колдонмолорун түзүүдө келип чыга турган тоскоолдуктардын алдын алууга жардам берет. Сөз түркүмдүк энтектештирүү, сөздөн уңгуну бөлуп көрсөтө турган колдонмону, б. а., лемматайзерди (lemmatizer) түзүүдө да өтө маанилүү.

Жогоруда айтылгандай, ТТИдеги ар бир процесс сөз түркүмдүк энтектештирүүнүн канчалык деңгээлде туура жана так аткарылгандыгынан көз каранды. Сөз түркүмдүк энтектештирүү – биз ойлогондон да кыйла татаалыраак процесс. Мындагы эң чоң көйгөйлөрдүн бири болуп кош маанилүүлүк (ambiguity) эсептелет. Англис тилиндеги кээ бир көп таралган сөздөр бир нече маанилерге ээ. Ошондуктан кээ бир учурда бир сөз бир канча эндер менен белгиленет. Ал эми сөз түркүмдүк энтектештирүүнүн негизги функциясын ошол сөздөрдүн кандай контекстте колдонулгандыгына жараша так энтектештирип, б. а., эн коюп, пайда болгон кош маанилүүлүктү жоюу түзөт. Мисалы, англис тилиндеги «shot» сөзү зат атооч да, этиш да болушу мүмкүн.

Энтектештирүүдөн мурда кандай процесстер орун алат?

Иш жүзүндө сөз түркүмдүк энтектештиргич (POS Tagger) киргизилген маалыматты (input = сүйлөм, фраза) энтектештирерден мурда, бир нече башка процесстер иштелип чыгат. Мисалы, сөз түркүмдүк энтектештиргич сөздөрдүн ырааттуулугун билиши үчүн токенизация (tokenisation) аркылуу биз киргизген маалымат = инпут сүйлөмдөргө, сөздөргө (токендерге) бөлүнөт. Андан тышкары, стоп сөздөрү, пунктуация сыяктуу элементтер жок кылынат же түшүрүлөт. Ошондой эле лемматизациялоо (lemmatization) ж.б.у.с. процесстер сөз түркүмдүк энтектештирүүдөн мурда жүргүзүлөт.

Бизге белгилүү болгон эндердин (tag) жыйындысы **эн топтому (tagset)** деп аталат. Браун корпусунун тегдеринен мисалдар: көптүк сандагы зат атооч үчүн **NNS**, өткөн чактагы этиш үчүн **VBD**, же сын атооч үчүн **JJ** ж.б.у.с. Эндердин топтому тыныш белгилерин да камтышы мүмкүн. Ал эми Turkic Lexicon Apertiumда жандуу зат үчүн **aa**, жансыз зат үчүн **nn**, экинчи жактын таандык мүчөсү үчүн **px2sg** тектери ж.б.у.с колдонулат. Азыркы учурдагы эң белгилүү корпустар: 87-tag Brown корпусу, 45-tag Penn Treebank корпусу, 61-tag C5 корпусу, же 146-tag C7 корпусу.

3 Сөз түркүмдүк энтектештиргичтин негизги түрлөрү кайсылар?

1. **Rule-Based tagger (**Эрежеге таянган энтектештиргич). Сөздүктө ар бир сөз үчүн мүмкүн болгон тегдер камтылат. Rule-Based tagger аты айтып тургандай, лингвистикалык эрежелерге таянуу менен кош маанилүүлүктү жоюуга жана белгисиз же түшүнүксүз сөзгө эн коюуга багытталат. Ошол эрежелерден бир мисал келтирсек: «Эгерде белгисиз же түшүнүксүз сөздүн, б. а., **Хтин** алдында аныктагыч (determiner) «the» болуп/туруп, андан кийин зат атооч келсе, аны сын атооч деп белгилеңиз». Анда эреженин формуласы төмөнкүдөй: **«DT X N».** Мисалы: «The **famous(adj)** writer» (белгилүү жазуучу). Бул эреже, албетте, англис тилиндеги сөздөрдү энтектөөдө колдонулат. Көрүнүп тургандай, бул жерде энтектештирүү үчүн контекст жана сүйлөмдөгү сөз орду абдан маанилүү. Бирок бул метод автоматтык эмес, б. а., энтектештирүү кол менен жасалат жана көп убакытты талап кылат.

2. **Statistic Tagger/Stochastic Tagger (**статистикалык/ стохастикалык энтектештиргич). Стохастикалык ыкма жыштыкты, ыктымалдуулукту же статистиканы камтыйт. Эң жөнөкөй стохастикалык ыкма белгилүү бир сөздү **(Y)** энтектештирүү үчүн бир корпустан ошол сөз **(Y)** үчүн эң көп (жыш) колдонулган тэгди аныктайт да, ушул маалыматты башка корпуста **Y** сөзүн энтектештирүү үчүн колдонот. Жашыруун Марков Модели (Hidden Markov Model) – стохастикалык энтектештирүүдө колдонулган ыкмалардын бири. Бирок айта кетчү нерсе, стохастикалык энтектештиргич тилдин грамматикасына туура келбеген тегдердин ырааттуулугун да колдонушу мүмкүн. Энтектештирүү кол менен эмес, автоматтык түрдө жасалат.

3. **Transformation-based Tagger** Rule-Based (эрежеге таянган) менен стохастикалык теггердин комбинациясы. Демек, бул энтектештиргич лингвистикалык эрежелерге да, статистикалык маалыматтарга да таянуу менен, ар бир сөзгө эн коёт [https://devopedia.org/part-of-speech-tagging].

4. Кыргыз жана англис тилдеринде сөз түркүмүк энтектештирүү учурунда пайда боло турган кыйынчылыктар

Эмнеге жогоруда белгиленген Rule-Based taggerде эскертилген эрежени кыргыз тилине ыңгайлаштырып, кыргызча сөздөрдү энтектөөдө колдонсок болбойт? Албетте, ар бир тилде эле корпус үчүн сөз энтектештирүү – абдан оор процесс. Бирок, өзгөчө, кыргыз

тилиндеги сөздөрдү энтектештирүү көп мээнетти талап кылат. Мисалы, сөз түркүмдүк энтектештирүү үчүн корпуска киргизилген алгачкы чыгармалар: «Манас» эпосунун 3 томунда, «Сынган кылычта» ж.б.у.с. чыгармаларда биз маанисин билбеген тарыхый сөздөр жана колдонуудан таптакыр чыгып калган архаизм сөздөр көп кездешет. Ал эми кыргыз тилинде англис тилине салыштырмалуу түшүндүрмө сөздүктөрдү жөнөкөй эле интернет булактарынан табуу өтө кыйын. Ошондуктан сөз түркүмдүк энтектештирүү процессин аткарып жаткан адам ошол белгисиз сөзгө кайсы энди ыйгаруу керектигин билбей, такалып калышы мүмкүн. Демек, жогоруда сөз болгон Rule-Based taggerде келтирилген эрежени кыргыз тилине ыңгайлаштырсак, биздин ишибиз жеңилдемек. Албетте, ал эреже кыргыз тилин бир аз өзгөрүүгө дуушар кылышы да мүмкүн. Мисалы, эрежеде айтылган аныктагыч (determiner) же артикль кыргыз тилинде дээрлик кездешпегендиктен, аны эрежеден алып салабыз. Демек, кыргыз тилиндеги сөздөрдү энтектештирүү үчүн: «Эгерде белгисиз Х сөзү сүйлөмдүн башында же ортосунан орун алып, аны зат атооч коштосо, ал белгисиз х сөздү сын атооч деп белгилешибиз керек. Формула: **X N = кызыл алма, кооз көйнөк, даамдуу тамак** ж.б. Ал эми азыр корпустан алынган «**тыталама»** сөзүн карап көрөлү. Сиз бул сөзгө кандай эн ыйгарат элеңиз?

*Көчтүн изин кууп бир аз жүрүп , алды жагы зуулдаган **тыталама** жардын боорунан өткөн жолго түшүштү эле, бир кыпчылга салынган жер көпүрө бар экен, эрегишчилер бузуп кетишиптир. («Сынган кылыч»)*

Эреже: «Эгерде белгисиз Х сөзү сүйлөмдүн башында же ортосунан орун алып, аны зат атооч коштосо, ал белгисиз х сөздү сын атооч деп белгилеңиз».

Хтен кийинки **«жардын»** сөзү – зат атооч. Демек, белгисиз сөз – «**тыталама» <adj>.**

Көчтүн_n_gen изин_n_px3sp_acc кууп_v_tv_prc_perf бир_num аз_adj жүрүп_v_iv_prc_perf, cm алды_n_acc жагы_n_px3sp_nom зуулдаган_adj,_cm **тыталама_adj,_**cm **жардын_n**_gen боорунан_n_px3sp_abl өткөн_adj_subst_nom жолго_n_dat түшүштү_v_iv_ger_pres_acc эле_postadv,_cm бир_num кыпчылга_n_dat салынган_v_iv_past_p3_pl жер_v_tv_gpr_fut көпүрө_n_nom бар_adj экен_cop_ger_past_nom.

Дагы да кайталап эскерте кетчү нерсе, сөз түркүмдүк энтектештирүүдө контекст жана сүйлөмдөгү сөз орду абдан маанилүү.

1. Бир эле сөз 1 (бир) же андан ашык тегге ээ (көп маанилүү сөздөр):

Мисалы, англис тилинде:

«The **back** door» (Арткы эшик) – Adjective - сын атооч (**JJ)**

«On my **back**.» (Менин аркамда) - Noun (**NN)**

«Promised to **back** the bill.» Verb (**VB)**

Кыргыз тилинде:

«Илгери көчмөндөр ***ашты*** кашык же аш айры колдонбостон, бир гана колу менен жешкен» (Зат атооч).

Ашты$<n><sg><acc>

«Биздин план төмөнкү системанын негизинде ишке ашты» (Этиш)

Ашты$<v><tv><pass><ifi>;

2. Сөздүктү карап эн (тег) коюу дайыма эле туура боло бербейт:

а. «**Акылдуу**»

«Анын ушунчалык акылдуу экендигине таң калып, баа бердим.» Бул жерде

«акылдуу» – сын атооч. **Акылдуу $<adj>**

Тигил «акылдуу» менен эч сүйлөшкүм келбейт.

Бул мисалда «акылдуу» сөзү жогорудагыдай адамдын кандайдыр бир сапатын көрсөтүп, сын атооч болбой, адамды өзүн, б. а., затташкан сын атооч болуп калды.

Акылдуу$<adj_subst>;

б. Кыргыз тилинин корпусунан алынган мисалдар:

«**Сакалчан**»

Жылаңайлак, жылаң баш, **Ак сакалчан дубана** Чекесинен үч сылап,

Дубана туруп муну айтат: «Кейибегин, сен катын, Кайгырган капаң жазылды, Келип колуңа шер тийди, Эми көөнүң ачылды.(*«Манас» эпосу).*

Жылаңайлак_adj,_cm жылаң_adj баш_n_nom,_cm **Ак_adj сакалчан_adj дубана_n_aa_hu_sg_nom,**_cm Чекесинен_n_px3sp_abl үч_num сылап_v_tv_prc_perf,_cm Дубана_n_aa_hu_sg_nom,_cm туруп_v_iv_prc_perf муну_prn_dem_acc айтат_v_tv_prc_irre

— Төрө, — деди акырын, — мен ыраазымын, төрө, кудай жалгагыр, жанагы **кара сакалчан** менен эле сөзүңөрдү бүтөрө бергиле, мен ыраазымын. (*«Сынган кылыч»)*

Төрө_n_nom,_cm —_guio деди_v_tv_ifi_p3_sg акырын_adj,_cm —_guio мен_prn_pers_p1_sg_nom ыраазымын_adj,_cm төрө_n_nom,_cm кудай_n_nom жалгагыр_unknown,_cm жанагы_adv_attr **кара_adj сакалчан_adj_subst**,_cm менен_cnjcoo эле_postadv сөзүңөрдү_n_px2pl_acc бүтүрө_v_iv_prc_impf,_cm бергиле_v_tv_opt_p2_frm_pl,_cm мен_prn_pers_p1_sg_nom ыраазымын_adj [https://corpora.clarin-d.unihttps://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/saarland.de/cqpweb/kyrgyz_20190418/].

Мисал: англис тилинде «call» сөзүнүн 41 мааниси бар (чалуу, атоо, чакыруу, ж.б.у.с).

Since he did not hear the call and missed it, he decided to call back. (Сөзмө сөз котормо*: Ал чалууну укпай калып, өткөрүп жибергендиктен, кайра чалууну чечти.*)

Nltk.pos_tag(text)[('since'),('he',‘PRP’),('have',‘VBP’),('not'),('heard','VBD'),('the','D

T'),(**'call','NN'**),('and','IN'),('missed','VBD),('it','PR'),('he','PRP),('decided','VBD),('to', 'TO'), (**'call','VB'**),('back','RB')]

Бул жерде 1- «call» -NN(noun) зат атооч, экинчи –«call»-VB (verb,base form) этиш.

Корутунду

Жыйынтыктап алсак, сөз түркүмдүк энтектештирүү (POS Tagging) тилдик корпустагы сөздөрдү маанисине жана контекстине жараша эң туура болгон энди тандап ага ыйгаруу болуп саналат. Сөз түркүмдүк энтектештирүү өз алдынча кандайдыр бир конкреттүү ТТИ көйгөйлөрүн чечпесе да, кийинчерээк башка ТТИ процедураларында пайда богон көптөгөн көйгөйлөрдү жөнөкөйлөтөт. Жогоруда берилген мисалдардагы «акылдуу» жана «сакалчан» сөздөр катышкан сүйлөмдөрдө байкалгандай, бир сөз контекстине жараша зат атооч же сын атоочтун да функциясын да аткарарына күбө болдук. Эрежеге таянган сөз түркүмдүк энтектештиргич тарабынан колдонулган эрежени кыргыз тилине ылайыкташтырдык жана «тыталама» сөзүнүн мисалында көрсөттүк. Жыйынтыктап айтканда, кыргыз тили үчүн эң ылайыктуу энтектештиргич катары Rule-Based tagger экенине ынандык.

Адабияттар жана шилтемелер:

1. A new Kyrgyz corpus: sampling, compilation, annotation. Aida Kasieva, Jörg Knappen, Stefan Fischer, Elke Teich. 1. Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan, 2 Universität des Saarlandes, Saarbrücken. 42. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS). 2020, März.

2. https://dgfs.de/de/cl/postersessions.html.

3. https://www.zfs.uni-hamburg.de/dgfs2020/programm/abstracts/dgfs2020-clp-kasieva.pdf

4. Частеречные разметки для нового корпуса кыргызского языка (инструментарий Turkic Lexicon Apertium) // «Вестник КРСУ». -2020. -Том 20. -№6. -Стр. 67-72.

5. http://vestnik.krsu.edu.kg/archive/154/6520

6. https://www.zfs.uni-hamburg.de/dgfs2020/programm/abstracts/dgfs2020-clphttps://www.zfs.uni-hamburg.de/dgfs2020/programm/abstracts/dgfs2020-clp-kasieva.pdfkasieva.pdf

7. https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/

8. https://devopedia.org/part-of-speech-tagging

9. https://en.m.wikipedia.org/wiki/Part-of-speech_tagging